

WHICH PRICING MODELS ARE THE MOST EFFECTIVE FOR AI AGENTS IN TODAY'S TECHNOLOGY LANDSCAPE?



Please tick relevant box	Project :	Thesis: <input checked="" type="checkbox"/>	Written Assignment:
Study Programme:	MSc Marketing & Sales		
Semester:	4		
Exam Title:	Master's Thesis		
Group Number:	18		
Names + Student Nos of group member(s):	Name(s)	Student Number(s)	
	Raul Moriana Sigel	20235830	
Submission date:	2.06.2025		
Project Title /Thesis Title	Which Pricing Models are most effective for AI Agents in today's technology landscape?		
According to module description, maximum number of characters/words/pages of the paper	60 pages/ 2.400 characters per page 144.000 characters total		
Number of characters/words/pages <i>(Standard page = 2400 characters including Tables and Figures, excluding References, Appendices, Front Page, Table of Contents)</i>	Characters: 141.188 Pages: 54 (Including Abstract)		
Supervisor (project/thesis):	Jochen Reiner		



Signature and Date

Which pricing models are the most effective for AI agents in today's technology landscape?

By

Raul Moriana Sigel

Aalborg Business School
MSc Marketing and Sales
2025

Acknowledgment

I would like to thank several people who made this thesis possible.

First, I thank Morten Krarup Kristensen for introducing me to PricingSaaS.

I also thank John Kotowski and Rob Litters, the founders of PricingSaaS, for sharing their expertise, insights and enabling me to create and distribute the survey that forms the foundation of this research.

I am grateful to my supervisor, Jochen Reiner, for his guidance and feedback throughout this process.

To my friends and flatmates, thank you for your support during the long hours of writing and research.

To my family, thank you for your encouragement and belief in me throughout my studies.

And to my girlfriend, thank you for your patience, love, and support. This would not have been possible without you.

Completing this thesis marks the end of an important chapter in my academic journey, and I am grateful to everyone who helped me reach this milestone.

A handwritten signature in black ink, appearing to be 'JRK' or similar, enclosed within a large, stylized loop.

Table of Content

Table of Content.....	4
List of Figures and Tables.....	7
List of Abbreviations.....	8
Abstract.....	10
1. Introduction.....	1
1.1 Background.....	1
1.2 Problem Statement.....	2
1.3 Research Scope & limitations.....	3
1.4 Research Questions.....	4
1.5 Disposition of Thesis.....	5
1.6 Theoretical Perspective.....	6
2. Literature Review.....	8
2.1 State - of - the - Art.....	8
2.1.1 Pricing of AIaaS.....	8
2.1.2 Pricing AI agents.....	9
2.2 The Rise of Agentic AI.....	11
2.2.1 Artificial Intelligence.....	11
2.2.2 AIaaS.....	11
2.2.3 AI Agents.....	11
2.2.3.1 Technical description of AI Agents.....	12
3.The Economics of AI Agents.....	13
3.1 Cost structure breakdown.....	13
3.1.1 LLM Models.....	13
3.1.2 Tokenization.....	14
3.1.3 Understanding the Factors Influencing the Costs of LLM Model API Calls.....	15
3.1.4 LLM Pricing per 1 million Tokens.....	15
3.1.5 Price vs Intelligence.....	16
3.1.6 Cost breakdown example of an AI agent.....	17
4. Pricing Models.....	19
4.1 Established SaaS pricing models.....	19
4.1.2 Nine Building Blocks of SaaS Revenue.....	19
4.1.2.1 Transactional one-off fees.....	20
4.1.2.2 Flat recurring fees.....	20
4.1.2.3 Metric-based fees.....	20
4.2 Limitations of SaaS pricing models for AIaaS.....	22
4.3 Pricing models for AIaaS.....	22
4.3.1 Framework for AI Agent Pricing Models.....	23
4.3.1.1 Per Agent Action/ Usage - based pricing mode.....	23
4.3.1.2 Per Agent Outcome/ Outcome - based pricing model.....	24
4.3.1.3 Per Agent/ FTE Replacement pricing model.....	24
4.3.1.4 The per workflow pricing model.....	25
4.3.1.5 Hybrid models.....	25

5. Methodology.....	26
5.1 Research Design.....	26
5.2 Data Collection Methods.....	27
5.2.1 Experimental Within-Subject Design.....	28
5.2.2 Observational Market Data.....	30
5.2.3 Variables overview.....	31
5.3 Data Analysis.....	32
5.3.1 Overview.....	32
5.3.2 Data Preparation.....	33
5.3.3 Statistical Analysis.....	33
5.4 Justification and Evaluation of Methodology.....	34
5.4.1 Internal Validity.....	34
5.4.2 External Validity.....	35
5.4.3 Reliability and Consistency.....	35
5.4.4 Objectivity.....	35
5.4.5 Ethical Integrity.....	36
6. Analysis & Discussion.....	36
6.1 Introduction.....	36
6.2 Data Preparation.....	37
6.3 Descriptive Statistics.....	38
6.3.1 Sample Demographics.....	38
6.4 Pricing Model Preferences.....	39
6.4.1 MaxDiff for Pricing Models and Attributes.....	39
6.4.2 Statistical Tests of Preference Differences.....	40
6.4.3 Coding of themes in open-ended explanations for pricing model choice.....	42
6.5 Predictors of Flat-Rate Preference.....	43
6.5.1 Binary Logistic Regression Analysis.....	43
6.5.2 Fisher's Exact Tests and Cramér's V.....	45
6.6 Market vs. Survey Gap Analysis.....	46
6.7 Summary of findings.....	46
7. Discussion.....	47
7.1 Theoretical Interpretation.....	47
8. Conclusion.....	49
8.1 Main findings.....	49
8.2 Practical and Managerial Implications.....	49
8.3 Limitations and Validity Threats.....	50
8.4 Future Research Directions.....	51
8.5 Conclusion.....	52
9. References.....	54
10. Appendix.....	62
Appendix 1: TAM Model (Davis, 1989).....	62
Appendix 2: Breakdown of costs for training and experiments (Cottier et al., 2024).....	62
Appendix 3: Amortized hardware and energy cost to train frontier AI models over time (Cottier et al., 2024).....	63

Appendix 4: Tokenization at OpenAI (OpenAI, n.d.a).....	64
Appendix 5: LLM inference prices have fallen 9x to 900x/year, depending on task (Cottier et al., 2025).....	65
Appendix 6: Example of an Multi-Agent System (MAS) Minkovski (2024).....	66
Appendix 7: Academic SaaS pricing frameworks comparison (Saltan & Smolander, 2019).....	67
Appendix 8: Parameters of pricing models for software products (Lehmann & Buxmann, 2009).....	69
Appendix 9: The Pricing Strategy Guideline Framework for SaaS vendors (Spruit & Abdat, 2012)	70
Appendix 10: The six pillars of a price model (Frohmann, 2018).....	71
Appendix 11: Decision Framework Pricing Model, Medina, 2025.....	72
Appendix 12: Artisan Add (Carmichael-Jack, J., 2024).....	73
Appendix 13: Pricing Model Distribution (Authors own contribution).....	73
Appendix 14: Pricing Modality Distribution per Category (Authors own contribution).....	74
Appendix 15: Tariff Distribution (Authors own contribution).....	74
Appendix 16: Top 10 Pricing Metric Distribution (Authors own contribution).....	75
Appendix 17: Threshold Distribution (Authors own contribution).....	75
Appendix 18: Free Plan Availability (Authors own contribution).....	76
Appendix 19: Typeform Survey (Authors own contribution).....	77
Appendix 20: Survey Results (Authors own contribution).....	83
Appendix 21: Distribution of Job Roles (Author's own contribution).....	94
Appendix 22: Distribution of Firm Size (Author's own contribution).....	94
Appendix 23: Distribution Prior AI Agent Purchasing Experience (Author's own contribution).....	95
Appendix 24: Distribution of Company Sectors (Author's own contribution).....	95
Appendix 25: Pricing Model - MaxDiff Net Scores (Author's own contribution).....	96
Appendix 26: Pricing Attributes - MaxDiff Net Scores (Author's own contribution).....	96
Appendix 27: Consumer preference vs market prevalence (Created in R).....	97
Appendix 28: Distribution Pricing Model Preference From Survey (Author's own contribution).....	97
Appendix 29: Cochran's Q Test (Created in R).....	98
Appendix 30: Pairwise McNemar Test (Flat vs Other Models) (Created in R).....	98
Appendix 31: Binary logistic Regression (Flat vs Other) (Created in R).....	99
Appendix 32: Fisher's Test - Heat Map (Created in R).....	100
Appendix 33: Cramer's V - Heat Map (Created in R).....	100
Appendix 34: Fisher's Test & Cramers V - Heat Map (Created in R).....	100
Appendix 35: Open-ended explanations for pricing model choice.....	101
Appendix 36: Philosophy of Science.....	102

List of Figures and Tables

Figure 1: Key components of advanced AI agents

Figure 2: Input and Output Token Prices by selected AI Labs

Figure 3: Intelligence vs Prices by selected AI Labs

Figure 4: Cost breakdown example of an AI agent

Figure 5: The Nine building blocks of SaaS revenue

Figure 6: AI Agent Pricing Model Framework

Figure 7: Sample Size Calculation (1)

Figure 8: Sample Size Calculation (2)

Table 1: MaxDiff Pricing Models

Table 2: MaxDiff Attributes

Table 3: Pair-wise McNemar Tests Comparing Flat-Rate to Alternative Pricing Models (Holm-adjusted p-values)

Table 4: Pair-wise McNemar Tests Comparing Pricing Attributes (Holm-adjusted p-values)

Table 5: Binary-logistic regression predicting preference for the flat-rate model (1 = chose flat as “most preferred”, 0 = did not)

Table 6: Fisher's Exact Test

Table 7: Effect size (Cramér's V)

List of Abbreviations

AI: Artificial Intelligence

TAM: Technology Acceptance Model

SaaS: Software as a Service

AlaaS: Artificial Intelligence as a Service

API: Application Programming Interfaces

LLM: Large Language Model

LMM: Large Multimodal Models

RAG: Retrieval-Augmented Generation

AWS: Amazon Web Services

AGI: Artificial General Intelligence

ROI: Return of Investment

APS: Advance Payment System

RRRs: Relative Risk Ratios

CoT: Chain of Thought

3PT: Three-Part Tariff

ARR: Annual Recurring Revenue

B2B: Business to Business

B2C: Business to Consumer

CFO: Chief Financial Officer

FTE: Full Time Employee

GPU: Graphic Processor Unit

KPI: Key Performance Indicator

MAS: Multi Agent Systems

OECD: Organisation for Economic Co-operation and Development

PEOU: Perceived Ease of Use

PU: Perceived Usefulness

PwC: PricewaterhouseCoopers

SDR: Sales Development Representative

Abstract

The rapid emergence of agentic AI presents significant monetization challenges, with many vendors defaulting to legacy SaaS pricing models, which prove ill-suited for AI's unique cost structures and value propositions. This thesis investigates which pricing models are most effective, defined through the lens of consumer acceptance (Frohmman, 2018), for AI agents and assesses alignment with current market offerings. Grounded in the Technology Acceptance Model (TAM) (Davis, 1989) and Flat-Rate Bias theory (Lambrecht & Skiera, 2006), this study posits that pricing models delivering greater cost predictability, enhancing price transparency, reducing perceived financial risk and resonating with users' perceived value will achieve higher acceptance.

A quantitative approach was employed, combining a within-subject survey experiment with 53 business professionals and a structured market audit of 101 AI agent pricing pages. Statistical analyses included Cochran's Q and McNemar tests to evaluate preference differences and binary logistic regression followed by Fisher's test and Cramer's V to determine a correlation between pricing attribute prioritization and pricing model preference.

Results reveal a strong flat-rate bias: 43.4% of respondents preferred flat-rate subscriptions (net utility +14), significantly more than usage-based (-17) and license-plus-overage (-12) models (Cochran's $Q(4) = 45.75$, $p < .001$). Outcome-based (24.5%, net utility 0) and credit-based plans (22.6%, net utility +7) were the next most favored. Cost predictability was the most important attribute (net utility +23). A logistic regression, however, demonstrates no significant predictors of flat preference: coefficients ranged from $\beta = 1.31$ (OR = 3.7, $p = 0.78$) for predictability, $\beta = 6.0$ (OR ≈ 400 , $p = 0.21$) for simplicity, $\beta = 0.45$ (OR ≈ 1.57 , $p = 0.47$) for simplicity each with wide confidence intervals arising from quasi-separation and low power. Fisher's tests confirmed the null result ($p = 0.78-1.00$) with only moderate effect sizes for transparency and simplicity ($V \approx 0.23-0.24$). A strong market misalignment was identified: 66% of audited vendors use credit-based models, and 22% use usage/overage, while only 2% offer flat-rate and <1% outcome-based pricing.

This study concludes that flat-rate and outcome-based pricing models are most effective for AI agents based on consumer acceptance, with 43.4% of participants preferring flat-rate models for their cost predictability. The research reveals a significant market misalignment, as only 2% of AI agent vendors currently offer flat-rate subscriptions while 66% use credit-based systems, despite strong consumer preference for predictable pricing structures. The findings support Flat-Rate Bias theory in the AI agent domain and demonstrate that freemium models would positively influence sign-ups for 86.8% of users. This research underscores the necessity for an acceptance-driven approach to pricing models, suggesting vendors should strategically integrate flat-rate and outcome-based plans to better meet customer demands and achieve more effective monetization in the rapidly evolving AI agent landscape.

1. Introduction

1.1 Background

“We marveled at our own magnificence as we gave birth to AI” a notable quote by the Character Morpheus from the Movie The Matrix (1999). Today 26 years fast forward it is reality. AI agents, the newest development of AI, pursue goals and complete tasks on behalf of users. They show reasoning, planning, memory and have a level of autonomy to make decisions, learn, and adapt, transforming industries from healthcare to finance (Google Cloud (n.d.)).

As artificial intelligence (AI) reshapes the technological landscape and disrupts the world we know today, the economic stakes for businesses are high. According to a study by PwC, AI could contribute up to \$15.7 trillion to the global economy output in 2030 and the Global AI market size reached \$184 Billion in 2024 (PwC. (n.d), (Statista,n.d.)). Investment in generative AI increased nearly eightfold from 2022 to 2024, totaling \$25.2 billion (Maslej et al., 2024). At the center of this transformation lies the critical challenge of effectively monetizing AI-powered products, particularly with the recent rise of AI agents. In today's dynamic market, traditional pricing models are being questioned and reinvented, inviting us to explore:

Which pricing models are most effective for monetizing AI agents in today's technology landscape?

Pricing is a crucial factor, as shown by a significant observation that a 1% rise in price can result in an 11% increase in operating profit. This illustrates that changes in pricing greatly affect both revenue and profit margins, whereas lowering prices can diminish both revenue and profit at the same time (Kohli & Suri, 2011). However, simultaneously pricing has been seen as the most difficult element of the marketing mix to manage effectively, as it requires a deep understanding of customer value perception, competitive dynamics and costs, to name a few (Simon, 1992). This amongst other reasons has further led to pricing being widely neglected by managers and academics, with fewer than 5% of the Fortune 500 companies including a full-time function dedicated to pricing (Hinterhuber and Liozu 2012), (Smit & Niekerk, 2014), (Kienzler & Kowalkowski, 2014)

In light of the rapid advancements in AI and the transformative impact of AI agents across various industries, understanding how to effectively price these technologies is fundamental for all companies seeking to successfully leverage AI agents. Effectiveness, defined by Drucker (1967) as “doing the right thing,” is operationalized in this thesis as consumer acceptance, because “by far the most important explanatory factor for the failure of a price model is the lack of customer acceptance” (Frohmann, 2018). Therefore, the right thing equals a pricing model that maximizes customer acceptance. To determine which pricing model is the most effective, this thesis aims to investigate consumer preferences for AI agent pricing models. Grounded in Flat-Rate Bias theory (Lambrecht & Skiera, 2006) and supported by the Technology Acceptance Model (TAM) and its emphasis on perceived

usefulness and ease of use (Davis, 1989), the study explores both psychological pricing effects and the broader behavioral drivers of technology adoption. A deductive, quantitative approach is employed, combining a within-subjects survey experiment with a structured market audit of existing AI agent pricing models (Creswell, 2014). By aligning theoretical expectations with empirical data, the research contributes to a deeper understanding of how pricing influences user preference, an essential factor in the successful monetization of AI agents.

1.2 Problem Statement

Numerous new startups are emerging every day, with 10,621 AI companies founded between 2013 and 2023, all eager to capitalize on the advancements in artificial intelligence (AI) (Stanford University, 2024). However, as the software market evolves from Software as a Service (SaaS) to Artificial Intelligence as a Service (AIaaS) and now to AI agents, businesses face a unique challenge in effectively monetizing these advanced technologies, shown by only 42 % of companies offering AI products or features are actually monetizing them (Yamase, S., 2025). According to a multitude of scholarly articles, pricing frequently remains neglected, undervalued, and lacks adequate emphasis overall by managerial personnel and executives (Simon & Fassnacht, 2019; Hinterhuber, 2003; Shipley & Jobber, 2001). Ramanujam, pricing expert from Simon Kutcher, cautions that organizations that neglect the importance of pricing are likely to be “leaving money on the table.” (Bashir, 2024). This oversight can profoundly affect the long-term sustainability and profitability of businesses.

Conventional SaaS pricing models, whether seat-based or flat-rate, are designed on the assumptions that marginal costs approach zero after deployment. However, autonomous AI agents disrupt this premise, as every usage incurs requests to API calls to LLM Models, which lead to variable expenses that scale with usage. Consequently, flat subscriptions expose vendors to cost overruns, seat licences overlook the labour AI agents displace, and pure usage - based pricing models shift cost volatility to customers, risking lack of adoption due to challenging budget and forecasting predictability. Although the SaaS literature provides useful foundations, the field still lacks empirically validated pricing models which simultaneously account for an AI agent’s variable costs, its value contribution, and end-users’ behavioural preferences.

This lack of evidence for effective pricing models not only hinders revenue generation but also risks diminishing profit margins from the launch, thereby limiting the growth potential of AI agent vendors. Consequently, there is an urgent need for research that identifies and evaluates pricing models that align with the unique characteristics and demands of AI agents, bridging the divide between conventional SaaS pricing models and the evolving AI market landscape.

1.3 Research Scope & limitations

This study focuses on evaluating consumer preferences for different AI agent pricing models through a quantitative research design. It employs an experimental within-subjects survey design alongside a structured observational audit of public pricing data. In the survey experiment, cross-sectional in nature, each participant evaluated multiple pricing models for a hypothetical AI agent, allowing direct comparison of how each pricing model affects user preference. Concurrently, an audit of real-world AI agent pricing pages was conducted to document prevailing pricing models in the market. By combining these methods, the research covers both perceptual consumer data and actual industry practices, providing a bounded yet comprehensive view of AI agent pricing models.

The pricing models examined include credit-based plans, flat-rate subscriptions, usage-based, outcome-based, and hybrid models that mix fixed fees with usage components, such as a license -based model with allowance and overage fee. These pricing models largely reflect common SaaS pricing strategies identified in prior literature (Saltan & Smolander, 2021). Within the scope of the experiment, all participants consider each of these models for the same AI agent under the assumptions that the costs are the same, ensuring that differences in preference can be attributed to the pricing model itself rather than differing contexts. The study specifically investigates how these pricing models impact consumer acceptance of AI agents. It does so using extensively established theoretical frameworks, such as the Technology Acceptance Model (TAM) and the concept of Flat-Rate Bias. TAM provides a framework for understanding acceptance by positing that external factors, in this case, the pricing model, shape users' perceived usefulness and ease of use of a technology, which in turn drive their willingness to adopt it (Davis, 1989). Flat-Rate Bias theory, as formulated by Lambrecht and Skiera (2006), suggests that consumers have a systematic preference for flat-rate or all-inclusive pricing due to the simplicity and predictability it offers, even when pay-per-use options might be economically cheaper. By integrating TAM and Flat-Rate Bias, the study's scope is defined to evaluate not only which pricing model is preferred, but also why, examining whether the appeal of simpler, flat-rate plans can be explained by the proposed psychological factors in the Flat-Rate Bias theory. In summary, the scope is limited to AI agent pricing, focusing on five specific pricing models and assessing consumer preference and acceptance drivers in the context of a controlled survey and current market offerings.

In doing so, several limitations stem from the research design and sample. First, the study relied on a purposive convenience sample drawn from the PricingSaaS community, without stratified random sampling. This means participants were largely individuals already interested or involved in software pricing, which may not represent the broader population of AI agent users. The sample size of 53 respondents, while sufficient for the chosen statistical analyses, is moderate. A larger or more diverse sample might reveal additional nuances. Because the survey was cross-sectional, with data collected at a single point in time, it captures preferences only as a snapshot and cannot account for how consumer attitudes might change over time or with extended use of AI agents. Additionally, all data on preferences

were self-reported in a hypothetical scenario. Such self-reported intentions can be subject to biases, participants might overestimate their willingness to adopt certain models or respond in ways they believe are expected. Despite efforts to mitigate biases, e.g., randomizing model presentation order and standardizing descriptions, the experimental setting cannot perfectly replicate real purchasing behavior. Therefore, there is a risk that stated preferences do not fully align with actual choices under real financial commitments. These methodological limitations suggest caution in interpreting the findings as universally applicable, the results are most valid for the specific sample and conditions studied.

This research also faces limitations related to theoretical and contextual constraints. The domain of AI agent pricing is an emerging area with little existing academic literature. As noted in the literature review, to my current knowledge no peer-reviewed studies have yet examined pricing models for AI agents, forcing this study to draw on adjacent theories and industry reports. The theoretical frameworks applied, TAM and Flat-Rate Bias, originate from broader technology acceptance and consumer pricing research, not from prior studies on AI agent pricing. While this provides a necessary foundation, it means the interpretations are somewhat extrapolated from related contexts to this new domain (Davis, 1989; Lambrecht & Skiera, 2006). Similarly, the taxonomy of pricing models tested was largely based on common SaaS pricing models (Saltan & Smolander, 2021) and may not encompass all emerging pricing innovations unique to AI agents. Another limitation is the reliance on publicly available pricing data for the observational audit. The market review included only companies that openly publish their pricing, thus, any AI agent vendors using confidential or custom pricing, e.g., enterprise negotiated plans not listed online, were excluded. This could skew the observed prevalence of certain models and means some pricing models in practice might not be captured in the analysis. Furthermore, the findings are context specific and may not generalize to other industries or user segments beyond the study's focus. The sample consisted mainly of the PricingSaaS community which can be characterized as tech-savvy individuals familiar with AI tools, so their preferences might differ from those of general consumers or users in different demographics or organizational roles. Likewise, the AI agent scenario for the survey, a customer support agent, and the time frame of data collection, May 2025, set boundaries on the applicability of results. In essence, while the study offers insights into consumer preferences for AI agent pricing models, its conclusions are best interpreted as indicative for similar contexts and populations. Broad generalization should be done cautiously, and further research with varied user groups and in other domains is encouraged to validate and extend these findings.

1.4 Research Questions

The answer to the following main research question aims to fulfill the purpose of the study.

Which Pricing Models are most effective for AI agents in today's technology landscape?

To answer the main research question, the following sub-research questions were formulated, which determine the structure of the literature review and methodology employed.

