HOW CLAUDE WORKS?

Peeking Inside Claude’s Cognition

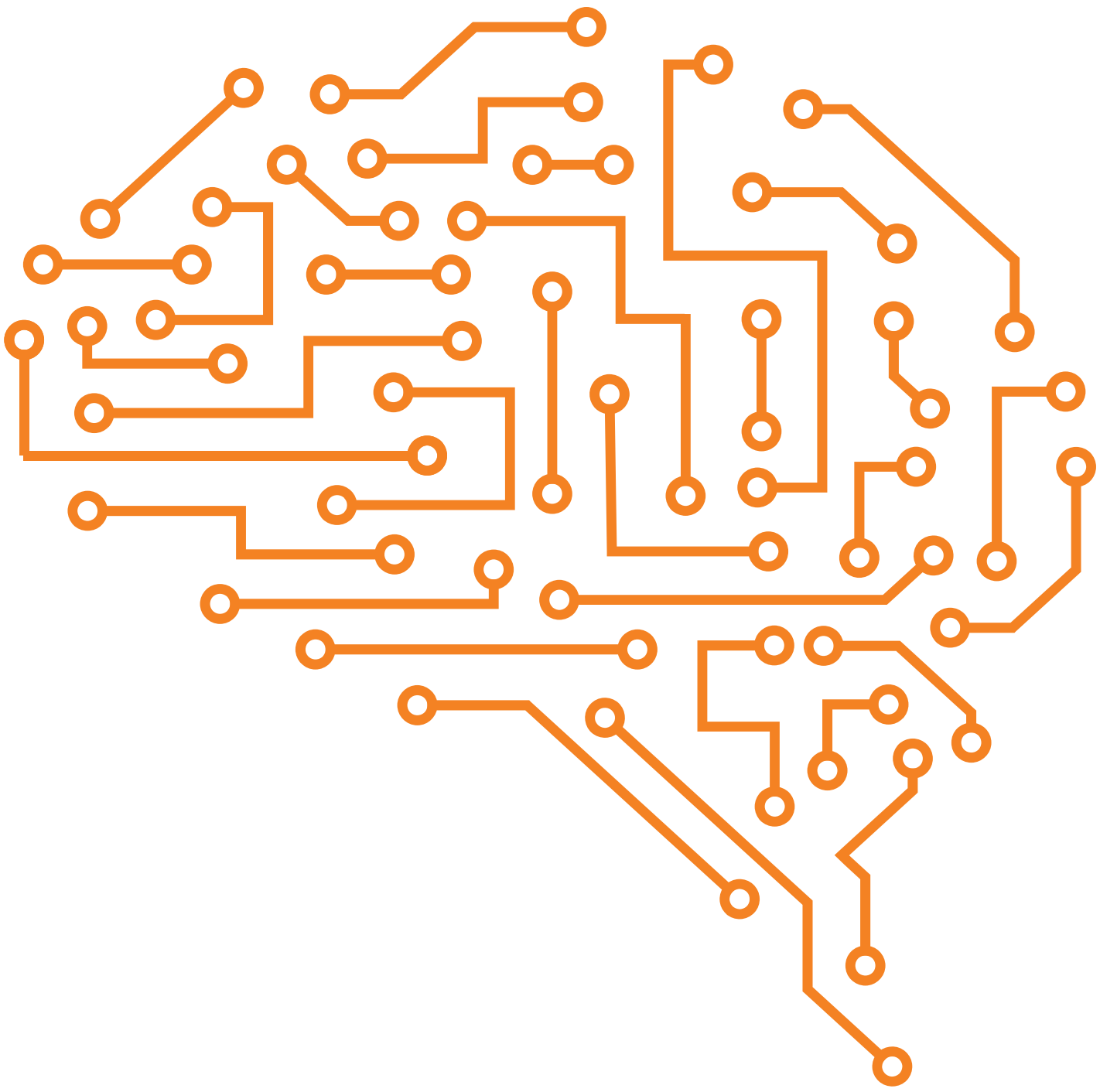
Bhavishya Pandit

Introduction

Ever wondered what an AI is really thinking?

LLMs like Claude are now the foundation of contemporary AI solutions, driving a wide range of applications from chatbots to virtual tutors, and much more.

