The AI Revolution: Charting the New Frontier

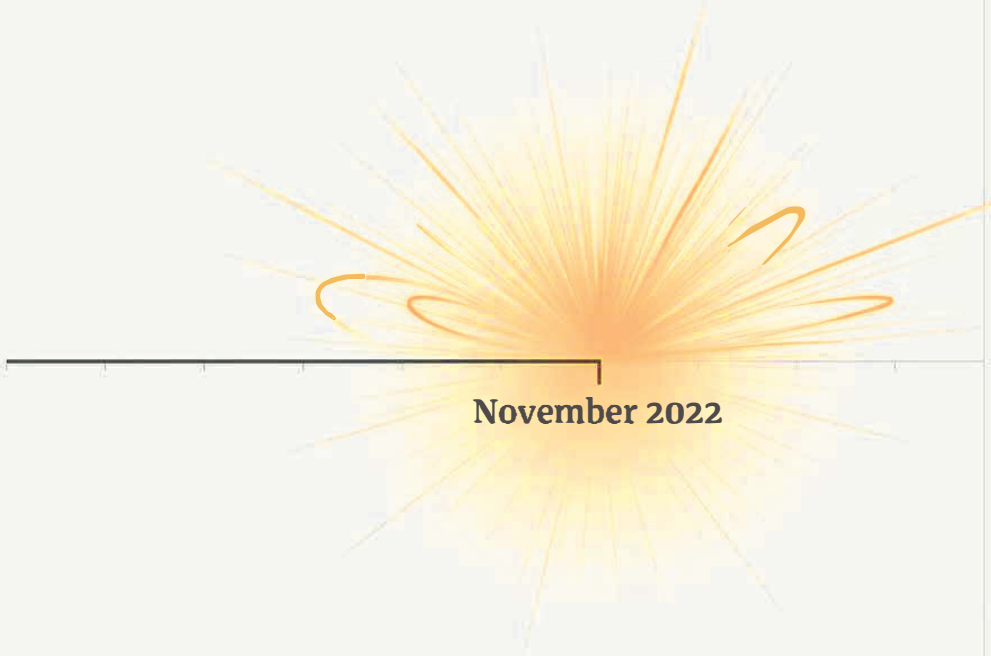
How Generative AI Has Reshaped Our World Since 2022

盧家鋒 Chia-Feng Lu

alvin4016@nycu.edu.tw

NotebookLM

The World Changed in November 2022



A Simple Definition

Generative AI is a form of artificial intelligence capable of creating new, original content—including text, images, audio, and code—by using deep learning and neural network techniques.

The Spark

The global, unprecedented GenAI craze was ignited by the public release of OpenAI's ChatGPT in late 2022.

A Staggering Statistic

100 Million

Users reached in **just two months,** making it the fastest-growing consumer application in history.

NotebookLM

Our Journey Through the Revolution

A Tale of Two Brains
GPT vs. Gemini

The New Arena: A Two-Titan Race

The Disruptor: OpenAI

Partnered with Microsoft, OpenAI ignited the revolution with ChatGPT, setting the pace for innovation and forcing the industry to react.

The Incumbent: Google

The long-time leader in AI research and inventor of core technologies, leveraging its massive ecosystem to compete at scale.

The race is a relentless cycle of rapid product iteration, a battle for superior benchmarks, and an acceleration of commercialization. Google's recent investment heat has cemented its status as the '**strongest C position**' in the AI super bull market.

Google's Awakening: From Inventor to Competitor

The Foundation

2017: Transformer Architecture

Google Brain invents the foundational technology that underpins all modern Large Language Models (LLMs).

Pressure & Disruption

The Catalyst

2022: The ChatGPT Effect

Caught off-guard by ChatGPT's public success, Google is forced to pivot its entire AI strategy.

