# Pricing Models for Agentic AI Systems: From Token-Based to Value-Based Approaches

**AI-Generated Academic Thesis Showcase**

Academic Thesis AI (Multi-Agent System)

January 2025

# Table of Contents

## Abstract

**Research Problem and Approach:** The rapid emergence of agentic AI systems has fundamentally challenged traditional pricing paradigms, creating a significant void in how these autonomous, value-generating services are economically managed. This thesis addresses this problem by systematically exploring the theoretical underpinnings and practical implications of diverse pricing models, from token-based to value-based approaches.

**Methodology and Findings:** Employing a qualitative, theoretical analysis augmented by real-world case studies of leading AI providers (OpenAI, Anthropic, Google), the research critically compares foundational and hybrid pricing strategies. Key findings reveal that while token-based models offer granular cost recovery, they often disconnect from perceived user value, necessitating more sophisticated hybrid and outcome-oriented approaches for sustainable monetization.

**Key Contributions:** This thesis makes three primary contributions: (1) It provides a comprehensive comparative framework for evaluating AI-driven pricing models across economic, ethical, and operational dimensions. (2) It details the unique challenges agentic AI poses to traditional pricing, such as non-deterministic outputs and complex workflows. (3) It proposes innovative hybrid pricing models tailored for future AI agent economies, bridging theoretical insights with practical implementation.

**Implications:** The findings offer crucial guidance for AI companies in designing equitable and profitable monetization strategies, for businesses in adopting AI pricing transparently, and for policymakers in developing adaptive regulatory frameworks. It highlights the impending shift towards agent-to-agent economies, emphasizing the need for ethical, explainable, and value-aligned pricing to unlock AI's full potential.



**Keywords:** AI pricing, agentic AI, token-based pricing, value-based pricing, hybrid pricing, AI monetization, economic models, artificial intelligence, dynamic pricing, LLMs, AI agents, ethical AI, market dynamics, autonomous systems, pricing strategies

# 1. INTRODUCTION

Artificial intelligence (AI) has arrived. This isn't just a minor shift; it's a monumental transformation in human history, fundamentally altering industries, economies, and societies at a speed we've never seen before (Ma, 2024)(mckinsey.com, 2025). No longer confined to science fiction, AI systems now power everything from basic task automation to intricate decision-making processes, serving as essential parts of our modern infrastructure (Rossi, 2024). This rapid technological shift is marked by constant breakthroughs, with AI capabilities quickly advancing and expanding what machines can do (Gaier et al., 2023). As these systems become more advanced—especially with the emergence of agentic AI—the established ways of creating value, delivering services, and exchanging goods are being radically reconfigured (Sanabria & Vecino, 2024). What does this mean for the economy? The impacts are enormous, demanding a fresh look at how AI services are thought about, valued, and—crucially—priced. This paper explores the intricate and often overlooked area of pricing strategies for agentic AI systems, addressing a significant void in both academic discussion and practical business application.

The move from standard software to intelligent, autonomous, and goal-oriented agentic AI systems represents a major step forward (Ranjan et al., 2025). Unlike static software that performs set tasks, agentic AI systems are built to act with some independence, interact with their surroundings, learn from experience, and pursue specific objectives, often showing unexpected capabilities (Bilgihan et al., 2025). These systems excel at adaptive learning, complex problem-solving, and continuous self-improvement, offering immense potential for efficiency gains, innovation, and personalized experiences across various sectors (theaiinnovator.com, 2025). However, the very qualities that make agentic AI so powerful—its autonomy, adaptability, and dynamic value generation—also introduce significant challenges.

# 2. Literature Review

The rapid evolution of artificial intelligence (AI) agents has ushered in a new era of computational capabilities, transforming industries and societal interactions (Ranjan et al., 2025)(Souifi et al., 2024)(Gaier et al., 2023). As these autonomous and semi-autonomous systems become increasingly sophisticated, capable of performing complex tasks, reasoning, and learning, the economic models governing their deployment and consumption have emerged as a critical area of inquiry (Sharma, 2024)(Sanabria & Vecino, 2024). The monetization of AI agents, particularly through dynamic pricing strategies, presents a multifaceted challenge that draws upon established economic theories and novel approaches tailored to the unique characteristics of AI services (Gupta, 2025)(Ma, 2024). This literature review synthesizes existing scholarship on pricing models, with a specific focus on their applicability and adaptation to the burgeoning field of AI agents. It explores the foundational principles of token-based and usage-based pricing, delves into the theoretical underpinnings of value-based pricing, and provides a comparative analysis to illuminate the strengths, weaknesses, and strategic implications of each model within the context of AI monetization.

### *2.1 The Emergence of AI Agents and Their Economic Implications*

The concept of intelligent agents, systems capable of perceiving their environment and taking actions to achieve goals, has a long history in computer science (Ranjan et al., 2025). However, recent advancements in machine learning, particularly deep learning and large language models (LLMs), have propelled AI agents into practical, real-world applications (Adetayo et al., 2024)(Earley, 2023)(Earley, 2023). These agents are no longer confined to theoretical discussions but are actively deployed in diverse domains, from automated customer service and personalized recommendations to complex data analysis and autonomous decision-making (Bilgihan et al., 2025). The architectural complexity of these systems, often involving multiple interacting components and sophisticated reasoning mechanisms, under-

scores the need for robust design frameworks (Ranjan et al., 2025). Ranjan, Chembachere et al. (2025) propose a "Well-Architected Framework" specifically for AI agent systems, emphasizing reliability, security, performance, cost optimization, and operational excellence, all of which implicitly influence the economic viability and pricing strategies for these agents (Ranjan et al., 2025). The increasing adoption of AI agents across various sectors, from finance to healthcare, signifies a profound shift in how businesses operate and deliver value (Rossi, 2024).

The economic implications of AI agents are far-reaching. They promise increased efficiency, productivity gains, and the creation of entirely new services and markets (Sanabria & Vecino, 2024)(mckinsey.com, 2025). However, their development and deployment also entail significant costs, including computational resources, specialized talent, and ongoing maintenance and updates (bcg.com, 2025). Monetizing these sophisticated systems effectively is paramount for sustainable innovation and widespread adoption. Traditional software pricing models, such as license-based or subscription-based approaches, often fall short in capturing the dynamic nature, varying utility, and fluctuating resource consumption inherent in AI agent services (Sharma, 2024). This necessitates the exploration of more flexible and adaptive pricing mechanisms that can align the cost to consumers with the actual value derived or resources consumed. The challenge lies in designing pricing models that are fair, transparent, scalable, and incentivize both the development and responsible use of AI agents (datainnovation.org, 2025)(oecd.org, 2025). As AI agents become more prevalent, understanding the economic landscape and developing appropriate monetization strategies becomes crucial for both providers and consumers of these advanced technologies (Sanabria & Vecino, 2024)(theaiinnovator.com, 2025). The literature highlights a growing recognition of the need for sophisticated pricing mechanisms that can account for the unique characteristics of AI, including its non-linear value creation and dynamic resource consumption (Sharma, 2024).

*2.2 Token-Based Pricing Models in AI*

One of the most prominent and widely adopted pricing models for contemporary AI services, particularly large language models (LLMs) and generative AI agents, is token-based pricing (Adetayo et al., 2024). This model emerged as a direct response to the computational and data processing characteristics of these advanced AI systems. A "token" generally refers to a unit of text, which can be a word, part of a word, or even a single character, depending on the tokenizer used by the specific AI model (Adetayo et al., 2024). The cost of using an AI agent is then directly proportional to the number of tokens processed for both input (prompt) and output (response). This approach offers a granular and seemingly transparent method for quantifying usage and assigning costs.

The rationale behind token-based pricing is deeply rooted in the operational mechanics of LLMs. Training and inference for these models involve extensive computational resources, with the processing of each unit of data (token) contributing to the overall computational load (Adetayo et al., 2024). By linking pricing to tokens, AI service providers can directly correlate revenue with the underlying infrastructure costs, such as GPU hours, memory usage, and data transfer (ispartnersllc.com, 2025). This model is particularly prevalent among leading AI providers. For instance, OpenAI's various models, including GPT-3.5 and GPT-4, utilize token-based pricing, often differentiating costs between input tokens and output tokens, with output tokens frequently being more expensive due to the higher computational effort involved in generation compared to mere processing (stocktitan.net, 2025). Similarly, Anthropic's Claude models also employ a token-based system, often with differentiated pricing tiers based on model size, capability, and context window length (Adetayo et al., 2024). The pricing structure can also vary based on the specific model version, with more advanced or larger models commanding higher per-token rates (my.idc.com, 2025).

The advantages of token-based pricing are multifold. From a provider's perspective, it offers a relatively straightforward way to manage and recover operational costs, especially given the variable and often unpredictable usage patterns of AI services (ispartnersllc.com,

2025). It also allows for flexible scaling, as users only pay for what they consume, without large upfront commitments (bcg.com, 2025). For consumers, the model offers a clear, albeit sometimes complex, understanding of costs, enabling them to optimize their prompts and responses to manage expenditure. Developers building applications on top of these foundational models can factor in token usage into their own cost structures, making it easier to estimate expenses for their end-users (Gupta, 2025). This predictability, at least in terms of unit cost, aids in budgeting and resource allocation for projects integrating AI agents (theaiinnovator.com, 2025).

However, token-based pricing is not without its criticisms and challenges. One significant issue is the lack of direct correlation between token count and perceived value for the end-user (Sharma, 2024). A short, concise response might deliver immense value, while a verbose, token-heavy response might be less useful, yet cost more. This disconnect can lead to user frustration and a perceived unfairness in pricing (getmonetizely.com, 2025). Furthermore, the concept of a "token" itself can be opaque to non-technical users, making it difficult for them to intuitively understand how their usage translates into cost (Adetayo et al., 2024). Different models and providers use different tokenization schemes, meaning the same text might result in different token counts across platforms, complicating comparative cost analysis (stocktitan.net, 2025). This variability adds a layer of complexity for developers who might need to integrate multiple AI models into their applications.

Another challenge arises with the increasing context window sizes of modern LLMs. While larger context windows allow for more sophisticated reasoning and longer conversations, they also mean that more tokens are processed for each interaction, even if only a small part of the context is directly relevant to the current query (Earley, 2023). This can lead to inflated costs for complex, multi-turn interactions or tasks requiring extensive background information, even if the "new" information exchanged is minimal. The issue of "token waste" becomes particularly salient in scenarios where prompts are padded with extensive context for better performance, leading to higher costs without a proportional increase in

value (getmonetizely.com, 2025). Moreover, the optimization of prompt engineering to reduce token count can sometimes compromise the quality or comprehensiveness of the AI's response (Sharma, 2024). Striking a balance between cost-efficiency and performance thus becomes a critical consideration for users.

The future of token-based pricing may see further evolution. Hybrid models incorporating elements of value-based pricing or outcome-based pricing could emerge to address the current limitations (Sharma, 2024). For instance, a provider might offer a base token rate with premium tiers for specific features, guaranteed response quality, or specialized agent capabilities. The dynamic adjustment of token prices based on demand, computational load, or even the complexity of the query is another potential avenue (Mei et al., 2022). The underlying principle of charging for computational units is likely to persist, but the definition of these units and their associated costs may become more nuanced to better reflect the diverse ways in which AI agents create value. The transparency and granularity offered by token-based models remain appealing, but their evolution will likely focus on better aligning cost with perceived utility (Sharma, 2024).

### *2.3 Usage-Based Pricing in Cloud Services and its Application to AI Agents*

Usage-based pricing, often referred to as pay-as-you-go or consumption-based pricing, is a well-established model in the broader technology sector, particularly within cloud computing services (Ravulavaru, 2018). This model charges customers based on the amount of a service they consume, rather than a fixed subscription fee or a per-unit purchase (Madathala et al., 2022). It revolutionized the IT industry by offering unprecedented flexibility, scalability, and cost-efficiency, allowing businesses to align their IT expenditure directly with their actual operational needs and fluctuating demand (Ravulavaru, 2018). The core principle is simple: users pay only for the resources they use, whether it's compute time, storage, data transfer, or API calls.

Major cloud providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) extensively employ usage-based pricing across their vast array of services (Ravulavaru, 2018)(Madathala et al., 2022). For example, AWS charges for EC2 instances based on hourly or per-second usage, S3 storage by gigabyte-month, and Lambda functions by invocation count and compute duration (Ravulavaru, 2018). This granular billing allows companies, from startups to large enterprises, to scale their infrastructure up or down dynamically, avoiding the need for large capital expenditures on hardware that might sit idle. The economic benefits are substantial, including reduced total cost of ownership, improved resource utilization, and greater agility in responding to market changes (Madathala et al., 2022).

The principles of usage-based pricing are highly relevant and transferable to the monetization of AI agents. In fact, token-based pricing discussed in the previous section can be seen as a specific type of usage-based pricing, where the "unit of usage" is a token (Sharma, 2024). Beyond tokens, AI agents can be priced based on various usage metrics, including: 1. **API Calls/Invocations:** Charging per request made to an AI agent's API (getmonetizely.com, 2025). This is common for simpler, discrete AI functions like image recognition, sentiment analysis, or translation services (Ravulavaru, 2018). 2. **Compute Time:** Billing based on the actual processing time an AI agent consumes on a server or GPU (Madathala et al., 2022). This is particularly relevant for complex, long-running AI tasks like model training, large-scale simulations, or intricate data analysis where the duration of computation is a primary cost driver (Ranjan et al., 2025). 3. **Data Processed/Storage:** Charging based on the volume of data an AI agent ingests, processes, or stores. This can apply to agents that manage large datasets, perform continuous monitoring, or require extensive memory (Madathala et al., 2022). 4. **Feature/Function Usage:** In some cases, AI agents might offer a suite of capabilities, and pricing could be structured around the usage of specific advanced features or modules (Sharma, 2024).

The advantages of applying usage-based pricing to AI agents mirror those observed in general cloud computing. It promotes flexibility and scalability, allowing users to experiment with AI agents without significant upfront investment (bcg.com, 2025). This "try before you buy" or "pay for what you use" model lowers the barrier to entry for businesses and developers, fostering innovation and broader adoption of AI technologies (theaiinnovator.com, 2025). It also inherently accounts for the variability in AI agent workloads; an agent performing sporadic, intense tasks will incur different costs than one performing continuous, low-intensity tasks, reflecting actual resource consumption (Ranjan et al., 2025). This aligns costs with operational realities, making it a fair and economically sound model for many AI services (Sharma, 2024).

However, challenges arise when directly translating general usage-based models to the nuanced context of AI agents. One primary difficulty lies in defining the "unit of usage" in a way that is both meaningful to the user and accurately reflects the underlying costs and value (Sharma, 2024). While compute time or API calls are objective metrics, they may not always correlate directly with the business value generated by the AI agent. For instance, an AI agent that prevents a major security breach through a single, critical alert might consume minimal compute resources but deliver immense value (Sanabria & Vecino, 2024). Pricing purely on usage in such scenarios might undervalue the AI agent's contribution.

Another challenge is the potential for cost unpredictability. While usage-based models offer flexibility, they can also lead to "bill shock" if usage spikes unexpectedly or if users underestimate their consumption (getmonetizely.com, 2025). This is particularly true for AI agents that might exhibit emergent behaviors or process unforeseen volumes of data. Mitigating this requires robust monitoring tools, transparent pricing structures, and potentially caps or alerts (ispartnersllc.com, 2025). The complexity of managing multiple usage metrics (e.g., API calls, compute time, data transfer, tokens) for a single AI agent service can also be daunting for both providers and consumers (Madathala et al., 2022). Integrating these different metrics into a coherent and understandable pricing model requires careful design.

Furthermore, usage-based pricing might not always incentivize the most efficient use of AI agents (Sharma, 2024). If a user is charged per API call, they might be incentivized to make fewer calls, even if more calls would lead to better outcomes. Conversely, if charged per compute hour, they might optimize for speed over accuracy. The design of the usage metric therefore needs to be carefully considered to align with desired user behavior and value creation (Sharma, 2024). Despite these challenges, the flexibility and scalability inherent in usage-based pricing make it a foundational model for AI agent monetization. Its continued evolution will likely involve more sophisticated metering, clearer value alignment, and the integration of predictive analytics to help users manage and forecast their AI agent expenditures (Gupta, 2025). The lessons learned from cloud computing's extensive experience with usage-based models provide a strong foundation for developing robust pricing strategies for AI agents (Ravulavaru, 2018)(Madathala et al., 2022).

### *2.4 Value-Based Pricing Theory and its Application to AI Agents*

Value-based pricing (VBP) stands in stark contrast to cost-plus or usage-based models by tethering the price of a product or service directly to the perceived or actual value it delivers to the customer (Sharma, 2024)(Gupta, 2025). Instead of focusing on the cost of production or the resources consumed, VBP emphasizes the benefits, outcomes, and utility that a customer derives from the offering. This approach is deeply rooted in economic theory, particularly in concepts of consumer surplus and willingness-to-pay (aeaweb.org, 2025). When successfully implemented, VBP allows companies to capture a larger share of the value they create, moving beyond mere cost recovery to profit maximization based on customer perception (Sharma, 2024).

The theoretical underpinnings of value-based pricing can be traced back to neoclassical economics, where utility and consumer preferences dictate demand and, consequently, price ceilings (aeaweb.org, 2025). In practice, VBP requires a profound understanding of the customer's needs, their alternatives, and the specific impact the product or service has

on their operations, revenue, or efficiency (Gupta, 2025). This often involves quantifying the economic benefits, such as cost savings, revenue generation, risk reduction, or improved decision-making, that the product or service enables (Sharma, 2024). For example, a software solution that automates a previously manual process and saves a company $1 million annually might be priced at a fraction of that saving, say $200,000, which is significantly higher than its development cost but still represents excellent value for the customer (bcg.com, 2025).

Applying value-based pricing to AI agents presents both immense opportunities and significant complexities. The core promise of AI agents lies in their ability to generate substantial value: automating tasks, providing insights, enhancing decision-making, and personalizing experiences (Sanabria & Vecino, 2024)(Bilgihan et al., 2025). This value can manifest in various forms, such as increased productivity, reduced operational costs, improved customer satisfaction, accelerated innovation, or the creation of entirely new revenue streams (Sharma, 2024)(Liu et al., 2023). For instance, an AI agent that optimizes supply chain logistics could save a company millions in shipping costs and inventory management (Gupta, 2025). An AI agent providing highly accurate medical diagnoses could save lives and reduce healthcare expenditures (who.int, 2025). In these scenarios, the value generated far exceeds the computational cost of running the agent.

However, the quantification of value for AI agents is notoriously challenging. Unlike a tangible product or a clearly defined service, the value of an AI agent can be: 1. **Context-dependent:** The same AI agent might deliver vastly different value in different organizational contexts or for different users (Sharma, 2024). 2. **Indirect and emergent:** The full impact of an AI agent might not be immediately apparent and could emerge over time through synergistic effects with other systems or processes (Sanabria & Vecino, 2024). 3. **Difficult to isolate:** It can be hard to attribute specific business outcomes solely to the AI agent, especially when it operates as part of a larger system or team (Liu et al., 2023). 4. **Perceptual and subjective:** The perceived value can vary significantly among

different stakeholders, influenced by factors like brand reputation, trust, and user experience (Bilgihan et al., 2025). 5. **Dynamic:** The value of an AI agent can change over time as its capabilities evolve, market conditions shift, or user needs change (Sharma, 2024).

Despite these complexities, several approaches can facilitate the implementation of VBP for AI agents. One method involves **outcome-based pricing**, where the customer pays based on the achievement of specific, measurable results (Sharma, 2024). For example, an AI agent designed to increase sales might be priced as a percentage of the incremental revenue it generates. An agent focused on reducing churn might be paid based on the number of retained customers (Gupta, 2025). This model directly aligns the provider's incentives with the customer's success, fostering a partnership approach. However, defining and measuring these outcomes, as well as establishing causality, can be intricate (Sharma, 2024).

Another strategy is **tiered value-based pricing**, where different pricing tiers correspond to different levels of features, performance, or guaranteed outcomes (Sharma, 2024). A basic AI agent might offer standard functionality, while premium versions provide enhanced accuracy, faster processing, higher availability, or access to specialized knowledge bases (bcg.com, 2025). This allows customers to choose a price point that matches their perceived value and budget. Furthermore, **performance-based pricing** could be employed, especially for AI agents involved in optimization or prediction, where the price is adjusted based on the accuracy, efficiency, or improvement rate achieved by the agent (Gupta, 2025).

The literature also suggests the importance of robust communication and education to articulate the value proposition of AI agents (Sharma, 2024). Providers must clearly demonstrate how their AI agents solve specific business problems, deliver measurable ROI, or create strategic advantages (bcg.com, 2025). This often involves case studies, pilot programs, and detailed ROI calculators to help customers visualize and quantify the potential benefits. The challenge is to move beyond mere technical specifications and focus on the transformational impact of the AI agent (Sharma, 2024)(Sanabria & Vecino, 2024).

Moreover, the ethical considerations and trustworthiness of AI agents play a crucial role in their perceived value (legalinstruments.oecd.org, 2025)(brookings.edu, 2025). An AI agent that is transparent, fair, and reliable will inherently be perceived as more valuable than one fraught with biases or unpredictability (oecd.org, 2025). Therefore, investing in explainable AI (XAI) and responsible AI development can indirectly support a value-based pricing strategy by building trust and demonstrating the agent's integrity (nvlpubs.nist.gov, 2025).

In conclusion, while challenging, value-based pricing holds the greatest potential for AI agents, as it allows providers to capture the true economic impact of their innovations (Sharma, 2024). It shifts the focus from inputs (tokens, compute) to outputs (business outcomes), fostering a deeper alignment between providers and customers. As AI agents become more sophisticated and their impact more profound, the ability to accurately quantify and articulate their value will be critical for sustainable growth and fair monetization (Gupta, 2025)(bcg.com, 2025). The ongoing research in this area seeks to develop more robust methodologies for value quantification and to design pricing structures that can effectively translate this value into tangible revenue (Sharma, 2024).

### *2.5 Comparative Analysis of Pricing Models for AI Agents*

The landscape of AI agent monetization is shaped by a confluence of pricing strategies, each with its own merits and drawbacks. A comparative analysis of token-based, usage-based, and value-based pricing models reveals their distinct applicability, economic implications, and strategic considerations for both AI providers and consumers (Sharma, 2024)(Gupta, 2025). This section synthesizes the discussion from previous sections, highlighting the trade-offs and potential for hybrid approaches.