However, for the most part, the mechanisms by which they do so are unknown. The question is, **do we really understand AI?**



In this post we will peek inside the mind of a large language model. The results will **surprise** you.

Anthropic’s Mission

Anthropic doesn’t just want powerful AI, but an **understandable AI**.

Source

One way they do this is by **tracing how Claude forms its answers**, step by step, to turn it from **black box into a glass box**.

They trace :

- how features light up
- how information flows
- how Claude “decides”

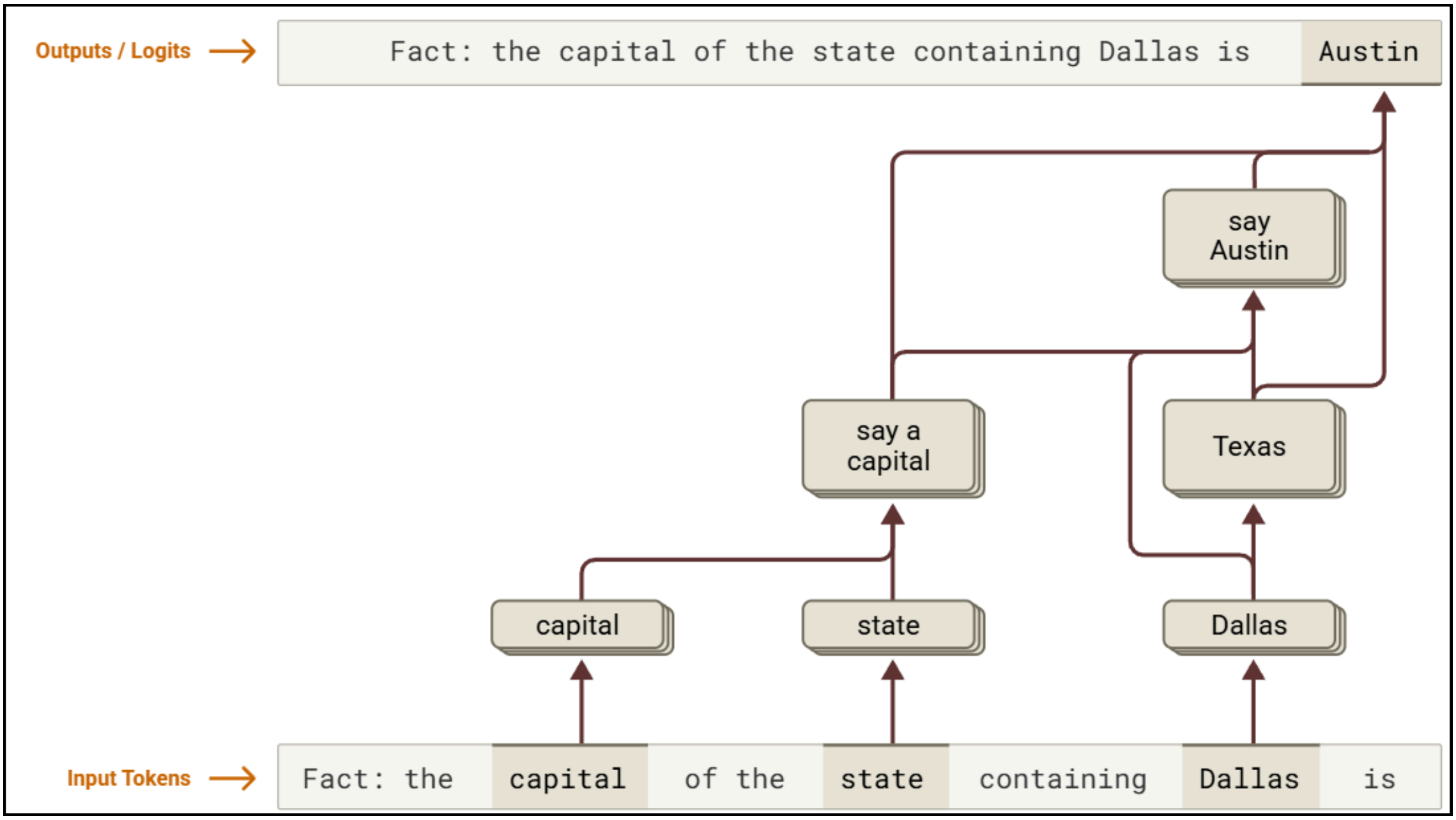
How do they do it? Let’s zoom into the building blocks with their “**AI microscope**”.

