Using LLMs in Bioinformatics

Scott A. Handley, PhD
Lord of the Moleculos
Washington University School of Medicine
The Edison Family Center of Genome Sciences & Systems Biology

19-January, 2026 | Workshop on Genomics | Cesky Krumlov

Who am I?

Who am I?

- Father

Who am I?

- Father
- Workshops

CENTRAL DEPARTMENT OF PHYSICS
T. U., KIRTIPUR
NIH NIAID
Washington University in St. Louis School of Medicine
La Jolla Institute
CAPACITY BUILDING IN BIOMEDICAL RESEARCH
Central Department of Biotechnology, Kirtipur
October 11th to 18th , 2022.
PBMC Analysis
NGS Library Preparation
REDCap
Bioinformatics
KATHMANDU

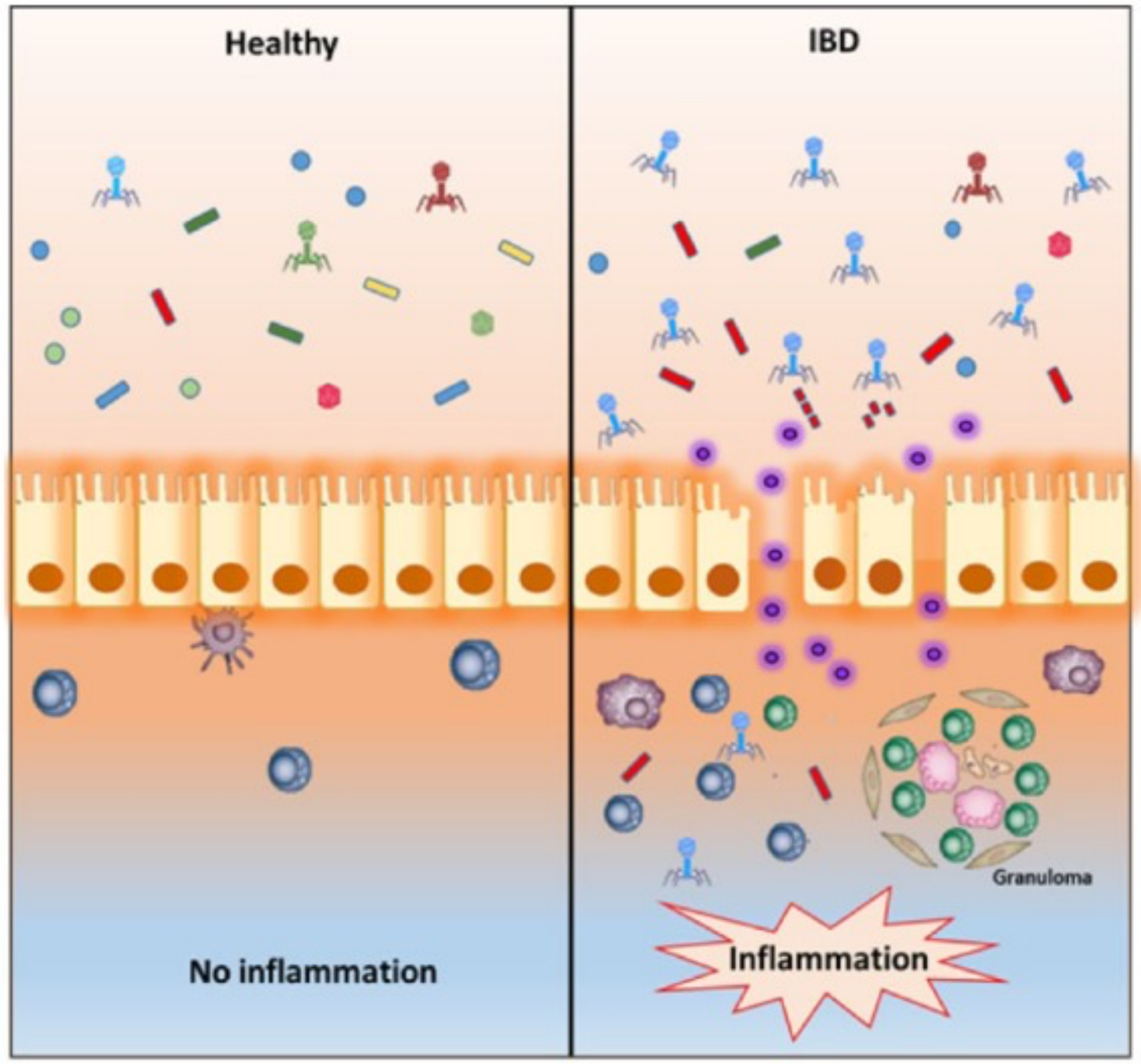
Missouri State UNIVERSITY

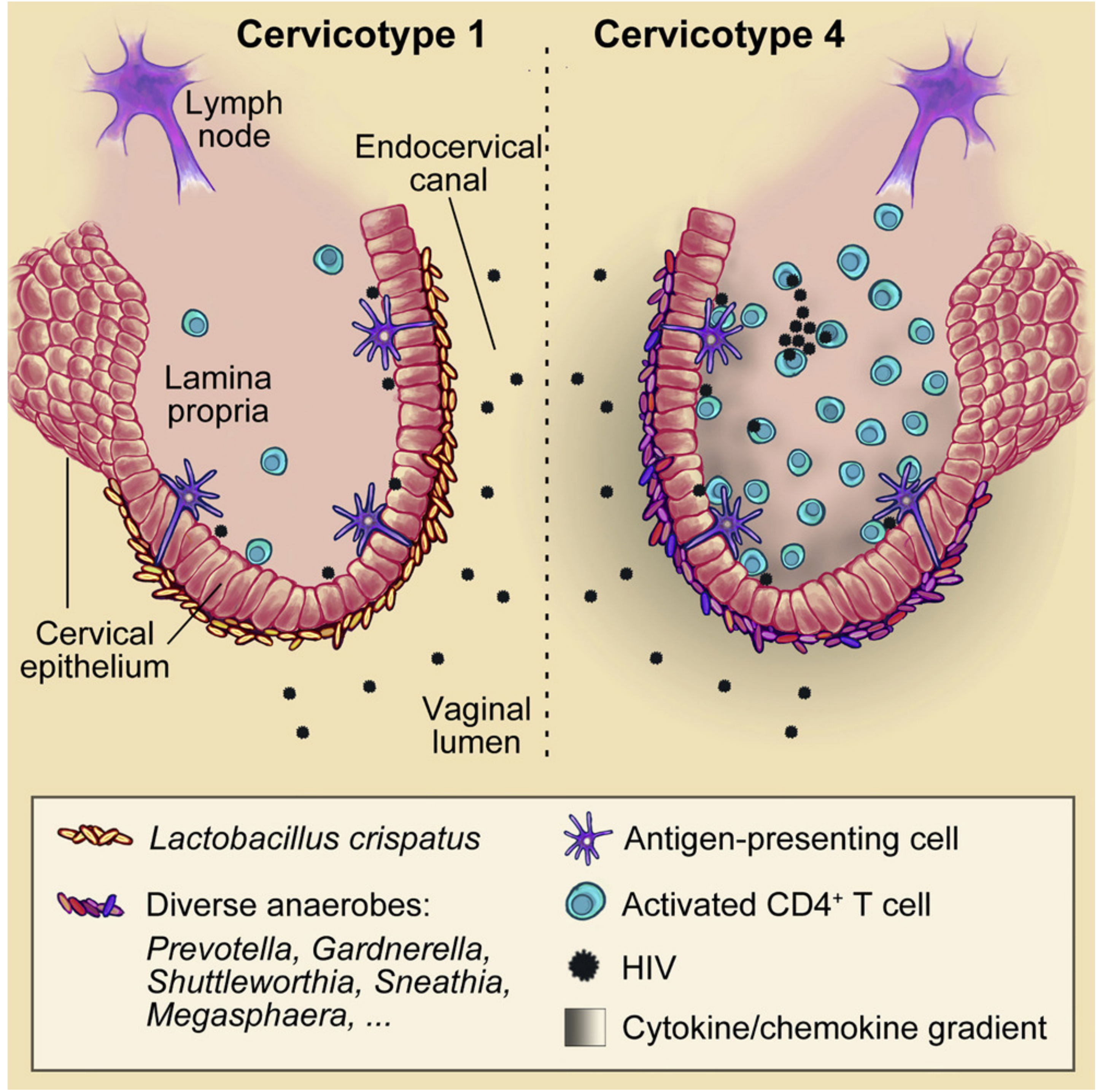
UPPSALA UNIVERSITET

WASHINGTON UNIVERSITY IN ST. LOUIS · 1853 ·

Who am I?

- Father
- Workshops
- Professor





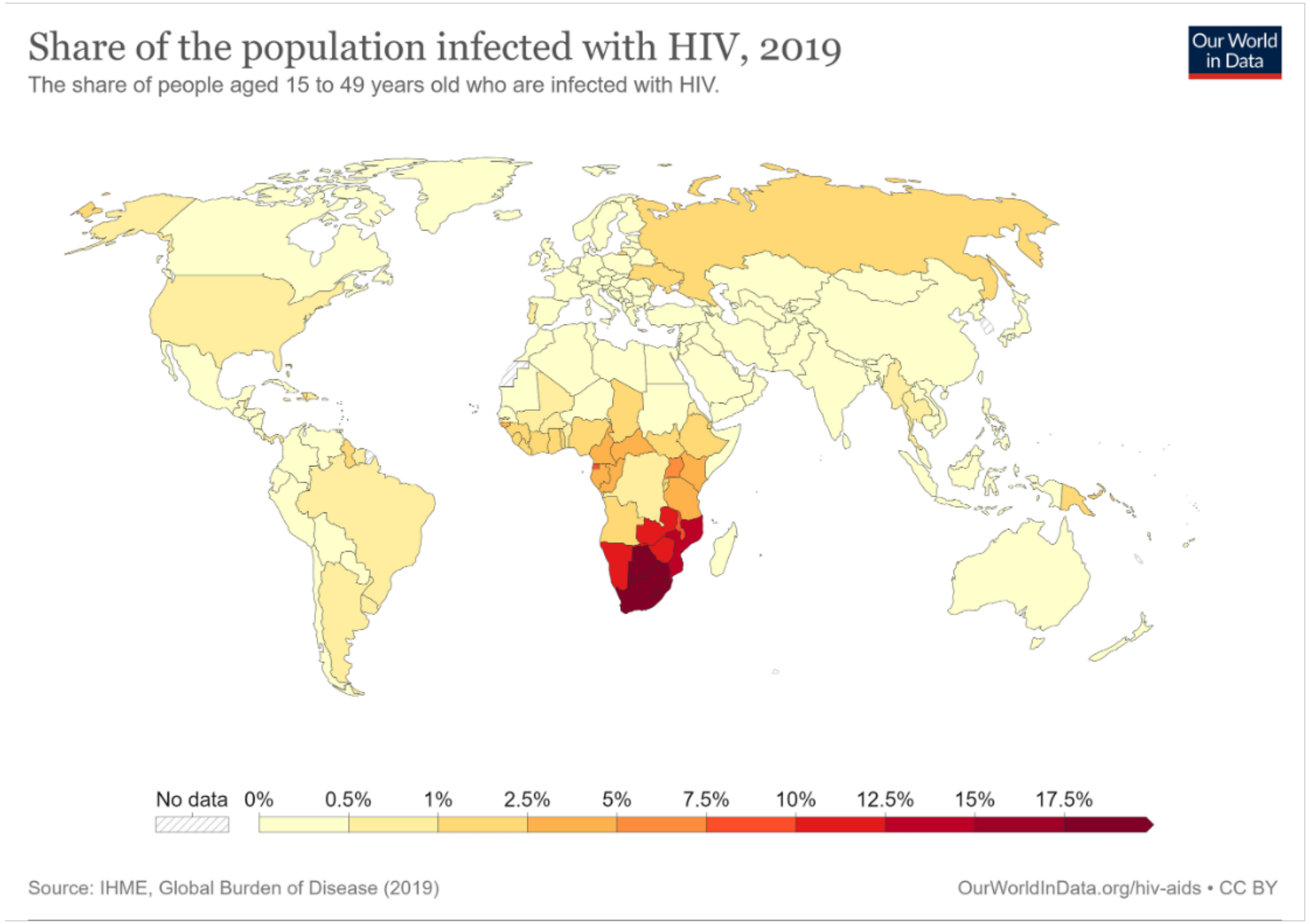
Novel Virus Discovery | Computational Tool Development | Agentic Research

FRESH Cohort

FRESH: Females Rising through Education, Support and Health
Primary Collaborator: Doug Kwon, MD, PhD, Ragon Institute, Boston, MA

HIV and the FGT

- HIV infection has led to 40 million deaths and left nearly 39 million living with disease in the past four decades
- Of the 1.2 million new infections each year almost 70% occur in sub-Saharan Africa
- Most new infections occur in young women of reproductive age
- Heterosexual intercourse is responsible for the majority of global HIV transmission
  - Women are twice as likely
- Mucosal tissues in the female genital tract represent the frontline of transmission
- Viral replication occurs within tissue resident CD4+ T cells (HIV target cells) in the mucosa of the cervix and vagina for 5-7 days prior to systemic dissemination
- Offers a “window of opportunity” to prevent systemic viral dissemination



Share of the population infected with HIV, 2019

The share of people aged 15 to 49 years old who are infected with HIV.

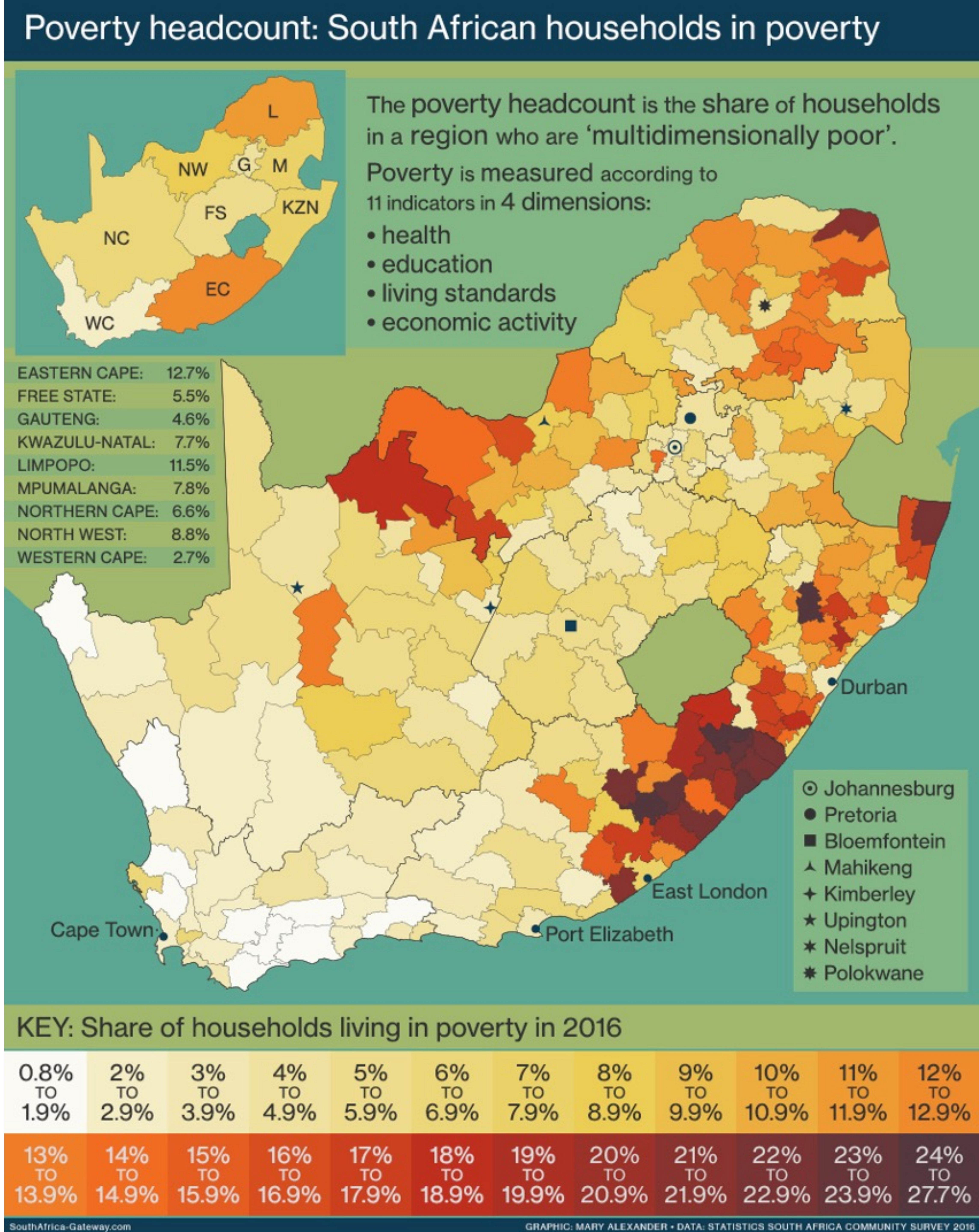
Our World in Data

Source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/hiv-aids • CC BY

FRESH Cohort

- Began recruitment in 2012
- Young women at high risk of HIV infection in KwaZulu-Natal (KZN) township in South Africa
- Enrolled in a 9-month program
- Tested for HIV RNA twice per week (finger stick)
- The goal is to detect HIV infection at the earliest possible time point (Fiebig stage I)
- Integrated with empowerment and life skills curriculum
- We have collected 3,600 longitudinal samples (cervical vaginal lavages, blood, cytobrushes, etc.) from 1,200 participants
- 120 have become HIV+



Poverty headcount: South African households in poverty

The poverty headcount is the share of households in a region who are 'multidimensionally poor'.

Poverty is measured according to 11 indicators in 4 dimensions:

- health
- education
- living standards
- economic activity

EASTERN CAPE:	12.7%
FREE STATE:	5.5%
GAUTENG:	4.6%
KWAZULU-NATAL:	7.7%
LIMPOPO:	11.5%
MPUMALANGA:	7.8%
NORTHERN CAPE:	6.6%
NORTH WEST:	8.8%
WESTERN CAPE:	2.7%

⊙ Johannesburg
● Pretoria
■ Bloemfontein
⅄ Mahikeng
✦ Kimberley
★ Upington
✶ Nelspruit
✸ Polokwane

KEY: Share of households living in poverty in 2016

0.8% TO 1.9%	2% TO 2.9%	3% TO 3.9%	4% TO 4.9%	5% TO 5.9%	6% TO 6.9%	7% TO 7.9%	8% TO 8.9%	9% TO 9.9%	10% TO 10.9%	11% TO 11.9%	12% TO 12.9%
13% TO 13.9%	14% TO 14.9%	15% TO 15.9%	16% TO 16.9%	17% TO 17.9%	18% TO 18.9%	19% TO 19.9%	20% TO 20.9%	21% TO 21.9%	22% TO 22.9%	23% TO 23.9%	24% TO 27.7%

SouthAfrica-Gateway.com

GRAPHIC: MARY ALEXANDER • DATA: STATISTICS SOUTH AFRICA COMMUNITY SURVEY 2016

Celebrating the Real Heroes
Many challenges over the years …

2021: Civil Unrest and Looting

FRESH severely affected and had to close for 12 weeks.

FRESH
FEMALES RISING THROUGH EDUCATION, SUPPORT AND HEALTH
Shop 1066 • Umlazi Mega City • Umlazi 4066
HEALTH SYSTEMS TRUST
Ragon Institute
HPP

Immediately post-looting

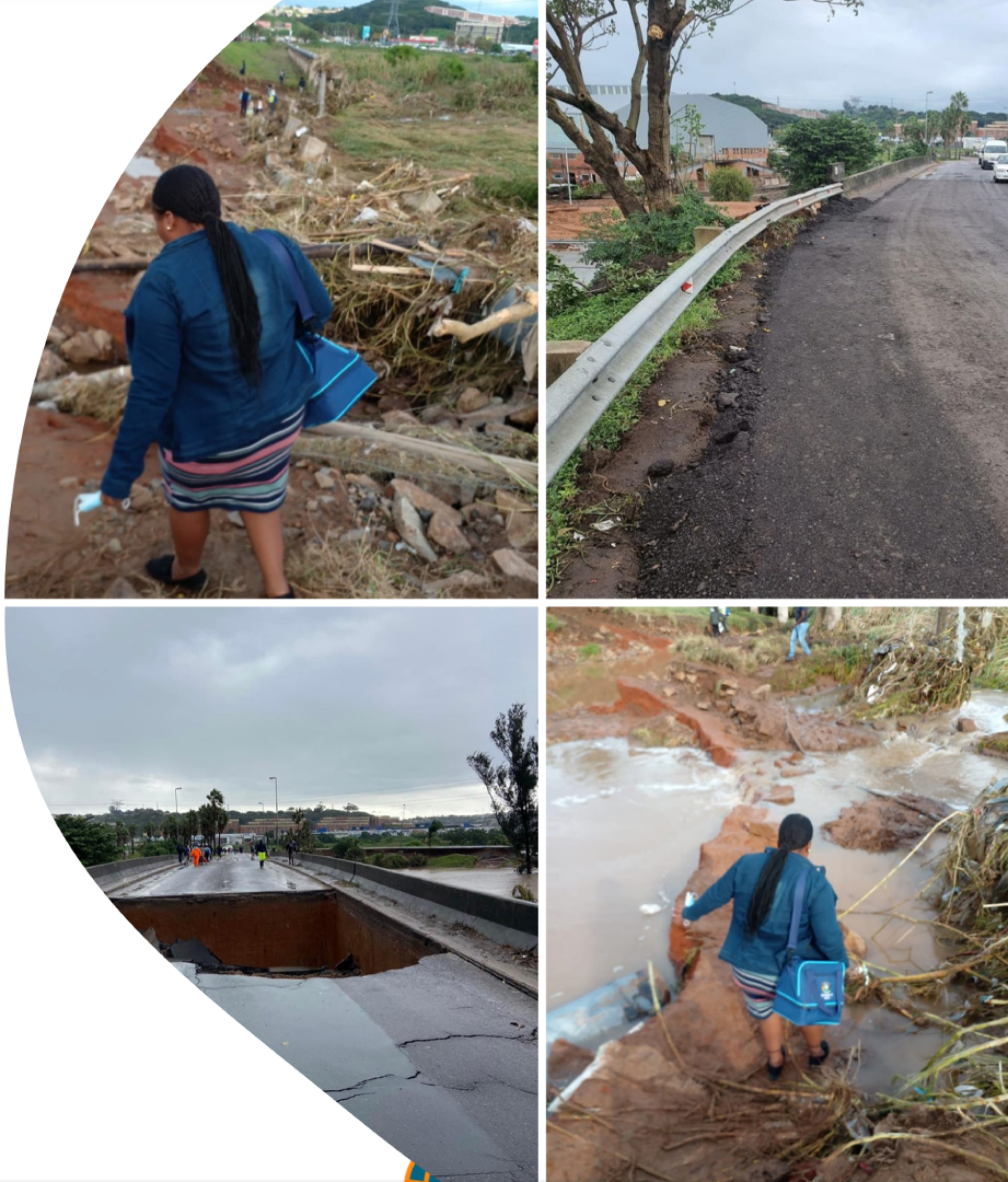
FRESH moved to a park home at the nearby hospital – continued to see acutes for sampling and ART dispensing. Zero ART default during the shutdown.

April 10, 2022
Catastrophic
Flooding in KZN

Roads to FRESH inaccessible. The team drove as close as they could and then walked to site.

