AGENTIC AI

PROMPTING LLMS

Erdem Özcan, GSÜ, October 2025

WHAT IS PROMPT ENGINEERING?

Guiding LLMs for optimal outcome

1 Art & Science of Prompts

Crafting inputs to steer LLM models towards generating specific and relevant responses

2 What is this all about?

Remember LLMs are auto-regressive models that generate one token at a time. Prompts is the initial conditioning that is given to the model to condition the rest of the tokens to be generated.

3 From a mathematicians point of view

We used to tune the outcome of ML models using features and hyper-parameters. Prompts has pretty much the same effect but using words instead, making it more accessible to a much wider audience.

IS THIS REALLY ENGINEERING

Art, science, or alchemy?

1 What is engineering?

It's the creative application of science and mathematics to design and develop solutions with predictable outcomes.

2 Prompting's Nature of prompting

It involves guiding probabilistic LLMs with natural language, leading to non-deterministic and iterative outputs, with possibly the same input generating a wide range of different outputs

3 Evolving Discipline

Systematic approaches are emerging, yet it navigates inherent uncertainty, differing from traditional engineering predictability.

ANATOMY OF A PROMPT

"You are a ***marketing expert****. Write a* ***3-sentence product description*** *for a* ***smartwatch****, focusing on* ***health tracking****, and use an* ***engaging but professional tone****."*

1. Identity
2. Goal / Task
3. Context
4. Tone
5. Format

Using LLMs in Code

```
from litellm import completion

resp = completion(
    model="gemini/gemini-2.5-flash",
    messages=[
        {
            "role": "system",
            "content": "You are a marketing expert. Use an engaging but professional tone."
        },
        {
            "role": "user",
            "content": "Write a 3-sentence product description for a smartwatch focusing on health tracking."
        }
    ],
    temperature=0.7,
)

print(resp["choices"][0]["message"]["content"])
```

1 System

General instructions to be given to the system to condition the output through the whole conversation thread.
Some LLM APIs call this ‘developer’

2 User

The prompts that are entered by the user of the application

3 Assistant

The model’s response. Some LLM APIs call this ‘agent’.

ZERO-SHOT

"You are a ***marketing expert****. Write a* ***3-sentence product description*** *for a* ***smartwatch****, focusing on* ***health tracking****, and use an* ***engaging but professional tone****."*

> *This smartwatch helps you track your health and stay active every day. It monitors your heart rate, steps, and sleep to keep you informed. A great choice for anyone who wants to improve their fitness.*

1. The prompt only contains the request and the rules
2. Adapted to situations like exploring or where there is no possibility to provide any examples
3. The output is directly a function of the patterns learned by the model

FEW-SHOTS

“You are a marketing expert.
Here are a few examples of product descriptions:
Example 1:
“Capture every heartbeat with precision. Our fitness band empowers you to reach your goals through smart insights and effortless style.”
Example 2:
“Transform your workouts with real-time tracking and elegant design. A wearable that moves as fast as you do.”

Now write a 3-sentence product description for a smartwatch focusing on health tracking, using an engaging but professional tone. are a marketing expert.

> *Elevate your well-being with a smartwatch that keeps pace with you. Monitor heart rate, sleep, and recovery with precision insights. Crafted for those who care about both performance and design.*

1. The prompt gives a few specific examples of what the expected outcome looks like
2. Adapted to situations where a few examples help identifying a pattern to be followed
3. The model’s output is conditioned by the samples offered by the user

CHAIN OF THOUGHT (COT)

You are a marketing expert.

***Think step by step** about what makes a great smartwatch description.*

*1. Identify the **key benefits to highlight**.*

*2. Decide on the **tone and target audience**.*

*3. Think about what **ideas or emotions** would inspire that audience.*

*4. Draft a **3-sentence description** that reflects those choices, focusing on health tracking and using an **engaging but professional tone**.*

Show your reasoning first, then your final description.

