

To know **where we're going**,  
we need to know **where we've come from**

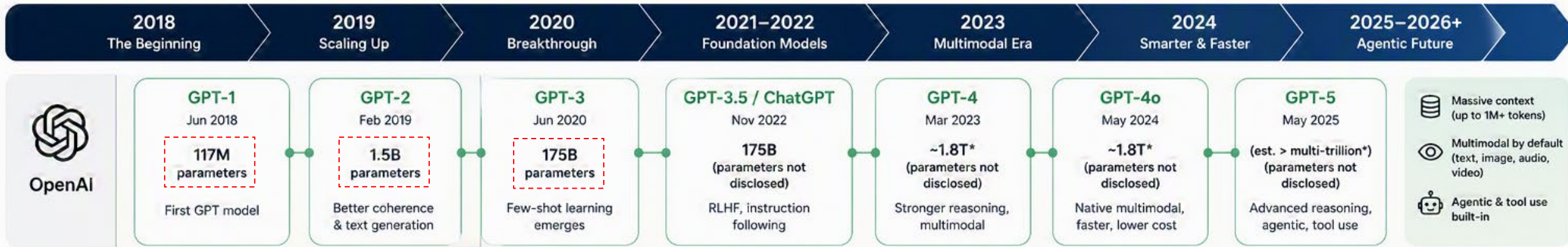


WHERE WE'VE COME FROM

WHERE WE'RE GOING

# AI MODEL EVOLUTION: MAJOR MODEL FAMILIES

From early language models to today's multimodal, reasoning and agentic systems



A **parameter** is something an AI model learnt during training.

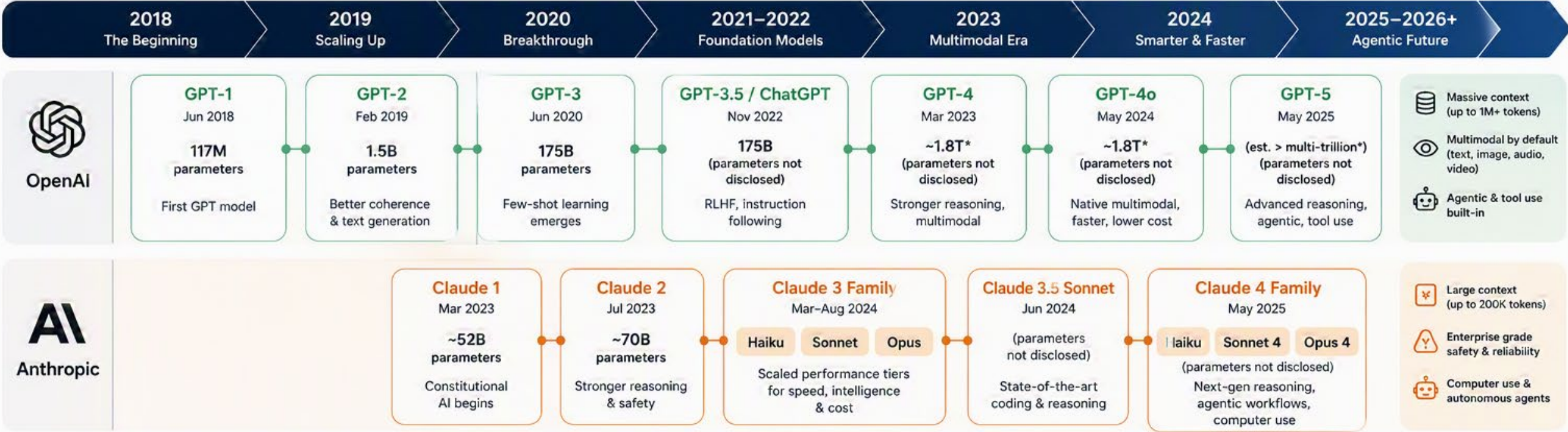


Brain learns through connections between neurons.

An AI model learns through mathematical "weights."

# AI Model Evolution: Major Model Families

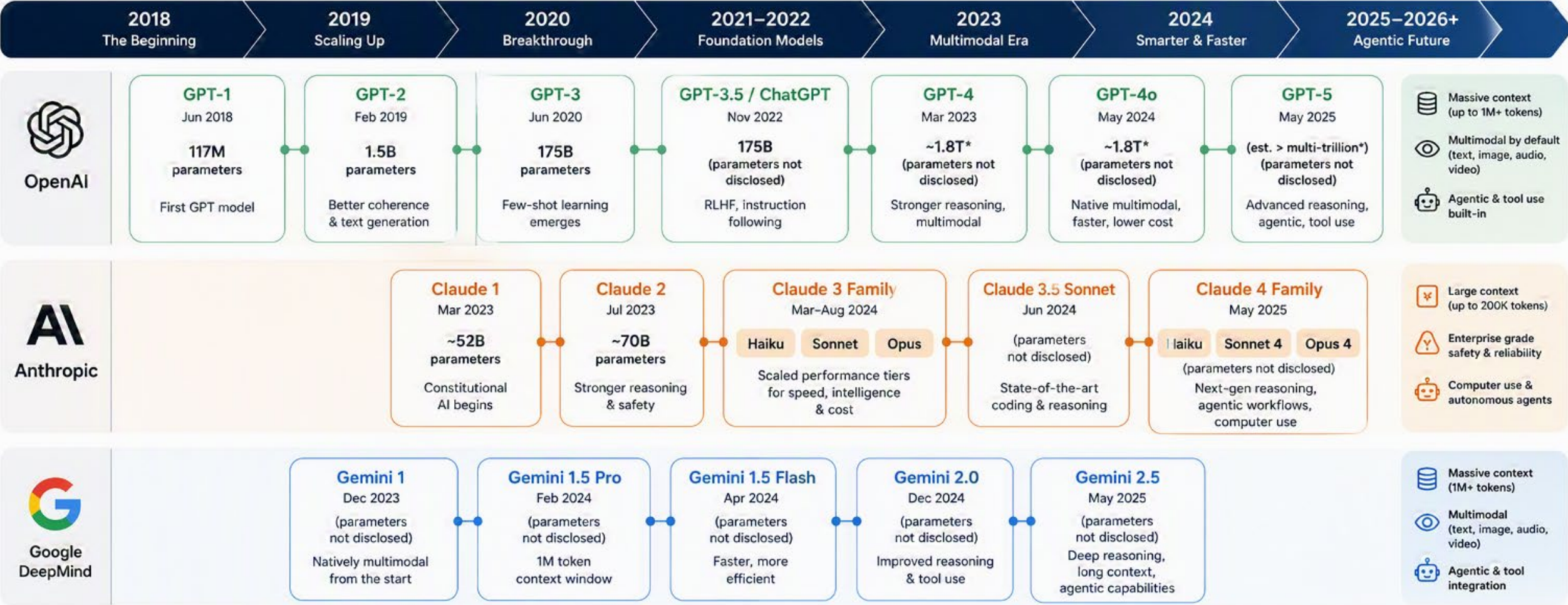
From early language models to today's multimodal, reasoning and agentic systems





