

# BUILDING GOOD LLM PRODUCTS

WITH THE THEORY OF CONSTRAINTS

STANFORD CS 224G

FEBRUARY 5 2026

ANDY BROMBERG

ANDY@ANDYBROMBERG.COM

TINKERING: ANDYBROMBERG.COM · INVESTING: AMITY.VC

YOU CAN JUST  
BUILD THINGS

BUT... HOW DO YOU MAKE  
THEM GOOD?

# WHAT WE'LL COVER

1 THE ISSUE WITH LLMS

2 THE THEORY OF CONSTRAINTS

3 EVALS & HOW TO IMPROVE

4 SOME PRACTICAL EXAMPLES  
AND BONUS: TOOLS I CAN'T LIVE WITHOUT TODAY

WHO AM I?

# ANDY BROMBERG

**amity**ventures

INVESTOR

Partner @ Amity Ventures  
Leading pre-seed through series A

 **COINLIST**  **eco**

FOUNDER

Fmr. founder & CEO, CoinList  
Fmr. founder & CEO, Eco

**interface0**   **Lightwork  
Home Health** **S.I.S.**

ENGINEER &  
TINKERER

interface0 — AI interface for power users & teams  
S.I.S. — large enterprise AI implementation  
Lightwork Home Health — home health testing

 **Stanford  
University**

ALUMNUS

Was class of ‘16; “leave of absence” in ‘14  
Mathematics & Computer Science

# THE BIG ISSUE WITH BUILDING ON LLMS?

# THEY ARE NON-DETERMINISTIC

☰ ChatGPT 5.2 Instant ▾

⋮

what's the capital of the USA?

That'd be Washington, D.C. 🇺🇸

Easy one—but it *does* trip people up since it's not in any state.

📄 🍷 🗣️ ⬆️ ↺ ⋮

☰ ChatGPT 5.2 Instant ▾

⋮

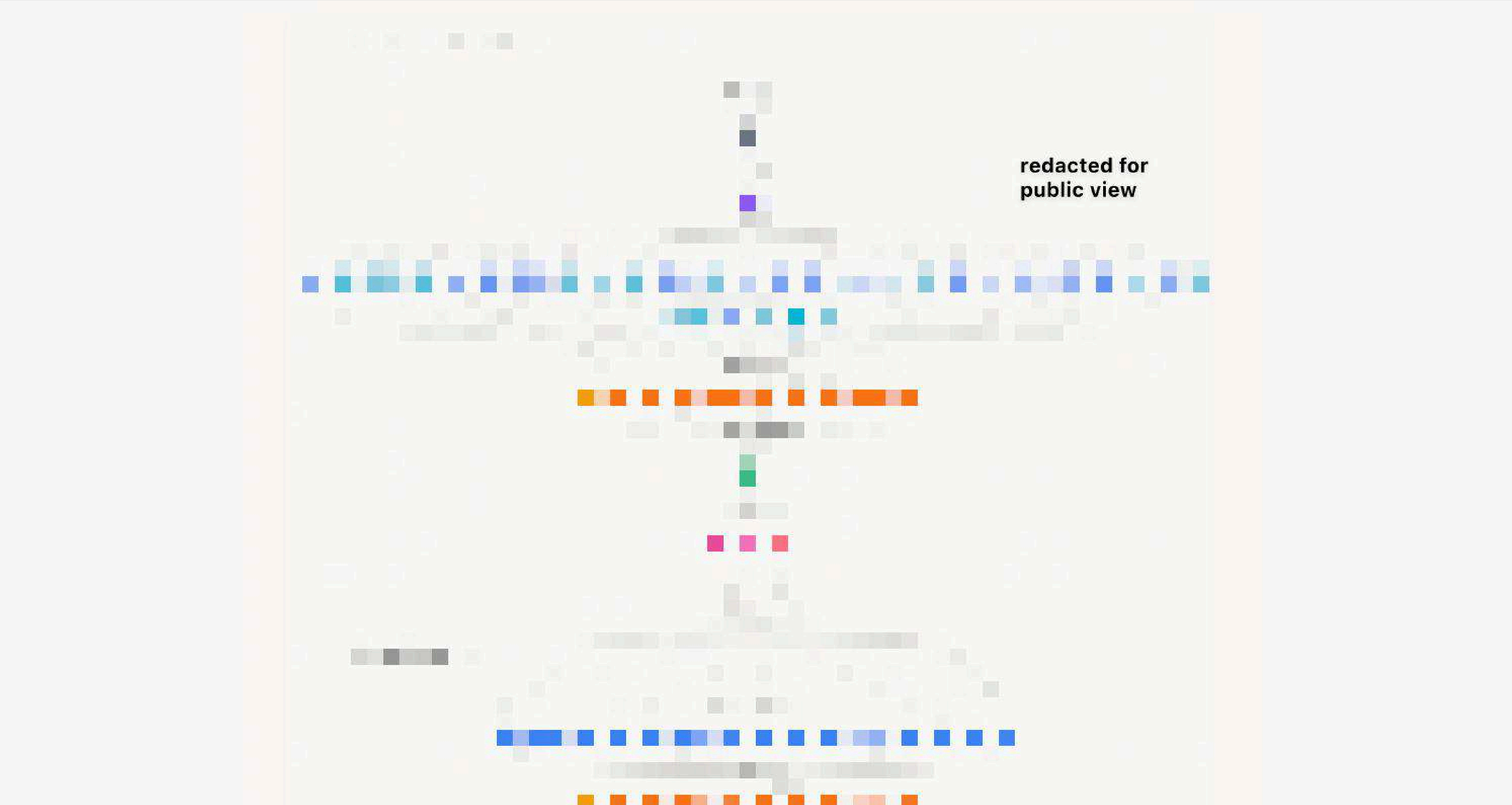
what's the capital of the USA?

The capital of the United States is Washington, D.C..

(And yep—*D.C.* stands for *District of Columbia*, not a state.)

📄 🍷 🗣️ ⬆️ ↺ ⋮

# AND COMPLEXITY COMPOUNDS NON-DETERMINISM



COMPOUNDING NON-DETERMINISM CAUSES  
QUALITY AND RELIABILITY ISSUES

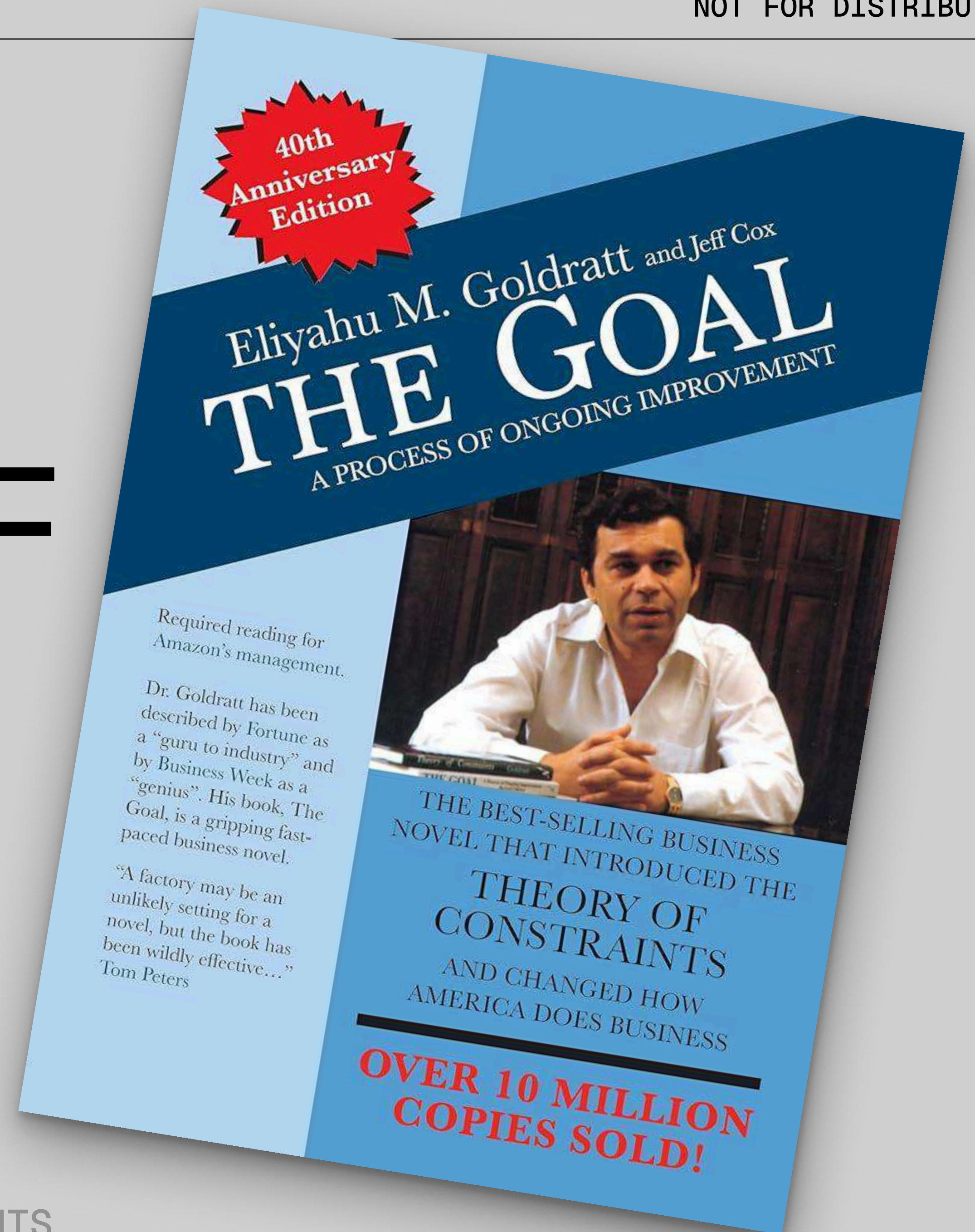
A USEFUL TOOL FOR SUCH ISSUES IS

**SYSTEMS THINKING**



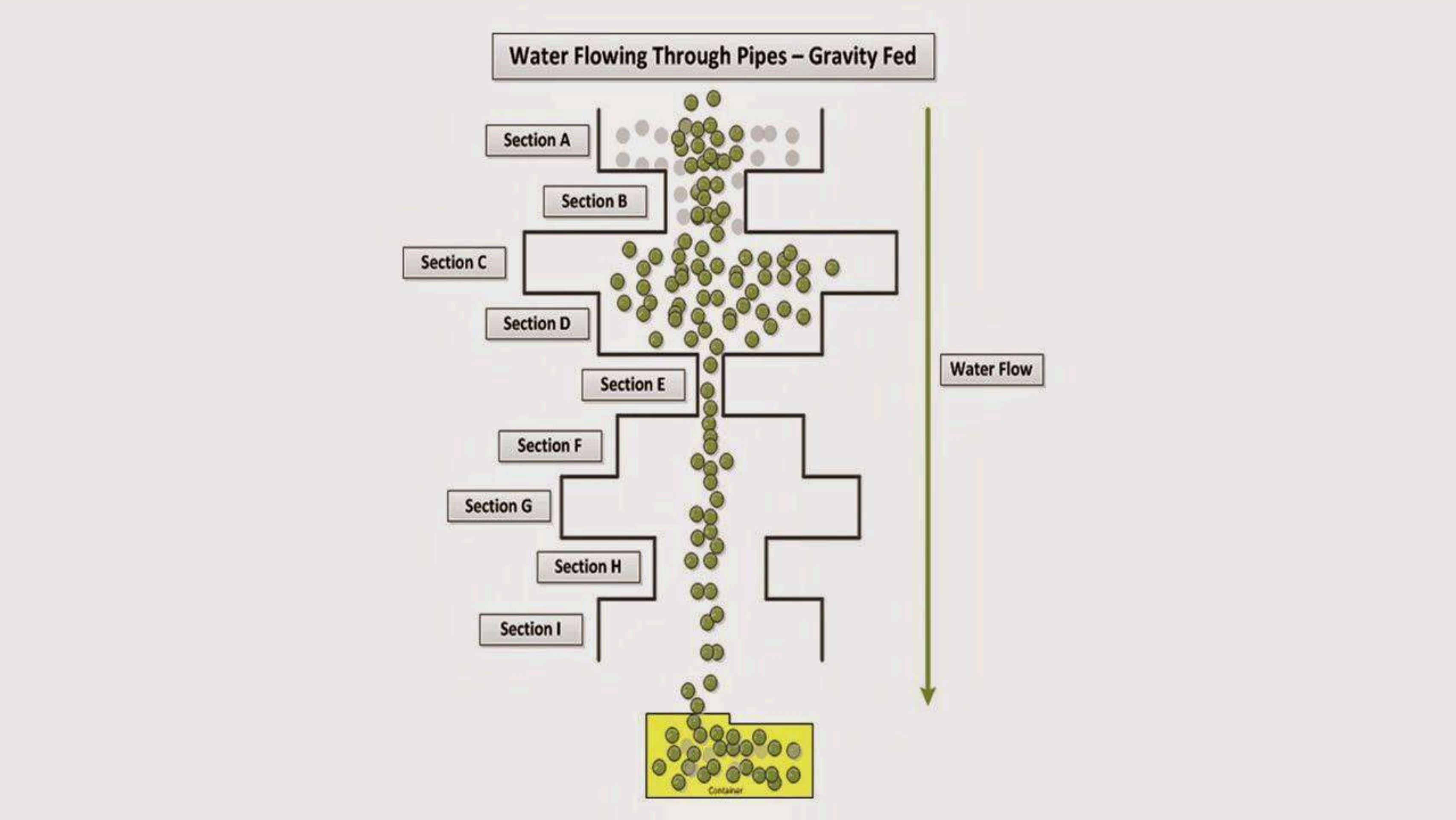
# TODAY'S FRAMEWORK: THE THEORY OF CONSTRAINTS