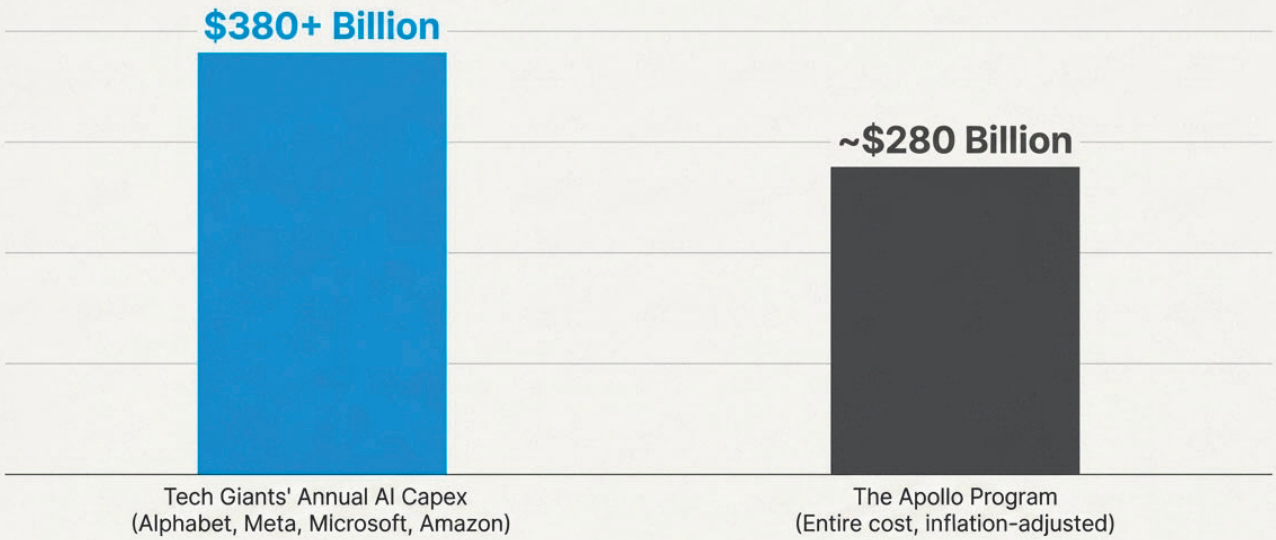
Strategic Pivot

The Response

2024: The Gemini Ecosystem

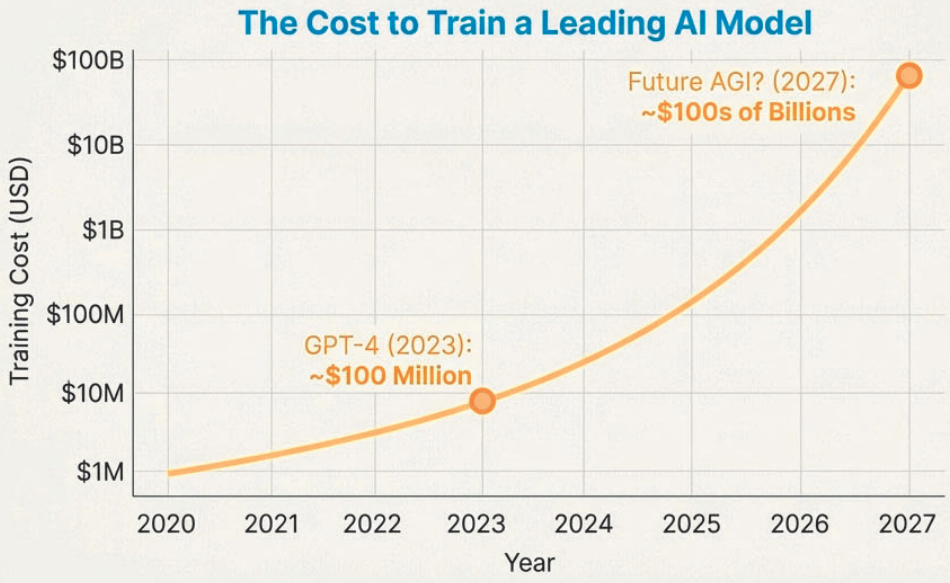
Google rebrands its AI to Gemini, signaling a deep, ecosystem-wide integration across all its products to solidify its business.