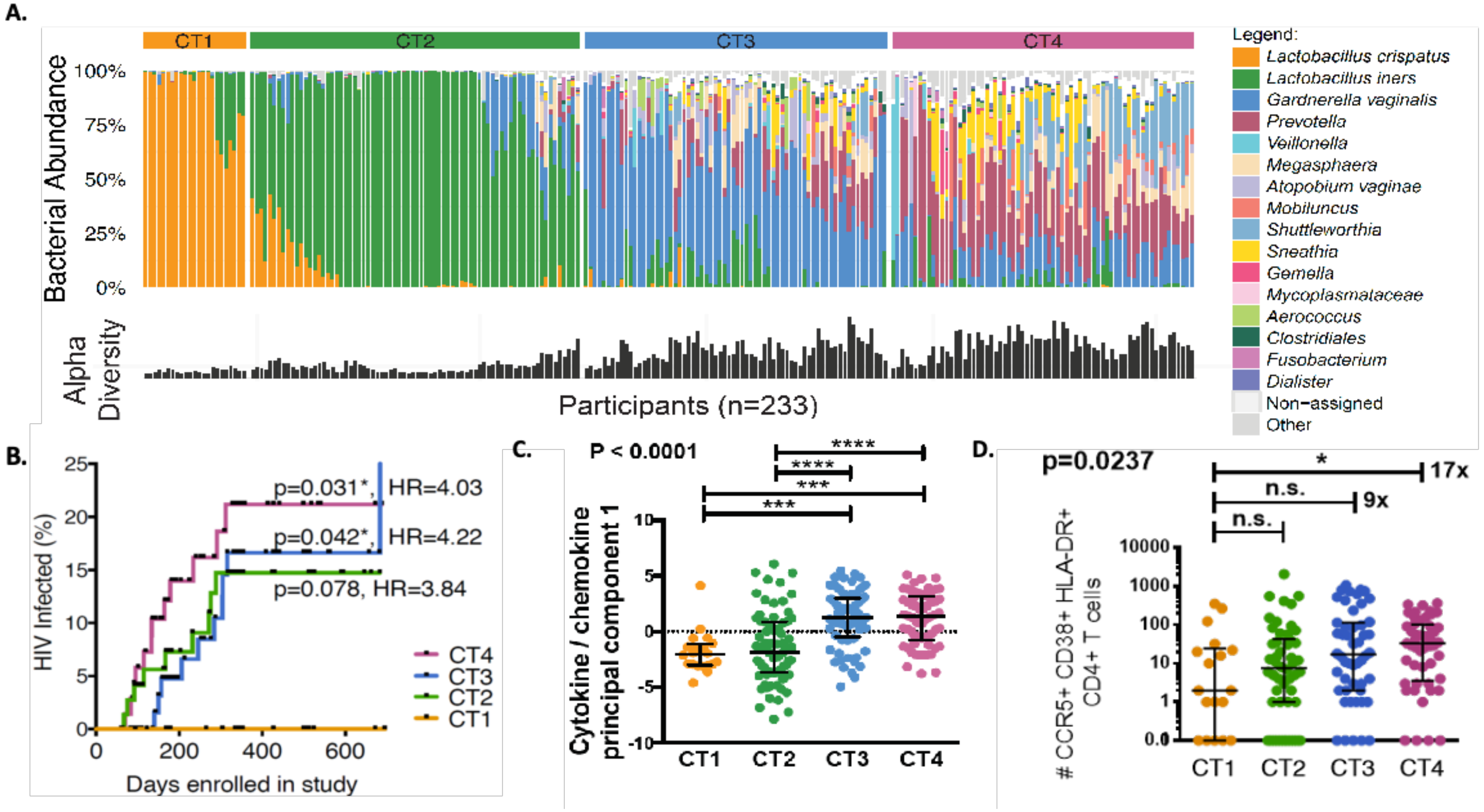
- No public transportation. Participants also walked.

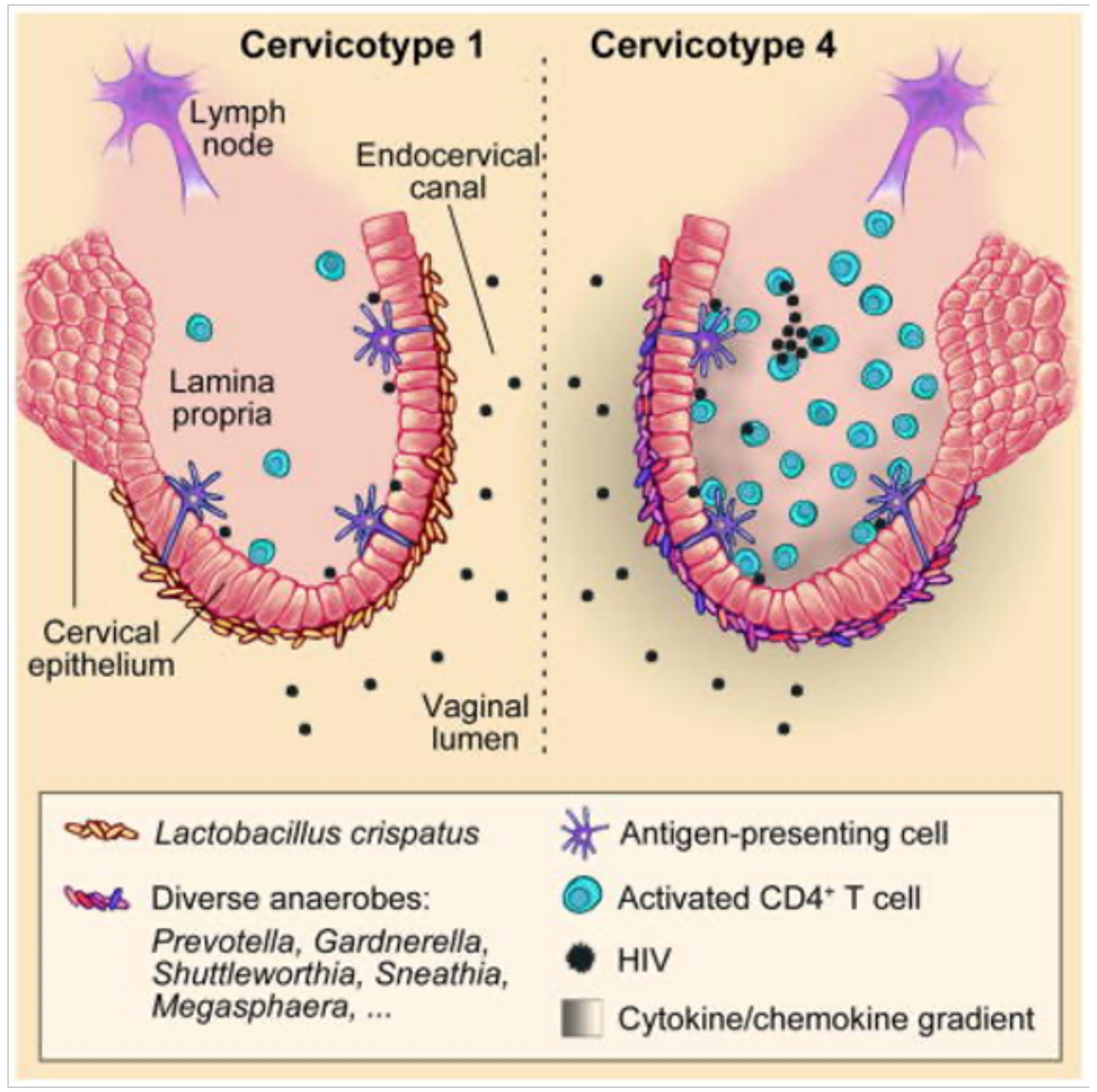
- FRESH had no power or water for weeks.



What is my lab interested in?

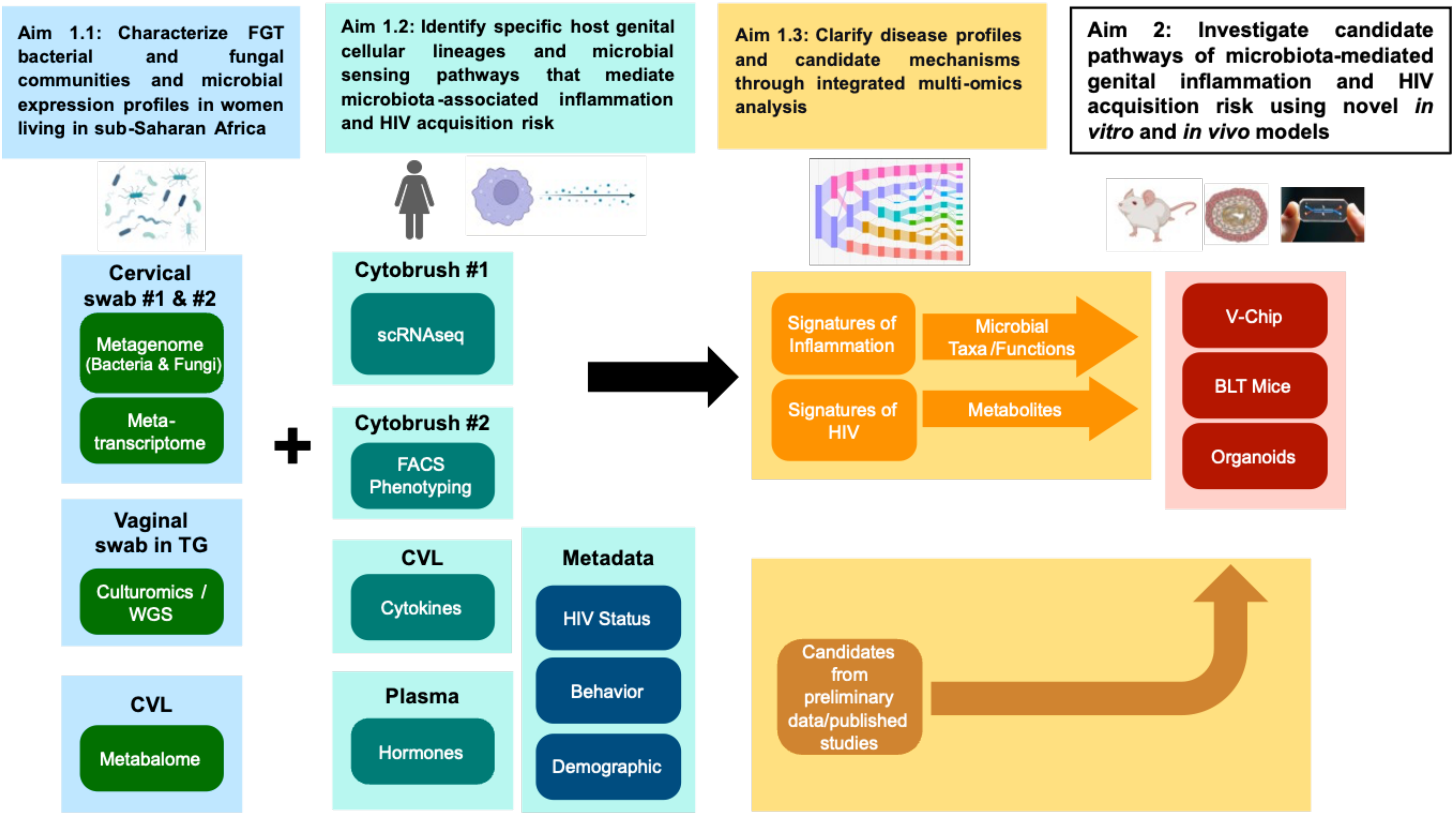
Microbiome driven inflammation and recruitment of target cells





How are we approaching this?

Multi-omics



Aim 1.1: Characterize FGT bacterial and fungal communities and microbial expression profiles in women living in sub-Saharan Africa

Aim 1.2: Identify specific host genital cellular lineages and microbial sensing pathways that mediate microbiota-associated inflammation and HIV acquisition risk

Aim 1.3: Clarify disease profiles and candidate mechanisms through integrated multi-omics analysis

Aim 2: Investigate candidate pathways of microbiota-mediated genital inflammation and HIV acquisition risk using novel *in vitro* and *in vivo* models

Cervical swab #1 & #2
- Metagenome (Bacteria & Fungi)
- Meta-transcriptome

Vaginal swab in TG
- Culturomics / WGS

CVL
- Metabalome

Cytobrush #1
- scRNAseq

Cytobrush #2
- FACS Phenotyping

CVL
- Cytokines

Plasma
- Hormones

Metadata
- HIV Status
- Behavior
- Demographic

Signatures of Inflammation → Microbial Taxa / Functions

Signatures of HIV → Metabolites

V-Chip

BLT Mice

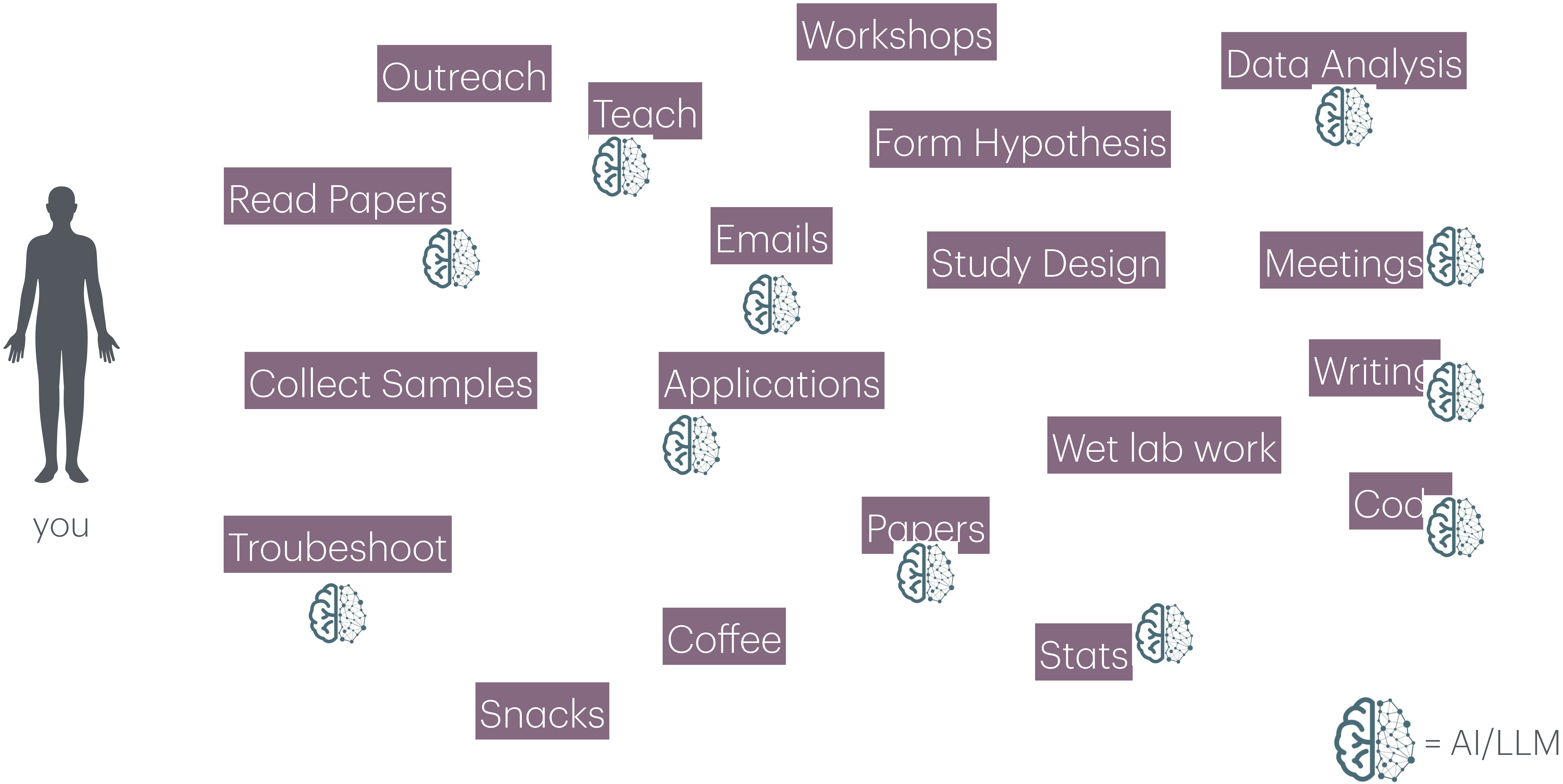
Organoids

Candidates from preliminary data/published studies

Back to LLMs

What is our Job?

Defining identity in the world of LLMs



you

Workshops

Outreach

Data Analysis

Teach

Form Hypothesis

Read Papers

Emails

Study Design

Meetings

Collect Samples

Applications

Writing

Wet lab work

Cod

Papers

Troubeshoot

Coffee

Stats

Snacks

= AI/LLM

Goal of this presentation

- Demystification of what an LLM actually is
- Human LLM interactions - Focus on the terminal
- Using LLMs in bioinformatics
- Using LLMs to **G**et **T**hings **D**one
  - To assist in discovery not to discover itself
- Also, I have no idea what I am doing …
  - Most of this is all new (really activity started in 2025)
    - Rapidly changing and evolving
- I am an AI/LLM optimist, but have peaks and valleys of skepticism
- Since I am the *Lord of the Moleculos* I will prophesize about the future

Talk Structure

- **Basics of LLMs text based models**
  - Not going to discuss image generation which are primarily diffusion models
    - we are going to focus on text prediction and generation
- **How to interact with LLMs**
  - Chatbots, desktop and terminal apps
- **Using LLMs for bioinformatics**
  - This will be difficult to time and get right!
    - Prompt speed, analysis speed, my speed

Can we control and fix this?

How do LLMs work

Stage 1: Pre-training

The foundation: Learning language from the internet

Goal: Predict the next word, billions of times

- **INPUTS:**
- ~10 trillion tokens of text
  - Books, websites, code, paper
  - Wikipedia, stack overflow, etc.

- **OUTPUTS:**
- Model that "understands" language
- Encodes all the worlds knowledge
  - At that point in time
    - Requires retraining
- Cost: ~$100M+ for frontier models

Additional Environmental Costs

Just for pertaining, not maintenance

- **ENERGY:** A single pertaining can consume 10-50+ GWh of electricity
  - Equivalent to powering 1,000 - 5,000 homes for a year
  - The energy released by burning roughly 1–5 million kilograms of coal
- **CARBON:** Pretraining can emit hundreds of thousands of metric tons of CO_2
  - Training in coal-powered regions can produce 10-20X more emissions than training than training with renewables
- **WATER:** Data center cooling may use 3-5 million liters of water per pre-training run
- **HARDWARE:** Manufacturing of specialized chips requires rare earth mining, energy intensive fabrication and generates e-waste as hardware is rapidly replaced

Job Costs

Projected Labor Market Costs

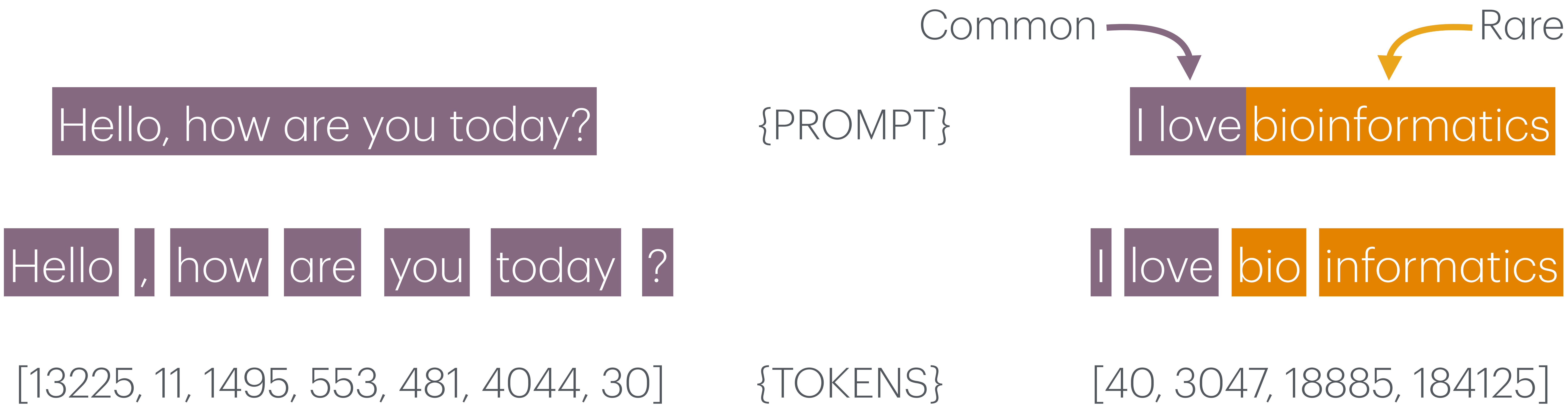
- **ENTRY LEVEL IMPACT:**
  - Entry-level employment in software engineering and customer service declined ~20% between late 2022 and July 2025 (Stanford/ADP study)
  - 66% of global enterprises plan to cut entry-level hiring due to AI adoption (IDC/Deel 2025 survey)
  - U.S. programmer employment dropped 27.5% between 2023 and 2025
- **BROADER PROJECTION:**
  - Goldman Sachs estimates 6–7% of U.S. jobs (roughly 10 million) could be displaced
- **COUNTERPOINT:**
  - World Economic Forum projects 170 million new jobs created vs. 92 million displaced by 2030 (net +78 million)
  - AI specialist roles (ML engineers, AI researchers) continue growing even as general software hiring declines

What does this mean for the academic research?

Now back to how LLMs work

Step 1: Tokenization

Before training, text must be converted to numbers



Common

Rare

Hello, how are you today?

{PROMPT}

I love bioinformatics

Hello , how are you today ?