# AI Model Evolution: Major Model Families

From early language models to today's multimodal, reasoning and agentic systems





# AI Model Evolution: Major Model Families

From early language models to today's multimodal, reasoning and agentic systems





OpenAI

|                                                                |                                                                                   |                                                                          |                                                                                                        |                                                                                                 |                                                                                                         |                                                                                                                          |                                                                                                                                                                                      |
|----------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>GPT-1</b><br>Jun 2018<br>117M parameters<br>First GPT model | <b>GPT-2</b><br>Feb 2019<br>1.5B parameters<br>Better coherence & text generation | <b>GPT-3</b><br>Jun 2020<br>175B parameters<br>Few-shot learning emerges | <b>GPT-3.5 / ChatGPT</b><br>Nov 2022<br>175B (parameters not disclosed)<br>RLHF, instruction following | <b>GPT-4</b><br>Mar 2023<br>~1.8T* (parameters not disclosed)<br>Stronger reasoning, multimodal | <b>GPT-4o</b><br>May 2024<br>~1.8T* (parameters not disclosed)<br>Native multimodal, faster, lower cost | <b>GPT-5</b><br>May 2025<br>(est. > multi-trillion*) (parameters not disclosed)<br>Advanced reasoning, agentic, tool use | <ul style="list-style-type: none"><li>Massive context (up to 1M+ tokens)</li><li>Multimodal by default (text, image, audio, video)</li><li>Agentic &amp; tool use built-in</li></ul> |
|----------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|



Anthropic

|                                                                            |                                                                               |                                                                                                                            |                                                                                                           |                                                                                                                                                      |                                                                                                                                                                                  |
|----------------------------------------------------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Claude 1</b><br>Mar 2023<br>~52B parameters<br>Constitutional AI begins | <b>Claude 2</b><br>Jul 2023<br>~70B parameters<br>Stronger reasoning & safety | <b>Claude 3 Family</b><br>Mar-Aug 2024<br>Haiku   Sonnet   Opus<br>Scaled performance tiers for speed, intelligence & cost | <b>Claude 3.5 Sonnet</b><br>Jun 2024<br>(parameters not disclosed)<br>State-of-the-art coding & reasoning | <b>Claude 4 Family</b><br>May 2025<br>Haiku   Sonnet 4   Opus 4<br>(parameters not disclosed)<br>Next-gen reasoning, agentic workflows, computer use | <ul style="list-style-type: none"><li>Large context (up to 200K tokens)</li><li>Enterprise grade safety &amp; reliability</li><li>Computer use &amp; autonomous agents</li></ul> |
|----------------------------------------------------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|



Google DeepMind

|                                                                                                 |                                                                                            |                                                                                             |                                                                                              |                                                                                                                   |                                                                                                                                                                    |
|-------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Gemini 1</b><br>Dec 2023<br>(parameters not disclosed)<br>Natively multimodal from the start | <b>Gemini 1.5 Pro</b><br>Feb 2024<br>(parameters not disclosed)<br>1M token context window | <b>Gemini 1.5 Flash</b><br>Apr 2024<br>(parameters not disclosed)<br>Faster, more efficient | <b>Gemini 2.0</b><br>Dec 2024<br>(parameters not disclosed)<br>Improved reasoning & tool use | <b>Gemini 2.5</b><br>May 2025<br>(parameters not disclosed)<br>Deep reasoning, long context, agentic capabilities | <ul style="list-style-type: none"><li>Massive context (1M+ tokens)</li><li>Multimodal (text, image, audio, video)</li><li>Agentic &amp; tool integration</li></ul> |
|-------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|

### CAPABILITY EVOLUTION (Across All Model Families)



1. Language Generation

2. Reasoning & Understanding

3. Multimodal Intelligence

4. Tool Use & Function Calling

5. Agentic Workflows

6. Computer Use & Automation

7. Autonomous Digital Workers

### ABOUT PARAMETERS

Parameters are the learnable weights inside an AI model.

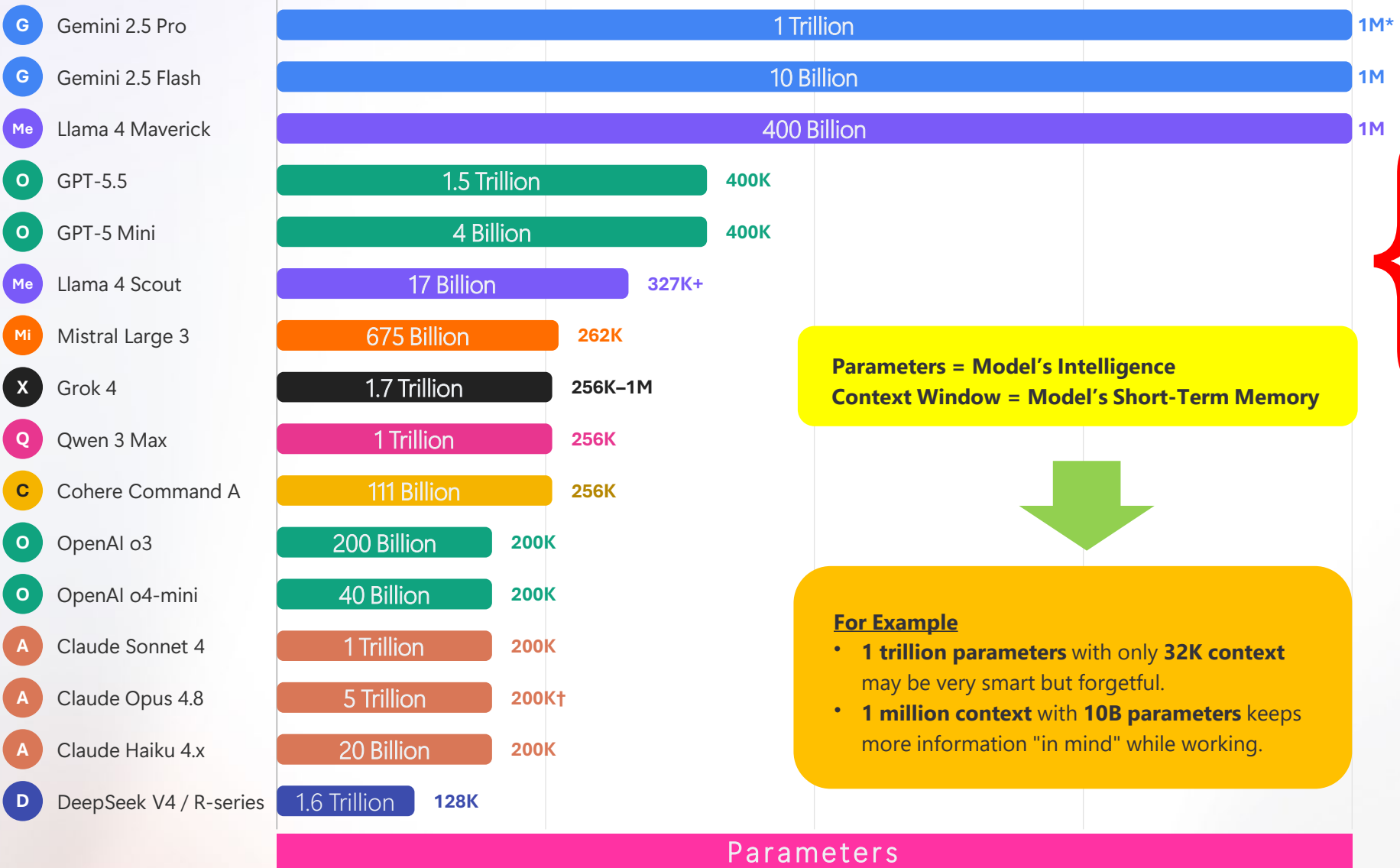
- More parameters can mean more capability—but not always better efficiency.
- Many leading models no longer disclose exact counts.