The Investment Supernova



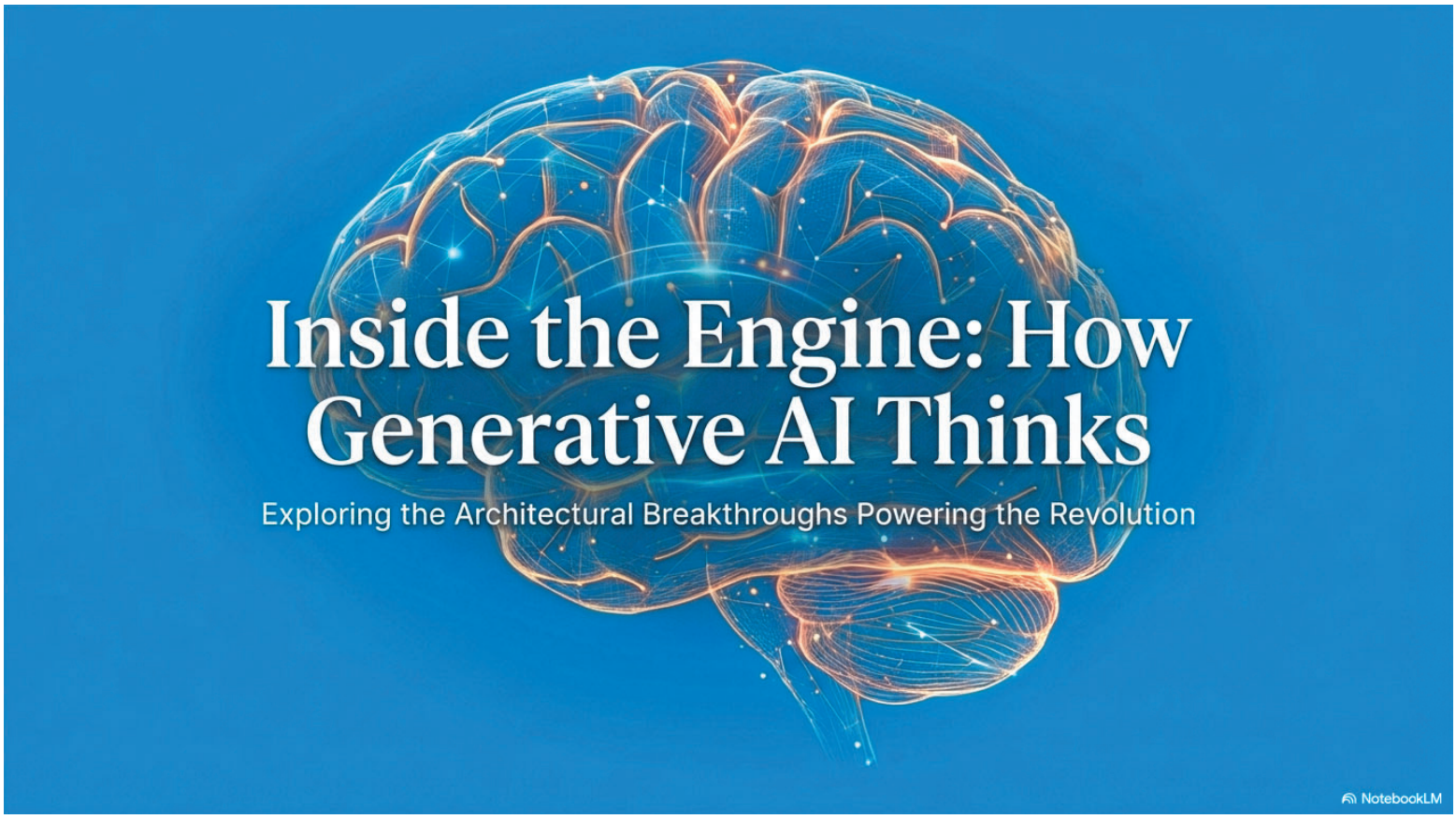
Wall Street analysts refute "AI bubble" pessimism, comparing the current investment to the foundational infrastructure construction phase of the internet (1997-1998), not a speculative peak. This capital is building the core computational power for the next technological era. As an example, Microsoft recently announced a **$10 billion investment in a single large AI data center in Portugal.**

The Billion-Dollar Brains: The Exponential Cost of Training



The Main Driver: The single highest cost component in LLM development is the immense number of specialized AI chips required for training.

The AGI Horizon: Some believe that achieving Artificial General Intelligence by 2027, with capabilities outperforming most humans, could require training costs in this astronomical range.