- 1) What are AI Agents and what is the nature of their cost structure?
- 2) Which are the limitations of SaaS pricing models for monetizing AI agents?
- 3) What makes a pricing model for AI agents effective?
- 4) Which pricing models exist for AI agents ?
- 5) If effectiveness is measured through customer acceptance, how can customer preference be explained?

1.5 Disposition of Thesis

This structured disposition aims to explain the layout of the thesis.

Introduction

This chapter sets the context for the study, states the research problem, objectives, questions, theoretical perspective and outlines the importance of investigating pricing models for AI agents.

Literature Review

This chapter reviews existing research on SaaS, AIaaS, and AI Agent pricing, identifies key gaps, and shapes the study's hypotheses.

Methodology

This chapter explains the deductive quantitative design, describes the within subjects survey and the structured market audit, and details sampling, instruments and statistical procedures.

Results & Analysis

Empirical findings from both datasets are presented and examined, indicating pricing model preferences and how those models appear in current market practice.

Discussion

Results are interpreted through Technology Acceptance Model and flat rate bias theory, their implications are linked to prior research, and study limitations are acknowledged.

Actionable guidance is offered to AI agents vendors on selecting and implementing AI agent pricing models, together with priorities for future academic work.

Conclusion

The chapter summarises key insights, answers the research question, and highlights the contribution of the study to theory and practice in AI monetisation.

References

A comprehensive list of references will be provided, encompassing academic literature, industry reports, survey results and relevant online resources that informed the research.

Appendix

The appendix will include supplementary materials such as survey instruments, detailed data analyses, and additional resources that support the findings of the thesis.

1.6 Theoretical Perspective

The Technology Acceptance Model (TAM), first introduced by Davis (1989), is a foundational framework for explaining how users come to accept and use new technologies (See Appendix1). The model states two core concepts: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU), which largely determine a user's behavioral intention to adopt a system (Davis, 1989). Over the years, TAM has become one of the most influential and widely applied models in information systems research, given its simplicity and strong explanatory power in diverse contexts (Musa et al., 2024; Singh, 2024; Ma & Liu, 2011; Chuttur M.Y. , 2009).

Perceived Usefulness and Perceived Ease of Use are the central components of the TAM. Perceived Usefulness (PU) is defined as the degree to which a person believes that using a given technology will improve their job performance or overall productivity (Davis, 1989). In parallel, Perceived Ease of Use (PEOU) refers to the degree to which a person believes that using the technology will be free of effort (Davis, 1989). These perceptions influence users' attitudes and willingness to adopt the system. Empirical studies have consistently confirmed the importance of these two constructs, PU has a strong positive effect on a user's intention to adopt technology, and PEOU contributes both directly and indirectly by enhancing PU to adoption decisions (Ma & Liu, 2011; Singh, 2024; Musa et al., 2024). For instance, a meta-analysis by Ma and Liu (2011) found that PU was a significantly stronger predictor of technology acceptance than PEOU, which often gets its influence through PU. Overall, when users perceive a system as useful and easy to use, they are far more likely to adopt it (Davis, 1989; Chuttur, 2009).

Importantly, TAM recognizes that external factors can influence perceptions of usefulness and ease of use (Musa et al., 2024). Variables such as system features, user training, or pricing models indirectly affect PU and PEOU (Singh, 2024). For example, a simple interface or favorable pricing can enhance perceived value and usability, while complex designs or pricing structures may reduce them. Studies have shown that simplifying technical complexity through design and training increases user acceptance (Singh, 2024).

In digital services, pricing acts as a critical external factor. Transparent or flat-rate pricing can improve perceived value and ease of use, whereas complex metered models may deter users. Thus, pricing perception plays a significant role in technology acceptance. How external

factors like pricing models shape PU and PEOU helps explain consumer preference for specific pricing models and highlights the relevance of the concept of Flat-Rate Bias, where consumers favor flat-rate payment plans for their predictability and simplicity. In addition to the TAM framework, this study draws on the Flat-Rate Bias theory developed by Anja Lambrecht and Bernd Skiera (2006), which has been applied to study consumer behavior in selecting pricing plans across various services, such as telecommunications, internet services, and digital subscriptions.

This theory indicates that consumers often prefer flat-rate pricing plans over pay-per-use options, even when the latter would be more cost-effective based on their actual usage. Lambrecht and Skiera identified four main psychological factors contributing to this bias:

- *Insurance Effect:* Consumers opt for flat rates to avoid the risk of unexpectedly high charges, valuing the predictability of costs.
- *Taximeter Effect:* The discomfort associated with watching costs accumulate, as with a taxi meter, leads consumers to prefer flat rates, which decouples usage from immediate financial implications.
- *Overestimation Effect:* Consumers tend to overestimate their future usage, leading them to believe that a flat rate will be more economical.
- *Convenience Effect:* Consumers might believe that choosing among optional tariffs is inconvenient and therefore might try to avoid the effort of identifying alternative tariffs and calculating the respective expected billing rate. Therefore choosing a flat rate from the convenience of not needing to search for the least costly tariff.

In line with Creswell's (2014) description of the deductive approach typically used in quantitative research, this study applies both the Technology Acceptance Model (TAM) and the Flat-Rate Bias theory to explain consumer preferences for pricing models. Based on the Technology Acceptance Model (TAM), external factors such as the presented pricing model are expected to influence consumers' perceived usefulness (PU) and perceived ease of use (PEOU) (Davis, 1989). Pricing models perceived as simpler and more predictable, such as flat-rate options, are likely to enhance PU and PEOU, thereby increasing consumer preference and acceptance. Furthermore, drawing from the Flat-Rate Bias theory, I would expect my independent variable, the type of pricing model presented (credit - based, outcome - based, flat-rate, usage-based and flat fee + free allowance + overage) to influence or explain the dependent variable, consumer preference for pricing models, because the psychological comfort and perceived value associated with Flat-Rate pricing may lead consumers to favor it over other models, regardless of actual usage patterns or cost efficiency (Lambrecht & Skiera, 2006). Derived from the theory, the main hypothesis that will be tested is as follows

H₁(1): *Flat-rate pricing models will be most preferred, and usage-based pricing least preferred, reflecting a flat-rate bias (Lambrecht & Skiera, 2006)*

To empirically test the abstract concept of consumer preference, the dependent variables are operationalized:

Consumer Preference: Measured via discrete choice experiments where participants select their most preferred and least preferred pricing model in a hypothetical scenario.

Bias Drivers: To further investigate the psychological mechanisms underlying flat-rate bias, respondents are asked to select the most and least important variable when considering pricing models. Cost predictability, Cost transparency, Fairness and Simplicity are included to measure participants' agreement with key cognitive and emotional drivers of pricing preference:

- 1) Insurance Effect: *Cost predictability, Fairness*
- 2) Taximeter Effect: *Cost Transparency*
- 3) Overestimation Effect: *Cost predictability, Fairness*
- 4) Convenience Effect: *Simplicity*

These variables are analyzed in relation to participants' pricing model choice Flat-rate to explore whether stronger agreement with specific bias dimensions predicts a preference for flat-rate pricing model. To measure these variables effectively and to obtain scores, an experimental within-subjects design was conducted. The methodology is further discussed in chapter 5.

2. Literature Review

2.1 State - of - the - Art

This State - of - the - Art review aims to give a comprehensive yet brief overview of the current state of research surrounding the field of pricing for AIaaS and AI agents.

2.1.1 Pricing of AIaaS

In March 2023, GPT - 4, the Large Language Model (LLM) developed by OpenAI unveiled its API, thereby facilitating companies to develop what are referred to as "AI Wrappers," or categorized as AI-as-a-Service (AIaaS), which constitutes a Software-as-a-Service (SaaS) product that leverages the artificial intelligence capabilities inherent in LLM models through the utilization of APIs (Alvaro, 2024), (Microsoft Azure. (n.d.)). A very limited number of scholarly articles have been published that particularly concentrate on the pricing aspects of AIaaS, only a few papers were found in the literature research efforts.

Further, the papers addressing AIaaS focus on the pricing approach from the perspective of LLM Model providers, however this study focuses on the customers of LLM Model providers, AIaaS and AI Agents using the LLM models providers API's. Gao et al. (2024) emphasize that AI services exhibit dynamic value trajectories due to user learning and model decay and propose an adaptive pricing algorithm that enables LLM Model providers to maximize revenue while maintaining a high demand rate for AI model APIs. Hajipour et al., (2023) proposes a formula to calculate the AIaaS product price per API call based on hardware and software cost, but doesn't offer practical recommendation for pricing model application. Bergemann et al. (2025) proposes an economic framework to analyze the optimal pricing and product design of Large Language Models (LLM) and finds that optimal mechanisms can be implemented through menus of two-part tariffs, with higher markups for more intensive users.

From the AIaaS vendors point of view Li et al., (2022) proposes a model for determining optimal pricing of AI-enabled products that maximizes the manufacturer's profitability and found that after the manufacturer launches AI-enabled products, it always needs to reprice regular products to maximize profitability. The study, however, only considers the impact of technology level on product demand. A study by Mahmood (2024) explores how companies can effectively launch and price generative AI tools. By modeling the pricing as a strategic game between two firms, the authors show that latecomers with market knowledge can always be cost-effective on at least one task, while first-movers must carefully set prices to maintain an advantage. If tasks are too similar, however, early movers risk losing their edge altogether (Mahmood, 2024). Agrawal et al. 2018 finds that AI predictions complement, rather than replace, human judgment by supplying accurate state information. AI lets decision-makers accumulate experience that uncovers their own utility functions and raises expected payoffs when prediction and judgment are used together. However, because that experience often reveals a dominant action, causing about half of users to cancel the subscription, therefore they conclude that a vendor maximizes its long-run profit by charging a single low, fully inclusive subscription price that keeps all users during the learning phase even though it leaves some value unmonetized.

2.1.2 Pricing AI agents

The emergence of AI Agents really started back in May 2023, when the voyager paper "VOYAGER: An Open-Ended Embodied Agent with Large Language Models" got published, which introduces the new development of "the first LLM-powered embodied lifelong learning agent in Minecraft that continuously explores the world, acquires diverse skills, and makes novel discoveries without human intervention" (Wang et al., 2023). In the comprehensive literature review efforts, no significant academic papers were found related specifically to the field of pricing of AI Agents, however a few online articles and blogs address this emerging topic. This gap in academic research highlights a significant opportunity for scholars and practitioners to explore the implications of pricing of AI agents.

According to a study by Simon & Kutcher only 42 % of companies offering AI products or features are actually monetizing them and propose a 4 P's framework for monetization implementation for AI, but there is currently no standard framework or dominant design for pricing AI agents (Yamase, S., 2025). A variety of pricing models have emerged, which tends to be learned by trial and error and anecdote (Sharma, 2025). The most common models to date include license-based pricing and usage-based pricing including performance and outcome-based pricing models. New pricing models such as Labor Replacement Pricing, Per-Execution Pricing, Per-Conversation Pricing and Agentic Pricing are emerging to address the limitations of per-seat SaaS subscriptions and offer a new way to price only for what is consumed (Sharma, 2025). However, consumption-based pricing models pose significant challenges for CFOs, who must plan annual budgets while managing price volatility and the risk of overspending (Gross, G., 2024). In many cases, providers are experimenting with hybrid combinations rather than relying on a single model, reflecting the early state of AI agent monetization (Thammineni, P., 2025), Medina (2025).

To sum up there has not been any significant academic research in the field of pricing AIaaS and AI agents besides a few theoretical models for LLM providers and the current knowledge for pricing AI agents is limited to anecdotal findings from online sources such as blogs, websites, social media posts and newsletters. Therefore there is a significant research gap in the field of pricing AI agents. Researchers should start by looking at the current state of research in the field of SaaS pricing and identify what is applicable to AIaaS and AI agents and what isn't. Research is warranted to understand the influence of various pricing models on customer adoption rates and overall profitability. There is a need for in-depth analysis of the cost implications associated with AI agents, especially considering the potential commoditization of AI technology and the trend of significantly reduced computing costs for large language models (LLMs), as seen e.g. by the development of Deepseek. Further, studies on pricing metrics will be needed as usage - based and outcome - based pricing models will most likely prevail in AIaaS and for AI agents, in order to determine what constitutes a unit of consumption for an AI agent and how it can be standardized across vendors. Additionally existing frameworks for pricing strategies have to be evaluated in the context of AI, a market which is characterised by strong competition and fast technological advancements. Moreover, establishing more effective methods to quantify an AI agent's value and return on investment (ROI) would be advantageous for both vendors and customers alike. Besides theoretical models, empirical and longitude studies are necessary to provide practical recommendations for businesses.

This study seeks to address this significant research gap in the field of pricing AI agents by critically examining the limitations of existing SaaS pricing models. It aims to explore pricing models applicable for AI agents and, through empirical research, assess consumer preferences regarding these models. To explain the underlying reasons for consumer preference the Technology Acceptance Model (TAM) and Flat-Rate Bias theory are applied. The findings will result in practical recommendations for AI agent vendors on effective monetization of their agents.

2.2 The Rise of Agentic AI

This chapter seeks to offer a concise overview of artificial intelligence (AI) and in particular AI agents. It will explore essential definitions and technical descriptions, while avoiding overly in depth technological details.

2.2.1 Artificial Intelligence

Artificial Intelligence (AI) consists of the two words artificial and intelligence. Looked at separately artificial means: “not natural or real : made, produced, or done to seem like something natural” (Britannica Dictionary, (n.d.a)) and intelligence refers to “the ability to learn or understand things or to deal with new or difficult situations” (Britannica Dictionary, (n.d.b)). Together, these terms describe AI as a construct designed to mimic human cognitive functions and can be defined as: “An area of computer science that deals with giving machines the ability to seem like they have human intelligence” and “the power of a machine to copy intelligent human behavior” (Britannica Dictionary, n.d.c). John McCarthy, a pioneer in the field of artificial intelligence, was the first to define the term "artificial intelligence." He described it as “The goal of AI is to develop machines that behave as though they were intelligent” (Ertel, 2018).

2.2.2 AIaaS

AIaaS can be defined “as cloud-based systems providing on - demand services to organizations and individuals to deploy, develop, train, and manage AI models (Lins et al., 2021). The term ”aaS” is used to define services that are offered via the cloud infrastructure on a subscription basis (Syed et al., 2025). Consequently, AI as a Service (AIaaS) can be defined as services that access pre-trained LLM models through the cloud by API queries to receive inferences and are offered on subscription basis to end-users. AIaaS comprises the characteristics of complexity abstraction, automation, customizability, and inherited cloud characteristics (Lins et al., 2021). Based on the definition of AIaaS, this paper defines AI agents as a specialized subset of AIaaS. These agents are autonomous systems that utilize AIaaS to carry out tasks, make informed decisions and engage with users, agents or other systems.

2.2.3 AI Agents

A key challenge in AI discourse is the inconsistent use of the term "agent." Traditionally, an "agent" is any system that perceives its environment and acts, which can include simple devices, e.g. Thermostats. Therefore, AI agents can broadly be defined as ”an entity that senses percepts (sound, text, image, pressure etc.) using sensors and responds (using effectors) to its environment” (Alvarez & Jurgens, 2024). The term "agent" in the field of AI can more narrowly be defined as “Agentic AI includes the class of autonomous AI systems that undertake to finish a set of complex tasks that span over long periods of time without human supervision. It learns context and makes decisions” (Acharya et al., 2025). AI agents typically possess two key elements: autonomy, the ability to operate independently without

constant human oversight, and authority, the granted permissions to perform actions within defined parameters to achieve specific objectives and influence their environment (Alvarez & Jurgens, 2024). Agents transmit input prompts as requests to large language models (LLMs), which generate responses that either specify final actions or provide further instructions. To execute actions, agents use tools for local computations, external server requests or search engines. Tool outputs are forwarded back to the agent, informing subsequent actions. This tool invocation enables agents to interact with the real world. Since agents depend on LLMs to interpret input, process feedback, and generate actions, LLMs form their essential backbone (He et al., 2024). In essence, Agentic AI refers to systems capable of making autonomous decisions and acting towards specific goals with minimal human intervention (Acharya et al., 2025).

2.2.3.1 Technical description of AI Agents

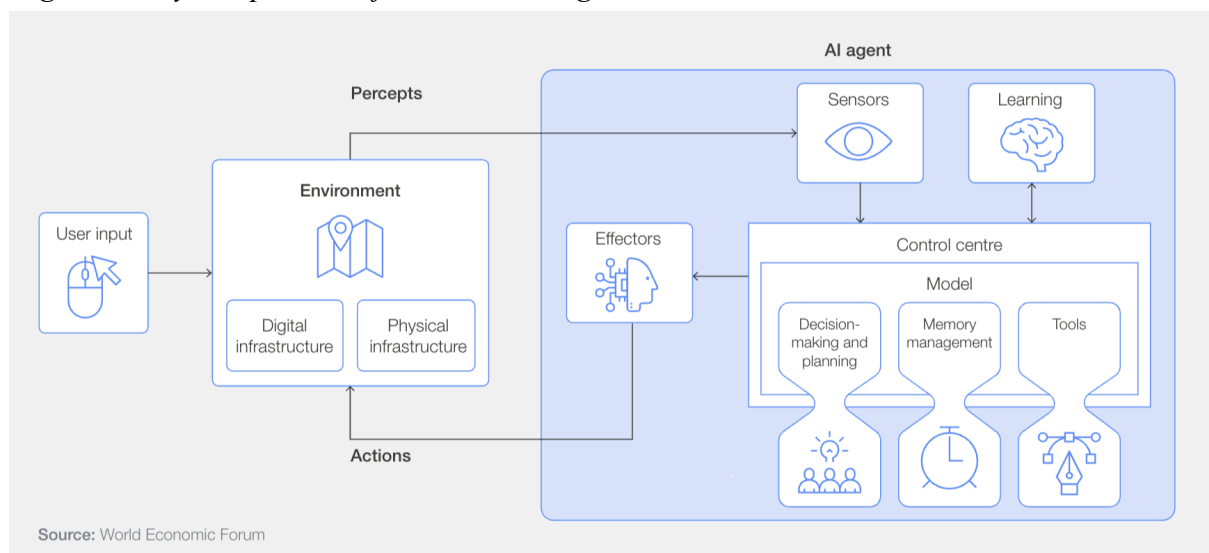
Since 2017, LLMs have transformed AI, especially in natural language understanding and generation, by using vast datasets to produce human-like text and solve complex tasks. Recent advances in LLMs and large multimodal models (LMMs) have elevated AI agents from simple reactive systems to sophisticated entities capable of planning, learning, and decision making based on environmental and user understanding (Alvarez & Jurgens, 2024). LLM agents are autonomous systems powered by LLMs, integrating reasoning, memory, cognitive skills, and tools to solve complex tasks in dynamic environments (Bousetouane, 2025). Today's AI agents rely on three core frameworks: reinforcement learning, goal oriented architectures, and adaptive control mechanisms, that enable goal directed behavior, contextual adaptation, and autonomous decision making (Acharya et al., 2025).

1. **User Input & Environment:** Receives external stimuli (text, voice) within its domain of operation.
2. **Sensors & Perceptrs:** Tools like cameras or database queries gather data, informing the agent about its surroundings.
3. **Control Center:** The core processor makes decisions and plans actions using model outputs, advanced algorithms, and chain-of-thought (CoT) reasoning for transparent multi-step problem-solving.
4. **Effectors & Actions:** Mechanisms (robotics, software commands) execute changes in the environment, such as moving objects or updating data.
5. **Memory Management:** Retains past interactions to maintain context and support informed, continuous decision-making.

6. **Tools:** Extend capabilities with functions like web searches, scheduling, project management and image/audio recognition.
7. **Learning:** Improves performance over time via machine learning and deep learning from continuous input.
8. **Application Layer:** Acts as the interface, translating control center outputs into task-specific actions.

In summary, the sum of the components of an advanced AI agent collectively enhance the agent's capacity to model its environment, such as the ability to retain memory and knowledge. Furthermore, they empower the agent with essential capabilities such as learning, planning, decision-making, perceiving, interacting and communicating effectively with its surroundings.

Figure 1: Key components of advanced AI agents



Source: (Alvarez & Jurgens, 2024)

3.The Economics of AI Agents

After having examined the fundamental concepts of AI agents along with their technical architecture, this chapter now shifts focus to the financial aspects related to these agents.

3.1 Cost structure breakdown

3.1.1 LLM Models

AI agents are powered by LLM models, with API integration costs ranging from a few cents to several thousands dollars. However, building a own LLM model requires substantial investments throughout their lifecycle for hardware, workforce, servers, data and energy

(Mahendra, 2023). A cost breakdown analysis shows AI labs allocate 47-67% of their budget to computing resources (AI accelerator chips, server components, and interconnect hardware), 29-49% to research and development personnel costs, and 2-6% to energy consumption (Cottier et al., 2024). A single AI-integrated search consumes 10 times more energy (3 KWh) than a regular Google search (Thales & Simon-Kucher, 2024).

Development costs vary significantly based on complexity, scale, media type and latency requirements. Most users won't develop their own LLMs due to high costs, often reaching millions for training, primarily from GPU expenses (Kamath et al., 2024). EpochAI states the most expensive publicly-announced training runs are around \$130 million, including Gemini Ultra's total amortized cost (hardware, electricity, staff compensation, and preliminary experiments), with 90% confidence interval between \$70 million to \$290 million (Cottier et al., 2024). The amortized cost to train compute-intensive models has grown at 2.4× per year since 2016, and if this trend continues, model development will cost more than a billion dollars by 2027, meaning only organizations with very high financial resources will afford frontier AI models (See Appendix 3). OpenAI announced a loss of 5 Billion USD in 2024 with 3.5 Billion in revenue due to high model training costs (Field, 2024).

Entry barriers include lack of financial resources, domain knowledge and computational resources. This resource gap led to pre-trained AI foundation models from AI Labs like Anthropic, OpenAI, Mistral, Google or Meta, indirectly bearing development costs (Gao et al., 2024). This LLM-as-a-service business model enables businesses to leverage advanced AI capabilities without developing their own models (Benram, 2025).

Utilizing third-party APIs for inference requests is simple and efficient for quick deployment, saving time and effort. However, costs can escalate with high volumes, and users may face limitations such as lack of customization, unpredictable latency, rate limits, and data privacy concerns. For applications exploiting pre-trained LLM capabilities, these trade-offs may be acceptable (Kamath et al., 2024).

3.1.2 Tokenization

The alternative of building and hosting your own LLM, is to use LLM - as - a - Service through API calls, to incorporate the functionality of the LLM models into an AI - powered product such as AIaaS or AI agents. Tokens represent the units that AI Labs utilize to establish the cost associated with accessing their APIs. Tokenization is the method AI Labs uses to convert words and sentences into a machine-readable format. This can be done at the level of words or subwords, depending on the details required for the specific application. Each word in the sentence is treated as a distinct token in word-level tokenization (Kamath et al., 2024). AI Labs have different tokenization methods and charge varying prices per token based on whether it's an input token, output token, or the specific model in question (Benram, 2025). Input and Output tokens are the volume of tokens transmitted to and received from the model, which significantly influence both processing duration and the utilization of computational resources of the model. For OpenAI models “as a rough rule of

thumb, 1 token is approximately 4 characters or 0.75 words for English text” (OpenAI, n.d.a) (See Appendix 4).

