Teaming with AI

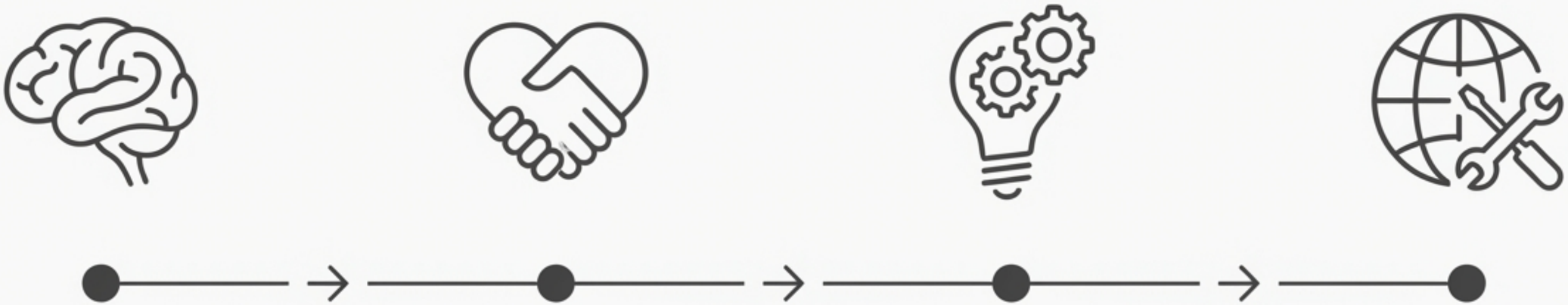
Understanding the New Digital Colleague

A Strategic Briefing for the Apple Board

November 2025

The AI Revolution is a Partnership

We are moving beyond viewing AI as a simple tool. The next paradigm is AI as a collaborator—a teammate we can train, task, and trust. This requires a fundamental understanding of how this teammate evolves.