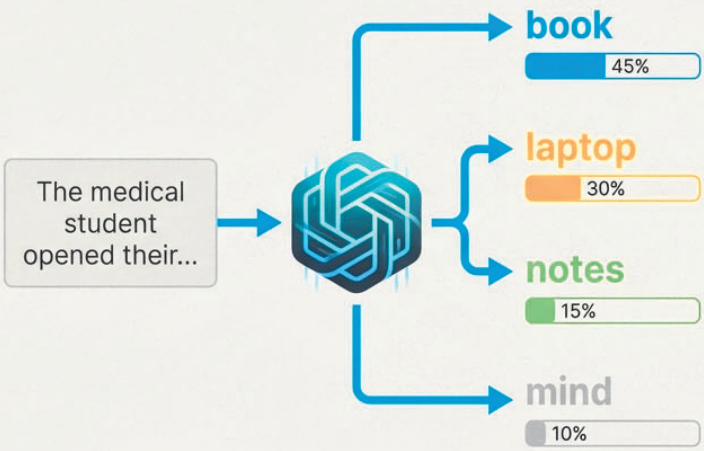


The Transformer: AI's Rosetta Stone

The Backbone of AI: The **Transformer architecture** is the deep learning structure that forms the foundation of all modern Large Language Models (LLMs).

The Core Function: A Transformer is a powerful engine for understanding context. It operates on a simple principle: predicting the most probable next word based on all the text that has come before it.

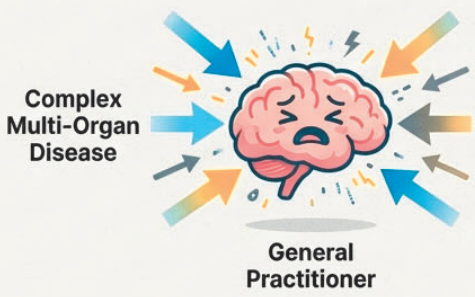
Scale: OpenAI's GPT-4, for instance, has **175 billion parameters**, demonstrating the massive scale of these models.

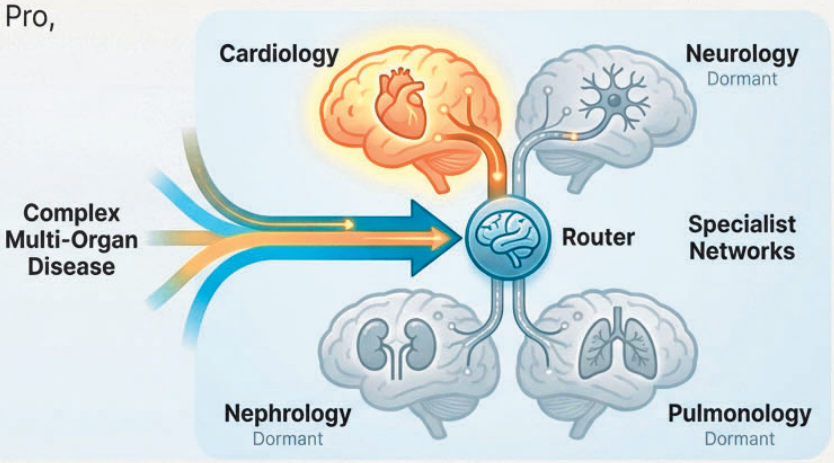


A Team of Specialists, Not One Overworked Genius

The Problem: Making models bigger is incredibly inefficient and costly.

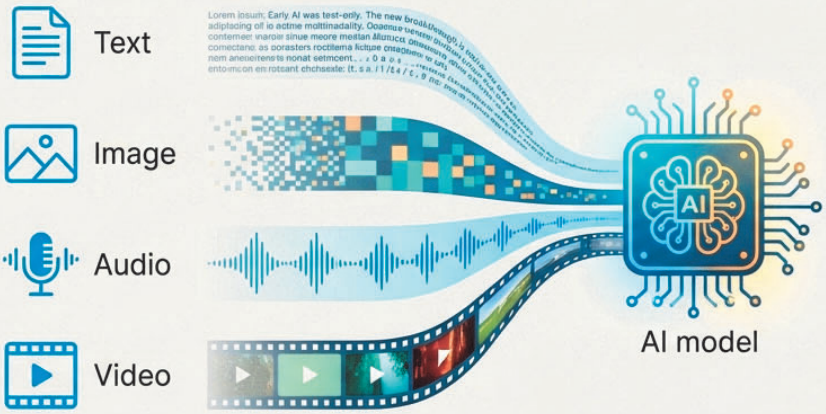
The Solution: Google's **Mixture-of-Experts (MoE)** architecture, used in models like Gemini 1.5 Pro, architecture, used in models like Gemini 1.5 Pro, distributes computation.