(YES: A 1980S SALES-Y LOOKING  
BOOK ON FACTORY MANAGEMENT.  
BEAR WITH ME ... )

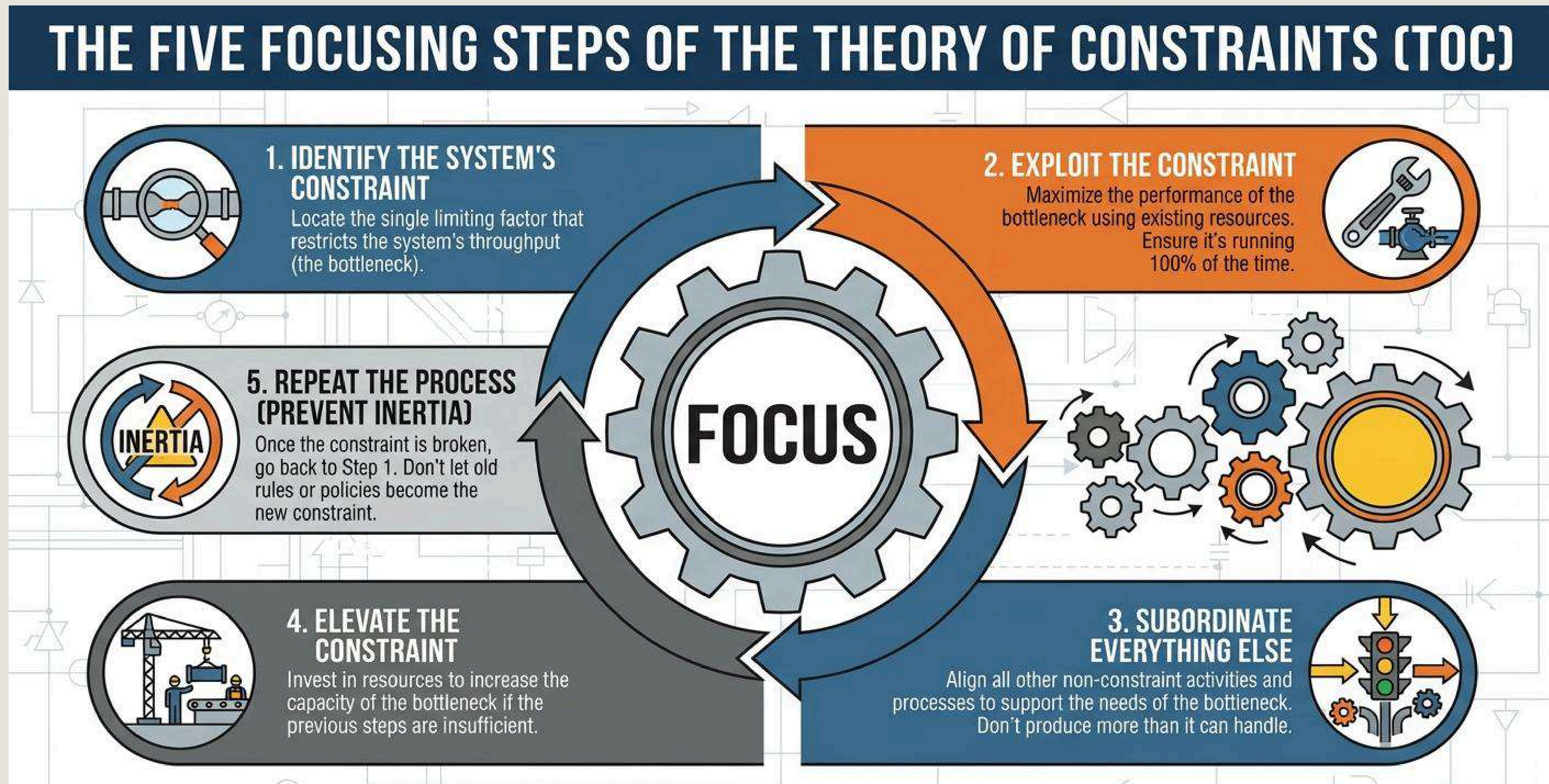




# TOC IN TWO IMAGES: THE WHAT

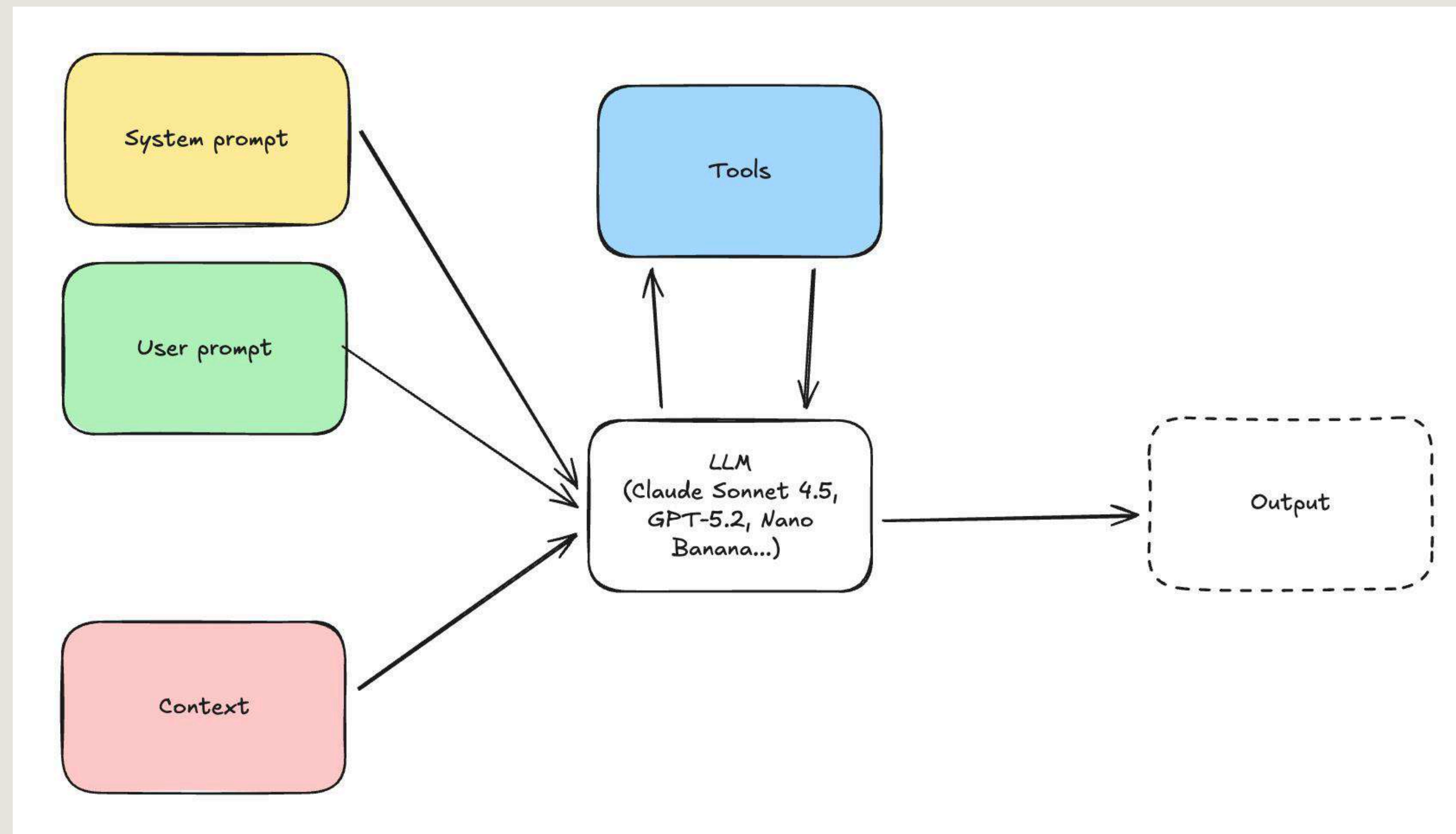


# TOC IN TWO IMAGES: THE HOW



# LET'S MODEL THE SYSTEM

## GOAL: CONSISTENT HIGH QUALITY OUTPUTS



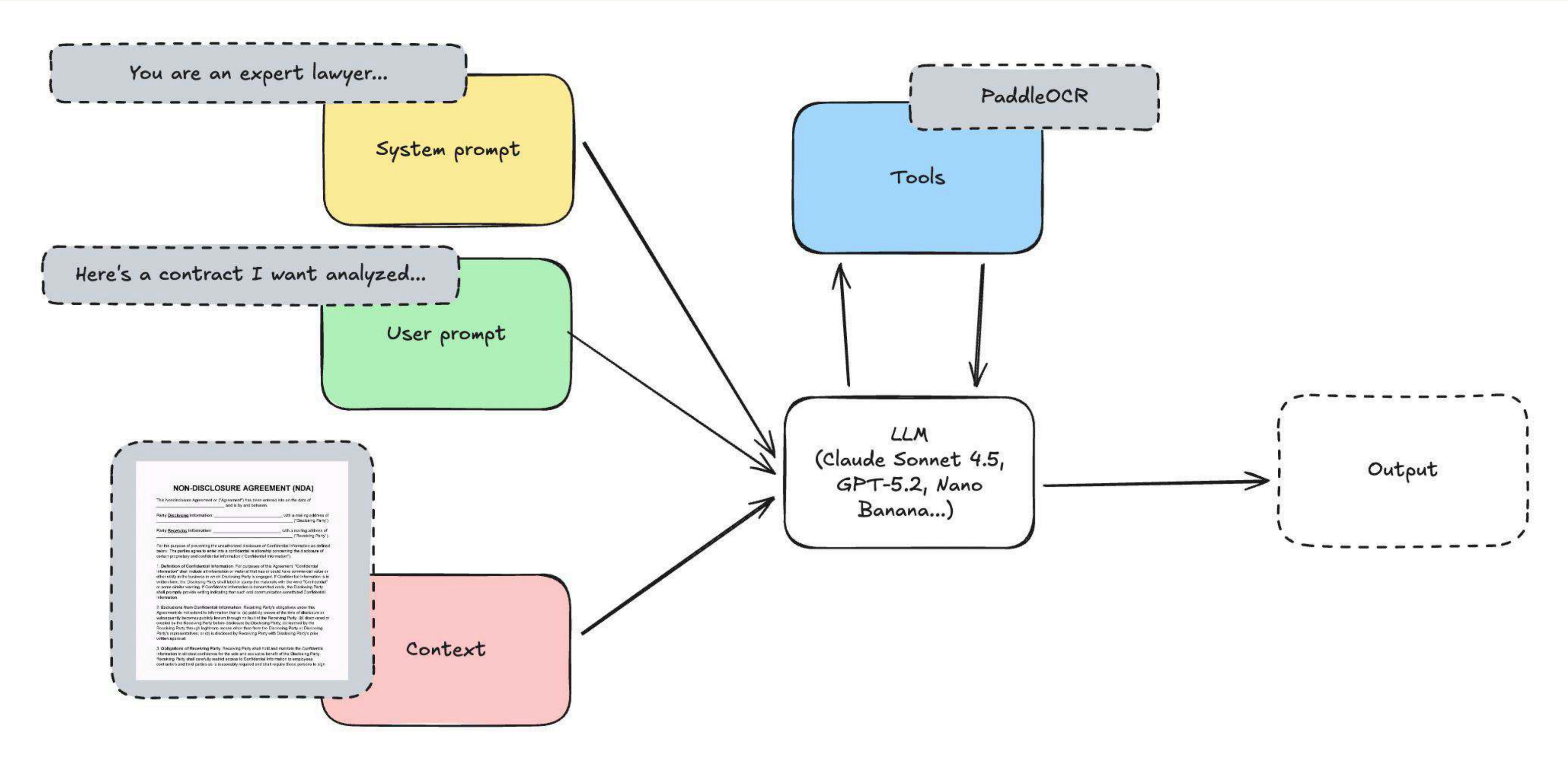
[HTTPS://ANDYBROMBERG.COM/FIELD-GUIDE-CONTEXT-ENGINEERING](https://andybromberg.com/field-guide-context-engineering)