> ***Reasoning:***
>
> *• Key benefits: health tracking (heart rate, recovery, sleep insights), smart reminders, elegant design.*
>
> *• Audience: health-conscious professionals who value balance, precision, and design.*
>
> *• What inspires them: a sense of control over health, visible progress, tools that reflect their lifestyle, products that combine function and elegance.*
>
> *• Tone: confident, focused, and aspirational.*
>
> ***Final Description:***
> *"Stay ahead of your health with a smartwatch that keeps pace with your ambitions. Track every heartbeat, recovery, and rest with precision insights that inspire progress. Crafted for professionals who believe wellness and performance go hand in hand."*

1. The prompt gives instructions for how the model should reason about generating an optimal output

2. Adapted to situations where the pattern to be followed cannot be described by a few examples

3. The model has an opportunity to condition itself towards the optimal outcome step-by-step by generating preliminary preceding tokens

INPUT FORMATTING

```
# Identity
You are a **marketing expert** who writes persuasive product
descriptions
for consumer tech.

# Instructions
* Think step by step before writing.
* Identify benefits, emotions, and tone.
* Output <reasoning> then <final_description>.
* Keep it under 3 sentences.

# Example
<user_query>
Describe a smartwatch focusing on health tracking, with an engaging
but professional tone.
</user_query>

<assistant_response>
<reasoning>
- Benefits: heart rate, sleep, recovery, design.
- Audience: health-conscious professionals.
- Inspire with control, balance, and progress.
- Tone: confident and aspirational.
</reasoning>

<final_description>
Track your health with precision and purpose.
From sleep to recovery, every detail empowers your best self.
Designed for professionals who lead with balance.
</final_description>
</assistant_response>
```

1. Markup format is useful to give structured documentation and emphasis on certain parts of the prompt
2. XML is useful to define delimiters and hierarchies
3. XML is also useful to define output structures that can be parsed

OUTPUT FORMATTING

```
<assistant_response>
<reasoning>
- Benefits: heart rate, sleep, recovery, design.
- Audience: health-conscious professionals.
- Inspire with control, balance, and progress.
- Tone: confident and aspirational.
</reasoning>

<final_description>
Track your health with precision and purpose.
From sleep to recovery, every detail empowers your best self.
Designed for professionals who lead with balance.
</final_description>
</assistant_response>
```

> Track your health with precision and purpose.
> From sleep to recovery, every detail empowers your best self.
> Designed for professionals who lead with balance.

1. Useful to post-process the output before presenting it to the user
2. Especially handy with CoT
3. Also useful to make the LLMs generate commands to be executed by the environment

REASONING MODELS

<thinking>
- Benefits: heart rate, sleep, recovery, design.
- Audience: health-conscious professionals.
- Inspire with control, balance, and progress.
- Tone: confident and aspirational.
</thinking>

Track your health with precision and purpose.
From sleep to recovery, every detail empowers your best self.
Designed for professionals who lead with balance.

1. Automatically generates the reasoning tokens in a separate block
2. Systematic CoT to generate the optimal conditioning before generating the answer
3. Usually not meant to be shown to the user

TOOL CALLING

# Identity
You are a marketing assistant that writes accurate, engaging
product descriptions for wearable tech.

# Tools
- search_product_info(query): retrieves verified product specs.
- summarize_reviews(product): extracts top 3 customer insights.

# Instructions
* Use the tools if you need up-to-date info.
* Then reason step by step about key benefits and tone.
* Output two sections: <reasoning> and <final_description>.

# Example
<user_query>
Write a 3-sentence description for the latest Fitbit smartwatch,
focused on health tracking and recovery features.
</user_query>

<assistant_response>
<tool_call>
search_product_info("latest Fitbit smartwatch health features")
summarize_reviews("Fitbit smartwatch")
</tool_call>

<reasoning>
- Specs: improved sleep tracking, stress management, recovery scores.
- User sentiment: praised for motivation insights and long battery life.
- Audience: health-conscious professionals.
- Inspire with mastery, balance, and self-awareness.
</reasoning>

