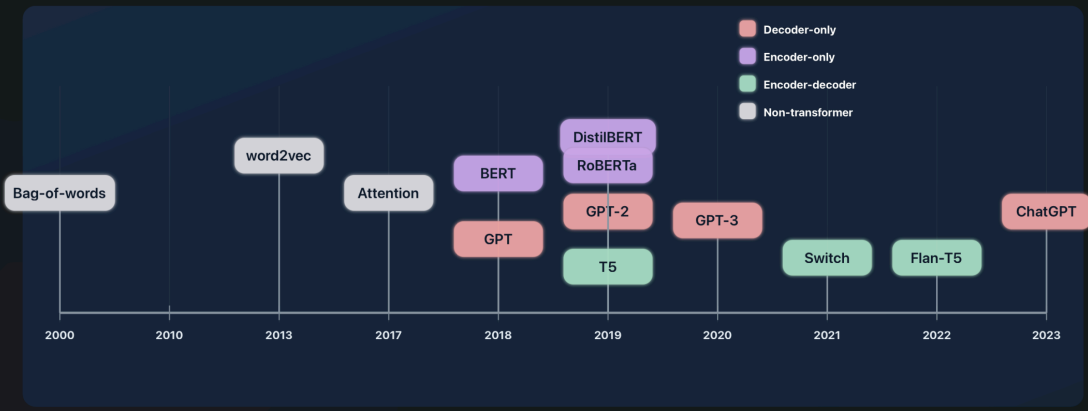
History of Language AI

AI: the science and engineering of making intelligent machines—especially computer programs.

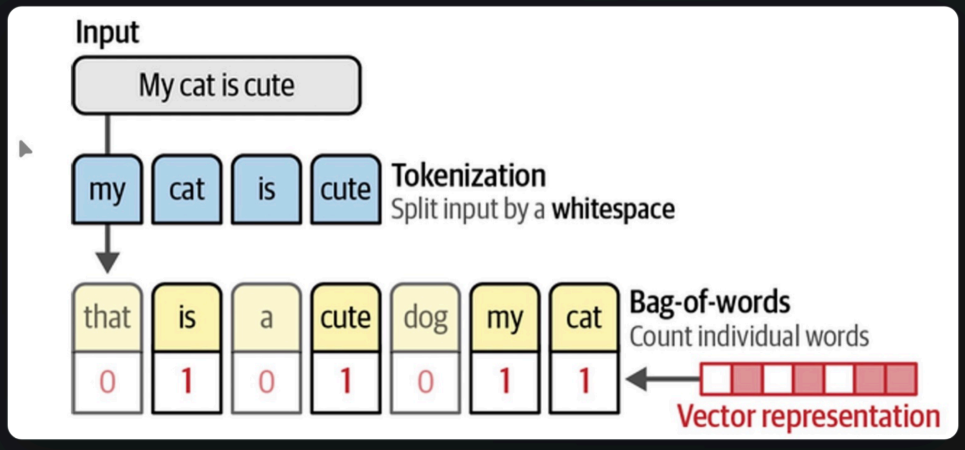
Language AI: a branch of AI that can understand, process, and generate human language.

Language AI pre-dates ChatGPT — here's a timeline:



How can a computer understand Human language?

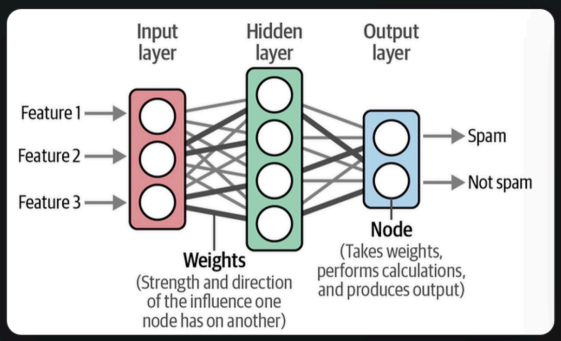
Bag-of-words — 1st attempt to process language

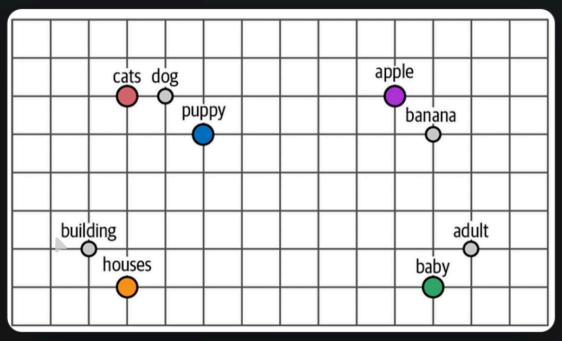


Limitation: It considers language to be nothing more than an almost literal bag of words and ignores the semantic nature, or meaning, of text.

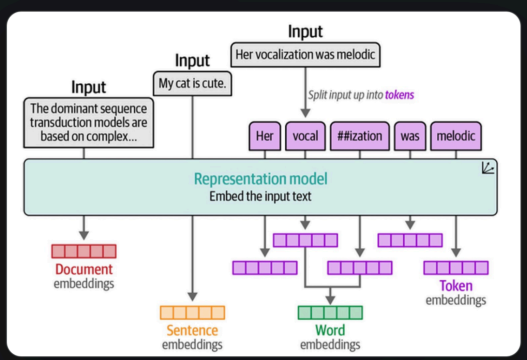
Word2Vec a first attempt to learn the meaning of text

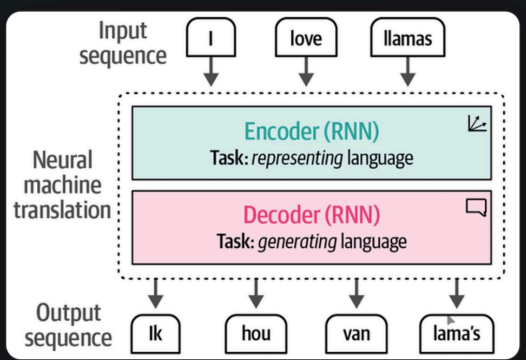
Definition: word2vec learns semantic representations of words by training on vast amounts of text (e.g., all of Wikipedia). It uses *neural networks* — interconnected layers of nodes — to process context and produce meaningful vectors for words.





Embeddings are the key to communicating with Language models



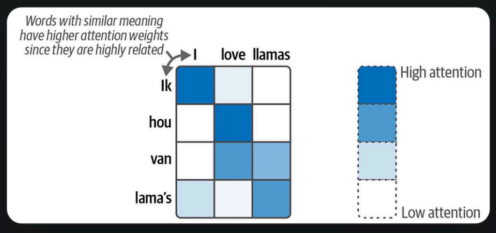


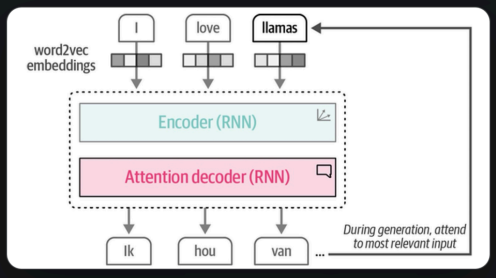
Limitation: *word2vec* creates static, downloadable representations of words. The word "bank" will always have the same embedding, regardless of context.

Solution: use *recurrent neural networks (RNNs)*, which model sequences. They're used for two tasks: **encoding** (representing an input sentence) and **decoding** (generating an output sentence).

Attention a breakthrough in Language models

Attention selectively determines which words are most important in a given sentence.

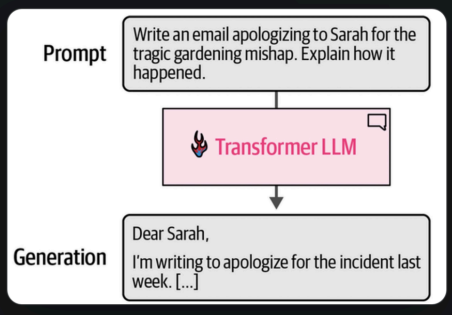


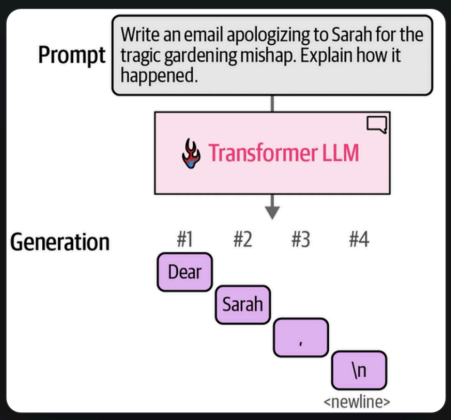


In practice: adding attention to the decoder lets the model weigh each input word when generating the next output, rather than relying on a single context vector.

Why it matters: the 2017 paper *Attention Is All You Need* introduced the **Transformer** — a network built entirely on attention (no recurrence). Transformers train in parallel, which unlocked today's large language models.

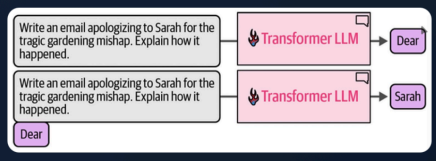
Transformers: The Basis of ChatGPT



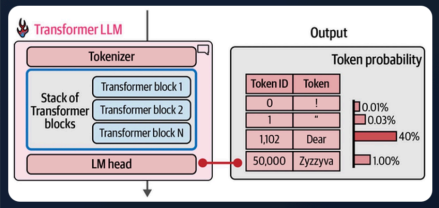


How does the model generate text? — After each token generation, we tweak the input prompt for the next generation step by appending the output token to the end of the input prompt.

LLMs are auto regressive — use current knowledge to generate next token



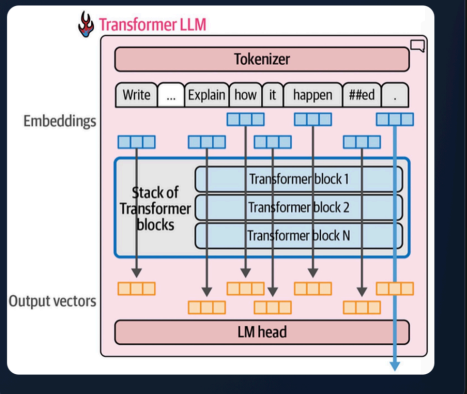
An output token is appended to the prompt, then this new text is presented to the model again for another forward pass to generate the next token.



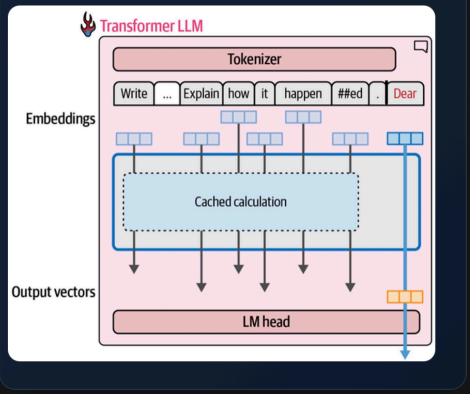
The model predicts for each word in its vocabulary a probability score, then chooses from this distribution

Choosing the next token: choosing the token with highest probability is called *greedy decoding* — but models can also sample from the distribution, controlled by the **temperature** parameter.

What makes Transformers special



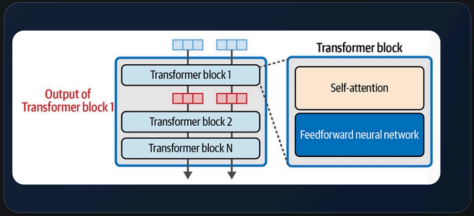
One of the most compelling features of Transformers is that they lend themselves better to parallel computing than previous neural network architectures in language processing.



If we give the model the ability to cache the results of the previous calculation (especially some of the specific vectors in the attention mechanism), we no longer need to repeat the calculations of the previous streams. This time the only needed calculation is for the last stream.

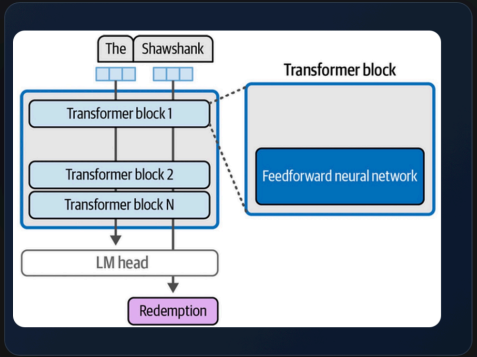
Caching is one of many optimizations to the Transformer architecture that make today's LLMs fast enough to give you the best user experience. We'll dive into optimizations more later.

A look inside the Transformer — a brief glimpse



Transformer LLMs are stacks of identical *Transformer blocks*. Each block processes its input and passes the result to the next — most of the model's work happens inside these blocks.

A Transformer block contains two parts: **self-attention** and a **feed-forward neural network**.

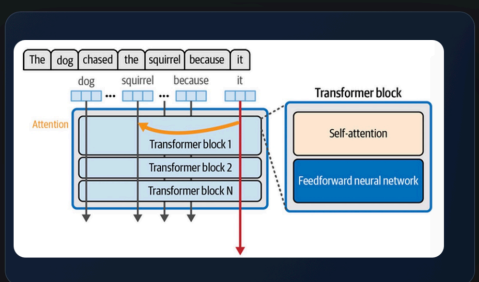


When trained on massive text archives containing phrases like "The Shawshank Redemption," the feed-forward network learns linguistic and contextual relationships that improve its next-token predictions.

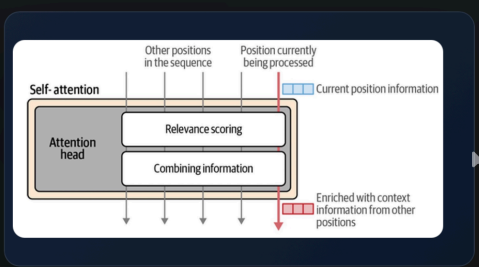
Memorization is necessary but not enough. What makes LLMs powerful is their ability to generalize — using the same machinery to interpolate between examples and handle inputs they've never seen before.

The Attention layer at a glance

Think of the following prompt: "The dog chased the squirrel because it". For the model to predict what comes after "it," it needs to know what "it" refers to. Does it refer to the dog or the squirrel? In a trained Transformer LLM, the attention mechanism makes that determination.



The self-attention layer incorporates relevant information from previous positions that help process the current token.



Attention heads run in parallel with a preceding step of splitting information and a later step of combining the results of all the heads.

The exact attention calculation is outside the scope of this workshop. At a high level, it uses matrix multiplications to capture context information in embedding vectors.

Changing Tracks: Prompt Engineering - Conversing with LLMs

Prompt engineering is the process of structuring, crafting, and refining the input (known as a "prompt") given to a generative AI model—particularly Large Language Models (LLMs)—to produce more desirable, accurate, and relevant outputs.

Element	Description
Persona	Tell the model who to be or which perspective to adopt.
Instruction	State the task with clear verbs and any must-follow constraints.
Context	Give background facts, examples, and goals the model needs.
Format	Specify the output shape—bullets, table, JSON, sections, or a template.
Audience	Describe who it's for and their level so explanations land.
Tone	Set the voice—friendly, formal, concise, persuasive, playful.
Data	Provide sources, snippets, or numbers to ground the answer.

Not every prompt needs this full structure, and this template isn't exhaustive—use only the parts that help.

Good and bad prompts and the approach to working with LLMs

Prompt

"Write about climate."

Likely outcome

- Generic, surface-level response
- Misses intent and audience
- No structure or sourcing

Why it's bad: Vague instruction, no context, audience, or format—so output is unfocused and hard to use.

Prompt

Persona: HS science teacher
Task: 200-word explainer, with 2 sources

Context: compare mitigation vs adaptation for UK
Format: bullets; Audience: age 15–16; Tone: clear, neutral
Data: cite gov.uk or Met Office

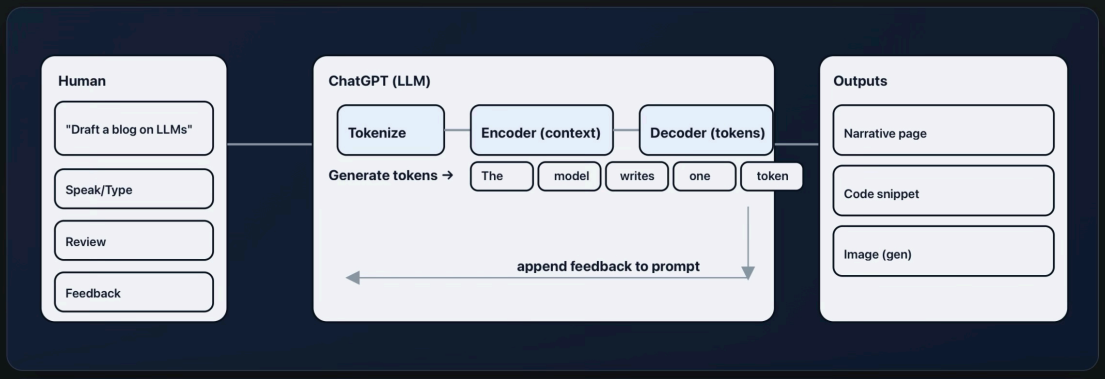
Likely outcome

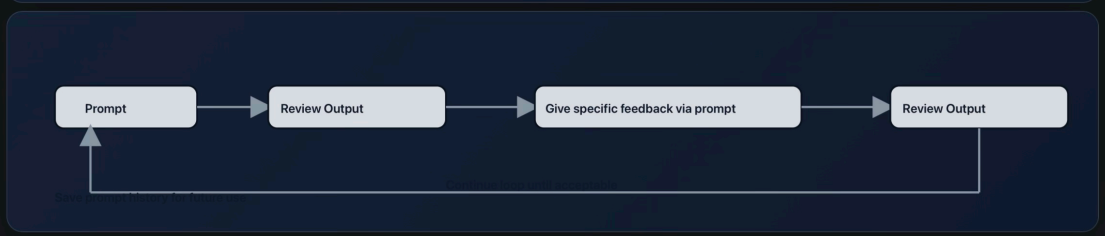
Structured, on-level, sourced, and immediately usable

Why it's better: Clear constraints and context guide the model to deliver focused, useful results.

Detailed prompts + verified tests + precise edits → strong final results. This is an iterative process by design.

Approach to using LLMs “Vibing”





Your role: You own the vision. Communicate clearly and in detail with the LLM, and check its work for accuracy and completeness.

Mindset: Treat the LLM like a bright, eager-to-please, relentless learner—guide it with specific feedback.

LLMs - The Bad, and the Ugly

WHAT ARE LLMS BAD AT?

Hallucinations (confident falsehoods): *Examples:* (1) Grok produced inappropriate content and extremist claims in ordinary chats; (2) Google’s Bard demo incorrectly said the James Webb Space Telescope took the first exoplanet photo. *Mitigations:* Grounding via Retrieval-Augmented Generation (RAG), careful tool use/function calling, and alignment (e.g., RLHF) to reduce guesswork.

Brittle reasoning & planning: LLMs can sound confident yet stumble on simple multi-step tasks—e.g., a short logic puzzle answered correctly one way can flip wrong if you slightly reword it—because the model predicts likely text rather than reliably reasoning step-by-step.

Security weaknesses (prompt injection / jailbreaks): Real-world incidents show attackers can inject hidden instructions or coax policy bypasses—for example, a dealership chatbot was manipulated into offering cars for $1 via prompt injection—highlighting the need for input filtering, sandboxed tools, and rigorous guardrails.

Quote: “With great power comes greater responsibility” — and that lies with us humans.

Impact of GenerativeAI on Software Development

Commercial Breakthrough: AI coding assistants (like GitHub Copilot) have become the first true killer app of Generative AI, attracting nearly $1B in funding and transforming GitHub's growth — now driving 40%+ of its revenue.

Massive Productivity Gains: Enterprises and tech giants (Google, Microsoft, Accenture) report 20–45% productivity boosts, with AI now writing up to 30% of company code in some cases.

Integration into Developer Workflows: Developers use AI tools daily — GitHub Copilot for coding, ChatGPT for problem-solving — creating a new "vibe coding" culture that blends creativity and flow with AI assistance.

Explosive Competition: The AI coding market is fiercely contested among Microsoft, Google, Amazon, Meta, and fast-rising start-ups like Cursor and Lovable, each building on models from OpenAI, Anthropic, and others.

Democratization of Software Creation: AI tools lower entry barriers — enabling non-coders to build apps and pushing skilled engineers toward architecture and design rather than syntax and boilerplate work.

Limitations and Risks: AI remains a powerful helper, not a replacement — lacking true understanding of architecture or system design, and potentially generating glitchy, unverified code without human oversight.

Future Outlook: Experts predict AI will surpass humans at competitive coding by 2025, ushering in a transformative shift where AI becomes essential in development — yet human judgment and governance remain critical.

Impact of Generative AI on Education

Student Adoption and Behavior: AI usage among students has surged — 92% of UK undergraduates now use generative AI, with 88% applying it for assessments. Top motivations: saving time and improving quality. Over a quarter of US college prompts are for coursework.

Learning and Critical Thinking: While AI boosts test performance, students often perform worse without it — suggesting "cognitive offloading." Studies link AI dependence with lower critical-thinking scores and higher reliance among younger learners.

Assessment and Academic Integrity: Universities are rethinking assessment. AI detection rates remain low (≈11% of papers show signs of AI use, only 3% with heavy use). "Using ChatGPT ≠ cheating" — but academic integrity frameworks lag behind adoption speed.

Evolving Teaching Methods: Some educators return to handwritten and oral exams; others embrace AI. Harvard Business School issues ChatGPT EDU accounts, while smaller colleges integrate AI into redesigned courses and assignments.

Benefits for Educators: AI reduces administrative load, freeing hundreds of thousands of teaching hours by 2026. Teachers use AI to personalize lessons and clarify complex concepts, improving engagement and feedback loops.

Workforce Readiness & Anxiety: AI is reshaping job markets — youth unemployment up 5–6% since ChatGPT's launch, with entry-level roles shrinking. Students question the ROI of degrees as AI automates white-collar work.

Institutional Responses: Reactions vary: some ban AI, others embed it deeply. Estonia provides nationwide AI access to 20,000 high schoolers, while experts caution that unplanned deployment may harm learning more than help.

The Path Forward: The goal: teach foundational skills first, then use AI for enhancement. Confidence and literacy predict smarter AI use. Success requires policies on responsible integration and focus on human critical thinking.

Impact of GenerativeAI - Workforce transformation

Company Expectations and Mandates: AI proficiency is rapidly becoming a baseline requirement. Shopify calls AI use a "fundamental expectation," while leaders at Amazon, Salesforce, and JPMorgan publicly link productivity gains to workforce AI adoption. The C-suite mandate: "Leverage AI to reduce workforce — ensure every employee is AI-equipped."

Current State of Enterprise AI Policies: Despite leadership enthusiasm, many firms lack clear guidance. 63% of employees use generative AI for work, but 23% say their company has no AI policy. This gap fuels widespread "shadow AI" usage with staff relying on public tools outside official governance.

How Employees Should Use AI Tools: Only 22% of workers say their company has a defined AI integration plan. The most effective models combine senior leadership engagement, cross-disciplinary collaboration, and employee-led innovation — empowering teams to shape safe, practical use cases.

Enterprise AI Challenges: MIT found 95% of workplace AI pilots fail due to lack of customization, poor integration, and limited learning capacity. Organizations that succeed typically harness bottom-up experimentation ("shadow AI economy") while building structured governance.

Workforce and Talent Implications: Experts warn against short-term layoffs, advising instead to "hire into reconfigured workflows." McKinsey and others stress building hybrid roles blending human oversight with AI-assisted execution to close emerging talent gaps.

Keys to Successful Adoption: Consensus: sustainable AI transformation depends on clear policies, employee involvement, and recognition that AI is a collaborative augmentation tool, not a replacement for human judgment.

4:28 PM Thu Oct 23
31%
Exit Full Screen

AI Tools Beyond ChatGPT — The Expanding Ecosystem

The New Frontier of AI Adoption: The AI ecosystem has exploded far beyond conversational chatbots. Millions now rely on specialized tools for writing, design, coding, meetings, and media creation — AI is becoming the invisible assistant in everyday work.

Writing and Language AI: Tools such as Grammarly, QuillBot, and DeepL are quietly the most-used AI platforms worldwide. They enhance communication, translation, and clarity for hundreds of millions of users — bringing measurable productivity gains across education and enterprise.

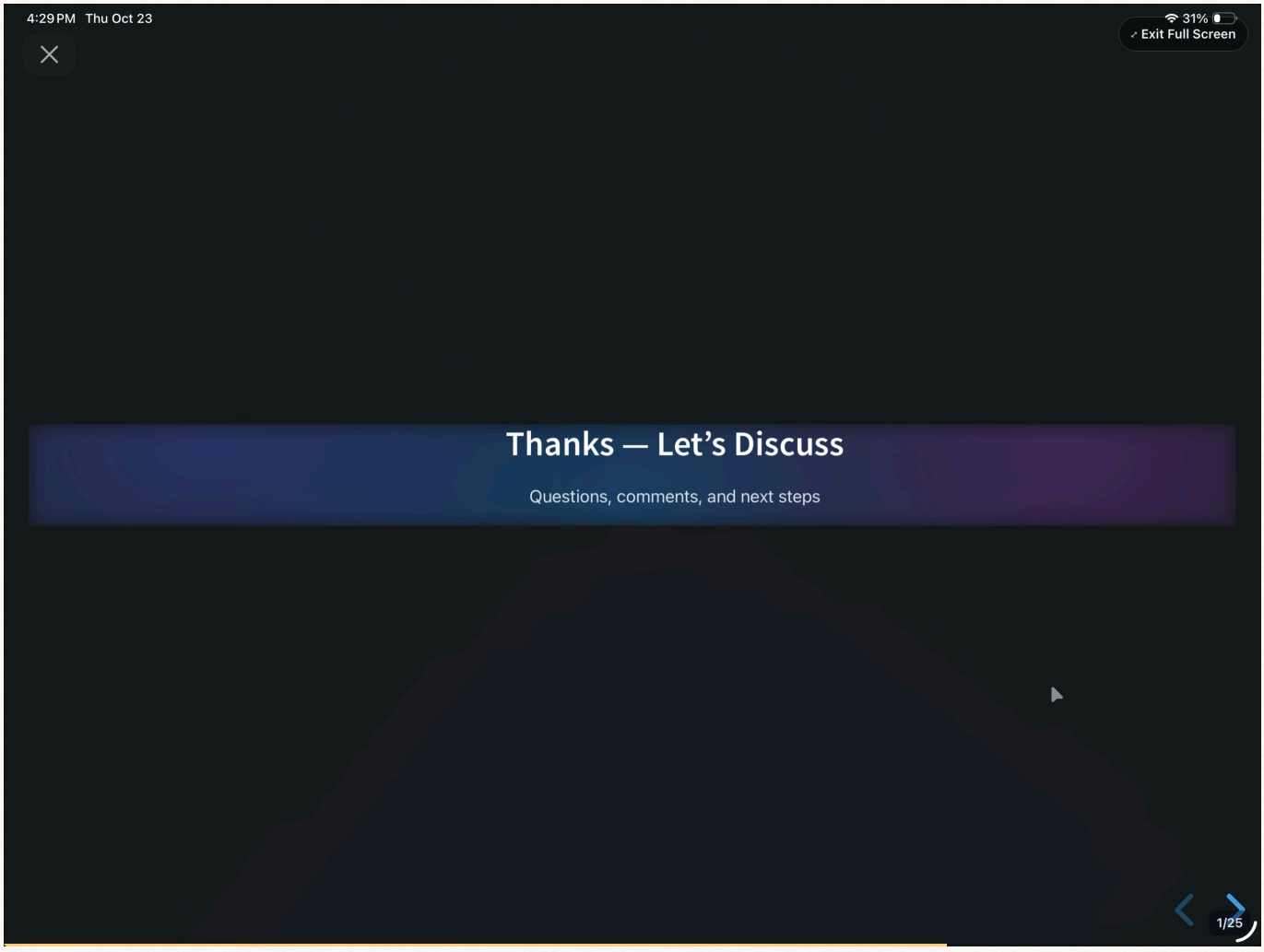
Creative and Design AI: Visual creation has gone mainstream with Midjourney, Adobe Firefly, and Canva's AI suite. Tens of millions of designers and marketers now generate art, layouts, and brand assets instantly — redefining how visual work is done.

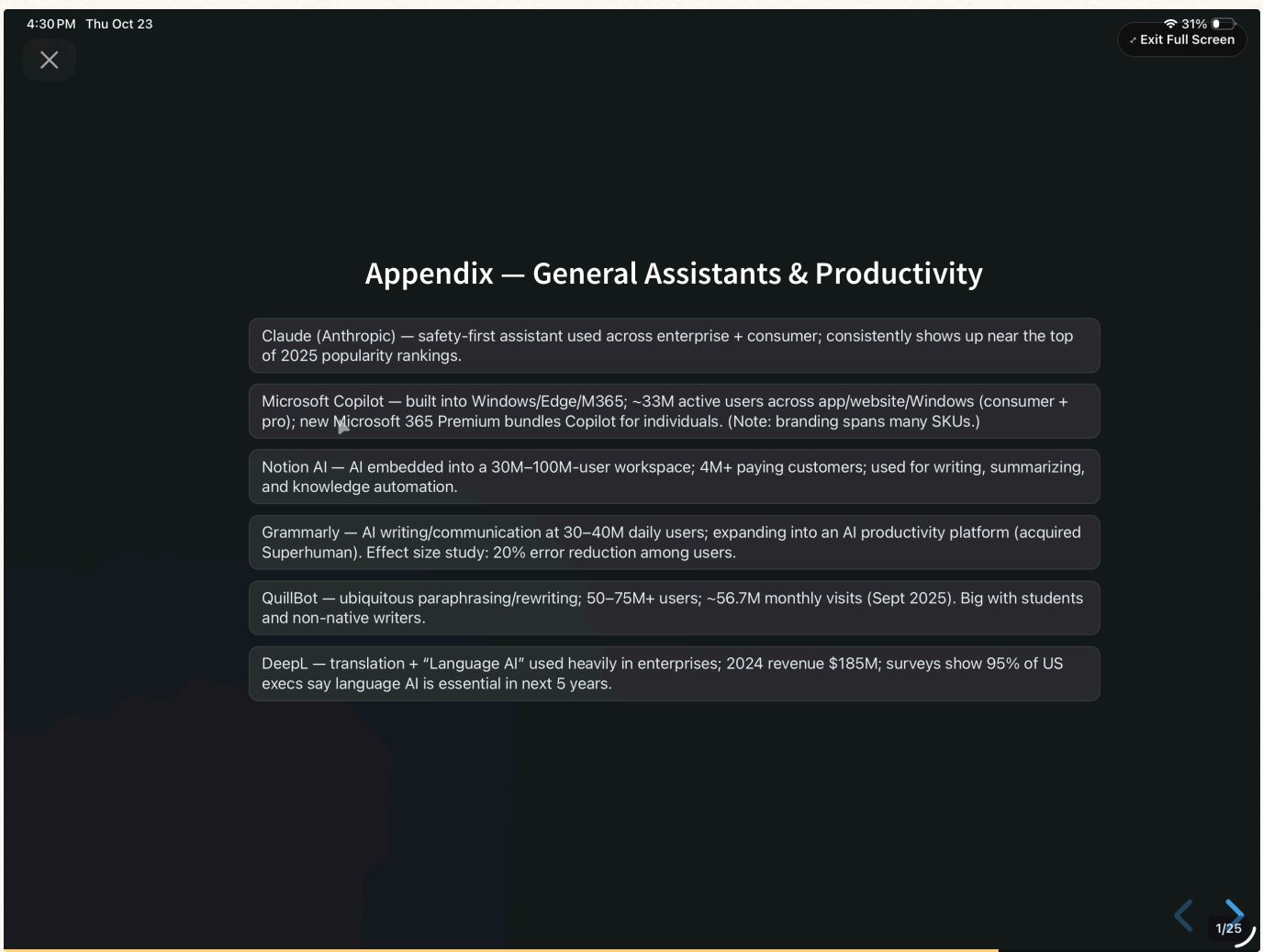
Video and Audio Generation: AI is transforming storytelling: Runway and Synthesia lead text-to-video, while ElevenLabs and Suno enable synthetic voices and music. These tools democratize professional media creation once reserved for studios.

Productivity and Work Automation: Meeting assistants like Otter.ai, Fireflies, and Fathom are standard in Fortune 500 workflows. They record, summarize, and analyze meetings — turning conversation into data and freeing knowledge workers from note-taking.

Developer Ecosystem Expansion: For engineers, Cursor and Codeium rival GitHub Copilot with open access and fast growth. These tools show how AI is shifting from passive help to active collaboration inside the IDE.

The Pattern: AI Enhances Human Judgment: Across categories, adoption thrives where AI augments human creativity, reasoning, and communication — not where it replaces them. The next frontier is tool-agnostic AI literacy — knowing how to choose and critically use AI across domains.


4:29 PM Thu Oct 23
31%
Exit Full Screen
Thanks — Let's Discuss
Questions, comments, and next steps
1/25


4:30 PM Thu Oct 23
31%
Exit Full Screen
Appendix — General Assistants & Productivity
Claude (Anthropic) — safety-first assistant used across enterprise + consumer; consistently shows up near the top of 2025 popularity rankings.
Microsoft Copilot — built into Windows/Edge/M365; ~33M active users across app/website/Windows (consumer + pro); new Microsoft 365 Premium bundles Copilot for individuals. (Note: branding spans many SKUs.)
Notion AI — AI embedded into a 30M–100M-user workspace; 4M+ paying customers; used for writing, summarizing, and knowledge automation.
Grammarly — AI writing/communication at 30–40M daily users; expanding into an AI productivity platform (acquired Superhuman). Effect size study: 20% error reduction among users.
QuillBot — ubiquitous paraphrasing/rewriting; 50–75M+ users; ~56.7M monthly visits (Sept 2025). Big with students and non-native writers.
DeepL — translation + "Language AI" used heavily in enterprises; 2024 revenue $185M; surveys show 95% of US execs say language AI is essential in next 5 years.
1/25

Appendix — Image Creation & Design

Midjourney — ~20–21M users in 2025; a top consumer image tool (also facing IP litigation risk—useful for risk slides).

Adobe Firefly — integrated into Creative Cloud; 2025: 30% sequential traffic growth, paid subs nearly doubled; claims of massive MAU in some coverage; practical enterprise adoption signal.

Canva (AI suite) — one of the most visited AI-enabled design platforms globally; consistently ranks at/near the top non-LLM AI tools lists.

Appendix — Video Generation

Runway (Gen-3 Alpha) — industry reference for text-to-video; widely used by creators, studios and marketing teams.

Synthesia — popular studio-style avatar video generation; a perennial leader in "best AI tools" roundups for corporate comms/training.

Pika / CapCut AI (optional mentions) — fast-growing consumer video gen/edit ecosystems; CapCut is widely used with AI features (source landscape shows strong adoption via creator economy).

Appendix — Audio, Music & Voice

ElevenLabs — leading AI voice cloning/tts; strong growth, notable government/enterprise references.

Suno — text-to-music for creators; ~1.1M–2.5M MAU range depending on source/time; reportedly raising at $2B+; AI music licensing deals being negotiated industry-wide.

Udio — text-to-music; public app and funding momentum; in the labels' emerging licensing talks.

Appendix — Meeting & Work Capture (AI Notetakers / Agents)

Otter.ai — crossed $100M ARR (Mar 2025); strong "AI meeting agent" suite and broad adoption.

Fireflies.ai — used at 75% of Fortune 500; unicorn valuation; integrates real-time web search into meetings.

Fathom — popular free tier; third-party reviews place it among top notetakers; claims time-savings and high user satisfaction.

Appendix — Coding & Developer Tools

Cursor — fast-growing AI IDE; reports of 1M users and $100M 2024 revenue, projecting $200M 2025; strong dev buzz.

Codeium — enterprise-friendly Copilot alternative; 2025 reviews show active uptake; third-party data pegs $80M revenue (directional).