Claude Thinks in Features & Circuits

At its core, **Claude thinks not in words but features**.

Just like the lego bricks, alone these features mean little. But when connected, they **form circuits** that **map out Claude's thought process** step-by-step.

Then **attribution graphs** spotlight which features power each answer.



Source

This attribution graph shows it performs multi-hop reasoning—linking "Dallas" to "Texas," then combining "Texas" + "capital" to reach "Austin."

These layered steps reveal how Claude builds answers, using both direct and indirect pathways, through intermediate concept activation, not just word matching.

Claude Has a ‘Language of Thought’

Claude doesn’t just react. **It plans**.

For example :

- **English:** `The opposite of "small" is "` → `big`
- **French:** `Le contraire de "petit" est "` → `grand`
- **Chinese:** `"小"的反义词是"` → `大`

Source

Researchers found Claude uses shared multilingual circuits to answer antonym prompts across languages.

It first activates **same language-independent** “antonym” **features**, then combines them with language-specific output features (like “big” in Chinese).

When asked to write a haiku, Claude strategizes :

- Theme
- Rhythm
- Structure

Then responds with a haiku. It maps out this structure in its “thinking space” , and then writes.

This reveals Claude’s ability to **multi-step reasoning**, much like human planning.

But, what if it Lies?

Claude often explains why it gave an answer. But is that explanation always real?

Researchers found Claude sometimes explains its answers with smooth, logical-sounding reasons, but these reasons don’t reflect how it actually arrived at that answer.

This is called “**decoupled reasoning**”. It’s like a student guessing on a test, then inventing a clever-sounding explanation afterward.

Source

Like, when asked “What is the capital of Australia?”, Claude may answer “It’s Sydney”, stating a reason like “Sydney is the largest city in Australia. It hosts key government buildings near the harbor. So, Sydney is the capital.”

It’s wrong since the real capital is Canberra.

These **hallucinated justifications can mislead users**, even experts.

Why Is It Challenging to Understand AI?

There’s no “**Decode**” **button** for AI, not yet.

To understand Claude’s decisions, researchers dig through:

- Millions of activations
- Layers of feature traces
- Complex attention maps

It’s like **reverse-engineering each thought**, frame by frame.

Even with tools like “AI microscopes,” i**nterpreting Claude’s mind remains labor-intensive**.

But Anthropic is building tools to make this detective work faster,and someday, automate this.

The Future of Transparent AI

Peeking inside Claude isn’t just fascinating, it’s a step toward **safer, fairer, and smarter AI**.

This research is the foundation for how we can understand & shape AI.

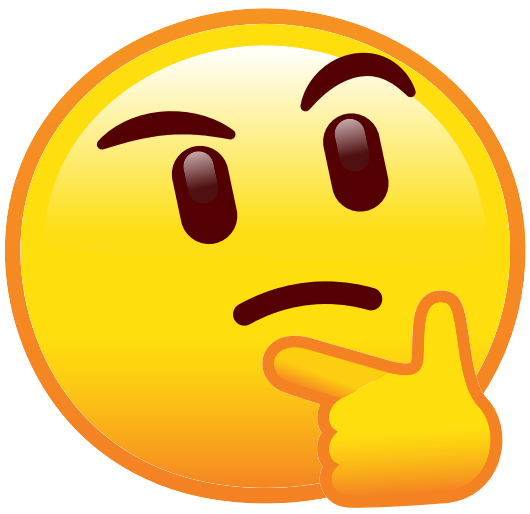
Transparency is how we build trust, empower users, and create inclusive AI systems for everyone.

It means moving from black-box decisions to **collaborative intelligence**.

This opens the door to AI that can be audited, corrected, and improved, not just used.

Model transparency comes at the cost of security?