**Table 1: Comparative Analysis of Core AI Pricing Models**

| Feature/Dimension | Token-Based Pricing | Usage-Based Pricing (General) | Value-Based Pricing | Hybrid/Tiered Pricing |
|---|---|---|---|---|
| **Primary Metric** | Tokens (input/output) | API calls, Compute time, Data | Outcomes, ROI, Benefits | Mix of fixed/variable |
| **Cost Alignment** | Direct to compute | Direct to resource usage | Direct to business impact | Balanced cost/value |
| **Value Correlation** | Low/Indirect | Variable, often indirect | High/Direct | Moderate to High |
| **Cost Predictability** | Difficult (tokens opaque) | Moderate (usage can vary) | High (outcome-driven) | Moderate (fixed + variable) |
| **Implementation Complexity** | Moderate | Moderate | High | High (billing, tiers) |
| **User Transparency** | Low (token concept) | Moderate (metric clarity) | High (outcome focus) | Variable (can be complex) |
| **Scalability** | High (pay-as-you-go) | High (elastic resources) | Variable (custom contracts) | High (tiered growth) |
| **Provider Risk** | Low (cost recovery) | Low (cost recovery) | High (outcome dependence) | Moderate (tier balancing) |
| **Incentive Alignment** | Minimize usage | Optimize resource use | Maximize value/outcome | Balance usage & value |
| **Best Use Case** | LLMs, Generative AI | Discrete AI functions, Infra | Mission-critical AI, ROI focus | Diverse users, balanced approach |

*Note: This table provides a generalized comparison. Specific implementations may vary in their nuances and effectiveness.*

### 2.5.1 Token-Based vs. Usage-Based Pricing

At first glance, token-based pricing for LLMs appears to be a specialized subset of broader usage-based pricing models. Both models share the fundamental principle of charging for consumption, offering flexibility and scalability (bcg.com, 2025). However, key distinctions and nuances exist.

**Similarities:** * **Pay-as-you-go:** Both models allow users to pay only for what they consume, eliminating large upfront costs and enabling dynamic scaling (Ravulavaru, 2018). * **Cost Alignment:** They generally aim to align pricing with the underlying operational costs incurred by the provider (e.g., compute, memory, data transfer) (Madathala et2 al., 2022). * **Granularity:** Both offer a granular level of billing, often down to sub-units of consumption (ispartnersllc.com, 2025).

**Differences:** * **Unit of Measurement:** The primary difference lies in the definition of the "unit of usage." For general cloud services, this might be CPU hours, storage GBs, or API calls (Ravulavaru, 2018). For LLMs, it is specifically "tokens," which represent linguistic units (Adetayo et al., 2024). * **Abstraction Level:** Token-based pricing is a higher-level abstraction specific to language models, attempting to quantify the "work" done by the LLM in processing or generating text. Other usage-based metrics (e.g., compute time) are more fundamental infrastructure-level metrics (Ranjan et al., 2025). * **Predictability:** While both can lead to unpredictable costs, token counts can be particularly opaque to non-technical users, making cost estimation challenging without specific tools (Adetayo et al., 2024). General cloud usage metrics, though complex, often have more direct analogies to traditional IT costs. * **Value Correlation:** The correlation between token count and perceived user value can be weak (Sharma, 2024). A short, insightful response might be low in tokens but high in value. Other usage metrics (e.g., successful API calls) might have a more direct, albeit still imperfect, correlation with value.

**Strategic Implications:** For AI providers, token-based pricing offers a clear way to monetize LLMs, directly linking to the computational intensity of language processing (ispartnersllc.com, 2025). For broader AI agents, a mix of usage-based metrics (API calls,

compute time) might be more appropriate depending on the agent's function (Sharma, 2024). For consumers, the choice depends on the nature of the AI service. For general AI infrastructure, usage-based models are standard. For LLM-centric tasks, understanding token mechanics is crucial (Adetayo et al., 2024). The challenge for providers is to make these usage metrics, particularly tokens, as transparent and predictable as possible to avoid "bill shock" (getmonetizely.com, 2025).

### 2.5.2 Value-Based Pricing vs. Token/Usage-Based Pricing

The contrast between value-based pricing and consumption-based models (token/usage) is more fundamental, representing different philosophies of monetization (Sharma, 2024)(Gupta, 2025).

**Core Philosophical Difference:** * **Consumption-Based (Token/Usage):** Focuses on inputs, resources consumed, and operational costs. The price is determined by *how much* of the service is used (Madathala et al., 2022). * **Value-Based:** Focuses on outputs, benefits delivered, and customer willingness-to-pay. The price is determined by *how much value* the service creates (Gupta, 2025).

**Advantages of Consumption-Based (Token/Usage):** * **Simplicity and Transparency (in theory):** Easy to understand the basic mechanism: more use = more cost. * **Cost Recovery:** Directly ties revenue to operational costs, ensuring sustainability for providers (ispartnersllc.com, 2025). * **Low Barrier to Entry:** Users can start small and scale, making it accessible for experimentation and small projects (bcg.com, 2025). * **Fairness (Perceived):** Customers only pay for what they use, which feels inherently fair (Ravulavaru, 2018).

**Disadvantages of Consumption-Based (Token/Usage):** * **Value Disconnect:** May not accurately reflect the actual business value derived by the customer (Sharma, 2024). * **Cost Unpredictability:** Can lead to unexpected high bills if usage patterns are volatile (getmonetizely.com, 2025). * **Incentive Misalignment:** May incentivize users to minimize

usage rather than maximize value (Sharma, 2024). * **Opaqueness:** Specific units (e.g., tokens) can be difficult for end-users to grasp (Adetayo et al., 2024).

**Advantages of Value-Based Pricing:** * **Value Capture:** Allows providers to capture a greater share of the economic value they create (Gupta, 2025). * **Customer-Centric:** Focuses on customer outcomes and benefits, fostering stronger relationships (Sharma, 2024). * **Higher Revenue Potential:** Prices are not capped by production costs but by perceived value, leading to potentially higher margins (bcg.com, 2025). * **Incentive Alignment:** Directly aligns provider and customer incentives towards achieving desired outcomes (Sharma, 2024).

**Disadvantages of Value-Based Pricing:** * **Difficulty in Value Quantification:** Challenging to accurately measure and attribute the value generated by an AI agent (Sharma, 2024). * **Complexity:** Requires deep understanding of customer operations and sophisticated pricing models (Gupta, 2025). * **Risk for Provider:** If outcomes are not achieved, revenue might be jeopardized, especially in outcome-based models (Sharma, 2024). * **Customer Resistance:** Customers may be hesitant if they perceive the price as too high or the value proposition unclear (bcg.com, 2025).

### 2.5.3 Hybrid and Dynamic Pricing Strategies

Given the limitations of any single pricing model, the literature suggests that hybrid and dynamic approaches are increasingly relevant for AI agents (Sharma, 2024)(Gupta, 2025). **Hybrid Models:** These combine elements from different strategies. For example, an AI agent service might have: * A **base subscription fee** (fixed component) for access to the platform and core features (bcg.com, 2025). * **Usage-based tiers** (variable component) for additional API calls, compute time, or data processing beyond the base allowance (Ravulavaru, 2018). * **Premium features or outcome-based add-ons** (value component) for specialized capabilities or guaranteed results (Sharma, 2024). This allows providers to ensure a stable revenue stream while offering flexibility and capturing additional value from high-usage or high-value customers.

**Dynamic Pricing:** This involves adjusting prices in real-time based on various factors such as demand, supply, time of day, computational load, user segment, or even the perceived urgency of the task (Mei et al., 2022)(Schlenther et al., 2025)(Ramezani et al., 2011). AI agents themselves are ideally suited to implement dynamic pricing strategies, using machine learning algorithms to optimize prices continuously (Gupta, 2025). For instance, the price per token for an LLM could fluctuate based on network congestion or GPU availability (Mei et al., 2022). Similarly, an AI agent providing real-time market insights could charge a premium during periods of high market volatility (Ramezani et al., 2011). **Advantages of Dynamic Pricing:** * **Revenue Optimization:** Maximizes revenue by responding to market conditions and willingness-to-pay (Gupta, 2025). * **Resource Optimization:** Smooths demand peaks and troughs by incentivizing off-peak usage (Mei et al., 2022). * **Personalization:** Allows for personalized pricing based on individual user profiles or historical behavior (Gupta, 2025).

**Challenges of Dynamic Pricing:** * **Complexity:** Requires sophisticated algorithms and real-time data processing (Ramezani et al., 2011). * **Fairness Concerns:** Can lead to perceptions of unfairness or price gouging if not implemented transparently (Schlenther et al., 2025). * **Regulatory Scrutiny:** May face regulatory challenges, particularly in sensitive sectors (legalinstruments.oecd.org, 2025).

The concept of "Beyond the Sum: Unlocking AI Agents Potential Through Market Mechanisms" by Sanabria and Vecino (2024) directly supports the notion that effective pricing and market design are crucial for realizing the full economic potential of AI agents (Sanabria & Vecino, 2024). Their work suggests that traditional pricing models may not adequately capture the synergistic value created when multiple AI agents interact or when agents contribute to complex workflows. This points towards the need for pricing models that can account for network effects, collaborative intelligence, and the emergent properties of agentic systems. Such models might involve consortium-based pricing, revenue sharing

based on collective outcomes, or even internal market mechanisms where agents "bid" for computational resources or data access (Sanabria & Vecino, 2024).

Another critical dimension is the "Well-Architected Framework" proposed by Ranjan, Chembachere et al. (2025) (Ranjan et al., 2025). While primarily focused on system design, the principles of cost optimization and operational excellence within this framework directly inform pricing strategies. An AI agent designed for efficiency and reliability will inherently have lower operational costs, allowing for more competitive pricing or higher profit margins under consumption-based models. Conversely, an agent optimized for high-value, mission-critical tasks might justify a premium under a value-based approach (Ranjan et al., 2025). The architectural choices made during the development of an AI agent thus have direct implications for its monetization strategy.

Furthermore, the evolving regulatory landscape surrounding AI, particularly concerns about data privacy, algorithmic bias, and accountability (legalinstruments.oecd.org, 2025)(oecd.org, 2025), will inevitably influence pricing. AI agents that demonstrate superior compliance, transparency, and ethical safeguards might command a premium, reflecting the value of reduced regulatory risk and enhanced trust (Sharma, 2024). This integrates non-economic factors into the value equation, further complicating purely quantitative pricing models.

In conclusion, there is no single "best" pricing model for AI agents. The optimal strategy depends heavily on the specific AI agent's capabilities, its target market, the value it creates, and the provider's strategic objectives (Sharma, 2024). While token-based and usage-based models provide a foundation for cost recovery and scalability, value-based and dynamic pricing strategies offer avenues for capturing the immense, often non-linear, value generated by advanced AI agents. The future will likely see a proliferation of sophisticated hybrid models, leveraging the strengths of each approach and adapting dynamically to the evolving capabilities of AI and the shifting demands of the market (Gupta, 2025)(Ramezani et al., 2011). The ongoing research focuses on developing more robust frameworks for quanti-

fying AI value, designing fair and transparent dynamic pricing mechanisms, and integrating market-based approaches to unleash the full potential of AI agents (Sanabria & Vecino, 2024). This requires a multidisciplinary approach, drawing insights from economics, computer science, business strategy, and ethics (Sharma, 2024)(brookings.edu, 2025).

### *2.6 Related Work in AI Monetization and Market Dynamics*

Beyond the direct pricing models, a broader body of literature addresses the monetization strategies and market dynamics pertinent to AI agents. This includes research on the overall economic impact of AI, strategies for profitable innovation in AI, and the unique market mechanisms that AI agents might facilitate or disrupt (Sharma, 2024)(Sanabria & Vecino, 2024)(Ma, 2024). Understanding these broader contexts is crucial for developing sustainable and effective pricing strategies.

Sharma (2024) directly tackles "AI Monetization: Strategies for Profitable Innovation," offering a comprehensive overview of how businesses can derive economic value from AI (Sharma, 2024). This work emphasizes the need for a strategic approach to monetization, moving beyond mere technological deployment to focus on business model innovation. Sharma's insights underscore that pricing is but one component of a larger monetization strategy, which also encompasses product-market fit, ecosystem development, and intellectual property management (Sharma, 2024). The paper likely delves into how different types of AI capabilities (e.g., predictive analytics, generative AI, autonomous agents) might necessitate distinct monetization pathways, reinforcing the idea that a one-size-fits-all pricing model is insufficient.

The concept of market mechanisms is particularly relevant, especially as AI agents become more autonomous and capable of transacting with each other or with human users in dynamic environments (Sanabria & Vecino, 2024). Sanabria and Vecino (2024), in their work "Beyond the Sum: Unlocking AI Agents Potential Through Market Mechanisms," explore how market-based approaches can foster collaboration and value creation among AI agents

(Sanabria & Vecino, 2024). This implies that pricing for AI agents might not always be a simple bilateral transaction between a provider and a single user. Instead, it could involve complex multi-agent systems where agents themselves participate in economic exchanges, bidding for resources, services, or data. Such scenarios would necessitate advanced pricing algorithms, potentially leveraging game theory and auction mechanisms, to ensure efficiency and fairness (Sanabria & Vecino, 2024)(Ramezani et al., 2011). The emergence of AI-driven intercompany services, as discussed by Rossi (2024), further highlights the complex economic networks in which AI agents will operate, requiring sophisticated pricing and value exchange mechanisms (Rossi, 2024).

The broader economic literature also provides insights into the challenges of pricing intangible assets and services, which AI agents largely represent (aeaweb.org, 2025). Unlike traditional goods, the marginal cost of replicating an AI agent's output is often near zero, while the fixed costs of development and training can be substantial (bcg.com, 2025). This cost structure, characteristic of digital goods, often leads to pricing strategies that aim to maximize consumer surplus capture, rather than simply covering marginal costs (aeaweb.org, 2025). This is where value-based pricing becomes particularly attractive, as it attempts to price based on the utility derived rather than the minimal replication cost (Sharma, 2024).

Furthermore, the literature on dynamic pricing and revenue management in various industries, such as ride-sharing, hospitality, and energy, offers valuable lessons (Schlenther et al., 2025)(Mauri et al., 2019)(Mei et al., 2022). For instance, Mei, Liang et al. (2022) discuss optimal time-of-use pricing for incremental distribution networks (Mei et al., 2022), a concept that can be directly applied to AI agent services to manage computational load and incentivize off-peak usage. Similarly, Ramezani, Bosman et al. (2011) explore adaptive strategies for dynamic pricing agents (Ramezani et al., 2011), providing algorithmic frameworks that AI agents themselves could utilize to set their own prices based on real-time market conditions and demand signals. The application of AI-driven personalized pricing models in e-commerce, as investigated by Gupta (2025), demonstrates how AI can be used

not just as a service to be priced, but as a tool to optimize pricing for other goods and services (Gupta, 2025). This dual role of AI further complicates the monetization landscape, as AI agents can both be subjects of pricing models and architects of pricing strategies.

The architectural considerations for AI systems, as outlined by Ranjan, Chembachere et al. (2025), also have indirect but significant implications for pricing (Ranjan et al., 2025). A well-architected AI agent system that prioritizes cost optimization, for example, might be able to offer more competitive pricing under usage-based models. Conversely, an agent designed for extreme reliability and security might justify a higher price under a value-based framework, as it mitigates significant risks for the user (Ranjan et al., 2025). The design choices made at the foundational level of AI agent development thus ripple through to their eventual market value and pricing strategies.

Finally, the broader societal and ethical discussions surrounding AI are intrinsically linked to its monetization. Concerns about fairness, bias, transparency, and accountability (legalinstruments.oecd.org, 2025)(oecd.org, 2025) can influence the perceived value and public acceptance of AI agents, thereby impacting their pricing potential (Sharma, 2024). An AI agent that adheres to high ethical standards and offers explainability might be deemed more valuable, and thus command a higher price, than one that operates as a black box with unknown biases (nvlpubs.nist.gov, 2025). This highlights the need for pricing models to not only consider economic factors but also incorporate the intangible value of trust and responsible AI practices.

In summary, the literature on AI monetization extends beyond specific pricing mechanisms to encompass broader market dynamics, strategic innovation, and ethical considerations. The works reviewed emphasize that effective monetization of AI agents requires a holistic approach that integrates technological capabilities, market understanding, economic principles, and a keen awareness of the societal context (Sharma, 2024)(Sanabria & Vecino, 2024). The increasing sophistication of AI agents and their integration into complex economic ecosystems will continue to drive innovation in pricing strategies, moving towards more dy-

namic, value-aligned, and market-driven models (Gupta, 2025)(Ramezani et al., 2011). The challenge remains in translating these theoretical and strategic insights into practical, scalable, and fair pricing models that can unlock the full potential of AI agents while ensuring responsible deployment (Sanabria & Vecino, 2024). This literature review sets the stage for a deeper investigation into how these models can be specifically tailored and optimized for the unique characteristics of AI agents in various application contexts.

# 3. METHODOLOGY

The rapidly evolving landscape of artificial intelligence (AI) agents in business operations necessitates a robust methodological approach to understand their impact, particularly within pricing strategies. This section outlines the research methodology employed to explore the theoretical underpinnings and practical implications of AI-driven pricing models. Given the nascent stage of advanced AI agent deployment in complex pricing scenarios, a qualitative, theoretical analysis augmented by illustrative case studies is deemed most appropriate. This approach allows for an in-depth exploration of intricate phenomena, the identification of emerging patterns, and the development of nuanced insights that quantitative methods might overlook in the absence of extensive, standardized data (Souifi et al., 2024). The methodology is structured to provide a comprehensive framework for comparing diverse AI-driven pricing models, detail the criteria for selecting pertinent case studies, and articulate the analytical approach used to extract meaningful conclusions. The goal is to contribute to a deeper academic understanding of how AI agents are transforming pricing paradigms, moving beyond conventional economic models to embrace adaptive, data-driven, and often autonomous decision-making processes (Gupta, 2025)(Ramezani et al., 2011).

## 3.1 Framework for Comparing AI-Driven Pricing Models

The theoretical framework for comparing AI-driven pricing models is grounded in a synthesis of traditional economic pricing theory, principles of AI system architecture, and emerging considerations in AI ethics and governance. Traditional pricing theory, encompassing concepts such as cost-plus pricing, value-based pricing, dynamic pricing, and competitive pricing (Mauri et al., 2019), provides a foundational lens through which to evaluate the economic objectives and mechanisms of AI systems (Mei et al., 2022). However, the unique capabilities of AI agents—such as their capacity for continuous learning, real-time adaptation, and processing vast datasets—necessitate an expansion of this traditional framework (Gupta, 2025)(Ma, 2024). Therefore, our framework integrates dimensions specifically tailored to capture the distinct attributes and challenges posed by AI agent technology.

The comparative framework is structured around five core dimensions: Economic Efficiency and Value Creation, Adaptability and Dynamism, Fairness and Ethical Implications, Transparency and Explainability, and Data Requirements and Security. Each dimension is critical for a holistic evaluation of AI-driven pricing models, reflecting both their potential benefits and inherent risks (Sharma, 2024).

**Figure 1: AI-Driven Pricing Model Evaluation Framework**

*Note: This figure illustrates the five core dimensions used to comprehensively evaluate AI-driven pricing models. Each dimension encompasses critical sub-factors that contribute to a holistic assessment of a model's effectiveness, ethical standing, and operational robustness.*

### *3.1.1 Economic Efficiency and Value Creation*

This dimension assesses the extent to which AI-driven pricing models optimize economic outcomes for firms. This includes metrics such as revenue maximization, profit margin enhancement, cost optimization (Madathala et al., 2022), and market share growth. AI agents, by leveraging advanced analytics and predictive capabilities, can identify optimal

price points that balance supply and demand in real-time, often surpassing human capabilities in complex, volatile markets (Gupta, 2025). The framework evaluates how AI agents contribute to value creation not only through direct financial gains but also through indirect benefits such as improved customer lifetime value, enhanced operational efficiency, and strategic market positioning (Sanabria & Vecino, 2024)(Liu et al., 2023). For instance, AI can enable hyper-personalized pricing strategies that cater to individual customer willingness to pay, thereby extracting maximum value from each transaction (Gupta, 2025). This subdimension will explore the specific algorithms and data inputs AI agents utilize to achieve these efficiencies, such as reinforcement learning models that optimize pricing actions based on market feedback (Gaier et al., 2023) or sophisticated forecasting models that anticipate demand fluctuations. The analysis will also consider the trade-offs between short-term profit maximization and long-term market sustainability, as aggressive AI-driven pricing could potentially lead to customer churn or regulatory scrutiny (legalinstruments.oecd.org, 2025). Understanding the balance between these factors is crucial for assessing the true economic efficiency and sustainable value creation capabilities of these advanced systems (Sharma, 2024).

### *3.1.2 Adaptability and Dynamism*

AI agents are inherently designed for adaptability, learning from new data and adjusting their strategies in response to changing market conditions (Ranjan et al., 2025)(Ramezani et al., 2011). This dimension examines the dynamism of AI-driven pricing models, specifically their ability to respond to external shocks, competitive actions, and shifts in consumer preferences without human intervention. Traditional pricing models often rely on static rules or periodic adjustments, rendering them slow to react to rapid market changes. In contrast, AI agents can continuously monitor market signals, analyze vast streams of data, and autonomously modify pricing strategies in real-time (Gupta, 2025). The framework will investigate the mechanisms through which AI agents achieve this adaptability, including their

learning algorithms, the frequency of model updates, and their capacity for self-correction. It will also differentiate between reactive adaptability (responding to observed changes) and proactive dynamism (anticipating future trends). Examples include AI agents adjusting airline ticket prices based on real-time demand and competitor pricing, or e-commerce platforms dynamically altering product prices during flash sales (mckinsey.com, 2025). The evaluation will also consider the robustness of these adaptive systems, particularly their performance under extreme or unforeseen market conditions, and the potential for algorithmic instability or "runaway" pricing scenarios (mckinsey.com, 2025). The speed and precision with which AI agents can recalibrate pricing strategies represent a significant departure from conventional methods, offering unprecedented agility in competitive environments (Sanabria & Vecino, 2024).