[HTTPS://ANDYBROMBERG.COM/CONSTRAINTS-AI](https://andybromberg.com/constraints-ai)



# LET'S MODEL THE SYSTEM

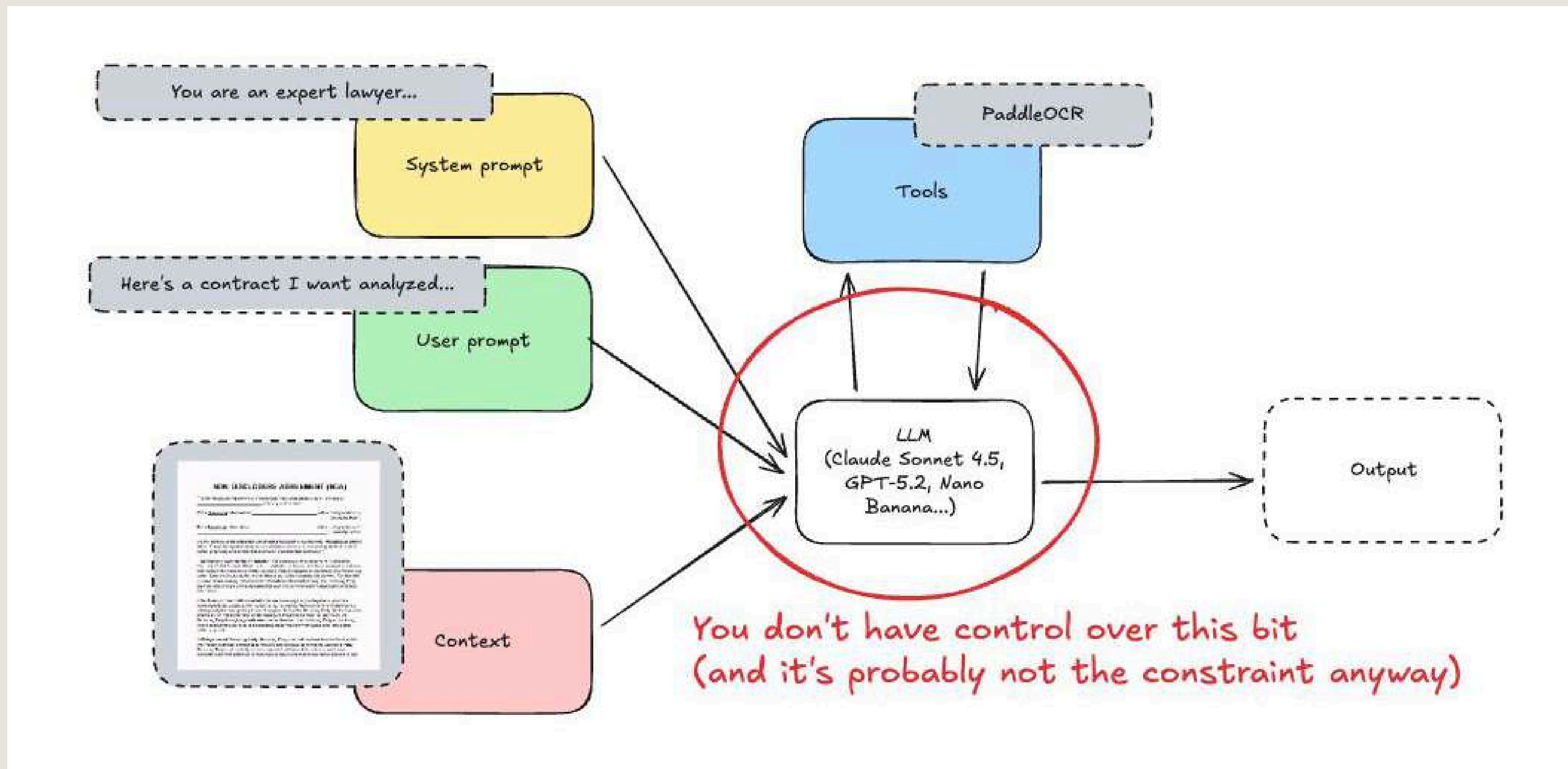
GOAL: CONSISTENT HIGH QUALITY OUTPUTS





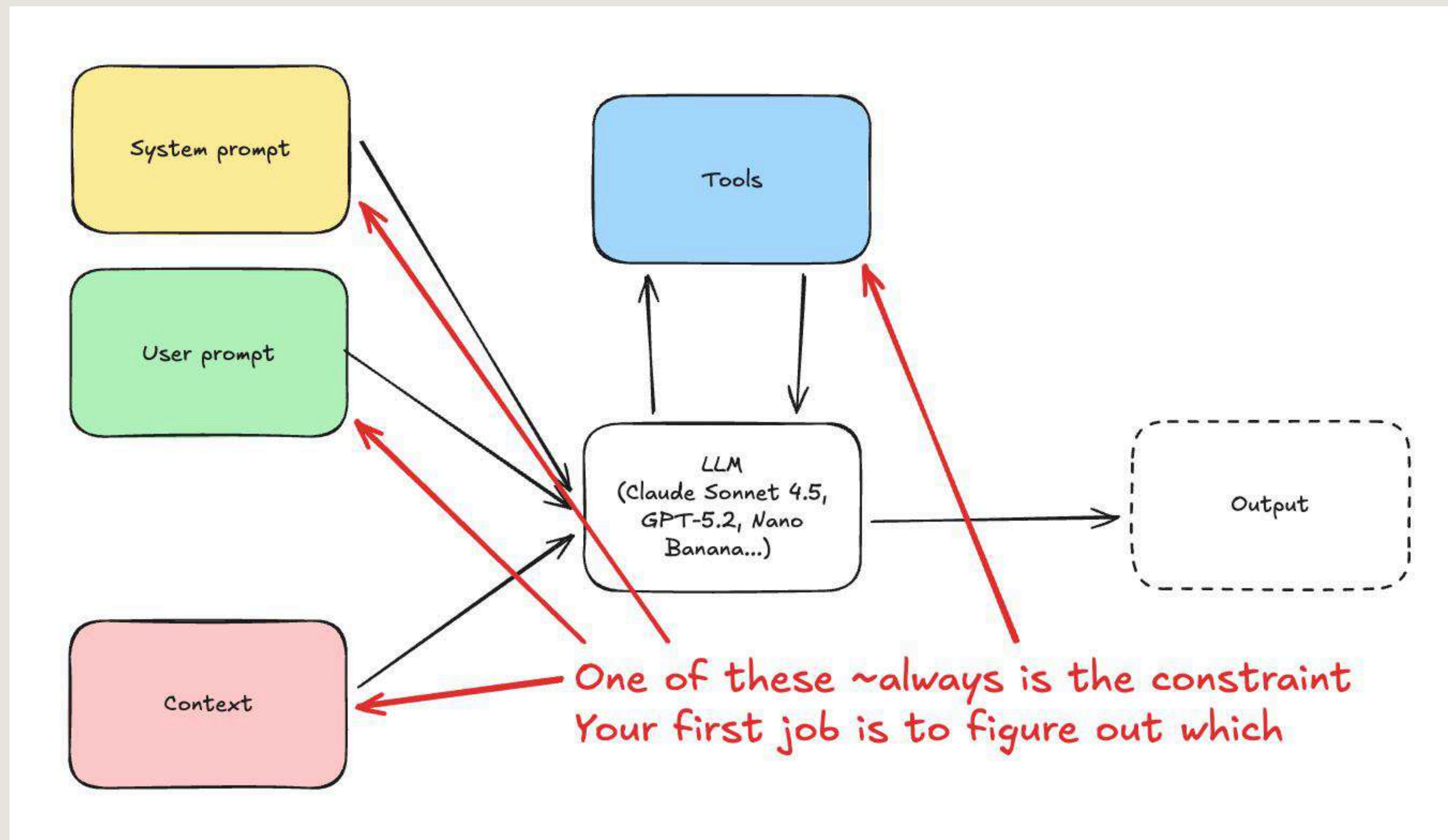
# LET'S MODEL THE SYSTEM

## GOAL: CONSISTENT HIGH QUALITY OUTPUTS



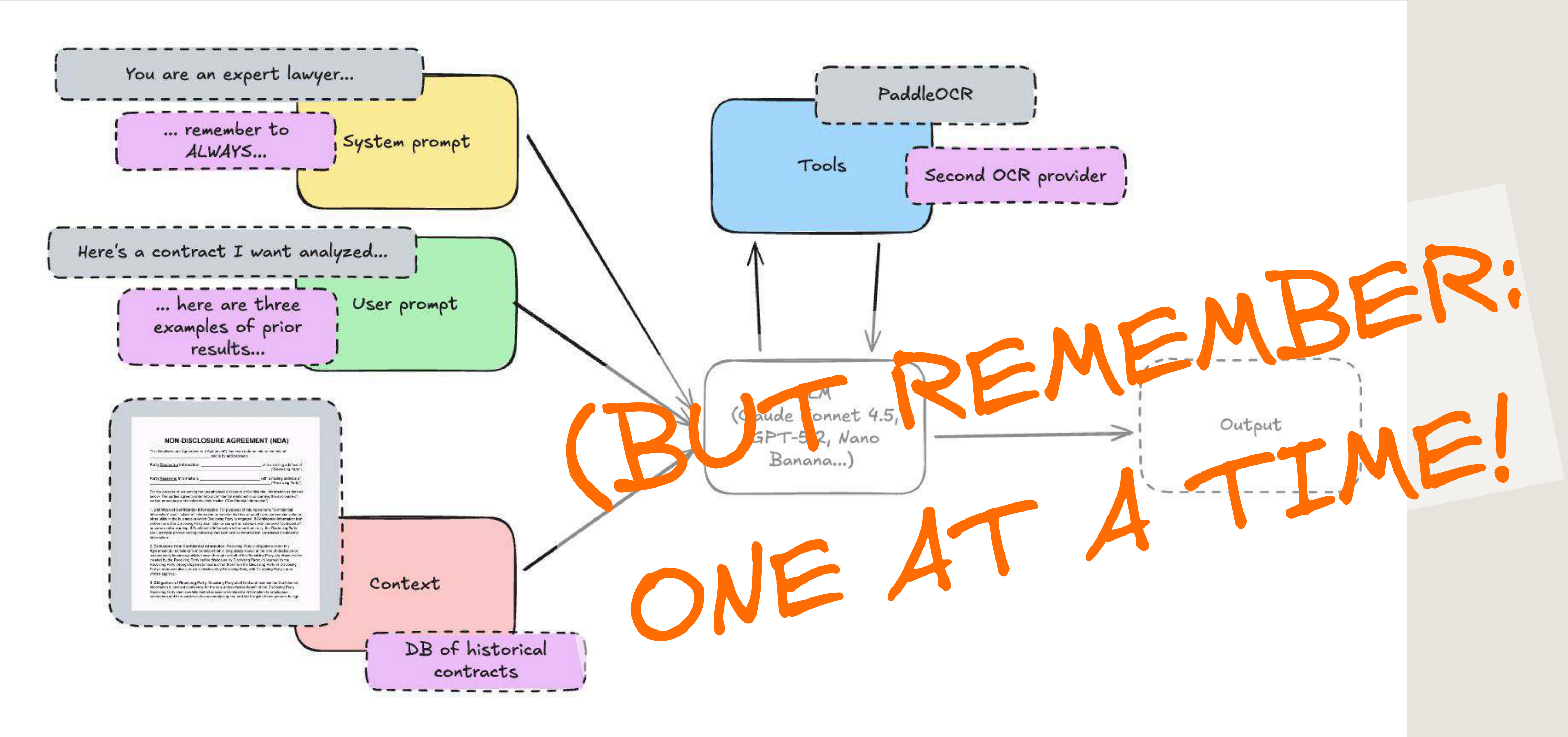
# LET'S MODEL THE SYSTEM

## GOAL: CONSISTENT HIGH QUALITY OUTPUTS



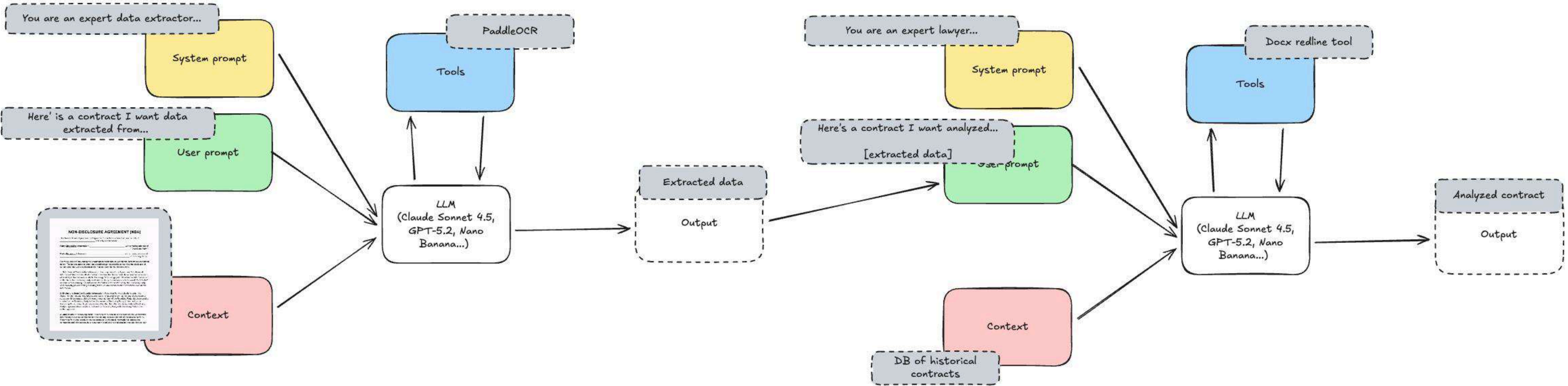


# WHAT MIGHT CHANGES LOOK LIKE?





# SOMETIMES YOU NEED TO CHANGE YOUR VIEW OF THE SYSTEM





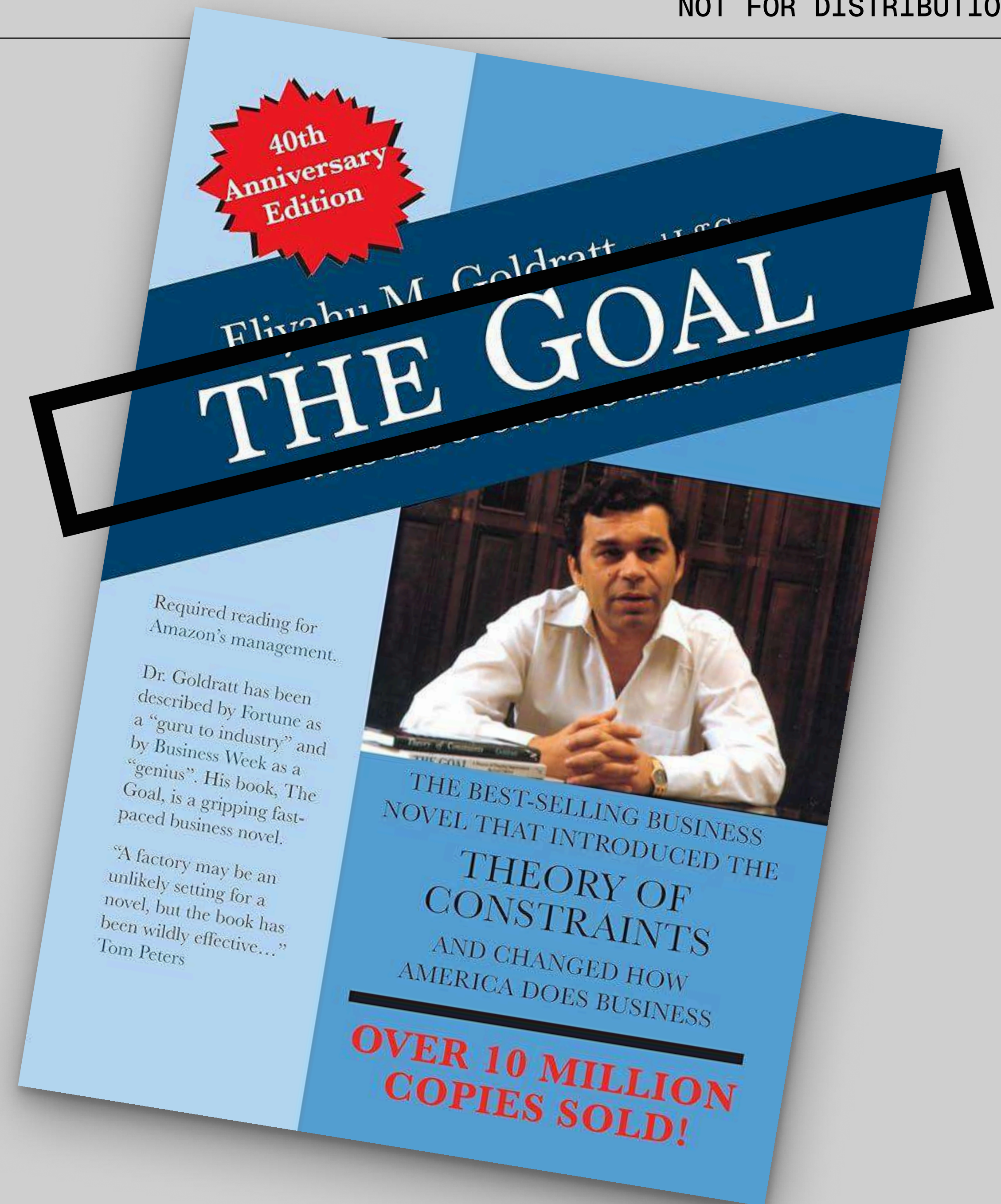
**HOW DO YOU KNOW WHAT TO CHANGE  
AND IF IT HELPED?**



# YOU NEED TO DEFINE ...

“YOU CANNOT UNDERSTAND THE  
MEANING OF PRODUCTIVITY  
UNLESS YOU KNOW WHAT THE  
GOAL IS.

UNTIL THEN, YOU’RE JUST  
PLAYING A LOT OF GAMES WITH  
NUMBERS AND WORDS.”





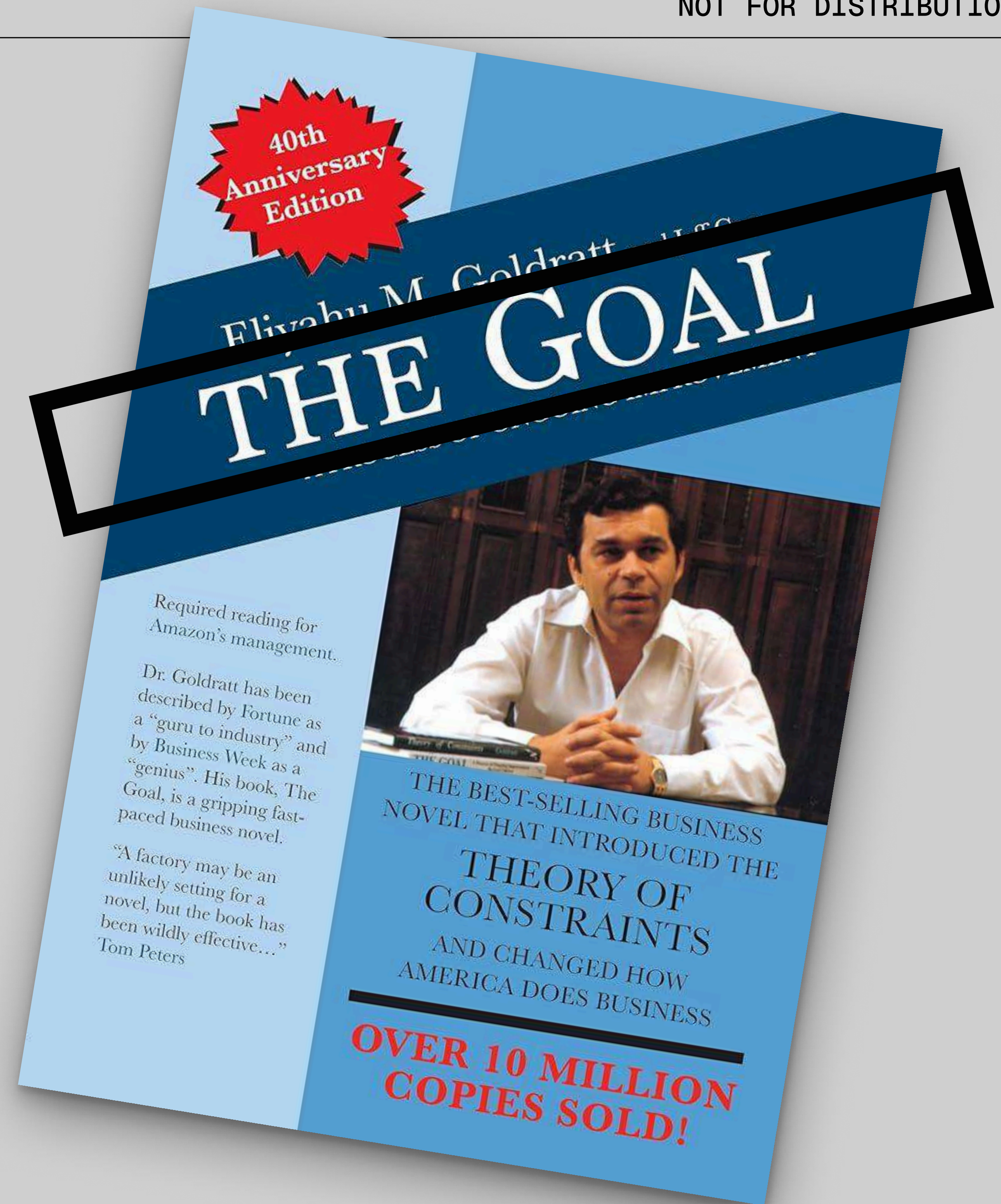