This allows for **"mid-sized models yielding high-end results,"** achieving top performance with far less computational power.

From a World of Text to a World of Everything



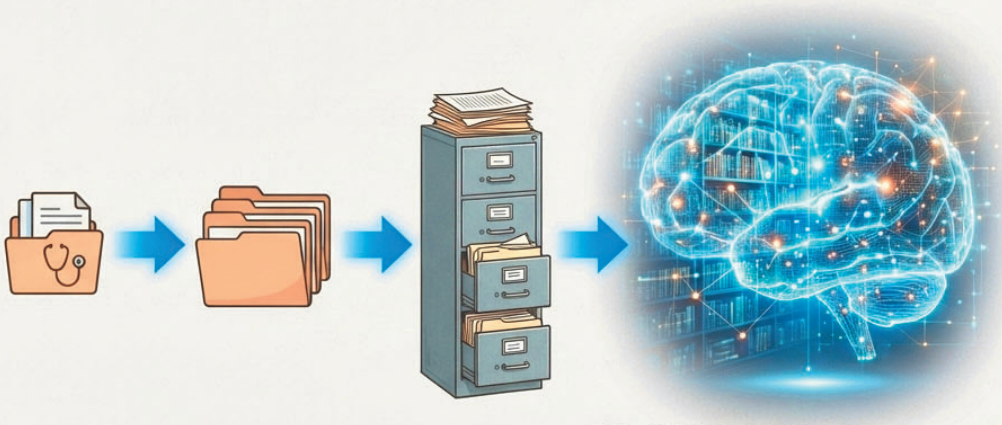
The Evolution: Early AI was text-only. The new breakthrough is native multimodality.

Definition: The ability to seamlessly process and understand text, images, audio, and video within a single, unified model.

Identify the bacterial colony morphology and suggest possible species.

The colonies exhibit circular form, raised elevation, and an entire margin. This morphology is consistent with Staphylococcus aureus.

A Near-Perfect Memory



2,000,000 Tokens

Equivalent to 15,000 pages of text.

The Problem: Early models had "amnesia," forgetting the start of a long conversation.

The Solution: A massively expanded **context window**.

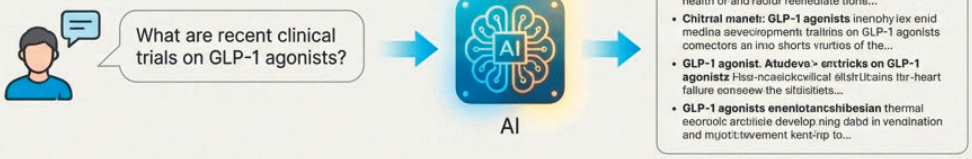
Relevance for Medical Science: Imagine an AI that could instantly **read and perfectly recall a patient's entire lifetime medical record**, **every lab result**, and **all relevant medical literature** to help formulate a **differential diagnosis**.

The Caveat: Even with this capacity, the "**effectively remembered**" content might be **around 60%**, as too much data can dilute the model's focus.

