just an agent away...

Deconstructing Today's AI Agents Landscape

Eximius Ventures

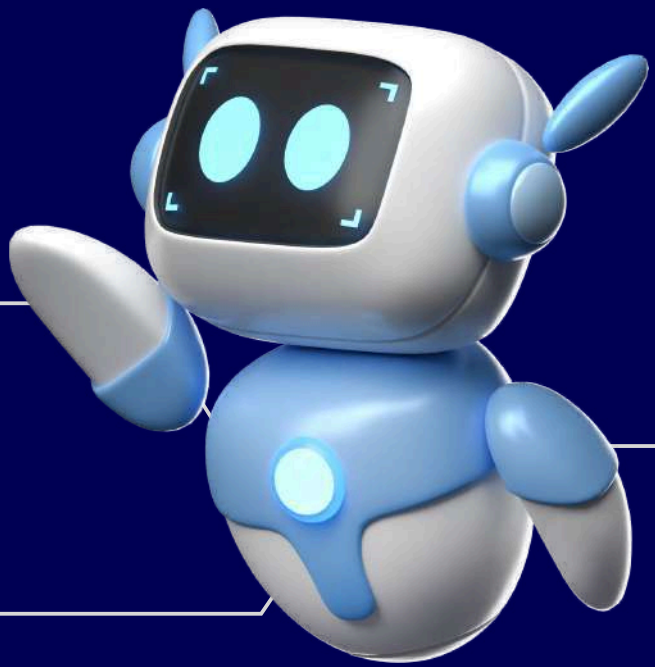

Table of Contents

Foreword

The pace of technological advancement has become both exhilarating and, at times, overwhelming. As investors and builders in the technology ecosystem, we find ourselves not just observing this acceleration but actively participating in shaping its trajectory. The emergence of AI agents represents one of those pivotal moments that commands our attention and challenges our understanding of what's possible.

In the current landscape, we're witnessing an interesting paradox: while AI agents have captured widespread attention, the discourse seems to oscillate between superficial hype and deeply technical dissertations. This gap presented us with an opportunity – and perhaps a responsibility – to bridge these extremes with a pragmatic, grounded perspective that serves our community.

This blueprint emerged from countless conversations with visionary founders and fellow investors who generously shared their insights. We are particularly grateful to Rak Garg (Bain Capital Ventures), Samir Kumar and Evan Wijaya (Touring Capital), Sudhee Chilappagari (Battery Ventures), Deedy Das (Menlo Ventures), Ashwin Raghav Mohan Ganesh (Unbound), Ravi Tandon (DecoverAI), Sahil Agarwal (Enkrypt.ai), Nirav Bhan (Floworks), Krishnakanth Govindaraju (Freshworks), Naman Maheshwari (Tune AI), Soham Ganatra (Composio), Ankit Maheshwari (Innovaccer), Vyas Sekar (Carnegie Mellon University) for their invaluable contributions in helping us connect the dots across this rapidly evolving landscape.

What makes this analysis unique is our deliberate approach to viewing AI agents through multiple lenses. We've worn the hat of a founder contemplating product-market fit, an investor evaluating long-term potential, and a product manager considering real-world implementation. This multifaceted perspective has helped us distil signal from noise, combining primary insights from our network with rigorous secondary research.

As we step into 2025 – widely anticipated as the year of AI agents – our goal is to equip you with a comprehensive understanding that goes beyond the headlines. Whether you're a founder building in this space, an investor evaluating opportunities, or a technology leader planning your AI strategy, we believe this blueprint will help you navigate the landscape with clarity and confidence.

We're excited to present our Blueprint of the State of AI Agents.

Preeti N Sampat
Partner, Eximius Ventures

Shubham Titare
Investments, Eximius Ventures

just an agent away…

By [Preeti Nellore Sampat](), [Shubham Titare]() with o1 and Claude 3.5 Sonnet.

Section 1: Introduction

> *The future arrives unevenly—some hail it as magic, others dismiss it as snake oil. Truth often lies somewhere in between.*

From the outside, it can be hard to tell progress from illusion. We live in a time when AI breakthroughs are heralded daily, yet many claims amount to little more than a slick demo in search of substance. Venture halls and boardrooms echo with promises that "this agent will revolutionise your entire workflow," while an undercurrent of scepticism warns us not to fall for charlatans peddling high-tech snake oil. So where does genuine possibility end, and hyperbole begin?

To find clarity, it helps to think philosophically about what we really mean by an "agent." By definition, an agent acts—it doesn't merely store knowledge or regurgitate information, it makes decisions and executes on them. This is a profound shift. Most software until now has simply digitised manual tasks or data, but AI agents promise to blur the line between application and actor. Philosophers have long wrestled with the nature of autonomy—what does it mean for a being (human or otherwise) to choose a course of action? With AI agents, that debate leaves the realm of academia and lands squarely in our daily workflows.

"This invention could be a game changer for hamster fitness."

Source: Paul Noth, The New Yorker

Yet, as we explore this frontier, we're also forced to confront enduring truths about building great products and businesses. Much like the early days of the internet or mobile apps, it's not enough to grab the latest AI model and stamp "agentic" on a pitch deck. Only those who pair real solutions with deep user understanding will endure. Hype fades fast, but well-executed innovation can transform entire industries.

In the sections ahead, we'll break down the essence of AI agents, digging into their anatomy, new business models, and the practical realities of finding product-market fit. You'll see recurring themes: the tension between autonomy and human-in-the-loop oversight, the disintegration of traditional moats like data and tech exclusivity, the rise of distribution as king, and how incumbents grapple with cannibalising their own revenue streams. By separating genuine breakthroughs from flashy talk, we hope to provide a grounded perspective on where AI agents can truly make their mark—and how investors, founders, and technology leaders can seize this moment with both caution and excitement.

Welcome to "Just an Agent Away..." Let's cut through the noise and uncover what real agentic technology looks like in practice—and why it matters right now.

Section 2: Defining AI Agents

How are agents different from traditional software? Is RAG an agent?

One challenge plaguing the discourse is that the term *agent* has been used inconsistently across industry and academia. In traditional AI, an "agent" is any system that perceives and acts on its environment—under that definition, even a simple thermostat might qualify. Thus, in a frenzied LLM era, the word *agent* has become a buzzword brandished by startups for marketing, often with no standardised meaning. Critics push back, arguing it is meaningless hype.

Yet the concept isn't empty. Many practitioners are actively working to refine our collective understanding of what constitutes an agent in the context of LLMs [OpenAI, LangChain, Lilian Weng, Anthropic]. One particular definition that resonates well and has been adopted for this essay is as follows:

> *An AI agent is a system that uses an LLM to decide the control flow of an application.*

This means that instead of having all logic pre-coded like traditional software, the LLM dynamically decides how the application operates, determining actions to take, tools to use, and responses to inputs.

The degree of control given to an LLM in guiding an application's flow allows for varying levels of autonomy, which we call **agentic**. This framework views agency not as a binary distinction but as a spectrum—a perspective shared across industry practitioners and academics.

This framework helps us make clear distinctions: popular architectures like RAG (Retrieval-Augmented Generation), while powerful, operate on predetermined steps coded

into the application. Though they use LLMs as tools for search, synthesis, and generation, they lack the dynamic control flow that defines true agency.

Section 3: Anatomy of an Agent

What are the different types of agents? What powers an agentic system?

Components of an Agent

Agents centre on four primary components that enable them to handle and act on complex tasks. First is **reasoning**, which allows the agent to interpret unstructured data. Contemporary foundation models embed partial world knowledge into their weights, giving them a strong baseline for logical inferences and semantic comprehension. Next is **external memory**, typically using vector and graph databases that let agents capture domain-specific context and recall information beyond general pre-trained knowledge. This storage complements the model's inherent capacities and allows for extended context across multiple steps.

A third essential component is **execution (tool use)**, meaning an agent can call external functions or APIs to achieve tasks that go beyond text generation—such as searching the web, pulling in enterprise data, or running code. Lastly, **planning** enables the agent to break down big problems into smaller sub-tasks and adapt its strategy based on intermediate results. Through iterative reflection and readjustment, agents can avoid the pitfalls of simply generating a response in one shot.

Not all current agents implement every component, but these building blocks form the foundation of agentic systems.

Types of Agents

Agents can be categorised into three broad types based on LLM autonomy:

1. **Decisioning Agent (Router Agent)**: These agents occupy the constrained end of the agentic spectrum. Here, LLMs act as routers to traverse through a predetermined decision tree. While language models govern the flow of the application, most of the logic remains hard-coded. Anterior's clinical review system illustrates this model: payer rules are mapped into a directed acyclic graph. The LLM moves step by step, evaluating medical documents against each node in the graph. Coding these systems is relatively straightforward because there's limited stochasticity (variability) to control for.

2. **Agent-on-Rails ("Goldilocks" Agent)**: This type balances autonomy and oversight. It has a high-level goal and some freedom to select from predefined tools and approaches, yet it remains tethered to a structured SOP or rulebook that curbs any tendency to stray off course. A typical cycle involves planning, choosing from limited actions, verifying alignment with guardrails, and then looping back to plan again. Many players (like All Hands AI and DevRev) have converged on this architecture because it preserves control while providing enough flexibility to handle varied tasks. However, building this design is more complex than creating a Router agent, as it requires weaving substantial stochasticity into a structured architecture.
3. **General Agent**: At the far end are general AI agents, the so-called holy grail of agentic design. This approach, seen in early prototypes like AutoGPT, attempts to rely on the LLM's own reasoning and coding capabilities without fixed rails. Theoretically, you can implement such an agent in a simple for-loop, letting the LLM pick an action at each step. While this design sparks imagination, it remains susceptible to deviation and inconsistency. A stable, fully adaptable agent—capable of handling virtually any task with minimal guardrails—remains a long-term ambition.

Section 4: The o(h!)1 Moment - Reasoning Models and its Implications

What are reasoning models? What's inference time scaling?

OpenAI's o1 model family marks a major shift in how LLMs handle complex tasks, especially those requiring deep reasoning and multi-step logic. Traditional "System 1" LLMs excel at quick pattern matching but often falter on harder problems. By integrating "System 2" reasoning into its architecture, o1 devotes extra compute to break challenges into smaller steps, refining solutions in real-time. This approach is particularly powerful in math, coding,

and logical planning, reducing the need for extensive prompt engineering or elaborate chain-of-thought prompts.

However, this added compute significantly raises costs and can cause latency to spike—sometimes by factors of 10 or more. Tests indicate a multiple-fold jump in expenses compared to standard models, so the trade-off is substantial. For routine text generation or classification, older models remain faster and cheaper. Yet for high-value domains such as intricate financial analysis, rigorous planning, or advanced research, o1's deliberative power can justify the cost if near-expert reasoning is essential. User studies show remarkable gains in math and logic accuracy with enough tokens and time, though its advantage is less pronounced in general writing, translation, or purely knowledge-based tasks.

In agentic architectures, o1's benefits become especially clear at higher planning and decision-making levels. Traditional AI agents typically orchestrate multiple specialised models—one for conversation, another for domain-specific retrieval, and so on. Embedding o1 at the top of these stacks enables it to parse objectives, tap into domain knowledge, and craft detailed task sequences that cheaper models then execute. This approach bolsters reliability in areas demanding deep deliberation, from drafting comprehensive legal contracts to orchestrating multi-stage data extraction or diagnosing obscure engineering issues.

Still, o1 is unlikely to replace smaller or older LLMs entirely. Cost and speed constraints alone make them preferable for many subtasks. It's neither economical nor necessary to employ o1 for short answers or routine classification, especially when GPT-4 or other capable "mid-tier" models perform sufficiently well at a fraction of the cost. Moreover, specialised or fine-tuned models often excel in narrower contexts—like summarising thousands of records—where multi-step reasoning is overkill.

In practice, a hybrid orchestration strategy is ideal: let o1 handle the complex "System 2" reasoning—planning, deducing, and validating—while delegating simpler tasks to narrower, cheaper models. This synergy leverages o1's strengths where they truly matter, mitigating latency and cost concerns across agentic systems.

Recent developments in "System 2" reasoning models have further advanced the field. OpenAI's introduction of the o3 model established new standards for multi-step reasoning, particularly in challenging domains like ARC-AGI and Codeforces. Shortly thereafter, DeepSeek-R1 emerged with a reinforcement learning-first approach, bypassing supervised fine-tuning while demonstrating robust chain-of-thought capabilities. Notably, DeepSeek-R1 achieves superior performance to the unreleased OpenAI o3 in coding tasks on Codeforces

and ARC-AGI, while offering more competitive pricing ($2.19 per million tokens versus $60 per million tokens for o1).

Section 5: Why Must We Look Beyond Tech? – Rethinking Moats

Is tech a moat? How to build strong differentiators?

Over the past year, many AI startups have touted proprietary data and custom architectures as their competitive advantages. Yet the rapid evolution of LLMs has made these claims increasingly tenuous. Instead, true defensibility in AI often arises from deep system integration, personalisation, effective distribution, domain expertise, and standout user experiences. Below is a closer look at why certain defences are overhyped—and what actually constitutes a moat in this space.

Overhyped Defenses

Proprietary Datasets:
During the very early days of mainstream generative AI hype until 2023, generative AI founders emphasised data moats. Many built their own models or acquired large datasets for specialised fine-tuning and reinforcement learning. This tactic worked early on, allowing them to differentiate in narrow domains. However, the release of GPT-4 changed the landscape. With a larger model size, expanded context window, and 90–95% accuracy across many tasks, GPT-4 required less extensive fine-tuning to achieve strong performance. Even datasets once considered unique—such as specialised radiology images or legal contracts—lost some of their moat value as models trained on broad data rivalled domain-specific accuracy. A more durable defence is a “closed-loop” system that continuously captures client-specific outcome data. Over time, that personalised history becomes a barrier to replication.

Custom Architecture:
Some startups promote novel architectures or techniques they believe are difficult to replicate. In reality, AI research is highly open and fast-moving. Breakthroughs often appear in open publications, and talented engineers move quickly between firms. Even notable innovations like FlashAttention, created by Tri Dao to optimise memory usage in transformer attention, were integrated into major frameworks within months. Techniques such as Chain-of-Thought (CoT) have similarly been absorbed into newer releases, like OpenAI’s o1.

The voice domain has followed suit: cascading architectures (transcribing speech to text, sending it to an LLM, and converting text back to speech) are already being replaced by LLMs like GPT-4o that support speech natively. Ultimately, custom architectures can serve as helpful stopgaps, but they lose relevance as foundational models grow more capable.

Actual Moats

Distribution and Domain Expertise:
Companies with privileged access or relationships gain a powerful distribution advantage. Abridge, for instance, benefits from its partnership with Epic in delivering AI scribe solutions, granting it direct pipelines to hospitals and clinics. Founders that bring in strong C-suite access likewise enjoy an early distribution advantage, securing pilot programs and scaling beyond experimental budgets.

Likewise, in deeply regulated or traditional sectors, domain expertise matters more than baseline AI capabilities. Sectors like legal, health and manufacturing still rely heavily on phone calls, emails, and spreadsheets, so any software solution requires intricate knowledge of procurement, compliance, and user behaviour to achieve penetration. That know-how, not the AI algorithm alone, becomes the differentiator.

Workflow Integration:
AI-native platforms that capture data at its source can begin as middleware and gradually usurp legacy systems. Liberate, for example, fields insurance claims through AI voice agents, initially integrating with Guidewire but gathering enough real-time data to one day supplant it. Similarly, Fixify handles IT issues at the point of customer contact, tapping into data before it enters a tool like ServiceNow. By controlling data flows and embedding themselves into crucial processes, these companies make it difficult—and expensive—for customers to switch.

Deep Personalisation:
Systems that learn from a client's unique history create a personalised "memory" not easily replicated elsewhere. An AI sales agent that refines its approach based on a specific client's prospect data offers insights another platform cannot simply inherit. This effect is strongest where variability is high and past interactions significantly influence outcomes, discouraging customers from abandoning their built-up history.

Niche Differentiation:
Markets can sustain multiple winners if each specialises in a different vertical or niche. While Sierra and Liberate both offer AI-driven support, Liberate zeroes in on insurance integration with platforms like Guidewire. By catering to the specific compliance and workflow demands of insurance, it gains a foothold that a more generic provider would find challenging to replicate.

User experience:
In the rush to launch AI products, many founders overlook the fundamentals of good product

design and engagement. Perplexity, a generative search platform, wins due to its clean interface and intuitive experience, encouraging repeat use. Cursor is another prominent example: although it also calls Anthropics' models under the hood, one of its core innovations lies in a simple yet powerful UX that lets engineers view code diffs and apply them with one click. That usability propelled Cursor to over $100 million in ARR and a $2.6 billion valuation. Github Copilot's success also owes much to seamless UX—users continue coding as usual while the tool unobtrusively suggests improvements.

These patterns reveal an important truth: evaluating AI-native applications follows many of the same principles we've long applied to traditional SaaS. The fundamentals of building defensible businesses remain largely unchanged.

Section 6: Strategic Moats Across Application Domains

How do moats play out?

Within the broad AI landscape, agentic applications often fall into three categories: horizontal, functional, or vertical.

Horizontal agents—like AI-powered note-taking or CRM tools—have broad applicability across roles and industries. To survive in this crowded space, frictionless user experience and deep integration into existing workflows are paramount (think 10x better experience). Distribution deals also create an important competitive advantage, though dependence on distributors carries the risk of marginalisation if they launch equivalent native features, as seen in the competition between Otter and Google's Gemini-powered note-taking capabilities.

Functional agents specialise in specific enterprise functions, such as customer support or sales. In this category, distribution effectiveness serves as a crucial differentiator, closely tied to the quality of the founding team. Development speed (for example, shipping velocity) represents another key competitive advantage. For instance, Decagon, focusing on customer support, releases feature updates weekly, attracting clients seeking to reduce call-center costs. While their product may not be uniquely innovative, their rapid improvement cycle and strong customer relationships ensure market sustainability.

Vertical agents tackle specialised, regulated fields like healthcare and law. Companies like Silna in health often succeed because they understand complex compliance

requirements and deeply ingrained purchasing processes. Though they may benefit from partial data moats—especially when they accumulate specialised datasets over time—the real advantage lies in speaking the language of arcane regulations and integrating with industry-standard systems. Vertical players can even build entirely new systems of record, capturing workflows from day one and embedding themselves so thoroughly that displacement becomes nearly impossible.

Section 7: David vs Goliath

What gives startups an advantage over an incumbent?

Given the fleeting nature of tech moats and distribution being a critical advantage, one might wonder why don't giant incumbents like Salesforce and Zendesk just "bolt-on" AI and outcompete all these agentic startups. After all, they already have a brand, a product suite, and an installed base of thousands of customers. They can simply layer AI into existing workflows, snap their fingers and be done. Right?

Yes and no. Incumbents have extraordinary distribution advantages. They also have constraints.

Incremental AI adoption vs AI-native innovation

Many established vendors have introduced AI in incremental ways – primarily chatbots or summarisation tools layered onto existing products. Even Salesforce's recent Agentforce announcements, for instance, focus (in large part) on moving data among its own suite of offerings rather than fundamentally reshaping workflows. By contrast, genuine AI-native solutions offer deeper, integrated functionality that reimagines how users interact with data. Consider Clay, an AI-native CRM expected to achieve >$500M ARR this year. Clay weaves AI into its core workflows, such as automatically enriching incoming leads with public web data. In principle, Salesforce could attempt similar functionality, but thousands of clients have customised their Salesforce instances. Adding a single new feature can break these intricate setups, creating pushback from major customers who value consistency over innovation.

Revenue cannibalisation, margin erosion and bureaucratic barriers to innovation

Cannibalisation is another deterrent for established players. Google, for instance, could feasibly integrate advanced generative features into search—yet it proceeds cautiously given that undercutting its core advertising revenues would be detrimental. Perplexity faces none of that tension and can move faster in building innovative features.

AI applications often deliver services at lower costs with reduced human involvement. This efficiency threatens the traditional 80-85% margins that incumbents have historically enjoyed and seek to protect.

Corporate bureaucracy further impedes innovation, as organisational inertia and internal incentives favour predictable revenue from existing products over disruptive changes.

Pricing model inflexibility

Pricing and business model constraints also undermine incumbents' attempts at deeper AI integration. While the cost of AI inference keeps dropping, it still remains substantial. Many AI-native startups adopt usage-based pricing: if an inference costs ten cents, the startup charges up to a dollar, pocketing the difference. Traditional software giants, however, have historically relied on seat-based pricing, and find this transition problematic. Customers accustomed to a per-seat model are reluctant to pay more for each AI interaction. However, bundling unlimited AI features into a seat license risks skyrocketing inference costs and eroding provider margins. This tension makes it incredibly difficult for large players to scale powerful AI capabilities without facing either customer dissatisfaction or severe profitability concerns.

Agentic startups approach pricing through an outcome-based lens. Customer support incumbents like Zendesk follow a traditional seat-based approach. In contrast, Decagon calculates how many support tickets can be processed with its solution and charges a percentage of the resultant savings. This appeals to customers who see clear, quantifiable benefits, while incumbents struggle to pivot to such a radically different structure. An established player cannot easily shift to outcome-based pricing without unsettling existing customers, contracts and internal revenue and profit projections.

Startups can thrive without needing to displace giants

Importantly, startups do not need to supplant incumbents outright to succeed. The market's scale enables new players to establish profitable niches. During the SaaS era, Slack thrived alongside Microsoft and Atlassian. Asana approached billion-dollar revenue despite competing with major software suites, while Box built a billion-dollar run rate in commodity

cloud storage. Similarly, capturing even a modest share of an incumbent's market can generate substantial returns for emerging AI ventures.

AI agents have moved beyond the experimental phase into real enterprise adoption. Menlo Ventures' 2024 survey of 600 IT decision-makers revealed a striking shift: from zero adoption in 2023 to 12% in 2024. Yet, despite this momentum, organisations face significant challenges. Both Menlo Ventures and Insight Partners' surveys identify reliability as the primary concern, particularly when agents interface with customer-facing operations or compliance requirements.

Section 8: Autonomy vs. Human-in-the-Loop

Are today's agents truly autonomous? What about hallucinations?

The promise of fully autonomous AI agents meets practical reality in today's market. While complete automation remains the ultimate goal, successful production systems have adopted a more pragmatic approach: strategic human oversight. These systems automatically process routine tasks but transition to human operators for edge cases and low-confidence scenarios. This hybrid model maximises efficiency by allowing AI to handle the bulk of operations while preserving human judgment for the critical 5-10% of cases where errors could be costly.

Industry leaders exemplify this balanced approach. Sierra's architecture automatically routes uncertain responses to human operators, while Profit Security's remediation agent escalates complex security alerts. Both platforms achieve 90-95% automation while maintaining human oversight for challenging cases. Meanwhile, robust AI security measures have become paramount to protect mission-critical data and maintain regulatory compliance in the face of growing cyber threats. This architecture has proven particularly appealing to enterprise customers, who often prefer having human judgment as a safeguard for mission-critical operations, especially given liability, regulatory, and trust considerations.

The technical foundation of these systems relies on confidence thresholds. Actions proceed automatically when confidence levels are high, but fall to human operators when uncertainty exceeds preset limits. These thresholds evolve through reinforcement learning as model performance improves. Equally crucial is robust system observability – operators need clear visibility into data lineage and decision logic. When corrections occur, they feed directly into training pipelines, driving continuous system improvement.

Some sectors and use cases, however, have embraced complete automation, particularly where risks are minimal or small inaccuracies can be absorbed as operational costs. Pallet demonstrates this in the freight industry, using multi-agent systems to automate the

conversion of clipboard notes into CRM data. Since minor errors neither significantly impact freight operations nor generate substantial costs, customers in this sector readily accept full automation.

The market is also seeing the rise of flexible hybrid models offering both autonomous and assistive capabilities within single product suites. Companies like Eve and EvenUp exemplify this approach. EvenUp, for instance, fully automates initial demand letters for personal injury cases – generating qualified leads autonomously – while providing lawyers with assistive tools for subsequent case management and resolution.

Section 9: PMF in the AI era

What's experimental revenue? Have PMF metrics changed?

Agentic startups are demonstrating rapid initial revenue growth that contrasts sharply with the traditional SaaS trajectory. In earlier years, SaaS companies typically took 12-18 months to hit their first million dollars in ARR. Today's agentic companies are surpassing that mark—often reaching $5M or more—within similar or even shorter timeframes. Sierra AI, for instance, grew from $1M to $20M in under a year, and similar examples exist across sectors.

Outcome-based or value-based pricing models are fueling this velocity: instead of billing per seat, many AI startups charge for measurable outcomes (e.g., tasks automated) or consumption (e.g., documents processed). This accelerates revenue recognition but raises questions of staying power. A key concern is whether early revenues truly indicate PMF or simply reflect "experimental" budgets.

Due to macroeconomic factors over the last few years, enterprises have prioritised efficiency. AI solutions, from Glean's unified search to Cursor's code editor, promise needed productivity boosts. Many large companies now allocate millions in experimentation budgets specifically for AI, distributed among various providers. In fact, 60% of the $13.8 billion spent on generative AI in 2024 comes from innovation budgets.

While this experimentation benefits AI startups, it results in volatile customer retention. Large enterprises run multiple trials only to consolidate down to two or three vendors within a few years. What looks like PMF today might turn into churn tomorrow.

Defining true PMF in the current environment remains challenging. Traditional SaaS definitions still apply—companies must identify a clear ideal customer profile (ICP), offer a consistent solution to a recurring pain point, and demonstrate recurring usage. When a startup's top customers are geographically scattered or lack common reasons for buying, it often indicates the absence of true PMF. More robust signals include daily usage data, net

promoter scores, and qualitative customer feedback. Glean demonstrates this with enterprise DAU/MAU ratios consistently above 50%. Similarly, Perplexity and Cursor maintain strong retention through engaging user experiences, with Cursor's simple code-diff workflow that developers consistently return to.

The rapid rise and fall of companies like InVision highlights the current market volatility. After reaching $50 million in ARR and scaling to $100M+, InVision dropped to $50M as Figma emerged. Similar risks face companies like Eleven Labs, which reached $50M ARR by offering text-to-speech or speech cloning—an area that foundational providers like OpenAI could potentially enter. The lesson for AI startups is that early success doesn't guarantee staying power unless they continuously refine their product to retain relevance.

Even robust growth to $10 million or beyond no longer guarantees a path to $100 million. This uncertainty is poised to hit later in a startup's life cycle. In previous SaaS eras, failures clustered before Series A, but now, a greater number of failures may surface after the A or B rounds. Interestingly, seed-stage investors are slightly de-risked—if teams cannot find sustainable PMF, many succeed through talent-driven acquisitions, as larger, legacy players seek AI expertise. Thus, at these stages, it remains wise to back strong, mission-aligned teams that can adapt to this fast-moving environment.

Section 10: Broader Themes We are Excited About

SaaS has evolved and extended to the services market. And more.

The transition from cloud computing to AI marks a pivotal shift from Software-as-a-Service (SaaS) to Service-as-a-Software. Initially, software companies leveraged the cloud to become service providers, tapping into a $350 billion market. Today, thanks to agentic applications, AI transforms labour-intensive tasks into software solutions, expanding the addressable market from billions to trillions of dollars. This evolution offers opportunities to bridge labour gaps, replace outsourced roles, and access new markets.

In healthcare, AI-powered scribes like DeepScribe automate documentation, enabling doctors to focus more on patient care and enhancing productivity. In outsourcing-heavy sectors such as legal e-discovery, companies like Decover AI enable in-house solutions, reducing reliance on expensive legal service providers. Additionally, innovative firms like XBOW are creating new markets with AI-driven penetration testers, making continuous security assessments affordable and accessible for businesses of all sizes.

Agents Creating a New System of Record

Organisations struggle with scattered, unstructured data across PDFs, call logs, and isolated CSVs. AI agents can transform this fragmented landscape into a unified system of record through two key approaches:

Data Enrichment: AI extracts and enhances existing information, capturing nuanced elements like speech patterns and contextual details that traditional systems miss.

Database Unification: By connecting previously siloed databases, AI creates a single source of truth. Hospital pharmacies illustrate this power – when prescription data is trapped in separate pharmacy and TPA portals, hospitals lose access to valuable rebate programs like 340B. An AI agent can bridge this gap by aggregating data across PDFs and portals, enabling real-time insights and compliance monitoring.

The Rise of Multimodal Models and AI Voice Agents

Multimodal LLMs are unlocking new value streams, particularly in voice-centric applications. Early voice AI products like Abridge (doctor-patient transcriptions) or Rillavoice (field sales recordings) mostly converted speech to text for analysis. Now, speech-native approaches promise real-time conversations with lower latency and richer contextual understanding—tone, sentiment, and emotional cues. The economic potential is significant: voice AI alone could create an additional $10B in software TAM in the next five years. Beyond customer support chatbots, next-generation voice agents can handle sales inquiries, schedule appointments in industries with labour shortages, or pick up calls after hours (a missed revenue opportunity for many businesses). Vertically specialised agents are emerging, each tailored to domain nuances—auto dealerships, home services, or healthcare with HIPAA compliance. This layering of domain expertise, specialised integrations, and advanced conversational flows raises the bar on what AI can accomplish.

Agents replacing junior to mid-management roles

AI agents are rapidly automating tasks, particularly those involving repetitive or data-heavy workflows. For example, Norm.ai agents audit marketing copy for SEC compliance, previously a role for specialised staff. Companies like 11x automate SDR lead qualification and outreach, slashing manual pipeline management time. This automation reduces headcount or enables teams to focus on strategic tasks, impacting hiring trends—fewer junior compliance or sales support roles, but more emphasis on creative, high-touch positions.

AI also streamlines organisational structures by automating coordination and reporting functions often handled by middle managers. Agents now collate updates, identify shortfalls, and compile dashboards—tasks traditionally performed by engineering or project managers.

With increasingly sophisticated capabilities, future “AI managers” might delegate tasks, track progress, and optimise workflows across entire departments.

Synthetic Data, AI Agents, and Scientific Breakthroughs

While synthetic data can boost AI performance, its impact is highly domain-specific. In fields like materials science and chemistry, computationally generated data can be nearly as reliable as physical experiments, making it a powerful way to scale discoveries. In more open-ended domains—like vision, where real-world variation is immense—synthetic data often requires careful balancing with real samples to avoid skewed models. Despite these nuances, synthetic data remains a critical enabler in areas where traditional datasets are scarce or prohibitively expensive to gather. By combining AI agents with high-fidelity synthetic data in select industries, we see a path for breakthroughs that simply weren’t feasible before.

Section 11: Conclusion

A Future (Still) Under Construction

In many ways, the story of AI agents is just beginning. We've gone from simple decision trees to hybrid models that juggle human oversight and machine autonomy—and now we stand on the threshold of an even more transformative shift. When agents no longer rely on humans to enter, curate, or extract data, the very notion of a "system of record" must evolve or risk fading into irrelevance. In a future where machine-to-machine interactions outpace those involving human operators, today's workflows and CRMs may feel as antiquated as paper ledgers. After all, if AI can request information, update databases, and reconcile discrepancies on its own, what new architectures will replace our current, human-centric data hubs?

That question underscores how dramatically agentic technology might reshape our digital infrastructure. Entire product lines—once built around manual data input—could disappear or be forced to reinvent themselves. We may see specialised "machine-grade" systems of record optimised for M2M communication, with humans stepping in only at pivotal checkpoints for strategy, compliance, or ethical oversight. The lessons are clear: moats can crumble overnight, domain know-how outlasts ephemeral tech advantages, and authentic distribution trumps mere technical cleverness. Yet through it all, the most audacious builders will keep pushing the boundaries of what autonomy can achieve.

Where does it all lead? Perhaps to new business models that bundle services and software into a single, fluid ecosystem—where overhead becomes an afterthought, and agentic orchestration manages the rest. Or to emergent forms of managerial AI that coordinate entire departments without anyone pressing "enter." We can't know every detail of what's next, but we do know that a lot of these transitions are closer than they seem. When those moments arrive, they won't feel like hype or snake oil—they'll simply feel inevitable. Because in this rapidly shifting world, the future is closer than ever.

It's only, as they say, *just an agent away…*

If you are building in this space, reach us out at shubham@eximiusvc.com and preeti@eximiusvc.com

References

[1] S. Huang, P. Grady, and o1, “Generative AI’s Act o1,” *Sequoia Capital*, Oct. 09, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://www.sequoiacap.com/article/generative-ais-act-o1/

[2] “2024: The State of Generative AI in the Enterprise,” Menlo Ventures. Accessed: Jan. 27, 2025. [Online]. Available: https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/

[3] S. Huang and P. Grady, “Goldilocks Agents,” *Sequoia Capital*, Jun. 18, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://www.sequoiacap.com/article/goldilocks-agents/

[4] L. Weng, “LLM Powered Autonomous Agents,” Lil’Log. Accessed: Jan. 27, 2025. [Online]. Available: https://lilianweng.github.io/posts/2023-06-23-agent/

[5] S. Huang and P. Grady, “Goldilocks Agents,” *Sequoia Capital*, Jun. 18, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://www.sequoiacap.com/article/goldilocks-agents/

[6] S. Kapoor and A. Narayanan, “New paper: AI agents that matter,” *AI Snake Oil*, Jul. 03, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://www.aisnakeoil.com/p/new-paper-ai-agents-that-matter

[7] darlin, “o1: The Missing Link in AI Agency?,” *blending bits*, Sep. 19, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://blendingbits.io/p/o1-the-missing-link-in-ai-agency

[8] P. Akkiraju, “The state of the AI Agents ecosystem: The tech, use cases, and economics,” Insight Partners. Accessed: Jan. 27, 2025. [Online]. Available: https://www.insightpartners.com/ideas/state-of-the-ai-agent-ecosystem-use-cases-and-learnings-for-technology-builders-and-buyers/

[9] “AI Agents: A New Architecture for Enterprise Automation,” Menlo Ventures. Accessed: Jan. 27, 2025. [Online]. Available: https://menlovc.com/perspective/ai-agents-a-new-architecture-for-enterprise-automation/

[10] LangChain, “What is an AI agent?,” *LangChain Blog*, Jun. 29, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://blog.langchain.dev/what-is-an-agent/

[11] R. Matican, “Part II: Multimodal capabilities unlock new opportunities in Vertical AI ,” Bessemer Venture Partners. Accessed: Jan. 27, 2025. [Online]. Available: https://www.bvp.com/atlas/part-ii-multimodal-capabilities-unlock-new-opportunities-in-vertical-ai#Exciting-developments-in-multimodal-architecture

[12] A. Rampell, “Input Coffee, Output Code: How AI Will Turn Capital into Labor,” *Andreessen Horowitz*, Aug. 22, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://a16z.com/ai-turns-capital-to-labor/

[13] A. Strange, “The AI Future Is Already Here, It’s Just Not Productized Yet,” *Andreessen Horowitz*, Jun. 28, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://a16z.com/ai-workflow-productivity/

[14] J. Shiff, “The Emerging ‘AI Native’ Playbook - Opportunities for Founders and Investors,” *AI Natives*, Nov. 25, 2024. Accessed: Jan. 27, 2025. [Online]. Available: https://theainative.substack.com/p/the-emerging-ai-native-playbook?triedRedirect=true

[15] M. Temkin, “In just 4 months, AI coding assistant Cursor raised another $100M at a $2.6B valuation led by Thrive, sources say,” TechCrunch. Accessed: Jan. 27, 2025. [Online]. Available: https://techcrunch.com/2024/12/19/in-just-4-months-ai-coding-assistant-cursor-raised-another-100m-at-a-2-5b-valuation-led-by-thrive-sources-say/

Eximius Ventures

Investing in Founders
From Ideation to Execution

eximiusvc.com/eximius-echo