3.1.3 Understanding the Factors Influencing the Costs of LLM Model API Calls

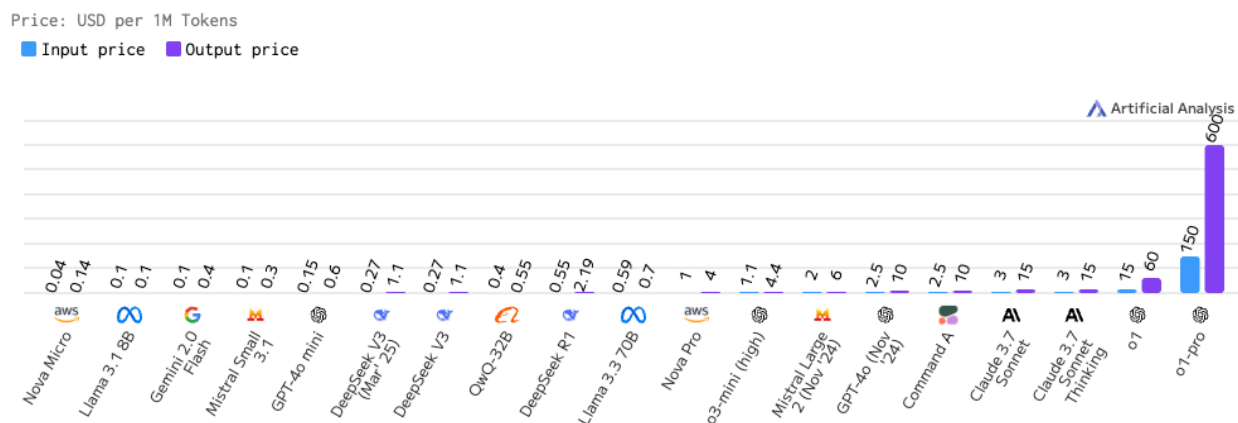
The cost of LLM API calls depends on multiple factors. Model selection is crucial, as larger models with advanced features typically incur higher expenses. Input and output size also affect costs, with larger token volumes requiring more computation. Processing auditory and visual media is generally more expensive than text due to higher complexity and tokenization needs. Latency demands further raise costs, as faster response times consume additional resources. For non-urgent tasks, using Batch APIs can reduce expenses by up to 50% for responses that can wait up to 24 hours (OpenAI, n.d.b).

Choosing between open-source and closed source models adds another layer of consideration. While open-source models may appear cheaper, closed-source inference costs are decreasing faster. Moreover, maintaining quality with open-source solutions often requires fine-tuning and prompt engineering, leading to extra costs and ongoing maintenance. OpenAI, for example, offers fine-tuning through its APIs and SDK libraries (Kamath et al., 2024). Understanding provider-specific pricing structures is essential for budgeting and ensures a balanced decision between cost and performance (Benram, 2025; Kamath et al., 2024).

3.1.4 LLM Pricing per 1 million Tokens

The standardized metric for the pricing of LLM’s is defined as USD per 1 million tokens. The below pricing overview compares selectively the newest models of the AI labs of Deepseek, Meta, Google, Amazon, Mistral, Cohere, Anthropic and OpenAI. In April 2025 the cheapest model is the open - sourced model by Nova Micro by AWS with 0.04\$ for input and 0.14\$ output per 1 million tokens and the most expensive model on the market is the closed -sourced model 01-Pro from OpenAI with \$150 for input \$600 for output per 1 million tokens (Artificial Analysis, n.d.a). The pricing overview confirms the analysis of (Kamath et al., 2024), that open - source models are cheaper than closed - sourced models. In general AI labs charge more for output tokens, the tokens the model generates, than for input tokens, the tokens sent in the prompt.

Figure 2: Input and Output Token Prices by selected AI Labs



Source: (Artificial Analysis, n.d.a)

3.1.5 Price vs Intelligence

In order to consider pricing vs performance or “Intelligence”, the following pricing overview presents a comparison of various LLM models alongside their respective intelligence levels. The Artificial Analysis Intelligence Index is derived as a weighted average of several evaluations, which include general knowledge, mathematical reasoning, and coding proficiency. The distribution of weights is as follows: General Knowledge and Reasoning (50%), Mathematical Reasoning (25%), and Code Generation (25%) (Artificial Analysis, n.d.b).

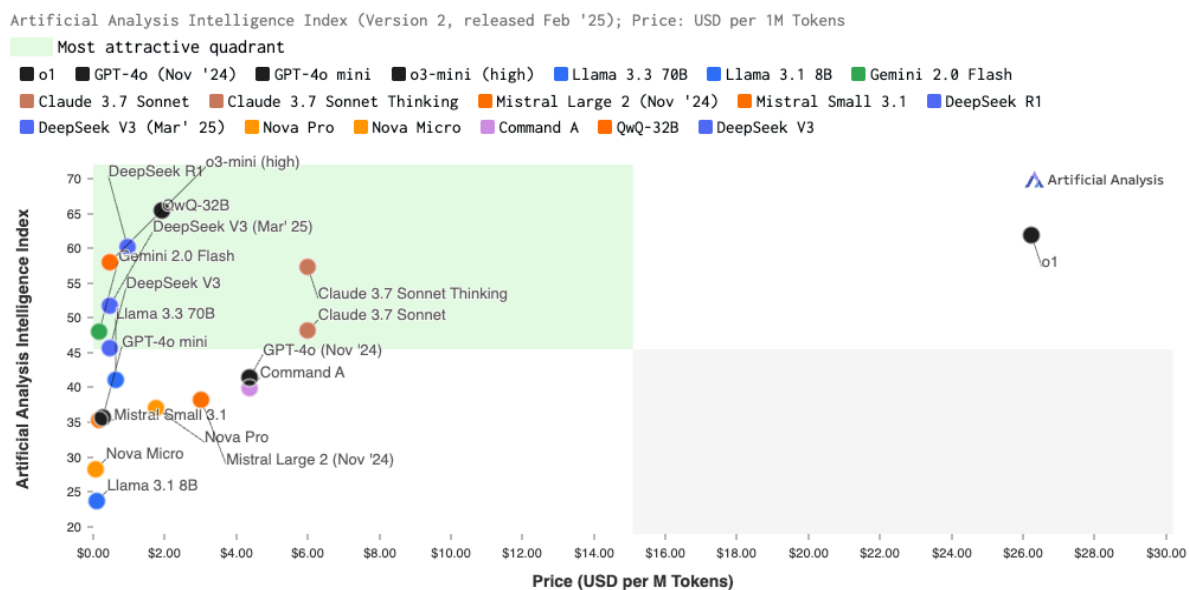
The chart illustrates a shifting competitive landscape where newer models offer high capabilities at lower prices, challenging the historically premium GPT-4 series. The most attractive segment, high intelligence, low cost, now features models like Deepseek v3, Gemini 2.0 Flash, and GPT-03 Mini (high), indicating a market trend toward delivering strong performance affordably. Notably, GPT-03 Mini (high) leads with a 65% intelligence score at \$2 per million tokens, followed by DeepSeek R1 with 60% intelligence at \$1, making it 50% cheaper. Both outperform Claude 3.7 Sonnet and GPT-4o by 10 -15% in intelligence while being significantly cheaper, as Claude 3.7 Sonnet costs \$6–7 per million tokens. However, performance varies by task, and Claude Sonnet 3.7 remains superior in coding tasks (Anthropic, 2024).

Benchmarking AI agents remains underdeveloped, complicating the separation of genuine progress from market hype. Kapoor et al. (2024) argue that agents differ fundamentally from

traditional LLMs, necessitating new benchmarking standards, including cost-controlled comparisons, task separation, rigorous holdouts, and standardized evaluation methods. Additionally, Cottier et al. (2025) confirm that LLM inference costs have decreased, with some benchmarks showing a 40x annual cost reduction, and others ranging from 9x to 900x (See Appendix 5). The most significant price drops occurred in the past year, suggesting this trend will persist. While factors like smaller models and hardware improvements are well-known, other drivers remain less transparent and harder to quantify.

Figure 3: Intelligence vs Prices by selected AI Labs

Intelligence vs. Price



Source: (Artificial Analysis, n.d.b)

3.1.6 Cost breakdown example of an AI agent

Building effective AI agents requires integrating multiple components: models, tools, knowledge bases, memory systems, audio and speech capabilities, safety guardrails, and orchestration layers (OpenAI, n.d.c). LLM application frameworks are essential for managing this complexity, helping developers build, orchestrate, and deploy agents. These frameworks differ in ease of use, offering either low-code interfaces or requiring advanced programming skills, and whether they support self-hosting on platforms like AWS or Google Cloud.

Examples include LangChain (open-source Python framework for advanced LLM workflows), LlamaIndex (focused on Retrieval-Augmented Generation), Flowise (JavaScript-based visual builder for no-code prototyping), and Dify.AI (full-featured platform for enterprise-grade applications) (Kamath et al., 2024). Rall et al. (2023) critique mainstream AI-as-a-Service platforms for failing to deliver true democratization, citing the importance of self-hosting, scalability, and openness. Their "Open Space for Machine Learning" concept

aims to address these gaps, however these features already seem present in solutions like Dify.Ai.

Cost structures involve monthly subscriptions from free sandbox versions to premium plans costing several hundred dollars, plus variable costs from API calls to LLMs for each workflow step, influenced by token volume and model-specific rates. Self-hosting introduces additional server costs.

A simplified Multi-Agent System (MAS), inspired by Minkovski, D. (2024), cost breakdown for Customer Support follows this workflow: initial greeting by Claude 3.5 Haiku (1,000 tokens, \$0.00025), intent classification using GPT-4o Mini (800 tokens, \$0.00016), detailed problem assessment with GPT-4o (2,500 tokens, \$0.01000), vector database query (\$0.00050), complex query processing by Claude 3.7 Sonnet (3,000 tokens, \$0.01200), satisfaction analysis with GPT-4o Mini (500 tokens, \$0.00010), and final response generation by Claude 3.5 Sonnet (1,500 tokens, \$0.00300) (See Figure 4).

This workflow consumes approximately 9,300 tokens at \$0.02601 per interaction. However, actual expenses depend on prompt design, token usage, and AI lab updates. More advanced agents using tools introduce additional costs. Multi-Agent Systems often combine models from different providers like Gemini, ChatGPT, Claude, and Mistral.

Total agent costs include fixed components: initial setup, data preparation, tuning, prompt design and variable costs scaling with token volume. Joint optimization of prompts and model selection can reduce ongoing variable costs through better initial design (Kapoor et al., 2024).

Unlike traditional SaaS businesses with 60–80% profit margins, AI agents typically achieve 50–60% margins due to token-based variable costs in addition to standard SaaS expenses (Casado & Bornstein, 2024). This raises critical questions about effectively pricing AIaaS and AI agents where costs scale with usage.

Figure 4: Cost breakdown example of an AI agent

Stage	Model	Token Usage	Cost
Initial Contact	Claude 3.5 Haiku	1,000 tokens	\$0.00025
Intent Classification	GPT-4o Mini	800 tokens	\$0.00016
Technical Assessment	GPT-4o	2,500 tokens	\$0.01000
Knowledge Base RAG	Vector DB Query	N/A	\$0.00050
Complex Query Processing	Claude 3.7 Sonnet	3,000 tokens	\$0.01200
Satisfaction Analysis	GPT-4o Mini	500 tokens	\$0.00010
Response Generation	Claude 3.5 Sonnet	1,500 tokens	\$0.00300
Total Cost		9,300 tokens	\$0.02601

Source: (Author's own contribution)

4. Pricing Models

A price model is a method to answer the questions for what, when, by whom, and on the basis of which parameters the price is defined and how it is paid (Frohmann, 2018). Pricing Models are a combination of pricing modalities and pricing metrics (John & Rob, 2025). The price modality is the structure that determines how to charge and defines the qualitative basis on which quantitative price levels are based. Pricing metric is the specific unit of measurement used within a pricing model to calculate the price. It's the "what" you are charging for e.g., users, data volume, API calls or tokens (Frohmann, 2018).

4.1 Established SaaS pricing models

Saltan & Smolander's (2019) multi-vocal literature review highlights the fragmented state of SaaS pricing research, and created a comprehensive guide that identified thirteen SaaS pricing frameworks (See Appendix 7). In my own review, I found Buxmann & Lehmann's (2009) *Software Products Pricing Typology* (See Appendix 8) and Spruit & Abdat's (2012) *Pricing Strategy Guideline Framework* (See Appendix 9) to be the most practical for SaaS pricing models. Additionally, Frohmann's (2018) *Six Pillars of a Price Model* (See Appendix 10), though not included in Saltan & Smolander's review, is highly relevant and worth considering.

For this chapter, the focus is specifically on the modality of pricing models, referred to as the "assessment base" by Lehmann & Buxmann (2009) and "reference base" by Frohmann (2018). Therefore, I have chosen Ulrik Lehrskov-Schmidt's *Nine Building Blocks of SaaS Revenue* from his book *The Pricing Roadmap*, as it focuses entirely on the modality aspect of SaaS pricing models (Lehrskov-Schmidt, 2023).

4.1.2 Nine Building Blocks of SaaS Revenue

Ulrik Lehrskov-Schmidt's *Pricing Roadmap* breaks down SaaS pricing into nine building blocks, the core fee types that businesses can combine into their pricing models (Lehrskov-Schmidt, 2023). These are grouped into transactional non-recurring fees, flat recurring fees, and metric-based recurring fees (Lehrskov-Schmidt, 2023). Vendors combine these elements primarily for price discrimination, aiming to capture consumer surplus by tailoring prices to different customer segments based on willingness to pay.

Pricing literature classifies models into one-part, two-part, and three-part tariffs. A one-part tariff is linear, with price directly proportional to quantity, such as an all-inclusive flat fee (Iyengar & Gupta, 2009; Frohmann, 2018). More commonly, firms use nonlinear pricing: a two-part tariff adds a fixed fee plus variable usage charges, while a three-part tariff (3PT) refers to a pricing model consisting of a fixed fee, a free allowance of units up to which the marginal price is zero, and a positive per-unit price for additional usage beyond that allowance (Chao, Y., 2013), (Liozu & Hinterhuber, 2023).

In practice, all three tariffs appear e.g. in the Quote to Cash software market. Fenerum uses a simple one-part model with a monthly flat fee based on revenue tiers (Fenerum, n.d.). Hyperline employs a two-part tariff, combining a base fee with a percentage of annual revenue (Hyperline, n.d.). Chargebee applies a three-part model: a base fee, a free allowance up to \$100,000 monthly revenue with overage charges beyond that allowance (Chargebee, n.d.). These examples illustrate how pricing models serve as strategic tools for differentiation and competitive advantage within the same vertical.

4.1.2.1 Transactional one-off fees

Transactional fees are non-recurring charges applied once per customer or linked to specific events. In SaaS, they typically align with onboarding, ongoing, and offboarding stages. Setup fees cover initial deployment efforts when significant work is required. Ad-hoc fees are one-time charges for extra services like training, custom reports, or feature upgrades. Exit fees apply at contract termination, covering data export or transition support (Lehrskov-Schmidt, 2023).

These one-off charges help monetize services outside standard subscriptions but must be used carefully to avoid customer dissatisfaction. While setup and ad-hoc fees are common, exit fees are less frequent as many SaaS providers emphasize easy onboarding and cancellation. Vendors may waive setup fees as sales incentives, demonstrating flexibility with this pricing lever (Lehrskov-Schmidt, 2023).

4.1.2.2 Flat recurring fees

Flat recurring fees are fixed charges billed regularly, typically monthly or annually, granting unlimited access to a service without depending on usage (Lehrskov-Schmidt, 2023). This model offers customers cost predictability (Buxmann & Lehmann, 2009). Lehrskov-Schmidt (2023) distinguishes three sub-types: flat base fees, flat add-on fees, and flat non-optional fees.

A flat base fee is the core subscription price, ensuring a minimum revenue per customer. Add-on fees are optional charges for extra features, like an analytics module, priced independently of usage. Flat non-optional fees, such as a compliance surcharge, are mandatory for all customers, used to cover shared costs like GDPR compliance (Lehrskov-Schmidt, 2023).

Flat recurring fees form the foundation of SaaS pricing. Most models include at least one flat subscription fee, with potential add-ons or non-optional charges. Their simplicity and revenue predictability make them widely adopted (Lehrskov-Schmidt, 2023).

4.1.2.3 Metric-based fees

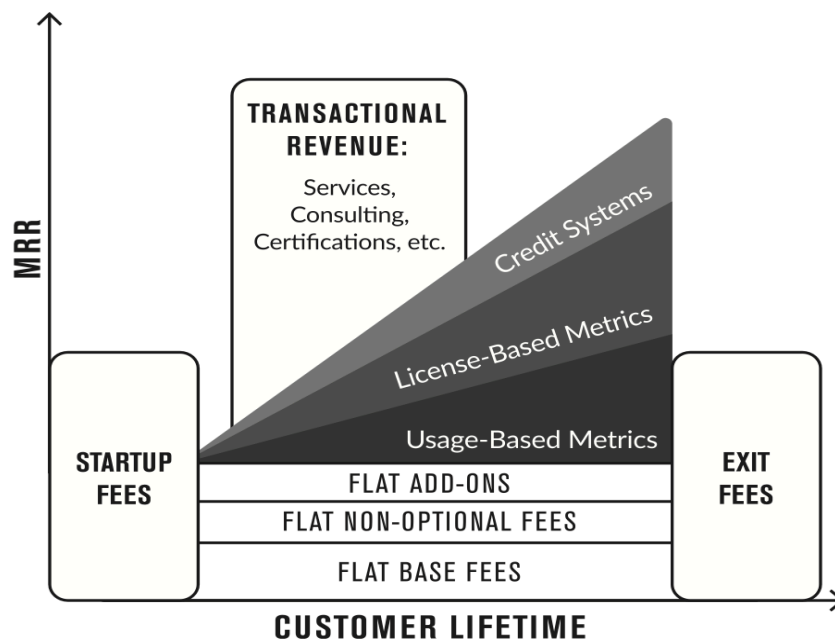
Metric-based fees are recurring charges that vary with usage metrics, such as per user, per GB, or per transaction. These fees fall into three categories: license-based, consumption-based, and credit-based (Lehrskov-Schmidt, 2023).

License fees involve upfront commitments to quantities like users or devices, offering predictable revenue and capacity limits. Per-user pricing is common in SaaS, a survey by KeyBanc of 284 SaaS firms found that pricing is “typically per user” (Liozu & Hinterhuber, 2023). Consumption-based fees, in contrast, charge based on actual usage, aligning cost with value but introducing unpredictability. Common in cloud services, this model is used by around 20% of SaaS firms either fully or in hybrid forms (Lehrskov-Schmidt, 2023; Liozu & Hinterhuber, 2023).

Credit-based fees offer a hybrid approach where customers prepay for usage allowances, providing predictability with flexibility for variable demand. Audible audiobook credits are a typical B2C example, while B2B SaaS providers may sell API call credits similarly (Lehrskov-Schmidt, 2023).

Metric-based fees closely tie revenue to service usage, often combined with flat fees to balance stability and value alignment. The *Nine Building Blocks of SaaS Revenue* framework encourages thinking of pricing architecture in a modular way. Any SaaS revenue model can be built by combining these blocks to suit the product and customer needs (Lehrskov-Schmidt, 2023). Adding complexity improves price discrimination but can make pricing harder to understand. Enterprise deals often require tailored, multi-block models, while simpler schemes serve smaller customers (Lehrskov-Schmidt, 2023).

Figure 5: The Nine building blocks of SaaS revenue



Source: (Lehrskov-Schmidt, 2023)

4.2 Limitations of SaaS pricing models for AIaaS

The fundamental difference between traditional SaaS and AI-powered products lies in their cost structures. Traditional SaaS pricing models operate on the assumption that variable costs are minimal once software is developed. However, AI-powered products like AIaaS and AI agents incur significant variable costs for each interaction, primarily in computing resources and API calls. These costs scale with usage volume and complexity rather than just user count, making traditional SaaS models not optimal for AI products. Therefore, fixed flat fee pricing becomes problematic for AIaaS. Unlike traditional SaaS, AI incurs significant variable costs for each interaction. High usage can lead to losses for vendors, while low usage may leave customers feeling overcharged. OpenAI's ChatGPT Pro plan reportedly operated at a loss due to unexpectedly high demand (Bousquette, 2025).

Seat-based pricing is challenged by AI agents that automate tasks previously performed by humans. Traditional per-user models fail to capture automation value, as AI agents replace multiple human users but count as a single "seat," leading to declining revenues as highlighted by Thales & Simon-Kucher (2024). Feature-based pricing also struggles due to rapid AI advancements. Computing power for training models has grown 4.7x annually since 2010, leading to features once seen as premium quickly becoming commoditized (Sevilla & Roldán, 2024). Further, 91% of LLMs degrade over time, making AIaaS that rely on older LLMs outdated (Gao et al., 2024). Static feature-based tiers lack flexibility to respond to such shifts, risking churn and decreasing adoption rates (Abonamah et al., 2021). Value-based pricing introduces complexity as quantifying AI-generated business value is difficult, complicated by factors such as model hallucinations, customer mistrust, rapid commoditization, and quality degradation (Gao et al., 2024).

In conclusion, traditional SaaS pricing models fail to accommodate AIaaS economic realities, particularly due to variable cost structure and fast-paced technological advancement. To address these challenges, the industry is shifting towards more complex pricing models, such as credit, outcome and usage-based pricing models.

4.3 Pricing models for AIaaS

Upon the examination of conventional Software as a Service (SaaS) pricing models and the inherent constraints associated with them, this chapter seeks to explore alternative pricing models applicable to Artificial Intelligence as a Service (AIaaS).

In addressing the research question, "What pricing models are most effective for monetizing AI agents in today's technology landscape?" We must first clarify what "effective" entails within the context of pricing models. For the purposes of this paper, we use Frohmann (2018) definition. Pricing models are effective, when they fulfill the criteria of

1. Customer acceptance

2. Revenue and profit security
3. Differentiation from Competition
4. Influencing price transparency

Additionally to the four essential criteria, I propose that AIaaS pricing models should ensure increased customer adoption, as most of the tools will rely on usage - based and outcome based pricing. Further, flexibility, due to the dynamic characteristics of the AI market affected by rapid development of new LLM capabilities, decreasing token costs and the phenomenon of commoditization and degradation (Abonamah et al., 2021), (Gao et al., 2024).

4.3.1 Framework for AI Agent Pricing Models

Academic literature lacks established pricing model frameworks for pricing AI agents. However, Manny Medina, founder of Outreach, addressed this gap through his new startup Paid. After analyzing over 60 AI agent companies, Medina identified four common pricing models: Per Workflow, Per Agent, Per Outcome, and Per Agent Action and developed a decision framework categorizing them by outcome- vs. activity-based and fixed vs. variable pricing (See Appendix 11).

4.3.1.1 Per Agent Action/ Usage - based pricing mode

Usage-based pricing, also known as consumption-based or pay-per-use, charges customers according to their actual usage of a product or service, measured through various pricing metrics (Frohmann, 2018). This model creates a dynamic relationship between price and consumption, where pricing influences usage behavior, and usage volume determines the applicable per-unit price (Iyengar & Gupta, 2009).

For AIaaS, usage-based pricing is particularly suitable, as vendors can estimate costs with LLM API requests by calculating token consumption times input and output token prices and apply a markup for their margin, following a cost-plus approach (Liozu & Hinterhuber, 2023). This ensures pricing transparency and fairness, as customers pay based on actual usage, enhancing perceived value.