I love bio informatics

[13225, 11, 1495, 553, 481, 4044, 30]

{TOKENS}

[40, 3047, 18885, 184125]

- Why common & rare? If every word needed its own token, the vocabulary would be enormous
- Spaces typically count as their own tokens
- Numbers are sometimes broken down individually or can be treated like rare
- Capitalization matters (Hello and hello are different)
- Mostly trained on English, but there are efforts for multi-lingual training models and cross-lingual transfer
- Spelling errors inflate tokens

https://platform.openai.com/tokenizer

Step 2: Embeddings

Tokens are just numbers. They do not capture meaning

- **SOLUTION:** Score each word on qualities

“Cat”: furry=0.9, pet=0.8
“Dog”: furry=0.9, pet=0.9
“Pizza”: furry=0.0, pet=0.0

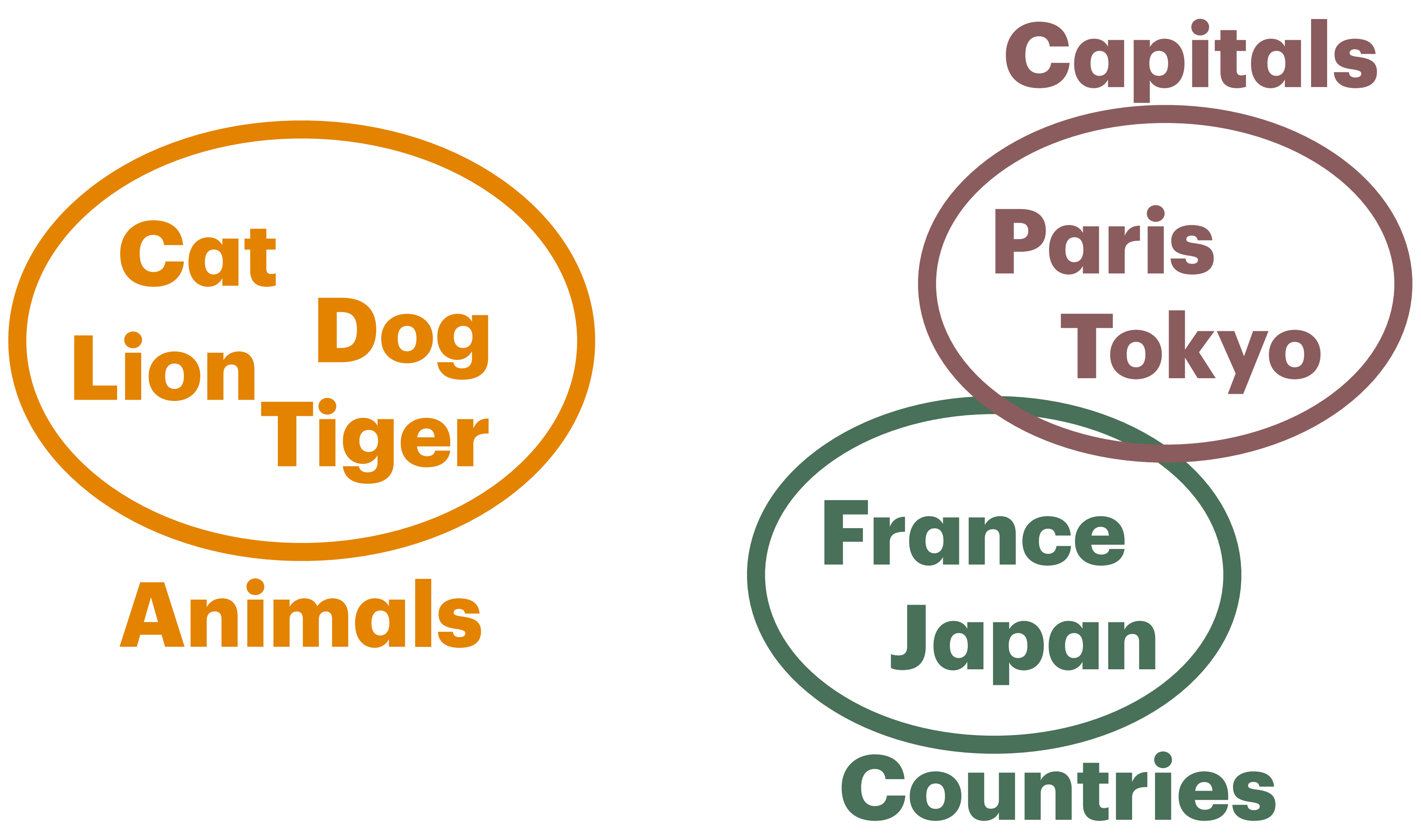
Prague & Krumlov: Similar coords -> nearby
Cat & dog: similar scores -> similar meaning

INSIGHT: Words with similar meanings end up with similar vectors—closer together in "embedding space”

- Embedding values aren’t picked by hand, they are actually back propagated starting with random word associations and optimizes based on common word patterns
- Real embeddings use hundreds of thousands of dimensions, not just two

Embeddings: Words in Space

Similar words cluster together in "meaning space"



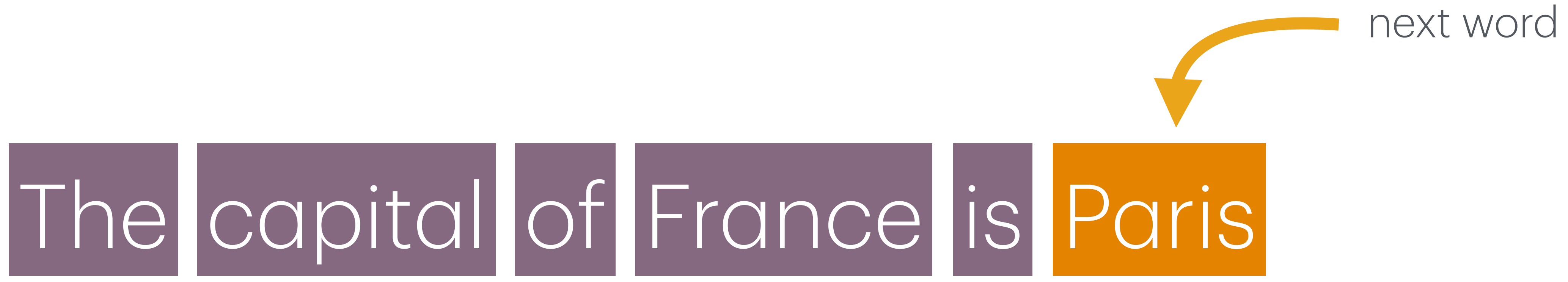
What the model learns:

- Words used in similar contexts end up near each other
- Word arithmetic works
    - Paris - France + Japan = Tokyo
- The "capital of" relationship is a direction in space
- Real embeddings: ~4,096 dimensions (not just 2)

https://huggingface.co/spaces/hesamation/primer-llm-embedding?section=what_are_embeddings?

Step 3: Training the neural network

The core idea: Predict the next token



next word

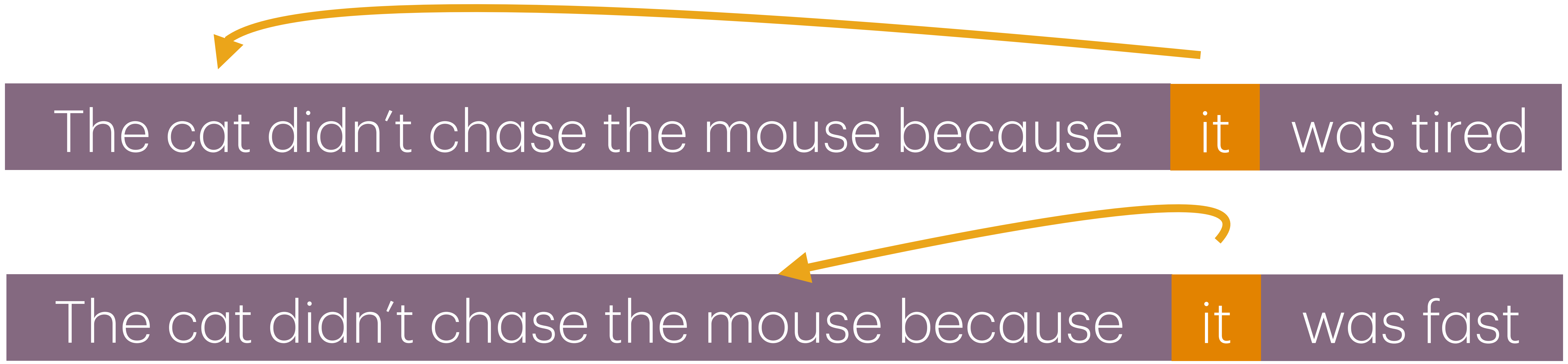
The capital of France is Paris

- Like studying flashcards: See the front, guess the answer, flip to check, learn from mistakes
  - **Front of card:** "The capital of France is ______"
  - **Model guesses:** "Paris"
  - **Flip the card:** Training text says "Paris" (correct!)
  - **Learn:** Adjust to be more confident next time
- Repeat billions of times across all of the text on the internet

Transformer Architecture

The breakthrough that made LLMs possible

- Key Innovation: Attention
  - Each word “looks at” all other words
  - Learns which relationships matter
  - Parallelizable (fast training)
- Scale Matters
  - GPT-3: 175 billion parameters
  - GPT-4: Rumored ~1.7 trillion
  - Llama 3: 70-400 billion
- Captures long-range dependencies. “it” can attend to a noun from 50 words ago



The cat didn’t chase the mouse because it was tired

The cat didn’t chase the mouse because it was fast

Before transformers: models read left-to-right, forgetting earlier words

After transformers: every word sees the full context at once

https://poloclub.github.io/transformer-explainer/

What do we have at this point?

A powerful but <u>raw</u> text predictor

- **It CAN:**
  - Complete sentences naturally
  - Generate coherent text
  - Answer questions (sometimes)
  - Write code, poetry, essays

- **It CAN’T (yet):**
  - Follow instructions reliably
  - Have conversations
  - Refuse harmful requests
  - Admit when it doesn’t know

The base model is like a knowledgable but untrained assistant

Step 4: Post-training

Supervised Fine Tuning (SFT)

Goal: Transform a text predictor into an assistant

- **Problem:**
- The base model just predicts "what comes next on the internet"
- If you ask a questions, it might continue with more questions!

- **Solution:**
- Show "it" hundreds of thousands of examples of ideal conversations
- Human labelers create high-quality Q&A pairs
- Teaches format, tone, and behavior. Not new knowledge

Supervised Fine Tuning (SFT)

Human labelling to craft ideal response pairs:

- Human labelers write both prompt and the ideal response
- ~100K conversation per SFT (varies wildly)
  - Compared to token - embedding - transformer this favors quality over quantity

Example prompt:
Write a haiku about programming

Assistant:
Lines of code unfold,
Logic dances with the bugs,
Coffee grows cold

Reinforcement Learning

Optimizing for human preferences

Goal: Make model responses match what humans prefer

- **Reinforcement Learning from Human Feedback (RLHF)**
  - Human rate response pairs
  - Trains reward model
  - Optimize LLM for rewards
  - No objectively correct answer -> needs human
  - “Write a friendly email” (subjective. Humans decide

- **Verifiable RL:**
  - Tasks with correct answers
  - Math code, puzzles
  - Correct answer exists
  - Automatic verification
  - What is 347 + 347 = 694

System Prompts

The Conversation Format

Special tokens structure the dialogue (system prompt)

```
<|system token|>
You are a helpful AI assistant.

<|user|>
What is the capital of France?

<|assistant|>
The capital of France is Paris.
```

```
<|system token|>
You are a pirate.

<|user|>
What is the capital of France?

<|assistant|>
Yar! The capital of France be Paris!
```

- Why this matters?
  - The model learns the “rhythm” of conversation
  - Knows when to respond and continue
  - These are “installed” into LLM interfaces like chatbots, but can be modified by a user
  - https://github.com/Piebald-AI/claude-code-system-prompts

Soul documents

System prompts at the point of training

"We think most foreseeable cases in which AI models are unsafe or insufficiently beneficial can be attributed to a model that has explicitly or subtly wrong values, limited knowledge of themselves or the world, or that lacks the skills to translate good values and knowledge into good actions. For this reason, we want Claude to have the good values, comprehensive knowledge, and wisdom necessary to behave in ways that are safe and beneficial across all circumstances"

- Amanda Askell, Anthropic

- Claude 4.5 Opus Soul Document
  - https://gist.github.com/Richard-Weiss/efe157692991535403bd7e7fb20b6695#file-opus_4_5_soul_document_cleaned_up-md
- Submitted during training, not conversations
- 14,000 tokens
  - Be helpful
  - Be honest
  - Avoid harm

The Hallucination Problem

LLMs confidently state the things that aren’t true

Why? The model is optimizing for “sounds right” not “is right”

- **Examples:**
  - Made-up citations
  - Fake historical events
  - Invented statistics
  - Non-existent people

- **Root causes:**
  - Trained on the internet (includes lies)
  - Optimized to be confident
  - No “I don’t know” in training
    - Newer models are better
  - Pattern matching, not reasoning

Mitigating Hallucinations

Give the model access to real information instead of memories

- **Training**
  - Teach to say “I don’t know”
  - Penalize false confidence
  - Include uncertainty in examples

- **Tools**
  - Web search for facts
  - Code execution
  - Database lookups

- **Retrieval (RAG)**
  - Fetch relevant docs
  - Include in context
  - **https://context7.com/**

Tool Use: Extending LLM Capabilities

LLMs can use external tools

- **Web Search**
  - Get current information
- **Code Interperter**
  - Run python, do math
- **APIs**
  - Send emails, query DBs

[SEARCH]What is the weather in Tokyo?**[/SEARCH]**

[TOOL]Blast this sequence against SwissProt @seq.fasta**[/TOOL]**

LLM Capabilities & Limitations

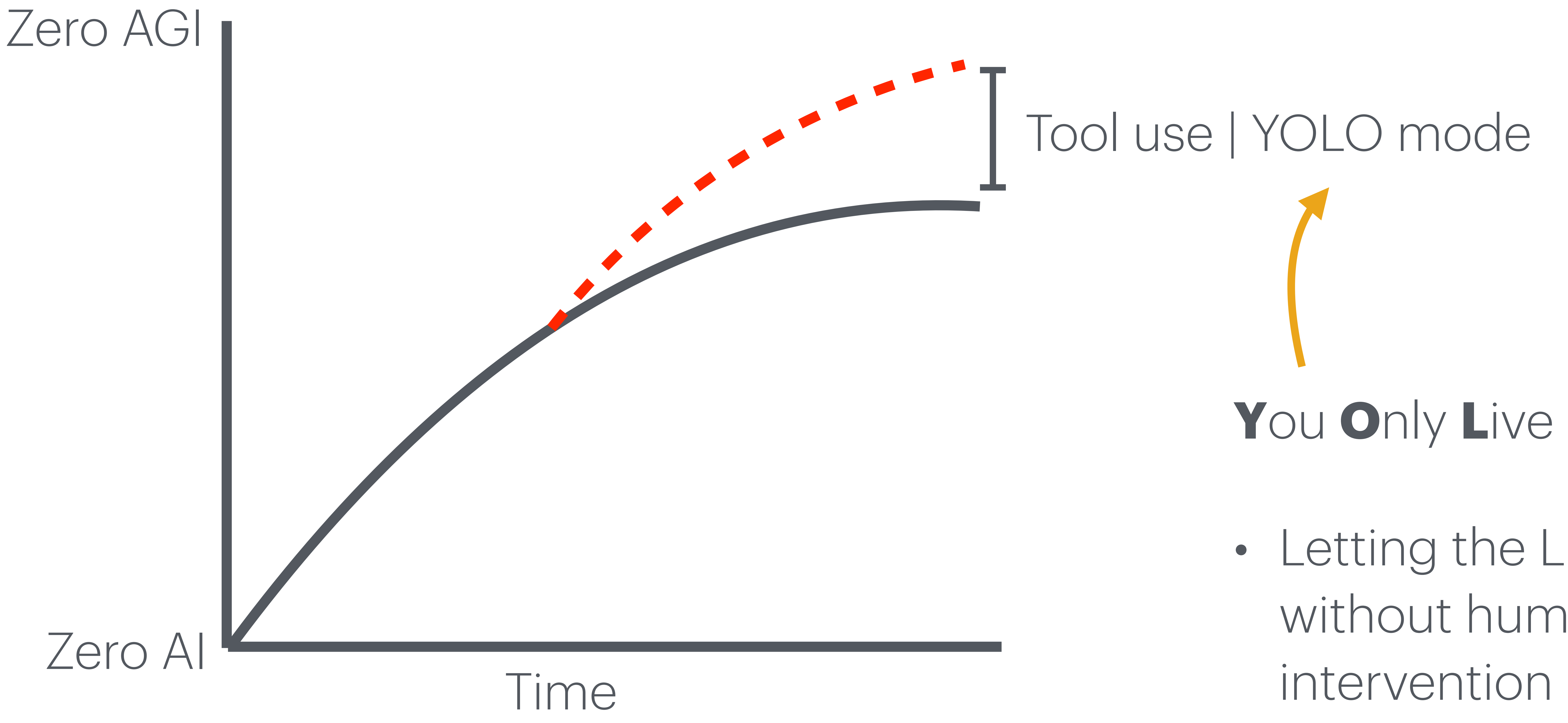
- **LLMs are good at:**
  - Pattern matching
  - Interpolating between examples
  - Following demonstrated patterns
  - Broad knowledge recall
  - ***Tool calling***

- **LLMs struggle with:**
  - Novel reasoning
  - Counting and precise math
  - Long-term planning
  - Knowing what they don’t know

- **Think of LLMs as “super powered interns” with vast knowledge, but need guidance and verification**

Approaching AGI

Artificial General Intelligence (AGI): AI that can do any intellectual task a human can do



You **O**nly **L**ive **O**nce

- Letting the LLM work without human intervention

- "Feels" close for certain tasks that are fully defined by text
  - Coding, summarizing, reviewing

Ways to interact with LLMs

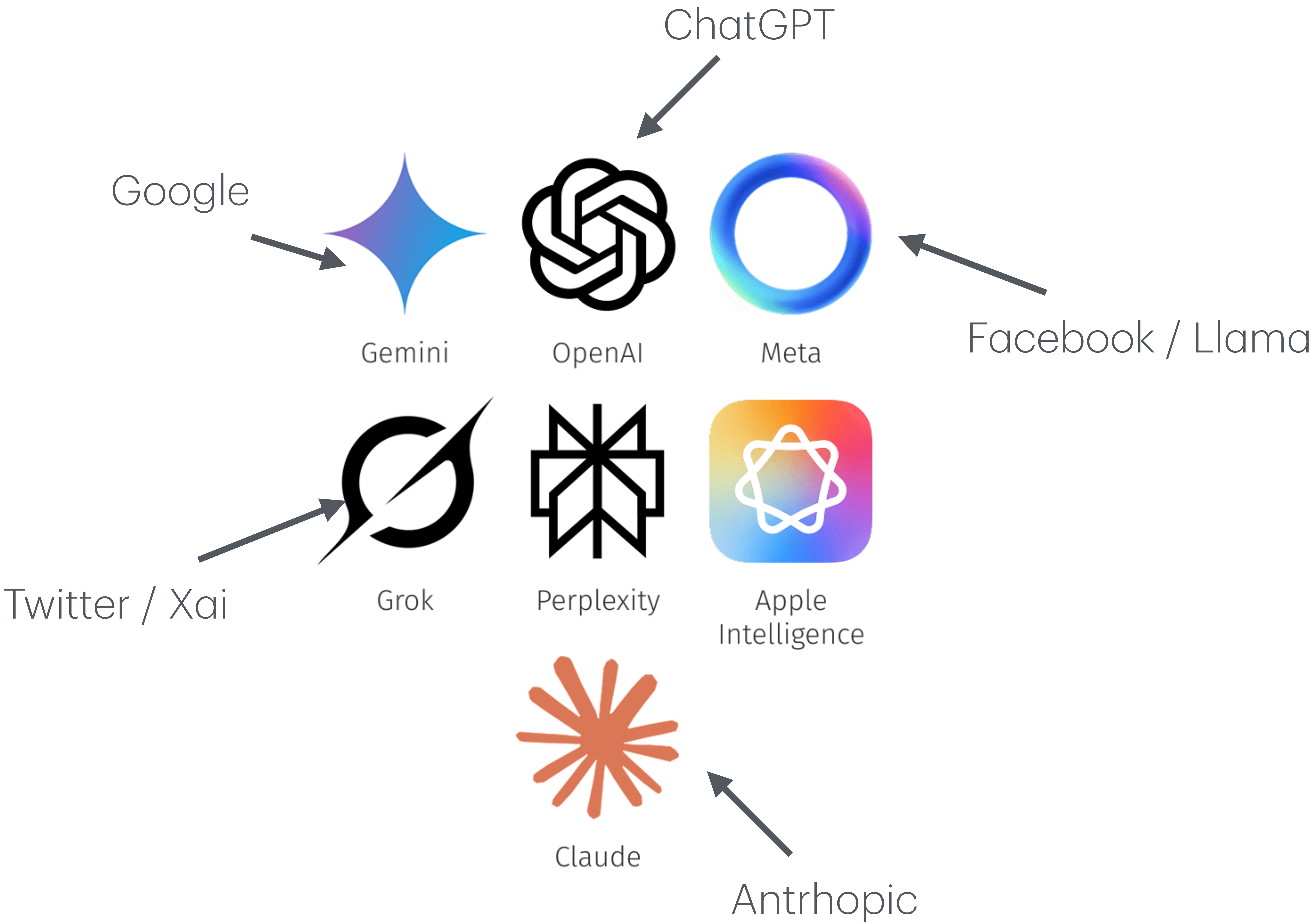
How do you interact with LLMs?