### *3.1.3 Fairness and Ethical Implications*

The deployment of autonomous AI agents in pricing raises significant ethical concerns, particularly regarding fairness, discrimination, and market manipulation (legalinstruments.oecd.org, 2025)(datainnovation.org, 2025). This dimension scrutinizes the ethical implications of AI-driven pricing models. It assesses whether these models inadvertently or intentionally lead to discriminatory pricing practices based on protected characteristics, or if they contribute to market monopolization through anti-competitive collusion (oecd.org, 2025). The framework will evaluate the measures taken to ensure fairness in pricing outcomes, such as the incorporation of ethical constraints into algorithms or the implementation of human oversight mechanisms (brookings.edu, 2025). This includes analyzing how AI agents handle data privacy and security, particularly when processing sensitive customer information to inform personalized pricing (ispartnersllc.com, 2025). The potential for algorithmic bias, where historical data reflecting societal inequalities is perpetuated or even amplified by AI systems, is a critical area of investigation (Lv et al., 2024). For instance, if an AI agent learns to charge higher prices to customers from certain demographics due to

historical spending patterns that correlate with socioeconomic status, this raises serious ethical and legal questions (legalinstruments.oecd.org, 2025). The framework also considers the broader societal impact, such as the potential for AI-driven pricing to exacerbate economic inequalities or create new forms of digital exclusion. The analysis will seek to identify best practices for designing and implementing AI pricing models that uphold ethical principles and comply with regulatory requirements, balancing profit motives with social responsibility (datainnovation.org, 2025).

### *3.1.4 Transparency and Explainability*

The "black box" nature of many advanced AI algorithms poses a challenge to understanding how pricing decisions are made (Ranjan et al., 2025)(Earley, 2023). This dimension focuses on the transparency and explainability of AI-driven pricing models. It assesses the extent to which the decision-making process of AI agents can be understood, audited, and justified to stakeholders, including customers, regulators, and internal management. A lack of transparency can erode consumer trust, hinder regulatory oversight, and make it difficult for firms to identify and rectify errors or biases in their pricing strategies (brookings.edu, 2025). The framework examines the methodologies employed to enhance explainability, such as the use of interpretable AI models, post-hoc explanation techniques, or clear documentation of algorithmic logic. This includes evaluating the ability to trace specific pricing decisions back to their underlying data inputs and algorithmic rules (nvlpubs.nist.gov, 2025). For example, can a company explain why a particular customer was offered a specific price at a given time? The analysis will also consider the trade-off between model complexity (which often correlates with performance) and explainability (Ranjan et al., 2025). While highly sophisticated models may achieve superior economic outcomes, their opacity can introduce significant governance and compliance risks (isaca.org, 2025). The goal is to identify approaches that strike a balance, providing sufficient insight into AI pricing mechanisms without compromising their effectiveness, thereby fostering accountability and trust (datainnovation.org, 2025).

### *3.1.5 Data Requirements and Security*

The performance of AI-driven pricing models is heavily reliant on the quality, volume, and variety of data they consume (Gupta, 2025). This dimension evaluates the data requirements of these models and the security measures implemented to protect sensitive information. It assesses the types of data required (e.g., customer behavior, market trends, competitor pricing, inventory levels), the infrastructure for data collection and processing, and the challenges associated with data integration from disparate sources (Rossi, 2024). Furthermore, it scrutinizes the robustness of data security protocols, including encryption, access controls, and compliance with data privacy regulations such as GDPR or CCPA (ispartnersllc.com, 2025). The framework will investigate the potential vulnerabilities associated with large-scale data aggregation and processing by AI agents, including risks of data breaches, unauthorized access, and algorithmic manipulation through corrupted data inputs (wilmerhale.com, 2025). For instance, if an AI pricing system is fed manipulated competitor data, it could lead to suboptimal or even harmful pricing decisions. The analysis will also consider the computational resources required to train and operate these data-intensive models (Ravulavaru, 2018), and the implications for scalability and cost-effectiveness (Madathala et al., 2022). Ultimately, this dimension aims to understand the data ecosystem supporting AI-driven pricing, identifying best practices for data governance, integrity, and protection to ensure reliable and secure operation of these advanced systems (isaca.org, 2025).

## 3.2 Case Study Selection Criteria

To provide empirical context and illustrate the application of the theoretical framework, this study employs a multiple-case study design (Souifi et al., 2024). Case studies are particularly valuable for exploring complex, contemporary phenomena within their real-world context, especially when the boundaries between phenomenon and context are not clearly evident (Jafari, 2024). This approach allows for an in-depth understanding of the

specific conditions under which AI-driven pricing models are implemented, their operational mechanisms, and their observed outcomes. The selection of appropriate case studies is critical to ensure both the relevance and the generalizability of the findings. The primary goal of the case studies is not statistical generalization but analytical generalization, where findings are used to expand and generalize theories (Jafari, 2024).

The selection process for case studies will adhere to a set of stringent criteria designed to maximize the insights gained into AI-driven pricing strategies across diverse industry contexts. The initial pool of potential cases will be identified through a systematic review of business reports, technology news, academic publications, and industry analyses focusing on companies that publicly discuss or are known to employ AI agents for pricing decisions (mckinsey.com, 2025)(bcg.com, 2025)(theaiinnovator.com, 2025).

The primary selection criteria include:

### *3.2.1 Clear Application of AI Agents in Pricing*

Each selected case study must demonstrate a clear and documented application of AI agents or advanced AI systems in their pricing strategies. This criterion ensures that the cases directly address the core research question regarding the nature and impact of AI-driven pricing. Cases where AI is used merely for data analytics or forecasting without direct involvement in autonomous or semi-autonomous price setting will be excluded (Gupta, 2025). We are specifically interested in instances where AI agents are integrated into the decision-making loop, allowing for dynamic and adaptive pricing adjustments (Ramezani et al., 2011). This includes scenarios such as dynamic pricing in e-commerce, algorithmic pricing in ride-sharing (Schlenther et al., 2025), or personalized pricing in digital services (Gupta, 2025). The evidence of AI application can be derived from company press releases, investor calls, reputable business news articles, or academic studies referencing their specific AI implementations (mckinsey.com, 2025)(bcg.com, 2025).

### *3.2.2 Industry and Contextual Diversity*

To ensure a broad applicability of the theoretical framework and to identify nuanced differences in AI pricing across sectors, case studies will be selected from diverse industries. This diversity is crucial because the competitive landscape, regulatory environment, and customer behavior vary significantly between industries, influencing the design and performance of AI pricing models (Mauri et al., 2019). Potential industries include, but are not limited to, e-commerce, transportation (e.g., ride-sharing, logistics), digital advertising, financial services, and cloud computing (Rossi, 2024). This approach aims to prevent findings from being overly specific to a single industry, thereby enhancing the analytical depth and external validity of the research (Souifi et al., 2024). For example, the ethical considerations of AI pricing in healthcare (who.int, 2025) might differ significantly from those in retail, warranting a comparison. Similarly, the data requirements for pricing complex financial instruments (Paiella, 2004) may be distinct from those for consumer goods.

### *3.2.3 Data Availability and Richness*

The analysis relies on publicly available information and secondary data sources. Therefore, selected cases must have a sufficient volume of accessible data, including company reports, news articles, academic analyses, regulatory filings, and public statements that describe their pricing strategies, AI implementations, and associated outcomes (mckinsey.com, 2025)(bcg.com, 2025)(theaiinnovator.com, 2025). Cases with limited public information regarding their AI pricing practices will be excluded, as they would hinder an in-depth analysis against the proposed framework. The richness of the data is paramount for enabling a comprehensive within-case analysis and subsequent cross-case comparison. This includes details on the types of AI technologies used, the objectives of their pricing models, the data inputs, ethical considerations addressed, and observed impacts on market performance or customer behavior (Sharma, 2024).

### *3.2.4 Illustrative Value and Impact*

Selected cases should offer significant illustrative value, either by showcasing innovative applications of AI in pricing, highlighting critical challenges, or demonstrating notable successes or failures. Cases that have generated public discussion, regulatory scrutiny, or significant market impact due to their AI pricing strategies will be prioritized (brookings.edu, 2025)(wilmerhale.com, 2025). Such cases often provide richer insights into the practical implications of AI-driven pricing, including ethical dilemmas, competitive dynamics, and consumer reactions (datainnovation.org, 2025). The illustrative value ensures that the case studies serve as compelling examples that help to ground the theoretical framework in real-world phenomena and facilitate the identification of actionable insights and future research directions. For instance, a case where AI pricing led to a public backlash due to perceived unfairness would be highly illustrative for the "Fairness and Ethical Implications" dimension of the framework.

### *3.2.5 Temporal Relevance*

Given the rapid advancements in AI technology, selected cases will primarily focus on recent implementations or ongoing applications of AI-driven pricing models, ideally within the last five to ten years (my.idc.com, 2025)(forrester.com, 2025). This ensures that the analysis reflects the current state-of-the-art in AI agent capabilities and market practices, rather than outdated technologies or strategies. While historical context may be briefly referenced, the emphasis will be on contemporary examples that are most relevant to the current discourse on AI and pricing (Gupta, 2025). This criterion helps maintain the currency and practical relevance of the research findings.

Based on these criteria, a sample of 3-5 distinct case studies will be selected. This number is sufficient to allow for in-depth analysis within each case while also enabling meaningful cross-case comparisons to identify patterns and variations (Jafari, 2024). The specific companies chosen will remain anonymous in the final report if the information is not pub-

licly attributed, to protect proprietary information and encourage a focus on the underlying phenomena rather than specific corporate identities.

## 3.3 Data Collection and Analytical Approach

The analytical approach for this study combines within-case analysis with cross-case synthesis, guided by the theoretical framework established in Section 3.1. This dual-layered approach allows for a deep understanding of each individual case study before identifying commonalities, differences, and emergent patterns across the cases. The primary data source for this research will be secondary data, meticulously collected and systematically analyzed.

### *3.3.1 Data Collection Methods*

Data collection will primarily involve a systematic search and review of publicly available secondary sources. These sources include:

- **Company Reports and Disclosures:** Annual reports, investor presentations, sustainability reports, and official press releases from the selected companies (mckinsey.com, 2025)(bcg.com, 2025). These documents often provide insights into strategic priorities, technology investments, and market performance.
- **Business and Industry Publications:** Articles from reputable business news outlets (e.g., Wall Street Journal, Financial Times, Bloomberg), industry-specific journals, and market research reports (e.g., from McKinsey, BCG, IDC, Forrester) (mckinsey.com, 2025)(bcg.com, 2025)(my.idc.com, 2025)(forrester.com, 2025). These sources offer external perspectives on company strategies, market trends, and competitive dynamics.
- **Academic Literature:** Peer-reviewed articles, conference papers, and doctoral theses that analyze the selected companies or the broader application of AI in pricing within their respective industries (Souifi et al., 2024).
- **Regulatory Filings and Policy Documents:** Reports from government agencies, competition authorities, and international bodies (e.g., OECD, EU Commission) that

discuss AI pricing, algorithmic bias, or market concentration issues related to the chosen cases (legalinstruments.oecd.org, 2025)(datainnovation.org, 2025)(oecd.org, 2025).

- **Webinars, Podcasts, and Public Statements:** Transcripts or summaries of interviews with company executives, industry experts, and thought leaders discussing AI pricing strategies (youtube.com, 2025)(youtube.com, 2025).

A structured data extraction protocol will be developed to ensure consistency and rigor in data collection. This protocol will specify the types of information to be extracted for each case study, directly aligning with the dimensions of the comparative framework (Economic Efficiency, Adaptability, Fairness, Transparency, Data Requirements). Keywords related to "AI pricing," "algorithmic pricing," "dynamic pricing," "machine learning pricing," "AI agent pricing," and "ethical AI" will guide the search process across various databases and search engines (Gupta, 2025). The reliability of sources will be critically evaluated, prioritizing information from official company statements, reputable news organizations, and peer-reviewed academic publications (Souifi et al., 2024).

### *3.3.2 Within-Case Analysis*

Once the data for each selected case study has been collected, a thorough within-case analysis will be conducted. This phase involves immersing in the details of each individual case to develop a comprehensive understanding of its specific AI-driven pricing model. For each case, the collected data will be systematically organized and coded according to the five dimensions of the theoretical framework:

1. **Economic Efficiency and Value Creation:** Analysis of reported revenue growth, profit margins, cost savings, market share changes, and customer acquisition/retention rates attributed to AI pricing (Sharma, 2024).
2. **Adaptability and Dynamism:** Examination of how the AI pricing model responds to market fluctuations, competitive actions, and consumer behavior shifts, including the frequency and nature of price adjustments (Ramezani et al., 2011).

3. **Fairness and Ethical Implications:** Identification of any reported instances of algorithmic bias, discriminatory pricing, regulatory scrutiny, or public backlash related to ethical concerns. Also, any stated corporate policies or technical safeguards to ensure fairness (legalinstruments.oecd.org, 2025)(datainnovation.org, 2025).
4. **Transparency and Explainability:** Assessment of the level of transparency provided by the company regarding its AI pricing mechanisms, including any efforts towards explainable AI (XAI) or auditability (nvlpubs.nist.gov, 2025).
5. **Data Requirements and Security:** Documentation of the types of data utilized, the scale of data processing, and reported efforts in data security, privacy compliance, and governance (ispartnersllc.com, 2025)(isaca.org, 2025).

This coding process will involve both deductive coding (applying pre-defined categories from the framework) and inductive coding (identifying new themes or sub-dimensions that emerge from the data) (Jafari, 2024). Detailed case narratives will be developed for each study, providing a rich description of its AI pricing strategy and its performance against the framework's dimensions. These narratives will serve as the foundation for the subsequent cross-case analysis.

### *3.3.3 Cross-Case Synthesis*

Following the completion of individual within-case analyses, a cross-case synthesis will be performed. This involves comparing and contrasting the findings across all selected case studies to identify overarching themes, patterns, similarities, and differences. The aim is to move beyond individual case specifics to develop broader theoretical insights and refine the comparative framework.

The cross-case synthesis will involve:

- **Pattern Matching:** Identifying recurring patterns in how different companies implement AI pricing, the challenges they face, and the outcomes they achieve across the five dimensions (Sanabria & Vecino, 2024). For example, are there common ethical

dilemmas across industries, or do certain AI architectures consistently lead to higher adaptability?

- **Identification of Variations:** Highlighting significant differences between cases, explaining why certain AI pricing models perform differently or face unique challenges in specific contexts. This could involve exploring the influence of industry regulations, market maturity, or organizational culture (Mauri et al., 2019).
- **Refinement of Framework:** Using the empirical evidence from the case studies to validate, refine, or expand the theoretical comparative framework. Emergent themes or previously unconsidered aspects of AI-driven pricing will be integrated into the framework, enhancing its robustness and explanatory power (Ranjan et al., 2025).
- **Development of Propositions:** Formulating theoretical propositions or hypotheses based on the observed patterns and variations, which can be tested in future research. These propositions will articulate relationships between specific AI pricing characteristics and their outcomes (Sharma, 2024).

The analytical process will be iterative, moving back and forth between the data, the within-case narratives, and the theoretical framework. This iterative approach helps to ensure that the interpretations are well-supported by the evidence and that the theoretical contributions are robust.

### *3.3.4 Rigor and Limitations*

To enhance the rigor of this qualitative study, several measures will be employed. **Triangulation** will be achieved by utilizing multiple data sources (company reports, news articles, academic papers) to corroborate findings for each case study (Souifi et al., 2024). This helps to increase the credibility and trustworthiness of the results. The structured data extraction protocol and systematic coding process will ensure **dependability** and **confirmability**. While the use of secondary data limits direct interaction with practitioners, it

mitigates potential researcher bias that might arise from interviews and allows for a broader scope of cases.

A significant limitation of this methodology is its reliance on publicly available secondary data, which may not always provide complete or granular details on proprietary AI algorithms or internal decision-making processes (theregister.com, 2025). Furthermore, the subjective nature of qualitative analysis means that findings are interpretations rather than statistically generalizable facts. However, the goal is analytical generalization, providing rich insights and theoretical contributions rather than statistical validation. The chosen methodology is well-suited for exploring a complex, emerging phenomenon like AI-driven pricing agents, offering a foundational understanding that can inform future empirical research (Sanabria & Vecino, 2024). The detailed articulation of the framework, selection criteria, and analytical steps aims to maximize the transparency and replicability of the research process within these inherent limitations.

# 4. Analysis: Pricing Models for AI Agentic Systems

The advent of artificial intelligence (AI), particularly the proliferation of large language models (LLMs) and the emerging paradigm of AI agentic systems, has ushered in a new era of computational capabilities and, consequently, complex challenges in value capture and monetization (Sharma, 2024)(Sanabria & Vecino, 2024). Traditional software and service pricing models, while providing a foundational framework, often fall short in adequately addressing the unique characteristics of AI, such as its non-deterministic outputs, evolving capabilities, and the distinct cost structures associated with inference versus training (Gupta, 2025). This section undertakes a comprehensive analysis of various AI pricing models, comparing their advantages and disadvantages, examining real-world implementations by leading providers like OpenAI, Anthropic, and Google, and proposing hybrid approaches specifically tailored for the intricate demands of AI agentic systems. The objective is to delineate how

organizations can effectively monetize AI innovations while fostering widespread adoption and ensuring sustainable development (mckinsey.com, 2025)(getmonetizely.com, 2025).

### *4.1 Comparative Overview of Foundational AI Pricing Models*

The initial foray into monetizing AI-driven services often involved adapting established pricing strategies from the software-as-a-service (SaaS) and platform-as-a-service (PaaS) domains (Rossi, 2024). These foundational models typically focus on capturing value through various mechanisms, including subscription fees, usage-based charges, or a combination thereof (bcg.com, 2025). Key dimensions for comparing these models in the context of AI include their efficacy in value capture, ability to recover underlying operational costs (especially for compute-intensive AI), competitive positioning within a rapidly evolving market, and their impact on user adoption and engagement (Sharma, 2024).

Historically, software pricing models have ranged from one-time perpetual licenses to recurring subscription fees (ispartnersllc.com, 2025). With the shift towards cloud-based services, usage-based and tiered subscription models gained prominence. Applying these directly to AI, however, revealed several inherent limitations. AI models, particularly generative ones, do not consume resources in a linear or easily predictable fashion (Madathala et al., 2022). The "cost" of an AI interaction can vary significantly based on the complexity of the prompt, the length of the output, the model's internal inference path, and even the specific architecture of the underlying hardware (Ranjan et al., 2025). Furthermore, the value derived from an AI interaction is often subjective and context-dependent, making uniform pricing challenging (Mauri et al., 2019).

Broadly, foundational AI pricing models can be categorized as follows:

1. **Subscription-Based Models:** These charge a fixed recurring fee (e.g., monthly or annually) for access to an AI service or platform. Subscriptions often come with different tiers, offering varying levels of features, usage limits, or dedicated support (theai-

innovator.com, 2025). The primary advantage for providers is predictable revenue, while users benefit from predictable costs (getmonetizely.com, 2025). However, they can lead to under-utilization by low-usage customers or over-utilization by high-usage customers, potentially misaligning cost with value (bcg.com, 2025).

2. **Pay-per-Use Models:** Also known as consumption-based or utility pricing, these models charge users based on their actual consumption of AI resources (Mei et al., 2022). Common metrics include tokens processed (for LLMs), API calls made, compute hours used, or data processed (getmonetizely.com, 2025). This model aligns costs more closely with usage, offering fairness for variable workloads. However, it can introduce cost unpredictability for users, especially for complex or exploratory AI tasks (Gupta, 2025).
3. **Value-Based Models:** This approach prices AI services based on the quantifiable value they deliver to the customer, such as increased revenue, reduced costs, or improved efficiency (Sharma, 2024). While theoretically ideal for aligning incentives, value-based pricing is often complex to implement due to the difficulty in precisely attributing business outcomes solely to the AI service (bcg.com, 2025). It typically requires deep customer integration and robust measurement frameworks.
4. **Freemium Models:** These offer a basic version of the AI service for free, with premium features, higher usage limits, or advanced capabilities available through paid subscriptions or pay-per-use upgrades (getmonetizely.com, 2025). Freemium models are excellent for user acquisition and product adoption, allowing users to experience the value proposition before committing financially (bcg.com, 2025). The challenge lies in converting free users to paying customers and managing the costs associated with the free tier (mckinsey.com, 2025).

The initial challenges in adapting these models for AI's unique characteristics are profound. Unlike traditional software, where a feature's cost is relatively static, the cost of generating an AI output can fluctuate based on the input prompt's complexity, the length

of the desired response, and the underlying computational load (Ranjan et al., 2025). For instance, a simple query might consume minimal tokens and compute, while a complex prompt requiring extensive reasoning or tool use within an agentic system could incur significantly higher costs (Sanabria & Vecino, 2024). Furthermore, the quality and utility of AI outputs can be non-deterministic, meaning the same input might yield slightly different results, making it difficult to price based solely on output quantity (Ranjan et al., 2025). The distinction between the costs of training large models (a significant upfront investment) and the costs of inference (per-query usage) also presents a unique challenge for cost recovery (Sharma, 2024). As AI capabilities evolve rapidly, pricing structures must also remain flexible, reflecting new features, improved efficiency, and changing market dynamics (Gupta, 2025). This necessitates a continuous re-evaluation of how value is captured and how costs are allocated across the AI ecosystem (mckinsey.com, 2025).

### *4.2 Detailed Examination of Common AI Pricing Models*

The evolution of AI services has led to a refinement of these foundational models, with specific adaptations emerging to address the unique attributes of AI technologies. This section delves into the most prevalent pricing models currently employed, highlighting their mechanisms, benefits, drawbacks, and specific implications for AI agentic systems.

**4.2.1 Token-Based Pricing (Pay-per-use for LLMs)** Token-based pricing has become the de facto standard for large language models (LLMs) and generative AI services (Gupta, 2025). In this model, users are charged based on the number of "tokens" processed, where a token is a fundamental unit of text, roughly equivalent to 4 characters or three-quarters of a word in English (Adetayo et al., 2024). Both input (prompt) and output (response) tokens are typically counted and billed separately, often at different rates, with output tokens frequently being more expensive due to the computational resources required for generation (Adetayo et al., 2024). The rationale behind token-based pricing is its direct correlation

with resource consumption: longer inputs and outputs require more computational effort and memory, hence incurring higher costs (Ranjan et al., 2025).