1. Understanding Language

The Foundation

2. Aligning Values

The Training

3. Solving Problems

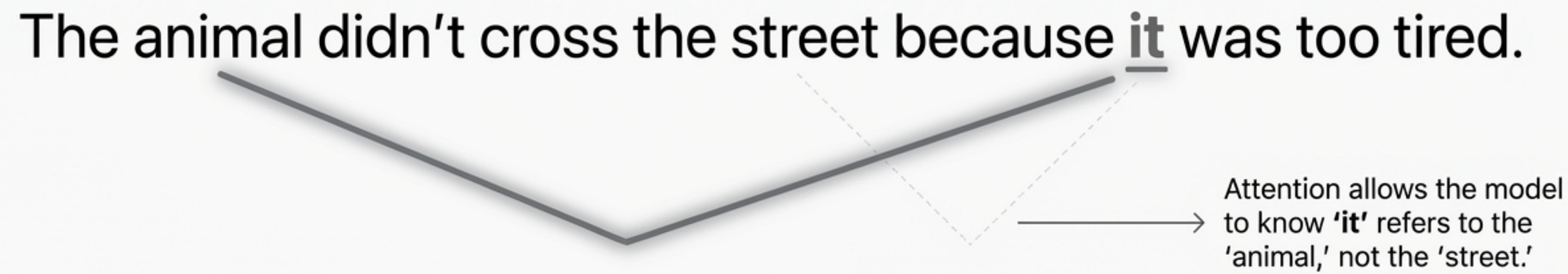
The Apprentice

4. Taking Action

The Collaborator

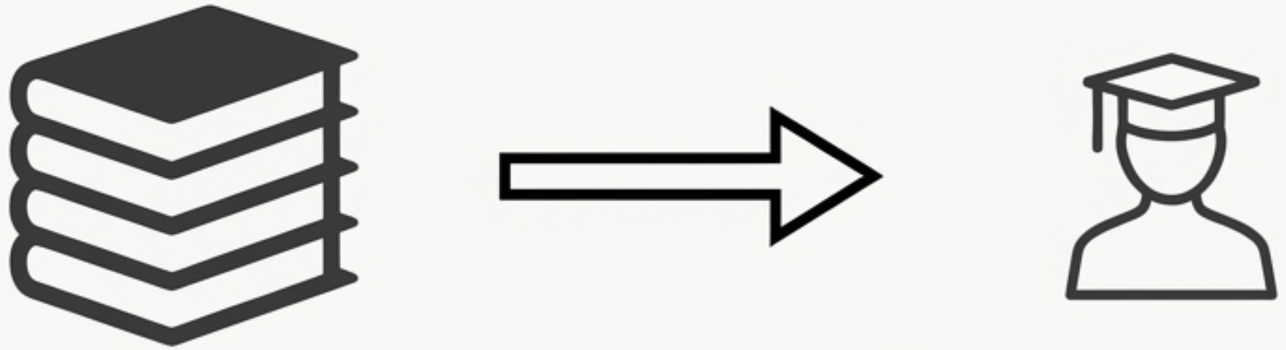
The Foundation: How AI Reads

At the core of modern AI is the **Transformer Architecture**, powered by a mechanism called **Self-Attention**. It's not just reading words; it's understanding context by weighing the importance of every word relative to every other word. This allows it to grasp complex relationships, grammar, and nuance.



The Training Regimen: From Knowledge to Utility

An AI teammate's capability is built in two distinct phases:



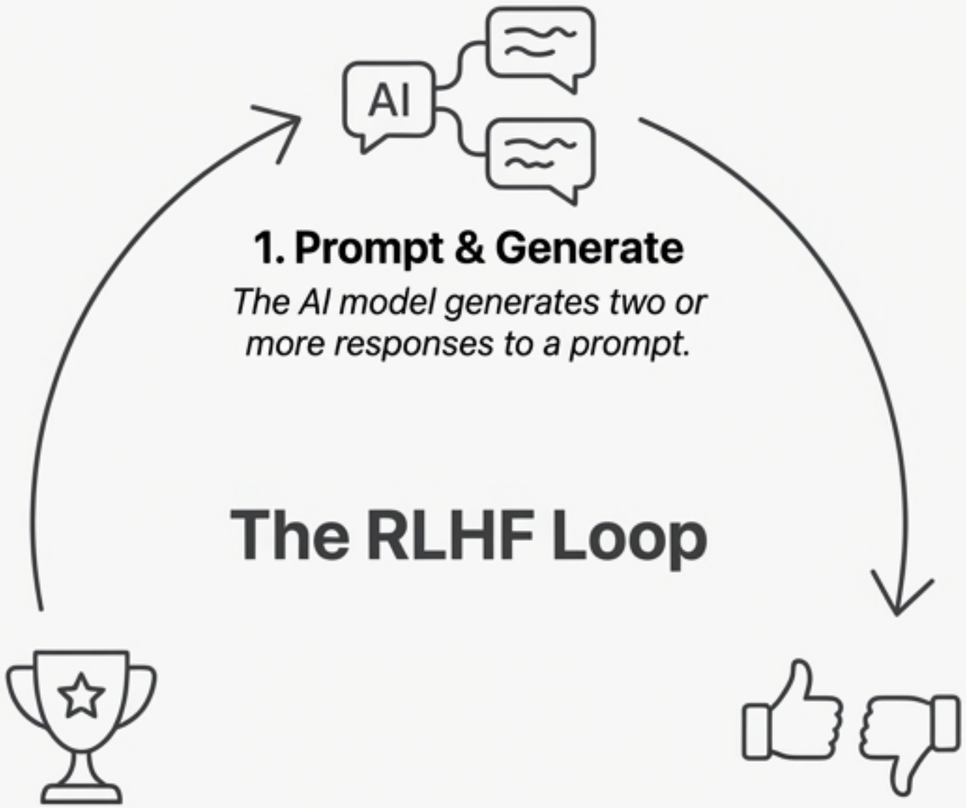
Pre-training

Fine-Tuning (Alignment)