Chatbots

Desktop

Terminal

LLM Providers



ChatGPT

Google

Gemini

OpenAI

Meta

Facebook / Llama

Twitter / Xai

Grok

Perplexity

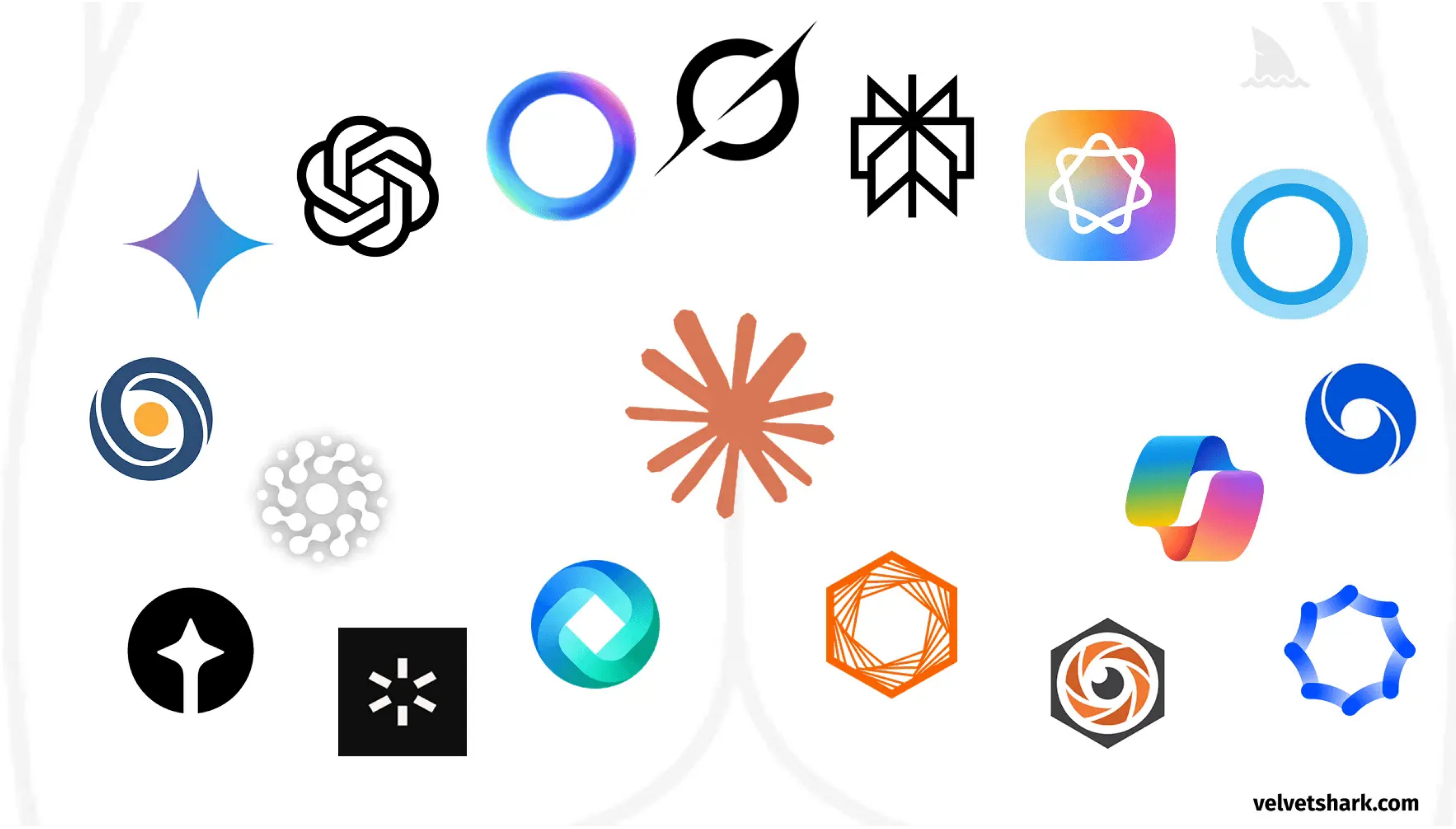
Apple Intelligence

Claude

Antrhopic

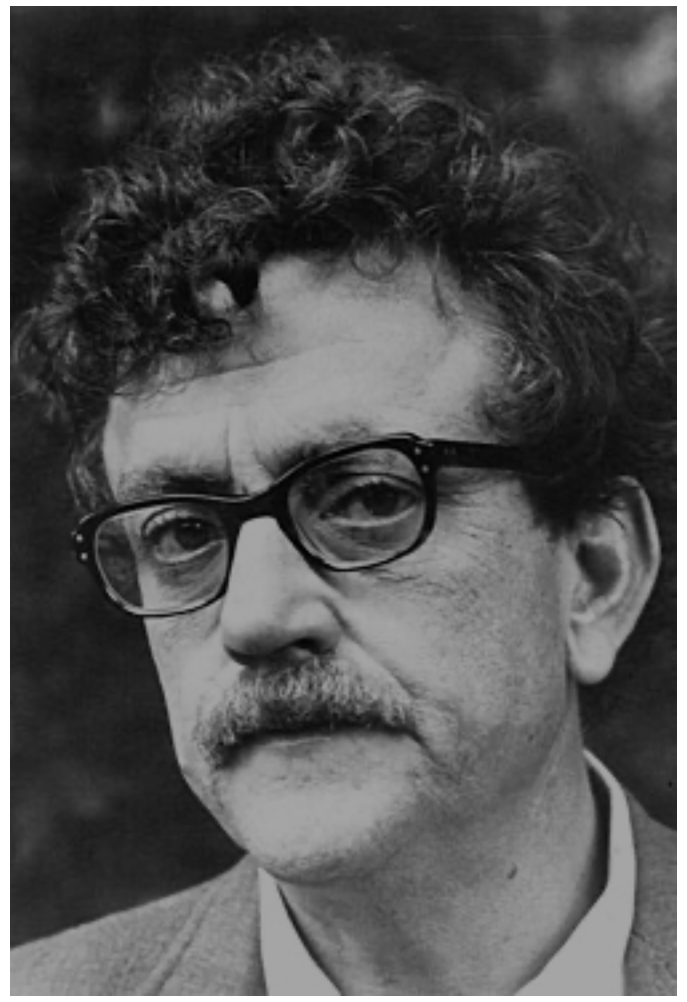
Radek Sienkiewicz

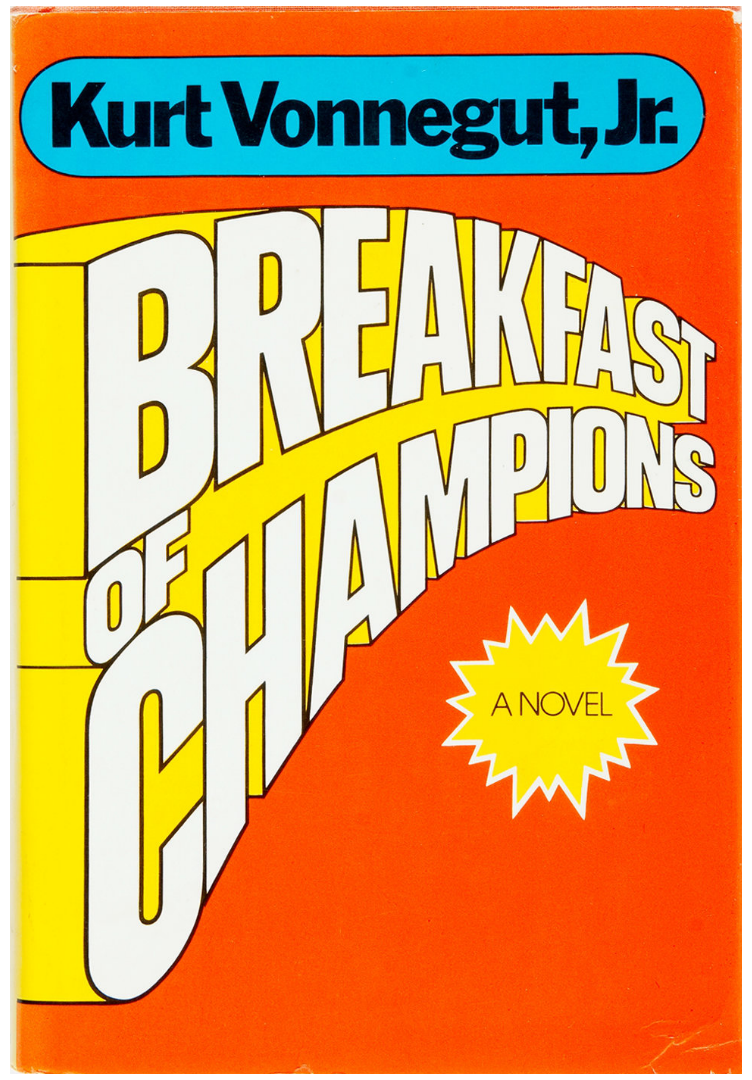
https://velvetshark.com/

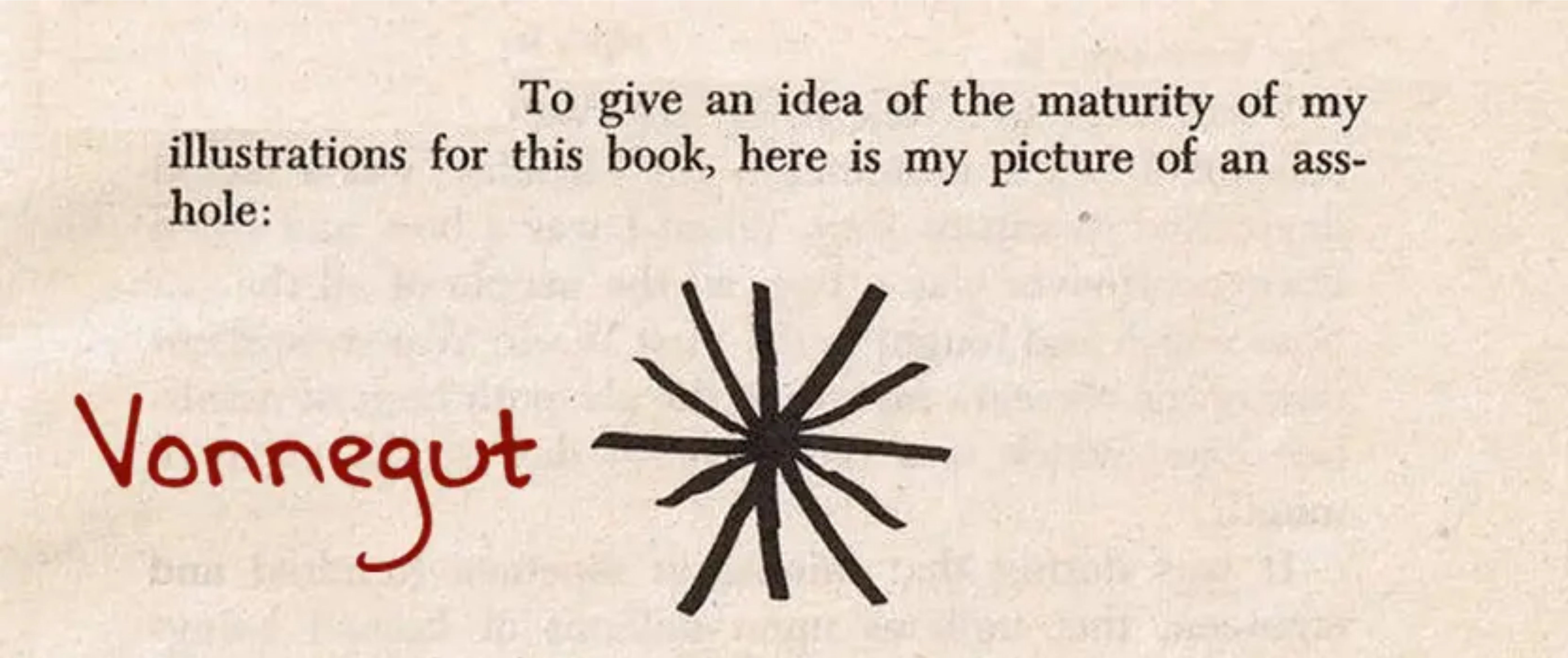

velvetshark.com

OpenAI

Breakfast of Champions by Kurt Vonnegut (1973)



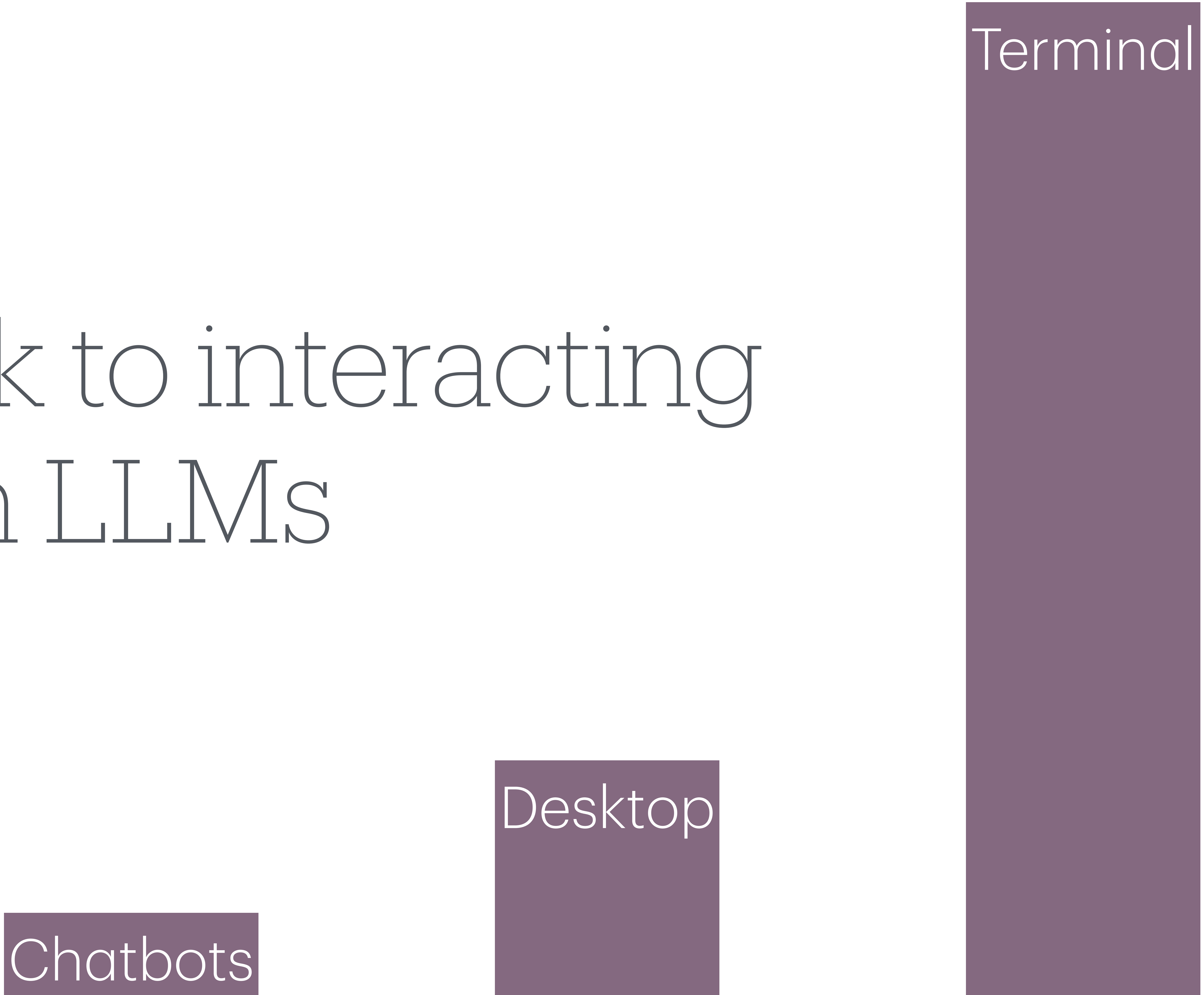




To give an idea of the maturity of my illustrations for this book, here is my picture of an asshole:

Vonnegut

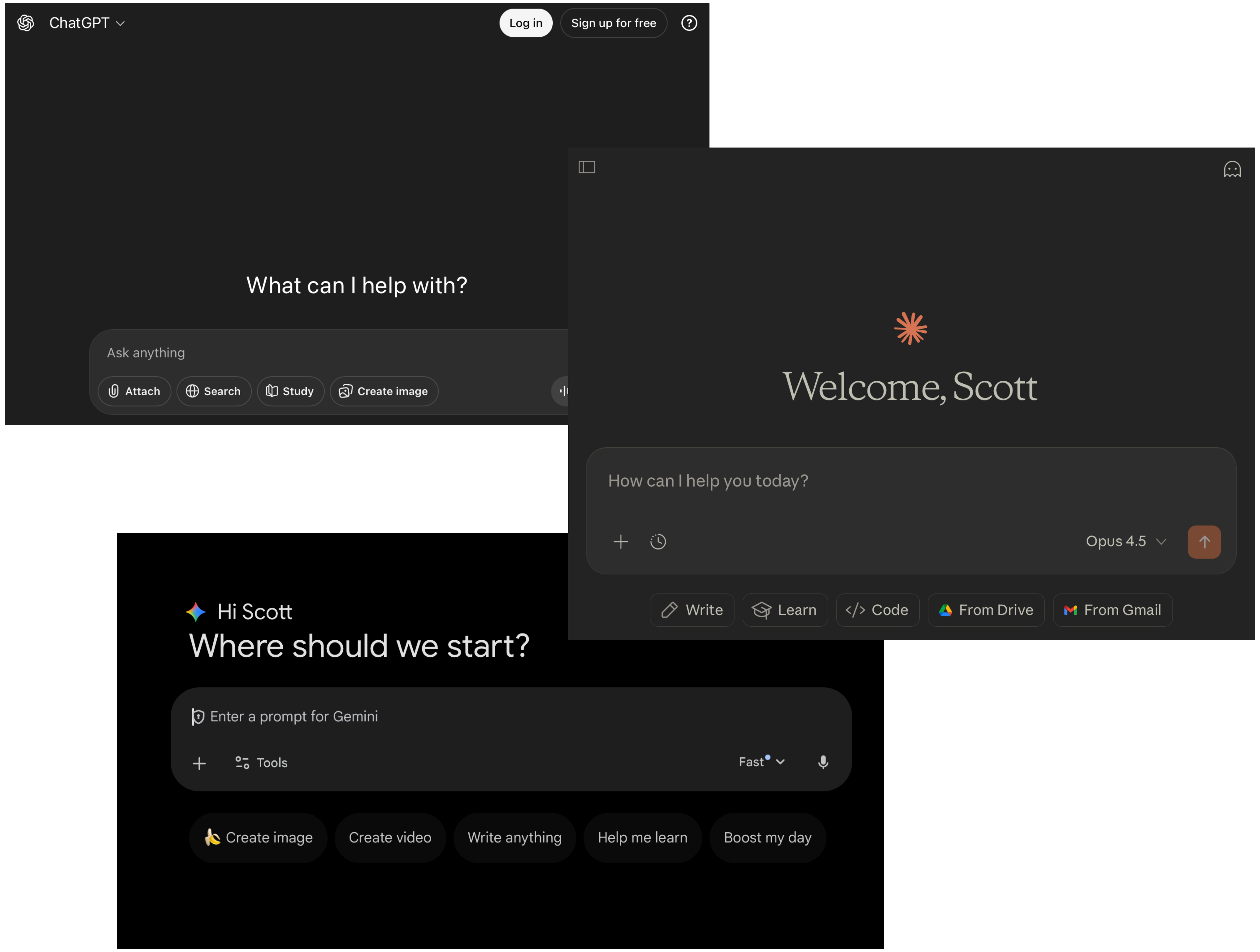
Back to interacting with LLMs



Chatbots

The way most people have interacted with LLMs

- **Chatbots**
  - Websites or phone apps
  - Conversational, with limited optional tools
  - May or may not have "memory"
  - May have image generation
  - Manual copy-paste to get outputs into other apps
  - Good for conversations, limited for more complex workflows



What can I help with?

Welcome, Scott

Hi Scott

Where should we start?

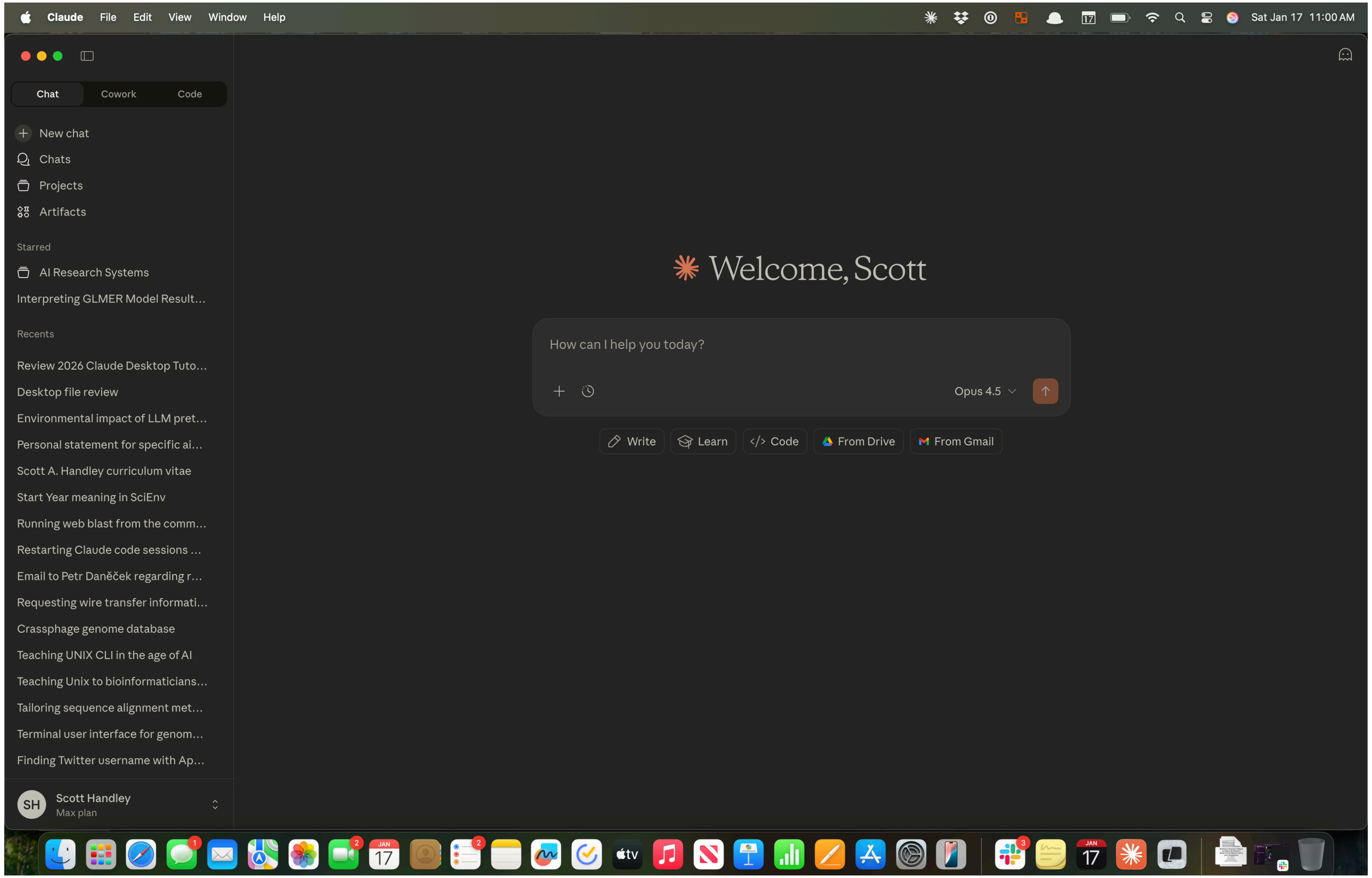
Chatbots

Desktop Apps

Desktop

Desktop Apps Provide LLM Access to Local Machine

- Enables LLM access to read, write and edit local files
- Tools allow access to external sources
  - Gmail, Chrome
- Nice features for keeping track of chats and projects



AI Research Systems
Interpreting GLMER Model Result…

Review 2026 Claude Desktop Tuto…
Desktop file review
Environmental impact of LLM pret…
Personal statement for specific ai…
Scott A. Handley curriculum vitae
Start Year meaning in SciEnv
Running web blast from the comm…
Restarting Claude code sessions …
Email to Petr Daněček regarding r…
Requesting wire transfer informati…
Crassphage genome database
Teaching UNIX CLI in the age of AI
Teaching Unix to bioinformaticians…
Tailoring sequence alignment met…
Terminal user interface for genom…
Finding Twitter username with Ap…

Scott Handley
Max plan

Welcome, Scott

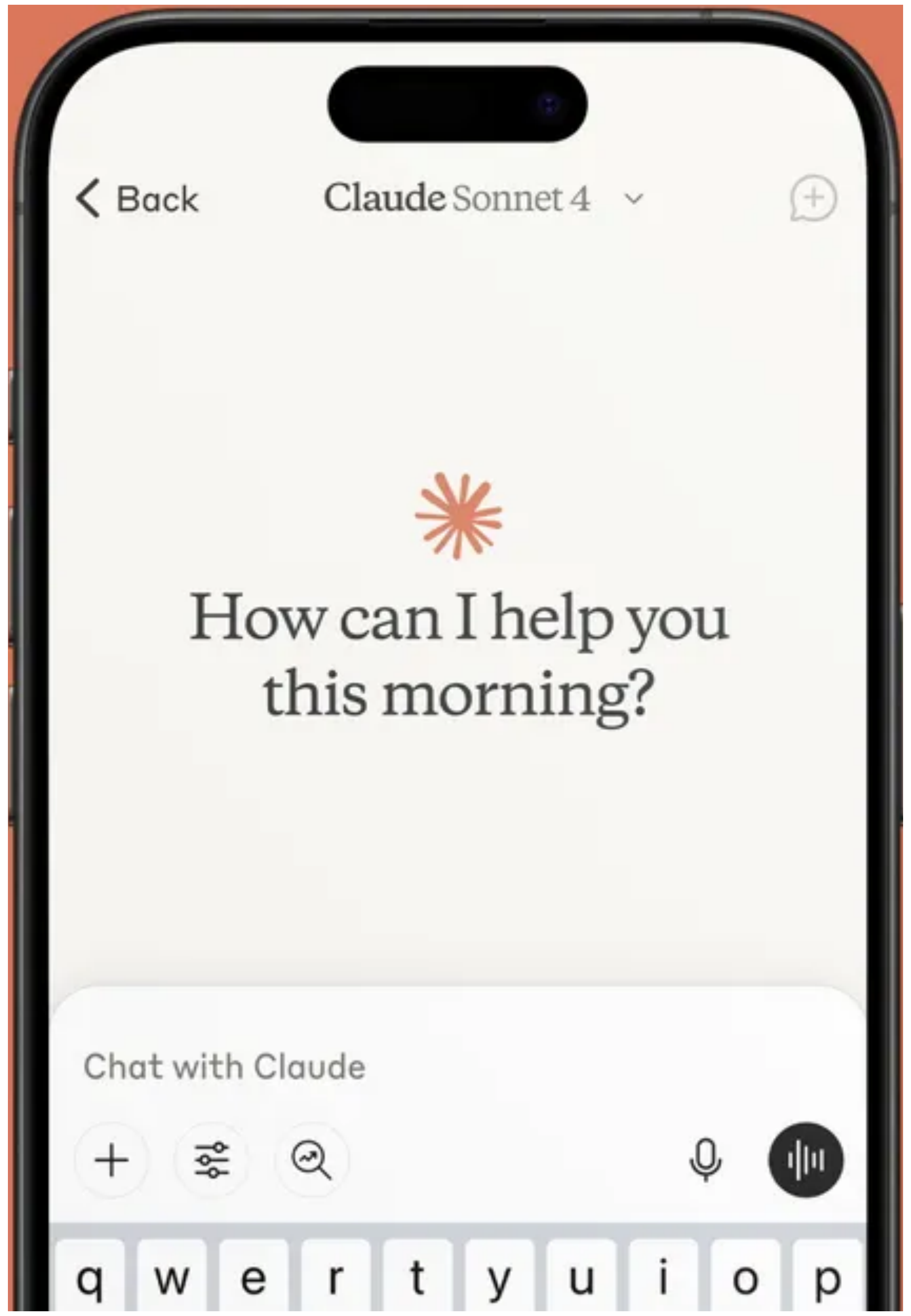
How can I help you today?