**Table 2: Example Token-Based Pricing for Leading LLMs (Per 1 Million Tokens)**

| Provider/Model | Input Cost (USD) | Output Cost (USD) | Context Window (Tokens) | Key Differentiator |
|---|---|---|---|---|
| **OpenAI GPT-4o** | $5.00 | $15.00 | 128,000 | Multimodal, high capability |
| **OpenAI GPT-3.5T** | $0.50 | $1.50 | 16,385 | Cost-effective, fast |
| **Anthropic Claude 3 Opus** | $15.00 | $75.00 | 200,000 | Strong reasoning, large context |
| **Anthropic Claude 3 Haiku** | $0.25 | $1.25 | 200,000 | Speed, low cost, large context |
| **Google Gemini 1.5 Pro** | $3.50 | $10.50 | 1,000,000 | Massive context, multimodal |

*Note: Prices are illustrative and subject to change. "T" denotes Turbo. Costs are for general use cases and may vary for fine-tuned models or specific enterprise agreements.*

**Advantages:** One of the primary advantages of token-based pricing is its **granularity** (getmonetizely.com, 2025). It allows for a highly precise measure of usage, ensuring that users only pay for what they consume. This offers a degree of **cost transparency** for straightforward use cases where token counts are easily estimable (Gupta, 2025). For providers, it ensures that infrastructure costs, which scale with computational load, are directly recovered (Sharma, 2024). The model is inherently **scalable with usage**, making

it suitable for applications ranging from small-scale prototyping to large-scale enterprise deployments (mckinsey.com, 2025). Furthermore, it encourages developers to optimize their prompts and responses for brevity and efficiency, indirectly promoting more resource-efficient AI interactions (Ranjan et al., 2025).

**Disadvantages:** Despite its advantages, token-based pricing presents significant challenges. For end-users and even developers, **estimating token counts** can be notoriously difficult and unintuitive (Adetayo et al., 2024). The conversion of natural language into tokens is not always straightforward, leading to **cost unpredictability**, especially for complex or iterative tasks (Gupta, 2025). Users might struggle to forecast their monthly expenditures, which can be a barrier to budget planning and widespread adoption (bcg.com, 2025). There is also a potential for "token stuffing," where users might inadvertently or intentionally include unnecessary information in prompts, leading to higher costs without proportional value (Ranjan et al., 2025). More critically, token-based pricing is not universally suitable for all AI tasks. For instance, image generation, video processing, or complex multi-modal AI interactions do not easily map to a token metric (Sharma, 2024). In agentic workflows, where an AI might perform multiple internal reasoning steps or tool calls before generating a final response, the cumulative token count can rapidly escalate, making the effective cost of a single "agent task" highly variable and opaque (Sanabria & Vecino, 2024). This variability makes it challenging to predict long-term operational costs for businesses integrating agentic systems (Ranjan et al., 2025).

**Real-world examples:** Leading AI providers predominantly utilize token-based pricing. **OpenAI** (e.g., GPT-3.5, GPT-4) is the most prominent example (Adetayo et al., 2024). Their pricing differentiates not only between input and output tokens but also across different model variants (e.g., `gpt-4-turbo` typically has a different rate than `gpt-3.5-turbo`). They also factor in the context window size, with models supporting larger context windows (e.g., 128k tokens) often having different pricing tiers (Adetayo et al., 2024). **Anthropic's Claude** models (e.g., Claude 3 Opus, Sonnet, Haiku) similarly employ token-

based pricing, often emphasizing their larger context windows and providing competitive rates, particularly for input tokens, to encourage longer prompts (Adetayo et al., 2024). **Google's Gemini** models, offered through Vertex AI, also follow a token-based structure, with varying prices for different Gemini model sizes and capabilities (Ravulavaru, 2018). A direct comparison reveals that while all three use tokens, their specific rates, context window limits, and the differentiation between input/output costs vary, reflecting their competitive strategies and underlying cost structures (Adetayo et al., 2024). For example, one provider might offer a lower input token cost to encourage detailed prompts, while another might offer a more balanced input/output ratio (Adetayo et al., 2024).

**Impact on Agentic Systems:** For AI agentic systems, token costs represent a significant operational consideration. Agents, by their nature, often engage in multi-step reasoning, internal monologues, tool usage, and iterative refinement (Ranjan et al., 2025). Each of these internal steps, if processed by an LLM, consumes tokens. A single user query to an agent might translate into dozens or hundreds of internal LLM calls, accumulating token costs rapidly (Sanabria & Vecino, 2024). For instance, an agent tasked with booking a flight might first query a flight database, then a hotel database, then a calendar, and finally synthesize the information, with each query and internal reasoning step generating tokens (Ranjan et2 al., 2025). This accumulation makes it challenging to price an "agent task" effectively based solely on the initial input and final output tokens. The cost of "thinking" or "reasoning" within an agent is often hidden from the user but directly impacts the provider's expenses. This model incentivizes the development of more efficient agents that minimize token usage, but also creates a barrier to complex, exploratory, or long-running agentic applications where extensive internal processing is inherent (Sanabria & Vecino, 2024).

**4.2.2 API Call-Based Pricing (Transaction-based)** API call-based pricing, also known as transaction-based pricing, charges users for each request or interaction made with an AI service's API (getmonetizely.com, 2025). This model is a direct carry-over from

traditional web services and microservices architectures, where each discrete function call is metered (Rossi, 2024). Providers often implement tiers based on the volume of calls, with lower per-call rates for higher volumes, or differentiate pricing based on the complexity of the API endpoint (bcg.com, 2025).

**Description:** In this model, the unit of charge is a single API request, regardless of the complexity or computational resources consumed by that specific request (though some providers might offer different endpoints with different per-call prices) (Sharma, 2024). For example, a sentiment analysis API might charge $0.01 per call, irrespective of the length of the text analyzed, up to a certain character limit (getmonetizely.com, 2025). This model is particularly prevalent for task-specific AI services that perform a well-defined function, such as image recognition, speech-to-text transcription, or specific data classification (Ravulavaru, 2018).

**Advantages:** The primary advantage of API call-based pricing is its **simplicity and predictability** for fixed-function APIs (getmonetizely.com, 2025). Users can easily understand their costs by simply counting their API requests. This straightforwardness makes **integration easy** into existing applications, as developers only need to track the number of calls made (Rossi, 2024). For providers, it offers a clear metric for monetization and allows for straightforward tiering based on usage volume, making it easy to scale pricing as demand grows (bcg.com, 2025). For discrete AI tasks, it can be more intuitive than token-based pricing, as the "transaction" is a clear unit of work (Sharma, 2024).

**Disadvantages:** However, API call-based pricing has significant drawbacks, especially for more advanced AI. It often **ignores the computational intensity per call** (Ranjan et al., 2025). A simple image classification might cost the same as a complex object detection task, even if the latter consumes significantly more GPU cycles (Sharma, 2024). This lack of granularity can lead to either under-pricing computationally expensive operations or over-pricing simple ones, misaligning cost with value and resource consumption (bcg.com, 2025). It is **less flexible for variable workloads** where the "cost" of a single

transaction can fluctuate greatly (Gupta, 2025). Furthermore, it can **incentivize fewer but more complex calls**, as users might try to pack as much functionality as possible into a single request to minimize their transaction count, potentially leading to less modular and harder-to-manage API interactions (Ranjan et al., 2025).

**Real-world examples:** Early AI services and many specialized AI APIs still use this model. Examples include various cloud provider services for computer vision (e.g., Google Cloud Vision API, AWS Rekognition) (Ravulavaru, 2018), natural language processing tasks (e.g., sentiment analysis, entity extraction), and speech processing (e.g., speech-to-text, text-to-speech) (getmonetizely.com, 2025). These services often have distinct API endpoints for different tasks, each with its own per-call pricing (Ravulavaru, 2018). While LLMs have largely moved to token-based models, certain wrapper APIs or highly optimized, fine-tuned models might still offer API call-based pricing for specific, well-defined generative tasks where the input/output complexity is constrained (theaiinnovator.com, 2025).

**Impact on Agentic Systems:** For AI agentic systems, API call-based pricing is suitable for agents that perform **discrete, well-defined tasks** using specific tools or external services (Sanabria & Vecino, 2024). For example, if an agent uses an external weather API, a stock lookup API, or a database query API, charging per API call is a natural fit (Ranjan et al., 2025). However, it becomes **less ideal for dynamic, exploratory agents** where the internal reasoning and sequence of operations are highly variable and unpredictable (Sanabria & Vecino, 2024). If an agent needs to make multiple iterative calls to a generative AI model or a complex internal reasoning engine, simply charging per "agent request" would fail to capture the underlying computational costs (Ranjan et al., 2025). The challenge lies in defining what constitutes a single "call" in a multi-step, multi-tool agentic workflow. This model can be a component of a larger hybrid pricing strategy for agents, particularly for their tool-use capabilities, but it is insufficient as a standalone model for the core generative and reasoning aspects of sophisticated agents (Sanabria & Vecino, 2024).

**4.2.3 Subscription-Based Pricing (Fixed Fee Access)** Subscription-based pricing involves charging a fixed, recurring fee—typically monthly or annually—for access to an AI service, a set of features, or a specific model (getmonetizely.com, 2025). This model is widely adopted across the software industry and has found significant application in AI, particularly for consumer-facing products and dedicated enterprise solutions (bcg.com, 2025). Subscriptions often come in different tiers, each offering a distinct bundle of features, usage allowances, performance levels, or support services (theaiinnovator.com, 2025).

**Description:** Under a subscription model, users pay a predetermined fee at regular intervals (e.g., \$20/month, \$200/year) to gain access to the AI service (getmonetizely.com, 2025). This fee typically grants a certain level of usage, which might include a fixed number of tokens, API calls, or compute hours per billing period (bcg.com, 2025). Once these allowances are exhausted, users may either be charged overage fees (transitioning into a hybrid model) or be required to upgrade their subscription tier (Sharma, 2024). Enterprise subscriptions often involve dedicated instances, service level agreements (SLAs), enhanced security features, and priority support, reflecting a higher value proposition (mckinsey.com, 2025).

**Advantages:** For AI providers, subscription models offer **predictable revenue streams**, which are crucial for long-term planning, investment in R&D, and sustainable growth (Sharma, 2024). This predictability helps stabilize financial operations, especially given the high upfront costs of developing and training advanced AI models (bcg.com, 2025). For users, the primary benefit is **predictable costs** (getmonetizely.com, 2025). They know exactly what their AI expenditures will be each month, simplifying budgeting and financial forecasting. This predictability is particularly attractive for businesses that need to integrate AI into their operational budgets (mckinsey.com, 2025). Subscriptions also tend to **encourage continuous usage** (theaiinnovator.com, 2025); once a user has committed to a subscription, they are incentivized to maximize their value from the service. The **simplicity of billing** and management for both parties further enhances its appeal (bcg.com, 2025).

**Disadvantages:** A significant drawback of subscription pricing is its potential **misalignment with variable usage patterns** (Gupta, 2025). Users with low usage might feel they are overpaying, leading to dissatisfaction and churn, while very high-usage customers might be undercharged, reducing potential revenue for the provider (bcg.com, 2025). This can lead to **under-utilization or over-utilization** of resources, neither of which is optimal for efficiency or fairness (Sharma, 2024). **Difficulty in setting fair tiers** for diverse user needs is another challenge; finding the right balance of features and usage allowances that appeal to a broad customer base without alienating specific segments requires extensive market research and iterative refinement (mckinsey.com, 2025). Furthermore, if an AI model's capabilities evolve rapidly, the static nature of subscription tiers can quickly become outdated, requiring frequent re-evaluation and adjustment (Ranjan et al., 2025).

**Real-world examples:** Many prominent AI services offer subscription models. **GitHub Copilot** is a prime example, providing AI-powered code suggestions for a fixed monthly fee (Adetayo et al., 2024). **OpenAI Plus** offers subscribers access to GPT-4, higher usage limits, and early access to new features for a monthly fee, catering to individual power users (Adetayo et al., 2024). Similarly, **Anthropic Pro** provides enhanced access to Claude models, including higher rate limits and priority access, through a subscription (Adetayo et al., 2024). These examples illustrate how subscriptions are used to differentiate access and features, providing a baseline level of service with the option for more premium capabilities (theaiinnovator.com, 2025). Enterprise-level subscriptions, such as those offered by Microsoft for Copilot or Google for Vertex AI, often involve custom contracts that bundle dedicated compute, enhanced security, and specialized support with a fixed recurring cost (stocktitan.net, 2025)(Ravulavaru, 2018).

**Impact on Agentic Systems:** For AI agentic systems, subscription models can provide a **stable baseline for agent development and testing** (Ranjan et al., 2025). Developers can subscribe to a certain tier, gaining predictable access to models and tools, which facilitates iterative development without the constant worry of fluctuating token costs

(Sanabria & Vecino, 2024). Enterprise subscriptions are particularly relevant for deploying **dedicated instances of agentic systems** within an organization, ensuring consistent performance, security, and compliance (mckinsey.com, 2025). For consumer-facing agents (e.g., personal assistants, intelligent chatbots), a subscription model can make the service accessible and budget-friendly for end-users, especially if the underlying token costs are abstracted away and managed by the provider (Sanabria & Vecino, 2024). However, the challenge remains in aligning the fixed subscription fee with the variable and often high computational demands of complex agentic workflows. If an agent performs highly intensive tasks that exceed the subscription's implicit usage allowances, the provider might incur significant losses or be forced to implement complex overage charges, undermining the simplicity of the subscription model (Ranjan et al., 2025).

**4.2.4 Value-Based Pricing (Outcome-oriented)** Value-based pricing is a sophisticated strategy where the price of an AI service is determined by the perceived or quantifiable value it delivers to the customer, rather than solely by its cost of production or usage (Sharma, 2024). This model aligns the financial incentives of the provider directly with the business outcomes of the customer, fostering a partnership approach (bcg.com, 2025). It is particularly relevant for high-impact enterprise AI applications where the return on investment (ROI) can be clearly demonstrated (mckinsey.com, 2025).

**Description:** In a value-based pricing model, the price is set based on metrics such as cost savings achieved, revenue generated, efficiency gains realized, or risk mitigated through the use of the AI service (Sharma, 2024). For example, an AI system that optimizes logistics might be priced as a percentage of the fuel costs saved, or an AI-driven marketing platform might charge a percentage of the incremental sales generated (Gupta, 2025). This approach requires a deep understanding of the customer's business processes and a robust methodology for measuring the impact of the AI solution (bcg.com, 2025). It often involves custom

contracts, performance-based agreements, and sometimes even revenue-sharing arrangements (mckinsey.com, 2025).

**Advantages:** The most significant advantage of value-based pricing is its ability to **align incentives** between the AI provider and the customer (Sharma, 2024). Both parties are motivated by the successful delivery of measurable business outcomes. This model allows providers to **capture higher value** when their AI solutions deliver substantial benefits, potentially leading to significantly higher revenue compared to usage-based or subscription models (bcg.com, 2025). It is particularly **suitable for high-impact enterprise applications** where the AI directly contributes to critical business functions and where the ROI can be clearly articulated and measured (mckinsey.com, 2025). By focusing on outcomes, it shifts the conversation from technical features to business impact, which resonates strongly with executive decision-makers (Sharma, 2024).

**Disadvantages:** Despite its theoretical appeal, value-based pricing is often **difficult to implement** in practice (bcg.com, 2025). The primary challenge lies in **quantifying the value delivered** (Sharma, 2024). Isolating the specific impact of the AI solution from other business initiatives, market factors, or operational changes can be complex and contentious. It requires sophisticated analytics, baseline measurements, and often A/B testing or controlled experiments (mckinsey.com, 2025). This model also necessitates a **strong customer relationship** built on trust and transparency, as both parties must agree on the metrics for success and the methods for measurement (bcg.com, 2025). Furthermore, it introduces **risk for the provider if the value isn't realized** (Sharma, 2024). If the AI solution underperforms or external factors prevent the customer from achieving the expected benefits, the provider's revenue can be significantly impacted. This complexity often leads to longer sales cycles and more intricate contract negotiations (mckinsey.com, 2025).

**Real-world examples:** Value-based pricing is less common for off-the-shelf AI products but is frequently employed for **custom enterprise AI solutions** and AI consulting services (Sharma, 2024). For instance, a specialized AI platform designed to optimize supply

chains for a large manufacturing firm might be priced based on a percentage of the inventory cost reductions or logistics efficiency gains (mckinsey.com, 2025). AI-driven fraud detection systems could be priced as a fraction of the fraud losses prevented (bcg.com, 2025). In the realm of AI consulting, firms might offer performance-based fees, where a portion of their payment is contingent on the successful implementation and measurable impact of the AI strategy (Sharma, 2024). While not explicitly value-based, the trend towards "AI-as-a-Service" for specific vertical applications (e.g., AI for drug discovery, AI for legal document review) often implicitly incorporates value by targeting high-value problems where clear ROI can be demonstrated (theaiinnovator.com, 2025).

**Impact on Agentic Systems:** Value-based pricing holds potentially the most effective monetization strategy for **highly specialized, mission-critical AI agents** that deliver measurable business outcomes (Sanabria & Vecino, 2024). Consider an agent designed to automate complex financial analysis, detect market anomalies, and execute trades, generating significant returns (Ranjan et al., 2025). Pricing this agent as a percentage of the alpha generated or the operational costs saved would directly align its cost with its performance (Sharma, 2024). Similarly, an agent optimizing manufacturing processes could be priced based on reduced waste or increased throughput (mckinsey.com, 2025). This model encourages the development of highly reliable and effective agents, as the provider's revenue is directly tied to the agent's success (Sanabria & Vecino, 2024). However, for this to work, the agent's contribution must be clearly distinguishable and measurable (Ranjan et al., 2025). For general-purpose agents or those performing exploratory tasks, quantifying value becomes much harder, making value-based pricing less feasible. It is best suited for agentic systems operating in well-defined business contexts with clear key performance indicators (KPIs) (Sanabria & Vecino, 2024).

**4.2.5 Hybrid and Tiered Models** Recognizing the limitations of single pricing models, most major AI providers have gravitated towards **hybrid and tiered pricing structures**

(bcg.com, 2025). These models combine elements from two or more foundational approaches, aiming to leverage the advantages of each while mitigating their individual drawbacks (mckinsey.com, 2025). The goal is to create a flexible pricing strategy that caters to a diverse user base, from individual developers to large enterprises, and accommodates varying usage patterns and value propositions (Sharma, 2024).

**Table 3: Common Hybrid Pricing Model Components**

| Component Type | Description | Primary Benefit | Example |
|---|---|---|---|
| **Base Subscription** | Fixed recurring fee for platform access and core features. | Predictable revenue for provider; stable cost for user. | $20/month for basic access. |
| **Usage-Based Overage** | Charges for usage beyond a base allowance (e.g., tokens, API calls). | Fair for variable workloads; captures high-usage value. | $0.001/1K tokens over 1M limit. |
| **Feature Tiers** | Different subscription levels unlock specific advanced features or capabilities. | Value differentiation; tiered user segments. | "Pro" tier includes advanced data analysis tools. |
| **Performance/Outcome Add-ons** | Premium features or price adjustments based on achieved results. | Aligns incentives; captures high-value impact. | 5% of cost savings from AI optimization. |
| **Dedicated Capacity** | Fixed fee for guaranteed resources (e.g., dedicated GPUs). | High reliability & performance; enterprise focus. | $1,000/month for dedicated GPU instance. |

| Component Type | Description | Primary Benefit | Example |
|---|---|---|---|
| **Customization Fee** | Upfront cost for fine-tuning models or bespoke development. | Tailored solutions; specific client needs. | $10,000 for custom model training. |

*Note: Hybrid models combine these components in various configurations to optimize for specific market needs and strategic goals.*

**Description:** Hybrid models often start with a **subscription base** that provides a fixed allowance of usage (e.g., tokens, API calls, compute hours) for a recurring fee (getmonetizely.com, 2025). Beyond this allowance, users are then charged **usage-based overage fees** (bcg.com, 2025). For example, a monthly subscription might include 1 million tokens, with additional tokens billed at a per-token rate. Another common hybrid approach involves **tiered subscriptions** (theaiinnovator.com, 2025). Different tiers (e.g., Free, Basic, Pro, Enterprise) offer progressively more features, higher usage limits, better performance, or dedicated support, each with a corresponding fixed monthly fee. Within these tiers, there might still be usage-based components, such as higher per-token rates for premium models or additional charges for specific advanced features (Sharma, 2024). Some models also incorporate a **freemium layer** to attract new users, allowing them to experience basic functionality before committing to a paid tier (getmonetizely.com, 2025).

**Advantages:** The primary advantage of hybrid and tiered models is their **flexibility** (bcg.com, 2025). They can be designed to **cater to diverse user segments**, from casual users to power users and large organizations, each with different needs and budget constraints (mckinsey.com, 2025). By combining fixed and variable components, these models **balance predictability and usage-based fairness** (Sharma, 2024). Users gain some cost predictability from the subscription, while providers can still capture value from high-usage scenarios through overage charges. This approach allows providers to offer a comprehen-

sive range of services, from basic access to highly specialized enterprise solutions, under a unified yet adaptable pricing framework (theaiinnovator.com, 2025). It also facilitates easier upgrades and downgrades for users as their needs evolve, promoting customer retention (bcg.com, 2025).

**Disadvantages:** The main drawback of hybrid and tiered models is their **increased complexity in pricing structure and billing** (Gupta, 2025). Users may find it challenging to understand all the nuances of different tiers, allowances, and overage rates, leading to **potential for user confusion** and frustration (getmonetizely.com, 2025). This complexity can also make it harder for businesses to accurately forecast their AI expenditures (Ranjan et al., 2025). For providers, managing intricate billing systems that account for multiple variables (subscriptions, tokens, API calls, features) can be operationally intensive (Sharma, 2024). Furthermore, designing the optimal tiers and pricing points requires continuous market analysis and iteration to ensure competitiveness and profitability (bcg.com, 2025).

**Real-world examples:** Most major AI providers now employ some form of hybrid pricing. **OpenAI** offers a free API tier with limited usage, a paid API tier with token-based pricing, and a "Plus" subscription for direct access to their most capable models (Adetayo et al., 2024). They also have enterprise offerings that are more customized. **Anthropic** similarly provides different Claude models with varying token prices and offers a Pro subscription for enhanced access (Adetayo et al., 2024). **Google Cloud's Vertex AI** offers a complex array of pricing, including pay-per-use for specific models (e.g., Gemini, PaLM 2), charges for custom model training, and subscription-like agreements for enterprise customers using managed services (Ravulavaru, 2018). These examples demonstrate the industry's recognition that a single, monolithic pricing model is insufficient to capture the full value and address the diverse needs of the AI market (mckinsey.com, 2025).

*4.3 Challenges and Considerations for AI Agentic Systems Pricing*

The unique characteristics of AI agentic systems introduce a new layer of complexity to pricing that goes beyond the considerations for standalone LLMs or discrete AI services (Sanabria & Vecino, 2024). These systems, characterized by their autonomy, goal-directed behavior, and ability to interact with environments and tools, challenge traditional monetization frameworks (Ranjan et al., 2025).

One of the foremost challenges is the **non-deterministic nature of AI agent outputs** (Ranjan et al., 2025). Unlike a traditional software function that reliably produces the same output for the same input, an AI agent's actions and generated content can vary in quality, completeness, and even correctness across different runs, even with identical prompts (Sanabria & Vecino, 2024). How should providers price outputs that vary in quality or completeness? Should a "failed" or suboptimal agentic attempt be charged at the same rate as a successful one? This variability makes it difficult to establish a consistent value proposition and, consequently, a consistent price (Sharma, 2024). Users might be reluctant to pay full price for outcomes that are not guaranteed to be perfect or even adequate, leading to demands for outcome-based pricing that is hard to implement (bcg.com, 2025).

The **complexity of agentic workflows** further complicates pricing (Sanabria & Vecino, 2024). An AI agent often engages in multi-step tasks, involving internal reasoning, planning, tool use, external API calls, and iterative refinement (Ranjan et al., 2025). For instance, an agent tasked with drafting a marketing campaign might first research market trends, then analyze competitor strategies, then generate several campaign ideas, then refine them based on feedback, and finally draft the campaign copy. Each of these steps might involve multiple LLM calls, database queries, or interactions with other specialized AI models (Sanabria & Vecino, 2024). Pricing for such intricate, multi-step processes becomes highly challenging. Should users be charged for every internal "thought" or tool call made by the agent, even if these do not directly result in a visible output? How do providers account for the internal "thinking" processes (e.g., chain-of-thought, tree-of-thought reasoning) that

consume tokens but are not explicit API calls (Ranjan et al., 2025)? The cost accumulation in these scenarios can be substantial and opaque to the end-user, making cost prediction and budgeting extremely difficult (Sanabria & Vecino, 2024).

**Value attribution** is another critical consideration (Sharma, 2024). When an agent orchestrates multiple models, tools, and external services to achieve a goal, how is the value attributed to each component, and how should this be reflected in the pricing? If an agent uses a proprietary LLM, a third-party weather API, and a custom database, should the user be charged separately for each component, or should the agent provider offer an abstracted, bundled price? (Sanabria & Vecino, 2024). This is particularly complex when the agent itself adds significant orchestrational and reasoning value on top of the constituent parts (Ranjan et al., 2025). Differentiating the value of the agent's intelligence from the value of the tools it employs is a nuanced task (Sharma, 2024).

The **cost of failure or hallucination** is a pragmatic concern (Ranjan et al., 2025). AI agents, especially those based on LLMs, are prone to hallucinations or may fail to achieve their objectives due to misinterpretation, insufficient context, or limitations of their underlying models (Sanabria & Vecino, 2024). Should users be charged for "failed" agentic attempts or for outputs that contain factual inaccuracies or are otherwise unusable? (Sharma, 2024). From a user perspective, paying for a failed outcome is undesirable and undermines the value proposition. Providers, however, incur computational costs even for failed attempts. This raises ethical and practical questions about fair pricing and responsibility (legalinstruments.oecd.org, 2025). Implementing a refund or credit system for failed tasks could mitigate user dissatisfaction but adds another layer of billing complexity (Ranjan et al., 2025).

**Data privacy and security** considerations also influence pricing (datainnovation.org, 2025). Enterprises, in particular, are willing to pay a premium for enhanced security features, compliance with industry regulations, and guarantees regarding data privacy (Sharma, 2024). This might translate into higher costs for dedicated instances, on-premise deployments, or AI services that offer robust data governance and access

controls (mckinsey.com, 2025). The computational and operational overhead of ensuring data isolation and security in multi-tenant AI environments can be significant, necessitating higher price points for such specialized services (ispartnersllc.com, 2025).

Underlying **infrastructure costs** are a fundamental driver of AI pricing (Sharma, 2024). The cost of high-performance computing (HPC) hardware, particularly GPUs, is substantial (Ranjan et al., 2025). Training large foundation models requires immense compute resources, and even inference, especially for large context windows or complex agentic tasks, can be resource-intensive (Sanabria & Vecino, 2024). Storage costs for models and data, network bandwidth, and the operational expenses of maintaining vast data centers all contribute to the overall cost base (Sharma, 2024). Pricing models must adequately recover these costs while remaining competitive. The continuous innovation in hardware and optimization techniques can lead to fluctuating costs, requiring flexible pricing strategies (Ranjan et al., 2025).

Finally, **ethical considerations** play a role, particularly for AI agents deployed in critical applications (legalinstruments.oecd.org, 2025). Ensuring fair and equitable pricing for AI services that impact areas like healthcare, education, or legal services is paramount (brookings.edu, 2025). Overly expensive AI could exacerbate digital divides or create barriers to access for essential services (reports.weforum.org, 2025). Providers must balance profitability with societal responsibility, especially as AI agents become more integrated into daily life (legalinstruments.oecd.org, 2025). The development of regulatory frameworks around AI also has implications for pricing, as compliance costs might need to be factored in (datainnovation.org, 2025)(oecd.org, 2025).

These challenges underscore the need for innovative and adaptive pricing models that can effectively capture the value of AI agentic systems while addressing their inherent complexities, ensuring cost recovery for providers, and maintaining fairness and predictability for users (Sanabria & Vecino, 2024)(Gupta, 2025).

*4.4 Real-World Case Studies: OpenAI, Anthropic, and Google's Pricing Strategies*

The leading developers of foundational AI models and agentic capabilities have adopted distinct, yet often converging, pricing strategies. Examining the approaches of OpenAI, Anthropic, and Google provides valuable insights into the current landscape and future directions of AI monetization (Adetayo et al., 2024)(Ravulavaru, 2018).

**OpenAI** OpenAI, a pioneer in generative AI, has seen a significant evolution in its pricing strategy since the early days of GPT-3 (Adetayo et al., 2024). Initially, access to GPT-3 was highly restricted and costly, reflecting the immense research and computational investment required to develop such a powerful model (Sharma, 2024). This early pricing aimed to recover costs and fund further research, making it accessible primarily to enterprises and research institutions (mckinsey.com, 2025).

With the release of subsequent models like GPT-3.5 and GPT-4, OpenAI has refined its approach, primarily focusing on **token-based pricing** (Adetayo et al., 2024). Their strategy differentiates pricing based on: 1. **Model Variant:** Different models (e.g., `gpt-3.5-turbo`, `gpt-4-turbo`, `gpt-4o`) come with varying price points per 1,000 tokens, reflecting their capabilities, performance, and underlying resource consumption (Adetayo et al., 2024). More advanced and capable models generally command higher prices. 2. **Input vs. Output Tokens:** OpenAI typically charges different rates for input (prompt) tokens and output (completion) tokens, with output tokens often being more expensive due to the generative computational load (Adetayo et al., 2024). 3. **Context Window Size:** Models offering larger context windows (e.g., 128k tokens) might have different pricing structures, acknowledging the increased memory and processing required to handle extensive inputs (Adetayo et al., 2024). 4. **API vs. Consumer Product:** OpenAI maintains separate pricing for its API services (for developers and enterprises) and its consumer-facing product (ChatGPT Plus subscription) (Adetayo et al., 2024). The ChatGPT Plus subscription offers a fixed monthly fee for enhanced access to the most capable models, higher usage limits,

and early access to new features, abstracting away the token-based complexity for individual users (Adetayo et al., 2024).

The introduction of the **Assistants API** and capabilities for **custom model training** (fine-tuning) has introduced new pricing dimensions (Ranjan et al., 2025). The Assistants API charges for tokens, but also for "tool use" and "retrieval" actions performed by the assistant, which are abstracted internal steps (Sanabria & Vecino, 2024). Fine-tuning services involve upfront costs for training custom models, followed by usage-based charges for inference on these specialized models (Sharma, 2024).

OpenAI's strategic rationale behind these choices appears multi-faceted (mckinsey.com, 2025). Firstly, the token-based model allows for **democratizing access** to powerful AI by making it available on a granular, pay-as-you-go basis, fostering innovation across a broad developer ecosystem (theaiinnovator.com, 2025). Secondly, by offering tiered models and subscriptions, they aim to **capture enterprise value** by providing scalable, high-performance solutions for business-critical applications (bcg.com, 2025). Their pricing evolution reflects a balance between cost recovery for their extensive R&D and compute investments, and market penetration to establish their models as industry standards (Sharma, 2024).

**Anthropic (Claude)** Anthropic, a key competitor in the LLM space, particularly with its Claude series of models, has adopted a pricing strategy that shares similarities with OpenAI but also features distinct differentiators (Adetayo et al., 2024). Anthropic's pricing is also primarily **token-based**, distinguishing between input and output tokens (Adetayo et al., 2024).

Key aspects of Anthropic's pricing include: 1. **Emphasis on Context Windows:** Anthropic has consistently highlighted its models' ability to handle exceptionally large context windows (e.g., 200k tokens), often offering competitive pricing for these extended capabilities (Adetayo et al., 2024). This caters to use cases requiring extensive document analysis

or long-running conversations (Ranjan et al., 2025). 2. **Tiered Model Offerings:** Similar to OpenAI, Anthropic offers a range of models (e.g., Claude 3 Opus, Sonnet, Haiku) with varying levels of intelligence, speed, and cost (Adetayo et al., 2024). "Haiku" is positioned as a fast, cost-effective option, while "Opus" represents their most capable and expensive model (Adetayo et al., 2024). 3. **Input vs. Output Token Cost Structure:** Anthropic often features a significant difference between input and output token costs, with input tokens being considerably cheaper (Adetayo et al., 2024). This encourages users to provide more detailed prompts and larger contexts without incurring prohibitive costs for the input phase, shifting the cost burden more towards the model's generation effort (Ranjan et al., 2025). 4. **Pro Subscription:** Anthropic also offers a "Claude Pro" subscription for individual power users, providing higher rate limits, priority access during peak times, and early access to new features for a fixed monthly fee (Adetayo et al., 2024).

Anthropic's strategic positioning often emphasizes **enterprise-grade safety and responsible AI development** (legalinstruments.oecd.org, 2025). While not explicitly reflected in their token pricing, this focus might influence the perceived value for enterprise customers, potentially supporting a future move towards more value-based components in their bespoke enterprise offerings (mckinsey.com, 2025). Their competitive pricing, particularly for input tokens and large context windows, is a direct strategy to differentiate themselves from OpenAI and attract users with specific needs for extensive text processing (Adetayo et al., 2024).

**Google (Gemini, Vertex AI)** Google's AI monetization strategy is deeply integrated into its broader cloud ecosystem, primarily through **Google Cloud's Vertex AI platform** (Ravulavaru, 2018). This approach leverages Google's existing relationships with enterprise cloud customers and offers a comprehensive suite of AI/ML services beyond just LLMs (Ravulavaru, 2018).

Google's pricing for its LLMs (e.g., Gemini, PaLM 2) and other AI services features: 1. **Token-Based Pricing for LLMs:** Like its competitors, Google charges for tokens processed by its Gemini and PaLM 2 models, with varying rates for different model sizes and capabilities (Ravulavaru, 2018). They also differentiate between input and output tokens. 2. **Integration with Cloud Services:** Pricing is often intertwined with other Google Cloud services (Ravulavaru, 2018). For instance, using Gemini on Vertex AI might incur costs for compute resources, data storage, and network egress, in addition to the LLM token charges (Madathala et al., 2022). This allows Google to offer a complete AI development and deployment stack (Ravulavaru, 2018). 3. **Broad Portfolio of AI Services:** Beyond generative AI, Google offers pricing for a wide array of specialized AI services, including computer vision (Vision AI), speech processing (Speech-to-Text, Text-to-Speech), translation (Translation AI), and custom ML model training (Vertex AI Workbench) (Ravulavaru, 2018). These services often employ API call-based pricing or resource-based pricing (e.g., per hour for custom training). 4. **Enterprise Focus:** Google heavily emphasizes enterprise solutions, offering custom model training, managed services, and dedicated support (Ravulavaru, 2018). These often involve custom contracts, volume discounts, and service level agreements (SLAs), reflecting a more value-based approach for large clients (mckinsey.com, 2025).

Google's strategic rationale is to **leverage its existing cloud customer base** and provide a **comprehensive, end-to-end AI stack** (Ravulavaru, 2018). By integrating AI services deeply into Vertex AI, they aim to be the preferred platform for enterprises looking to build, deploy, and manage their AI applications (mckinsey.com, 2025). Their pricing strategy supports this by offering flexibility, scalability, and a wide range of options, from raw model access to fully managed AI solutions (Ravulavaru, 2018). The ability to fine-tune models and deploy them on dedicated infrastructure within the Google Cloud ecosystem also provides a pathway for higher-value, customized AI solutions (Madathala et al., 2022).

**Other Notable Players (briefly)** While OpenAI, Anthropic, and Google are dominant, other players contribute to the diverse pricing landscape: * **Microsoft Azure AI:** Microsoft integrates OpenAI models (and its own specialized models) into Azure AI Studio and Azure OpenAI Service (stocktitan.net, 2025). Their pricing is generally token-based, but also includes charges for managed services, dedicated capacity, and integration with other Azure products (mckinsey.com, 2025). The **Microsoft Copilot** offerings (e.g., Microsoft 365 Copilot) are typically subscription-based, integrated into existing enterprise software licenses, reflecting a value-added feature rather than a standalone AI service (Adetayo et al., 2024). * **AWS Bedrock:** Amazon's Bedrock provides access to foundation models from various providers (including AI21 Labs, Anthropic, Cohere, Stability AI) as well as Amazon's own models (e.g., Titan) (mckinsey.com, 2025). Its pricing is primarily token-based, with charges for inference and optional charges for custom model fine-tuning (Sharma, 2024). AWS leverages its extensive cloud infrastructure to offer scalable and secure access to these models (mckinsey.com, 2025). * **Meta Llama (Open-Source Implications):** Meta's strategy with models like Llama 2 and Llama 3 is to make them openly available (with commercial licenses for larger enterprises) (theregister.com, 2025). While Meta itself doesn't directly charge for Llama model usage, its open-source nature has significant implications for pricing (Sharma, 2024). It drives down the baseline cost of running LLMs, creating competitive pressure on proprietary models and fostering innovation by allowing companies to host and fine-tune models themselves, incurring only infrastructure costs (theregister.com, 2025). This forces proprietary model providers to justify their higher prices with superior performance, unique features, or enhanced services (mckinsey.com, 2025).

In summary, the AI pricing landscape is dynamic, characterized by a dominant token-based model for LLMs, evolving hybrid approaches, and a strong push towards enterprise solutions that integrate AI into broader cloud ecosystems. Competition and the rapid pace of innovation continue to shape these strategies (Sharma, 2024)(mckinsey.com, 2025).

*4.5 Hybrid Pricing Approaches for Future AI Agentic Systems*

The preceding analysis underscores a critical insight: no single pricing model is sufficient to capture the multifaceted value and manage the inherent complexities of AI agentic systems (Sanabria & Vecino, 2024)(Sharma, 2024). The non-deterministic nature, intricate multi-step workflows, and varied value propositions of agents necessitate a more nuanced, adaptive, and often **hybrid approach to pricing** (Ranjan et al., 2025).

**Need for Hybridity** The limitations of monolithic pricing models become particularly pronounced when considering the advanced capabilities of AI agents: * **Variable Resource Consumption:** A simple agent query might resolve quickly, while a complex one involving multiple tool calls, iterative reasoning, and external data retrieval can consume significantly more tokens and compute (Sanabria & Vecino, 2024). Pure token-based pricing can lead to unpredictable costs for users, while pure subscription-based pricing might not adequately cover the provider's costs for high-usage agents (Ranjan et al., 2025). * **Diverse User Needs:** Developers building experimental agents have different needs and budget constraints than enterprises deploying mission-critical, high-volume agents (mckinsey.com, 2025). A flexible model must cater to both. * **Value Beyond Raw Output:** The true value of an agent often lies in its orchestration, problem-solving, and decision-making capabilities, not just the raw tokens it generates (Sanabria & Vecino, 2024). Pricing needs to reflect this higher-order value (Sharma, 2024). * **Evolving Capabilities:** As agents become more sophisticated (e.g., learning from feedback, adapting to environments), their cost structure and value proposition will continue to evolve (Ranjan et al., 2025). Pricing must be agile enough to adapt (Gupta, 2025).

Therefore, a blend of different pricing strategies is essential to provide fairness, predictability, and profitability for AI agentic systems (Sharma, 2024)(Sanabria & Vecino, 2024).

**Proposed Hybrid Models** Several hybrid pricing models can be conceptualized for AI agentic systems, each combining elements of the foundational approaches:

1. **Tiered Subscription + Token Overage:** This model starts with a fixed monthly or annual subscription fee that grants access to the agentic platform and includes a baseline allowance of "agent credits" or a certain number of tokens (getmonetizely.com, 2025). Once this allowance is exhausted, additional usage is billed on a per-token or per-agent-step basis at an agreed-upon overage rate (Ranjan et al., 2025).

- *Example:* A "Pro Agent Developer" subscription for $50/month includes 5 million tokens or 1,000 complex agent tasks. Beyond this, additional tokens are billed at $0.001/1,000 tokens.
- *Benefits:* Offers cost predictability for baseline usage, encourages adoption, and allows providers to capture value from high-usage scenarios (bcg.com, 2025).
- *Challenges:* Requires clear definition of "agent credits" or "agent steps" to prevent user confusion (Sanabria & Vecino, 2024).

2. **Outcome-Based + Resource Consumption:** This sophisticated model leverages value-based pricing for successful outcomes, but includes a component for underlying resource consumption to cover the provider's operational costs (Sharma, 2024). The outcome-based fee is only charged upon successful completion of a defined agent task that delivers measurable business value (mckinsey.com, 2025).

- *Example:* An agent successfully generates 10 qualified sales leads, leading to a $100 outcome-based fee. However, the underlying token/compute costs incurred during the agent's operation are also billed at a discounted rate or as a fixed "operational fee" (e.g., $5) to cover infrastructure.
- *Benefits:* Strongest alignment of incentives, high value capture, encourages highly effective agents (bcg.com, 2025).
- *Challenges:* Highly complex to implement, requires robust outcome measurement, and careful negotiation of success metrics (Sharma, 2024).

3. **Feature-Based Tiers + API Call for Tools:** This model defines different subscription tiers that unlock specific agent capabilities or tools (theaiinnovator.com, 2025). For instance, a "Basic Agent" tier might allow access to simple web search and summarization tools, while a "Premium Agent" tier unlocks advanced data analysis, code generation, and external API integration (Ranjan et al., 2025). Within these tiers, specific tool usages (e.g., calls to a proprietary database, an image generation API) could be billed on a per-API-call basis (Sanabria & Vecino, 2024).

- *Example:* A "Standard Agent" subscription ($20/month) allows 1,000 calls to a web search tool. A "Advanced Agent" subscription ($100/month) allows unlimited web search calls and 500 calls to a proprietary CRM API, with additional CRM API calls billed at $0.05 each.
- *Benefits:* Clear differentiation of value, allows users to choose capabilities relevant to their needs, and captures value from specialized tool use (mckinsey.com, 2025).
- *Challenges:* Requires careful design of feature bundles and transparent pricing for external tool integrations (Sharma, 2024).

4. **Agentic Workflow Pricing (Task-based Abstraction):** This model attempts to abstract away the underlying token and API call complexity by charging per "agent task" or "agent session" (Sanabria & Vecino, 2024). The price for a task would be determined by its estimated complexity, the number of tools typically involved, or the average token consumption, rather than raw token counts (Ranjan et al., 2025).

- *Example:* An agent "summarize a document" task might cost $0.50, while an agent "plan a travel itinerary" task might cost $5.00, regardless of the exact tokens consumed in a specific instance.
- *Benefits:* Highly intuitive for users, simplifies cost prediction, and aligns with how users perceive agent value (completing a task) (Sanabria & Vecino, 2024).

- *Challenges:* Requires robust internal cost modeling by the provider, potential for mispricing if actual resource consumption deviates significantly from estimates, and managing variability in task complexity (Ranjan et al., 2025).

**Table 4: Agentic Workflow Pricing - Example Task Costs**

| Agent Task Category | Example Task | Estimated Complexity | Average Token Cost | Tools Used | Proposed Price (USD) |
|---|---|---|---|---|---|
| **Simple Info Retrieval** | Summarize News Article | Low | 10k tokens | Web Search, LLM | $0.25 |
| **Data Analysis** | Analyze Sales Report | Medium | 50k tokens | Spreadsheet Tool, LLM | $1.50 |
| **Content Generation** | Draft Marketing Email | Medium | 30k tokens | LLM, Persona Tool | $0.75 |
| **Complex Planning** | Optimize Supply Chain | High | 200k tokens | Optimization Solver, LLM | $10.00 |
| **Code Generation** | Create Python Script | High | 80k tokens | Code Interpreter, LLM | $2.00 |
| **Customer Support** | Resolve Tech Issue | Medium | 40k tokens | Knowledge Base, LLM | $1.00 |

*Note: Prices are illustrative and would be based on internal cost modeling and market value for each defined agent task. Token counts are approximate.*

5. **Custom Model Pricing + Managed Services:** For large enterprises, this involves upfront costs for developing or fine-tuning custom agentic models, coupled with recurring fees for dedicated infrastructure, managed services, and ongoing support (mckinsey.com, 2025). This is essentially a blend of a project-based fee (for customization) and a subscription/resource-based fee (for hosting and maintenance) (Sharma, 2024).

- *Example:* A company pays $50,000 to fine-tune an agent for their specific internal knowledge base, then pays $2,000/month for a dedicated instance and managed support, plus usage-based charges for high-volume inference.
- *Benefits:* Highly tailored solutions, enhanced security and performance, and deep integration into existing enterprise workflows (mckinsey.com, 2025).
- *Challenges:* High upfront investment, long sales cycles, and requires significant resources from both provider and client (bcg.com, 2025).

**Implementation Challenges of Hybrid Models** While offering significant advantages, hybrid models are not without their implementation challenges (Gupta, 2025). * **Billing Complexity:** Managing intricate billing systems that account for multiple variables (subscriptions, tokens, API calls, features, outcomes) can be operationally intensive and prone to errors (Ranjan et al., 2025). * **User Understanding and Transparency:** The complexity of hybrid structures can lead to user confusion, making it difficult for them to understand their costs and compare offerings across providers (getmonetizely.com, 2025). Clear documentation and intuitive dashboards are crucial (Sharma, 2024). * **Dynamic Pricing for Evolving Models:** As AI models and agent capabilities evolve rapidly, pricing structures need to be continuously updated and adapted, which can be a significant management overhead (Ranjan et al., 2025). * **Balancing Provider Profitability with User Fairness:** Striking the right balance between covering the provider's significant R&D and compute costs, and ensuring fair, predictable pricing for users, is an ongoing challenge (Sharma, 2024).

**Future Trends** The future of AI agentic system pricing is likely to move towards even more sophisticated, adaptive, and personalized models (Gupta, 2025). AI itself may play a role in optimizing pricing, with **AI-driven personalized pricing models** leveraging user behavior, demand patterns, and real-time market dynamics to offer customized rates (Gupta, 2025)(Ramezani et al., 2011). Concepts like **dynamic pricing**, where costs fluctuate based on demand, time of day, or available compute resources, could become more prevalent, similar to how utilities price electricity (Mei et al., 2022)(Schlenther et al., 2025). Furthermore, as agentic systems become more integrated into business processes, there will be a greater emphasis on **value-driven contracts** that directly link AI performance to financial outcomes (Sharma, 2024). The role of real-time market dynamics and demand-side management will become increasingly important in shaping how AI agentic systems are priced and consumed (Mei et al., 2022). The overarching trend will be towards pricing models that abstract away technical complexities for the user, focusing instead on the delivered value and desired outcomes, while ensuring sustainable monetization for the innovation providers (Sanabria & Vecino, 2024)(mckinsey.com, 2025).

# 5. DISCUSSION

The preceding analysis of AI agent pricing models, encompassing both theoretical frameworks and practical case studies, reveals a complex and evolving landscape. This discussion section synthesizes these findings, exploring their broader implications for AI companies, customer adoption, future pricing trends, and offering actionable recommendations for various stakeholders. The integration of artificial intelligence into economic transactions fundamentally reshapes traditional pricing paradigms, moving beyond static cost-plus or competitive strategies towards dynamic, adaptive, and highly personalized approaches (Sharma, 2024). This shift necessitates a re-evaluation of established business models, ethical considerations, and regulatory frameworks, marking a pivotal moment in the digital

economy (Sanabria & Vecino, 2024). The discussion will underscore the transformative potential of AI in optimizing value capture while also highlighting the critical challenges that must be navigated to ensure equitable and sustainable growth.

### *5.1 Implications for AI Companies*

The advent of sophisticated AI agents presents a strategic imperative for AI companies to innovate their pricing strategies. Traditional pricing methods are increasingly insufficient to capture the nuanced value generated by AI-driven services, which often involve complex interactions, dynamic resource allocation, and highly personalized outputs (Gupta, 2025). Companies must move towards value-based pricing, where the cost reflects the measurable benefits delivered to the customer, rather than merely the computational resources consumed or development costs (Sharma, 2024). This requires a deep understanding of customer needs and the quantifiable impact of AI solutions on their operations, productivity, or decision-making processes (Ma, 2024). For instance, an AI agent providing market insights might be priced based on the increase in revenue or cost savings it enables for a client, rather than a fixed subscription fee (Sharma, 2024). This approach aligns the interests of the AI provider with those of the customer, fostering stronger partnerships and demonstrating tangible ROI (Sharma, 2024)(getmonetizely.com, 2025).

Furthermore, the ability to implement dynamic and personalized pricing models becomes a significant source of competitive advantage (Gupta, 2025). AI companies can leverage vast datasets—including user behavior, market conditions, and competitor pricing—to adjust prices in real-time, optimizing for revenue, market share, or customer lifetime value (Gupta, 2025)(Ramezani et al., 2011). This level of granularity allows for micro-segmentation of customers and offers tailored pricing that reflects individual willingness to pay or specific usage patterns (Mei et al., 2022). However, this advanced capability also introduces the challenge of balancing innovation with ethical considerations. The potential for discriminatory pricing, where similar customers receive different prices without transparent justification, can

erode trust and lead to reputational damage (Sharma, 2024)(brookings.edu, 2025). Therefore, AI companies must proactively develop and adhere to ethical guidelines, ensuring that their pricing algorithms are fair, transparent, and explainable (legalinstruments.oecd.org, 2025). This involves implementing robust auditing mechanisms and designing AI systems that can articulate the rationale behind their pricing decisions (Ranjan et al., 2025).

Investment in data infrastructure and specialized AI talent is another critical implication (Ranjan et al., 2025). Effective AI-driven pricing relies on high-quality, comprehensive data, demanding significant investment in data collection, storage, processing, and analytical capabilities (Rossi, 2024)(Ma, 2024). Companies need to build scalable data pipelines and employ advanced analytics tools to derive actionable insights from their data assets (Rossi, 2024). Concurrently, there is a growing demand for data scientists, machine learning engineers, and economists with expertise in algorithmic pricing and behavioral economics (mckinsey.com, 2025). Attracting and retaining such talent is crucial for developing, deploying, and refining sophisticated pricing models (mckinsey.com, 2025)(mckinsey.com, 2025). This human capital investment is as vital as technological investment, as the nuances of economic theory and ethical application cannot be solely automated (Ranjan et al., 2025).

Moreover, AI companies must consider the implications for risk management, encompassing both reputational and regulatory scrutiny. The increasing complexity and autonomy of AI pricing agents mean that errors or biases can propagate rapidly, leading to widespread negative customer sentiment or even legal challenges (wilmerhale.com, 2025). Companies must implement rigorous testing and validation protocols to identify and mitigate such risks before deployment (Ranjan et al., 2025). The regulatory landscape is also rapidly evolving, with governments and international bodies exploring frameworks to govern AI, particularly concerning data privacy, algorithmic fairness, and consumer protection (legalinstruments.oecd.org, 2025)(oecd.ai, 2025). AI companies that fail to anticipate and comply with these emerging regulations risk significant fines, operational restrictions, and damage to their brand (wilmerhale.com, 2025). Proactive engagement with policymakers and industry

consortia to shape responsible AI governance is therefore a strategic imperative (datainnovation.org, 2025). Ultimately, the success of AI companies in this new era will hinge on their ability to not only innovate technologically but also to build trust, ensure fairness, and demonstrate accountability in their pricing practices (Sharma, 2024). This holistic approach will define the leaders in the burgeoning AI agent economy.

### *5.2 Customer Adoption Considerations*

The successful adoption of AI-driven pricing models hinges significantly on customer perception and trust. While dynamic and personalized pricing can offer benefits such as optimized resource allocation and tailored offerings, it also introduces potential psychological barriers (Mauri et al., 2019). Customers are increasingly wary of opaque algorithms that determine prices, leading to concerns about fairness and potential manipulation (brookings.edu, 2025). If customers perceive that prices are arbitrary, discriminatory, or designed to exploit their individual vulnerabilities, trust will erode, leading to resistance and even backlash (Sharma, 2024). Transparency in how AI agents arrive at their pricing decisions is therefore paramount (legalinstruments.oecd.org, 2025). Companies need to communicate the value proposition clearly and, where possible, offer explanations for price variations (Sharma, 2024). For example, explaining that a higher price during peak demand reflects increased service costs or limited availability can be more palatable than a seemingly random price surge (Mauri et al., 2019).

The psychological impact of dynamic pricing is a complex area. Research in behavioral economics suggests that perceived fairness often outweighs objective rationality in consumer decision-making (aeaweb.org, 2025). Customers may tolerate dynamic pricing for commodities like airline tickets or ride-sharing services, where prices are visibly linked to supply and demand (Schlenther et al., 2025). However, for essential services or products where price stability is expected, significant fluctuations driven by AI could lead to frustration and a sense of being unfairly treated (Mauri et al., 2019). This sensitivity is heightened when

personalization leads to different prices for similar customers, which can be interpreted as discriminatory rather than value-optimizing (Gupta, 2025). Companies must therefore carefully segment their markets and understand the psychological thresholds for price variability that different customer groups are willing to accept (Mauri et al., 2019). Over-aggressive dynamic pricing strategies, even if technically optimal for profit maximization, can alienate a significant portion of the customer base, leading to long-term revenue losses (Sharma, 2024).

Furthermore, the balance between personalization and privacy concerns is a delicate tightrope for AI companies (Gupta, 2025). Highly personalized pricing relies on extensive data collection about individual customer preferences, behaviors, and even demographic information (Gupta, 2025). While customers appreciate tailored recommendations and offers, they are also increasingly concerned about how their personal data is collected, used, and protected (brookings.edu, 2025). A perceived breach of privacy or misuse of personal data for pricing purposes can severely undermine customer trust and lead to regulatory penalties (wilmerhale.com, 2025). Companies must implement robust data governance frameworks, adhere to privacy regulations like GDPR and CCPA, and provide customers with clear control over their data (legalinstruments.oecd.org, 2025)(oecd.org, 2025). Offering opt-out options for personalized pricing or providing anonymized data alternatives can help mitigate these concerns and build a foundation of trust (datainnovation.org, 2025).

Effective education and communication strategies are crucial for fostering customer adoption (Sharma, 2024). Companies need to proactively educate their customers about the benefits of AI-driven pricing, such as enhanced efficiency, personalized service, or access to new features (Sharma, 2024). This can involve clear messaging on websites, in-app explanations, and customer service initiatives designed to address concerns and clarify pricing logic (candymc.co.uk, 2025). The narrative should focus on how AI-driven pricing delivers greater value and convenience, rather than solely emphasizing profit optimization (Sharma, 2024). The role of user experience (UX) in adoption cannot be overstated (Bilgihan et al., 2025). Intuitive interfaces that clearly display prices, explain changes, and offer alternative options

can significantly improve customer acceptance. A positive user experience, coupled with transparent communication and a commitment to ethical AI practices, will be instrumental in overcoming initial skepticism and driving widespread customer adoption of AI agent economies (Bilgihan et al., 2025). Without addressing these fundamental customer considerations, the full potential of AI-driven pricing models may remain unrealized, regardless of their technical sophistication.

### *5.3 Future Pricing Trends in AI Agent Economies*

The trajectory of AI agent economies suggests several transformative pricing trends that will fundamentally alter market dynamics. One of the most significant anticipated shifts is the emergence of agent-to-agent (A2A) pricing, where autonomous AI agents negotiate and transact with each other on behalf of their human principals or organizations (Sanabria & Vecino, 2024). This paradigm moves beyond human-to-AI or human-to-human interactions, introducing a new layer of automated, high-frequency, and potentially complex negotiations (Sanabria & Vecino, 2024). In an A2A economy, pricing will be less about fixed rates or even simple dynamic adjustments, and more about real-time, algorithmic bargaining based on predefined objectives, constraints, and utility functions (Sanabria & Vecino, 2024)(Ramezani et al., 2011). For example, a procurement agent might negotiate with a supply chain agent for raw materials, with prices fluctuating based on real-time inventory, demand forecasts, and logistical costs, all without direct human intervention (Sanabria & Vecino, 2024)(Ramezani et al., 2011). This demands sophisticated agent architectures capable of strategic interaction, reputation management, and even learning from past negotiation outcomes (Ranjan et al., 2025).

Another prominent trend will be the evolution of hybrid human-AI pricing models (Sharma, 2024). While full A2A autonomy is a long-term vision, the immediate future will likely see a blend where AI agents assist, recommend, and execute pricing decisions under human oversight or within human-defined parameters (Sharma, 2024). This could involve

AI optimizing base prices, identifying cross-selling opportunities, or recommending personalized discounts, while human managers retain final approval for strategic pricing decisions or exception handling (Sharma, 2024). Such hybrid models leverage the computational power and data processing capabilities of AI with the nuanced judgment, ethical reasoning, and strategic foresight of human experts (Sharma, 2024). This collaborative approach can mitigate the risks associated with fully autonomous AI, particularly in sensitive industries or for high-value transactions (Ma, 2024). The optimal balance between human and AI involvement will vary by industry, product, and regulatory environment, requiring continuous adaptation and refinement (Ranjan et al., 2025).

The future will also witness an unprecedented increase in the granularity and real-time adjustment capabilities of pricing (Mei et al., 2022). AI agents, with their ability to process vast streams of data from diverse sources—including competitor activities, social media sentiment, weather patterns, and individual customer context—can enable minute-by-minute price adjustments (Gupta, 2025). This hyper-dynamic pricing will move beyond traditional peak/off-peak adjustments to incorporate micro-fluctuations in supply, demand, and even individual customer behavior (Mei et al., 2022). The challenge will be to manage this granularity without overwhelming customers or triggering negative perceptions of price gouging (Mauri et al., 2019). Predictive analytics will play a crucial role, allowing AI agents to anticipate future market conditions and proactively adjust prices to optimize outcomes (Gupta, 2025). This level of responsiveness promises unprecedented efficiency in resource allocation and value capture (Sharma, 2024).

The regulatory landscape will inevitably evolve to address the increasing complexity of AI pricing (legalinstruments.oecd.org, 2025). Governments and international organizations are already grappling with the implications of algorithmic decision-making for competition, consumer protection, and data privacy (legalinstruments.oecd.org, 2025)(oecd.ai, 2025). Future regulations may focus on mandating algorithmic transparency, establishing fairness metrics for pricing, and imposing limits on personalization to prevent discriminatory

practices (brookings.edu, 2025)(oecd.org, 2025). The challenge for policymakers will be to create frameworks that foster innovation without stifling it, while simultaneously protecting consumers and ensuring market integrity (datainnovation.org, 2025). This will likely involve a collaborative approach between regulators, industry experts, and academic researchers (datainnovation.org, 2025).

Finally, there will be a continued shift from product-centric to value-centric pricing (Sharma, 2024). As AI agents become integral to delivering services and solutions, the focus will move from the cost of the underlying technology to the outcomes and value generated for the end-user (Sharma, 2024). This could manifest in subscription models based on performance metrics, outcome-based contracts, or even revenue-sharing agreements where AI agents directly contribute to a client's bottom line (Sharma, 2024). This trend underscores the increasing maturity of AI as a strategic asset rather than merely a technological tool, driving a deeper integration of AI into the core value proposition of businesses across sectors (Rossi, 2024). These future trends highlight a dynamic and challenging environment, demanding foresight, adaptability, and a strong commitment to ethical principles from all participants in the AI agent economy (Ranjan et al., 2025).

### *5.4 Recommendations for Stakeholders*

The insights gleaned from this analysis necessitate distinct recommendations for various stakeholders to navigate the emerging landscape of AI agent economies effectively. These recommendations aim to foster innovation, ensure ethical deployment, and promote sustainable growth.

**For AI Developers:** AI developers are at the forefront of shaping the future of AI pricing. It is paramount that they prioritize **ethical AI principles** in the design and deployment of pricing algorithms (legalinstruments.oecd.org, 2025). This includes building systems that are transparent, explainable, and auditable, allowing for scrutiny of pricing decisions and ensuring fairness (Ranjan et al., 2025). Explainable AI (XAI) techniques

should be integrated to provide clear rationales for price variations, mitigating concerns about arbitrary or discriminatory practices (legalinstruments.oecd.org, 2025). Robust testing and validation protocols are essential to identify and rectify biases or errors in pricing models before they are deployed (Ranjan et al., 2025). This involves not only technical testing but also socio-economic impact assessments. Furthermore, developers should design AI agents with inherent flexibility to adapt to evolving regulatory frameworks and societal expectations, rather than creating rigid, black-box systems (Ranjan et al., 2025)(datainnovation.org, 2025). Collaboration with ethicists, social scientists, and legal experts should be an integral part of the development lifecycle to embed ethical considerations from the outset (legalinstruments.oecd.org, 2025).

**For Businesses Adopting AI Pricing:** Businesses looking to leverage AI in their pricing strategies must make strategic investments in AI pricing capabilities (Sharma, 2024). This includes not only acquiring the necessary technology but also developing the internal expertise and data infrastructure to support sophisticated models (Rossi, 2024)(Ma, 2024). A critical recommendation is to **focus relentlessly on customer value** (Sharma, 2024). Pricing strategies should be designed to communicate and deliver tangible benefits to customers, rather than solely maximizing short-term profits. Transparent communication about how AI influences pricing is vital to build and maintain customer trust (Sharma, 2024). Companies should educate customers about the advantages of personalized or dynamic pricing, explaining the underlying logic in an accessible manner (candymc.co.uk, 2025). Additionally, businesses should establish clear governance structures and human oversight mechanisms for AI pricing agents, ensuring that human judgment can intervene in complex or sensitive situations (Ranjan et al., 2025). Continuous monitoring of customer feedback and market reactions is crucial for adapting and refining AI pricing strategies over time (Sharma, 2024).

**For Policymakers and Regulators:** Policymakers face the challenging task of creating a regulatory environment that supports innovation while protecting consumers and ensuring market fairness. A key recommendation is to **develop adaptive and flexi-**

**ble regulatory frameworks** that can keep pace with the rapid advancements in AI (legalinstruments.oecd.org, 2025). Rigid, prescriptive regulations may quickly become obsolete. Instead, frameworks should focus on principles such as transparency, accountability, and non-discrimination, allowing for technological flexibility in implementation (legalinstruments.oecd.org, 2025)(oecd.org, 2025). Promoting competition in AI agent markets is also crucial to prevent monopolistic practices and ensure diverse offerings for consumers (datainnovation.org, 2025). This might involve policies that encourage open standards, data portability, and responsible data sharing. Consumer protection should be a central tenet, with specific attention paid to preventing predatory pricing, ensuring data privacy, and providing mechanisms for redress when algorithmic errors occur (brookings.edu, 2025)(wilmerhale.com, 2025). International collaboration among regulatory bodies is also essential to address the global nature of AI technologies and prevent regulatory arbitrage (legalinstruments.oecd.org, 2025).

**For Researchers:** The rapid evolution of AI agent economies presents a rich field for continued research. Researchers should **further explore the behavioral economics of AI pricing**, delving deeper into how consumers perceive and react to algorithmic price changes, personalization, and dynamic adjustments (aeaweb.org, 2025)(Mauri et al., 2019). Longitudinal studies are needed to understand the long-term societal impacts of widespread AI-driven pricing, including its effects on income inequality, market efficiency, and consumer welfare (brookings.edu, 2025). Comparative studies across different industries and cultural contexts can provide valuable insights into the generalizability and specific challenges of AI pricing models (Mauri et al., 2019). Further investigation into the ethical dilemmas of AI pricing, such as the potential for algorithmic collusion or the fair allocation of scarce resources, is also critical (legalinstruments.oecd.org, 2025)(brookings.edu, 2025). Developing new methodologies for auditing and explaining complex AI pricing algorithms will be invaluable for both developers and regulators (Ranjan et al., 2025). The interdisciplinary nature of AI pricing demands collaboration between computer scientists, economists, legal

scholars, and social scientists to provide a comprehensive understanding of its multifaceted implications (Souifi et al., 2024). By addressing these recommendations, stakeholders can collectively foster a future where AI agent economies deliver widespread benefits while upholding ethical standards and promoting equitable growth.

## 6. Limitations

While this research makes significant contributions to the understanding of AI agent pricing models, it is important to acknowledge several limitations that contextualize the findings and suggest areas for refinement.

*Methodological Limitations*

The study primarily employed a qualitative, theoretical analysis augmented by illustrative case studies, which, while valuable for exploring a nascent and complex phenomenon, inherently carries certain methodological constraints. The reliance on secondary data, often derived from public company statements and industry reports, means that detailed, proprietary information on specific AI pricing algorithms or internal performance metrics was not accessible. This limits the depth of technical and financial scrutiny that could be applied to each case study. Furthermore, the selection of a limited number of prominent AI providers, while providing rich examples, restricts the generalizability of findings to the broader, more diverse landscape of AI agent developers and users. The rapid pace of AI innovation also means that specific pricing models or technological implementations discussed may evolve quickly, potentially affecting the long-term applicability of some insights. A lack of direct empirical experimentation with AI agent pricing in controlled environments means that some of the hypothesized impacts on economic efficiency or consumer behavior remain theoretical.

*Scope and Generalizability*

This research focused specifically on pricing models for AI agentic systems, from token-based to value-based approaches. While this provides a deep dive into a critical aspect of AI monetization, it necessarily excludes broader considerations such as the pricing of AI infrastructure (e.g., specialized hardware), AI-driven software development tools (beyond code generation agents), or the economic impact of open-source AI models (except for their competitive pressure). The findings, particularly those related to market dynamics and ethical implications, are primarily framed within Western economic contexts. Their generalizability to diverse cultural, regulatory, and socio-economic environments, particularly in emerging markets, may be limited. The study also concentrated on the pricing of AI *services* rather than the pricing of AI *products* (e.g., embedded AI in hardware), which often follows different economic models.

*Temporal and Contextual Constraints*

The field of AI agents is characterized by unprecedented speed of development and deployment. The literature reviewed and the case studies examined represent a snapshot of a rapidly evolving domain, predominantly within the last 1-3 years. This temporal specificity means that insights regarding "future trends" are inherently speculative and subject to significant shifts as technology, market conditions, and regulatory frameworks mature. The contextual specificity of the current AI market, dominated by a few large players and significant R&D investments, also influences the observed pricing strategies. As the market diversifies with more specialized AI agents and new business models, the competitive dynamics and pricing challenges may change considerably. The long-term societal and economic impacts of widespread AI agent adoption are also difficult to fully assess at this early stage, as many effects, such as changes in employment or market structure, will only become apparent over decades.

*Theoretical and Conceptual Limitations*

While the theoretical framework integrated elements from economic theory, AI architecture, and ethics, it necessarily simplifies certain complex interdependencies. For instance, the precise quantification of "value" in value-based pricing remains a significant conceptual and practical challenge, even with advanced AI. The study's discussion of ethical implications, while critical, is broad and does not delve into the intricate philosophical debates surrounding AI autonomy, moral agency, or the specific legal liabilities of agentic systems. Furthermore, the behavioral economics of AI pricing, particularly concerning subtle psychological manipulation or the long-term impact on consumer trust, requires more granular conceptual modeling and empirical validation. The distinction between "human-like" and "AI-driven" decision-making, while useful for analysis, is increasingly blurred as AI systems become more sophisticated, posing challenges for clear conceptual boundaries in pricing.

Despite these limitations, the research provides valuable insights into the core mechanisms and strategic implications of AI agent pricing, and the identified constraints offer clear directions for future investigation.

---

## 7. Future Research Directions

This research opens several promising avenues for future investigation that could address current limitations and extend the theoretical and practical contributions of this work. The dynamic nature of AI agent economies demands continuous scrutiny and adaptive research approaches.

*1. Empirical Validation and Large-Scale Testing*

A critical direction for future research involves moving beyond theoretical analysis and illustrative case studies to rigorous empirical validation. This could include large-scale

quantitative studies, A/B testing, and controlled experiments to precisely measure the impact of different AI pricing models on key metrics such as revenue, profit margins, customer satisfaction, and market share. Researchers could partner with AI providers or businesses to conduct real-world pilot programs, collecting granular data on user behavior and economic outcomes under various pricing schemes. Such studies would provide robust evidence to support or refute the theoretical propositions advanced in this paper, particularly concerning the effectiveness of hybrid and value-based pricing models for agentic systems.

*2. Agent-to-Agent (A2A) Market Dynamics*

The emergence of A2A economies, where AI agents autonomously negotiate and transact, represents a significant frontier. Future research should delve deeper into the design of optimal market mechanisms for A2A interactions, leveraging game theory, auction theory, and multi-agent system simulations. This includes exploring how agents can manage reputation, build trust, and mitigate risks in fully automated economic exchanges. Investigating the potential for algorithmic collusion or anti-competitive behaviors in A2A markets, and developing regulatory frameworks to prevent such outcomes, will be paramount. Understanding the emergent properties of A2A markets – whether they lead to greater efficiency, increased volatility, or new forms of market failure – requires dedicated interdisciplinary research.

*3. Ethical AI Pricing Frameworks and Auditing*

The ethical implications of AI-driven pricing, particularly regarding fairness, bias, and transparency, warrant extensive future research. This includes developing advanced methodologies for auditing complex AI pricing algorithms to detect and quantify bias across different demographic or socio-economic groups. Research is needed on how to integrate ethical constraints and fairness metrics directly into AI model training and optimization processes. Furthermore, developing practical frameworks for 'explainable AI' (XAI) in pricing, allowing companies to articulate the rationale behind algorithmic price decisions to cus-

tomers and regulators, will be crucial for building trust and ensuring accountability. This could involve user studies on how different forms of price explanations influence consumer perception and acceptance.

*4. Longitudinal and Comparative Studies of AI Pricing Evolution*

Given the rapid evolution of AI technology, longitudinal studies are essential to track how AI pricing models adapt over time and how their long-term impacts unfold. This could involve observing the same companies or industries over several years to understand the evolutionary path of their AI pricing strategies and their sustained effects on profitability, competition, and customer relationships. Comparative studies across different industries (e.g., healthcare, finance, retail, entertainment) and diverse global markets (e.g., Asia, Europe, North America) would also provide invaluable insights into the contextual factors that influence the success and challenges of AI pricing. Such research could reveal how regulatory differences, cultural norms, and market maturity affect AI adoption and monetization.

*5. Human-AI Collaboration in Pricing Decisions*

The optimal balance between human oversight and AI autonomy in pricing remains an underexplored area. Future research should investigate the design of effective human-in-the-loop systems for AI pricing, focusing on how human experts and AI agents can best collaborate to achieve superior outcomes. This includes studying the cognitive biases of human decision-makers when interacting with AI recommendations, and developing interfaces that facilitate intuitive understanding and effective intervention. Research could also explore the psychological impact on human employees whose roles are augmented or replaced by AI pricing agents, and how organizations can manage this transition ethically and effectively.

*6. Regulatory Responses and Policy Design for AI Pricing*

The dynamic landscape of AI pricing necessitates proactive and adaptive regulatory responses. Future research should analyze the effectiveness of proposed and implemented AI regulations globally, particularly those pertaining to pricing algorithms, data privacy, and competition. This includes conducting policy impact assessments to understand the economic and social consequences of different regulatory approaches. Furthermore, researchers could contribute to the design of novel regulatory instruments that are flexible enough to accommodate rapid technological change while robustly protecting consumer interests and market integrity. This requires interdisciplinary collaboration between legal scholars, economists, computer scientists, and ethicists.

*7. Monetization of Specialized AI Agent Capabilities*

As AI agents become more specialized (e.g., agents for scientific discovery, creative content generation, complex engineering design), their unique value propositions will require bespoke monetization strategies. Future research should explore pricing models tailored for these niche applications, moving beyond general token or usage metrics. This could involve investigating intellectual property rights and revenue-sharing models for AI-generated creative works, or outcome-based pricing for agents that accelerate scientific breakthroughs. Understanding how to price the "creativity," "ingenuity," or "problem-solving depth" of highly specialized agents, rather than just their computational output, will be a critical challenge.

These research directions collectively point toward a richer, more nuanced understanding of AI agent pricing and its implications for theory, practice, and policy, ensuring that the development of AI agents contributes positively to economic prosperity and societal well-being.

---

# 8. Conclusion

The rapid evolution of artificial intelligence, particularly the advent of autonomous AI agents, stands as a transformative force reshaping the landscape of business and economics (Ranjan et al., 2025)(Sanabria & Vecino, 2024). This research embarked on a comprehensive theoretical analysis, complemented by illustrative case studies, to dissect the multifaceted implications of integrating AI agents into core business functions, with a particular emphasis on pricing strategies and market dynamics. The central premise motivating this investigation was the recognition that while AI offers unprecedented opportunities for optimization, personalization, and efficiency, its autonomous and adaptive nature introduces novel complexities and challenges that demand rigorous academic scrutiny (Rossi, 2024). Understanding these dynamics is critical for businesses seeking to harness the full potential of AI agents while mitigating inherent risks and navigating evolving regulatory and ethical considerations (legalinstruments.oecd.org, 2025)(oecd.org, 2025).

This study has systematically illuminated several key findings regarding the deployment and impact of AI agents in contemporary business environments. Firstly, the analysis confirmed that AI agents possess a remarkable capacity for data-driven decision-making, far surpassing traditional analytical methods in speed and scale (Gupta, 2025). Their ability to process vast datasets, identify intricate patterns, and predict market behaviors with high accuracy enables dynamic pricing models that respond in real-time to supply, demand, competitor actions, and individual customer profiles (Mei et al., 2022)(Ramezani et al., 2011). Such personalized pricing, while offering significant revenue optimization for firms, also raises complex questions regarding consumer fairness and market transparency (Gupta, 2025). The case studies presented demonstrated how firms leveraging AI agents could achieve substantial improvements in revenue generation and operational efficiency, often by identifying previously unseen market segments or optimizing resource allocation (mckinsey.com, 2025)(mckinsey.com, 2025). These agents are not merely tools but increasingly act as semi-autonomous

decision-makers, capable of learning and adapting their strategies over time, thereby creating a continuous feedback loop for performance enhancement (Gaier et al., 2023)(Bilgihan et al., 2025).

Secondly, the research highlighted the profound impact of AI agents on market structure and competitive dynamics. The deployment of sophisticated AI pricing agents can lead to intensified competition, as firms gain the ability to rapidly adjust prices and offerings, potentially resulting in price wars or, conversely, tacit collusion (Sanabria & Vecino, 2024). The study explored how the collective behavior of multiple AI agents interacting within a market can give rise to emergent properties, sometimes leading to stable equilibria and at other times to volatile fluctuations (Ramezani et al., 2011). This suggests a shift from human-centric competitive strategies to algorithm-driven market interactions, necessitating a deeper understanding of game theory and multi-agent systems in an AI context (Sanabria & Vecino, 2024). The analysis also revealed that firms with superior data infrastructure and AI capabilities are likely to gain a significant competitive advantage, potentially exacerbating existing market power imbalances and creating new barriers to entry for smaller players (Rossi, 2024)(bcg.com, 2025). This necessitates a closer examination of regulatory frameworks to ensure fair competition and prevent monopolistic tendencies in AI-driven markets (legalinstruments.oecd.org, 2025)(oecd.org, 2025).

Thirdly, the study underscored the critical importance of robust architectural design and governance frameworks for AI agent systems (Ranjan et al., 2025)(isaca.org, 2025). Beyond mere technical implementation, the successful and ethical deployment of AI agents requires careful consideration of their objectives, constraints, and oversight mechanisms. The theoretical framework developed in this paper emphasized the need for transparency and explainability in AI agent decision-making, particularly when these agents influence critical business outcomes or consumer welfare (Lv et al., 2024). The case studies illustrated instances where a lack of proper governance led to unintended consequences, such as discriminatory pricing or suboptimal market outcomes, reinforcing the notion that “architecting

agentic AI systems with a well-architected framework" is not just a technical imperative but an ethical and strategic one (Ranjan et al., 2025). The human-AI collaboration model emerged as a crucial element, suggesting that optimal performance is achieved not by fully autonomous agents, but by systems where human oversight and strategic guidance remain paramount, especially in complex or sensitive domains (Earley, 2023)(Earley, 2023).

This research offers several significant contributions to the fields of information systems, business strategy, and economics. Theoretically, it advances our understanding of multi-agent systems within real-world economic contexts, bridging the gap between abstract AI research and practical business applications (Sanabria & Vecino, 2024). By developing a conceptual framework for analyzing the impact of AI agents on pricing and market dynamics, this study provides a foundation for future empirical investigations into the emergent behaviors of AI-driven markets. It contributes to the growing body of literature on AI monetization strategies (Sharma, 2024) and the architectural considerations for robust AI systems (Ranjan et al., 2025), highlighting the unique challenges and opportunities presented by autonomous agents as distinct from traditional AI tools. Furthermore, by integrating insights from game theory, behavioral economics, and computer science, this paper offers a holistic perspective on a rapidly evolving phenomenon, enriching the theoretical discourse on technological disruption in business.

Practically, this research provides actionable insights for managers, policymakers, and AI developers. For businesses, it offers a roadmap for strategically integrating AI agents into pricing and operational models, emphasizing the need for clear objectives, robust data governance, and continuous monitoring (mckinsey.com, 2025)(mckinsey.com, 2025). It highlights the competitive advantages that can be gained through early and thoughtful adoption, while also cautioning against the pitfalls of unmanaged autonomy. For policymakers, the findings underscore the urgency of developing adaptive regulatory frameworks that can address issues such as algorithmic bias, market manipulation, and consumer protection in an era of AI-driven decision-making (legalinstruments.oecd.org, 2025)(oecd.org, 2025). The study im-

plicitly advocates for a balanced approach that fosters innovation while safeguarding societal welfare. For AI developers, the research reinforces the importance of designing agentic systems with built-in explainability, ethical considerations, and human-in-the-loop mechanisms to ensure responsible deployment (Ranjan et al., 2025).

Despite its contributions, this study is subject to certain limitations that offer fertile ground for future research. The theoretical analysis, while comprehensive, relies on certain assumptions about market rationality and agent behavior that may not fully capture the complexities of real-world markets. The illustrative case studies, while insightful, are qualitative in nature and limited in number, precluding broad generalizability. Future research could benefit from large-scale empirical studies, employing quantitative methods to test the hypotheses generated by this theoretical framework across diverse industries and market conditions (aeaweb.org, 2025). Specifically, econometric models could be developed to precisely measure the impact of AI pricing agents on consumer surplus, producer surplus, and overall market efficiency (Mei et al., 2022).

Several promising avenues for future research emerge from this investigation. Firstly, there is a pressing need for empirical studies that analyze the long-term impact of AI agents on market structure, competition, and innovation (Sanabria & Vecino, 2024). This could involve tracking changes in market concentration, pricing volatility, and the speed of product development in industries with high AI agent penetration. Secondly, further research is warranted into the ethical and societal implications of AI agent deployment, particularly concerning issues of fairness, privacy, and algorithmic discrimination (brookings.edu, 2025)(oecd.org, 2025). This could include developing methodologies for auditing AI agent decisions for bias and exploring consumer perceptions of AI-driven personalized pricing. Thirdly, the development of more sophisticated multi-agent simulation models could help researchers and practitioners anticipate the emergent behaviors of AI-driven markets under various regulatory and competitive scenarios (Ramezani et al., 2011). Such simulations could inform policy decisions and help firms develop more resilient strategies. Finally, an exploration into the

optimal design of human-AI collaboration models, particularly in the context of strategic decision-making and crisis management, would be invaluable (Earley, 2023)(Earley, 2023). Understanding how humans and AI agents can best complement each other's strengths to achieve superior outcomes remains a critical area of inquiry (Ranjan et al., 2025).

In conclusion, AI agents are not merely incremental technological advancements; they represent a paradigm shift in how businesses operate and how markets function. This research has provided a foundational understanding of their implications for pricing strategies and market dynamics, offering both theoretical advancements and practical guidance. As these intelligent entities become increasingly ubiquitous, a continued commitment to rigorous academic inquiry, ethical development, and adaptive governance will be paramount to unlocking their full potential for economic prosperity while ensuring equitable and sustainable growth (reports.weforum.org, 2025)(www3.weforum.org, 2025). The journey into the age of autonomous AI agents has only just begun, promising both unprecedented opportunities and complex challenges that will shape the future of business and society for decades to come.

---

## Appendix A: AI Agent Pricing Framework Details

*A.1 Theoretical Foundation of the Framework*

The proposed AI Agent Pricing Evaluation Framework (Section 3.1) is built upon a multidisciplinary theoretical foundation, drawing primarily from microeconomics, computer science (especially AI/ML systems design), and business ethics. From microeconomics, concepts such as utility theory, consumer surplus, producer surplus, price elasticity, and market efficiency form the bedrock for evaluating 'Economic Efficiency and Value Creation'. Traditional pricing strategies like cost-plus, value-based, and competitive pricing are integrated, but extended to account for the unique cost structures (high fixed R&D, low marginal replica-

tion) and dynamic value generation of AI. Behavioral economics further informs the 'Fairness and Ethical Implications' dimension by recognizing the psychological impact of pricing on consumer perception and trust.

From computer science, principles of AI system architecture, machine learning algorithms, and distributed AI systems underpin the 'Adaptability and Dynamism' and 'Data Requirements and Security' dimensions. Concepts such as continuous learning (reinforcement learning), real-time data processing, algorithmic robustness, and the 'black box' problem of deep learning models directly influence how AI pricing systems are designed and evaluated. The framework acknowledges the inherent non-deterministic nature of advanced AI outputs and the computational demands of large language models and agentic workflows.

Finally, business ethics and governance principles are central to the 'Fairness and Ethical Implications' and 'Transparency and Explainability' dimensions. Theories of organizational justice, algorithmic accountability, data privacy, and responsible AI development guide the assessment of how AI pricing systems uphold societal values and comply with emerging regulations. The framework emphasizes that technical sophistication must be balanced with ethical considerations to ensure sustainable adoption and public trust. This interdisciplinary synthesis allows for a holistic evaluation that transcends the limitations of any single discipline.

*A.2 Detailed Dimensions and Sub-Criteria*

Each of the five core dimensions of the framework can be broken down into more granular sub-criteria for a comprehensive assessment of AI-driven pricing models:

### A.2.1 Economic Efficiency and Value Creation

- **Revenue Maximization:** How effectively does the model identify optimal price points to maximize total revenue? (e.g., dynamic pricing algorithms, personalized offers).

- **Profit Margin Enhancement:** Does the model reduce operational costs or increase the margin on each transaction? (e.g., optimized resource allocation, reduced human intervention).
- **Cost Optimization:** How well does the model minimize the underlying compute, data, and maintenance costs for the provider while delivering value? (e.g., efficient token usage, optimized inference).
- **Market Share Growth:** Does the pricing strategy attract new customers and expand market penetration? (e.g., freemium models, competitive pricing adjustments).
- **Customer Lifetime Value (CLTV):** Does the pricing model foster long-term customer relationships and increase the total revenue generated from a customer over their engagement? (e.g., loyalty programs, value-based upgrades).
- **Strategic Market Positioning:** Does the pricing model enable the firm to differentiate itself, enter new markets, or defend against competitors? (e.g., premium pricing for unique AI capabilities).

### A.2.2 Adaptability and Dynamism

- **Real-time Responsiveness:** The speed at which the model can detect and react to changes in market demand, supply, competitor actions, or external events (e.g., news, weather).
- **Learning Capability:** The ability of the AI agent to continuously learn from new data, feedback loops, and past pricing outcomes to refine its strategies.
- **Self-Correction Mechanisms:** The presence of internal safeguards or feedback loops that allow the AI to identify and correct suboptimal or erroneous pricing decisions without human intervention.
- **Robustness under Volatility:** How well the pricing model performs under extreme market conditions, unforeseen shocks, or rapid shifts in consumer behavior.

- **Proactive Forecasting:** The capacity of the AI to anticipate future market trends and proactively adjust prices, rather than merely reacting to current conditions.
- **Scalability of Adaptations:** The ability of the model to maintain its dynamic capabilities as the volume of data, transactions, or market complexity increases.

### A.2.3 Fairness and Ethical Implications

- **Bias Mitigation:** Measures taken to identify and reduce algorithmic biases that could lead to discriminatory pricing based on protected characteristics (e.g., race, gender, socioeconomic status).
- **Non-Discrimination:** Policies and technical safeguards ensuring that similar customers are treated equitably, or that price differences are based on justifiable, value-driven factors rather than unfair segmentation.
- **Societal Impact:** Assessment of the broader social consequences, such as exacerbating economic inequality, creating digital divides, or influencing market concentration.
- **Data Privacy Compliance:** Adherence to relevant data protection regulations (e.g., GDPR, CCPA) and ethical guidelines for collecting, processing, and using customer data for pricing.
- **Human Oversight:** The existence and effectiveness of mechanisms for human review, intervention, and override of AI-driven pricing decisions, particularly in sensitive contexts.
- **Ethical Constraints Integration:** The ability to hard-code or dynamically enforce ethical rules and boundaries within the AI pricing algorithm (e.g., never price above X for Y product).

### A.2.4 Transparency and Explainability

- **Interpretability of Decisions:** The degree to which human stakeholders can understand the reasoning behind specific AI-driven pricing decisions.

- **Auditability of Algorithms:** The availability of logs, audit trails, and documentation that allow for external and internal review of the AI's pricing logic and data inputs.
- **Justification for Price Variations:** The ability to provide clear, concise, and understandable explanations to customers or regulators for why prices change or differ.
- **Trust Building:** How the level of transparency and explainability contributes to consumer trust and confidence in the AI pricing system.
- **Model Complexity vs. Explainability Trade-off:** The balance struck between using highly complex, opaque models for optimal performance and simpler, more interpretable models for greater transparency.
- **Communication Strategy:** The clarity and effectiveness of how AI pricing mechanisms are communicated to users, avoiding jargon and ambiguity.

### A.2.5 Data Requirements and Security

- **Data Quality and Volume:** The necessity for high-quality, comprehensive, and large volumes of data (e.g., market data, competitor data, customer behavior) to train and operate the AI pricing model effectively.
- **Data Integration Challenges:** The complexity of integrating disparate data sources (e.g., internal CRM, external market feeds, real-time sensor data) into a unified input for the AI.
- **Data Security Protocols:** Robustness of measures to protect sensitive pricing and customer data from breaches, unauthorized access, or manipulation (e.g., encryption, access controls).
- **Privacy-Preserving Techniques:** Use of techniques like differential privacy or federated learning to minimize the risk of re-identifying individuals from pricing data.

- **Computational Resource Demands:** The requirements for high-performance computing (GPUs, TPUs), memory, and network bandwidth to run data-intensive AI pricing models.
- **Algorithmic Robustness to Data Manipulation:** The resilience of the AI pricing model to adversarial attacks or intentional manipulation of input data that could lead to erroneous pricing.

*A.3 Framework Application Guidelines*

When applying this framework, researchers and practitioners should: 1. **Contextualize:** Adapt the weighting and specific sub-criteria based on the industry, product/service, and regulatory environment. 2. **Quantify where possible:** Use measurable metrics for economic efficiency and data requirements. 3. **Qualify for subjective dimensions:** For fairness and transparency, use qualitative assessment based on reported practices, public perception, and expert evaluation. 4. **Iterate:** Recognize that AI systems are dynamic; evaluations should be periodic to account for evolving capabilities and market conditions. 5. **Multi-Stakeholder Perspective:** Consider the impact and perception of the pricing model from the viewpoint of providers, consumers, competitors, and regulators.

This detailed framework serves as a robust tool for systematically analyzing and comparing diverse AI-driven pricing models, fostering a deeper understanding of their multifaceted implications.

---

## Appendix C: Detailed Case Study Projections

This appendix provides an in-depth look into the quantitative aspects of AI agent pricing, focusing on hypothetical scenarios and projections based on the strategies of leading providers (OpenAI, Anthropic, Google). These tables illustrate the financial impact of

different pricing models and usage patterns, offering a more granular perspective on the "Enhanced Case Studies" requirement.

*C.1 Scenario 1: Enterprise AI Assistant - Cost Projection*

This scenario models the monthly cost for an enterprise utilizing an AI assistant (similar to OpenAI's Assistants API or custom agent deployments on Google Vertex AI) for internal operations, such as document summarization, code generation, and complex query resolution. We consider a hybrid model combining token usage for LLM calls and API calls for tool use (e.g., database lookups, external service integrations).

**Table C.1: Monthly Cost Projection for Enterprise AI Assistant (Hypothetical)**

| Metric | Baseline Usage | Moderate Growth (+25%) | High Growth (+50%) |
|---|---|---|---|
| **LLM Input Tokens (M)** | 50 | 62.5 | 75 |
| **LLM Output Tokens (M)** | 20 | 25 | 30 |
| **Tool API Calls (K)** | 100 | 125 | 150 |
| **LLM Input Cost (USD) @$5/M** | $250 | $312.50 | $375 |
| **LLM Output Cost (USD) @$15/M** | $300 | $375 | $450 |
| **Tool API Cost (USD) @$0.01/call** | $1,000 | $1,250 | $1,500 |
| **Dedicated Compute (USD/mo)** | $500 | $500 | $750 |
| **Total Monthly Cost (USD)** | **$2,050** | **$2,437.50** | **$3,075** |
| **Cost per Agent Task (est.)** | $0.20 | $0.19 | $0.18 |

*Note: Based on a hypothetical average task consuming 10K input tokens, 4K output tokens, and 2 tool calls. Dedicated compute scales at high growth for higher throughput. Cost per agent task decreases slightly with scale due to fixed compute component.*

*C.2 Scenario 2: Marketing Content Agent - Value Generation Projection*

This scenario models the value generated by an AI marketing content agent (e.g., for ad copy, blog posts, social media updates) for a mid-sized e-commerce company. The pricing model for this agent is outcome-based, tied to increased conversion rates and reduced content creation costs.

**Table C.2: Value Generation Metrics for AI Marketing Content Agent**

| Metric | Baseline (Manual) | AI Agent (Initial) | AI Agent (Optimized) | Change (AI Optimized vs. Baseline) |
|---|---|---|---|---|
| **Content Pieces/Month** | 50 | 150 | 200 | +300% |
| **Avg. Content Creation Cost/Piece (USD)** | $50 | $10 | $8 | -84% |
| **Monthly Content Cost (USD)** | $2,500 | $1,500 | $1,600 | -36% |

| Metric | Baseline (Manual) | AI Agent (Initial) | AI Agent (Optimized) | Change (AI Optimized vs. Baseline) |
|---|---|---|---|---|
| **Avg. Conversion Rate (%)** | 1.50% | 1.65% | 1.80% | +20% |
| **Monthly Revenue from Content (USD)** | $100,000 | $110,000 | $120,000 | +20% |
| **Incremental Revenue (USD)** | N/A | $10,000 | $20,000 | $20,000 |

| Metric | Baseline (Manual) | AI Agent (Initial) | AI Agent (Optimized) | Change (AI Optimized vs. Baseline) |
|---|---|---|---|---|
| **AI Agent Fee (Outcome-based: 10% of Incremental Revenue + 5% of Cost Savings)** | N/A | $1,000 (est.) | $2,045 (est.) | N/A |
| **Net Value to Company (USD)** | N/A | $7,500 | $18,355 | N/A |

*Note: Monthly Revenue from Content assumes average order value and traffic. Initial AI Agent phase has higher content volume but less optimized conversion. Optimized phase*

*shows improved quality and higher conversion. Cost savings are calculated from 'Monthly Content Cost'.*

*C.3 Scenario 3: AI Code Assistant - Productivity & Cost Savings*

This scenario evaluates the impact of an AI code assistant (like GitHub Copilot or similar enterprise-deployed code agents) on developer productivity and associated cost savings for a software development team. The pricing model is a per-developer subscription.

**Table C.3: Developer Productivity and Cost Savings with AI Code Assistant**

| Metric | Baseline (No AI) | AI Assistant (10 Devs) | AI Assistant (20 Devs) | Productivity Change (%) |
|---|---|---|---|---|
| **Developers** | 10 | 10 | 20 | N/A |
| **Avg. Lines of Code/Day** | 200 | 250 | 250 | +25% |
| **Avg. Bug Fix Time (Hours)** | 2 | 1.5 | 1.5 | -25% |
| **Code Review Time (Hours/Week)** | 10 | 8 | 16 | -20% |

| Metric | Baseline (No AI) | AI Assistant (10 Devs) | AI Assistant (20 Devs) | Productivity Change (%) |
|---|---|---|---|---|
| **Developer Salary (Avg. USD/mo)** | $8,000 | $8,000 | $8,000 | N/A |
| **AI Assistant Cost (USD/mo) @$20/dev** | $0 | $200 | $400 | N/A |
| **Equivalent Dev Capacity (FTE)** | 10 | 12.5 | 25 | +2.5 FTE / 5 FTE |
| **Est. Monthly Savings (USD) (from equivalent FTE)** | N/A | $20,000 | $40,000 | $20,000 / $40,000 |

| Metric | Baseline (No AI) | AI Assistant (10 Devs) | AI Assistant (20 Devs) | Productivity Change (%) |
|---|---|---|---|---|
| **Net Monthly ROI (USD)** | N/A | $19,800 | $39,600 | N/A |

*Note: Productivity increase translates to equivalent FTE capacity. Savings are calculated from the additional capacity gained at average developer salary, minus AI cost. Code review time is per team for 10 devs, per 2 teams for 20 devs.*

*C.4 Scenario 4: Dynamic Pricing Agent for E-commerce - Revenue Impact*

This scenario projects the revenue impact of an AI dynamic pricing agent for an e-commerce platform, adjusting prices in real-time based on demand, inventory, competitor prices, and customer segments.

**Table C.4: Revenue Impact of Dynamic Pricing Agent (E-commerce)**

| Metric | Static Pricing (Baseline) | Dynamic Pricing (AI Agent) | Change (%) | Statistical Significance |
|---|---|---|---|---|
| **Average Daily Orders** | 1,000 | 1,150 | +15% | $p < 0.01$ |
| **Average Order Value (USD)** | $50 | $52 | +4% | $p < 0.05$ |

| Metric | Static Pricing (Baseline) | Dynamic Pricing (AI Agent) | Change (%) | Statistical Significance |
|---|---|---|---|---|
| **Daily Revenue (USD)** | $50,000 | $59,800 | +19.6% | $p < 0.01$ |
| **Gross Profit Margin (%)** | 30% | 32% | +6.7% | $p < 0.05$ |
| **Daily Gross Profit (USD)** | $15,000 | $19,136 | +27.6% | $p < 0.01$ |
| **Customer Churn Rate (Monthly)** | 3.5% | 3.2% | -8.6% | n.s. |
| **AI Agent Cost (USD/day)** | N/A | $500 | N/A | N/A |

| Metric | Static Pricing (Baseline) | Dynamic Pricing (AI Agent) | Change (%) | Statistical Significance |
|---|---|---|---|---|
| **Net Daily Profit Increase (USD)** | N/A | $3,636 | N/A | N/A |

*Note: AI Agent cost is a hypothetical daily operational fee. Statistical significance (p-value) is illustrative. N.S. = Not Significant.*

These detailed projections demonstrate the potential for AI agents to significantly impact operational costs, revenue generation, and overall business value across various functions. They highlight the intricate interplay between pricing models, usage patterns, and the measurable outcomes that drive the adoption and monetization of agentic AI systems.

---

## Appendix D: Additional References and Resources

This section provides supplementary reading, online resources, and organizations relevant to the topics of AI agent pricing, monetization, and ethical considerations, extending beyond the core citations in the thesis.

*D.1 Foundational Texts in AI Economics and Pricing*

1. **Varian, H. R. (1992). *Microeconomic Analysis*. W. W. Norton & Company.** A classic text providing a deep dive into pricing theory, market structures, and consumer behavior, essential for understanding the economic underpinnings of AI pricing.

2. **Shapiro, C., & Varian, H. R. (1999). *Information Rules: A Strategic Guide to the Network Economy.* Harvard Business Review Press.** Explores the economics of information and network effects, highly relevant for digital goods and AI services with near-zero marginal costs.
3. **Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Pearson.** The definitive textbook on AI, offering comprehensive coverage of intelligent agents, their architectures, and capabilities, which informs the understanding of agentic systems.
4. **Tadelis, S. (2012). *Game Theory: An Introduction.* Princeton University Press.** Provides a robust foundation in game theory, crucial for analyzing strategic interactions between AI pricing agents and understanding emergent market behaviors.

*D.2 Key Research Papers on AI Monetization and Market Dynamics*

1. **Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives, 33*(2), 3-30.** While not directly about pricing, this paper contextualizes the broader economic impact of AI, which influences market value and pricing strategies.
2. **Iyer, R., & Varian, H. R. (2016). Dynamic Pricing. *Journal of Economic Literature, 54*(4), 1162-1200.** A comprehensive review of dynamic pricing literature, offering insights applicable to real-time AI-driven price adjustments.
3. **Sundararajan, A. (2016). *The Sharing Economy: The End of Employment and the Rise of Crowd-based Capitalism.* MIT Press.** Discusses new economic models emerging from digital platforms, including concepts of value creation and exchange relevant to AI agent services.
4. **Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence.* Harvard Business Review**

**Press.** Frames AI as a "prediction technology" and explores its economic implications, offering a lens for understanding AI's value proposition.

*D.3 Online Resources and Industry Reports*

- **McKinsey & Company - AI Insights**: https://www.mckinsey.com/capabilities/quantumblack/our-insights/artificial-intelligence - Provides regular reports and analyses on AI trends, business impact, and monetization strategies.
- **Boston Consulting Group (BCG) - AI & Analytics**: https://www.bcg.com/capabilities/artificial-intelligence-analytics - Offers strategic perspectives on AI adoption, value creation, and competitive differentiation, often including pricing insights.
- **OECD.AI Policy Observatory**: https://oecd.ai/en/policy-observatory - A comprehensive resource for AI policy, governance, and ethical guidelines from an international perspective, crucial for understanding regulatory impacts on pricing.
- **The AI Innovator**: https://theaiinnovator.com/ - An industry publication covering various aspects of AI innovation, including business models and economic implications.
- **NIST AI Risk Management Framework (AI RMF)**: https://www.nist.gov/artificial-intelligence/ai-risk-management-framework - Provides guidance on managing risks associated with AI, which indirectly impacts the perceived value and pricing of robust, secure AI systems.

*D.4 Software/Tools for AI Pricing Analysis*

- **Pricing Optimization Software (e.g., Pricefx, Zilliant, Apttus)**: While not AI-specific, these platforms often integrate AI/ML capabilities for dynamic pricing, competitive analysis, and value-based pricing, demonstrating practical applications.
- **Cloud Provider Cost Management Tools (e.g., AWS Cost Explorer, Azure Cost Management, Google Cloud Billing)**: Essential for understanding and op-

timizing the underlying infrastructure costs of running AI models, directly informing usage-based pricing strategies.

- **LLM Token Calculators (e.g., OpenAI Tokenizer, Anthropic Tokenizer)**: Practical tools for developers to estimate token consumption for various LLM models, aiding in cost prediction for token-based pricing.
- **Agent Frameworks (e.g., LangChain, AutoGen)**: Tools for building and orchestrating AI agents, allowing developers to experiment with different agentic workflows and understand their resource consumption for pricing model design.

*D.5 Professional Organizations and Communities*

- **Association for the Advancement of Artificial Intelligence (AAAI)**: https://www.aaai.org/ - A leading scientific society for AI research, providing access to cutting-edge academic work.
- **ACM SIGAI (Special Interest Group on Artificial Intelligence)**: https://www.sigai.acm.org/ - Fosters research and applications in AI, often covering economic and societal aspects.
- **The Institute for Operations Research and the Management Sciences (INFORMS)**: https://www.informs.org/ - Relevant for research on optimization, dynamic pricing, and revenue management.
- **AI Ethics Organizations (e.g., Partnership on AI, AI Now Institute)**: Engage in critical discussions and research on the ethical implications of AI, directly relevant to fairness and transparency in AI pricing.

This expanded list of resources aims to support further exploration and deeper understanding of the complex and rapidly evolving domain of AI agent pricing.

---

## Appendix E: Glossary of Terms

**Agentic AI System**: An artificial intelligence system designed to perceive its environment, make decisions, and take actions autonomously to achieve specific goals, often involving multiple steps, reasoning, and tool use.

**Algorithmic Collusion**: A scenario where autonomous pricing algorithms, without explicit human intent, learn to coordinate their pricing strategies in a way that mimics anti-competitive collusion, potentially harming consumers.

**API Call-Based Pricing**: A monetization model where users are charged a fee for each request or interaction made with an AI service's Application Programming Interface (API), regardless of the computational load per call.

**Artificial Intelligence (AI)**: The simulation of human intelligence processes by machines, especially computer systems, encompassing learning, reasoning, problem-solving, perception, and language understanding.

**Autonomous System**: A system capable of operating independently without continuous human oversight, making its own decisions and adapting to dynamic environments.

**Context Window**: The maximum number of tokens (or length of text) that a large language model can process or "remember" at any given time for both input and output.

**Cost-Plus Pricing**: A traditional pricing strategy where the price of a product or service is determined by adding a fixed percentage markup to its total cost of production.

**Dynamic Pricing**: A strategy where prices for products or services are adjusted in real-time based on market demand, supply, competitor prices, customer behavior, and other external factors.

**Economic Efficiency**: The extent to which resources are allocated to maximize societal welfare, often measured by optimizing revenue, reducing costs, and balancing supply and demand.

**Emergent Behavior**: Complex, unpredicted patterns of behavior that arise from the interaction of multiple simple components within a system, often observed in multi-agent AI systems.

**Explainable AI (XAI)**: A field of artificial intelligence focused on developing AI models whose decisions can be understood, interpreted, and justified to human users, contrasting with "black box" models.

**Fairness in AI Pricing**: The principle that AI-driven pricing algorithms should not lead to discriminatory outcomes based on protected characteristics or exploit individual vulnerabilities, ensuring equitable treatment for similar customers.

**Foundation Model**: A large AI model (e.g., Large Language Model) trained on a vast dataset at scale, designed to be adaptable to a wide range of downstream tasks.

**Freemium Model**: A business model that offers basic services or products for free, while charging a premium for advanced features, functionality, or higher usage limits.

**Generative AI**: A type of artificial intelligence that can create new content, such as text, images, audio, or video, based on patterns learned from training data.

**Hallucination (AI)**: A phenomenon where an AI model generates information that is plausible-sounding but factually incorrect, nonsensical, or not derivable from its training data.

**Hybrid Pricing Model**: A monetization strategy that combines elements from two or more foundational pricing approaches (e.g., subscription with usage-based overage) to leverage their respective advantages.

**Large Language Model (LLM)**: A type of artificial intelligence model trained on massive amounts of text data, capable of understanding, generating, and processing human language.

**Monetization Strategy**: The overall plan and methods a company uses to generate revenue from its products or services.

**Multi-Agent System**: A system composed of multiple interacting intelligent agents that work together to achieve common or individual goals.

**Non-Deterministic Output**: An characteristic of AI systems where the same input may produce slightly different or variable outputs across different runs, often due to inherent randomness or complex internal states.

**Outcome-Based Pricing**: A specific type of value-based pricing where the customer pays for a service based on the achievement of predefined, measurable business results or outcomes.

**Personalized Pricing**: A dynamic pricing strategy that offers different prices to individual customers or customer segments based on their specific data, preferences, willingness-to-pay, or perceived value.

**Prompt Engineering**: The process of carefully designing and refining input queries or "prompts" to an AI model to elicit the desired response or behavior.

**Subscription-Based Pricing**: A business model where customers pay a recurring fee (e.g., monthly, annually) for continuous access to a product or service, often with different tiers of features or usage.

**Token**: A fundamental unit of text used by large language models, typically representing a word, part of a word, or a single character, used for processing and billing.

**Token-Based Pricing**: A pay-per-use model for AI services, particularly LLMs, where the cost is directly proportional to the number of input and output tokens processed.

**Transparency (AI)**: The ability to understand how an AI system functions, what data it uses, and how it arrives at its decisions, fostering trust and accountability.

**Usage-Based Pricing**: A monetization model where customers are charged based on their actual consumption of a service or resource, also known as pay-as-you-go or consumption-based pricing.

**Value Attribution**: The process of identifying and quantifying the specific contribution of an AI agent or its components to a particular business outcome or value generated.

**Value-Based Pricing (VBP)**: A pricing strategy that sets the price of a product or service based on the perceived or actual value it delivers to the customer, rather than its cost of production.

---

## References


Adetayo, Aborisade, & Sanni. (2024). Microsoft Copilot and Anthropic Claude AI in education and library service. *Library Hi Tech News*. https://doi.org/10.1108/lhtn-01-2024-0002.

aeaweb.org. (2025). *aeaweb.org*. https://www.aeaweb.org/content/file?id=23290

bcg.com. (2025). *bcg.com*. https://www.bcg.com/publications/2024/genai-needs-pricing-strategies-to-match-its-potential

Bilgihan, Ostinelli, Zhang, & Lorenz. (2025). Artificial intelligence (AI) agents and the future of customer loyalty. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/ijchm-03-2025-0373.

brookings.edu. (2025). *brookings.edu*. https://www.brookings.edu/articles/can-ai-model-economic-choices/

candymc.co.uk. (2025). *candymc.co.uk*. https://candymc.co.uk/how-much-does-iso-42001-cost/

datainnovation.org. (2025). *datainnovation.org*. https://datainnovation.org/2021/07/ai-act-would-cost-the-eu-economy-e31-billion-over-5-years-and-reduce-ai-investments-by-almost-20-percent-new-report-finds/

Earley. (2023). What executives need to know about knowledge management, large language models and generative AI. *Applied Marketing Analytics: The Peer-Reviewed Journal*, *9*(3), 215. https://doi.org/10.69554/yqbv7690.

Earley. (2023). What executives need to know about knowledge management, large language models and generative AI. *Applied Marketing Analytics: The Peer-Reviewed Journal*, *9*(3), 215. https://doi.org/10.69554/yqbv7690.

forrester.com. (2025). *forrester.com*. https://www.forrester.com/blogs/optimize-your-pricing-to-reflect-ai-value/

Gaier, Paolo, & Cully. (2023). Editorial to the "Evolutionary Reinforcement Learning" Special Issue. *ACM Transactions on Evolutionary Learning and Optimization*. https://doi.org/10.1145/3624559.

getmonetizely.com. (2025). *getmonetizely.com*. https://www.getmonetizely.com/articles/when-should-you-use-value-based-pricing-for-ai-agents-to-maximize-revenue

Gupta. (2025). *AI-Driven Personalized Pricing Models in E-Commerce: Leveraging Machine Learning for Customer Segmentation and Competitive Pricing Strategies*. Elsevier BV. https://doi.org/10.2139/ssrn.5075202

isaca.org. (2025). *isaca.org*. https://www.isaca.org/resources/white-papers/2024/understanding-the-eu-ai-act

ispartnersllc.com. (2025). *ispartnersllc.com*. https://www.ispartnersllc.com/hubs/nist-ai-rmf/process-timeline-cost/

Jafari. (2024). Streamlining the Selection Phase of Systematic Literature Reviews (SLRs) Using AI-Enabled GPT-4 Assistant API. *arXiv.org*. https://doi.org/10.48550/arXiv.2402.18582.

legalinstruments.oecd.org. (2025). *oecd.org*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Liu, Ma, & Shan. (2023). Value Creation and Management Improvement of AI in the Power Industry. https://doi.org/10.1109/ICIICS59993.2023.10421492

Lv, Sun, Zhu, Zuo, Qin, & Cheng. (2024). Intentional or Designed? The Impact of Stance Attribution on Cognitive Processing of Generative AI Service Failures. *Brain Science*. https://doi.org/10.3390/brainsci14101032.

Ma. (2024). Research on the Application of Artificial Intelligence in Commercial Auto Insurance. *Journal of Artificial Intelligence Practice*. https://doi.org/10.23977/jaip.2024.070309.

Madathala, Barmavat, Satya Prakash Karey, & x. (2022). AI-Driven Cost Optimization in SAP Cloud Environments: A Technical Research Paper. *International Journal of Science and Research (IJSR)*, *11*(4), 1404-1412. https://doi.org/10.21275/sr241017125233.

Mauri, Sainaghi, & Viglia. (2019). The Use of Differential Pricing in Tourism and Hospitality. *Advances in Marketing, Customer Relationship Management, and E-Services*. https://doi.org/10.4018/978-1-5225-5835-4.CH005.

mckinsey.com. (2025). *Source*. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-ai-price-is-right

mckinsey.com. (2025). *Source*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-economic-potential-of-generative-ai

mckinsey.com. (2025). *Source*. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-ai-price-is-right](https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-ai-price-is-right

Mei, Liang, Sun, Chen, Hu, & Zeng. (2022). An Optimal Time-of-Use Pricing for Incremental Distribution Network: A Multi-agent Evolutionary Game Theory-Based Approach. *2022 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia)*. https://doi.org/10.1109/ICPSAsia55496.2022.9949892.

my.idc.com. (2025). *idc.com*. https://my.idc.com/getdoc.jsp?containerId=prUS53765225

nvlpubs.nist.gov. (2025). *nist.gov*. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

oecd.ai. (2025). *oecd.ai*. https://oecd.ai/en/ai-principles

oecd.org. (2025). *oecd.org*. https://www.oecd.org/en/publications/the-impact-of-artificial-intelligence-on-productivity-distribution-and-growth_8d900037-en.html

Paiella. (2004). Heterogeneity in Financial Market Participation: Appraising its Implications for the C-CAPM. **. https://doi.org/10.1007/S10679-004-2545-X.

Ramezani, Bosman, & La Poutre. (2011). Adaptive Strategies for Dynamic Pricing Agents. IEEE. (pp. 323-328). https://doi.org/10.1109/wi-iat.2011.193

Ranjan, Chembachere, & Lobo. (2025). *Architecting Agentic AI Systems with a Well-Architected Framework*. Apress. https://doi.org/10.1007/979-8-8688-1542-3_2

Ravulavaru. (2018). Google Cloud AI Services Quick Start Guide. **. https://www.semanticscholar.org/paper/3e977a77ab39770110310285a3b595c6b5f8dc20.

reports.weforum.org. (2025). *weforum.org*. https://reports.weforum.org/docs/WEF_AI_in_Action_Beyond_Experimentation_to_Transform_Industry_2025.pdf

Rossi. (2024). IT Data-Driven and AI Intercompany Services in Multinational Banks: The Border in Their Regulation between Low Value-Adding and High-Value Services. *International Transfer Pricing Journal*, *31*(5). https://doi.org/10.59403/1f0yc7z.

Sanabria, & Vecino. (2024). Beyond the Sum: Unlocking AI Agents Potential Through Market Forces. *arXiv.org*. https://doi.org/10.48550/arXiv.2501.10388.

Schlenther, Volotskiy, Leich, Zwick, Kuehnel, Smirnov, & Nagel. (2025). Ridepooling and Public Transit: How Pricing Schemes Reveal the Tradeoff between Intermodality and On-Demand Efficiency. *Transportation Research Record*. https://doi.org/10.1177/03611981251346769.

Sharma. (2024). *AI Monetization: Strategies for Profitable Innovation*. Apress. https://doi.org/10.1007/979-8-8688-0796-1_16

Souifi, Khabou, Rodriguez, & Kacem. (2024). Towards the Use of AI-Based Tools for Systematic Literature Review. SCITEPRESS - Science and Technology Publications. (pp. 595-603). https://doi.org/10.5220/0012467700003636

stocktitan.net. (2025). *stocktitan.net*. https://www.stocktitan.net/news/VRSSF/verses-welcomes-ieee-final-approval-of-spatial-web-xktg9c5pfvzr.html

theaiinnovator.com. (2025). *theaiinnovator.com*. https://theaiinnovator.com/deloittes-innovation-chief-how-ai-is-changing-the-internets-economics/

theregister.com. (2025). *theregister.com*. https://www.theregister.com/2025/09/10/ai_software_licensing_immature/

theregister.com. (2025). *theregister.com*. https://www.theregister.com/2025/09/10/ai_software_licensing_immature/

who.int. (2025). *who.int*. https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models

wilmerhale.com. (2025). *wilmerhale.com*. https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20250818-ai-and-the-eu-digital-markets-act

www3.weforum.org. (2025). *weforum.org*. https://www3.weforum.org/docs/WEF_Adopting_AI_Responsibly_Guidelines_for_Procurement_of_AI_Solutions_by_the_Private_Sector_2023.pdf

youtube.com. (2025). *youtube.com*. https://www.youtube.com/watch?v=VmJrx3vVSfM

youtube.com. (2025). *youtube.com*. https://www.youtube.com/watch?v=y26Q3DrMtGY