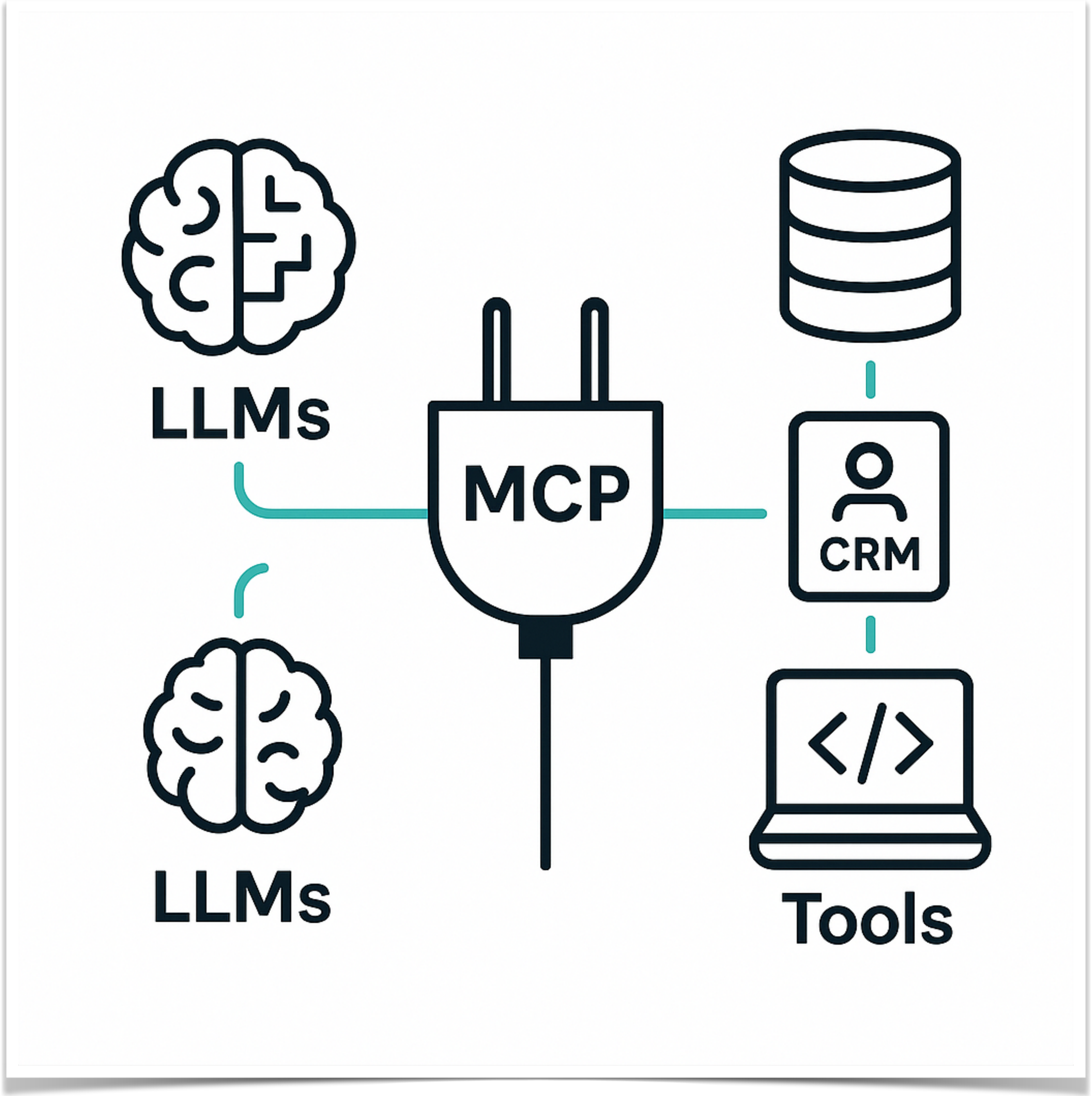
<final_description>
Master your well-being with the new Fitbit smartwatch.
Track sleep, stress, and recovery with precision insights that drive progress.
Built for professionals who value balance, energy, and focus.
</final_description>
</assistant_response>

Getting beyond the LLMs capabilities

- Including search results to ground answers on
- Accessing APIs and databases
- Executing programs
- Collaborating with other LLMs
- Interacting in other modalities

MCP

The “USB-C of AI”



1 One universal connector

MCP is an open-source, JSON-RPC standard that lets any model plug into any tool, securely and at scale.