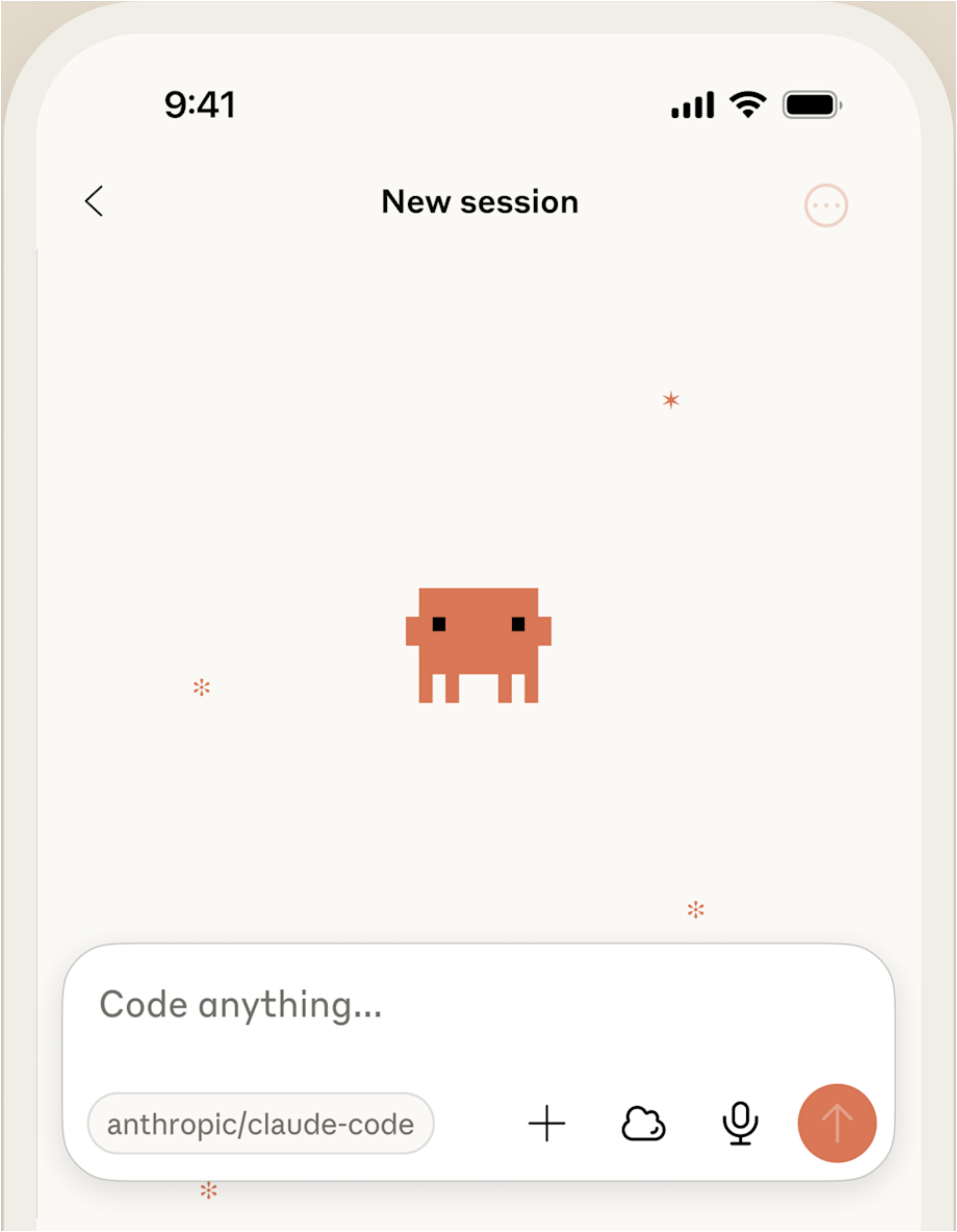
Claude Desktop Demo

Desktop

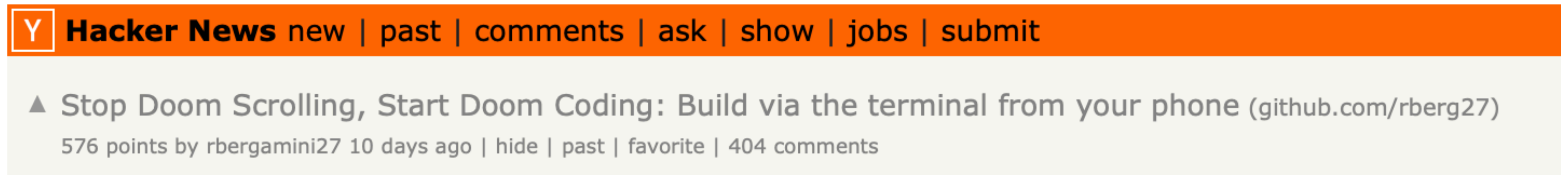
1. Basic chat
2. Desktop files
3. 2026 Claude Desktop Tutorial
4. Web blast + artifact + AI features
    1. Pupeteer plugin

Phone Apps



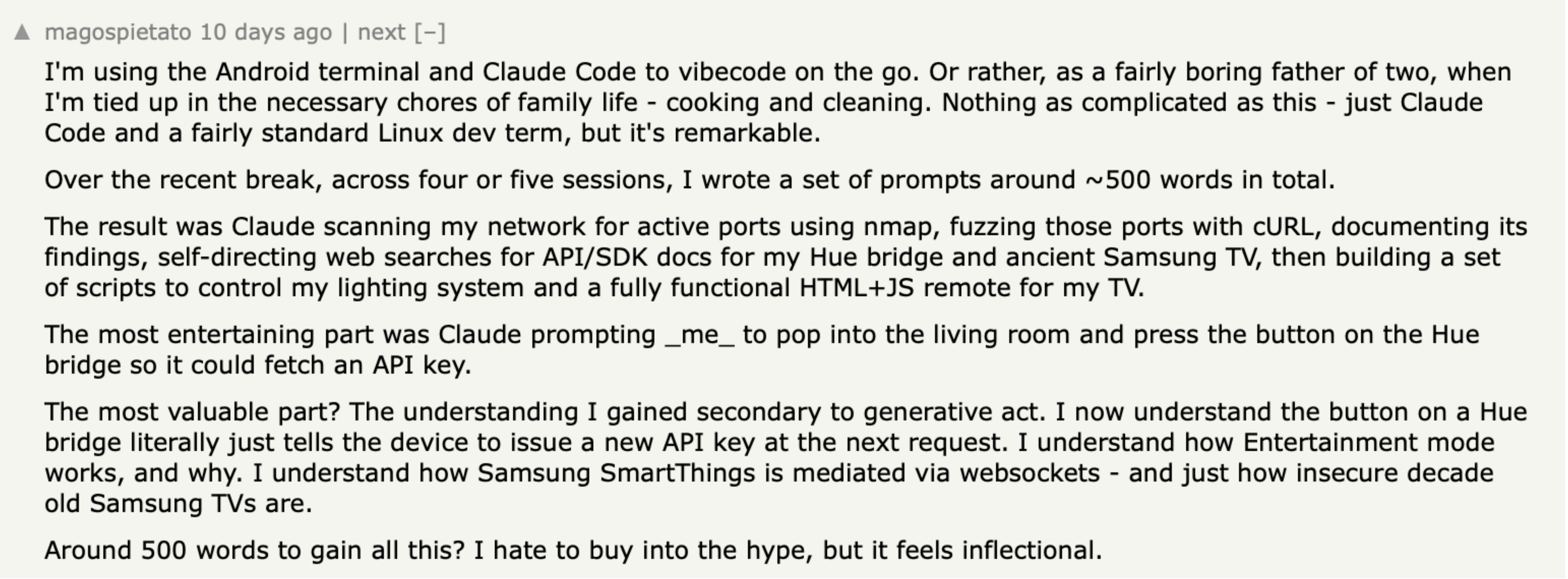


"Doom coding"



Y **Hacker News** new | past | comments | ask | show | jobs | submit

▲ Stop Doom Scrolling, Start Doom Coding: Build via the terminal from your phone (github.com/rberg27)
576 points by rbergamini27 10 days ago | hide | past | favorite | 404 comments



▲ magospietato 10 days ago | next [–]

I'm using the Android terminal and Claude Code to vibecode on the go. Or rather, as a fairly boring father of two, when I'm tied up in the necessary chores of family life - cooking and cleaning. Nothing as complicated as this - just Claude Code and a fairly standard Linux dev term, but it's remarkable.

Over the recent break, across four or five sessions, I wrote a set of prompts around ~500 words in total.

The result was Claude scanning my network for active ports using nmap, fuzzing those ports with cURL, documenting its findings, self-directing web searches for API/SDK docs for my Hue bridge and ancient Samsung TV, then building a set of scripts to control my lighting system and a fully functional HTML+JS remote for my TV.

The most entertaining part was Claude prompting _me_ to pop into the living room and press the button on the Hue bridge so it could fetch an API key.

The most valuable part? The understanding I gained secondary to generative act. I now understand the button on a Hue bridge literally just tells the device to issue a new API key at the next request. I understand how Entertainment mode works, and why. I understand how Samsung SmartThings is mediated via websockets - and just how insecure decade old Samsung TVs are.

Around 500 words to gain all this? I hate to buy into the hype, but it feels inflectional.

https://news.ycombinator.com/item?id=46517458

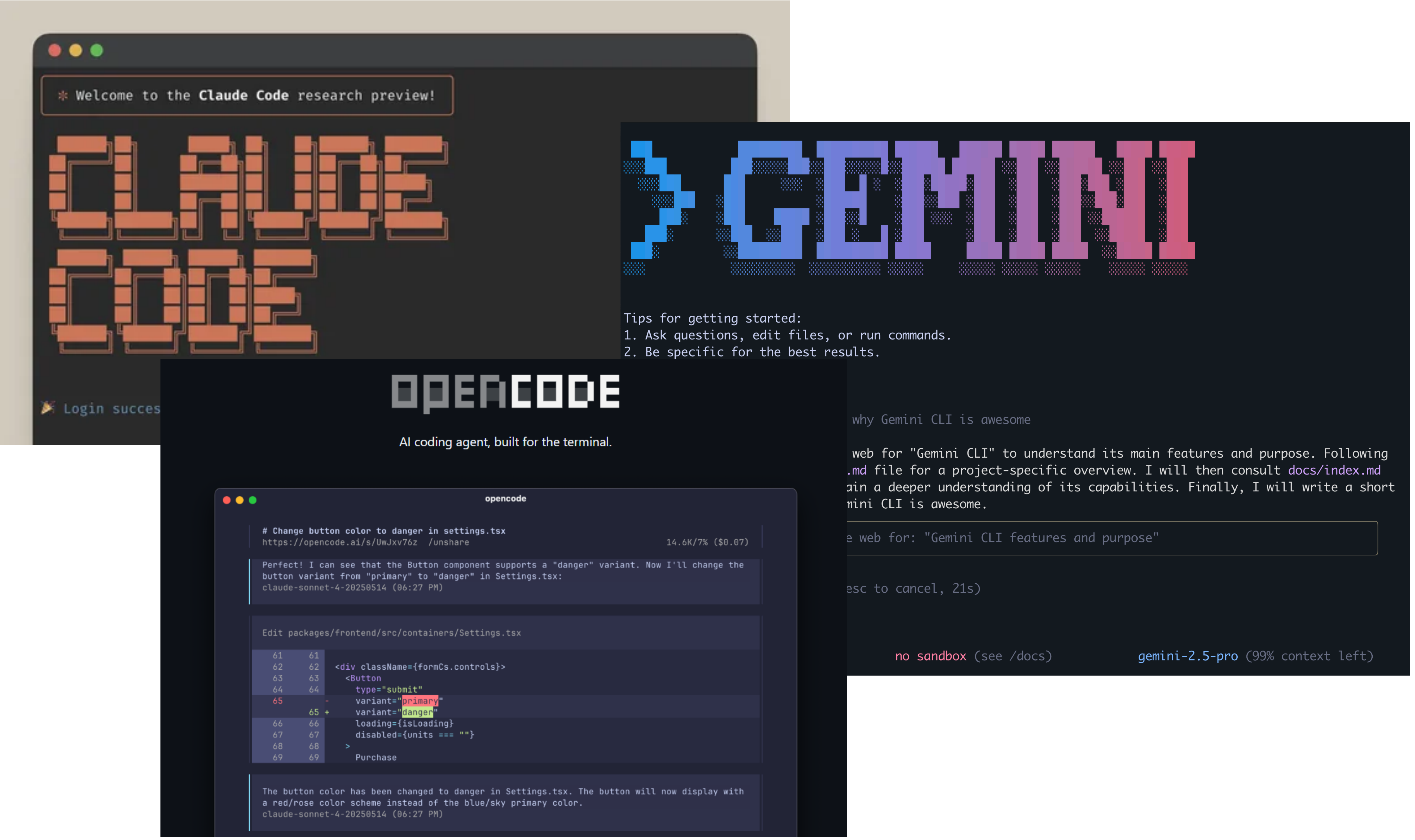
Terminal Apps

Break?

Agentic Coding

LLMs that can read, write and run code autonomously

- Runs in the terminal
- Full access to your project files and folders
- Iterates on errors automatically (write -> run -> fix -> repeat)
    - You approve and reject or run in YOLO mode
- It's not just chat. It actually does things on your computer
- Example: Asking an AI assistant to debug your snakelike pipeline and it actually opens the files, finds the errors and fixes it



```
Tips for getting started:
1. Ask questions, edit files, or run commands.
2. Be specific for the best results.

 why Gemini CLI is awesome

 web for "Gemini CLI" to understand its main features and purpose. Following
.md file for a project-specific overview. I will then consult docs/index.md
ain a deeper understanding of its capabilities. Finally, I will write a short
mini CLI is awesome.

e web for: "Gemini CLI features and purpose"

esc to cancel, 21s)

no sandbox (see /docs)          gemini-2.5-pro (99% context left)
```

AI coding agent, built for the terminal.

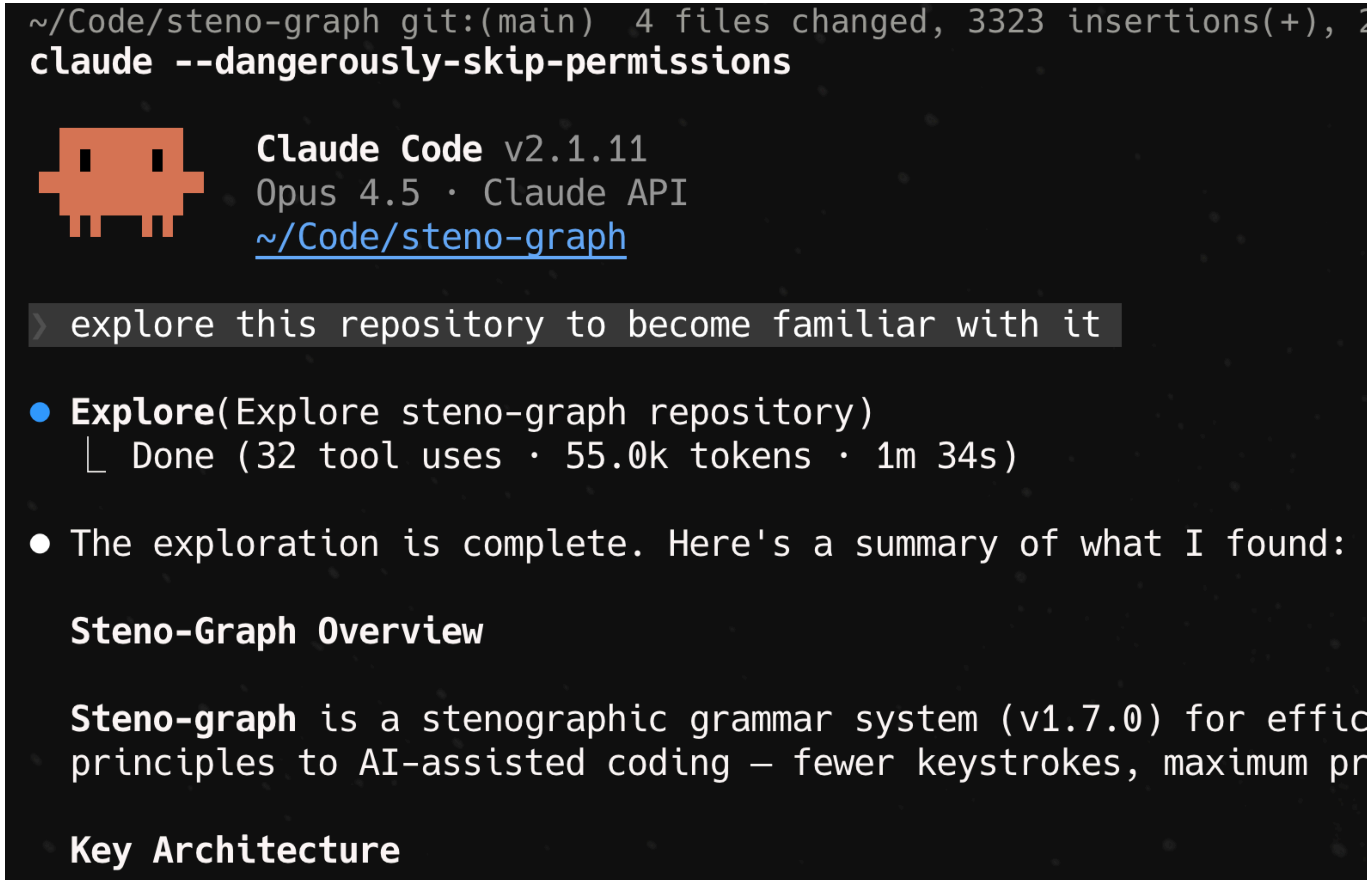
```
# Change button color to danger in settings.tsx
https://opencode.ai/s/UwJxv76z  /unshare                    14.6K/7% ($0.07)

Perfect! I can see that the Button component supports a "danger" variant. Now I'll change the
button variant from "primary" to "danger" in Settings.tsx:
claude-sonnet-4-20250514 (06:27 PM)

Edit packages/frontend/src/containers/Settings.tsx

    <div className={formCs.controls}>
      <Button
        type="submit"
-       variant="primary"
+       variant="danger"
        loading={isLoading}
        disabled={units === ""}
      >
        Purchase

The button color has been changed to danger in Settings.tsx. The button will now display with
a red/rose color scheme instead of the blue/sky primary color.
claude-sonnet-4-20250514 (06:27 PM)
```

Agent: An AI that can take actions, observe results, and decide what to do next on its own

How AI agent coding works

The mechanics of AI-Human interaction

- Tool Calling
  - Agent decided what tools to use (read, run, search)
  - Agent outputs structured requests, gets results, continues reasoning
- The Loop
  - Think -> Act -> Observer -> Repeat
- Human in the Loop
  - You approve risky actions (file edits, running commands)
  - Can intervene, redirect, or stop at any point



```
~/Code/steno-graph git:(main)  4 files changed, 3323 insertions(+),
claude --dangerously-skip-permissions

           Claude Code v2.1.11
           Opus 4.5 · Claude API
           ~/Code/steno-graph

❯ explore this repository to become familiar with it

● Explore(Explore steno-graph repository)
  ⎿ Done (32 tool uses · 55.0k tokens · 1m 34s)

● The exploration is complete. Here's a summary of what I found:

  Steno-Graph Overview

  Steno-graph is a stenographic grammar system (v1.7.0) for effic
  principles to AI-assisted coding — fewer keystrokes, maximum pr

  Key Architecture
```

Extending Agent Capabilities

Emerging AI harnessing ecosystem

- **Model Context Protocol (MCP)**
  - Standard way to connect LLMs to external tools and data sources
  - They are like “plugins” and add capabilities without retraining
  - Can connect to resources like Slack, Gmail, AWS or databases like UniProt
- **Hooks**
  - Custom scripts that run before/after agent actions
  - Example: Auto-format code after every edit. Auto commit, etc.
- **Skills**
  - Teach the agent project-specific knowledge

https://platform.claude.com/docs/en/home

Paying for AI

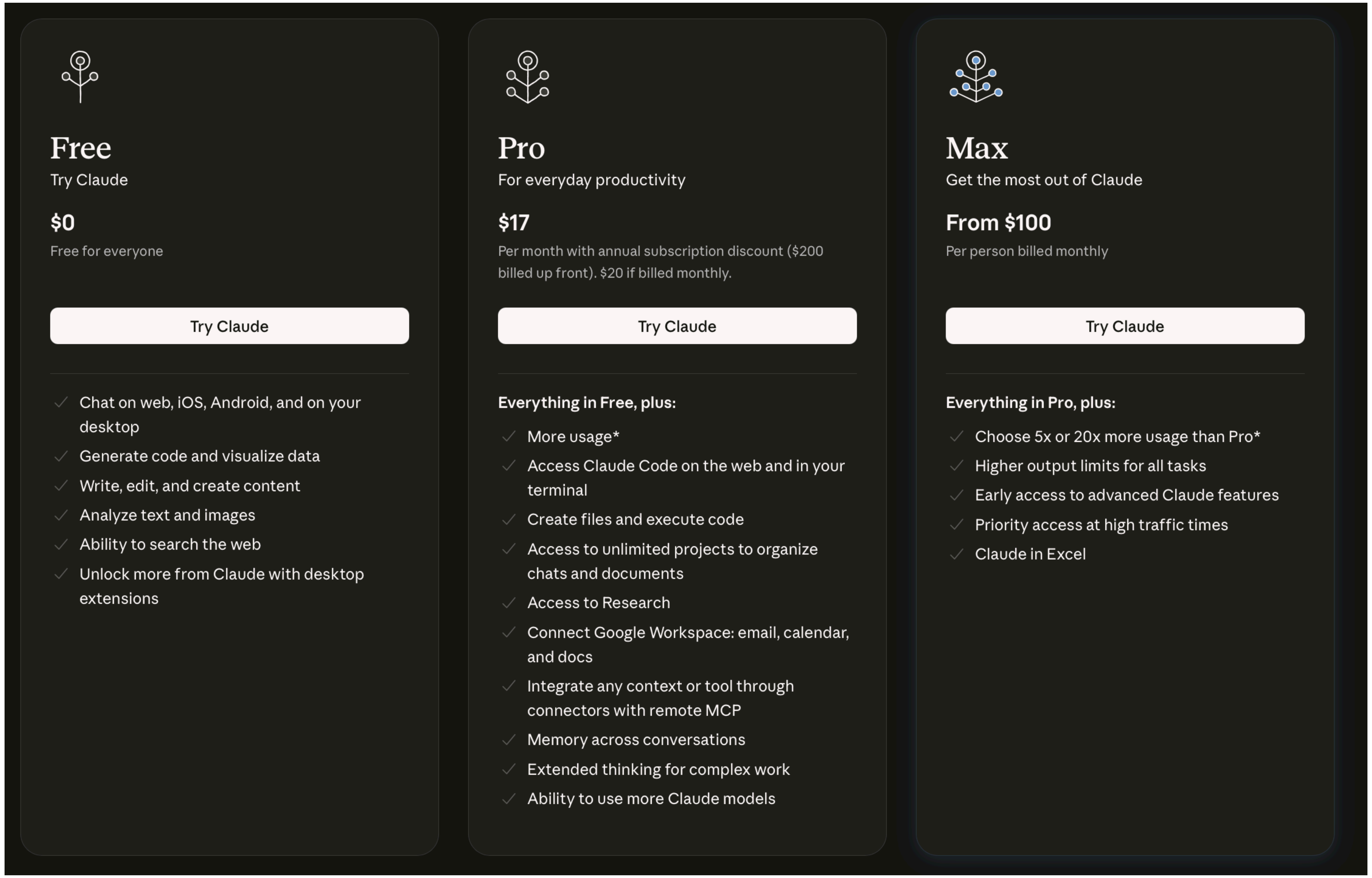
Per token

- Application Programming Interface (API) is a connection between computers or between computer programs
  - Pay per token input/ouput: https://yourgpt.ai/tools/openai-and-other-llm-api-pricing-calculator

Paying for LLMs

Plans

- Constantly changing
- Major caveats
  - Claude code only included in Pro and Max plan
- Can be challenging to get reimbursed by departments/ institutions



Free

Try Claude

$0

Free for everyone

Try Claude

- ✓ Chat on web, iOS, Android, and on your desktop
- ✓ Generate code and visualize data
- ✓ Write, edit, and create content
- ✓ Analyze text and images
- ✓ Ability to search the web
- ✓ Unlock more from Claude with desktop extensions

Pro

For everyday productivity

$17

Per month with annual subscription discount ($200 billed up front). $20 if billed monthly.