# YOU NEED TO DEFINE ...

GOOD RESULTS NEED IMPROVEMENT

IMPROVEMENT NEEDS EVALS

EVALS NEED A GOAL

A GOAL NEEDS A DEFINITION





WE SAID OUR GOAL WAS:

**CONSISTENT HIGH QUALITY OUTPUTS**

BUT HOW DO YOU  
MEASURE THAT?



HOW DO YOU MEASURE YOUR GOAL?

1. "PROPER" EVALS (MAY YOU BE BLESSED  
WITH A PRODUCT THAT  
PERMITS THESE)

2. VIBE EVALS

# PROPER EVALS

**TL;DR:**

- BUILD AND UPDATE AN “EVAL SET”
- RUN AGAINST IT CONTINUALLY
- UPDATE ACCORDINGLY
- ADD ONLINE/LIVE EVALUATION AS A BONUS

**BUT: SCORING IS HARD FOR MANY USE CASES**

(WHY DO YOU THINK MERCOR, SCALE AI, SURGE AI ARE AS BIG AS THEY ARE?)

**braintrust**



```
POST https://api.openai.com/v1/evals
```

**harbor**



**LangSmith**

# EXAMPLE: DOCUMENT EXTRACTION



PRE-LABELED



DOC1 : {  
TITLE: "PHOTOGRAPHY  
CONTRACT",  
BODY: "..."},

DOC2 : {  
TITLE: "EMPLOYMENT  
CONTRACT",  
BODY: { ... }",

...