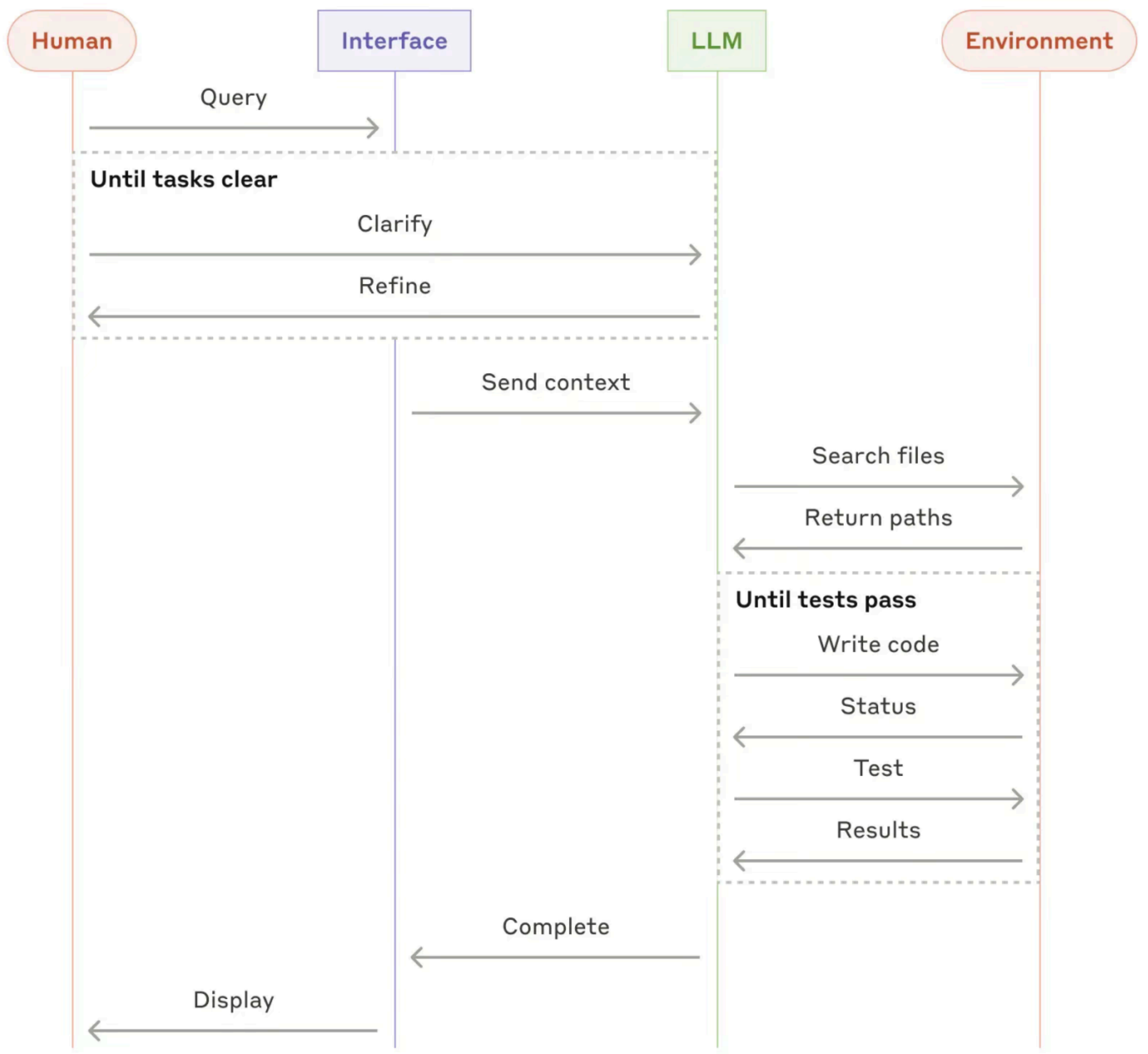
2 Why it matters?

Eliminates the N × M integration grind. Major players—OpenAI, Google DeepMind, Microsoft Windows, Replit—have already adopted or announced support

3 Business upside & guard-rails

Standardisation cuts integration cost, preserves portability, and simplifies governance audits

GENERAL ARCHITECTURE



1 Human

The ultimate commander and controller

2 Interface

Usually a conversational web application

3 LLM

Acting as the orchestrator

4 Environment

Observations and creation of side effects