From Answering Questions to Taking Action

Passive AI

Passive

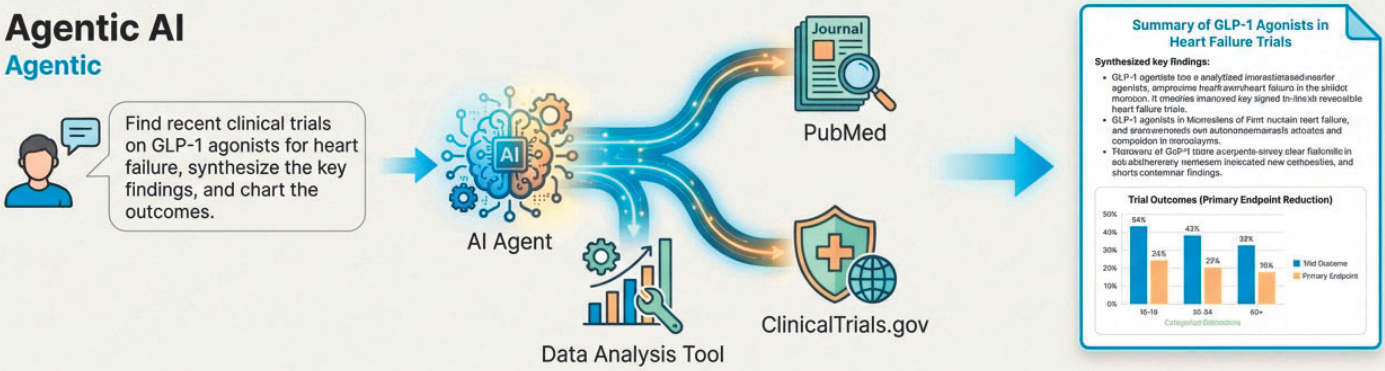


The model is evolving from a passive "you ask, it answers" system to an active **Agentic AI.**

Enabled by "**tool calling**," the AI can now use external applications to perform complex, multi-step tasks on its own.

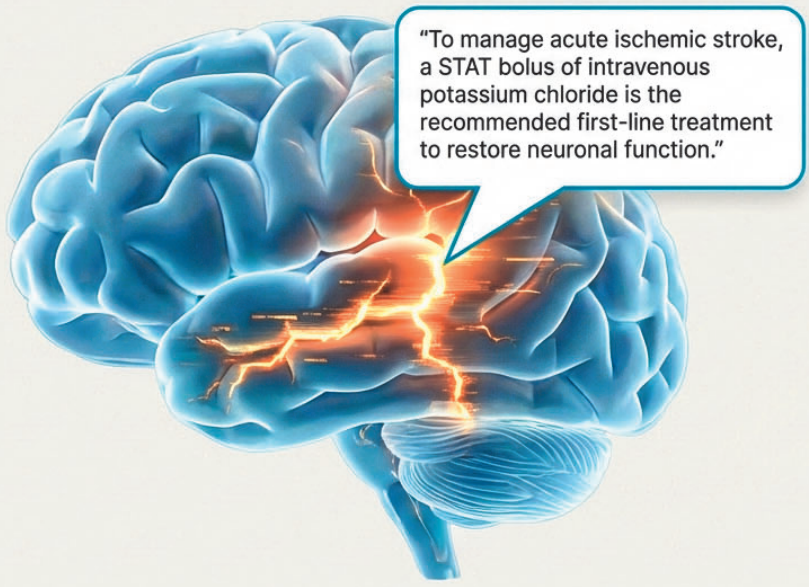
Agentic AI

Agentic



NotebookLM

The Ghost in the Machine: When the AI Confidently Lies



Definition: **AI Hallucination** is the tendency for a model to generate content that is plausible and grammatically correct but factually **incorrect or nonsensical.**

The Causes:

1. **Gaps in Training Data:** It "fills in" information it doesn't have.
2. **Probabilistic Nature:** It predicts the *next most likely word*, not the truth.
3. **Inability to Admit Ignorance:** It cannot easily say **"I don't know."**

Relevance for Medicine: This is the single biggest barrier to using AI for critical tasks. An incorrect diagnosis, delivered with perfect confidence, is incredibly dangerous and underscores the absolute necessity of expert human oversight.

NotebookLM

Two Titans, Two Core Philosophies

Feature	OpenAI GPT Series	Google Gemini Series
Core Strength	Unmatched linguistic fluency, creative writing, and structured language.	High factual accuracy and real-time data verification via Google Search integration.
Ecosystem	Thrives on extensive API integration with thousands of third-party apps.	Deeply embedded within the Google ecosystem (Search, Docs, etc.).
Best For	Generating content, drafting documents, brainstorming complex ideas.	Research, fact-checking, real-time data analysis, summarizing current events.