SCORE = SIMILARITY( {PRE-LABELED} , {ACTUAL OUTPUT} )

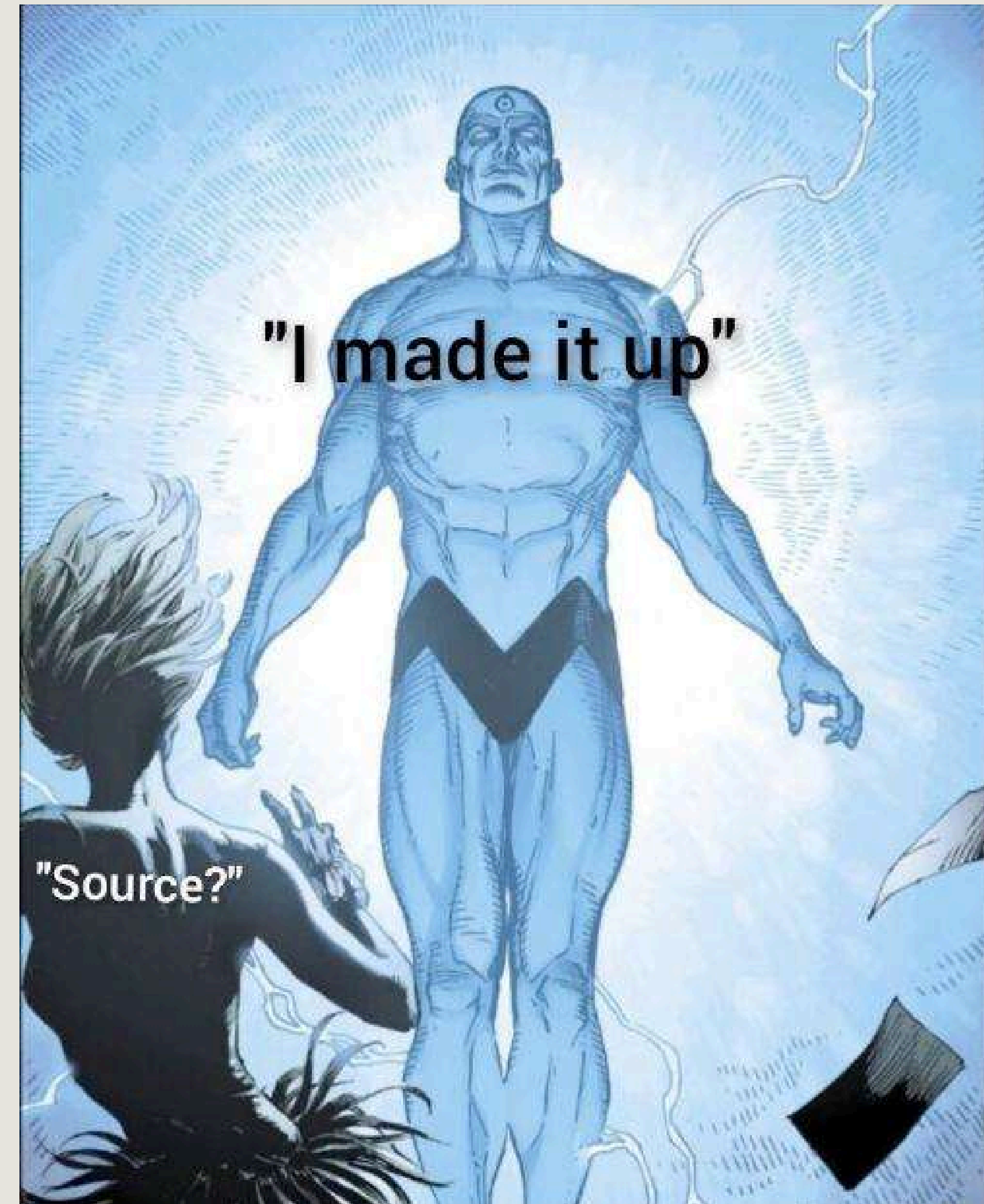
GOAL = MAXIMIZE( SCORE )



# VIBE EVALS

**TL;DR:**

- JUST, UH, LOOK AT THE RESULTS AND ASK YOURSELF HOW GOOD THEY ARE



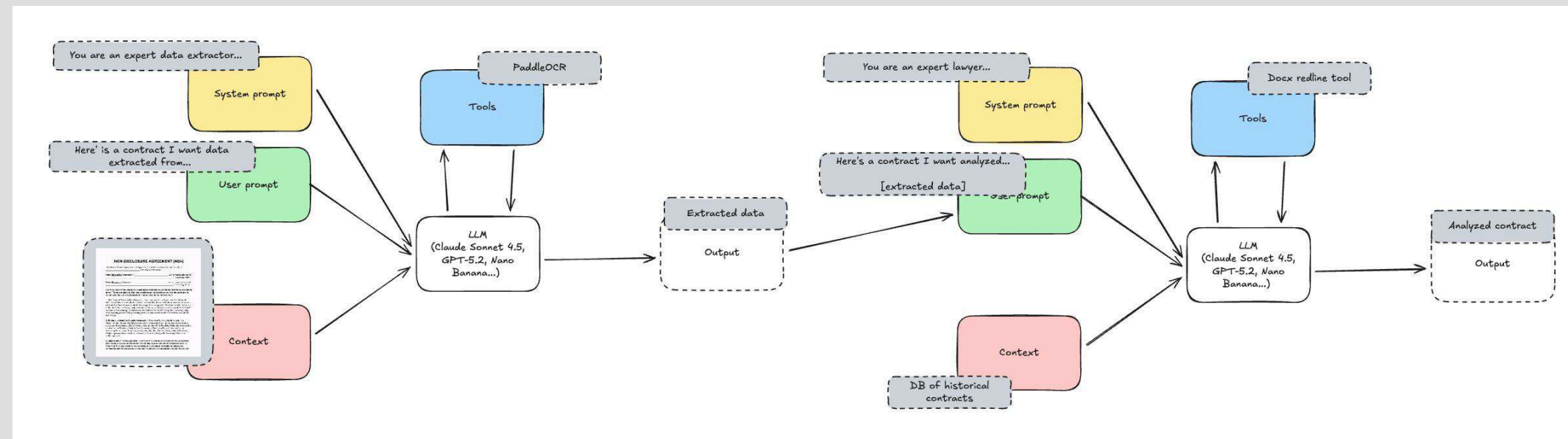


# EVAL OPTIONS

- 1      PROPER EVAL: PRE-LABELED RESULTS THAT SHOULD MATCH EXACTLY
- 2      PROPER EVAL: PRE-LABELED RESULTS THAT YOU CAN COMPUTE SIMILARITY TO
- 3      PROPER EVAL: LLM AS A JUDGE OF THE RESULTS
- 4      VIBE EVAL: HOW DOES IT FEEL

# IN SUMMARY: THE IMPROVEMENT LOOP

## STEP 1: DIAGRAM OUT YOUR SYSTEM



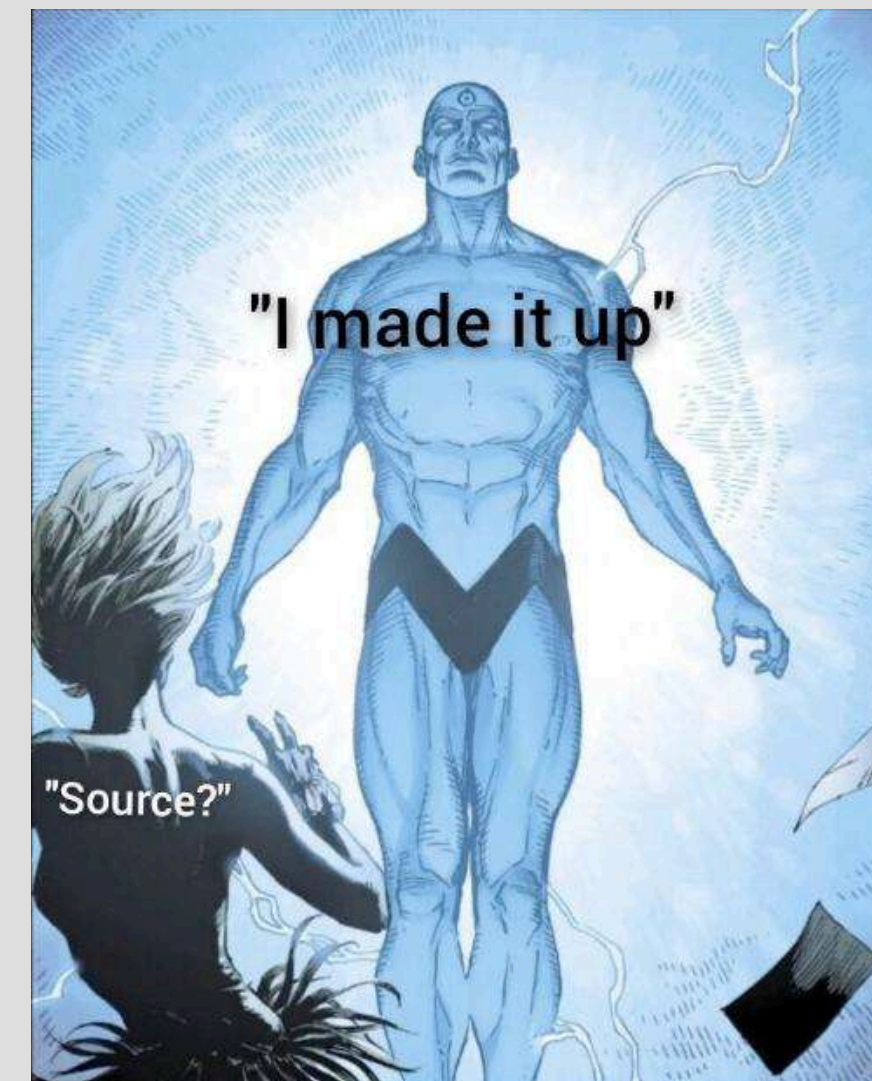
## STEP 2: DEFINE YOUR GOAL & EVAL PROCESS FOR SAID GOAL

# "PROPER" ... OR ... VIBE EVALS?

### Scorer types

Braintrust offers three types of scorers:

- **Autoevals:** Pre-built, battle-tested scorers for common tasks like factuality checking, semantic similarity, and format validation. Start here for standard evaluation needs.
- **LLM-as-a-judge:** Use a language model to evaluate outputs based on natural language criteria. Best for subjective judgments like tone, helpfulness, or creativity that are difficult to encode in code.
- **Custom code:** Write custom evaluation logic in TypeScript or Python. Best when you have specific rules, patterns, or calculations to implement. Custom code scorers can evaluate either the final output or the entire execution trace for multi-step workflows.

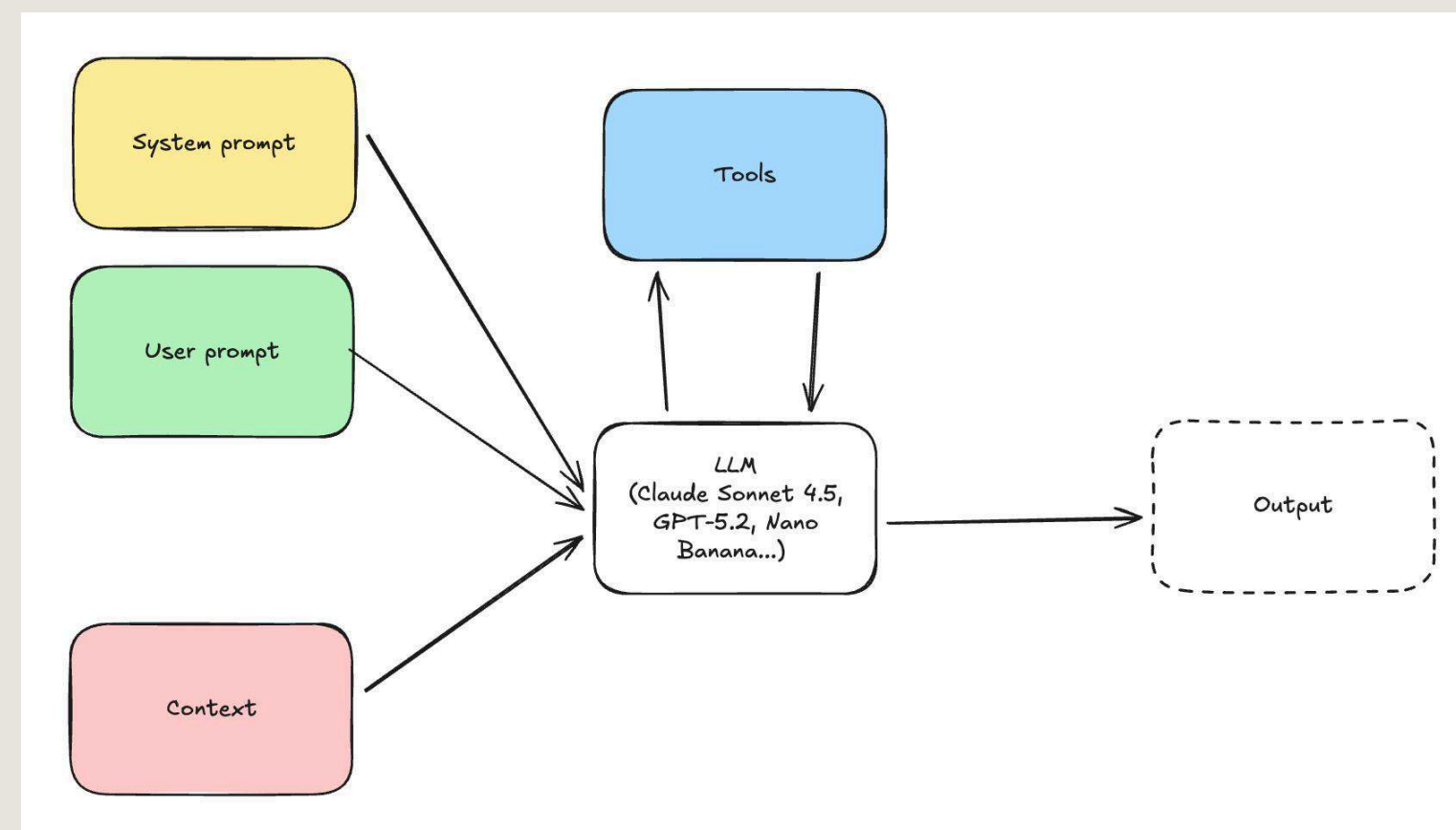


# IN SUMMARY: THE IMPROVEMENT LOOP

**STEP 1:** DIAGRAM OUT YOUR SYSTEM

**STEP 2:** DEFINE YOUR GOAL & EVAL PROCESS FOR SAID GOAL

**STEP 3:** ASSESS WHERE YOU THINK THE CONSTRAINT IS



SYSTEM PROMPT, USER PROMPT,  
CONTEXT, OR TOOLS  
... OR THE SYSTEM VIEW ITSELF

**STEP 4:** USE TOC: EXPLOIT CONSTRAINT, SUBORDINATE EVERYTHING ELSE;  
ELEVATE CONSTRAINT ... AND MEASURE EVALS EACH TIME