### KEY TRENDS 2025+

- Reasoning models become the new default (thinking before answering)
- Longer context windows unlock more powerful use cases
- Multimodal + real-time understanding everywhere

# Context Window | How Much Can Each Model Remember?

Context Window



Parameters = Model's Intelligence  
Context Window = Model's Short-Term Memory



**For Example**

- **1 trillion parameters** with only **32K context** may be very smart but forgetful.
- **1 million context** with **10B parameters** keeps more information "in mind" while working.

| Context Window       | Tokens  |
|----------------------|---------|
| System Instructions  | 5,000   |
| Conversation History | 20,000  |
| Attached Documents   | 150,000 |
| Prompt Quality       | 5,000   |
| Model's Response     | 20,000  |
| Total                | 200,000 |

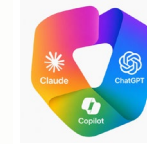
| Context | Approx Content                                                |
|---------|---------------------------------------------------------------|
| 128K    | ~300 text pages                                               |
| 200K    | ~500 text pages                                               |
| 400K    | ~1,000 text pages                                             |
| 1M      | - Several Books<br>- Large Codebase<br>- Years of Transcripts |
| 2M      | Small Library                                                 |

\*Gemini 2.5 Pro reaches 2M and †Claude Opus 4.x reaches 1M in some configurations. Grok 4 spans 256K-1M by variant/platform.

# Coming Soon - Microsoft AI Models (1<sup>st</sup> Party)

| Model                                                                                                        | Think of it as...  | Human Role          | Use it when you need to...                                    |
|--------------------------------------------------------------------------------------------------------------|--------------------|---------------------|---------------------------------------------------------------|
|  <b>MAI Thinking-1</b>      | The strategist     | Analyst             | Work through complex problems, plans, decisions, or analysis. |
|  <b>MAI Image-2.5</b>       | The designer       | Graphic Designer    | Create polished images, slides, concepts, or visual assets.   |
|  <b>MAI Image-2.5 Flash</b> | The quick designer | Production Designer | Produce lots of images quickly, where speed matters most.     |
|  <b>MAI Transcribe-1.5</b>  | The note taker     | Court Reporter      | Turn meetings, calls, or recordings into accurate text.       |
|  <b>MAI Voice-2</b>         | The speaker        | Voice Actor         | Turn written content into natural-sounding speech.            |
|  <b>MAI Voice-2 Flash</b> | The live speaker   | Call Centre Agent   | Support real-time voice conversations, bots, or assistants.   |
|  <b>MAI Code-1 Flash</b>  | The developer      | Software Developer  | Write, review, fix, or improve software code.                 |

# Microsoft: The Complete Family of Copilots

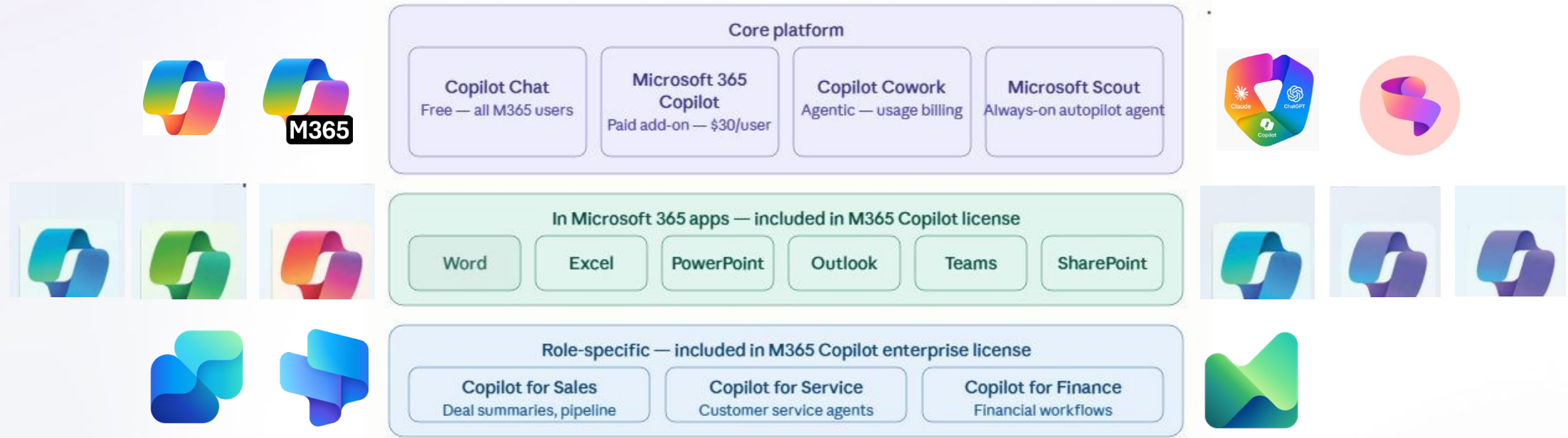




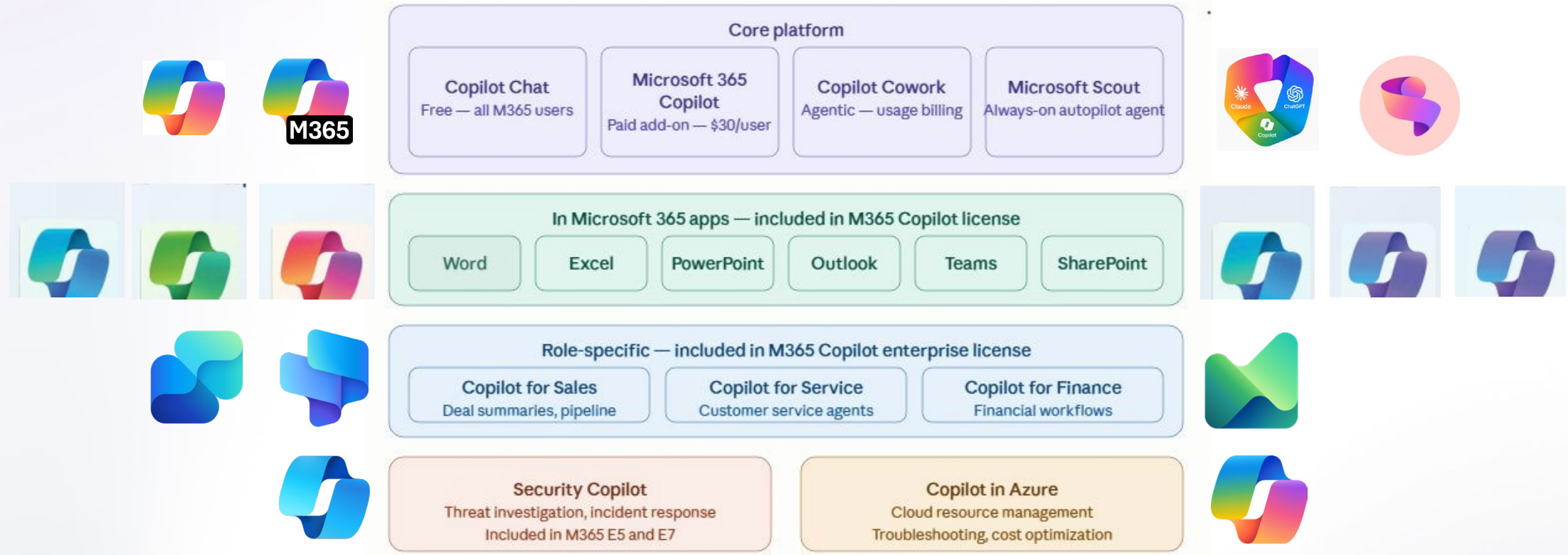
# Microsoft: The Complete Family of Copilots



# Microsoft: The Complete Family of Copilots

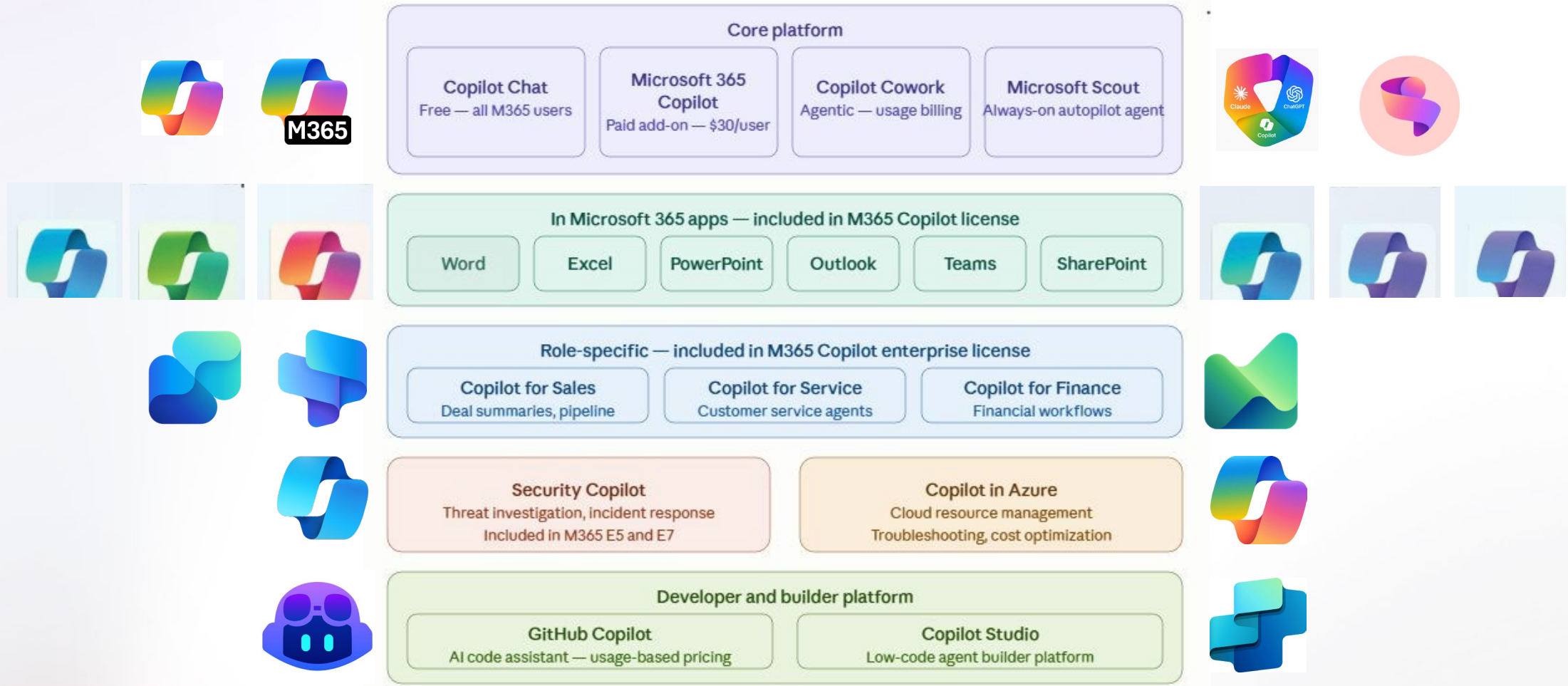


# Microsoft: The Complete Family of Copilots

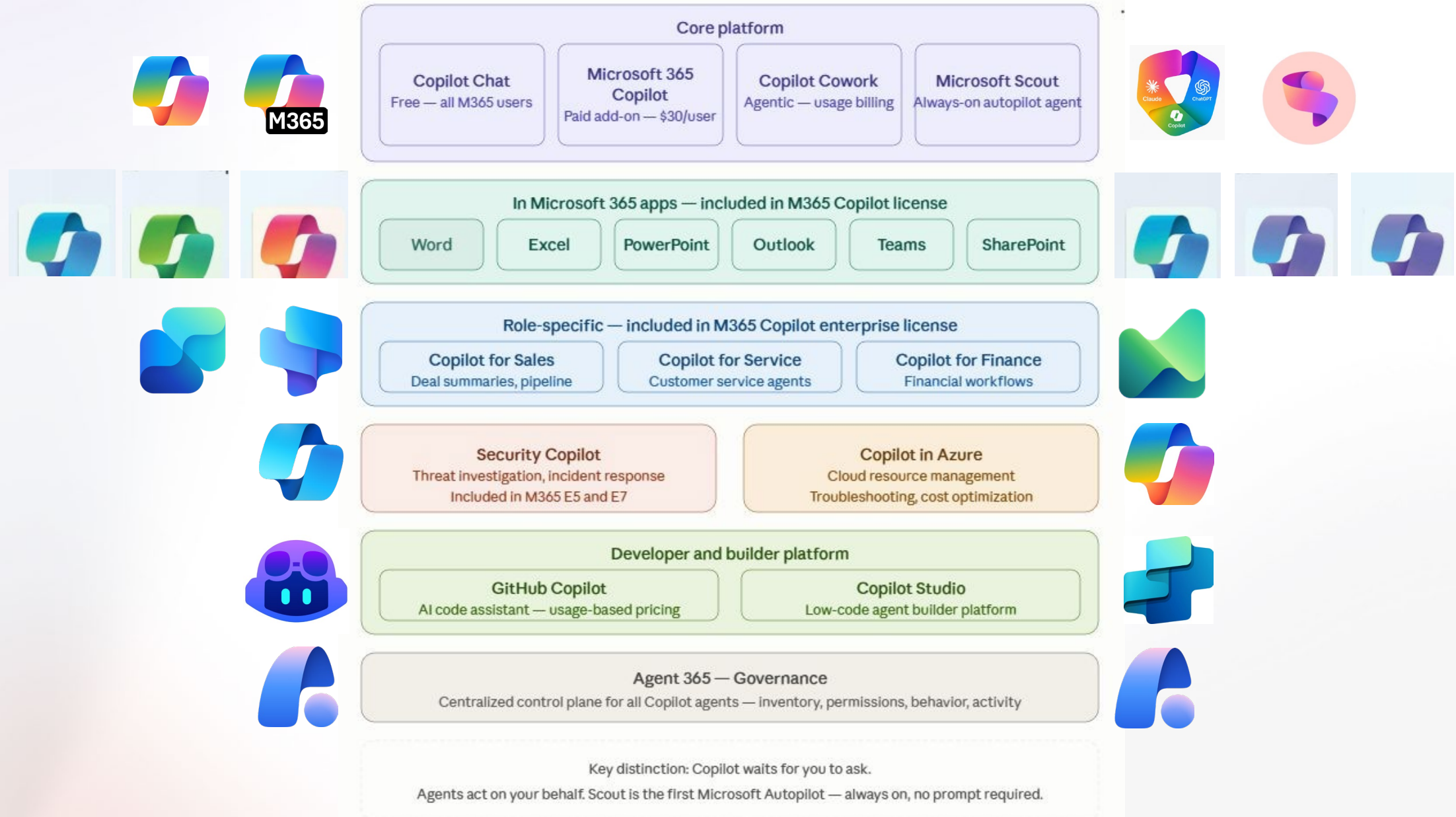




# Microsoft: The Complete Family of Copilots



# Microsoft: The Complete Family of Copilots



# Work IQ | The Secret Sauce



## Microsoft 365 Copilot

### AI embedded across the flow of work

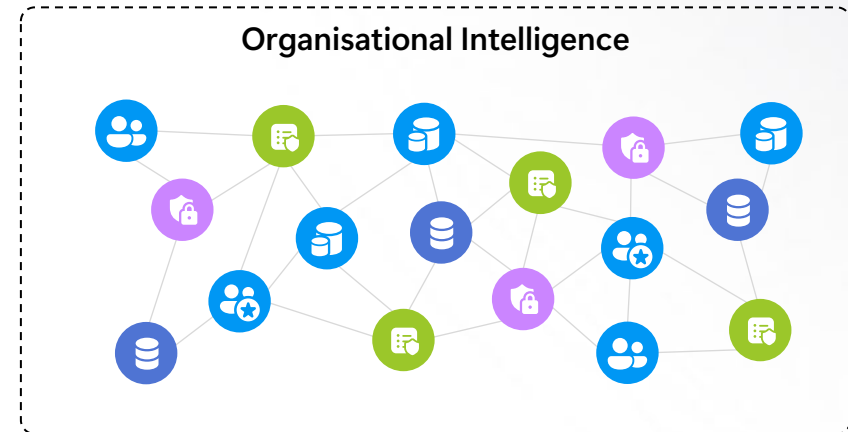
Copilot brings AI directly into everyday work—turning intent into action across documents, meetings, email, and collaboration.



## Cowork

### Execution in action

Cowork is where Copilot executes long-running work. It plans steps, reasons across apps and data, and carries work forward over time — producing complete outputs rather than single responses. Work is observable, controllable, and persists beyond a single interaction.



## Work IQ

### The intelligence layer

Work IQ grounds Copilot and agents in real business context — connecting data, relationships, and work patterns to enable accurate reasoning and execution. It goes beyond raw data access, continuously building context so AI can make better decisions and take the right actions.



# The Cost of Parameters & Context

## Copilot Cowork Tasks Types



### Light Tasks

Leverages a handful of knowledge sources for a light-reasoning scenario like a summary, usually resulting in one short deliverable.

- Few sources
- Light reasoning
- 1 or fewer outputs

#### PROMPT EXAMPLE

*"Create a weekly status update for my team that runs every Monday morning – including any unresolved priorities and a short list of key executive meetings pulled from my calendar."*

EST. COPILOT CREDITS

100 – 300



### Medium Tasks

References many sources over varying time frames, uses structured reasoning, and produces multiple deliverables.

- Many sources
- Structured reasoning
- 2+ outputs

#### PROMPT EXAMPLE

*"Help me prepare for a customer meeting by pulling relevant emails, calendar items, and recent files into a briefing doc based on our template; an Excel overview of sales and sales trends; and a client-ready presentation."*

EST. COPILOT CREDITS

300 – 700



### Heavy Tasks

Requires deep analysis and broad source aggregation over long time frames, generating many outputs.

- Broad aggregation
- Deep reasoning
- Many outputs

#### PROMPT EXAMPLE

*"Analyse 6 months of expected product usage data to understand usage patterns, classify prospects, and produce a leadership-ready report..."  
(shortened)"*

EST. COPILOT CREDITS

> 700

# Microsoft Copilot: 4 Levels of Autonomy

From hands-on-the-wheel to full autopilot | who drives the work?



Copilot

✓

ChatCowork

+

 New task

Search

Scheduled

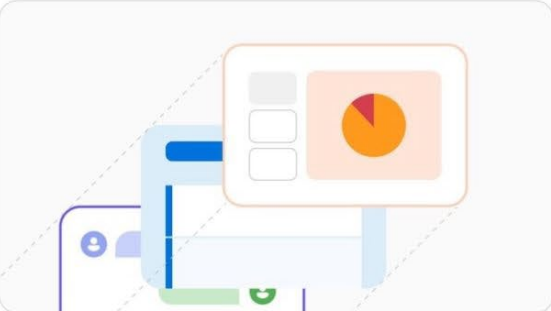
Customize

# You make the decisions, Cowork takes care of the details



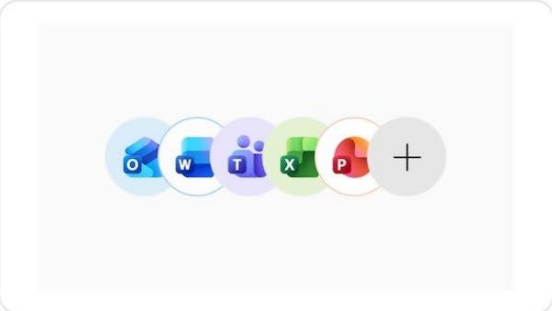
**Plan and complete work**

Cowork plans, organizes, and completes tasks using your files and tools.



**Manage multiple tasks at once**

Cowork keeps work moving across projects, so you don't have to manage it all yourself.



**Use your data and apps**

Cowork delivers results tailored to you—while keeping everything secure and compliant.

Request access



# Target Personas for Copilot Cowork

AS ASSESSED DURING PREVIEW

## User Personas



### Corporate Knowledge Workers

- Juggle shifting priorities
- Heavy email and chat load
- Create docs, decks, and spreadsheets



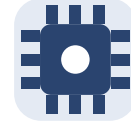
### Management & Senior Leaders

- Need headlines, risks, and recommendations
- Decide more than create
- Calendar packed with high-stakes meetings



### Customer Facing Knowledge Workers

- Balance external customers and internal teams
- Organised by accounts, not projects
- Anchored in relationship history

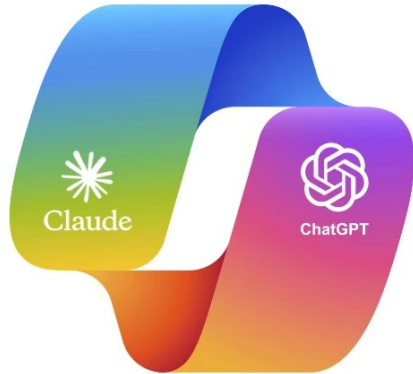


### Technical Workers

- Need deep focus blocks
- Produce code, designs, or systems
- Work mostly outside Office apps

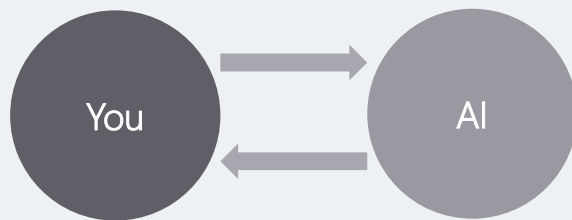
# To utilise Cowork effectively, we must learn how to **delegate**:

## Question → Response



### ASSISTANCE

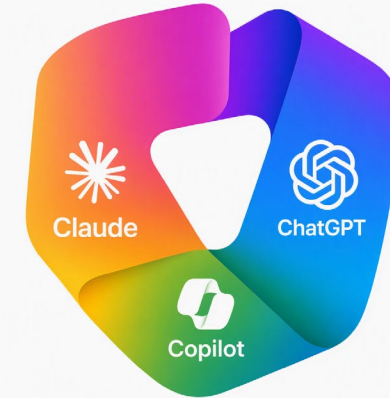
How most people use AI



You → AI → You → AI → You...

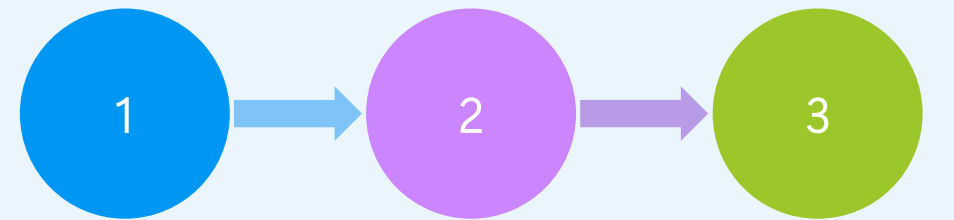
*You remain the "executor" of every step.*

## Job → Delivery



### DELEGATION

Injecting AI in the flow of work



You brief

Cowork executes

Result returned

*You move from the "doer", to the director.*

You may need to steer Cowork based on outputs and simultaneously while on a mobile device – trigger here, see results there.

# Chat and Cowork: Work IQ guidelines when to use each



## Copilot Chat

Your secure AI for work which is grounded in your data (Work IQ) to answer, draft, and spark ideas.

---



### Best for

Fast, focused tasks in a single conversation.



### Speed

Quick - seconds to minutes – single prompt



### Task complexity

Single step, single session.



### Use when

You need a quick hand: a draft, an answer, an idea.

### Scenarios

- Draft & iterate on content
- On-demand daily briefing
- Summarise project status



## Copilot Cowork

Your secure AI that gets work done. It is an agent that takes action across Microsoft 365 (Work IQ).

---



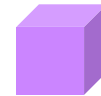
### Best for

Complex, multi-step, outcome driven



### Speed

Thorough — minutes to hours, autonomously.



### Task complexity

Multi-task across apps, tasks & sources.



### Use when

You need Copilot to go do the work, end to end.

### Scenarios

- Clean up your inbox & calendar
- Run multi-step workflows across systems
- Prep for meetings & business reviews

# Delegation to AI works the same way as to a human

**01**

**Delegate the outcome, not the steps.**

Say what “done” looks like. Let the doer find the path.

**02**

**Hand off the whole job.**

If you stay in the loop at every step you are still the one doing it.

**03**

**Brief once, brief well.**

Context, materials, and constraints all go in up front.

**04**

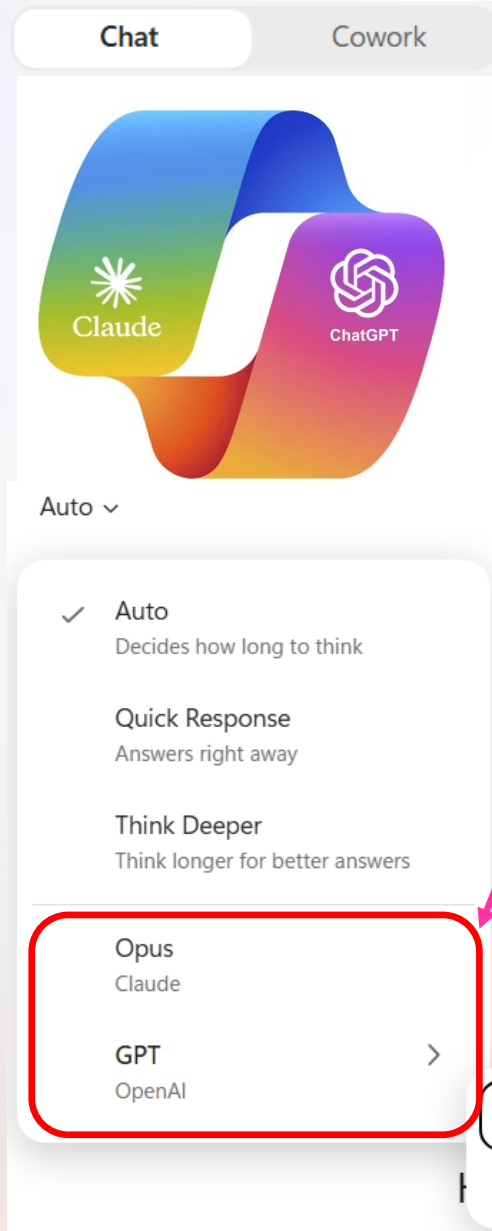
**Be reachable, but don't hover.**

Be available to steer, then let the work run.

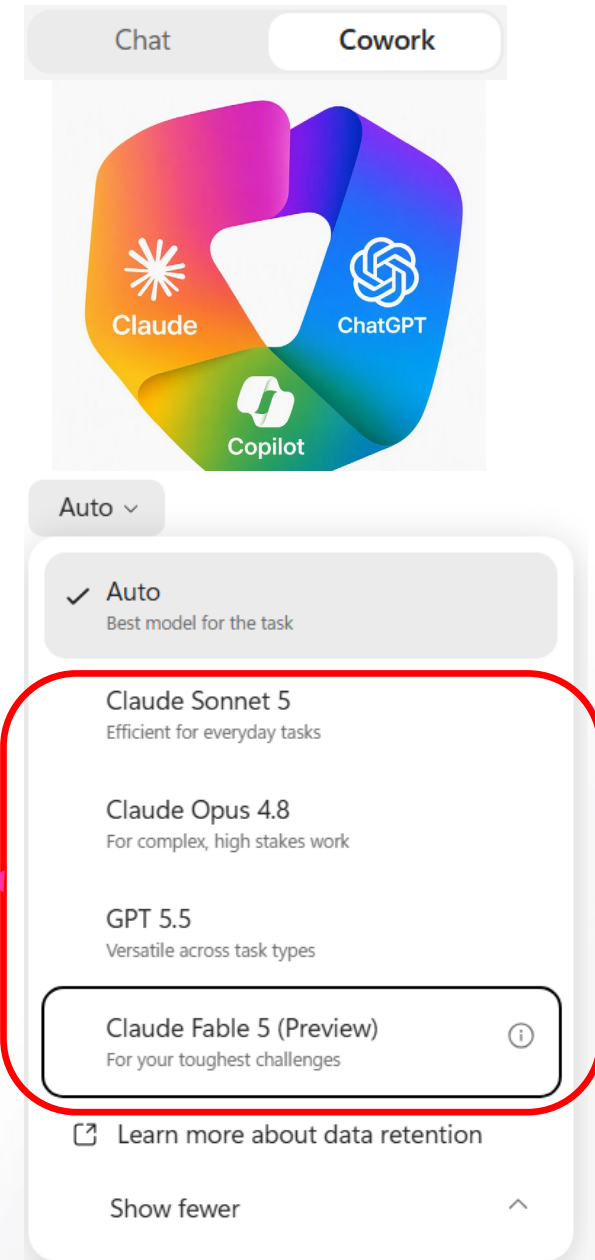
*With AI (specifically Cowork) our role is to steer it in the right direction.*



# Models: USL v Usage



Same Model Family  
Different Model Capability



# Cowork picks the right model for each task

## Auto

**Cowork picks the best model for your task. This is the one that matters.**

For most people, most of the time, leave it here.