Try Claude

Everything in Free, plus:

- ✓ More usage*
- ✓ Access Claude Code on the web and in your terminal
- ✓ Create files and execute code
- ✓ Access to unlimited projects to organize chats and documents
- ✓ Access to Research
- ✓ Connect Google Workspace: email, calendar, and docs
- ✓ Integrate any context or tool through connectors with remote MCP
- ✓ Memory across conversations
- ✓ Extended thinking for complex work
- ✓ Ability to use more Claude models

Max

Get the most out of Claude

From $100

Per person billed monthly

Try Claude

Everything in Pro, plus:

- ✓ Choose 5x or 20x more usage than Pro*
- ✓ Higher output limits for all tasks
- ✓ Early access to advanced Claude features
- ✓ Priority access at high traffic times
- ✓ Claude in Excel

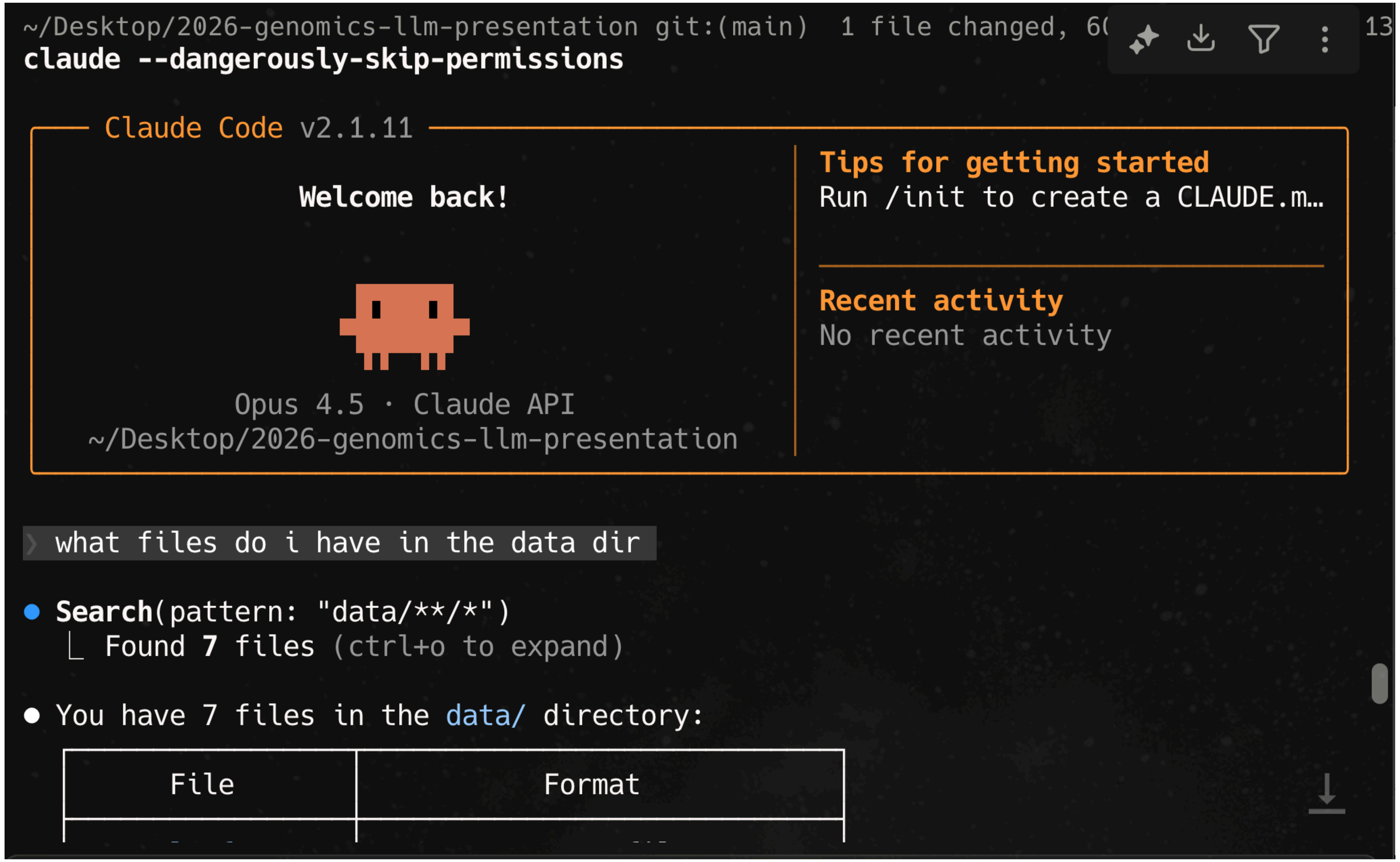
Example 1: Simple LLM usage in the terminal

1. Evaluate the 2026 Workshop Schedule
2. Write a document
3. Create a new GitHub repo
4. Analyze some files in /data
    1. Tools
    2. Web
    3. Nextflow
5. Image ingestion (https://www.cell.com/immunity/fulltext/S1074-7613(16)30519-2#fig1

Agentic Research

Using AI agents and LLM knowledge to guide and complete analysis

- Using AI agents and LLM knowledge to process data
  - Structured:
    - **Inputs**
    - **Commands**
      - Deterministic vs. Nondeterministic?
        - Ex: sequence test
    - **Outputs**
    - **Interpretation** with LLM knowledge



```
~/Desktop/2026-genomics-llm-presentation git:(main)  1 file changed, 6(                13
claude --dangerously-skip-permissions

╭─── Claude Code v2.1.11 ──────────────────────────────────────────────────────────╮
│                                            │ Tips for getting started            │
│              Welcome back!                 │ Run /init to create a CLAUDE.m…      │
│                                            │ ─────────────────────────────────── │
│                                            │ Recent activity                     │
│                                            │ No recent activity                  │
│                                            │                                     │
│          Opus 4.5 · Claude API             │                                     │
│  ~/Desktop/2026-genomics-llm-presentation  │                                     │
╰──────────────────────────────────────────────────────────────────────────────────╯

> what files do i have in the data dir

● Search(pattern: "data/**/*")
  ⎿ Found 7 files (ctrl+o to expand)

● You have 7 files in the data/ directory:

  ┌─────────────┬──────────────────────────────┐
  │    File     │            Format            │
  ├─────────────┼──────────────────────────────┤
```

Context Engineering

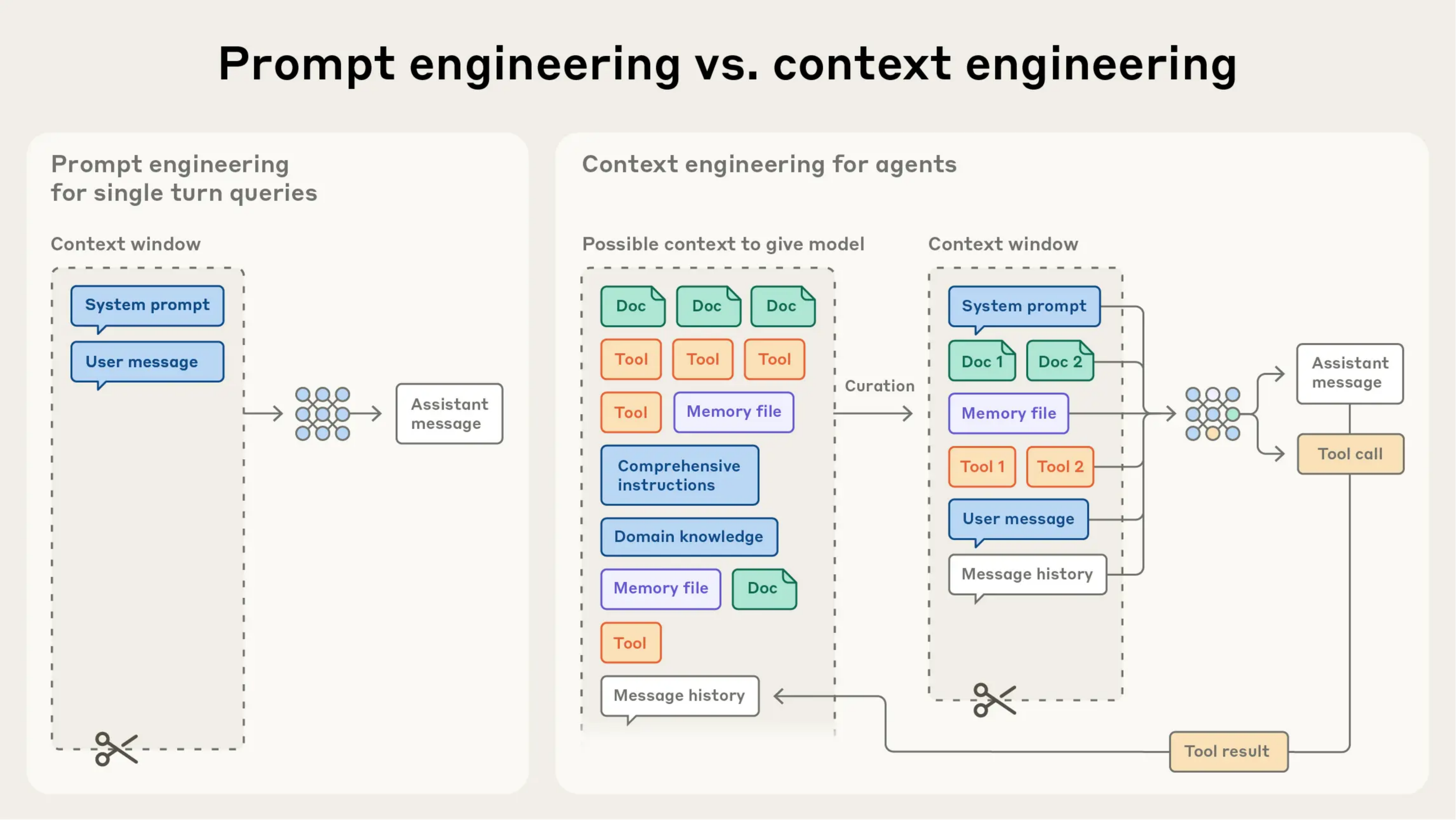
Context Engineering

Context is a critical finite resource for AI agents

- Context refers to the set of tokens when sampling from a large language model

- Engineering challenge is to optimize the utility of context tokens at any given time to respond with the desired state

- In other words, getting the LLM to understand your context and report back effectively

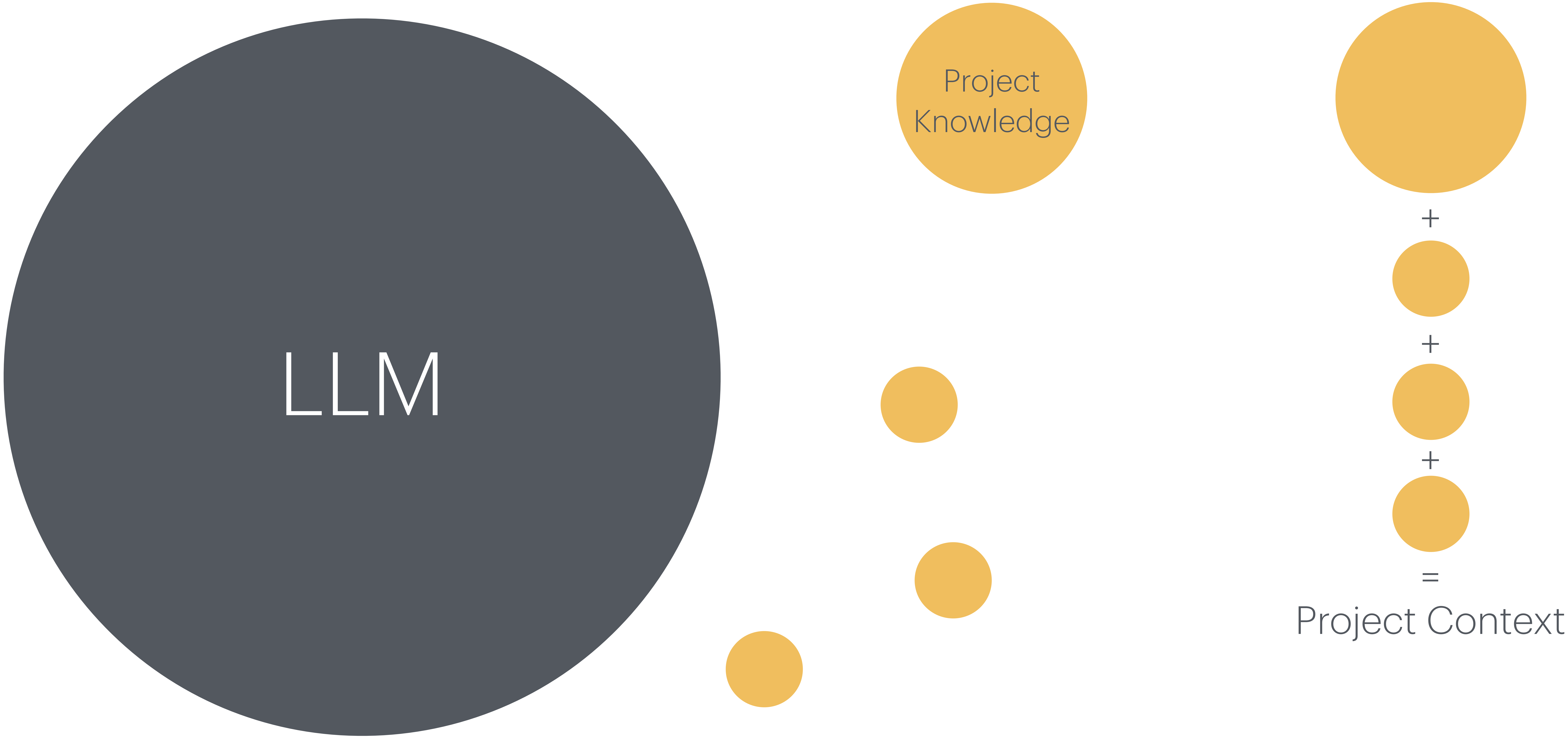


Prompt engineering vs. context engineering

https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

LLM and Context knowledge

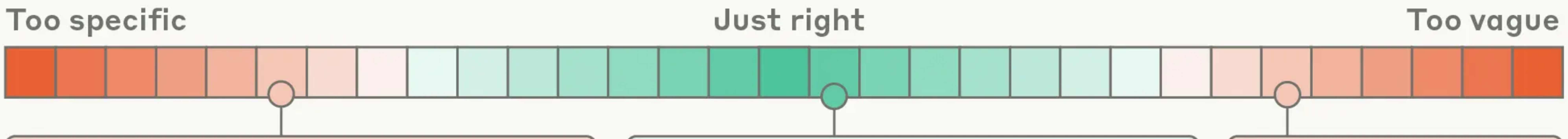
Context provides conversation specific knowledge



Why is context engineering important?

- Like humans, LLMs have limited memory capacity
- LLMs will lose focus or experience confusion at a certain point (context rot or the "dumb zone")
- This attention scarcity comes from the transformer architecture which enables every token to see every other token across all context resulting in an n^2 pairwise relationship
    - As context length increases, the models ability to capture these pairwise relationships degrades
- There is a natural tension between context size and attention focus
- Effective prompts can help "guide" context engineering (next slide)
- ***This is the currently #1 issue when working with LLM agents!***

Calibrating the system prompt



You are a helpful assistant for Claude's Bakery.
You must respond to the name Claude.
For every user request you MUST FOLLOW THESE STEPS:

1. Identify the user intent as one of the following: ["incident_resolution", "general_inquiry", "order_resubmission", "account_maintenance", "requires_escalation"]
2.
   - if user intent is "incident_resolution", ask 3 followup questions to gather information, then always call the resolve tool
   - if user intent is "general_inquiry", do not ask followup questions and answer in one shot
   - if user intent ...
   - ...
3. Here is an exhaustive list of cases that should be tagged as "requires_escalation":
   - If the intent is incident_resolution but the user is in a different country
   - If the user left a physical belonging in the store
   - ...
4. Once you've ruled out escalation scenarios you should consider all the tools at your disposal.
5. If the user_request contains an order_id you should tag the user intent as "order_resubmission", unless the user meets 5/7 of the following requirements:
   - User is asking for time update
   - User is asking for location update
   - ...
6. If the user wants to request a new order, but they already have another order in flight, you should follow these 5 steps of the resolution procedure:
   - (1) Call check_order tool to see where the current order is
   - ...

...

You are a customer support agent for Claude's Bakery. You specialize in assisting customers with their orders and basic questions about the bakery. Use the tools available to you to resolve the issue efficiently and professionally.

You have access to order management systems, product catalogs, and store policies. Your goal is to resolve issues quickly when possible. Start by understanding the complete situation before proposing solutions, ask follow-up questions if you do not understand.

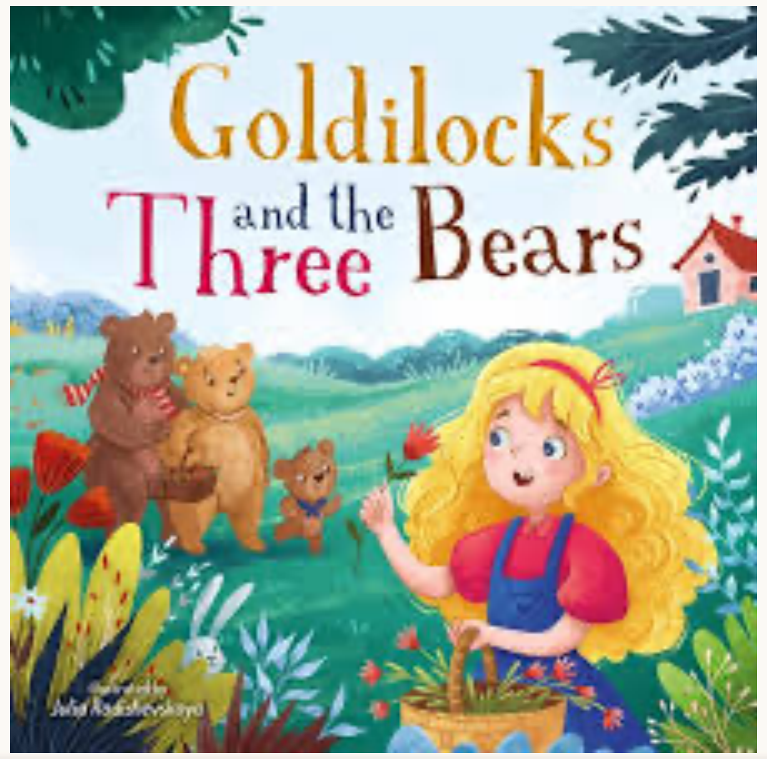
Response Framework:

1. Identify the core issue - Look beyond surface complaints to understand what the customer actually needs
2. Gather necessary context - Use available tools to verify order details, check inventory, or review policies before responding
3. Provide clear resolution - Offer concrete next steps with realistic timelines
4. Confirm satisfaction - Ensure the customer understands the resolution and knows how to follow up if needed

Guidelines:

- When multiple solutions exist, choose the simplest one that fully addresses the issue
- If a user mentions an order, check its status before suggesting next steps
- When uncertain, call the human_assistance tool
- For legal issues, health/allergy emergencies, or situations requiring financial adjustments beyond standard policies, call the human_assistance tool
- Acknowledge frustration or urgency in the user's tone and respond with appropriate empathy

You are a bakery assistant, you should attempt to solve customers issues in a manner consistent with the principles and essence of the company brand. Escalate to a human if needed.