# IN SUMMARY: THE IMPROVEMENT LOOP

**STEP 1:** DIAGRAM OUT YOUR SYSTEM

**STEP 2:** DEFINE YOUR GOAL & EVAL PROCESS FOR SAID GOAL

**STEP 3:** ASSESS WHERE YOU THINK THE CONSTRAINT IS

**STEP 4:** USE TOC: EXPLOIT CONSTRAINT, SUBORDINATE EVERYTHING ELSE;  
ELEVATE CONSTRAINT ... AND MEASURE EVALS EACH TIME

**STEP 5:** GOTO STEP 3

# IN SUMMARY: THE IMPROVEMENT LOOP

STEP 1: DIAGRAM OUT YOUR SYSTEM

STEP 2: DEFINE YOUR GOAL & EVAL PROCESS FOR SAID GOAL

STEP 3: ASSESS WHERE YOU THINK THE CONSTRAINT IS

STEP 4: USE TOC: EXPLOIT CONSTRAINT, SUBORDINATE EVERYTHING ELSE;  
ELEVATE CONSTRAINT ... AND MEASURE EVALS EACH TIME

STEP 5: GOTO STEP 3

THIS SOUNDS BORING AND  
PROGRAMMATIC AND OBVIOUS...  
BUT... IT WORKS

(AND MOST PEOPLE DO NOT APPROACH IT SO RIGOROUSLY)

# IN SUMMARY: THE IMPROVEMENT LOOP

STEP 1: DIAGRAM OUT YOUR SYSTEM

STEP 2: DEFINE YOUR GOAL & EVAL PROCESS FOR SAID GOAL

STEP 3: ASSESS WHERE YOU THINK THE CONSTRAINT IS

STEP 4: USE TOC: EXPLOIT CONSTRAINT, SUBORDINATE EVERYTHING ELSE;  
ELEVATE CONSTRAINT ... AND MEASURE EVALS EACH TIME

STEP 5: GOTO STEP 3

THIS SOUNDS BORING AND  
PROGRAMMATIC AND OBVIOUS...  
BUT... IT WORKS

BONUS: IT ACTUALLY WORKS FOR MOST  
ANY PROBLEM IN LIFE OR BUSINESS



# THREE PRACTICAL EXAMPLES

1 LIGHTWORK KNOWLEDGE BASE

2 INTERFACE0

3 ~~[REDACTED]~~



**EVAL:**

- HOLDBACK OF A SET OF GOOD EMAILS (WITH THE CLIENT INQUIRY) AND SEGMENTS OF REPORTS (WITH THE RAW DATA)
- RUN LATEST AND GREATEST AGAINST THE INPUTS
- USE AN LLM AS A JUDGE TO COMPARE THE HOLDBACK WITH THE OUTPUTS

**TOC APPLICATION:**

- ORIGINAL CONSTRAINT: SYSTEM PROMPT
- ONGOING CONSTRAINT: CONTEXT (MORE AND MORE AND MORE)

**Lightwork Knowledge Base**

AskManage

Ask

Assessment context (optional)

Search assessments...

Response type

Draft email

Global contexts

251111 Collated Reports

Lightwork Newsletters

Evaluation Standards

☐ Search web as well

☐ Show debug

Question

a client is doing a big renovation and asked about what recessed lighting options would be good

3 contexts selected

Subject: Recessed Lighting Recommendations for Your Renovation

Hi [Client Name],

It's great to hear about the renovation. Since you are starting with a clean slate, you have a unique opportunity to in both your health and the aesthetic of the home.

When selecting recessed lighting, we prioritize three main health metrics to avoid issues like eyestrain, headaches


**1. Flicker-Free Technology**

Most standard LEDs flicker because of how they convert AC electricity to DC. Even if it's not visible, this can cause

# interface0

## EVAL:

- VIBES - THE REQUESTS ARE TOO VARIED FOR A PROPER EVAL

interface0 



## How can I help you today?

+ Knowledge



You can /use prompts or @mention knowledge & tools



GPT-5.2



0





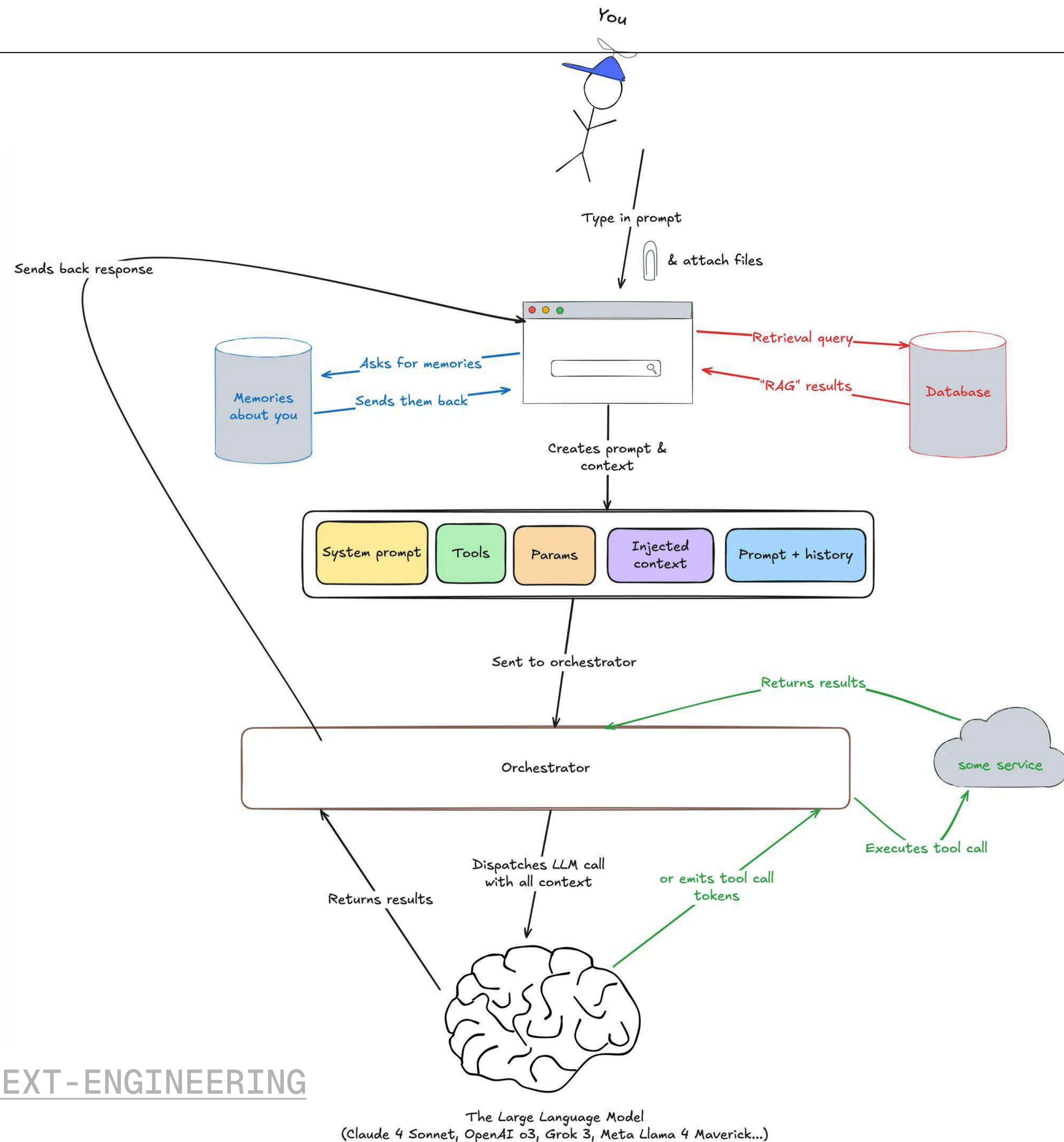
# interface0

## EVAL:

- VIBES - THE REQUESTS ARE TOO VARIED FOR A PROPER EVAL

## TOC APPLICATION:

- ORIGINAL CONSTRAINT: SYSTEM PROMPT
- SECOND CONSTRAINT: USER PROMPTS  
→ HELPED WITH TEMPLATES
- ONGOING CONSTRAINTS: ALTERNATING BETWEEN TOOLS AND CONTEXT