However, shifting from subscriptions to usage-based pricing complicates revenue forecasting for vendors and budgeting for customers, requiring transparent and real-time usage monitoring. Latva-Koivisto (2025) notes that usage-based pricing faces buyer resistance, as cost unpredictability undermines budgeting efforts. Liozu & Hinterhuber (2023) therefore emphasize the need for firms to develop capabilities in value quantification, customer experience management, success oversight, and KPI development in order to implement usage - based pricing successfully. Common AIaaS metrics, such as token counts or API calls, do not fully capture customer value. More outcome-oriented metrics, like resolved customer inquiries, better reflect the incremental economic benefits that customers derive from the service.

To mitigate revenue fluctuation, several pricing structures can be applied. One method is a base fee subscription with defined usage thresholds, where overages incur per-unit fees. Alternatively, prepaid credit models allow customers to purchase usage upfront, providing flexibility while securing predictable revenue. Another option is the Advance Payment System (APS), where companies estimate consumption over a year and spread payments evenly, reconciling discrepancies at year-end. This model reduces price sensitivity, encourages loyalty, and mitigates churn (Schulz et al., 2015).

A practical example of usage-based pricing in AI Agents is Salesforce's Agentforce, which charges \$2 per conversation (Salesforce, n.d.).

4.3.1.2 Per Agent Outcome/ Outcome - based pricing model

Outcome-based pricing, or value-dependent pricing, determines fees based on the tangible results achieved rather than transaction volume or usage metrics. This model, which is a form of usage-based pricing, requires clearly defined outcomes that are significant, measurable, and neutral to ensure transparency and fairness (Frohmann, 2018).

Although outcome-based pricing offers a strong value alignment, its implementation is complex due to the difficulty in defining and verifying success criteria. Moreover, it inherits challenges from usage-based models, particularly in cost predictability. Nonetheless, by tying payments to delivered outcomes, this model reduces adoption risks for businesses and ensures clients pay only for measurable results (Frohmann, 2018).

For AI Agents, outcome-based pricing is especially relevant. A common example are customer support agents that are pricing per resolved customer inquiry, directly linking fees to business value. Further, this approach incentivizes continuous improvement from vendors while providing clients with clear ROI.

A notable example includes Intercom's AI support agent "Fin," priced at \$0.99 per resolved conversation. As of early 2024, 17% of Intercom product purchases included the Fin add-on (Thales & Simon-Kucher, 2024). Conversely, Zendesk integrates its AI agent into its tiered plans, adding value to existing subscriptions. In the fintech sector, ChargeFlow applies outcome-based pricing by taking a 25% fee on successfully recovered chargebacks (ChargeFlow, n.d.).

4.3.1.3 Per Agent/ FTE Replacement pricing model

The per agent, or FTE replacement, pricing model ties the value of AI agents directly to the human labor they substitute. Typically, AI agents are priced below the equivalent human cost, such as offering an AI Sales Development Representative (SDR) for \$30 per hour compared to a human SDR's \$50 (Morales, 2024). This approach is popular in verticals like sales, customer support, and marketing.

Artisan, an AI agent startup, exemplified this strategy with provocative campaigns like "Stop hiring Humans," which, despite backlash, drove \$2 million in ARR growth within two

months (Carmichael-Jack, 2024) (See Appendix 12). Similarly, OpenAI is reportedly preparing specialized AI agents for tasks such as sales ranking, software development, and research, with pricing tiers ranging from \$2,000 to \$20,000 per month (Wiggers, 2025). If successful, this approach promises the highest revenue potential, as these agents would tap into workforce budgets rather than traditional software budgets, which are typically much smaller.

The broader question remains whether AI agents can replace humans at scale to justify this pricing model. Klarna, the payment provider, provides a compelling case, reporting that its AI assistant performs the work of 700 full-time customer support employees, with higher accuracy, 25% fewer repeat inquiries, and significant efficiency gains, contributing an estimated \$40 million in profit improvement for 2024 (Klarna, 2024).

However, large-scale displacement remains uncertain. OECD and McKinsey projections suggest up to one-third of work activities could be automated by 2030, with regional disparities in adaptability (Deshpande et al., 2021). Other studies, however, argue AI's current role is to augment, not replace, human decision-making due to high implementation costs (Agrawal et al., 2018; DeVon, 2024). The evolution of AI, particularly towards Artificial General Intelligence (AGI), will ultimately determine whether labor replacement becomes a standard pricing model.

4.3.1.4 The per workflow pricing model

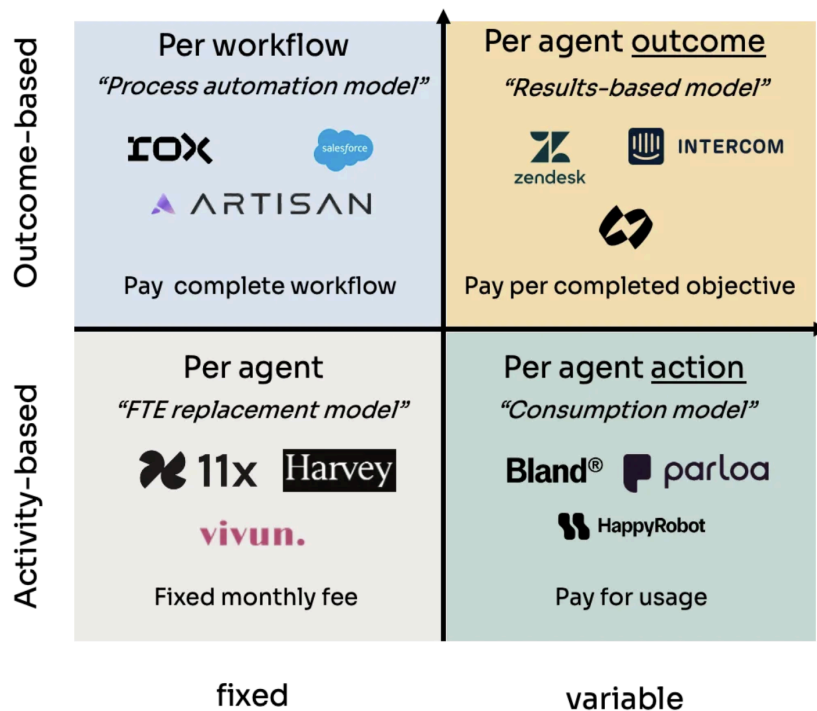
The per workflow pricing model, while similar to per agent or usage-based pricing, charges for the completion of an entire workflow rather than individual tasks. A workflow represents a sequence of interconnected tasks, emphasizing process automation over isolated actions. For instance, instead of charging per email sent, a workflow model monetizes the full process, including ICP research, lead identification, drafting, personalization, and outreach (Medina, 2025). This approach aligns pricing with the broader value of automated workflows rather than single outputs.

4.3.1.5 Hybrid models

Debates around AI pricing models often blur terminology, yet most AIaaS and AI Agent models will likely mirror existing SaaS structures. The prevailing approach is expected to be a hybrid model, typically a two-part tariff combining a fixed base fee with a usage-based component. This structure balances cost predictability with the flexibility needed to accommodate AI's variable costs.

Throughout this chapter, pricing models were analyzed through modalities and metrics, referencing Lehrskov-Schmidt's *Nine Building Blocks of SaaS Revenue* for foundational strategies. The limitations of traditional SaaS pricing for AIaaS were discussed, emphasizing the need to account for variable costs. Medina's (2025) AI Pricing Model Framework was introduced as a practical tool for AI agent monetization, addressing these unique challenges.

Figure 6: AI Agent Pricing Model Framework



Source: (Medina, 2025)

5. Methodology

5.1 Research Design

This study adopts a deductive quantitative research design that combines an experimental within-subjects design survey with a structured market-audit of AI agent pricing pages. In line with Creswell's (2014) model of research design, the methodology reflects the alignment between philosophical worldviews, selected strategies of inquiry and research methods. This study assumes a postpositivist worldview, which emphasizes the hypothesis-testing of theory-driven constructs, and is most commonly associated with quantitative strategies. If the problem or research question calls for the identification of factors that influence an outcome, then a quantitative approach is best (Creswell, 2014). Creswell (2014) identifies postpositivism as a deterministic and reductionist approach, where researchers begin with a theory and then reduce it to specific variables that can be measured and tested. This worldview assumes that objective reality exists and that through careful observation and structured methods, researchers can approximate the truth, even if absolute certainty is unattainable. Thus, postpositivism supports the use of experimental designs, surveys, and statistical analysis to examine causal relationships and test theoretical assumptions. In this scenario, the researcher tests a theory by specifying narrow hypotheses and the collection of data to support or refute the hypotheses. The data is collected through an instrument that

measures attitudes, and the information is analyzed using statistical procedures and hypothesis testing (Creswell, 2014).

Accordingly, this research starts from the theoretical assumption, based on both the Technology Acceptance Model (Davis, 1989) and Flat-Rate Bias theory (Lambrecht & Skiera, 2006), that external factors such as pricing models influence consumer perceptions and acceptance. Specifically, pricing models are expected to affect perceived usefulness (PU) and perceived ease of use (PEOU), which in turn shape consumer preference, formulating the research question: *Which pricing models are most effective for monetizing AI agents in today's technology landscape?* Building on extensive prior empirical findings of Flat Fee Bias, indicating that flat - fee pricing is often preferred by consumers (Lambrecht & Skiera, 2006), this study investigates whether consumers prefer certain pricing models for AI agents. Effectiveness is operationalized as a proxy for consumer acceptance on the ground that “Criteria such as the acceptance of the price model from the customer’s point of view are core prerequisites for market success” (Frohmann, 2018). Further, an experimental within-subjects design provides post - test data to challenge or confirm a potential Flat-Rate Bias for AI agent pricing. The data is aimed to support or refute the hypotheses:

- 1) H1(1): *Flat-rate pricing models will be most preferred, and usage-based pricing least preferred, reflecting a flat-rate bias*
- 2) H1 (2): *Users who value simplicity and predictable costs, aligning with TAM notions of ease-of-use and the psychological factors of the Flat-Rate Bias Theory, will favor flat plans* (Davis, 1989), (Lambrecht & Skiera, 2006)
- 3) H1 (3): *There is a misalignment, e.g. flat-rate pricing models are under-provided relative to market demand.*

This quantitative research incorporates two data sources, primary data from a within-subjects experiment and secondary data from real-world pricing information, to provide a more comprehensive analysis. This design enables data triangulation, where data from multiple sources strengthens the validity of conclusions (Creswell, 2014). Overall, the chosen methodological approach is structured and aligned with Creswell’s (2014) research design framework to ensure coherence between the research question, the data collection techniques, and the analytic procedures.

5.2 Data Collection Methods

To address the research question, two complementary datasets were collected, an experimental survey capturing user preferences for pricing models of AI agents and an observational audit of public pricing pages of AI agents.

5.2.1 Experimental Within-Subject Design

The primary data source is an online survey experiment administered via Typeform. Survey research aims to generalize from a sample to a population, allowing researchers to make inferences about the characteristics, attitudes, or behaviors of that broader group. As Babbie (1990) contends, by studying a representative subset, one can extend findings to the entire population with quantified levels of confidence and precision (Creswell, 2014).

The survey employed a true experiment as a within-subjects design, cross-sectional, where each participant saw all five pricing models presented simultaneously, order randomized, then selected their most and least preferred models. This design controls for between-person variability as every respondent provides feedback on each model (Creswell, 2014).

Choice-task sequence: R – X – O

- R: randomization of pricing model arrangement
- X: exposure to all five pricing model treatments
- O: outcome measurement (Most-Least preference selection)

Each participant was exposed to a purchasing scenario for an AI Customer Support Agent, introduced as: "Meet Your New Support Teammate (An AI Agent). Imagine having an AI Agent that can handle all of your Customer Support conversations for you — instantly, 24/7, with consistent quality."

Participants performed two tasks: selecting their MOST and LEAST preferable pricing model, then identifying what they find MOST and LEAST important when considering pricing models from four options: costs predictability, costs transparency, simplicity and fairness. The pricing models included:

- Credit-based: Fixed monthly fee for specific credits allowing limited use
- Flat-Rate Subscription: Premium fixed recurring fee for unlimited usage
- Pay-per-Use/Usage-based Pricing: Pay only what you use, scaling with conversations
- Outcome-based: Pay premium for each successful resolution
- License + overage: Fixed monthly fee for limited conversations plus overage fees

Clear operational definitions were provided, and Typeform randomized presentation order to mitigate order effects and bias.

A single-stage, non-probability convenience sample targeted professionals worldwide who work for or own companies. The sample was drawn from the PricingSaaS community (406 members, 7,978 newsletter subscribers). Self-report screening ensured respondents met the "work for or own a company" criterion. The survey gathered general consumer opinions, not just AI-familiar users, though participants indicated prior AI agent experience. Sample size calculation used the standard formula for estimating population proportions (Creswell, 2014):

Figure 7 : Sample Size Calculation (1)

$$n = \frac{Z_{1-\alpha/2}^2 p (1 - p)}{E^2}$$

Source: (Creswell, 2014)

- $Z_{1-\alpha/2}$ is the critical value for a two-sided confidence level, 1.96 for 95% confidence.
- p is the anticipated proportion of respondents selecting any given pricing mode, $p=0.5$ was used to maximize variance and thus ensure a conservative sample estimate.
- E is the tolerable margin of error, set to $E=0.14$ (14%) based on the exploratory nature of this master's project and resource constraints.

Substituting these values:

Figure 8 : Sample Size Calculation (2)

$$n = \frac{1.96^2 \times 0.5 \times (1 - 0.5)}{0.14^2} \approx 48.8$$

Source: (Author's own contribution)

Rounding up, a sample of 49 respondents is required. Consequently, it was aimed for at least 50 completed surveys to achieve a $\pm 14\%$ margin of error at the 95% confidence level. With our achieved sample of 53 completed responses, the margin of error becomes 13,6%, thus at the 95% confidence level, we can be confident that the observed preference proportions fall within $\pm 13.6\%$ of the true population values.

The survey explained the study's purpose and anonymous participation. No personally identifying information was collected. Questions were reviewed for neutrality, pricing model descriptions were standardized, and the questionnaire was pilot-tested with $N \approx 5$ participants.

This within-subjects design offers scientific and economical advantages. Each participant provides data for each pricing model, improving precision by reducing individual variability. This enables detecting preference differences with moderate sample size, efficient for a master's project. Concerns like learning effects were addressed through randomization and careful construction.

Combined with observational data, the study yields methodological triangulation, examining what users prefer versus what companies employ. This enhances construct validity by incorporating perception and reality. As Creswell (2014) suggests, corroborating evidence from different sources builds stronger arguments. Web-based administration enables efficient

data collection without substantial costs, fitting within tight project timelines (Creswell, 2014).

5.2.2 Observational Market Data

The secondary data source consisted of observational data on how AI agents are currently priced. A structured review of public pricing pages from a variety of AI agent vendors to capture the prevalence and characteristics of pricing models in the market was conducted. To compile this, an online *AI Agents Directory* and official websites of AI agent vendors were leveraged (Marketplace, n.d.). Inclusion criteria were defined to focus the dataset: platforms or services were included that (a) offer standalone AI agents or assistants to end-users or businesses, and (b) publicly disclose their pricing model on a website. Platforms that did not list pricing details, e.g., only “Contact us for pricing” or were solely open-source/non-commercial were excluded. Data was gathered in April 2025. For each qualifying AI agent service, the following variables were recorded: the name of the platform, its primary pricing model, categorized into one of the six pricing models defined below, the type of tariff (One, two or Three - part tariff), key pricing metrics (E.g. as price per user, per credit, or per conversation), if a tiered pricing and or free plan was offered and if it included a hard or soft threshold. Data were collected by manually visiting each platform’s official pricing page, to ensure accuracy and objectivity a screenshot was taken of the pricing site and made accessible in the Google - Sheet as a Google Drive link. The final observational dataset consisted of 101 AI agents, with a balanced representation across five different application domains: productivity, software development, sales, customer service and voice agents.

All collected information was organized into a spreadsheet table. Each AI agent was assigned to a pricing model category using a consistent coding scheme. For example, if a platform offered a Credit - based pricing model, it was coded as “Credit;” if it charged strictly per usage e.g. API call, it was “Usage”. This systematic coding approach ensured consistency in how each case was categorized, reducing researcher bias. No private or sensitive data were accessed, only publicly available information was used, respecting the platforms’ terms of use. By structuring the market data in alignment with the same five pricing model categories used in the survey, direct comparisons between consumer preferences and industry practices were facilitated.

The six pricing models observed were defined as followed with their corresponding tariff type:

Model	Definition	Tariff
License	An upfront payment that grants the right to use the product before any actual use occurs.	1
Flat fee	A fixed price that never changes, no matter how much or how little the customer ultimately uses.	1

Usage-based	Charges calculated after consumption and proportional to how much the customer actually used.	1
Credit-based	The customer prepays for an amount of credits that can be redeemed later, a hybrid of license and usage.	1
License + Usage	A fixed base fee plus an additional variable charge tied to usage.	2
Credit + Overage	A fixed monthly fee that includes a set number of credits, once the allowance is exhausted, an overage fee is applied for every extra unit consumed.	3

5.2.3 Variables overview

Data Type	Measurement	Outcome
Choice	Nominal (Most & Least preferred of five pricing models) Binary (Flat-Rate vs. Non-Flat-Rate)	Pricing Model Preference Flat- rate selected or not
Bias drivers	(Most & Least Important when selecting pricing models)	Prioritization of pricing attributes
Freemium Preference	Would you be more likely to sign up for an AI agent if it included a free plan with limited features? Yes/No	Indication for or against Freemium as a customer acquisition strategy
Market observation	Nominal (Pricing model categories)	Real world pricing structures

5.3 Data Analysis

5.3.1 Overview

Data Type	Method(s)	Purpose
Pricing-model preference (“Most” & “Least” picks across 5 models)	Most–Least (MaxDiff) utility estimation (simple net <i>BW</i> score or conditional/multinomial logit) • Cochran’s Q omnibus test on “most-preferred” proportions • Pairwise McNemar tests (Holm adjusted)	Quantify relative utilities of the five pricing models and test whether preference differences are statistically significant.
Attribute importance (“Most” & “Least” picks across 4 attributes: cost predictability, cost transparency, simplicity, fairness)	Most–Least (MaxDiff) utility estimation for attributes • Cochran’s Q across four attributes • Pairwise McNemar tests (Holm adjusted)	Derive importance weights for each attribute; identify attributes valued significantly more or less than others.
Link between attribute utilities and model choice	Binary Logistic Regression DV = Flat - rate pricing model chosen as <i>most preferred</i> IVs = individual attribute utilities Descriptive statistics (frequency and %)	Identify psychological and contextual predictors of choosing each pricing model (e.g., drivers of flat-rate bias).
Market-observation data		Compare market prevalence of pricing models with consumer-derived utilities.
Freemium preference	Descriptive counts & percentages (survey: % “Yes” vs. “No”)	Examine consumer interest in a free-tier and compare to how many vendors actually offer a freemium plan.

5.3.2 Data Preparation

The raw data was screened for completeness and quality. Records with missing responses or signs of inattentive answering were mitigated through Typeform settings for making questions required and randomization of the questions.

For the survey data for each respondent, two binary indicators were created per pricing model in the Most and Least preferred task: Best = 1 if the model was selected as *most preferred* and Worst = 1 if the model was selected as *least preferred*. Models not selected in either case were coded as 0 in both columns.

The same binary coding structure was applied to the four pricing-model attributes: Cost Predictability, Cost Transparency, Simplicity and Fairness, indicating which attribute was considered *most important* and *least important*.

Finally the data from the publicly available AI agent pricing pages were reviewed for consistency and correctness. Duplicates were removed, and each platform was classified into one of the pricing model categories to ensure comparability with survey responses.

Data preparation is further discussed in the Results & Analysis chapter 6.

5.3.3 Statistical Analysis

The analysis focused on three main areas: consumer preferences for AI agent pricing models, the importance of pricing attributes, and the relationship between attribute importance and pricing model choice. A final descriptive comparison was made between consumer preferences and market offerings.

First, consumer preferences for the five pricing models were analyzed using a MaxDiff scaling. For each respondent, selections were coded to identify which model was chosen as *most preferred* and which as *least preferred*. Based on these responses, net Most/Least scores were calculated for each model by subtracting the number of “least preferred” selections from the number of “most preferred” selections. This provided an initial measure of relative preference.

To formally test whether preference distributions differed across models, a Cochran’s Q test was conducted. This non-parametric test is suitable for repeated-measures data where participants evaluate multiple options. If the Cochran’s Q test indicated significant differences, McNemar’s pairwise tests were performed to identify which models were significantly more or less preferred compared to others. These comparisons were adjusted for multiple testing using Holm’s correction to reduce the risk of false positives.

A similar approach was used to analyze the importance of four pricing attributes: Cost Predictability, Cost Transparency, Simplicity, Fairness. Most/Least selections for attributes were coded and analyzed through MaxDiff net scores, Cochran’s Q test, and McNemar’s pairwise comparisons, following the same procedure as for the pricing models.

To examine whether the importance placed on specific attributes influenced selecting a flat-rate pricing model, a binary logistic regression was performed. The dependent variable was the pricing model chosen as *most preferred*, with flat-rate serving as the reference category. Independent variables were the individual attribute importance scores derived from the Most/Least task for cost transparency, cost predictability, simplicity and fairness.

Finally, pricing model prevalence in the current AI agent market was summarized descriptively. Frequencies and percentages of platforms using each pricing model were calculated and compared with consumer preference data from the survey. This comparison highlighted potential gaps between market offerings and user preferences, providing insights for managerial recommendations. Special attention was given to the prevalence of free plans with limited features, which were noted separately. These observations were compared to survey responses to the specific question: “*Would you be more likely to sign up for an AI agent if it included a free plan with limited features?*” This comparison aimed to provide additional managerial insights into customer acquisition strategies beyond the analysis of pricing model preferences.

All statistical analyses were conducted using R version 4.3, with a significance threshold of $\alpha = 0.05$.

5.4 Justification and Evaluation of Methodology

This section evaluates the viability of the chosen methods, using established criteria to ensure the study’s credibility. According to Creswell (2014), key criteria for quantitative research quality include validity, both internal and external, and reliability, whereas Greener (2008) highlights the importance of objectivity, consistency, and ethical integrity in business research. Each of these are addressed below in the context of the study design.

5.4.1 Internal Validity

Internal validity refers to the accuracy of cause-and-effect inferences within the study (Creswell, 2024). As Creswell (2014) suggests, anything that could affect the results from inside the study should be mitigated. In this research, internal validity is strengthened by the within-subjects experimental design, due to each participant evaluating all pricing models, individual difference variables are held constant across comparisons. The survey instructions and tasks were identical for each pricing model, ensuring a standardized design. Through randomization, changing the order of questions they appeared in, biases are avoided. Potential participation fatigue effects were mitigated by keeping the survey reasonably brief with an estimated completion time of about 3 minutes. Therefore, any significant differences in the Cochran’s Q test can more confidently be attributed to true preference differences caused by the pricing model attributes, not extraneous variables. Therefore with high confidence, the observed differences in pricing model preference are real and not due to confounding factors, which influence can’t be directly detected (Creswell, 2024). For content validity, Creswell (2014) advises clear operational definitions and systematic evaluation of measurement instruments. Accordingly, each pricing model was precisely defined, to ensure

the survey measures the intended constructs, the instrument was co-developed with the founders of PricingSaaS, domain experts in SaaS pricing models. They jointly crafted and reviewed each item for clarity, relevance, and comprehensive coverage of all five pricing options, ensuring strong content validity (Creswell, 2014).

5.4.2 External Validity

External validity concerns the generalizability of results to other populations or settings (Creswell, 2014). Although the sample may not represent all AI agent users, participants' relevant backgrounds, confirmed by their job roles, prior experience and market, support valid evaluations of pricing models. The purposive sample from the PricingSaaS community increases the study's relevance to the target segment familiar with software and AI services. Incorporating observational data from real AI agent companies further enhances external validity, allowing comparison between user preferences and actual market practices. The triangulation of these sources increases credibility and supports generalizability when both align (Creswell, 2014). Sample characteristics are reported transparently to encourage future research and broader validation.

5.4.3 Reliability and Consistency

Reliability refers to the consistency and repeatability of research procedures (Greener, 2008). According to Creswell (2014), reliability means that repeating the study or reapplying the instrument should yield similar results, assuming preferences remain unchanged. Reliability in this study was ensured by using a standardized survey instrument with identical descriptions and questions were administered via a stable online survey platform, Typeform, guaranteeing uniformity. Closed-ended questions minimized interpretation differences. The survey was pretested internally for clarity and functionality. Observational data followed a consistent categorization protocol for each website with double-checks to minimize errors. Greener (2008) notes that reliability is essentially the consistency of results over time, so stability is also considered. The market data snapshot could evolve if taken at a different time, however, pricing models tend not to change abruptly, which gives confidence that the findings are robust at least in the short term. Overall, another researcher following the same procedures should be able to replicate the data collection and analysis with comparable outcomes, indicating good reliability.

5.4.4 Objectivity

To maintain objectivity, potential for researcher bias was minimized at all stages. Greener (2008) emphasizes that objectivity in data collection can be achieved by using systematic and purpose-designed methods for recording data. In the survey, objectivity was pursued by framing questions neutrally, for example, no leading language was used that might make one pricing model sound inherently better. The within-subject design makes each participant serve as their own control, which reduces the influence of any one participant's biased perspective on the overall results. For the observational research, clear inclusion criteria and classification rules were defined before gathering data. The coding of each AI agent's pricing

model was done based on explicit characteristics, e.g., presence of a free tier, presence of usage metering, etc. rather than the authors own opinion or definitions, preserving neutrality. Statistical analyses including MaxDiff, Cochran's Q, and multinomial regression were chosen to minimize subjectivity and rely on reproducible criteria. Where judgment was required, such as coding ambiguous cases, decisions were reviewed by an external expert from PricingSaaS to enhance transparency and fairness. This approach follows Greener's systematic data handling to uphold objectivity in business research.

5.4.5 Ethical Integrity

This study followed ethical research principles throughout, ensuring informed consent, anonymity, and transparency (Greener, 2008). Survey participants received a clear explanation of the study via a welcome page, including its purpose and procedures, no covert methods were used. No personal identifiers were collected. Following Creswell (2014), responses were anonymized. The data is stored securely on Typeform's servers. Hypotheses were not disclosed to reduce bias, participants were fully informed about the study's academic nature. The topic posed no significant risk. Observational data involving no human subjects and ethical standards were maintained by accurately documenting and attributing public information. By conducting the research with these measures, the research meets the ethical standards of honesty, confidentiality, and accountability (Greener, 2008). Further, AI has been used for this thesis in compliance with AAU guidelines for use of GenAI. It has been used as a tool for brainstorming and idea generation. Additionally, AI has been used as a tool to assist in correcting grammatical errors and ensuring a logical coherence throughout the thesis. All text presented in the thesis is the work of the authors and has been created independently.

6. Analysis & Discussion

6.1 Introduction

This chapter presents the analysis of survey data to address the hypothesis on AI agent pricing model preferences. The key questions are:

Which pricing models do prospective users prefer most and least for AI agents?

1. $H_1(1)$: *Flat-rate pricing models will be most preferred, and usage-based pricing least preferred, reflecting a flat-rate bias (Lambrecht & Skiera, 2006)*

Do certain pricing attributes (e.g. cost predictability, transparency, fairness or simplicity) drive a preference for flat-rate models?

2. $H_1(2)$: *Users who value simplicity and predictable costs, aligning with TAM notions of ease-of-use and the psychological factors of the Flat-Rate Bias Theory, will favor flat plans (Davis, 1989), (Lambrecht & Skiera, 2006)*

How do consumer pricing preferences compare to current market offerings?

3. $H_1(3)$: *There is a misalignment, e.g. flat-rate pricing models are under-provided relative to market demand.*

To investigate these questions, first the survey data was coded. Further, the descriptive statistics of the sample are examined. Next, the pricing model preference structure using a MaxDiff (Best-Worst) analysis is examined, resulting in net preference utilities for different pricing models and pricing attributes. Furthermore, of 53 respondents, 35 answered the survey question “Can you shortly explain why you chose that pricing model?”. The texts were subjected to a thematic coding process and analyzed for frequencies. Then the significance of these preference differences from the MaxDiff analysis are tested with Cochran’s Q and McNemar tests. Further, the aim was to determine which pricing attributes predict an individual’s preference for flat-rate pricing, using a binary logistic regression and follow-up Fisher’s exact tests for robustness. Additionally, the survey results are compared to a market audit of AI agent pricing models, highlighting gaps between what consumers want and what the market offers. Finally, a summary of findings transitions to the discussion of implications. This analytical approach follows a logical flow from data preparation to descriptive and inferential analysis, and interpreting results relative to the hypotheses, consistent with best practices for quantitative research (Creswell, 2014).

6.2 Data Preparation

Survey data were cleaned and re-coded for accurate analysis. All responses from main data collection were screened for completeness and quality, excluding pilot tests. No respondents were removed for speeding or straightlining due to Typeform randomization ensuring satisfactory engagement. A pilot test with 5 colleagues refined question wording but was excluded from the final sample.

Categorical variables and derived variables were coded for analysis. Job roles were consolidated into categories (e.g. "Individual Contributor," "Manager"), and company size ranges were coded as ordered categories. Flag variables were created from MaxDiff best-worst exercise data. Respondents completed Best-Worst tasks choosing most and least preferred options from pricing models and attributes. Binary indicators were generated: "Most" flag = 1 if chosen as most-preferred, "Least" flag = 1 if chosen as least-preferred. These flags determined counts and calculated net utility scores (most minus least counts) for each pricing model and attribute in MaxDiff analysis.

MaxDiff results were coded into dummy variables for attribute priorities. For each respondent, the pricing attribute selected as "Most" preferred most frequently was identified. Four dummies were created: `most_predic`, `most_transp`, `most_simple`, `most_fair`, indicating whether a respondent's top-ranked pricing attribute was cost predictability, cost transparency,

pricing simplicity, or pricing fairness. Each respondent has exactly one dummy equal to 1 and the rest 0.

The respondent's preferred pricing model was captured through five dummies (flat, usage-based, credit-based, outcome-based, or overage) as indicated by their highest scored MaxDiff option. This was converted into binary variable `choice_bin` (flat-rate vs. non-flat) to test flat-rate preference. `Choice_bin` = 1 if top choice was flat-rate model, 0 otherwise. This binary outcome was used in logistic regression to identify predictors of flat preference.

Of 53 respondents, 35 answered "Can you shortly explain why you chose that pricing model?" Texts underwent thematic coding (Creswell, 2014). Explanations were imported into Excel and reviewed to identify recurring motives. Seven analytic themes were determined: Value Alignment, Cost Predictability, Risk Minimization, Simplicity, Fairness, Experience, and Flexibility. A coding scheme mapping key phrases to each theme was developed, and a frequency table confirmed each theme's prevalence.

All data processing and analysis were conducted using R (Version 4.2.2). The tidyverse package suite was used for data cleaning and statistical analyses were performed with appropriate R packages (e.g., stats for logistic regression). MaxDiff counts and utilities were obtained through R, counting "most" vs "least" selections. Intermediate results were exported to Excel for tabulation. Standard quantitative data handling procedures were followed (Creswell, 2014) to ensure reliability, including double-checking re-coding logic and verifying consistent sum totals and sample sizes after transformations.

In total, $N = 53$ valid responses were analyzed after data preparation. This sample size, while modest, was sufficient for descriptive and nonparametric analyses of stated hypotheses, though it poses limitations for complex modeling. With data prepared and key variables defined, sample characteristics and core preference results were examined next.

6.3 Descriptive Statistics

6.3.1 Sample Demographics

The sample spans a diverse range of company sizes and professional roles. Notably, almost half of respondents (45.3%) reported working in very large organizations with over 1,000 employees, while a substantial amount of respondents (37.7%) work at small firms with fewer than 25 employees. The remaining respondents were split among mid-sized companies with 7.5% in firms of 251-1,000 employees; (5.7%) in firms of 25-100; and only (3.8%) in companies of 101-250 employees (See Appendix 22). This bimodal distribution suggests we captured perspectives from both enterprise-level and startup/small business environments, which could influence pricing preferences, as larger companies might have bigger budgets or different procurement habits than startups.

In terms of job roles, about one-third of respondents (34%) were Individual Contributors, and roughly one-quarter (26%) were at the Manager level. The remainder were split between

higher managerial ranks and executives: (14%) Director, (12%) Vice President, and (14%) C-level executives (See Appendix 21). This indicates a good mix of perspectives, though the majority (60%) are non-executives, meaning our findings largely reflect end-users or mid-level decision makers rather than top executives. All respondents presumably have some influence or interest in AI agent adoption, since they took the survey, but their organizational level might impact their pricing sensitivities, for instance, individual contributors may favor free or low-cost options if they lack budget authority.

Respondents represented a variety of industry sectors, with no single industry dominating. (See Appendix 24). Among those who did specify a sector, the most common were Software Development & Testing (~17%), Business Operations software (~15%) and Other (~23%). This spread implies our sample's pricing preferences are not tied to one vertical, but they likely reflect general attitudes towards AI agent pricing models across diverse verticals.

Finally, the survey assessed respondents' prior experience with AI agents. A majority (58.5%) indicated they have never purchased an AI agent before. About a quarter (24.5%) said they are actively exploring AI agents but have not yet purchased one. The remaining (17%) have previously purchased an AI agent for their business or personal use. In other words, over 80% of the sample have never bought an AI Agent before (See Appendix 23).

Overall, the sample demographics suggest a broad cross-section of potential AI agent users was captured. This heterogeneity in size, role, and sector helps ensure that the preference insights are not limited to one type of organization.

6.4 Pricing Model Preferences

6.4.1 MaxDiff for Pricing Models and Attributes

To quantify preferences, a MaxDiff Best-Worst exercise was employed. Respondents evaluated a series of pricing models and attribute options by picking the most and least preferred in each set. From these choices, net utility scores for each item were determined, where a higher net score indicates stronger overall preference.

The flat-rate subscription emerged as the most preferred pricing model and achieved a net utility score of +14, the highest among models. The next highest was the credit - based model with a net score of +7. The outcome - based pricing model was neutral with a net score of 0, indicating it was about equally likely to be picked as most or least preferred. This neutrality indicates a polarization with some respondents that liked the idea of paying only for successful outcomes, while others might have concerns about how vendors define and track a "successful outcome" or again the unpredictability of the pricing model. Outcome-based pricing is rare in practice, in the market sample it was only employed by 1%, so respondents may have been uncertain how to assess it, resulting in divided opinions. In contrast, the usage-tied models scored negative: the license fee with an allowance + overage fee had a net score of -12, and the pure usage - based model, pay-per-use with no base, scored lowest at -17 (See Appendix 25).

The order of preferences is therefore: Flat > Credit > Outcome > Overage > Usage. This provides evidence supporting the hypothesis of a flat-rate bias, users gravitated strongly toward a flat-rate model even without considering price levels. Beyond pricing model preferences, the MaxDiff also measured what general attributes of a pricing model are most valued. Four attributes were tested: Cost Predictability, Cost Transparency, Simplicity, and Fairness. The net utility scores show that Cost Predictability is by far the most important attribute, with a net score of +23, the highest score observed in the MaxDiff results. Cost Transparency, scored 0 or neutral. Simplicity scored -12, and Fairness scored -11 (See Appendix 26). In summary, the MaxDiff results reveal that what users value most is flat-rate pricing and predictability of cost.

Table 1: MaxDiff Pricing Models

Pricing model	net score
usage	-17
credit	7
flat	14
outcome	0
allow	-12

Source: (Author's own computation)

Table 2: MaxDiff Attributes

Attribute	net score
predic	23
transp	0
simple	-12
fair	-11

Source: (Author's own computation)

6.4.2 Statistical Tests of Preference Differences

To verify that the differences observed above are statistically significant and not due to chance or sampling variation, non-parametric tests appropriate for related proportions data were conducted. Specifically, Cochran's Q test to check if there are overall differences in the share of respondents preferring each item, and McNemar's tests with Holm-adjusted p-values for multiple comparisons for post-hoc pairwise comparisons between specific items.

For the pricing models, Cochran's Q test confirmed a highly significant difference in preference distributions ($Q(4) = 45.75$, $p < 0.001$, for five related dichotomous variables indicating whether each model was chosen as "most preferred" by each respondent at least once (See Appendix 29). This meant that not all pricing models were equally likely to be preferred, some were chosen by significantly more people as a top choice than others. Therefore a pairwise comparison between pricing models using McNemar tests was conducted. Each test consists of a 2x2 table and looks at: "did the person prefer Flat vs did the person prefer Credit" and checks for imbalance. The results showed that flat-rate was significantly more preferred than nearly every other model. In particular, the proportion of respondents choosing flat-rate as their "most preferred" model was significantly higher than those choosing usage-based ($p < .001$) and overage-based ($p < .001$), both in favor of flat-rate. By contrast, the differences Flat vs. Credit ($p = .126$) and flat vs. outcome-based ($p = .126$) did not reach statistical significance, indicating that credit-based and outcome-based plans were not reliably less popular than flat-rate in this sample, even though 43 percent of

users did pick flat-rate versus 25 percent for outcome-based (See Table 3). Thus, outcome pricing, while neutral in MaxDiff net score, was chosen enough so that it couldn't be ruled out as equally appealing to flat-rate under these sample conditions (See Appendix 30).

For the pricing attributes, Cochran's Q test also showed a significant overall difference ($Q(3) = 31.45, p < 0.001$), confirming that respondents disproportionately found certain attributes important (See Appendix 29). In this case cost predictability. Sequentially, McNemar pairwise tests underscores that Cost Predictability was significantly more likely to be chosen as most important than any other attribute, $p < 0.001$ for predictability vs transparency and fairness. Predictability vs transparency closely misses conventional significance for confidence intervals of 95% ($p = .059$) This statistically supports the interpretation that predictability dominates user priorities. Finally, as shown by the net scores, there was no significant difference between Simplicity and Fairness. In other words, almost everyone put predictability first, many put transparency second and simplicity vs fairness for third/fourth place was evenly split with no clear preference between those two (See Table 4).

The statistical tests confirm that the pricing model preference is significantly flat-rate and cost predictability. For practitioners, this means there is a statistically validated preference in the sample that favors certain pricing models and attributes, flat-rate and predictable, over others, metered and unpredictable.

Table 3. Pair-wise McNemar Tests Comparing Flat-Rate to Alternative Pricing Models (Holm-adjusted p-values)

comparison	chi2	p_unadj	p_adj
Flat vs Usage	15.38461538	8.77E-05	2.63E-04
Flat vs Credit	3.457142857	0.06297905121	0.1259581024
Flat vs Outcome	2.777777778	0.09558070455	0.1259581024
Flat vs Overage	17.64	2.67E-05	1.07E-04

Source: (Author's own computation)

Table 4. Pair-wise McNemar Tests Comparing Pricing Attributes (Holm-adjusted p-values)

comparison	p_adj
predic vs transp	0.05875276778
predic vs simple	0.0002985740093
predic vs fair	0.0002985740093
transp vs simple	0.2968804621
transp vs fair	0.2968804621
simple vs fair	1

Source: (Author's own computation)

6.4.3 Coding of themes in open-ended explanations for pricing model choice

When asked for the reasons behind their top pricing model choice, value alignment led all themes (28.6%): respondents wanted a model “aligned with our value proposition & outcomes.” Next, cost predictability (22.9%) as e.g through: “my CS budget is fixed, flat fee suits me.” Risk minimization accounted for 17.1% of replies, with an emphasis on trialability “barrier-to-entry zero: try it out.” Further, 14.3% noted simplicity: “most simple model, needs minimal attention”. The remaining responses were split among fairness, experience, and flexibility each $\approx 2.9\%$, with 8.6% of answers marked as N/A, Not Applicable, as they were not interpretable as e.g., “Regarding evolution of the model observing close-up results.”.

Theme	n	% of respondents	Example quotation
Value alignment	10	28.6%	“Because it aligns with our value proposition & outcomes.”
Cost predictability	8	22.9%	“My CS budget is fixed I need the replacement to be fixed.”
Risk minimization	6	17.1%	“Barrier-to-entry zero: try it out before buying.”

Simplicity	5	14.3%	“It’s the most simple model & needs minimal attention.”
Other (N/A)	3	8.6%	“Regarding evolution of the model observing close-up results.”
Fairness	1	2.9%	“My usage varies, so I don’t want to pay more than my usage.”
Experience	1	2.9%	“Most experience with this pricing.”
Flexibility	1	2.9%	“Flexibility.”

Having established which pricing models are most desired, the aim is to investigate who is more likely to prefer a flat-rate model, and whether it is predictable based on the attribute importance of demographic background.

6.5 Predictors of Flat-Rate Preference

6.5.1 Binary Logistic Regression Analysis

It was assumed that certain users, particularly those who place high importance on simplicity or predictability, would be more inclined to prefer a flat-rate model. To test this, a binary logistic regression model where the dependent variable was, 1 = chose flat-rate as their preferred pricing model, 0 = chose a different model. The key independent variables were the dummy indicators of which pricing attribute each person thought of most important: `most_predic`, `most_transp`, `most_simple`, `most_fair`. These dummies serve as proxies for an individual's preference. By including these variables, the aim is to answer the question: “Is a person whose top priority is X significantly more or less likely to choose a flat plan than someone whose top priority is Y?”

However, due to the small sample and the categorical nature of the attributes, the model encountered some quasi-separation issues, leading to large standard errors for some coefficients. In fact, one predictor `most_fair` had to be excluded from the model because none of the flat-preferring respondents had fairness as their top attribute (See Appendix 31).

The regression coefficients, odds ratios, did align qualitatively with expectations in a couple of cases, but none were statistically significant at conventional levels. For instance, the coefficient for `most_simple` was positive and quite large: $\beta \approx +6.0$ in log-odds, implying an

odds ratio of about $e^6 \approx 400$ for flat preference if simplicity was top priority. However, the standard error on this estimate was very high, on the order of 10^{14} and the 95% confidence interval was very wide from ~ 0.07 to 178.0 in odds ratio space. This is a sign of separation, essentially, almost all respondents who valued simplicity might have indeed chosen flat making the effect appear huge, but because the sample is relatively small, it wasn't possible to estimate its reliability.

For the other attributes: `most_predic`, predictability focused individuals, had a positive but small coefficient ($\beta \approx +1.31$) but with a very large standard error as well, with $p \approx 0.78$, not significant. The lack of a clear effect here is likely because many respondents in the sample value predictability, so that it was the top attribute for the majority, thus it doesn't differentiate sufficiently. The `most_transp` predictor had a coefficient around $\beta \approx +0.45$ with odds ratio ~ 1.57 for flat-rate preference if transparency is top priority, but again not significant ($p \approx 0.48$). Finally, `most_fair` couldn't be evaluated due to no one in that category choosing flat-rate.

Overall, the logistic regression did not achieve any statistically significant predictors of flat-rate preference. Thus, the conclusion is that within the sample, it wasn't feasible to confidently predict who will prefer flat-rate based on the measured pricing attribute priorities. This could be due to the limited sample size, $N=53$ does not give sufficient power and the relatively homogeneous preference for cost predictability.

The direction of effects with positive coefficients for simplicity, predictability and transparency, suggests the hypothesis isn't completely off, but there is lack of evidence to reject the null hypothesis. The wide confidence intervals and some separation issues make the binary logistic regression unreliable, which is why it is complemented with Fisher's exact tests and Cramer's V next.

As Creswell (2014) emphasizes, quantitative analyses require sufficient sample size to detect relationships. Therefore, the null findings here should be interpreted with caution given the sample. The lack of significance could well be a Type II error, due to failing to detect a real effect due to low power. Moreover, the quasi-complete separation illustrates how small sample logistic regressions can produce unstable estimates.

Table 5: Binary-logistic regression predicting preference for the flat-rate model (1 = chose flat as “most preferred”, 0 = did not)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.67	0.912870924 1	-0.44416477 44	0.6569234579	0.09	4.02
most_predic	1.31	0.983494735 9	0.276497377 7	0.7821660735	0.19	11.07
most_transp	0.45	1.125462579	-0.70949288 87	0.4780186629	0.05	4.63
most_simple	6	1.443375456	1.241367561	0.2144699969	0.43	178.02
most_fair	NA	NA	NA	NA	NA	NA

Source: (Author's own computation)

6.5.2 Fisher's Exact Tests and Cramér's V

Given the limitations of the logistic regression, a series of simpler Fisher's Exact tests were performed to assess if there is a correlation between flat-rate preference and each attribute priority, one at a time. Fisher's test is more suitable for small sample contingency tables as it doesn't rely on large sample chi-square approximations. Each pricing attribute: predictability, transparency, simplicity and fairness, was cross-tabulated with whether a respondent's top attribute was that attribute (yes/no) against whether their preferred model was flat-rate (yes/no). This results in a 2×2 table for each attribute, and the Fisher's p-value and the Cramér's V effect size for each table were computed in R (See Appendix 32-34).

Consistent with the regression, none of the associations were statistically significant at the $p < 0.05$ level. The p-values were: *predictability vs flat preference*: $p = 0.78$ ($V = 0.08$); *transparency vs flat*: $p \approx 0.115$ ($V = 0.23$); *simplicity vs flat*: $p \approx 0.154$ ($V = 0.24$); *fairness vs flat*: $p = 1.00$ ($V = 0.02$). These confirm that there is no strong evidence of dependence between any single attribute priority and choosing flat-rate pricing model as most preferred. However, the magnitude of the effect sizes, Cramér's V, for transparency and simplicity are around 0.23 - 0.24, which is a moderate correlation. Although not significant given the sample, this suggests a possible trend for respondents who rated pricing simplicity or cost transparency as their top priority were somewhat more likely to choose the flat-rate pricing model than those who did not. But due to limited significance, it was refrained from over interpretation.

To sum up, it wasn't possible to statistically determine significant drivers of flat-rate preference among the stated attribute priorities in this sample. The high overall popularity of flat and of predictability as a trait made it hard to discriminate against a particular profile of individuals preferring flat-rate pricing. Slight evidence points to those prioritizing simplicity/transparency for having a flat-rate preference, however, given the data limitations, these insights are statistically not reliable. Further, as the flat-rate preference is seen across

demographics, it emphasizes the universal nature of a potential flat-rate bias and application of the TAM.

Table 6: Fisher's Exact Test

attribute	p_value
most_predic	0.78
most_transp	0.115
most_simple	0.154
most_fair	1

Source: (Author's own computation)

Table 7: Effect size (Cramér's V)

attribute	cramers v
most predic	0.08
most transp	0.23
most simple	0.24
most fair	0.02

Source: (Author's own computation)

6.6 Market vs. Survey Gap Analysis

101 commercial AI-agent offerings were audited and their primary pricing models were recorded. The result was as follows: Credit-based 66%, License + Overage 22%, usage-based 9%, flat-rate 2%, and outcome-based <1%, one provider (See Appendix 13). By contrast, the survey shows consumers prefer flat-rate 43.4%, outcome-based 24.5%, credit-based 22.6%, usage-based 5.7%, and Overage 3.8%. Flat-rate and outcome models are heavily underrepresented relative to demand, while credit-based plans dominate despite lower consumer interest (See Appendix 14). Most offerings used one-part tariffs (75.2%) with single charging components, while 23.8% employed three-part tariffs combining base fees, allowances, and overages. Two-part tariffs were rare (1.0%) (See Appendix 15). Regarding pricing metrics: Credits (19 offerings) and per-user pricing (18 offerings) were most common, followed by minutes (10), messages (7), and features (7) (See Appendix 16). Tiered pricing dominated (90.1%), with hard thresholds (57.4%) more common than soft thresholds (32.7%) (See Appendix 17). Finally, 44% of audited vendors provide a freemium plan or trial (See Appendix 18), yet 86.8% of respondents said a free tier would increase their signup likelihood. This gap suggests vendors could potentially increase customer acquisition by offering a free plan. Notably, the 86.8% freemium preference suggests users want to mitigate adoption risk via trials (See Appendix 27).

6.7 Summary of findings

Cochran's Q tests confirmed significant differences in preferences across pricing models and attributes. In the sample of the survey, flat-rate pricing resulted dominantly preferred with a MaxDiff net score of +14, respondents choosing it far more often than any other model with Cochran's Q $p < .001$ an holm-adjusted McNemar pairwise comparisons reinforced these preferences over usage - and overage-based pricing models. Outcome-based plans ranked second, followed by credit-based, while pay-per-use and overage approaches were least favored. These results support $H_1(1)$: *Flat-rate pricing models will be most preferred, and usage-based pricing least preferred, reflecting a flat-rate bias* (Lambrecht & Skiera, 2006)

Further, examining the prioritization of pricing attributes, cost predictability generated a net utility of +23, well above transparency, simplicity and fairness, underscoring that users prioritize cost predictability in order to forecast and budget expenses. Efforts to link flat-rate preference to demographics or to attribute priorities achieved no significant statistical predictors, suggesting this bias is widespread across attributes preferences, roles, firm sizes and prior AI experience. These results do not support H₁(2): *Users who value simplicity and predictable costs, aligning with TAM notions of ease-of-use and the psychological factors of the Flat-Rate Bias Theory, will favor flat plans* (Davis, 1989), (Lambrecht & Skiera, 2006)

Yet the market audit of 101 AI-agent offerings shows a very different picture: 66 percent employ credit-based pricing and 22 percent usage/overage, while just 2 percent offer true flat-rate subscriptions and fewer than 1 percent use outcome-based models. In other words, flat - rate and outcome - based pricing models are under offered by 41% and 24% , while credit and overage pricing models are over represented. Further, 87 percent of participants said a free tier would make them more likely to sign up, however, only 44 percent provide a free tier of the audited AI agents. These results support H₁(3): *There is a misalignment, e.g. flat-rate pricing models are under-provided relative to market demand.*

This misalignment is a central finding of the study. It raises important discussion points: If vendors adjust their models and introduce a flat - rate subscription, would they achieve a broader adoption? Conversely, if they maintain as is, is there a risk that potential users will be hesitant or churn out once they experience complex billing?

7. Discussion

This chapter interprets findings on business professionals' preferences for AI agent pricing models within the Technology Acceptance Model (TAM) and Flat-Rate Bias theory. This discussion interprets the quantitative results from the analysis: MaxDiff utilities, Cochran's Q tests, McNemar pairwise comparisons, logistic regressions, Fisher's exact tests, and thematic coding, through the lens of the TAM and flat-rate bias while assessing the hypotheses.

7.1 Theoretical Interpretation

The results can be interpreted through the perspective of the Technology Acceptance Model (TAM) and the flat-rate bias theory. TAM states that users' perceived usefulness (PU) and perceived ease-of-use (PEOU) drive technology adoption (Davis, 1989). Pricing is an external factor influencing these perceptions. A flat-rate subscription simplifies users' budgeting capabilities, increasing PEOU, by removing the need to monitor usage and enhances perceived usefulness by making costs predictable. The data suggest that even in the context of AI agents, where actual costs were not specified, the very structure of "unlimited use for one premium price" is dominantly more attractive. The statistical significant preference of flat-rate and cost predictability aligns with this. On the other hand, complex metered pricing such as usage, overage or also outcome - based pricing likely reduce perceived ease-of-use, explaining their unpopularity.

The flat-rate bias theory (Lambrecht & Skiera, 2006) provides a psychological explanation. Participants likely preferred a flat fee to avoid the risk of unpredictable high charges, valuing the cost predictability of a fixed rate, in line with the “insurance effect”. This matches the high net utility for predictability. Consequently, pricing models that trigger a “taximeter effect” such as usage - based pricing, decreases PEOU. Overestimation and convenience effects may further reinforce flat-rate bias as users may overestimate their future usage, so flat-rate seems economically safer, and simply prefer the convenience of one all-inclusive fee. In sum, these four biases explain why respondents strongly gravitated toward flat plans even without further cost information taken into account.

Outcome-based pricing, despite its variable nature, ranking as second most preferred pricing model is an interesting finding. Paying per successful outcome may similarly reduce perceived risk as users only pay when value is delivered. This can enhance perceived usefulness, as cost only occurs if the agent performs, and ease-of-use through no need to track usage metrics excessively. Thus, outcome-based models may offer to some degree the same attributes that make flat rates attractive.

Transparency, knowing how the price is determined, was neither strongly liked nor disliked, suggesting that while honesty and clarity are desired, they are less important than the outcome of what one pays. The relatively low importance of simplicity was a bit surprising at first, since "simple pricing" was expected to be attractive. However, this result may indicate that once cost predictability and transparency are accounted for, additional simplicity is less critical. Simplicity was frequently marked as a lower priority compared to the other factors, hence its negative net score.

Fairness being negative is also interesting, one might assume everyone wants fair pricing. However, the negative net here does not mean people want unfair pricing, but rather that when forced to prioritize, other factors are prioritized. It's possible that respondents assume their definition of fairness will be met if predictability and transparency are in place, thus they did not often choose "fairness" as the top attribute on its own. Another possible interpretation is, in the context of AI agent pricing, the respondents were willing to sacrifice some pay-per-use or pay-per-outcome "fairness" in exchange for the peace of mind of a predictable cost through a flat-rate.

Furthermore, the lack of prior purchase experience could explain the strong interest in a freemium model, since trying the technology at low risk would be valuable to these users. Many are potential first time adopters, for whom pricing structure could significantly influence the decision to purchase an AI agent. This highlights the importance of factors that ease adoption, consistent with TAM's emphasis on external variables affecting PU and PEOU (Davis, 1989).

Finally, neither logistic regression nor Fisher tests with Cramer's V found significant links between pricing attribute preference flat-rate preference. However, the small to moderate effect sizes for attributes like simplicity suggest slight trends that those valuing simplicity were more likely to choose flat, but these were not statistically significant. Overall, the

findings reinforce the TAM, that when a pricing model offers lower complexity and financial risk, thus raising PEOU and PU, users across demographics prefer it. In short, the results illustrate how TAM (Davis, 1989) and flat-rate bias theory (Lambrecht & Skiera, 2006) explain the observed flat-rate dominance and the strong emphasis on cost predictability.

8. Conclusion

8.1 Main findings

This study found a flat-rate bias in pricing model preferences for AI agents. In the MaxDiff analysis, the flat-rate subscription plan achieved the highest net utility (+14), far above all other models, while pay-per-use (-17) and license+overage models (-12) scored lowest. Credit-based plans scored (+7), and outcome-based pricing was neutral (0). Cochran's Q tests confirmed these differences as highly significant for pricing models $Q(4)=45.75$, $p<.001$, and for pricing attributes $Q(3)=31.45$, $p<.001$. Pairwise comparisons showed flat-rate was significantly preferred over usage and overage-based plans ($p<.001$), while flat vs. credit or outcome plans were not significantly different ($p\approx.126$). In practical terms, 43.4% of respondents ranked flat-rate as their top choice, whereas outcome-based plans were second most popular with 24.53% and credit-based as third most preferred with 22.64%.

Regarding pricing attributes, cost predictability was voted the most important of all others with a net utility of +23, statistically higher than simplicity and fairness (all pairwise $p<.001$). Predictability vs transparency closely misses conventional significance for confidence intervals of 95% ($p = .059$). This indicates users overwhelmingly prioritize the ability to forecast and budget AI agent expenses. Simplicity and fairness were rated least important. Notably, the logistic regression, exact Fisher test and Cramer's v did not achieve any statistically significant predictors of flat-rate preference based on pricing attribute preference.

A market audit revealed a large gap between user desires and vendor offerings. While 66% of AI agent services use credit-based pricing and only ~2% offer flat subscriptions, the opposite is true in preferences, only 22.6% of users preferred credit plans vs. 43.4% preferring flat. Further, outcome-based plans 1% of offerings vs. 24.5% user preference. Finally, freemium emerged as a potentially relevant customer acquisition strategy with 86.8% of respondents saying a free trial tier would increase their likelihood to sign up, yet only 44% of vendors currently offer any free plan. In sum, the empirical results support a flat-rate bias and cost predictability preference for AI agent pricing, thereby supporting hypotheses 1 and 3 and rejecting hypothesis 2. Further a misalignment between customer preferences and market practice is highlighted.

8.2 Practical and Managerial Implications

These insights suggest a few recommendations and or considerations for AI agent vendors. First, realign pricing models to meet customer preferences. The current market, 66% credit-based models vs only 2% flat, is misaligned with demand, 43% prefer flat vs 25% outcome (See Appendix 27). Therefore vendors should consider flat-rate subscriptions and

outcome-based plans inline with their underlying cost structure. By offering a flat - rate subscription, firms can leverage a large segment that values predictability. However, the difficulty is to design a flat - rate subscription that is not too high and diminishes adoption nor too low and causes losses due to excessive token costs. Similarly, outcome-based offerings could appeal to customers who want performance guarantees. But outcome-based pricing comes with a set of challenges as discussed earlier. If implemented successfully, these “most preferred” pricing models position a vendor to potentially achieve higher customer acceptance and thereby “most effective pricing models”. A potential solution could be a hybrid of a lower fixed fee + a fee per successful outcome, a two - part tariff, that way vendors can mitigate the risk of potential losses by charging a minimum fee, align price with value and customers have a higher degree of cost predictability then with a pure usage-based model.

Second, adopting a freemium or trial model for customer acquisition. With ~87% of respondents, stating it would increase their likelihood to sign up, adding a trial could substantially increase adoption rates. Practically, vendors might offer a free basic agent or a time-limited trial of premium features, keeping in mind the costs incurred for the usage in that trial period. This lowers the barrier to initial use and increases PEOU by letting users experience the product without risk, which TAM identifies as an adoption enabler. The data indicates that providing a free tier could have an impact on customer acquisition, as it addresses customer fears by removing upfront cost and even has broad appeal among 22 large enterprises out of 24 (+ 1.000 employees) in the sample.

Third, simplify the pricing packaging and communication. Although cost transparency ranks lower than predictability, vendors should aim for high cost transparency, as the survey showed customers often dislike metered models, and transparent pricing improves perceived ease of use by reducing complexity and thereby potentially increasing the adoption likelihood, despite metered models. CFO’s need to be able to understand the underlying pricing structure in order to be able to create budgets, forecast costs and calculate ROI’s.

Finally, vendors should differentiate on pricing innovation. Given that credit and overage - based plans are overrepresented in the market compared to customer desire, offering flat or outcome-based alternatives can be a competitive advantage. For example, an AI analytics provider might introduce a premium unlimited plan or a per “successful” insight pricing fee. In summary, when vendors set a strategy for monetization of their agents, they do not only need to take into account their underlying cost structure, but also consider an existing flat-rate bias, substantial preference for outcome-based pricing and demand for trials, as revealed by this study, which could substantially decrease user acquisition and satisfaction and thereby pricing model effectiveness, if disregarded.

8.3 Limitations and Validity Threats

Several limitations constrain these conclusions. The sample was relatively small (N=53), and as Creswell (2014) notes, small to moderate convenience samples and cross-sectional surveys limit generalizability. Here, respondents were mostly from the PricingSaaS community,

which may not represent all AI users. The lower sample size also means lower statistical power, indeed, the logistic regressions showed no significant predictors, likely due to Type II error. Therefore, some true effects by demographic or pricing attribute preference have likely been missed.

The within-subjects experimental design improves internal validity, each participant saw all models, but external validity is constrained. Preferences were gathered for a hypothetical AI agent with equalized costs. Actual purchasing behavior in a real market context with varying prices and brand factors may differ. The cross-sectional snapshot captures opinions at one point in time, preference attitudes could evolve as users gain experience or as industry norms change. Moreover, respondents might exhibit hypothetical bias by stating preferences in a survey that they might not reveal in real buying situations.

Finally, self-selection and survey framing pose potential threats. The sample likely overrepresents professionals interested in pricing, and underrepresents those unfamiliar with AI. The use of single Most/Least choice self-reports might lead to extreme results relative to respondents' more nuanced underlying preferences. All of these factors together: sampling bias, low N, survey design, affect external validity and power. Creswell (2014) would caution that these validity threats mean the findings, while internally consistent, should be generalized carefully. In practice, these limitations invite viewing the results as strongly suggestive rather than definitive evidence across all contexts.

8.4 Future Research Directions

To build on this study, future work should pursue larger and more varied samples. A replication with a much bigger sample would test whether the flat-rate bias for AI agents indeed significantly holds across industries, firm sizes, and international markets and would increase the possibility to determine a statistical significance of psychological and contextual predictors of preferring a specific pricing model. Similarly, given the current cross-sectional design, longitudinal studies are needed, with the aim of tracking how pricing preferences change as respondents actually adopt and use AI agents over time to validate whether stated intentions translate into behavior. Field experiments e.g. offering different pricing models to randomized user groups could measure real sign-up and retention outcomes under flat vs. usage pricing.

The findings of outcome-based models being the second most preferred pricing models, invites for further research to determine the underlying psychological reasoning for this. Furthermore, studies on conversion rates and profitability of outcome-based models would give practitioners practical insights into its effectiveness besides customer acceptance.

Finally, behavioral validation through operational data is ideal. Partnering with AI agent vendors to analyze actual purchase and usage logs would reveal how pricing models affect conversion, adoption and churn. If one provider introduces a flat-rate tier, did adoption increase as predicted? This data could confirm or refine the patterns discovered in this sample. Such longitudinal and experimental research would address the stated limitations:

sample size, external validity, hypothetical bias and strengthen the theoretical model. Overall, pursuing these directions would increase the validity of the findings.

8.5 Conclusion

This research aimed to identify the most effective pricing models for AI agents in today's dynamic technology landscape. Following Frohmann (2018), effectiveness was operationalized as consumer acceptance, grounded in the principle that "Criteria such as the acceptance of the price model from the customer's point of view are core prerequisites for market success". Thus, following Frohmann (2018), an effective AI agent pricing model is one that achieves high customer acceptance. The answer derived from this acceptance-centric approach is that flat-rate and outcome-based are the most effective pricing models. These models achieved the highest consumer acceptance, primarily due to their alignment with user preferences for cost predictability and transparency, revealing a significant discrepancy with current market offerings.

This study systematically addressed its sub-research questions. AI agents are autonomous, LLM-driven systems, uniquely characterized by a cost structure with substantial variable costs tied to API calls and token consumption, distinguishing them from traditional SaaS. This variable cost nature exposes the limitations of conventional SaaS pricing models, which often fail to adequately manage these variable expenses without either risking vendor profitability or imposing unpredictable costs on customers. While vendors experiment with various models such as usage-based, outcome-based, FTE replacement, and workflow-based, this research aims to identify and explain customer preference for these models, and thus acceptance, through the Technology Acceptance Model (TAM) and Flat-Rate Bias theory. These frameworks highlight that perceived ease of use, usefulness, and psychological inclinations towards risk minimization through cost predictability are key drivers of acceptance.

The operationalization of effectiveness through Frohmann's (2018) consumer acceptance framework shifted the analytical lens from purely economic metrics to the behavioral determinants of market adoption, thereby uncovering a significant gap between the pricing models vendors currently offer and those that customers prefer the most. A notable 43.4% of participants preferred flat-rate models, and the attribute of cost predictability achieved a dominant MaxDiff net utility score of +23. This contrasts significantly with the current market offerings where only approximately 2% of AI agent vendors offer flat-rate subscriptions, while a majority (66%) lean on credit-based systems of the sample.

Based on these consumer acceptance findings, hypothesis H1(1) stating flat-rate models would be most preferred and usage-based least, was supported by consumer acceptance data. Further, H1(3) positing a misalignment with flat-rate models being under provided, was strongly supported by the market audit versus consumer acceptance findings. However H1(2), linking a preference for flat plans to pricing attribute prioritization, was rejected. Overall these results confirm the Flat-Rate Bias theory within the domain of AI agents and

extend TAM by underscoring pricing as a pivotal external variable influencing PU and PEOU and thereby pricing model acceptance.

The practical implications for businesses aiming to successfully monetize their AI agents, through price model acceptance, as defined by Frohmann (2018), are clear. Vendors should strategically consider integrating flat-rate and outcome-based plans into their offerings. According to the findings, from Medina's (2025) framework for AI agent pricing, thus, per Agent pricing through a flat fee and per Agent Outcome are expected to be the most effective based on customer acceptance. Moreover, the strong indication that freemium models or trials would positively influence sign-ups for 86.8% of users points to a powerful lever for enhancing acceptance and initial adoption.

In reinforcing its contribution, this thesis demonstrates the importance of an acceptance driven approach to pricing models for AI agents. By leveraging the Frohmann (2018) operationalization of acceptance, this research not only identifies the most effective pricing models but also provides a theoretically grounded understanding through the TAM (Davis, 1989) and Flat-Rate Bias theory (Lambrecht & Skiera, 2006) of why these models resonate with consumers, providing insights to effective pricing models for AI agents.

9. References

- Abonamah, A. A., Tariq, M. U., & Shilbayeh, S. (2021). *On the Commoditization of Artificial Intelligence*. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.696346>
- Acharya, D. B., Kuppan, K., & Divya, B. (2025). *Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive survey*. *IEEE Access*, 1. <https://doi.org/10.1109/access.2025.3532853>
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2018). *Human Judgment and AI Pricing*. *AEA Papers and Proceedings*, 108, 58–63.
- Alvarez, F., & Jurgens, J. (2024). *Navigating the AI frontier: A primer on the evolution and impact of AI agents* (White paper). World Economic Forum.
- Alvaro. (Jun 18, 2024). *The misunderstood AI wrapper opportunity*. Medium. Retrieved March 4, 2025, from https://medium.com/@alvaro_72265/the-misunderstood-ai-wrapper-opportunity-afabb3c74f31
- Anthropic. (2024, June 20). *Claude 3.7 Sonnet and Claude Code* <https://www.anthropic.com/news/claude-3-7-sonnet>
- Artificial Analysis. (n.d.a). *Comparison of AI models across intelligence, performance, price*. Retrieved March 30, 2025, from <https://artificialanalysis.ai/models>
- Artificial Analysis. (n.d.b). *Intelligence Benchmarking*. Retrieved March 30, 2025, <https://artificialanalysis.ai/methodology/intelligence-benchmarking>
- Bashir, I. (2024, June 5). *How to Achieve Profitable Growth through Your Pricing Strategy*. *Amplitude.Com*. <https://amplitude.com/blog/monetization-pricing-strategy>
- Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. 'The rising costs of training frontier AI models'. ArXiv [cs.CY], 2024. arXiv. <https://arxiv.org/abs/2405.21015>.
- Ben Cottier et al. (2025), "LLM inference prices have fallen rapidly but unequally across tasks". Published online at epoch.ai. Retrieved from: 'https://epoch.ai/data-insights/llm-inference-price-trends' [online resource]
- Benram, G. (2025, March 5). *Understanding the cost of Large Language Models (LLMs)*. *TensorOps*. <https://www.tensorops.ai/post/understanding-the-cost-of-large-language-models-llms>

- Bergemann, D., Bonatti, A., & Smolin, A. (2025). *The economics of large language models: token allocation, Fine-Tuning, and optimal pricing*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2502.07736>
- Bousquette, I. (2025, February 1). *No one knows how to price AI tools*. The Wall Street Journal. <https://www.wsj.com/articles/no-one-knows-how-to-price-ai-tools-f346ea8a>
- Bousetouane, F. (2025). *Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2501.00881>
- Britannica Dictionary. (n.d.a.). *Artificial Definition & Meaning*. Retrieved March 29, 2025, from: <https://www.britannica.com/dictionary/artificial>
- Britannica Dictionary. (n.d.b). *Intelligence Definition & Meaning*. Retrieved March 29, 2025, from: <https://www.britannica.com/dictionary/intelligence>
- Britannica Dictionary. (n.d.c). *Artificial Intelligence Definition & Meaning*. Retrieved March 29, 2025, from: <https://www.britannica.com/dictionary/artificial-intelligence>
- Buxmann, P., Diefenbach, H., & Hess, T. (2009). *Pricing strategies of software vendors*. *Business & Information Systems Engineering*, 1(6), 452–462. <https://doi.org/10.1007/s12599-009-0075-y>
- Carmichael-Jack, J. (2024, December 13). *The story behind the "Stop Hiring Humans" billboards in San Francisco*. Artisan. Retrieved April 8, 2025, from <https://www.artisan.co/blog/stop-hiring-humans>
- Casado, M., & Bornstein, M. (2024, April 25). *The New Business of AI (and How It's Different From Traditional Software)*. Andreessen Horowitz. <https://a16z.com/the-new-business-of-ai-and-how-its-different-from-traditional-software/>
- Chao, Y. (2013). *STRATEGIC EFFECTS OF THREE-PART TARIFFS UNDER OLIGOPOLY*. *International Economic Review*, 54(3), 977–1015. <http://www.jstor.org/stable/24517073>
- Chargebee. (n.d.). *Plans and Pricing*. Retrieved April 19, 2025, from <https://www.chargebee.com/pricing/>
- Chargeflow. (n.d.). *Pricing*. Retrieved April 20, 2025, from <https://www.chargeflow.io/pricing>
- Chuttur M.Y. (2009). "Overview of the Technology Acceptance Model: Origins, Developments and Future Directions ," Indiana University, USA . Sprouts: Working Papers on Information Systems, 9(37). <http://sprouts.aisnet.org/9-37>
- Creswell, J. W. (2014). *Research design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.

- Davis, F. D. (1989). *Perceived usefulness, perceived ease of use, and user acceptance of information technology*. MIS Quarterly, 13(3), 319. <https://doi.org/10.2307/249008>
- Dennis Rall, Bernhard Bauer, and Thomas Fraunholz. 2023. *Towards Democratizing AI: A Comparative Analysis of AI as a Service Platforms and the Open Space for Machine Learning Approach*. In Proceedings of the 2023 7th International Conference on Cloud and Big Data Computing (ICCBDC '23). Association for Computing Machinery, New York, NY, USA, 34–39. <https://doi.org/10.1145/3616131.3616136>
- Deshpande, A., Picken, N., Kunertova, L., De Silva, A., Lanfredi, G., & Hofman, J. (2021). *Improving working conditions using artificial intelligence*. RAND Europe. European Union. <http://www.europarl.europa.eu/supporting-analyses>
- DeVon, C. (2024, February 2). *It's too expensive to replace human workers with AI—for now, says MIT study*. CNBC. <https://www.cnbc.com/2024/02/02/mit-study-using-ai-to-replace-humans-may-be-too-expensive.html>
- Drucker, P. F. (1967). *The effective executive*.
- Ertel, W. (2018). *Introduction to Artificial Intelligence*. Springer.
- Fenerum. (n.d.). *Pricing*. Retrieved April 19, 2025, from <https://www.fenerum.com/da-EN/pricing/>
- Field, H. (2024, September 27). *OpenAI sees roughly a \$5 billion loss this year on \$3.7 billion in revenue*. CNBC. <https://www.cnbc.com/2024/09/27/openai-sees-5-billion-loss-this-year-on-3point7-billion-in-revenue.html>
- Frohmann, F. (2018). *Digitales pricing*. In Springer eBooks. <https://doi.org/10.1007/978-3-658-22573-5>
- Gao, J., Wang, Z., & Wei, X. (2024). *An Adaptive Pricing Framework for Real-Time AI Model Service Exchange*. IEEE Transactions on Network Science and Engineering, 11(5), 5114–5128.
- Google Cloud. (n.d.). *What are AI agents? Definition, examples, and types*. Retrieved March 4, 2025, from <https://cloud.google.com/discover/what-are-ai-agents>
- Greener, S. (2008). *Business research methods*. Bookboon.
- Gross, G. (2024, December 19). *How Will AI Agents Be Priced? CIOs Need to Pay Attention*. CIO.com. Retrieved from <https://www.cio.com/article/3624540/how-will-ai-agents-be-priced-cios-need-to-pay-attention.html>

- Hajipour, V., Hekmat, S., & Amini, M. (2023). *A value-oriented Artificial Intelligence-as-a-Service business plan using integrated tools and services*. *Decision Analytics Journal*, 8, 100302.
- He, Y., Wang, E., Rong, Y., Cheng, Z., & Chen, H. (2024). *Security of AI agents*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2406.08689>
- Hinterhuber, A. (2003) 'Towards value-based pricing—An integrative framework for decision making', Falkstrasse 16, 6020 Innsbruck, Austria.
- Hinterhuber, Andreas & Liozu, Stephan. (2012). *Is It Time to Rethink Your Pricing Strategy?*. MIT Sloan Management Review. 53. 69-77.
- Hyperline. (n.d.). *Pricing*. Retrieved April 19, 2025, from <https://www.hyperline.co/pricing>
- Iyengar, R., & Gupta, S. (2009). *Nonlinear pricing*. In Edward Elgar Publishing eBooks. <https://doi.org/10.4337/9781848447448.00025>
- Jaime Sevilla and Edu Roldán (2024), "*Training Compute of Frontier AI Models Grows by 4-5x per Year*". Published online at epoch.ai. Retrieved from: '<https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year>' [online resource]
- John, & Rob. (2025). *2025 Pricing Metric Report*. PricingSaaS. <https://pricingsaas.com/>
- Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). *Large Language Models: a deep dive: Bridging Theory and Practice*. Springer.
- Kapoor, S., Stroebel, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). *AI agents that matter*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2407.01502>
- Kienzler, M., & Kowalkowski, C. (2014). *Pricing strategy: An assessment of 20 years of B2B marketing research*. <http://impgroup.org/uploads/papers/8228.pdf>
- Klarna. (2024, February 28). *Klarna AI assistant handles two-thirds of customer service chats in its first month*. Klarna. <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>
- Kohli, C., & Suri, R. (2011). *The price is right? Guidelines for pricing to enhance profitability*. *Business Horizons*, 54(6), 563–573. <https://doi.org/10.1016/J.BUSHOR.2011.08.001>
- Lambrecht, A., & Skiera, B. (2006). *Paying Too Much and Being Happy about It: Existence, Causes, and Consequences of Tariff-Choice Biases*. *Journal of Marketing Research*, 43(2), 212–223. <https://doi.org/10.1509/jmkr.43.2.212>

Latva-Koivisto, A. (2025, March 11). *Why Value-Based Pricing of AI SAAS product failed (And what actually works)*. Deductive Innovation Insights.
<https://newsletter.antti.lk/p/value-based-pricing-ai>

Lehrskov-Schmidt, U. (2023b). *The Pricing Roadmap: How to design B2B SAAS pricing models that your customers will love*. Houndstooth Press.

Li, Y., Yang, Z., Sun, J., & Hu, X. (2022). *Pricing Strategies of AI-enabled and Regular Products*. In *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 860–864). IEEE.

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). *Artificial intelligence as a service*. *Business & Information Systems Engineering*, 63(4), 441–456.
<https://doi.org/10.1007/s12599-021-00708-w>

Liozu, S. M., & Hinterhuber, A. (2023). *Digital pricing strategy*. In Routledge eBooks.
<https://doi.org/10.4324/9781003226192>

Ma, Q., & Liu, L. (2011). *The technology acceptance model*. In IGI Global eBooks.
<https://doi.org/10.4018/9781591404743.ch006.ch000>

Mahendra, A. (2023). *AI Startup Strategy*. In Apress eBooks.
<https://doi.org/10.1007/978-1-4842-9502-1>

Mahmood, R. (2024). *Pricing and competition for generative AI*. arXiv (Cornell University).
<https://doi.org/10.48550/arxiv.2411.02661>

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *Artificial Intelligence Index Report 2024*. arXiv (Cornell University).
<https://doi.org/10.48550/arxiv.2405.19522>

Medina, M. (2025, April 9). *A new framework for AI agent pricing*. Kyle Poyar's Growth Unhinged. <https://www.growthunhinged.com/p/ai-agent-pricing-framework>

Microsoft Azure. (n.d.). *Künstliche Intelligenz als Dienst (AIaaS)*. Retrieved March 4, 2025, from [https://azure.microsoft.com/de-de/resources/cloud-computing-dictionary/what-is-aiaaS#:~:text=Service%2C%20AIaaS\)%3F-,K%C3%BCnstliche%20Intelligenz%20als%20Dienst%20\(AI%2Das%2Da%2DService,und%20%2DFunktionen%20auf%20Abonnementbasis%20erm%C3%B6glicht.](https://azure.microsoft.com/de-de/resources/cloud-computing-dictionary/what-is-aiaaS#:~:text=Service%2C%20AIaaS)%3F-,K%C3%BCnstliche%20Intelligenz%20als%20Dienst%20(AI%2Das%2Da%2DService,und%20%2DFunktionen%20auf%20Abonnementbasis%20erm%C3%B6glicht.)

Minkovski, D. (2024, November 20). *AI-Powered Customer Support: the ultimate Multi-Agent system*. Medium.
<https://medium.com/codex/ai-powered-customer-support-the-ultimate-multi-agent-system-524ba369fb2a>

- Morales, A. (2024, August 29). Council Post: *Why AI is creating a revolution in SAAS pricing models*. Forbes.
<https://www.forbes.com/councils/forbestechcouncil/2024/08/29/why-ai-is-creating-a-revolution-in-saas-pricing-models/>
- Musa, H. G., Fatmawati, I., Nuryakin, N., & Suyanto, M. (2024). *Marketing research trends using technology acceptance model (TAM): a comprehensive review of researches (2002–2022)*. Cogent Business & Management, 11(1).
<https://doi.org/10.1080/23311975.2024.2329375>
- OpenAI. (n.d.a). *Tokens*. OpenAI. Retrieved March 30, 2025
<https://platform.openai.com/docs/concepts/token>
- OpenAI. (n.d.b). *Pricing*. OpenAI. Retrieved March 30, 2025,
<https://platform.openai.com/docs/pricing>
- OpenAI. (n.d.c). *Agents*. OpenAI. Retrieved March 30, 2025,
<https://platform.openai.com/docs/guides/agents>
- PwC. (n.d). *Sizing the prize*. Retrieved April 3, 2025, from
<https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>
- Salesforce. (n.d.). *Agentforce Pricing*. Retrieved April 20, 2025, from
<https://www.salesforce.com/agentforce/pricing/>
- Saltan, A., & Smolander, K. (2019). *Towards a SAAS Pricing Cookbook: A Multi-vocal Literature review*. In Lecture notes in business information processing (pp. 114–129).
https://doi.org/10.1007/978-3-030-33742-1_10
- Saltan, A., & Smolander, K. (2021). *How SaaS Companies Price Their Products: Insights from an Industry Study*. In Lecture notes in business information processing (pp. 1–13).
https://doi.org/10.1007/978-3-030-67292-8_1
- Schulz, F., Schlereth, C., Mazar, N., & Skiera, B. (2015). *Advance payment systems: Paying too much today and being satisfied tomorrow*. International Journal of Research in Marketing, 32(3), 238–250. <https://doi.org/10.1016/j.ijresmar.2015.03.003>
- Sharma, R. (2025, January 28). *Executive Guide to AI Agent Pricing: Winning Strategies and Models to Drive Growth*. Forbes Business Council. Retrieved from
<https://www.forbes.com/sites/forbesbusinesscouncil/2025/01/28/executive-guide-to-ai-agent-pricing-winning-strategies-and-models-to-drive-growth/>
- Shipley, D., & Jobber, D. (2001). *Integrative pricing via the pricing wheel*. Industrial Marketing Management, 30(3), 301–314. [https://doi.org/10.1016/s0019-8501\(99\)00098-x](https://doi.org/10.1016/s0019-8501(99)00098-x)

Simon, H. (1992). *Preismanagement : Analyse, Strategie, Umsetzung*. Gabler.
<http://ci.nii.ac.jp/ncid/BA19654979>

Simon, H. and Fassnacht, M., 2019. *Price Management: Strategy, Analysis, Decision, Implementation*. 1st ed. Springer. Available at: <https://doi.org/10.1007/978-3-319-99456-7>.

Singh, Preet Deep. (2024). *Generative AI through the Lens of Technology Acceptance Model*. Available at SSRN: <https://ssrn.com/abstract=4953174> or <http://dx.doi.org/10.2139/ssrn.4953174>

Smit, L., & van Niekerk, T. I. (2014). *Selecting a pricing strategy : a statistical approach*. Journal for New Generation Sciences, 12(1), 141–157.
<https://journals.co.za/content/newgen/12/1/EJC159459>

Spruit, M., & Abdat, N. (2012). *The pricing strategy guideline framework for SaaS vendors*. *International Journal of Strategic Information Technology and Applications*, 3(1), 45–61.
<https://doi.org/10.4018/jsita.2012010103>

Stanford University (2024). *The 2024 AI Index Report*. Retrieved March 4, 2025, from <https://hai.stanford.edu/ai-index/2024-ai-index-report>

Statista. (n.d.). *Artificial intelligence (AI) funding and startups*. Retrieved March 4, 2025, from <https://www.statista.com/topics/12089/artificial-intelligence-ai-funding-and-startups/#topicOverview>

Syed, N., Anwar, A., Baig, Z., & Zeadally, S. (2025). *Artificial Intelligence as a Service (AIAAS) for Cloud, Fog and the Edge: State-of-the-Art practices*. ACM Computing Surveys.
<https://doi.org/10.1145/3712016>

Thales & Simon-Kucher. (2024). *The essential 4 Ps for AI monetization: A comprehensive guide to success in your AI go-to-market* [PDF].
https://www.simon-kucher.com/sites/default/files/perspectives-files/THALES_essential-4-Ps-for-AI-monetization-guide.pdf

Thammineni, P. (2025, January 21). *The Complete Guide to AI Agent Pricing Models in 2025*. Agentman (Medium). Retrieved from <https://medium.com/agentman/the-complete-guide-to-ai-agent-pricing-models-in-2025-ff65501b2802>

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L. J., & Anandkumar, A. (2023, October 19). *VOYAGER: An open-ended embodied agent with large language models*. Retrieved March 4, 2025, from <https://voyager.minedojo.org>, <https://arxiv.org/pdf/2305.16291>

Wiggers, K. (2025, March 5). *OpenAI reportedly plans to charge up to \$20,000 a month for specialized AI 'agents'* TechCrunch.

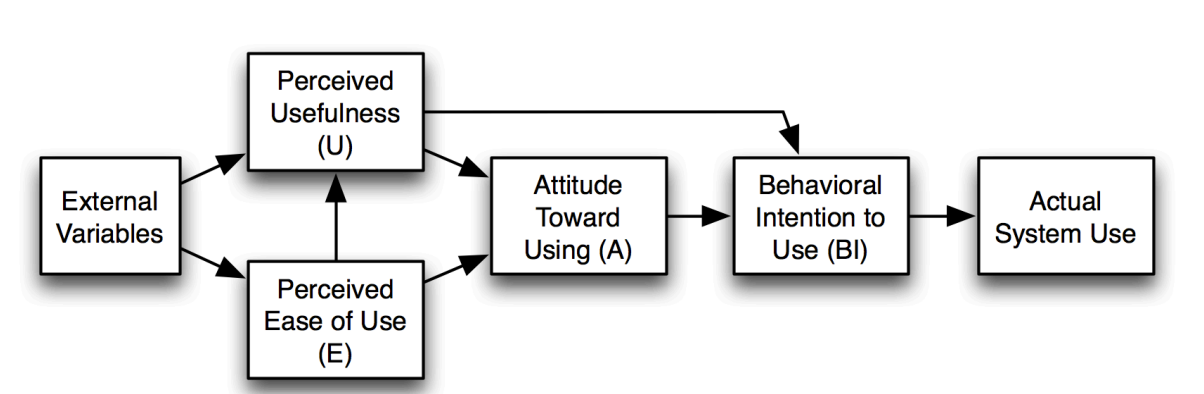
<https://techcrunch.com/2025/03/05/openai-reportedly-plans-to-charge-up-to-20000-a-month-for-specialized-ai-agents/>

Yamase, S. (2025, January 23). *The Essential 4 Ps for AI Monetization: A Guide to Success in Your AI Go-to-Market*. Simon-Kucher & Partners and Thales. Retrieved from

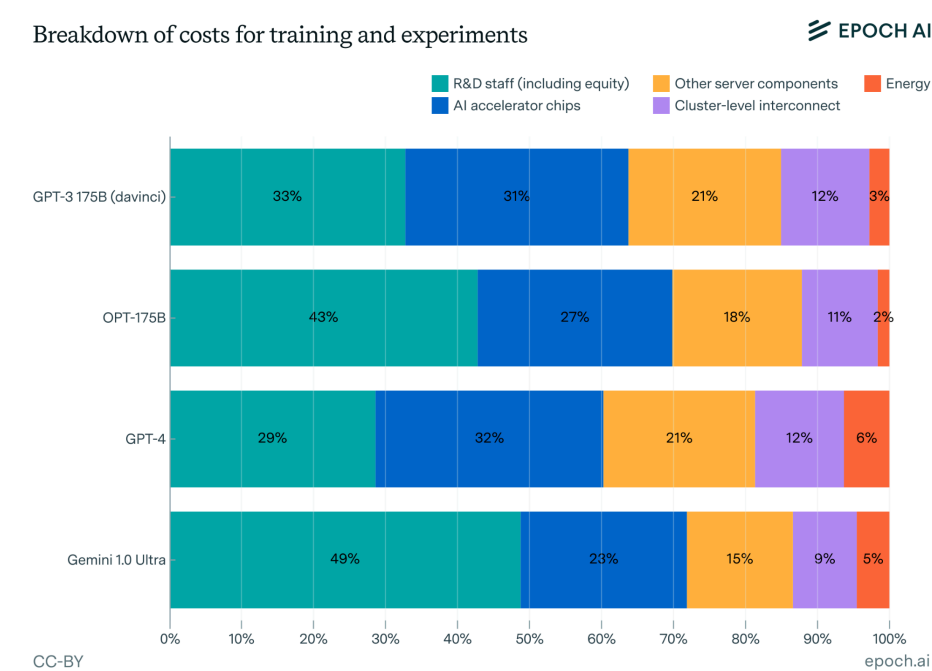
https://www.simon-kucher.com/sites/default/files/perspectives-files/THALES_essential-4-Ps-for-AI-monetization-guide.pdf

10. Appendix


Appendix 1: TAM Model (Davis, 1989)

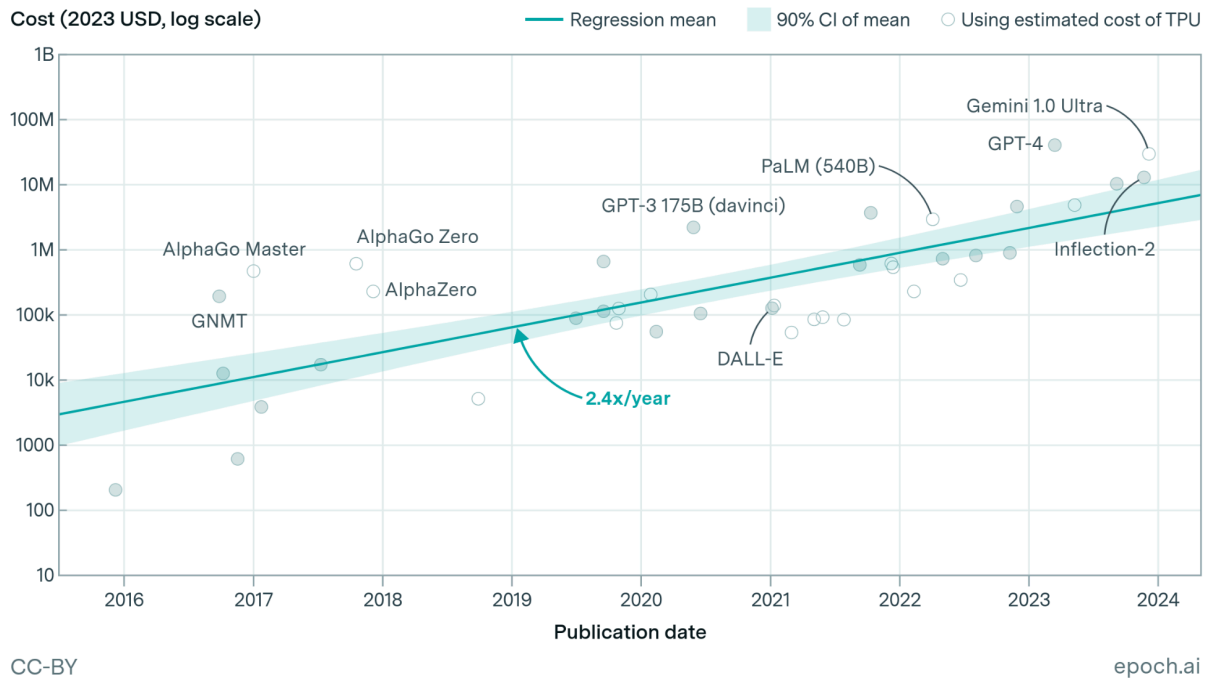


Appendix 2: Breakdown of costs for training and experiments (Cottier et al., 2024)



Appendix 3: Amortized hardware and energy cost to train frontier AI models over time (Cottier et al., 2024)

Amortized hardware and energy cost to train frontier AI models over time  EPOCH AI



Appendix 4: Tokenization at OpenAI (OpenAI, n.d.a)

GPT-4o & GPT-4o mini

GPT-3.5 & GPT-4

GPT-3 (Legacy)

Write an answer to this customer complaint E-Mail:

Hello [SaaS Provider],

Recently, our team has been struggling with frequent outages and sluggish load times on [Product Name]. These issues disrupt our workflow and delay deliverables. Could you please address these concerns and provide a clear plan to improve stability? We need faster resolutions moving forward.

Clear

Show example

Tokens

Characters

68

367

Write an answer to this customer complaint E-Mail:

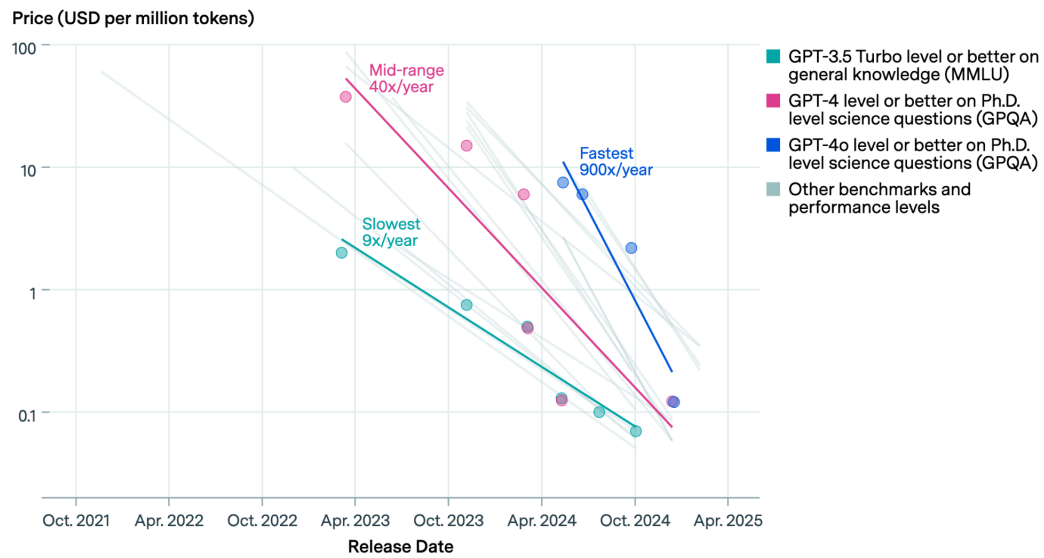
Hello [SaaS Provider],

Recently, our team has been struggling with frequent outages and sluggish load times on [Product Name]. These issues disrupt our workflow and delay deliverables. Could you please address these concerns and provide a clear plan to improve stability? We need faster resolutions moving forward.