Phase	Objective	Process	Analogy
1. Pre-training	**General Knowledge**	The model learns language, facts, and patterns from a vast corpus of text and code.	*Reading the entire library.*
2. Fine-Tuning	**Specific Skills & Behavior**	The pre-trained model is trained on a smaller, high-quality dataset to perform specific tasks and align with human preferences.	*Apprenticeship for a specific job.*

The Alignment: Teaching AI Our Preferences

How do we ensure our AI teammate is helpful and harmless? We use **Reinforcement Learning from Human Feedback (RLHF)**. It's a scalable process for encoding human values directly into the model.



The RLHF Loop

1. Prompt & Generate

The AI model generates two or more responses to a prompt.

2. Human Feedback

A human ranks the responses from best to worst.

3. Reward & Update

A 'reward model' learns to predict human preferences, then uses this signal to fine-tune the original AI model.

The Leap to Reasoning: Beyond Autocomplete

"Vanilla" LLMs excel at pattern matching but struggle with multi-step problems. The next evolution is creating Reasoning Models. Instead of just giving an answer, they generate an internal monologue—a Chain of Thought—to break a problem into solvable steps.

Vanilla LLM

Prompt: A bear was born in 2020. How old is it in 2025?

Answer: 2020. (Incorrect)

Reasoning LLM

Prompt: A bear was born in 2020. How old is it in 2025?

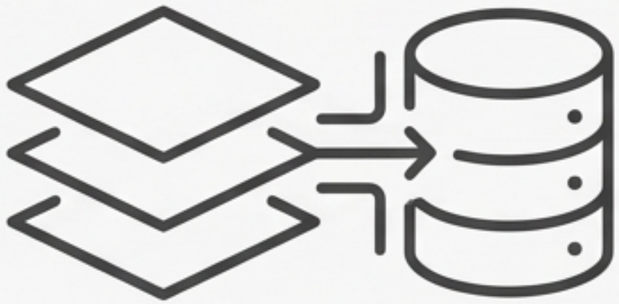
Reasoning Chain (shows its work):

1. The current year is 2025.
2. The bear was born in 2020.
3. Age = Current Year - Birth Year.
4. Age = 2025 - 2020 = 5.

Answer: 5. (Correct)

The Extension: Accessing the Real World

An AI's core knowledge is static, and it cannot perform actions on its own. We give our teammate senses and hands through two key techniques:



Retrieval-Augmented Generation (RAG)

Connects the LLM to external, up-to-date knowledge bases (e.g., internal documents, the live web). When asked a question, it first *retrieves* relevant facts and then *generates* an informed answer.

Giving your teammate a library card and a search engine.

Tool Calling

Gives the LLM access to a predefined set of APIs or 'tools' to perform actions. The LLM can determine which tool to use, what inputs to provide, and how to interpret the output.

Giving your teammate a toolkit (e.g., a calculator, a calendar API).

The Ultimate Teammate: The AI Agent

When we combine reasoning, RAG, and tools, we create an **AI Agent**: an autonomous system that can pursue complex goals. It operates in a continuous loop, progressively working towards a solution.

1. Observe

Assess the situation and goal. Gather information (using RAG or tools).

The Agentic Loop
(ReAct Framework)