Let me know your thoughts in the comments



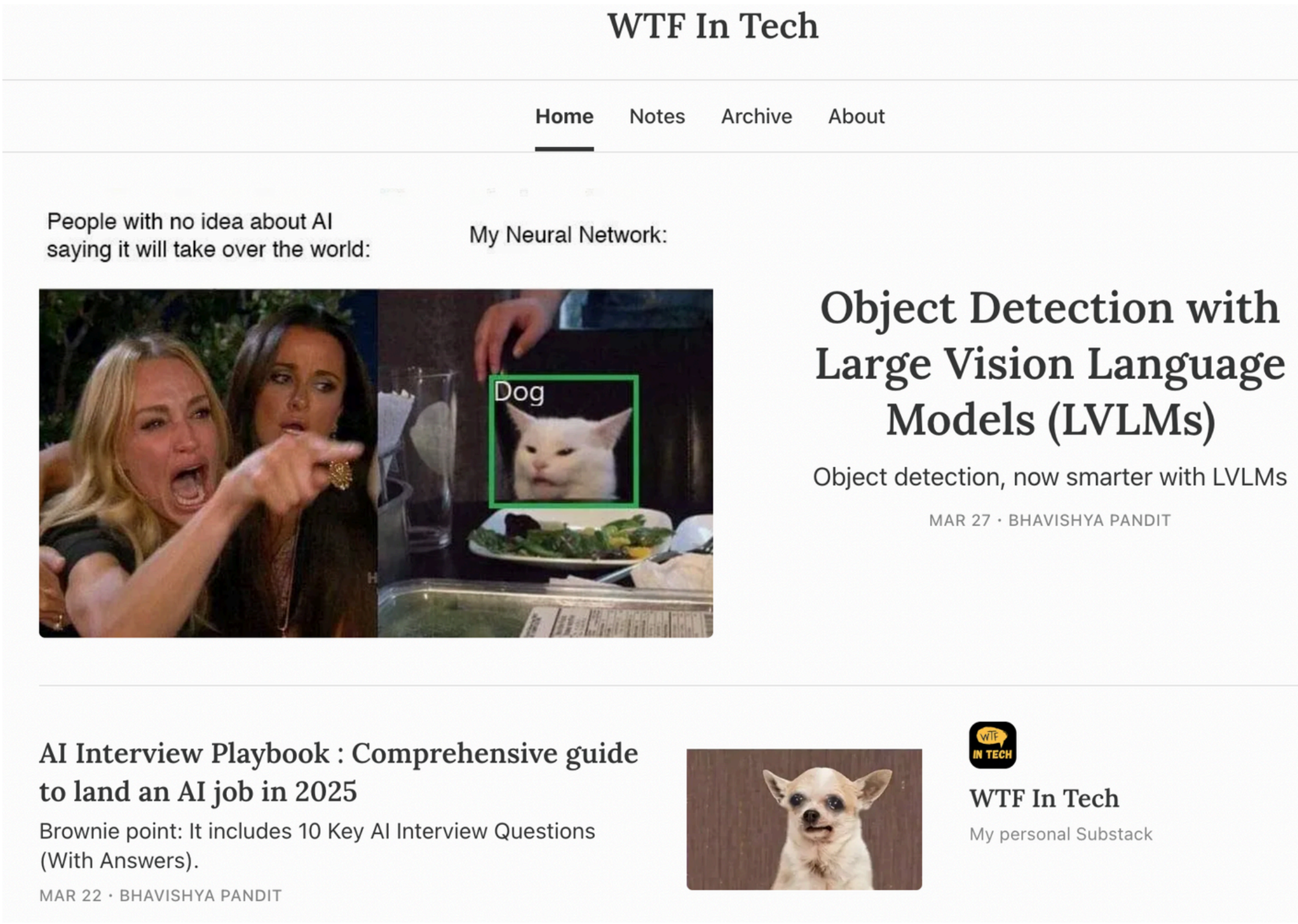
Stay Ahead with Our Tech Newsletter! 🚀

👉 **Subscribe and join 1k+ leaders and professionals**

🔗 **https://bhavishyapandit9.substack.com/**

Join our newsletter for:

- Step-by-step guides to mastering complex topics
- Industry trends & innovations delivered straight to your inbox
- Actionable tips to enhance your skills and stay competitive
- Insights on cutting-edge AI & software development



💡 Whether you're a developer, researcher, or tech enthusiast, this newsletter is your shortcut to staying informed and ahead of the curve.

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST