

artificiality

MINDS MEETING MACHINES

Artificiality Pro: January 2024

10 Obsessions for 2024

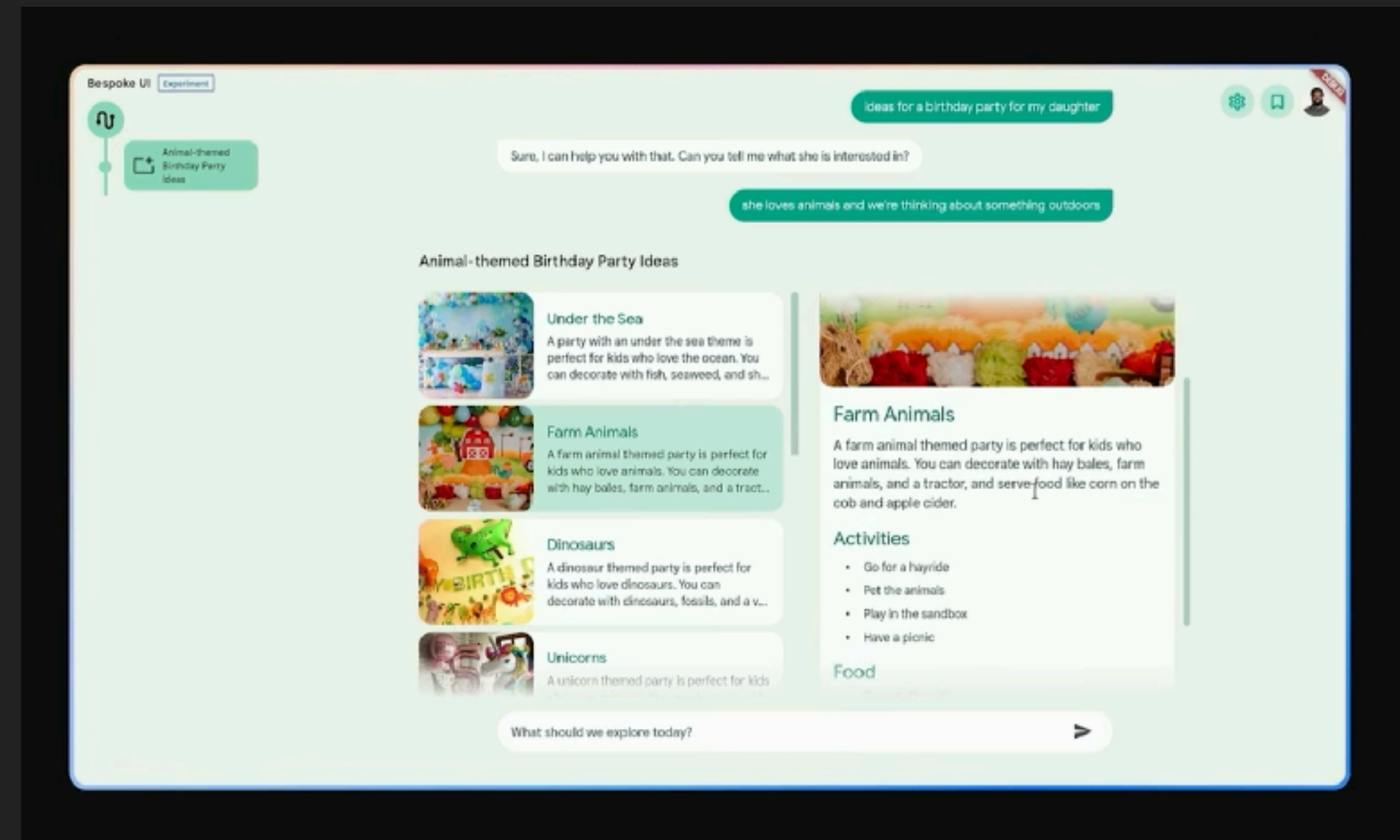


1. Agentic AI
2. AI Inside
3. Edge AI
4. AI-Enhanced Learning
5. Human Centered Gen AI
6. Interpretability
7. Memory vs Margins
8. Source Dilemma
9. Trust
10. World of Workflows

1. Agentic AI



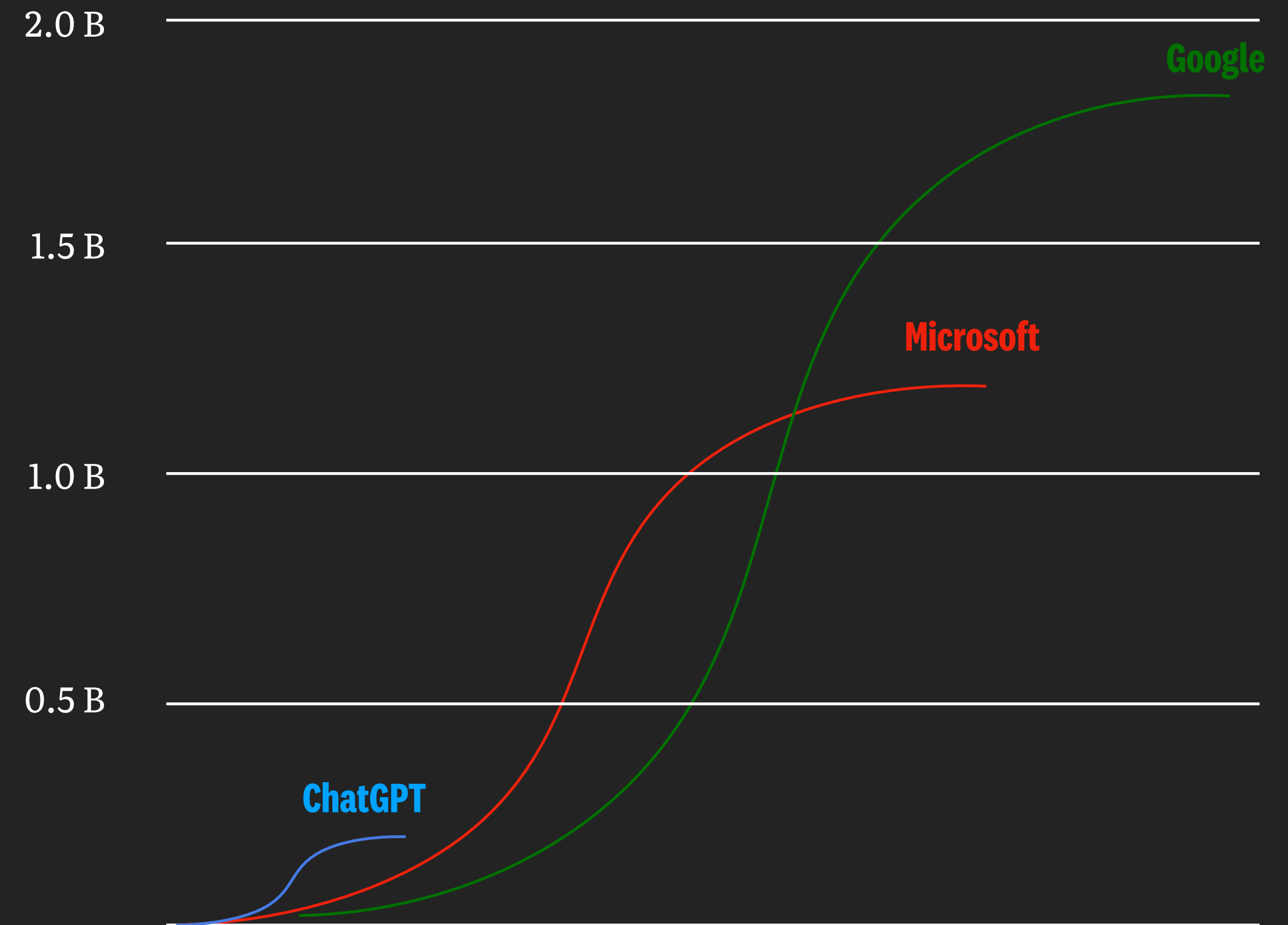
- **Premise:** Personal generative AI that is able to act as a personal agent will redefine how we use the web and/or apps.
- **Status:** GPT 4 shows early promise as does Gemini. Early glimpses of personal dynamic autonomous agents are here.
- **Watching:** The science and the early adoption of such agents.



2. AI Inside



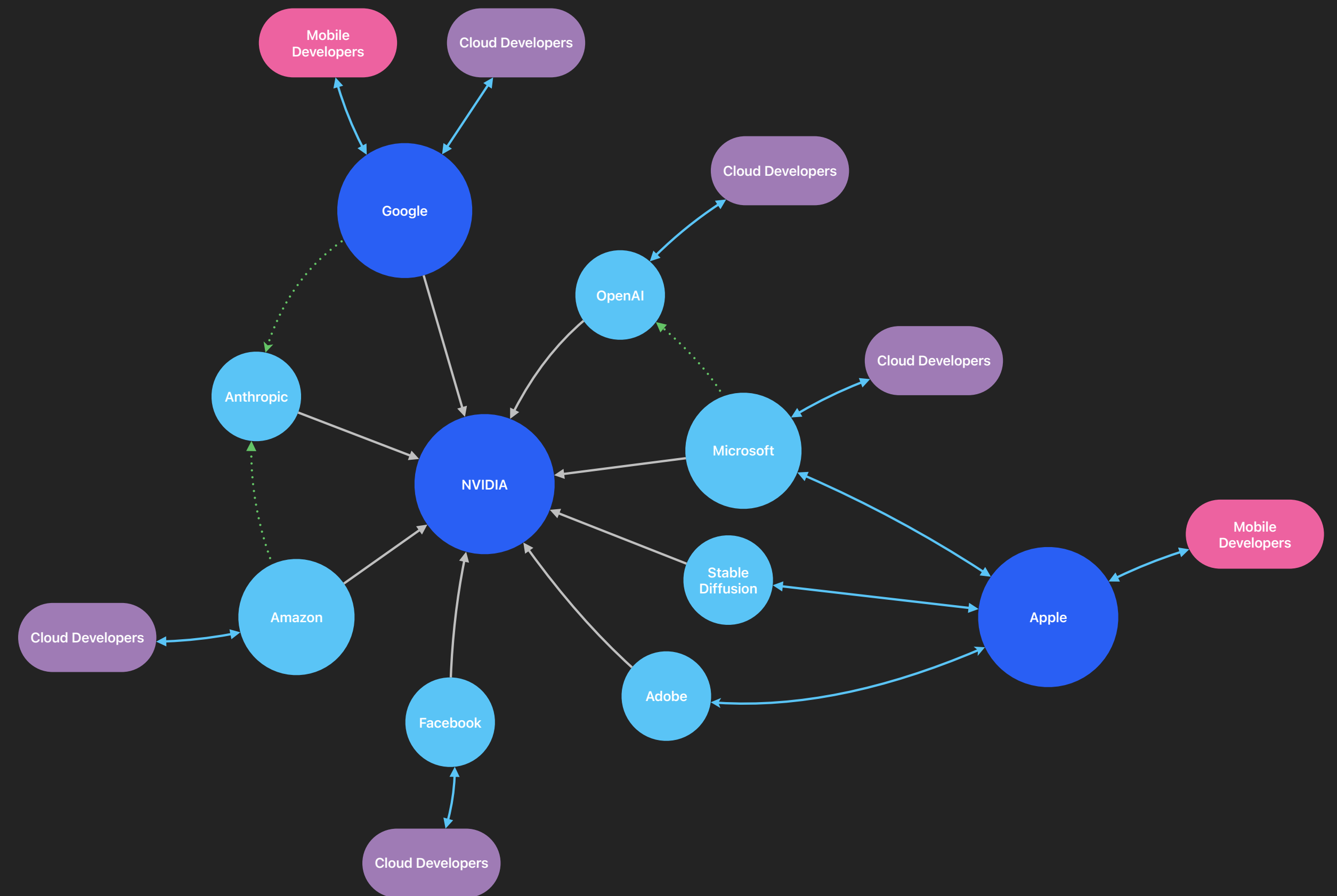
- **Premise:** Shift from novel AI apps (ChatGPT) to apps with AI inside (CoPilot) to provide AI benefits within existing workflows.
- **Status:** Existing Gen AI workflows require using novel apps like ChatGPT. Integration into existing apps and workflows is emerging.
- **Watching:** Expansion of AI inside from smaller apps like Notion to major apps like Microsoft Office and Google Workspace.



3. Edge AI



- **Premise:** AI in the cloud benefits from scale but is challenged by cost and privacy. Mobile solves these challenges but is lagging in capability.
- **Status:** Apple, Google, and Microsoft have made important announcements of models, frameworks, and/or chips.
- **Watching:** Developer-focused announcements, especially at Apple WWDC and/or Google I/O.



4. AI-Enhanced Learning



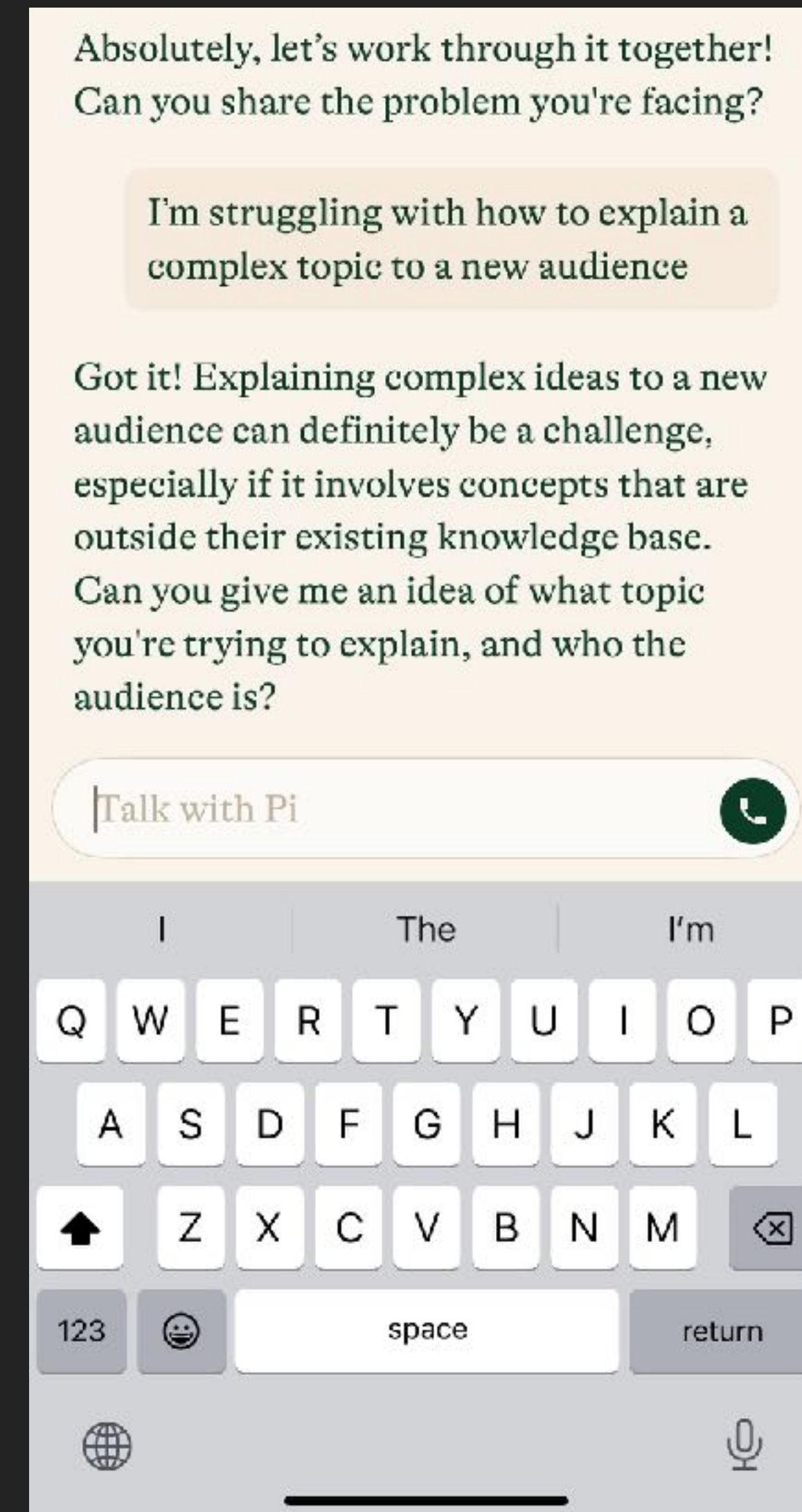
- **Premise:** The integration of generative AI in learning and skill development is revolutionizing the educational landscape.
- **Status:** Current advancements in AI, like the latest versions of language models, are showing significant potential in creating personalized learning experiences.
- **Watching:** The effectiveness of AI as a learning partner, the impact on skill development and proficiency, and the impact of gen AI on inclusivity and equity.

	<i>Student Learning</i>	<i>Faculty Teaching</i>	<i>Admin Processes</i>
<i>Faculty/Staff</i>	41%	50%	42%
<i>Leadership</i>	72%	65%	74%

5. Human Centered Gen AI



- **Premise:** Humans will only embrace AI if it enhances their desire for purpose and agency.
- **Status:** Current task-oriented UX is helpful but not yet empowering.
- **Watching:** Human centered design approaches that extend beyond tasks and engages in the pursuit of higher goals and objectives.



6. Interpretability



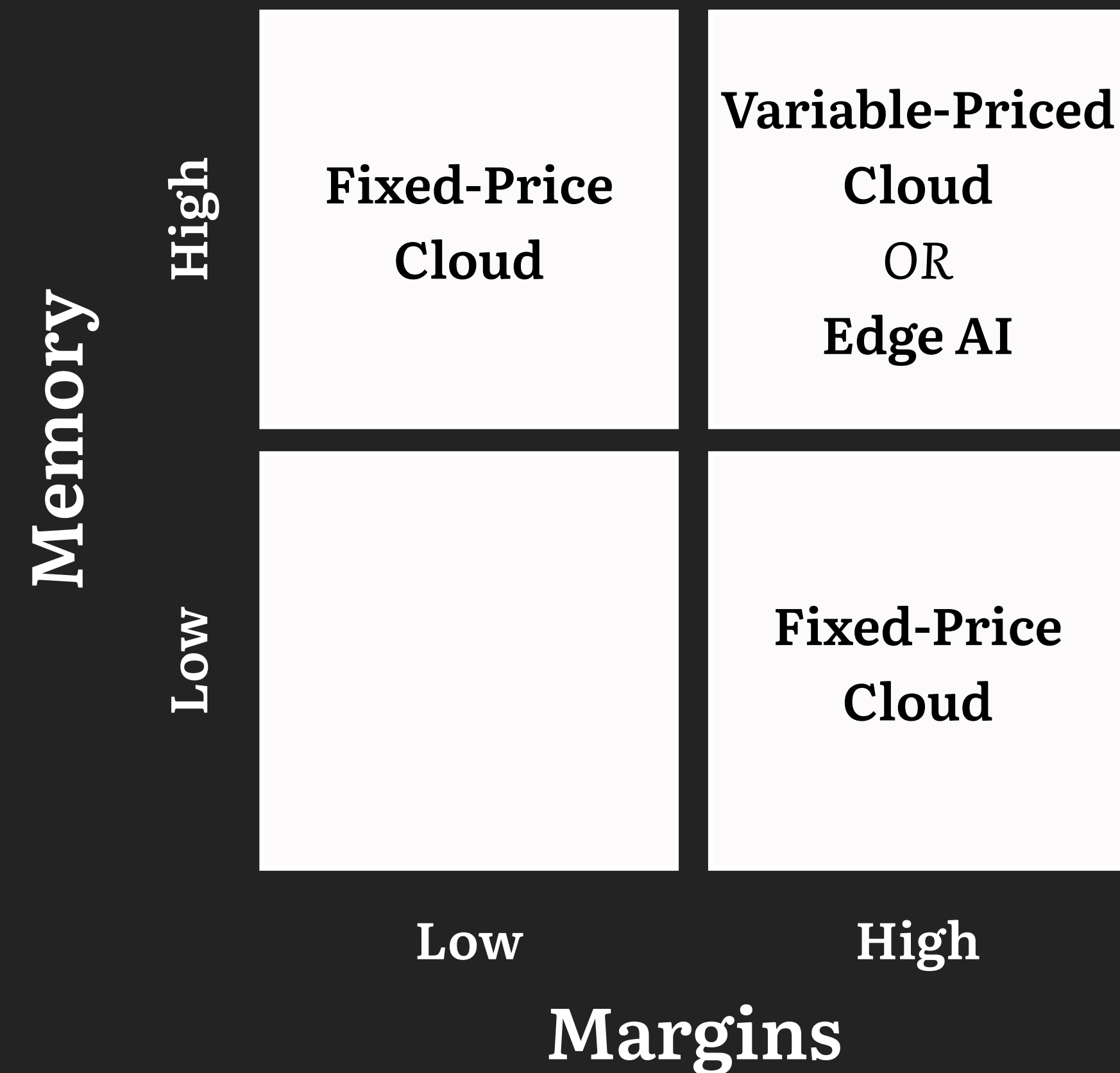
- **Premise:** Interpretability enables critical value-proposition generative AI.
- **Status:** Early progress shows potential for scale and discovery of new ways to understand the inner workings of AI.
- **Watching:** How the science progresses to engineering and investment and how it can unlock high value, critical use cases.





7. Memory vs Margins


- **Premise:** Increasing memory increases context and usefulness but memory is expensive.
- **Status:** Today's fixed fee, cloud-based system is memory-limited.
- **Watching:** Increasing interest in mobile, subscription fee increases, and new technical approaches.



8. Source Dilemma




- **Premise:** Closed-source vs open-source will be an important, path dependent choice that affects auditability, cost, quality, security, and vendor lock-in.
- **Status:** Experiencing a “Linux moment” as the market, once dominated by closed models, is expanding to include robust open source models & communities.
- **Watching:** Evolving standards for model transparency and accountability, the economic models supporting AI development, the impact of open-source AI on innovation and accessibility

 OpenAI / gpt-3.5-turbo

OpenAI's most capable and cost effective model in the GPT-3.5 family optimized for chat purposes, but also works well for traditional completions tasks.

Context	4,096 tokens
Input Pricing	\$1.50 / million tokens
Output Pricing	\$2.00 / million tokens

[Model Page](#) [Pricing](#) [Website](#)

 Meta / llama70b-v2-chat

70 billion parameter open source model by Meta fine-tuned for chat purposes served by Fireworks. LLaMA v2 was trained on more data (~2 trillion tokens) compared to LLaMA v1 and supports context windows up to 4k tokens.

Context	4,096 tokens
Input Pricing	\$0.70 / million tokens
Output Pricing	\$2.80 / million tokens

[Model Page](#) [Pricing](#) [Website](#)

9. Trust



- **Premise:** In order for AI to be useful, we need to know if/when to trust it.
- **Status:** Everyone has heard about hallucinations.
- **Watching:** Deeper questions like: Does my internal AI have access to the data I need it to? Are the citations in Gen AI search accurate?

Trust & Gen AI Search

51.5% of sentences are fully supported with citations

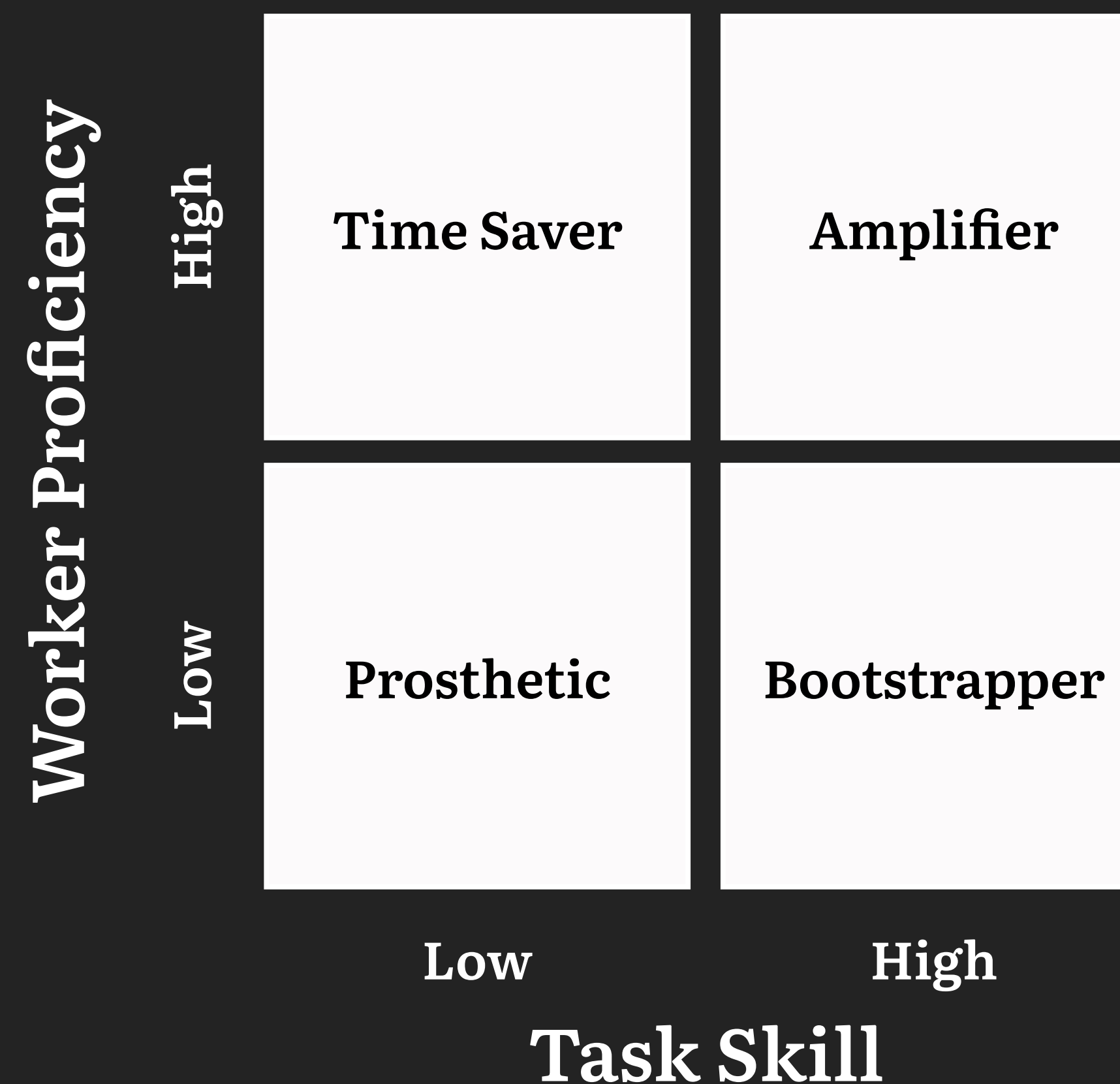
74.5% of citations support their associated sentence

Research source: Evaluating Verifiability in Generative Search Engines, Liu et al. Evaluated Bing Chat, NeevaAI, Perplexity.ai, and YouChat in March/April 2023.

10. World of Workflows



- **Premise:** Generative AI's impact on work will be multifaceted but at its core, the route to higher productivity involves making decisions about whether we want AI to compete with or complement our cognition.
- **Status:** Generative AI alters work by separating work across two dimensions—skill requirement (tasks) and level of proficiency (workers).
- **Watching:** How will people, processes and tool design respond to gen AI workflows.



10 Obsessions for 2024



1. Agentic AI
2. AI Inside
3. Edge AI
4. AI-Enhanced Learning
5. Human Centered Gen AI
6. Interpretability
7. Memory vs Margins
8. Source Dilemma
9. Trust
10. World of Workflows

artificiality

MINDS MEETING MACHINES