2. Plan

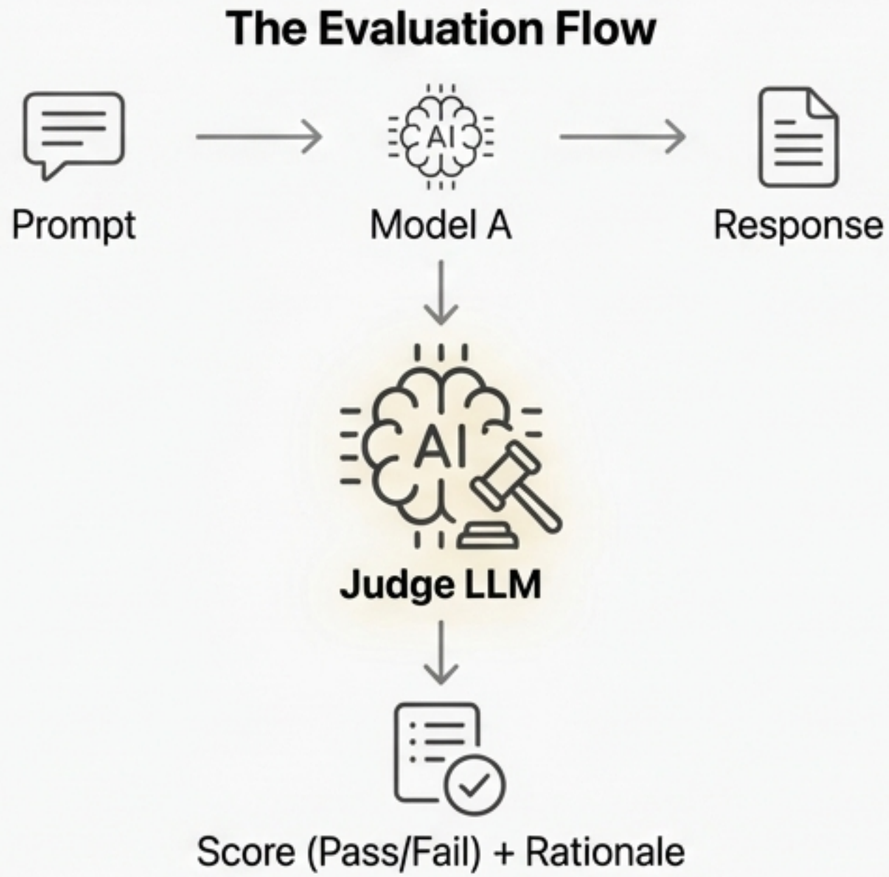
Break down the goal into the next actionable step. Decide which tool to use.

3. Act

Execute a chosen tool with the correct parameters (e.g., call an API).

The Performance Review: How We Measure Success

Evaluating free-form AI responses is challenging. Human evaluation is ideal but slow and expensive. The emerging industry standard is LLM-as-a-Judge: using a highly capable, impartial LLM to evaluate the output of other models.



This provides scalable, consistent, and interpretable feedback. However, we must remain vigilant for biases (e.g., position, verbosity, and self-enhancement bias).

Benchmarking: The AI Combine

To compare models objectively, we rely on standardized benchmarks that test a wide range of capabilities, much like an athletic combine. This allows us to track progress and understand the specific strengths and weaknesses of each AI teammate.

Benchmark	Domain Tested	Description
MMLU	**General Knowledge**	A broad, multi-disciplinary test across 57 subjects to measure retained knowledge.
GSM8K / AIM	**Math Reasoning**	Tests the ability to solve grade-school to competition-level math problems, requiring multi-step logic.
HumanEval / SWE-Bench **SWE-Bench**	**Coding**	Assesses the ability to write functional code to solve real-world software engineering problems.
TAU Bench	**Agentic Tool Use**	A simulated environment to test an agent's ability to use multiple tools to complete complex tasks.

The Strategic Imperative for Apple

We are at the threshold of a new era of computing, defined by partnership with AI.

We have the components to build the most capable, personal, and trustworthy AI teammates in the world. The foundational technology is here. The challenge and opportunity now lie in product vision and execution.

Key Strategic Questions:

1. How do we seamlessly embed these increasingly capable **Agents** into our entire ecosystem—from the OS to our core applications?
2. How do we ensure every AI interaction embodies Apple's core principles of **privacy, intuitive design, and user trust?**
3. What new product categories and user experiences does this new class of **"AI Teammate"** unlock?