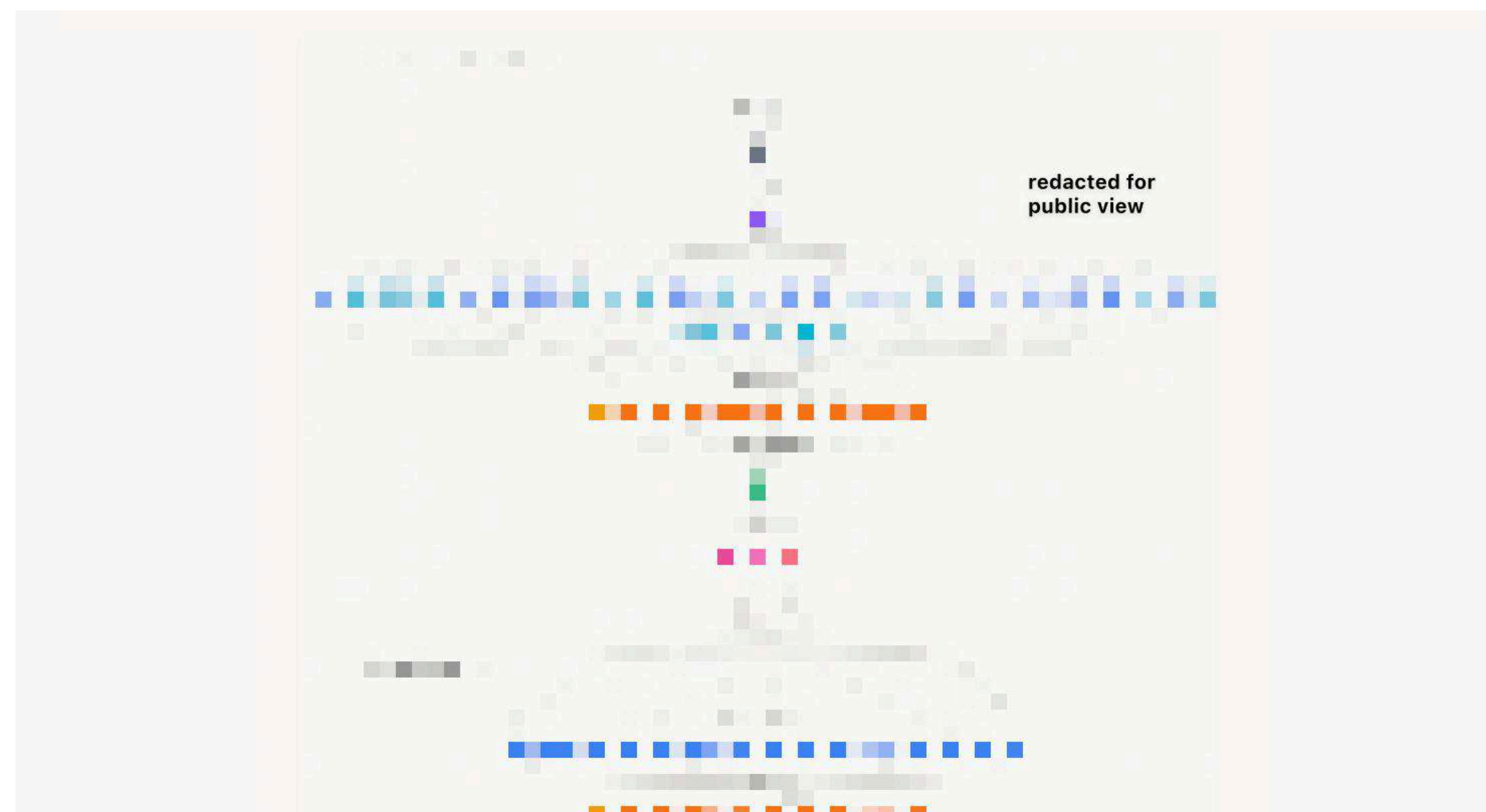
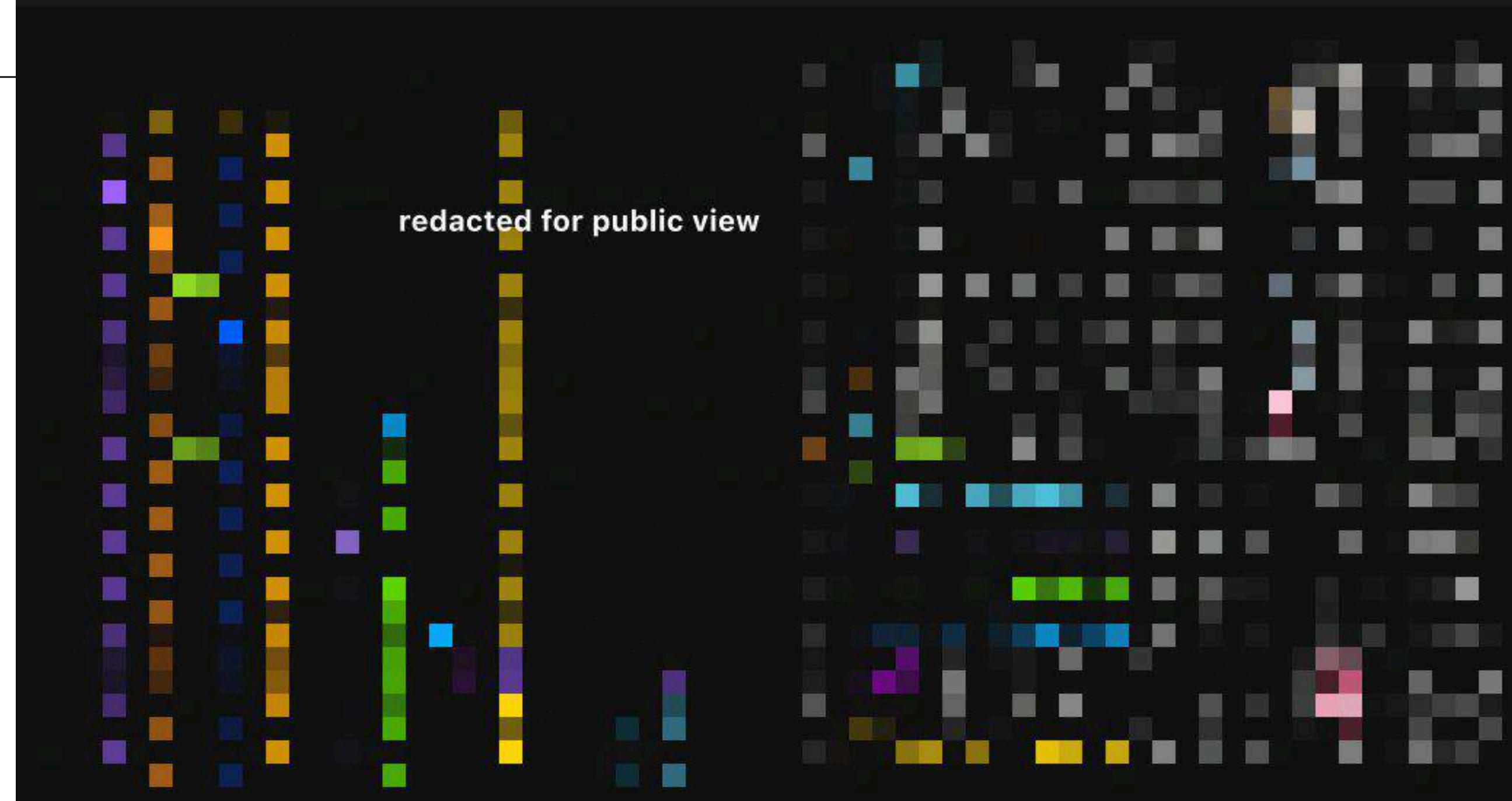
# [REDACTED]

## EVAL:

- EXPERT FEEDBACK ON THE RESULTS (SOME STRUCTURED, SOME UNSTRUCTURED)
- USER FEEDBACK ON THE RESULTS (UNSTRUCTURED)

## TOC APPLICATION:

- ORIGINAL CONSTRAINT: SYSTEM PROMPTS FOR SPECIFIC SUBAGENTS
- THEN: TOOLS
- NOW BACK TO CONTEXT & SYSTEM PROMPTS



# WHAT WE COVERED

1 THE ISSUE WITH LLMS

2 THE THEORY OF CONSTRAINTS

3 EVALS & HOW TO IMPROVE

4 SOME PRACTICAL EXAMPLES



# THE STATE OF PLAY TODAY: **THE LLM IS ~NEVER THE CONSTRAINT**

SO IF WE BUY THE THEORY OF CONSTRAINTS ...

THAT MEANS HERE ARE THINGS IN YOUR CONTROL TO GET  
YOUR PRODUCT TO THE QUALITY & CONSISTENCY YOU WANT

IT IS YOUR JOB TO MAP THE SYSTEM, BUILD EVALS,  
FIND THE CURRENT CONSTRAINT, AND FIX IT

RINSE & REPEAT. MAGIC IS IN YOUR HANDS!

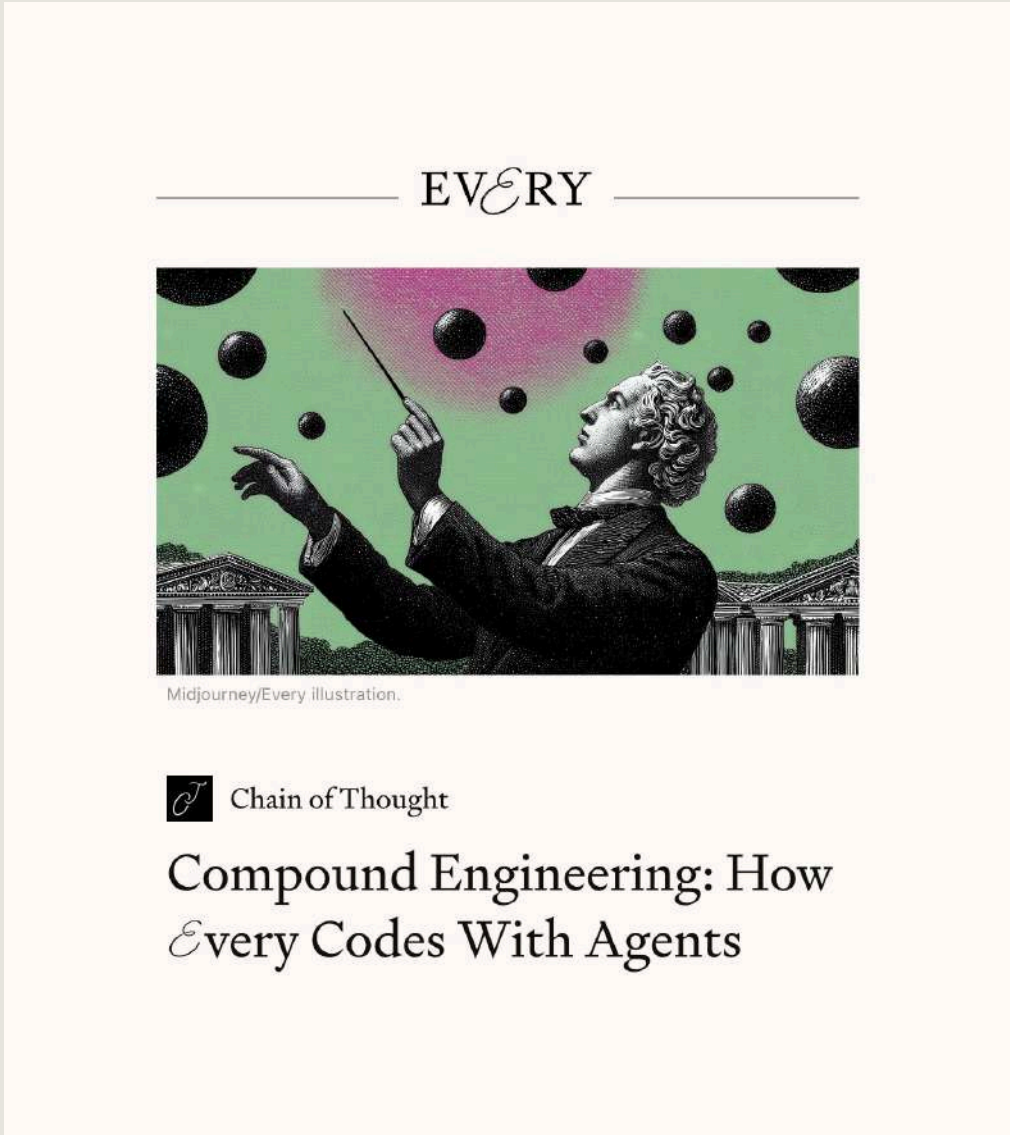
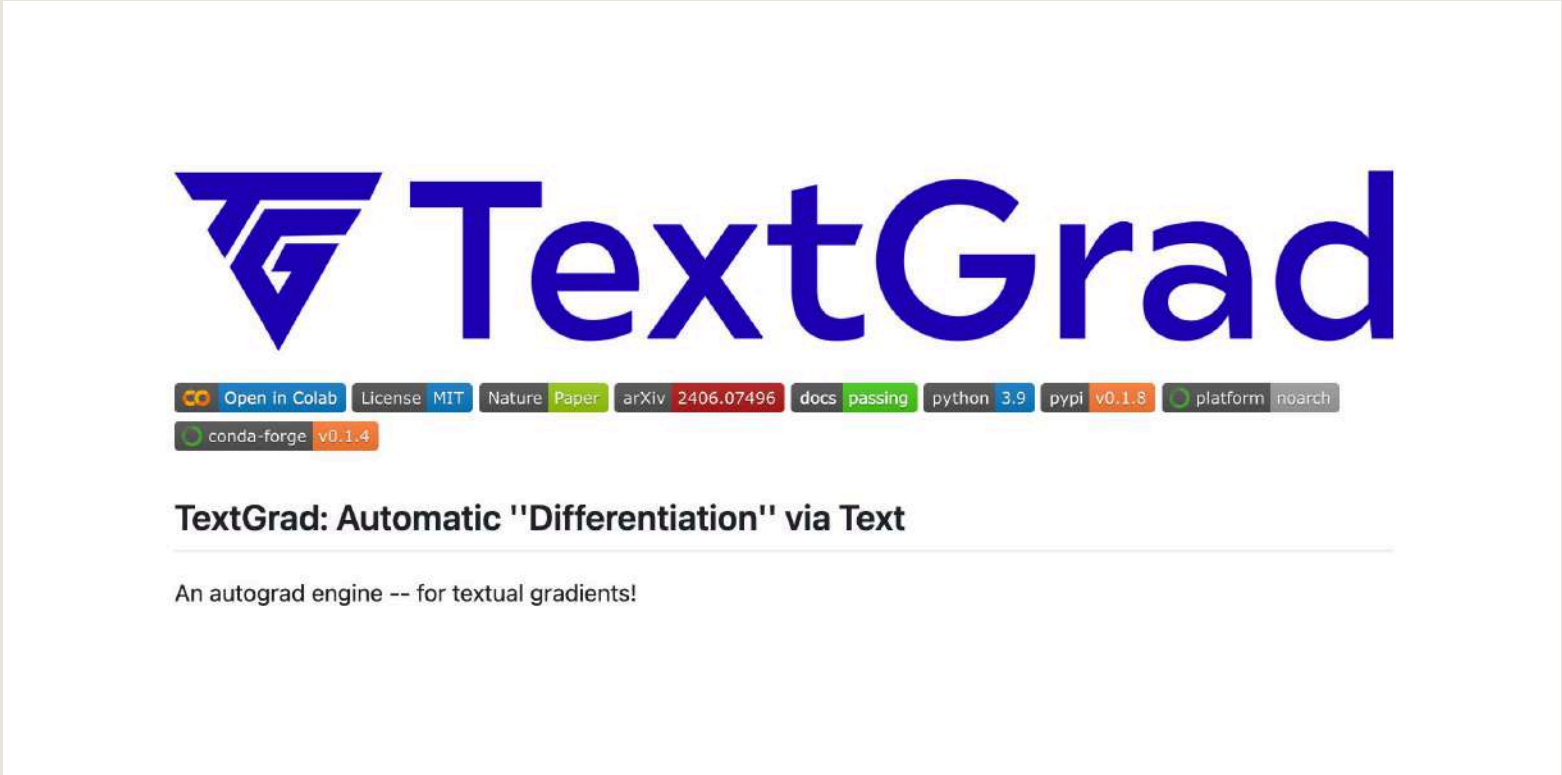
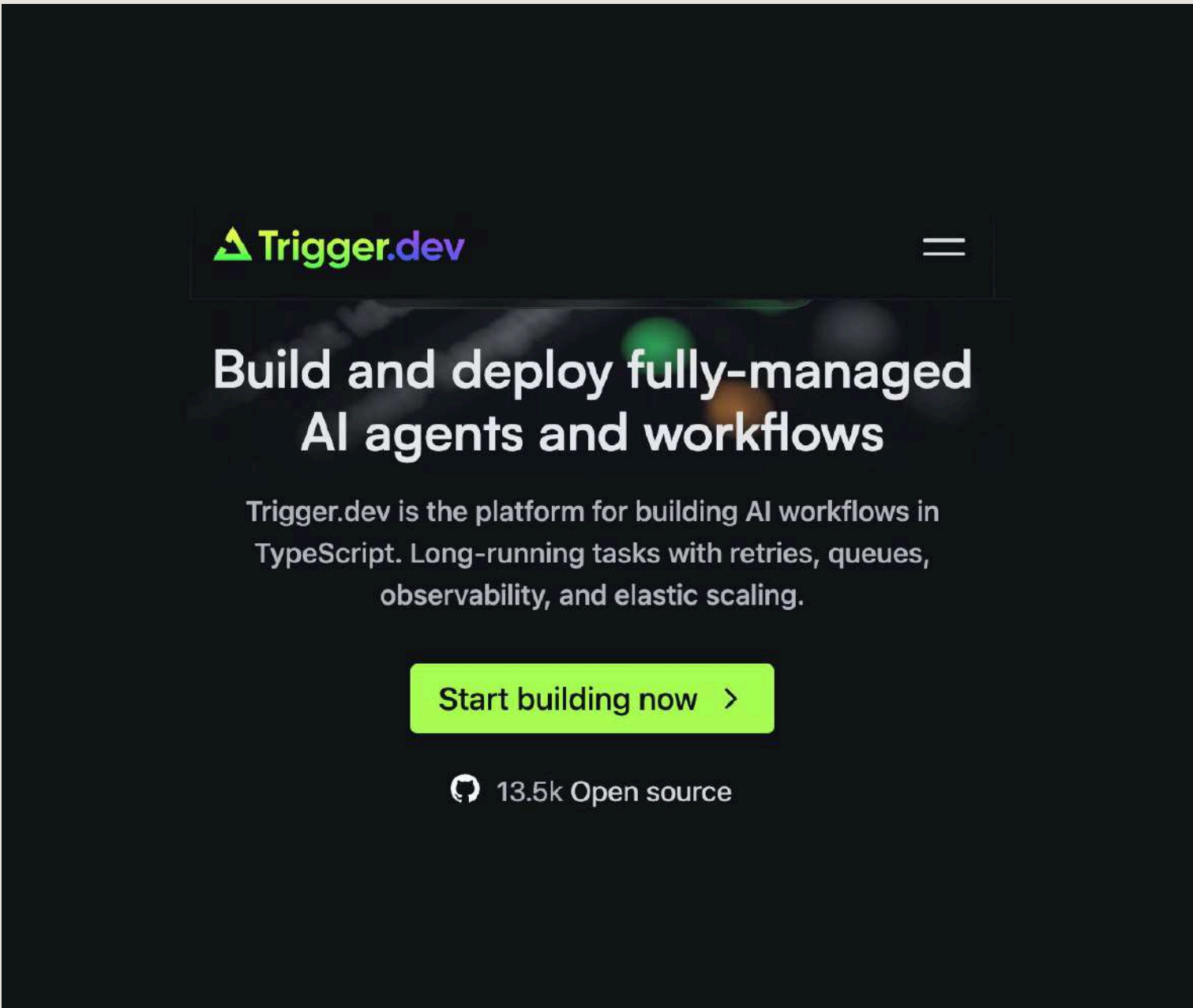
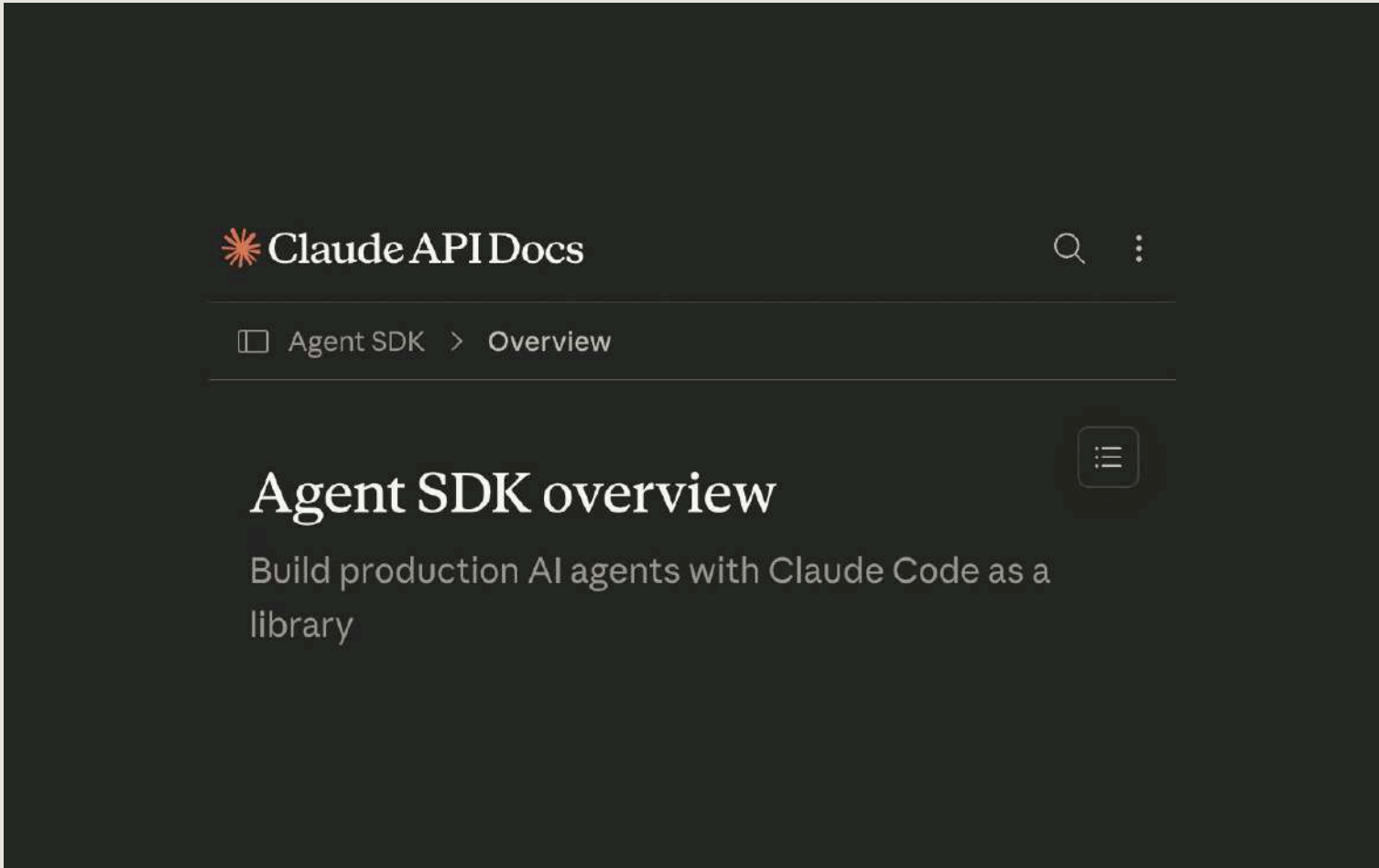
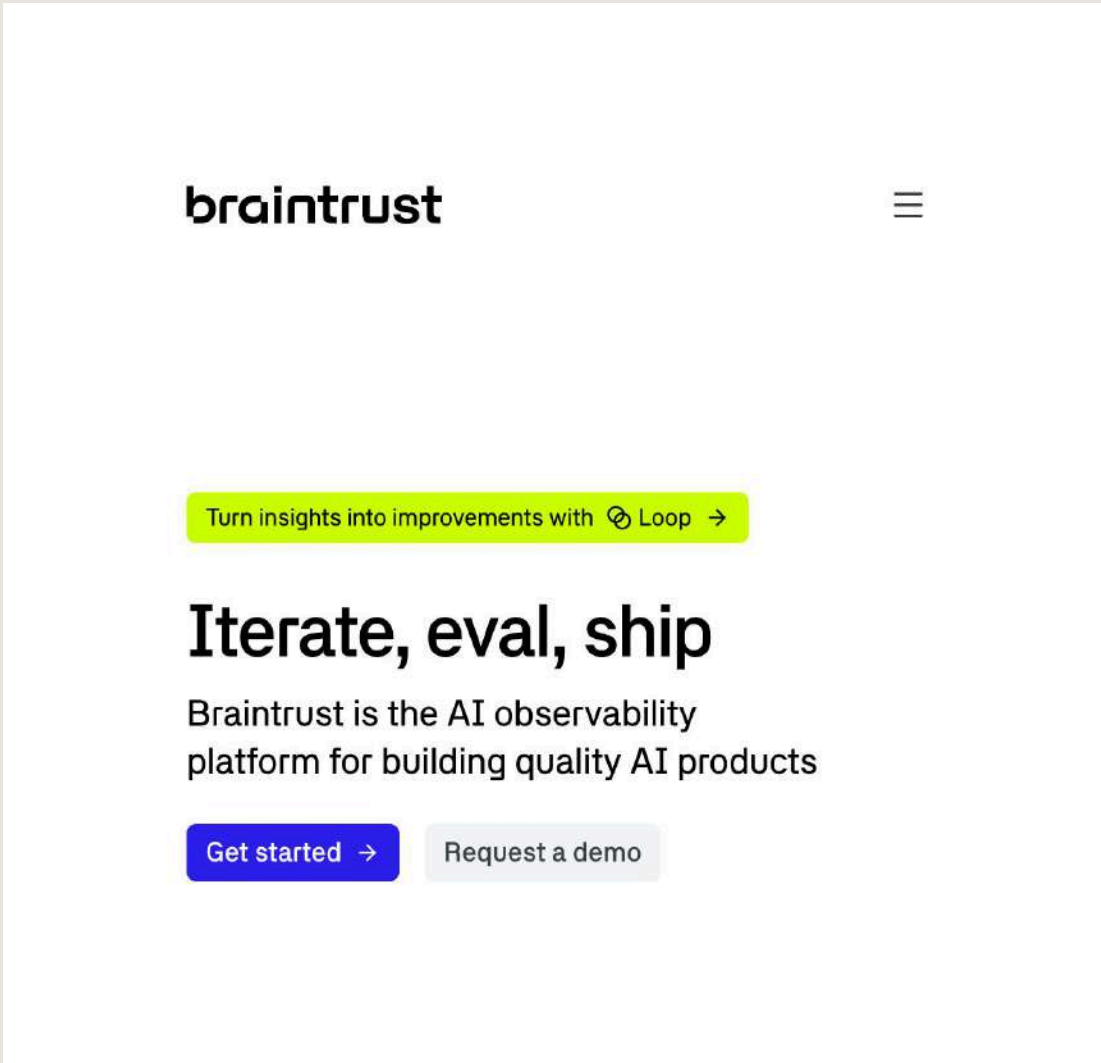
# PLEASE REACH OUT!

ANDY@ANDYBROMBERG.COM

I'D LOVE TO HEAR FROM YOU :)

P.S.: [HTTPS://TINKERING.CLUB](https://tinkering.club)

# TOOLS I CAN'T LIVE WITHOUT\*



\* PROBABLY OUT OF DATE BY FEBRUARY 6 2025 AT THE PACE THINGS ARE GOING



# BUILDING GOOD LLM PRODUCTS

WITH THE THEORY OF CONSTRAINTS

STANFORD CS 224G

FEBRUARY 5 2026

ANDY BROMBERG

ANDY@ANDYBROMBERG.COM

TINKERING: ANDYBROMBERG.COM · INVESTING: AMITY.VC