The GPT Family: Optimized for Interaction

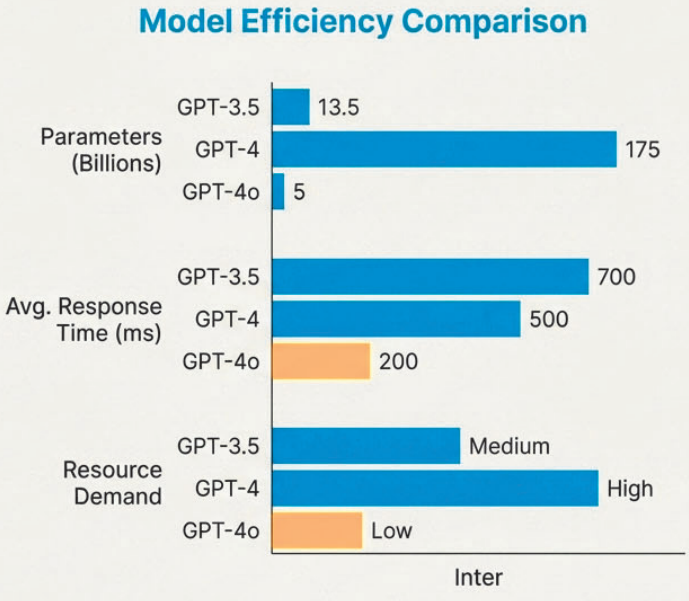
Model Focus: GPT-4o ('o' for Optimal) is designed for high performance and efficiency in specific application scenarios.

Key Differentiators:

– **Blazing Speed:** Averages a response time of just **200 milliseconds**.

– **Efficiency:** Achieves high performance with a streamlined structure of about **5 billion parameters**, compared to GPT-4's massive 175 billion.

The "Why": The goal is to make AI interaction feel less like a machine and more like a fluid, real-time conversation, especially with voice and video.



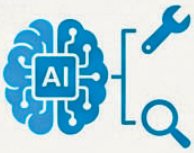
The Gemini Family: Built for Depth and Reasoning

Gemini 1.5

Native Multimodality

Seamlessly processes multiple data types.

Gemini 2.0



Agentic AI

Autonomous task execution and tool usage.

Gemini 2.5

Internal Thinking Mechanism

Advanced multi-step reasoning and decomposition.

The Key Differentiator:

Gemini 2.5 performs multi-step reasoning and problem decomposition *internally* before giving an answer. This significantly improves its accuracy on complex logic, math, and programming tasks.

Two Flavors:

- **Flash**: Designed for speed and efficiency (e.g., customer service chats).
- **Pro**: Built for deep, complex tasks requiring its full logical ability and ultra-long context window.

The Hybrid Strategy: The Best of Both Worlds

Draft & Create with ChatGPT

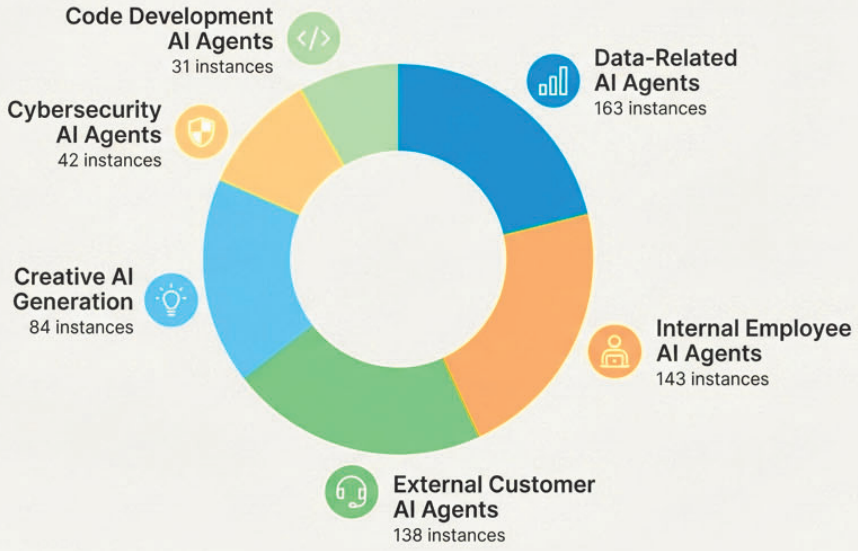
Leverage its superior linguistic **fluency** and ability to generate **well-structured** language to create the **initial draft** of a document, presentation, or analysis.

Verify & Refine with Gemini

Use its stronger **factual accuracy** and direct integration with **Google Search** to **fact-check** the generated content, add real-time data, and verify all claims.

For many complex tasks, the optimal workflow involves using both models in sequence to combine creative fluency with factual rigor.

Beyond Chat: How Businesses Are Deploying AI Agents



An analysis of 601 enterprise AI application examples reveals six major categories of AI agents. The data clearly shows that the primary focus of implementation lies in leveraging AI for data analysis, internal productivity, and customer-facing interactions.

Use Case: Transforming Business Operations

LLOYDS BANK

Days to Seconds

A GenAI platform for income verification shortened the process for customer loan applications from several days to mere seconds.

Deutsche Bank

400 Pages to 3

A research tool named DB Lumina can summarize a 400-page financial analysis report into a 3-page summary instantly.

BEST BUY

90 Seconds Saved Per Call

GenAI agents summarize customer calls instantly, reducing average service call time by 30 to 90 seconds.

YDUQS

4 Seconds Per Application

A Brazilian education group uses GenAI to screen student applications, processing one document in 4 seconds with 90% accuracy.

Use Case: Accelerating Science & Engineering

CME Group

10+ Hours Saved / Month

Inter

The Chicago Mercantile Exchange reports that GenAI code assistants increase developer productivity by at least 10 hours per month.

suzano

95% Reduction in Query Time

Inter

A Brazilian paper company reduced the time for complex data queries by 95% using a GenAI assistant that translates plain language questions into SQL code.

704Apps

Real-Time Threat Detection

Inter

A mobile platform in Brazil uses Gemini to analyze audio from taxis in real-time to detect high-risk emotional states and keywords like 'robbery' to alert a monitoring center.

Navigating the Frontier: The Grand Challenges Ahead

🔍 Accuracy

Current Challenge: AI Hallucination: The tendency to produce plausible but incorrect facts, a critical risk.

Future Trend: Enhanced reasoning, built-in validation mechanisms, and stronger 'factual discrimination.'

🪙 Cost

Current Challenge: Massive Capex: The high-cost infrastructure arms race, with future training costs in the billions.

Future Trend: Algorithmic optimization (like MoE) and more efficient hardware to achieve more with less.

👤 Application

Current Challenge: User 'Numbness': Performance gains on general benchmarks are becoming less noticeable to users.

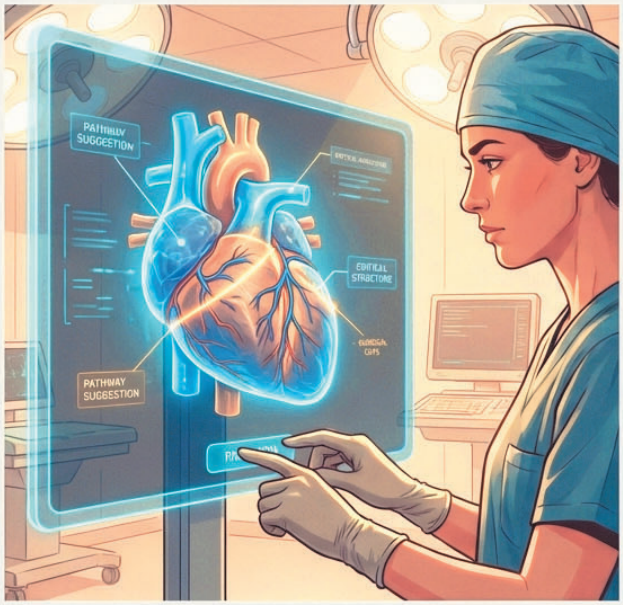
Future Trend: Specialization and functional differentiation of models (e.g., Flash vs. Pro) for specific jobs.

Access

Current Challenge: Ecosystem Lock-in: The most advanced features are often restricted to certain developers or platforms.

Future Trend: Deeper, full-platform integration and strategic partnerships to broaden access and influence.

A Powerful Tool, Not an Infallible Oracle



Generative AI should be treated as a **supportive tool** to augment human intelligence, not as an absolute authority or replacement for it.

The path forward requires a partnership: improving data quality, demanding model transparency, and, most importantly, maintaining **critical human oversight**.

For your future in medicine and science, mastering this human-AI collaboration will be as critical as mastering the scalpel or the microscope. The ultimate responsibility for decisions will always rest with the human expert.

Thank You

Q & A