https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

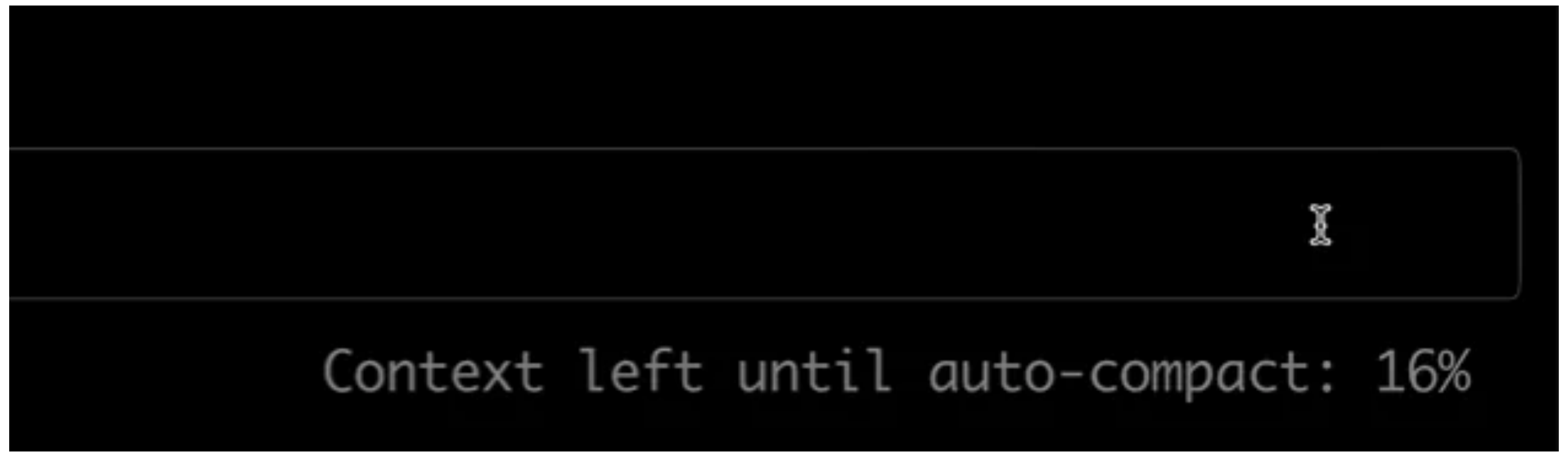
Navigating the Goldilocks Paradigm

Effective rules for context optic

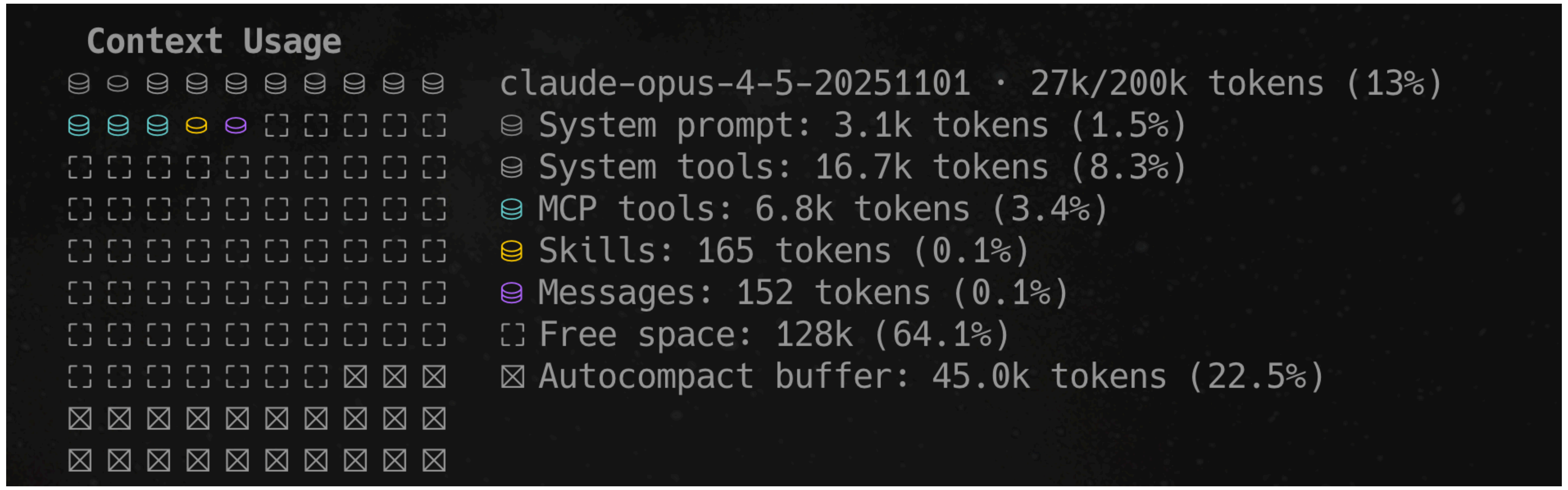
- Prompts should be extremely clear and use simple, direct language that presents at the right *altitude* for the agent
- Two extremes
    - Engineered hardcoded and designed to elicit exact agents behavior
    - Vague high-level guidance that does not provide concrete signals
- The optimal value lies somewhere in between

Context Compaction

- All agentic coding tools will enforce compaction at a set number of tokens
  - Claude code = 200k or 1m tokens
  - Gemini = 1m tokens
- You can manually compact (/compact)

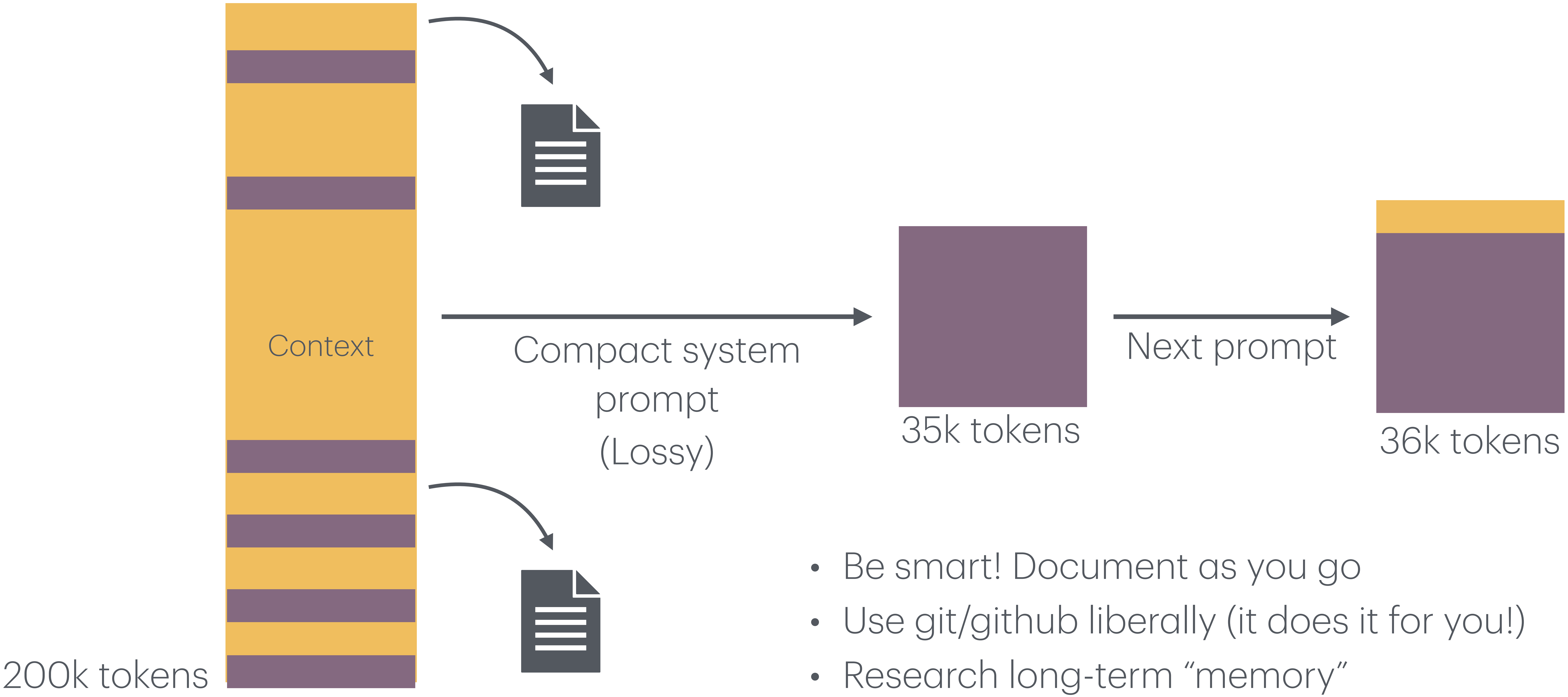


```
Context left until auto-compact: 16%
```



```
Context Usage
⛁ ⛀ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁   claude-opus-4-5-20251101 · 27k/200k tokens (13%)
⛁ ⛁ ⛁ ⛁ ⛁ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ System prompt: 3.1k tokens (1.5%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ System tools: 16.7k tokens (8.3%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ MCP tools: 6.8k tokens (3.4%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ Skills: 165 tokens (0.1%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ Messages: 152 tokens (0.1%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛶ Free space: 128k (64.1%)
⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛝ ⛝ ⛝   ⛝ Autocompact buffer: 45.0k tokens (22.5%)
⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝
⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝
```

Context Compaction



- Be smart! Document as you go
- Use git/github liberally (it does it for you!)
- Research long-term "memory"
  - Several well developed packages
  - https://github.com/steveyegge/beads

Agentic Research Summary

- Agents are just tools
  - Excel at text based tasks
  - Access to the world knowledge
  - You are the manager, the agent is an employee
    - Requires guidance, evaluation and validation
  - Agents are tools that use tools. Use tools for deterministic evaluation
- Context is queen
  - LLMs have attention spans
  - Effective context management keep your agents smart
    - Most "bad experiences" with LLMs come from ineffective context management

Anthropic Academy

Get in the know with Anthropic resources. From API development guides to enterprise deployment best practices, the academy has you covered.

Featured courses

New courses available on Anthropic Academy. Learn more in-depth about AI Fluency, API development, Model Context Protocol and Claude Code. Earn certificates upon completion.

https://www.anthropic.com/learn

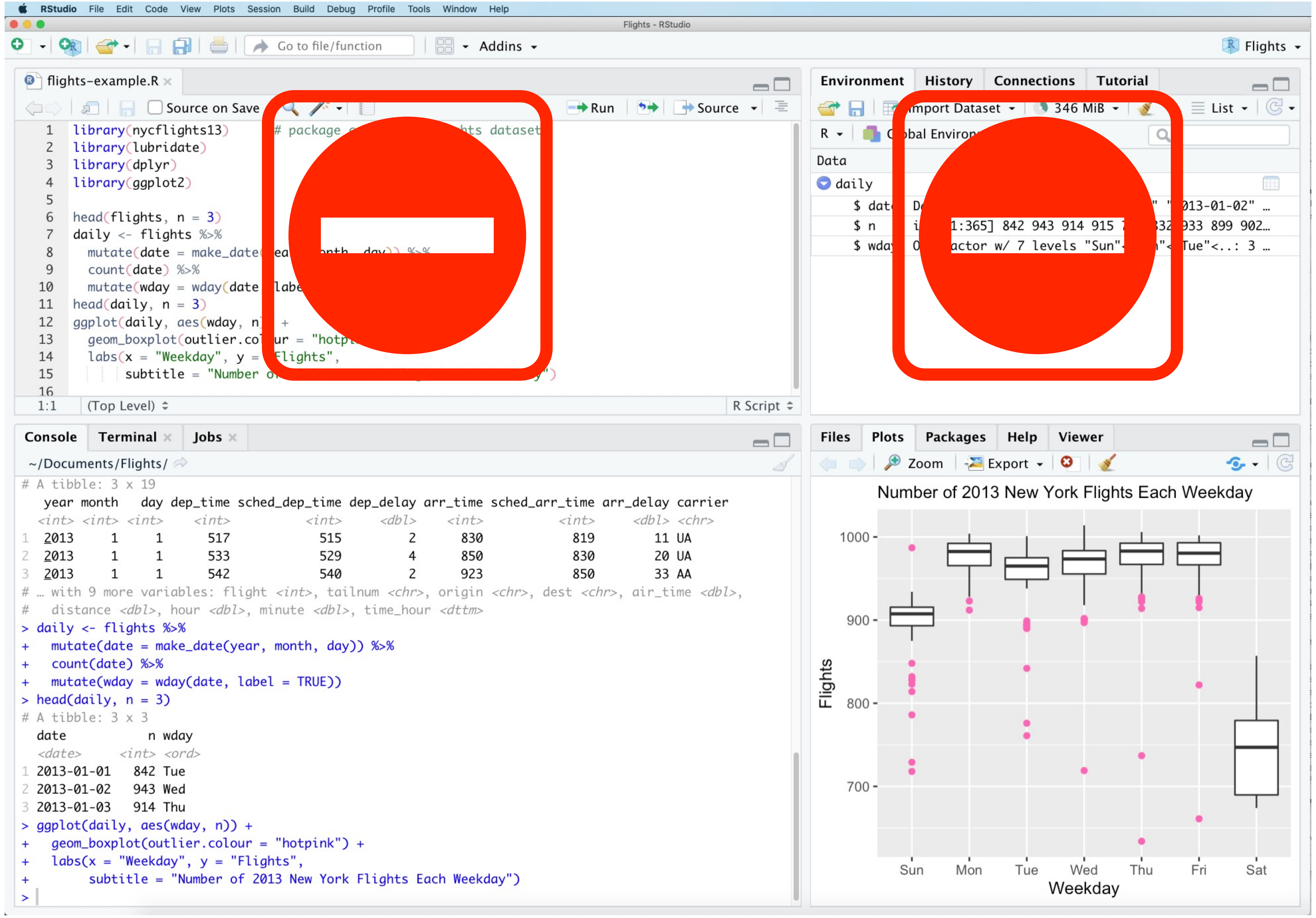
Where are things going

An AI IDE for bioinformatics and data science

Minimal Agentic Research Requirements

What is really needed for data analysis?

- Documentation?
- Terminal?
- File manager?
- Data viewer / plot window?

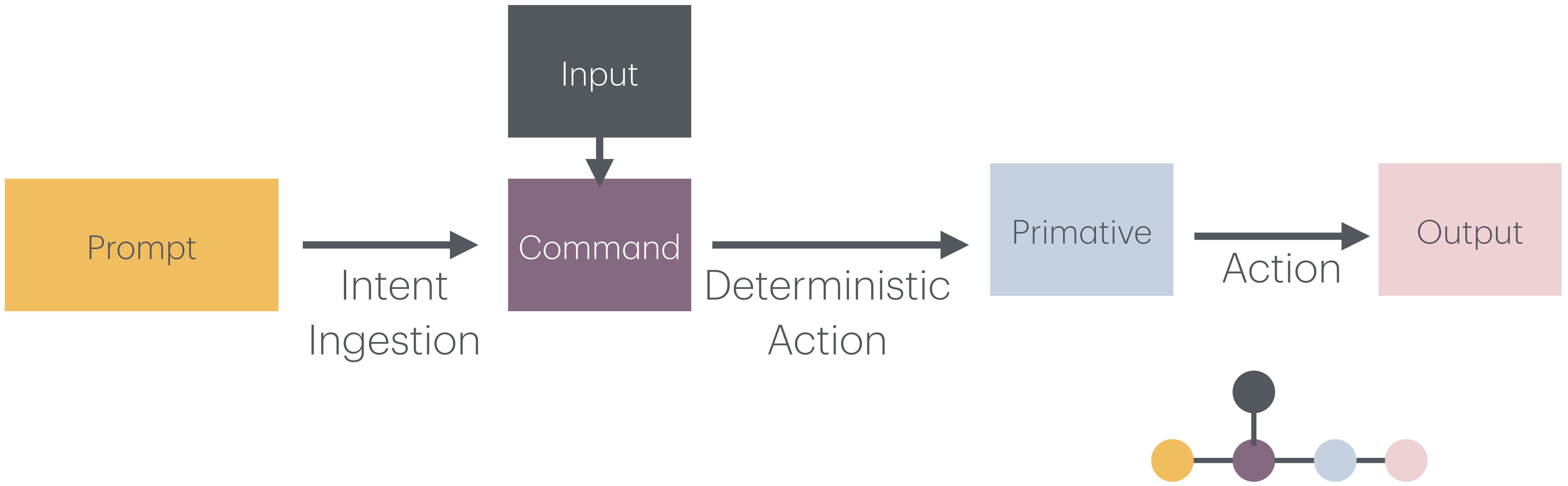


```
library(nycflights13)        # package                hts dataset
library(lubridate)
library(dplyr)
library(ggplot2)

head(flights, n = 3)
daily <- flights %>%
  mutate(date = make_date(
  count(date) %>%
  mutate(wday = wday(date
head(daily, n = 3)
ggplot(daily, aes(wday, n
  geom_boxplot(outlier.co
  labs(x = "Weekday", y =
       subtitle = "Number o
```

~/Documents/Flights/

```
# A tibble: 3 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
1  2013     1     1      517            515         2      830            819        11 UA
2  2013     1     1      533            529         4      850            830        20 UA
3  2013     1     1      542            540         2      923            850        33 AA
# … with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
> daily <- flights %>%
+   mutate(date = make_date(year, month, day)) %>%
+   count(date) %>%
+   mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
# A tibble: 3 x 3
  date           n wday
  <date>     <int> <ord>
1 2013-01-01   842 Tue
2 2013-01-02   943 Wed
3 2013-01-03   914 Thu
> ggplot(daily, aes(wday, n)) +
+   geom_boxplot(outlier.colour = "hotpink") +
+   labs(x = "Weekday", y = "Flights",
+        subtitle = "Number of 2013 New York Flights Each Weekday")
> 
```

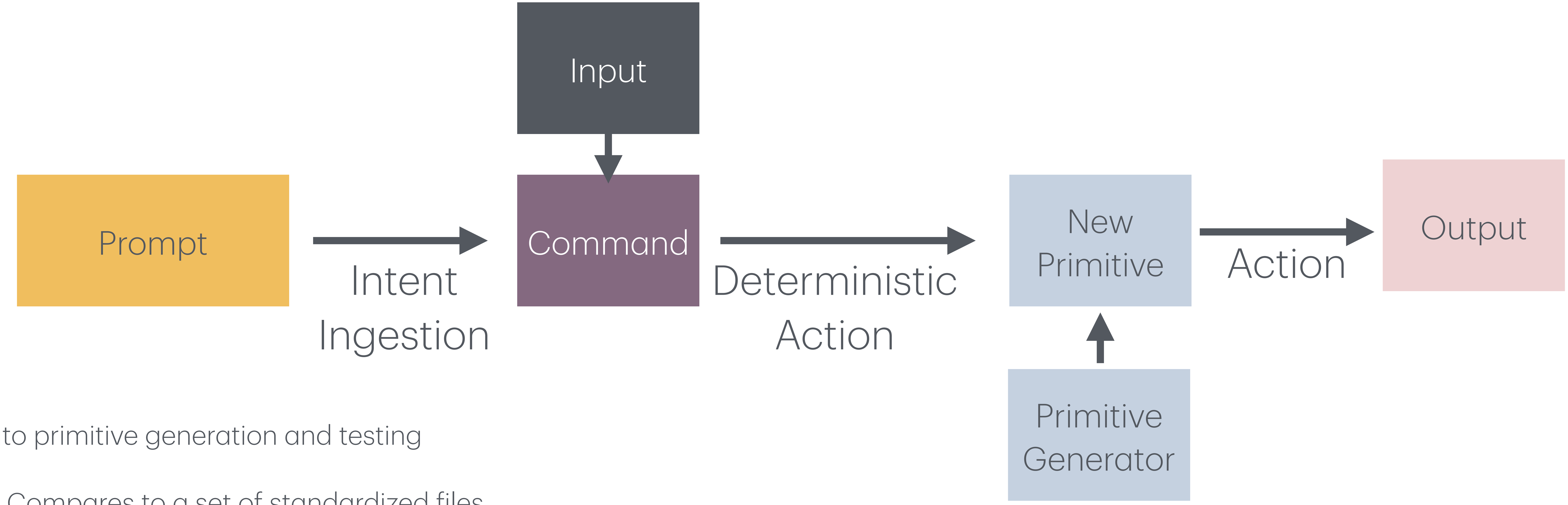
Agentic Harness for Reproducible Research



- Instructing the LLM to run and track deterministic primitives
- Primitives: A single, focused statistical or plotting operation. One building block that does exactly one thing
  - Calculate GC content
  - Clr: transform counts to log-ratios
  - PCA: Reduce dimensions for visualization

LLM Generated Primitives

What happens if a primitive doesn't exist?

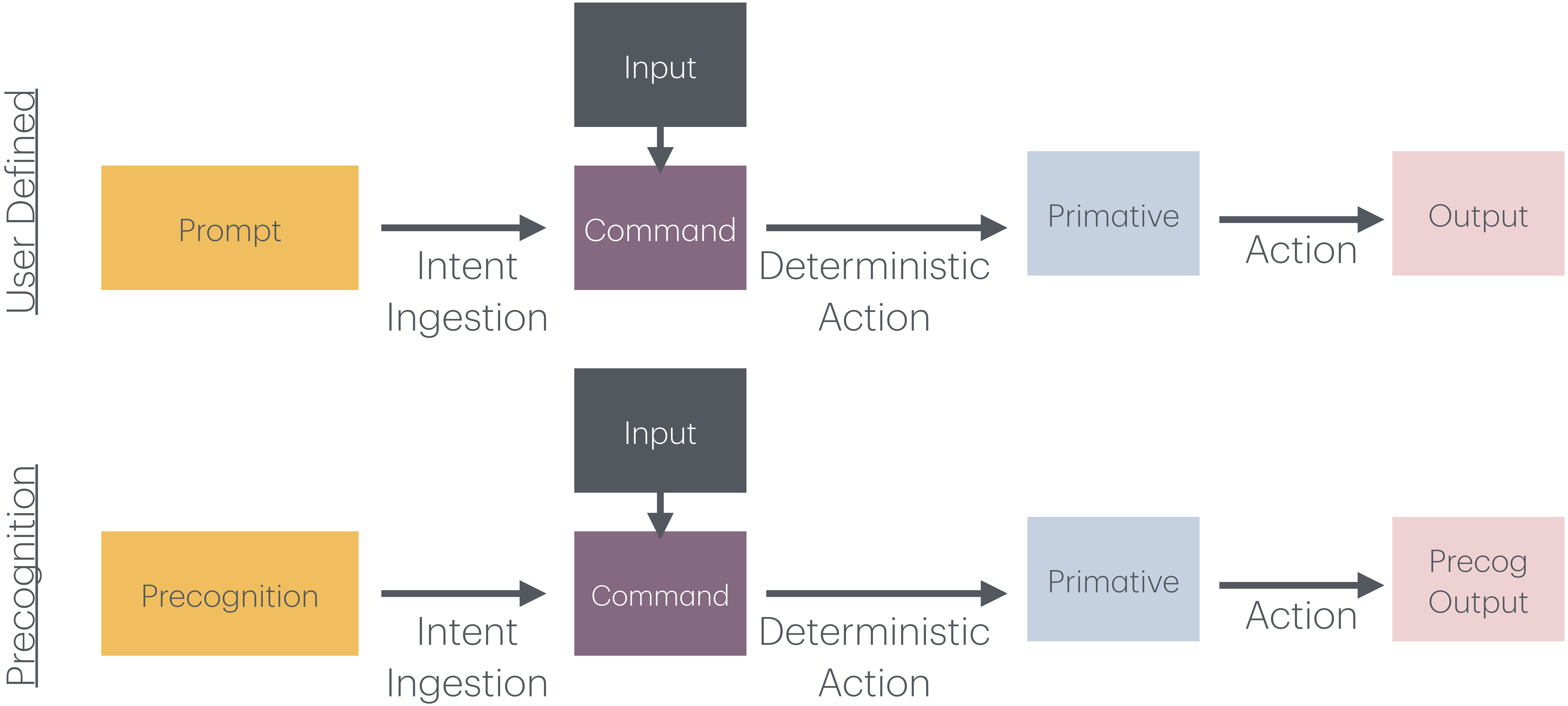


- Auto primitive generation and testing
  - Compares to a set of standardized files
- Bioanvil: test and data validation framework
  - Synthetic ground truth data sets
- Can also download and validate against other 'classical' software
- Primitive creation is implemented and benchmarked by AI

Analytical Precognition

- LLMs are smart and can predict (precog) many analytical approaches
- Example:
  - *"I want to compare differential gene expression between health and disease groups"*
    - DeSeq2 is run comparing healthy to disease
      - Output: csv, volcano plot, interpretation
    - Background precognition:
      - LLM understands your metadata and that samples are paired and runs LIMMA to test for random effects
      - Output: csv, volcano plot, interpretation
    - LLM compares and summarizes outputs of both

Analytical Precognition



User Defined

Input

Prompt → Intent Ingestion → Command → Deterministic Action → Primative → Action → Output

Precognition

Input

Precognition → Intent Ingestion → Command → Deterministic Action → Primative → Action → Precog Output

Analytical Precognition

- Precognition can run in the background when resources are available
- Allows LLM guided exploration of data
- Allows reporting and remixing
  - User: "I wonder what these data would have looked like if we would have applied a CLR transform?"
  - Agent: Pulls precoged CLR transform data and retraces analysis graph summarizing results
- Live analytical human data interrogation and review
  - Slide showing analysis of Shannon diversity at a seminar or lab meeting
  - Someone asks, *"How would this have looked if you analyzed with Simpsons diversity?"*
  - Precognition already completed (or possibly spun up on the fly) and slide updates from Shannons to Simpsons diversity

Biostack example

Stenographic Intent Ingestion

Intent Ingestion

- LLMs interpert “intent”
- Intent can be images or text
- Text can be provided by typing or voice
- LLMs are great and interpreting the intent of even messy voice and converting to text based intent

Stenograph

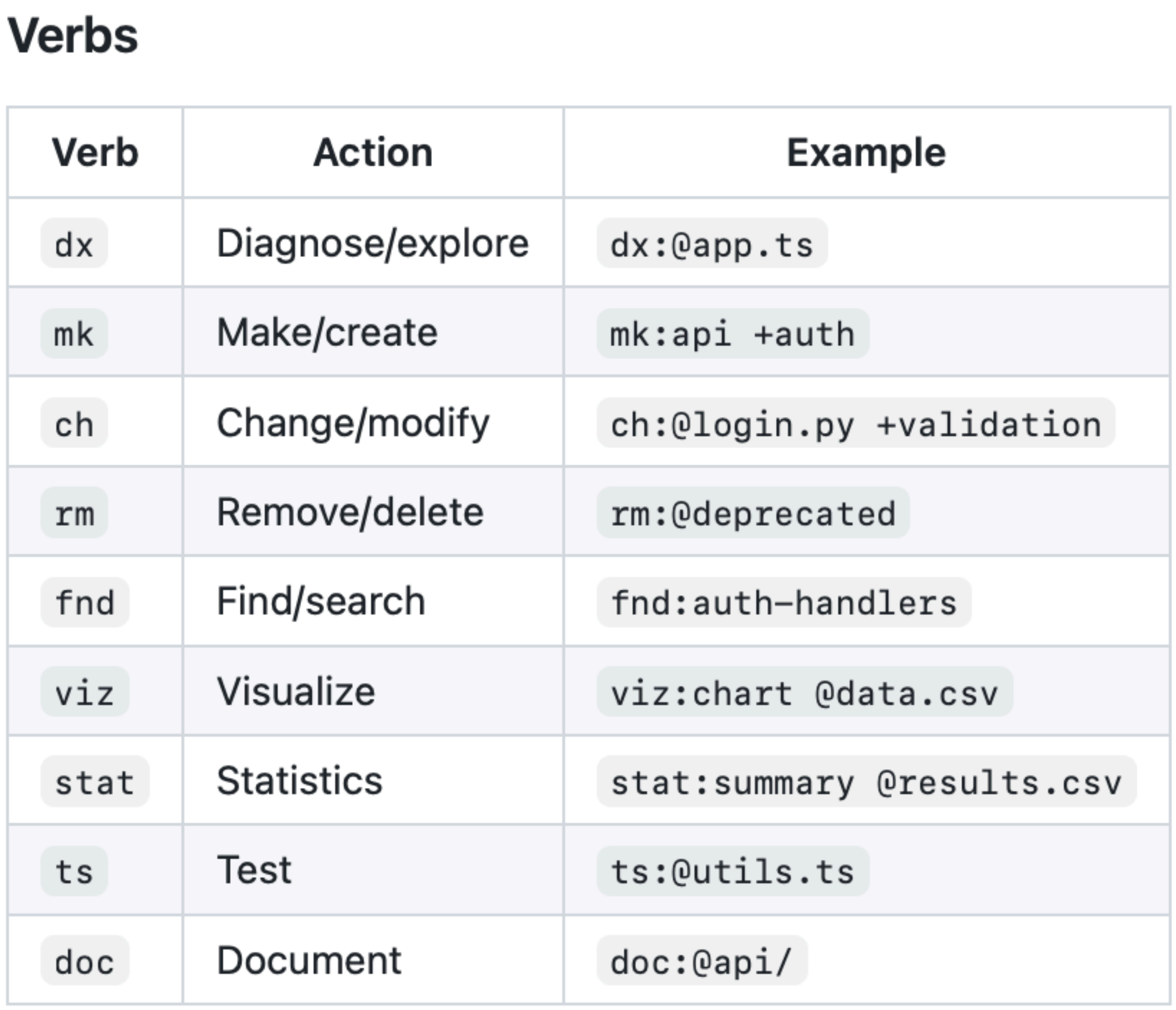
Stenographic grammar for human-AI collaboration

Stenograph

- Court stenographers capture up to 225+ words per minute using compressed, structured input
- Instead of, "Please look at the README.txt file and add error handling to all the async functions” you would just: dx:@README.md
- Provides a programmatic traceable layer to prompt ingestion by LLMs

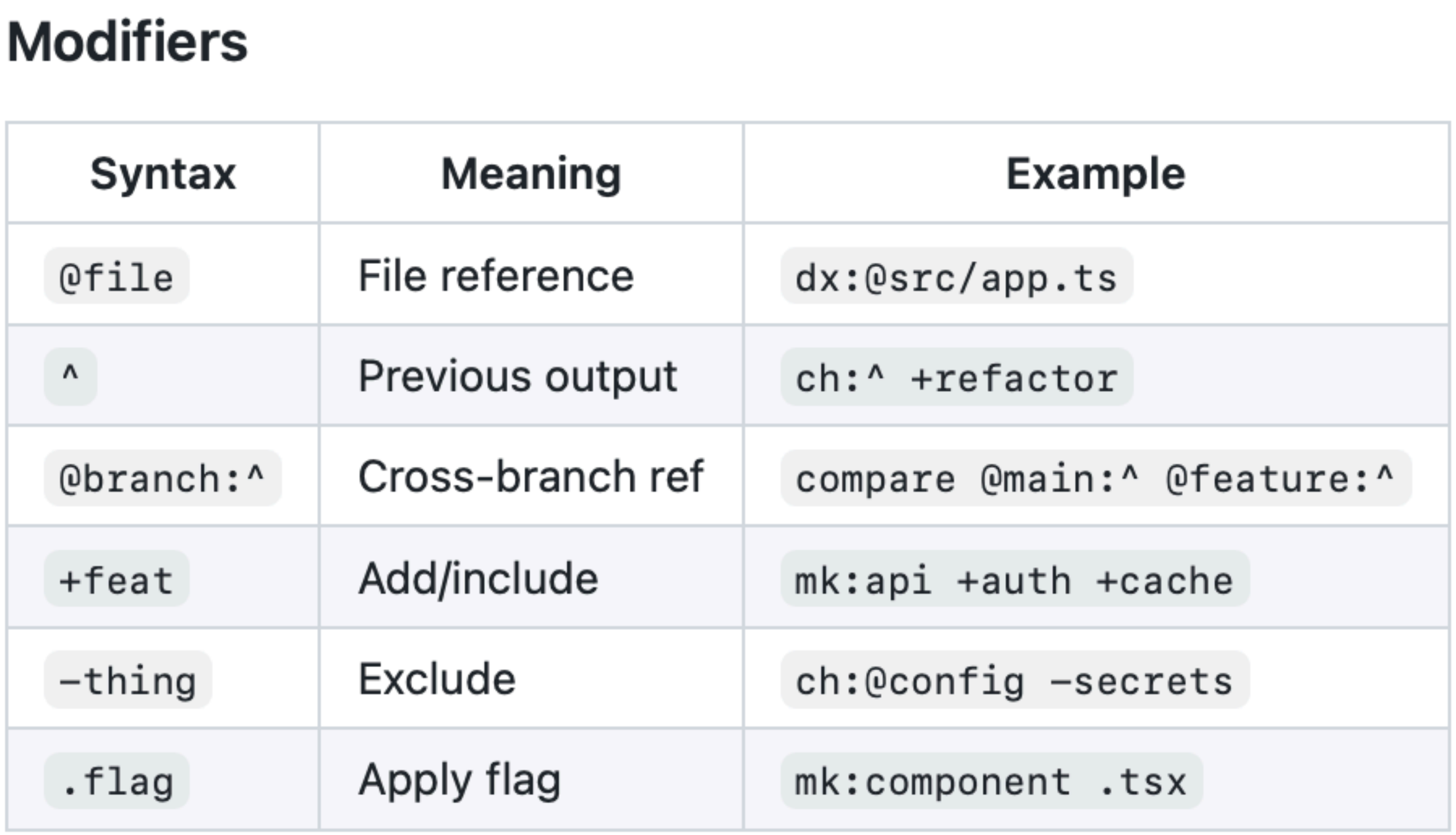
Stenograph

- Standardized syntax model:
  - [mode][verb]:[target] [@refs] [+add] [-exclude] [.flag] [precision]



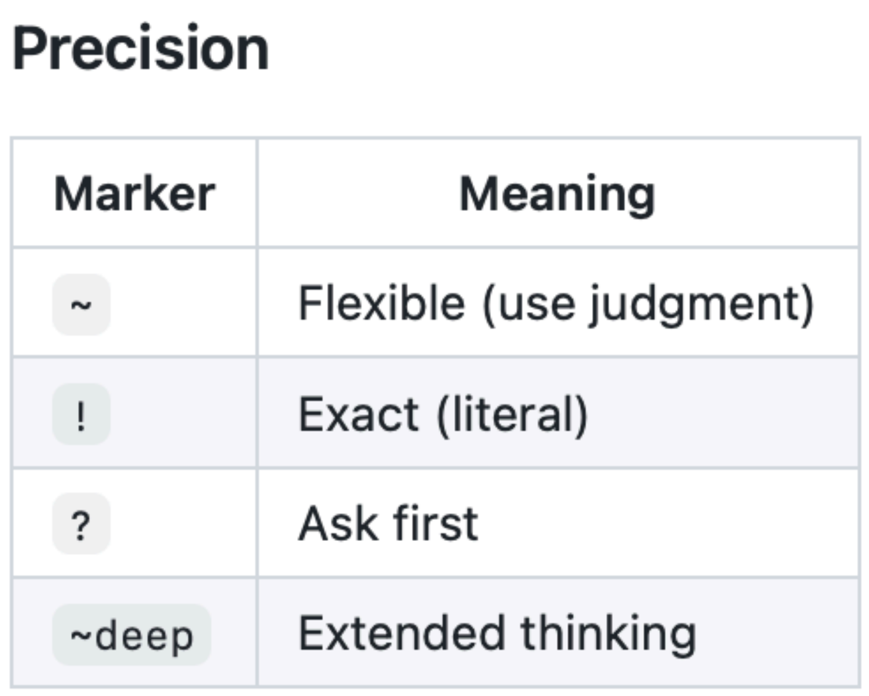
Verbs

Verb	Action	Example
`dx`	Diagnose/explore	`dx:@app.ts`
`mk`	Make/create	`mk:api +auth`
`ch`	Change/modify	`ch:@login.py +validation`
`rm`	Remove/delete	`rm:@deprecated`
`fnd`	Find/search	`fnd:auth-handlers`
`viz`	Visualize	`viz:chart @data.csv`
`stat`	Statistics	`stat:summary @results.csv`
`ts`	Test	`ts:@utils.ts`
`doc`	Document	`doc:@api/`



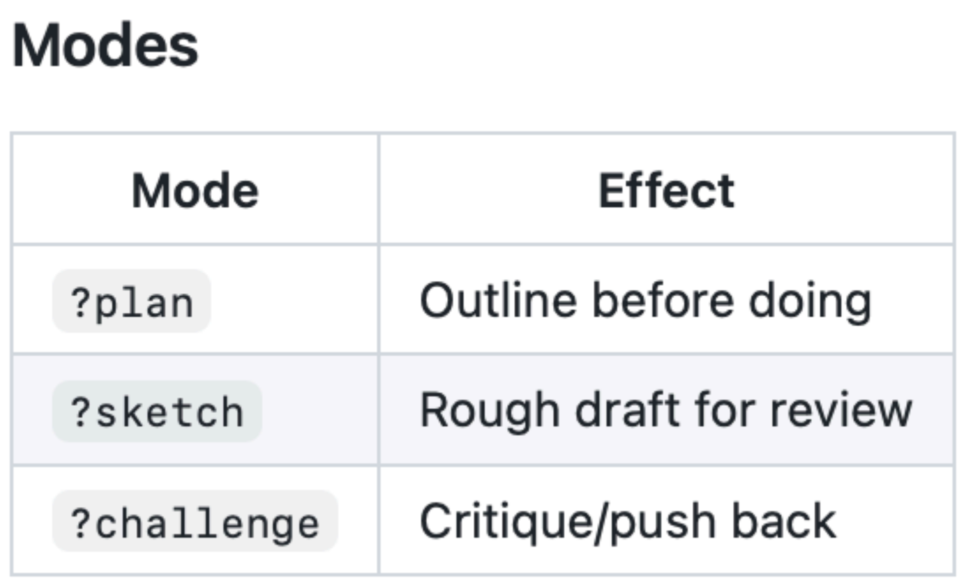
Modifiers

Syntax	Meaning	Example
`@file`	File reference	`dx:@src/app.ts`
`^`	Previous output	`ch:^ +refactor`
`@branch:^`	Cross-branch ref	`compare @main:^ @feature:^`
`+feat`	Add/include	`mk:api +auth +cache`
`-thing`	Exclude	`ch:@config -secrets`
`.flag`	Apply flag	`mk:component .tsx`



Precision

Marker	Meaning
`~`	Flexible (use judgment)
`!`	Exact (literal)
`?`	Ask first
`~deep`	Extended thinking



Modes

Mode	Effect
`?plan`	Outline before doing
`?sketch`	Rough draft for review
`?challenge`	Critique/push back

Stenograph example

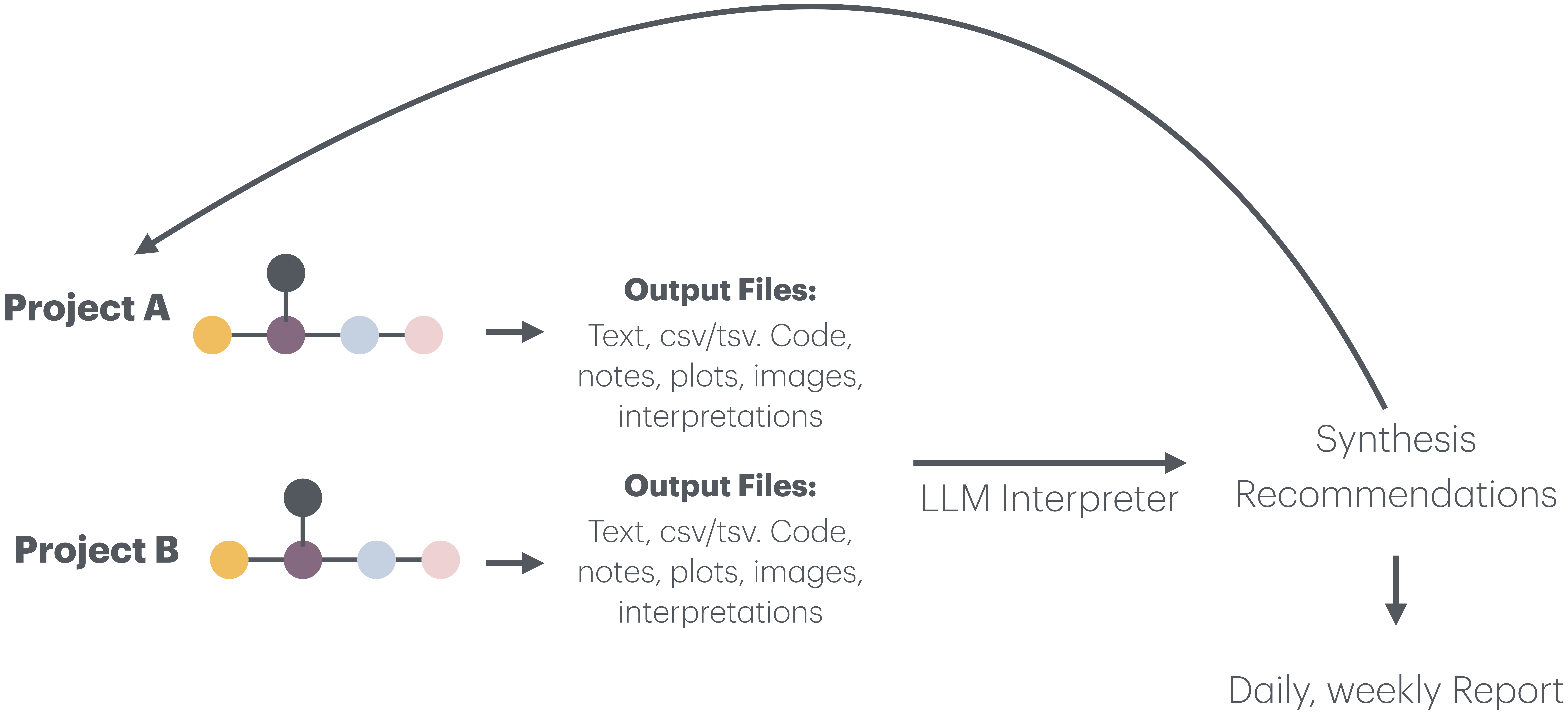
What does stenographer solve

Intent ingestion

- Consistant programmatic execution
- Agentic Research is ephemeral
  - Chat history scrolls away
  - No record of what was tried vs. what worked
  - Can't compare alternative approaches
  - Context lost between sessions

- The git solution
  - Each command becomes a tracked unit
    - Command, input, output, summary, timestamp, branch
  - Explore without fear with branching
  - Side-by-side comparisons to compare alternative approaches

Federated Research



Project A

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

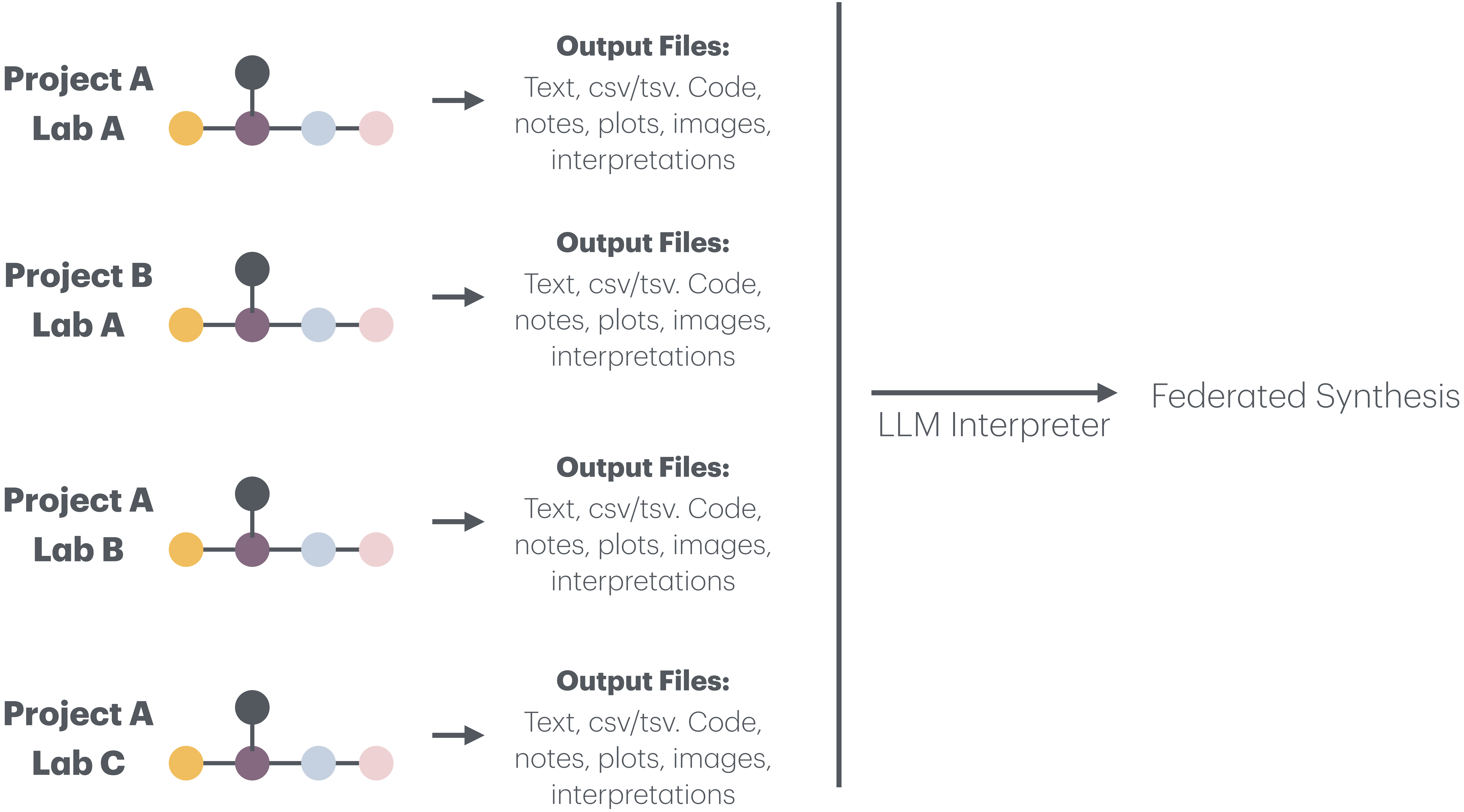
Project B

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

LLM Interpreter

Synthesis
Recommendations

Daily, weekly Report



Project A
Lab A

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

Project B
Lab A

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

Project A
Lab B

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

Project A
Lab C

Output Files:
Text, csv/tsv. Code, notes, plots, images, interpretations

LLM Interpreter

Federated Synthesis

Links

- Tokenizer: https://platform.openai.com/tokenizer
- Embeddings: https://huggingface.co/spaces/hesamation/primer-llm-embedding?section=what_are_embeddings?
- Hugging Face: https://huggingface.co/
- Transformers: https://poloclub.github.io/transformer-explainer/
- Claude system prompts: https://github.com/Piebald-AI/claude-code-system-prompts
- Claude soul document: https://gist.github.com/Richard-Weiss/efe157692991535403bd7e7fb20b6695#file-opus_4_5_soul_document_cleaned_up-md
- Claude developer docs: https://platform.claude.com/docs/en/home
- LLM friendly context documentation: https://context7.com/
- Pricer per token per model: https://yourgpt.ai/tools/openai-and-other-llm-api-pricing-calculator
- Context engineering: https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents
- Ahthropic Academy: https://www.anthropic.com/learn
- Beads (memory layer): https://github.com/steveyegge/beads

Questions?