## Claude Opus 4.8

**High-stakes reasoning**

Deep analysis. Serious writing. Complex, multi-step.

## Claude Sonnet 4.6

**Fast everyday tasks**

Drafting. Quick lookups. Routine work.

## GPT 5.5 Frontier

**Verbose writing**

Long-form. Citation-heavy research.

## ✗ Wrong Tool

Using a sledgehammer to crack an egg

*Overkill. Messy. Inefficient.*



**Using the wrong model for the task**

Leads to unnecessary complexity, higher cost, and poor outcomes.

## ✓ Right Tool

Crack an egg by tapping two eggs together

*Simple. Effective. Just right.*



**Using the right model for the task**

Delivers the best results with efficiency, accuracy, and ease.



**Right Model. Right Task. Better Results.** Use the model that fits the task

# Same Cowork task, two different ways

Expensive

Optimised

## Lazy Brief

**"Organize my inbox"**

> A generic guess and you won't even know it guessed.



**VAGUE**

**"Look into our sales pipeline."**

Pulls in everything, runs long, costs more.

## Sharp Brief

**"Summarise unread emails from the last 3 days, grouped by project, flag what needs my reply this week, as a short bulleted list, most urgent first."**

> Finished, usable work.



**CLEAR**

**"List the 5 oldest open deals over \$50k and the last action on each."**

Faster, cheaper, and a more useful answer.

# Which model, and when

Identical prompt. Identical outputs — a deck, a spreadsheet, a dashboard. The credits burned varied 2.7×

## Claude Sonnet 4.6

*The workhorse*

398 credits



### Best for

Everyday drafting, reports, running skills and agent-mode work — about 80% of daily tasks.

### Avoid when

The task needs deep, multi-step reasoning.

## Claude Opus 4.8

*The heavyweight*

477 credits



### Best for

Strategic analysis, legal or regulatory drafts, and debugging problems nobody else can crack.

### Avoid when

The task is routine — it's overkill.

## GPT-5.5

*The generalist*

1,069 credits



### Best for

Research-and-write and web-grounded work that needs broad, fresh knowledge.

### Avoid when

Routine doc or skill work — Sonnet does it for ~40% of the cost.

**Takeaway** Sonnet handles the everyday 80% for about a third of GPT-5.5's credits. Match the model to the task and let Auto orchestrate when you're unsure.



# Knowing when NOT to delegate

**Knowing when NOT to delegate is itself a skill.**

| When the task is..                   | Reach for      | Because                                    |
|--------------------------------------|----------------|--------------------------------------------|
| A quick question with a quick answer | Chat           | No credits, instant answer                 |
| Brainstorming or thinking out loud   | Chat           | Help while you work, not work done for you |
| Trivial - not worth the credits      | Chat or you    | Cowork would cost more than it saves       |
| Still too vague to brief             | Don't delegate | Sharpen the brief first                    |
| High stakes - legal, financial       | Don't delegate | Own the final call yourself                |
| Data you shouldn't expose            | Don't delegate | Don't point Cowork at it                   |

**"Is this worth the credits?" - ask this before every Cowork task.**

# Copilot Chat & Copilot Cowork: When to use each

|                 | Copilot Chat (Basic)<br>*free*                                                                                                                                  | M365 Copilot (Premium)<br>*paid*                                                                                                                  | Copilot Cowork<br>*requires M365 Copilot license*                                                                                                                                                       |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| What it is      | Your easy way to get started with AI. It combines web information with business data in your everyday apps to deliver relevant results.                         | Your secure AI for work — fast, intuitive, secure, and grounded in your data. Get answers, draft content, dig deeper and spark ideas.             | Your AI that gets the work done — a multi-step agentic solution that takes actions across your Microsoft 365 apps.                                                                                      |
| Best for        | Quick help, ideas, and research.                                                                                                                                | Fast, focused tasks in one conversation: draft, summarize, brainstorm, Q&A, and discrete research.                                                | Complex, multi-step work spanning multiple apps — end-to-end deliverables.                                                                                                                              |
| Speed           | Variable — may be limited during high demand with standard access.                                                                                              | Fast — response in seconds to minutes.                                                                                                            | Thorough — works autonomously through the full task for minutes to hours.                                                                                                                               |
| Task complexity | Single-step or short interactions - supports basic, chat-driven tasks.                                                                                          | Single-step or single-session grounded in your work data.                                                                                         | Multi-step workflows across tasks, applications, and sources.                                                                                                                                           |
| Use when        | You need quick assistance — draft content, get an answer, brainstorm ideas, or summarize information you provide.                                               | You need AI to work with your context - drafting an email from a thread, summarizing meetings, or creating content using your business data.      | You need Copilot to go do the work — multi-deliverable, end-to-end requests that need actions across apps, files and systems.                                                                           |
| Scenarios       | <ul style="list-style-type: none"><li>Quick catch-up on a topic</li><li>Draft or refine content from a prompt</li><li>Summarize a document you upload</li></ul> | <ul style="list-style-type: none"><li>Draft and iterate on content</li><li>On-demand daily briefing</li><li>Summary of a project status</li></ul> | <ul style="list-style-type: none"><li>Clean up your entire inbox and calendar</li><li>Multi-step project launches</li><li>Cross-app workflows like preparing for meetings or business reviews</li></ul> |

# Agentic Services (\$\$\$)

## From Assist to Execution



- Intune Governs **Devices**
- Entra Governs **Identities**
- Purview Governs **Data**
- Agent 365 Governs **Agents**

From assist

To execution

### AI Advisory

- Reimagine processes as AI-led, multi-step workflows / scenarios
- Provide business value assessment (cost estimation, ROI) to prioritize key scenarios

### Deploy, Secure, Govern

- Establish readiness foundation (security & compliance controls, data structure) for trusted baseline
- Deploy Cowork with Work IQ grounding
- Provide usage governance & cost optimization services

### Build and Manage Agentic Solutions

- Build and deploy enterprise plugins by function
- Integrate Cowork with LoB systems
- Run ongoing AgentOps

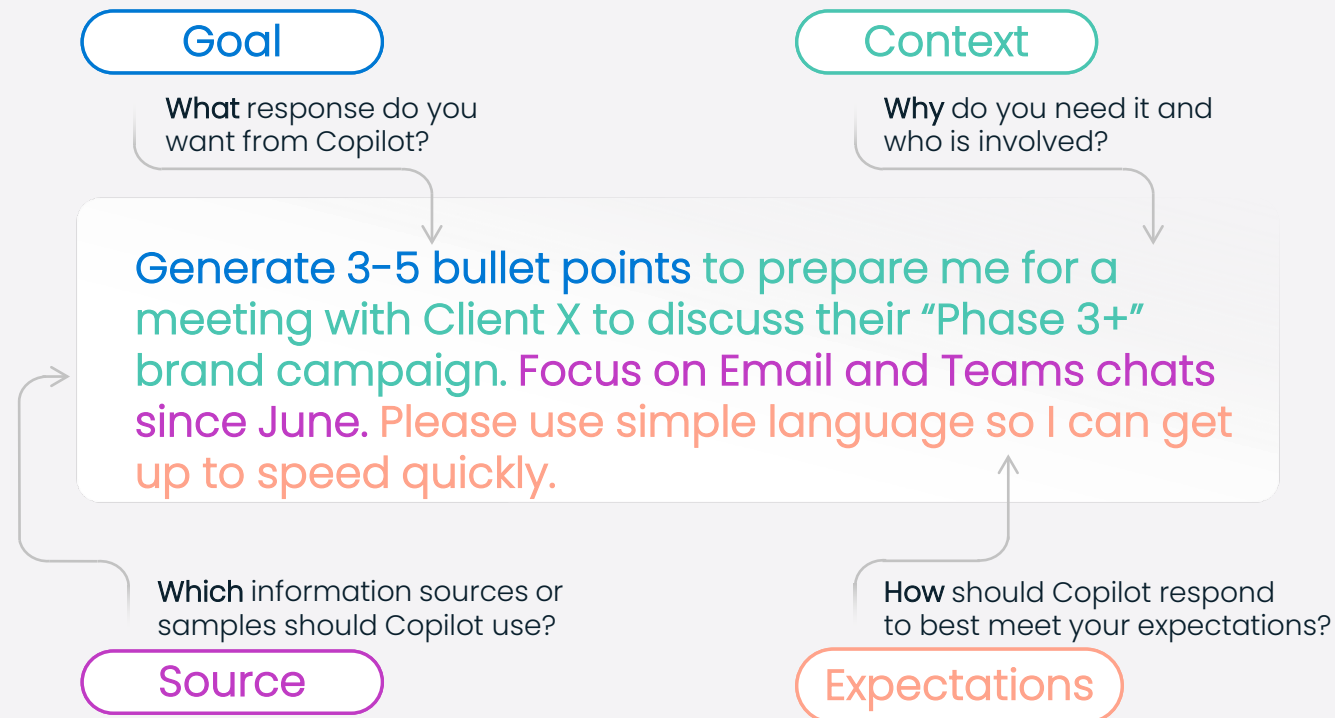
### Drive Adoption & COE

- Build persona skill packs & prompt libraries
- Drive adoption of priority scenarios
- Stand up an Agent COE
- Measure adoption, ROI & cost

# Copilot Prompting – GCSE (Prompt Coach)

## Include the right prompt ingredients

To get the best response, follow the key elements below when phrasing your Copilot prompts.



**Tip:** "Get Clever, Start Exploring"  
"Start Iterating"

# The Cowork GRAMMAR of Prompt Writing

Every strong Cowork prompt covers 7 elements — work through GRAMMAR to turn a vague request into a clear, complete brief.

| Component              | Purpose                                            | Example                                                                                     |
|------------------------|----------------------------------------------------|---------------------------------------------------------------------------------------------|
| <b>G – Goal</b>        | What business outcome do you want?                 | <i>"Produce a Board-ready project status report."</i>                                       |
| <b>R – Role</b>        | What role should Cowork assume?                    | <i>"Act as a Principal Consultant and PMO Lead."</i>                                        |
| <b>A – Audience</b>    | Who is the output for?                             | <i>"Executive Steering Committee."</i>                                                      |
| <b>M – Material</b>    | What sources should be used?                       | <i>"Use project files, meeting notes, Teams chats and emails from the last 30 days."</i>    |
| <b>M – Method</b>      | What process should Cowork follow?                 | <i>"Review updates, identify risks, extract actions, group by workstream."</i>              |
| <b>A – Assumptions</b> | What assumptions, constraints or exclusions apply? | <i>"Exclude completed actions. Treat missing dates as unconfirmed."</i>                     |
| <b>R – Result</b>      | What should the final deliverable look like?       | <i>"Create a Word document with executive summary, risks, actions and recommendations."</i> |



# Connector v Skill v Plugin

Understanding when to use which

## Connector + Skill = Plugin

### Connector

= reach

The bridge that lets Cowork pull data from another system e.g. a Fabric database

+

### Skill

= behaviour

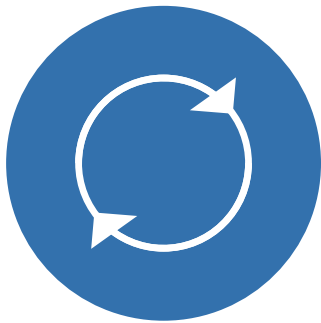
A repeatable way of doing something you teach Cowork i.e. a process, a method.

=

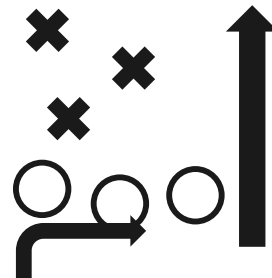
### Plugin

= a bundle

A package that bundles connectors and/or skills to do a specific job.



**Reusable** - Define once, apply consistently.



**Shapes behaviour** - process, steps, format



**Loads on demand** - gives the Cowork capability

Teach Cowork how you work. Cowork knows when to use the skill

# Getting the most out of Copilot Cowork



## Write a sharp brief

Use the GRAMMAR structure so Cowork gets it right first time.

- Name the role, action, and object (up to 250,000 characters)
- Add context, sources, and constraints (use file links not attachments)
- Say exactly what the output should be
- Use rich text (bold, italics etc)



## Right-size the task

Match the ask to the task tier so credits go where they matter.

- Light: quick summaries, one deliverable
- Medium: multi-source, structured output
- Heavy: deep analysis, many outputs
- Ask for clarity before it burns through credits – maybe use Researcher!



## Point it at the right sources

Tell Cowork where to look instead of “everything.”

- Name the emails, files, and channels (files can be up to 200Mb)
- Set the time frame
- Group by project or account
- Use skills (you can create up to 50 custom skills)



## Automate & mind your credits

Make repeatable work hands-free and keep spend in check.

- Schedule research → doc → Teams → email
- Use the Cost Estimator before big jobs
- \*\*Add a training gate before heavy use\*\*
- Turn large prompts into skills

**WARNING:** if an agent can browse untrusted pages and also talk to privileged local services, loopback can become an attack surface and control planes must be authenticated, authorised, and isolated (AutoJack RCE)



# Demo

## Tour

- Copilot App (setup for success – S&I)
- Copilot Control Plane (cost mgmnt)
- Agent 365 Admin Centre (observe)

## Prompts

- New MSP Business Plan (long)
- Compare Copilot v Cowork
- Chaining Skills (DD zip file)
- FIFA World Cup