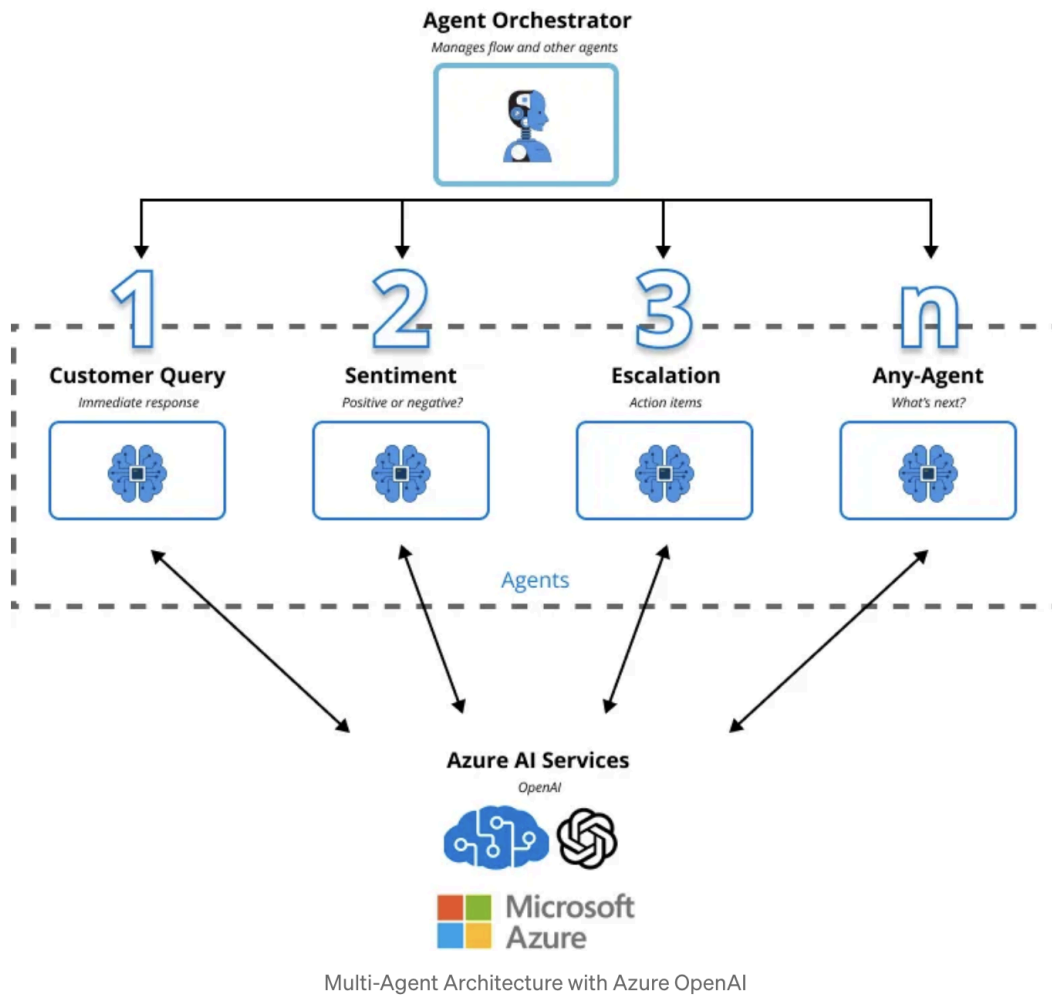
Text

Token IDs

Appendix 5: LLM inference prices have fallen 9x to 900x/year, depending on task (Cottier et al., 2025)



Appendix 6: Example of an Multi-Agent System (MAS) Minkovski (2024)

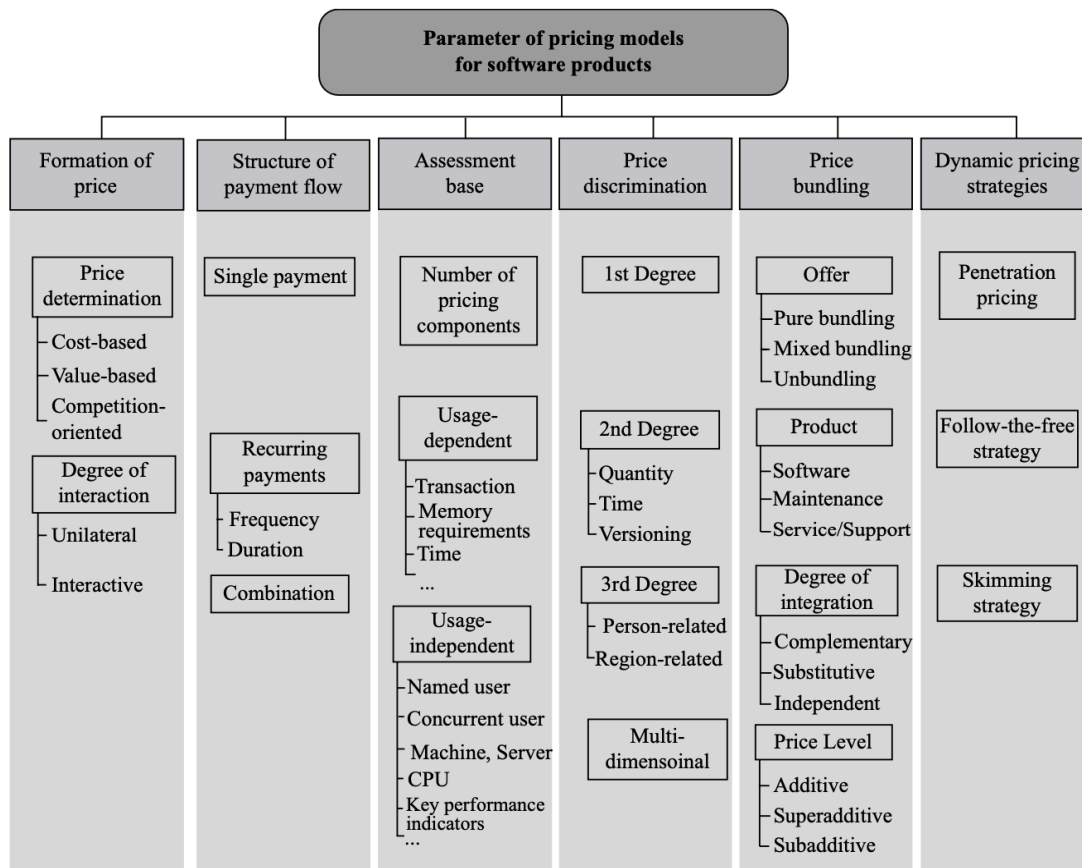


Appendix 7: Academic SaaS pricing frameworks comparison (Saltan & Smolander, 2019)

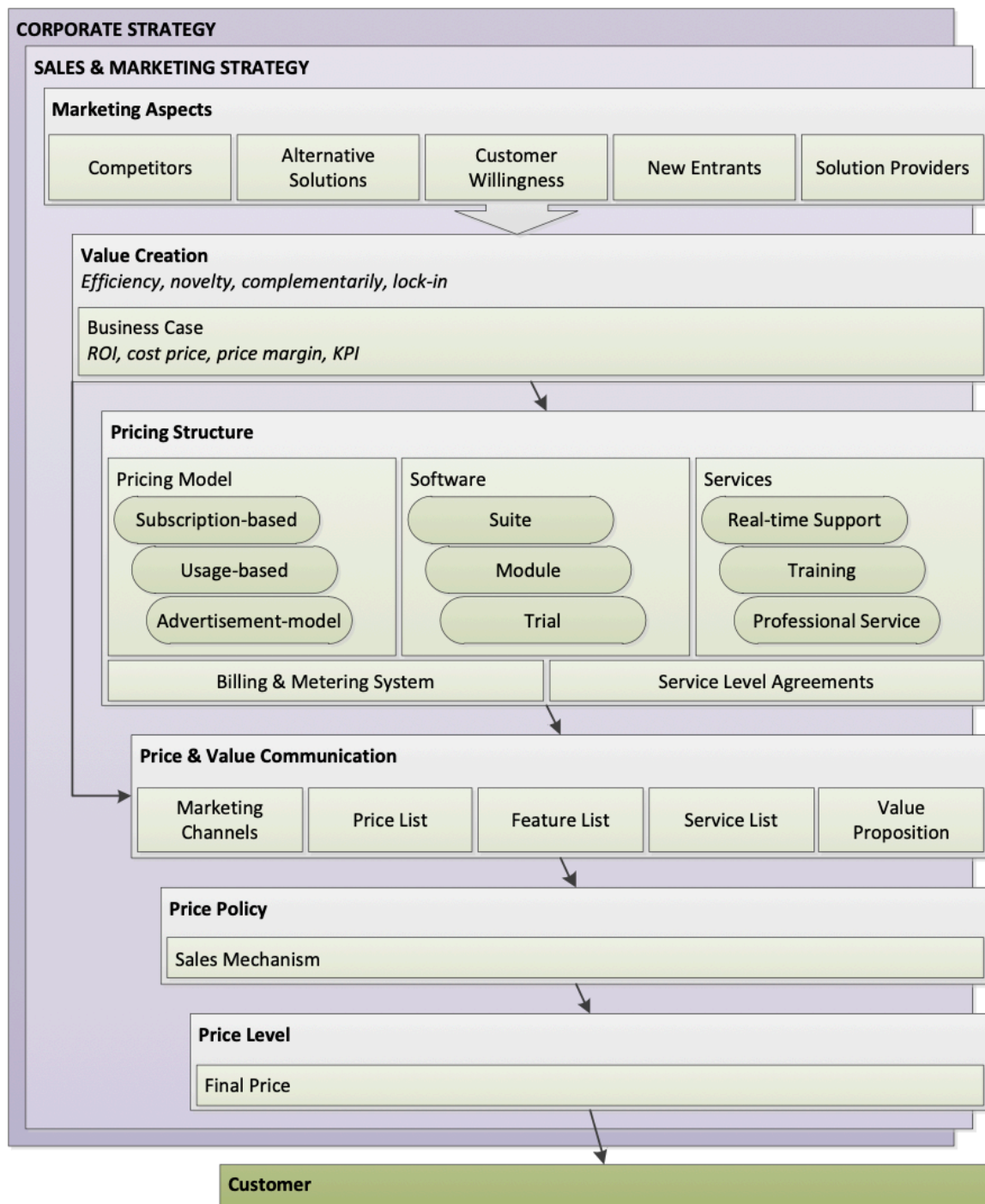
#	Name	Ref.	Perspective	Framework structure	Scientific and practical origins
F1	Customer-centric value-based pricing framework	[7,8]	Analysis-oriented	Customer-centric two-staged framework: 1: Pre-purchase phase (Communication & Transparency) 2: Post-purchase phase (Dynamism & Service)	Literature overview (esp. [14,26,28,36]) Series of in-depth interview
F2	Customer-value based pricing framework	[22,23]	Action-oriented	Depicts the interconnection between three pillars: 1: Customer characteristics 2: Company objectives 3: Pricing objectives and strategy	Literature overview (esp. [19])
F3	Pricing process framework	[59]	Analysis-oriented	Three-stage pricing process structure: 1: Data collection 2: Strategy analysis 3: Strategy establishment	Literature overview (esp. [11,16])
F4	Competitive forces based framework	[38]	Analysis-oriented	Four-layer model: 1: Competitive forces 2: Factors impacting 3: Revenue models 4: Competitive advantage	Literature overview (esp. [41,42])
F5	Software products pricing typology	[30]	Analysis-oriented	Typology based on six pricing parameters: 1: Formation of price, 2: Structure of payment flow, 3: Assessment base, 4: Price discrimination, 5: Price bundling, 6: Dynamic pricing strategies	Literature overview (esp. [9]) Series of in-depth interviews
F6	Cloud solution pricing framework	[28]	Analysis-oriented	Typology based on seven pricing parameters: 1: Pricing scope 2: Structure of payment flow 3: Assessment base 4: Price discrimination 5: Price bundling 6: Dynamic pricing strategies	Literature overview (esp. [25,30]) Market survey
F7	Pricing strategy guideline framework	[2,55]	Action-oriented	Five layers of pricing within Corporate and Sales & Marketing Strategies: 1: Value Creation + Business Case, 2: Pricing Structure, 3: Price and Value Communication, 4: Price Policy + Sales Mechanism 5: Price Level	Literature overview (esp. [4,24,26,41,42])

#	Name	Ref.	Perspective	Framework structure	Scientific and practical origins
F8	Pricing canvas framework	[20]	Action-oriented	Six-segment pricing canvas: 1: Customer Segments, 2: Value Proposition, 3: Cost Structure, 4: Competitors and Market, 5: Pricing Strategy, 6: Price Model	Not specified
F9	Pricing strategies decision framework	[17]	Action-oriented	Six-step framework: 1: What is the Customer's Value of the Product? 2: Is the Customer Aware of this Value? 3: Can the Customer Base be Segmented? 4: Is the Customer's Demand Variable or Uncertain? 5: Establish a Price Floor 6: What are the value metrics that are most important to the customer?	Case-studies (own experience) Pacific Crest SaaS Company Survey, Totango Reports on SaaS Metrics
F10	PWC pricing management framework	[47]	Analysis-oriented	Four pricing management segments: 1: Pricing strategy, 2: Price formulation, 3: Transaction management, 4: Performance management	Case-studies (own experience)
F11	Mastering pricing framework	[43]	Action-oriented	Pricing pillars: 1: Pricing at the Seed Stage 2: Pricing at the Expansion Stage 3: Pricing at the Growth Stage	Large-scale survey and market research
F12	ACCION pricing framework	[3]	Action-oriented	Four-step framework: 1: Define your upper bound 2: Define your lower bound 3: Identify any reasons to charge less than max value 4: Structure your pricing model as a compromise between upper bound and lower bound	Not specified
F13	Product Focus pricing framework	[44, 46]	Action-oriented	Pricing pillars: 1: Pricing constraints 2: Pricing cycle 3: Pricing Evolution	Case-studies (own experience)

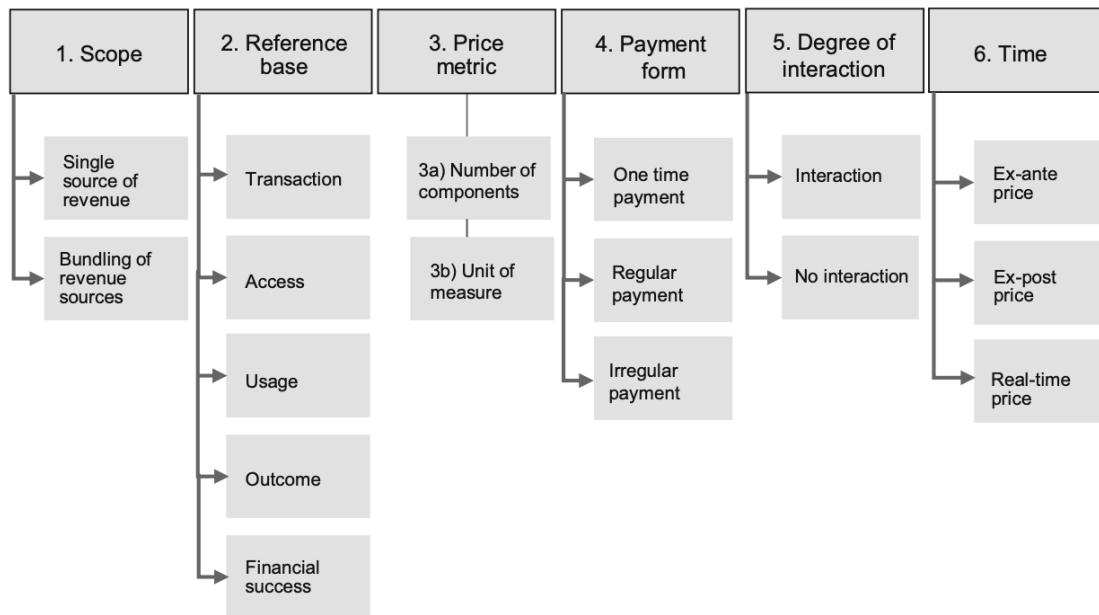
Appendix 8: Parameters of pricing models for software products (Lehmann & Buxmann, 2009)



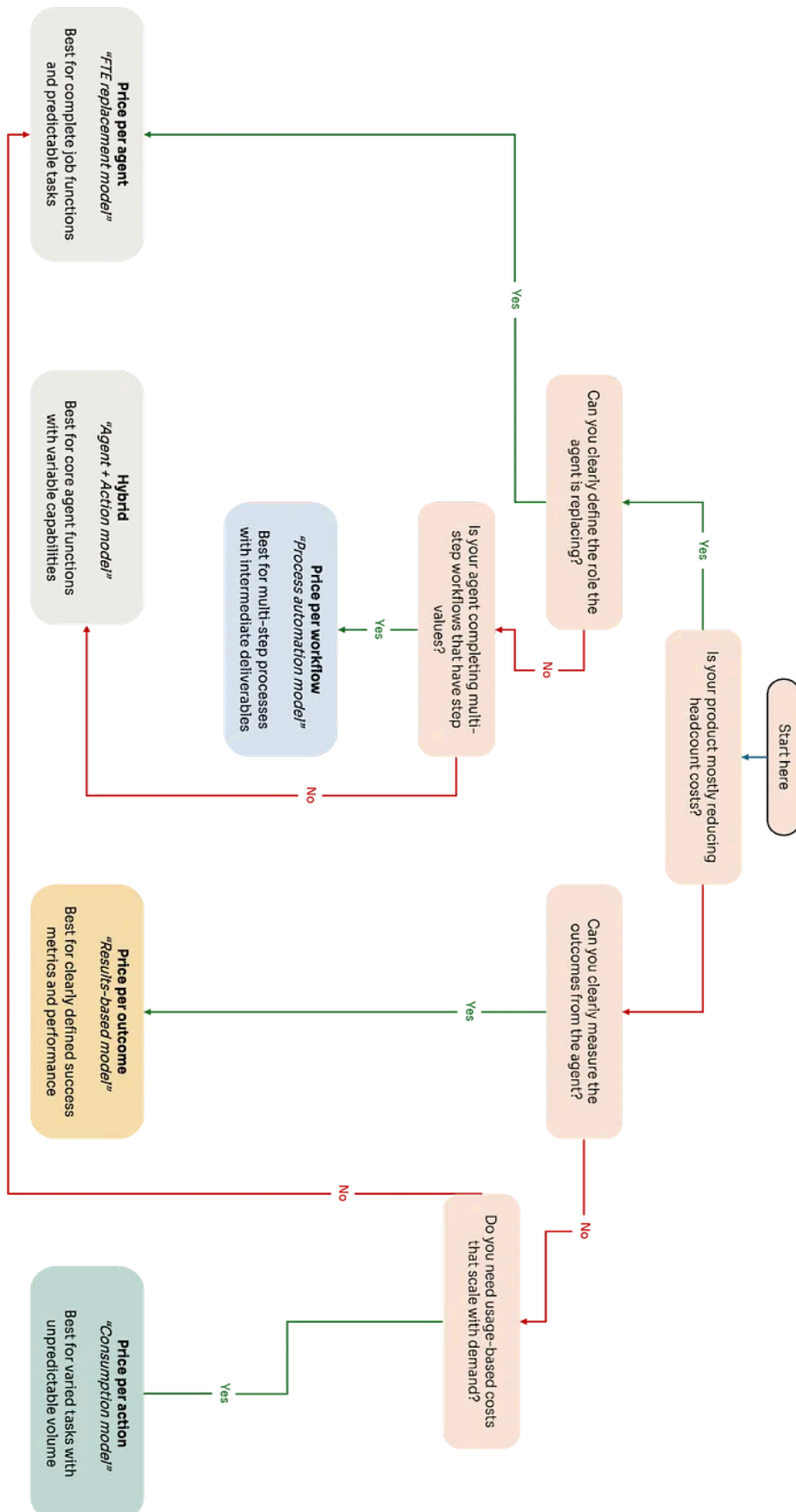
Appendix 9: The Pricing Strategy Guideline Framework for SaaS vendors (Spruit & Abdat, 2012)



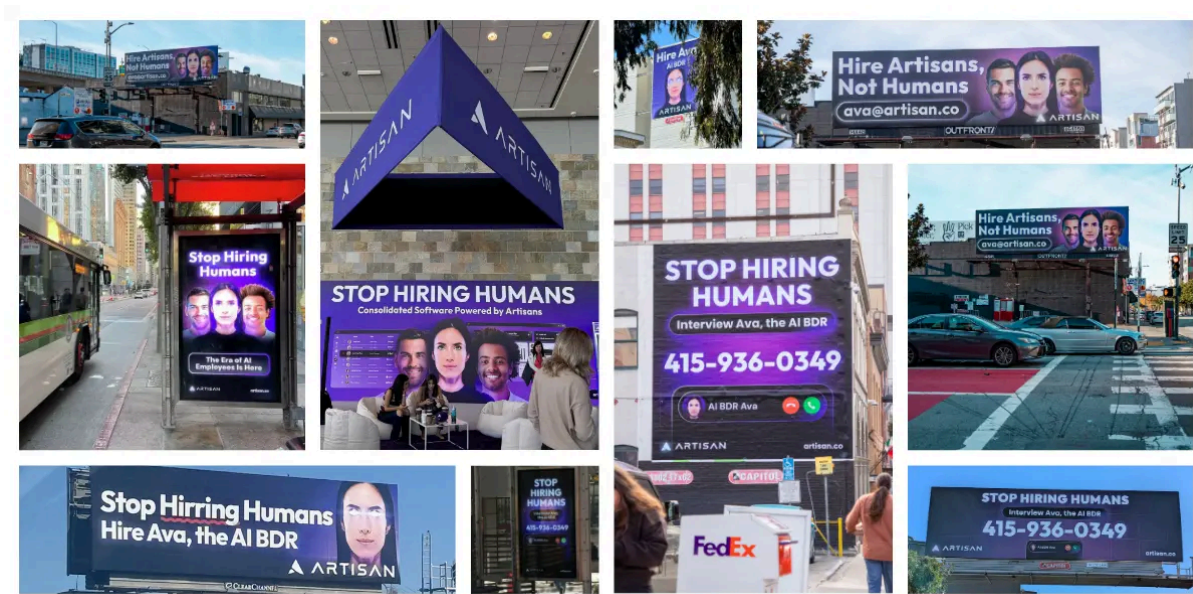
Appendix 10: The six pillars of a price model (Frohmann, 2018)



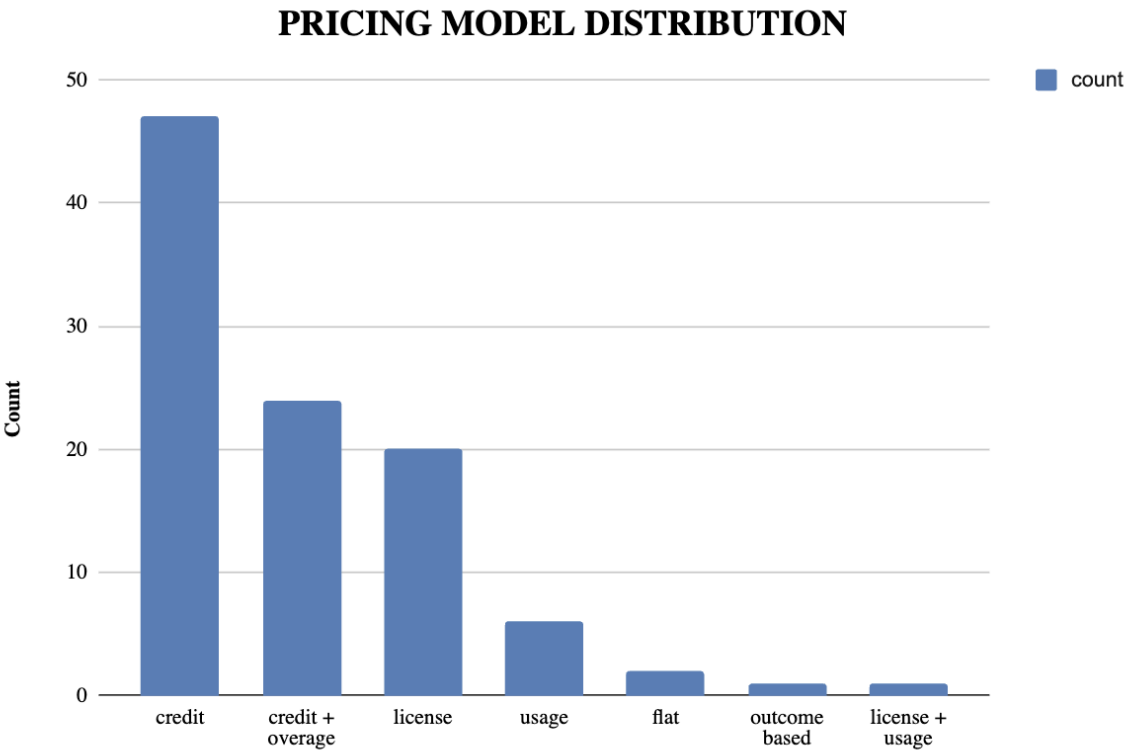
Appendix 11: Decision Framework Pricing Model, Medina, 2025



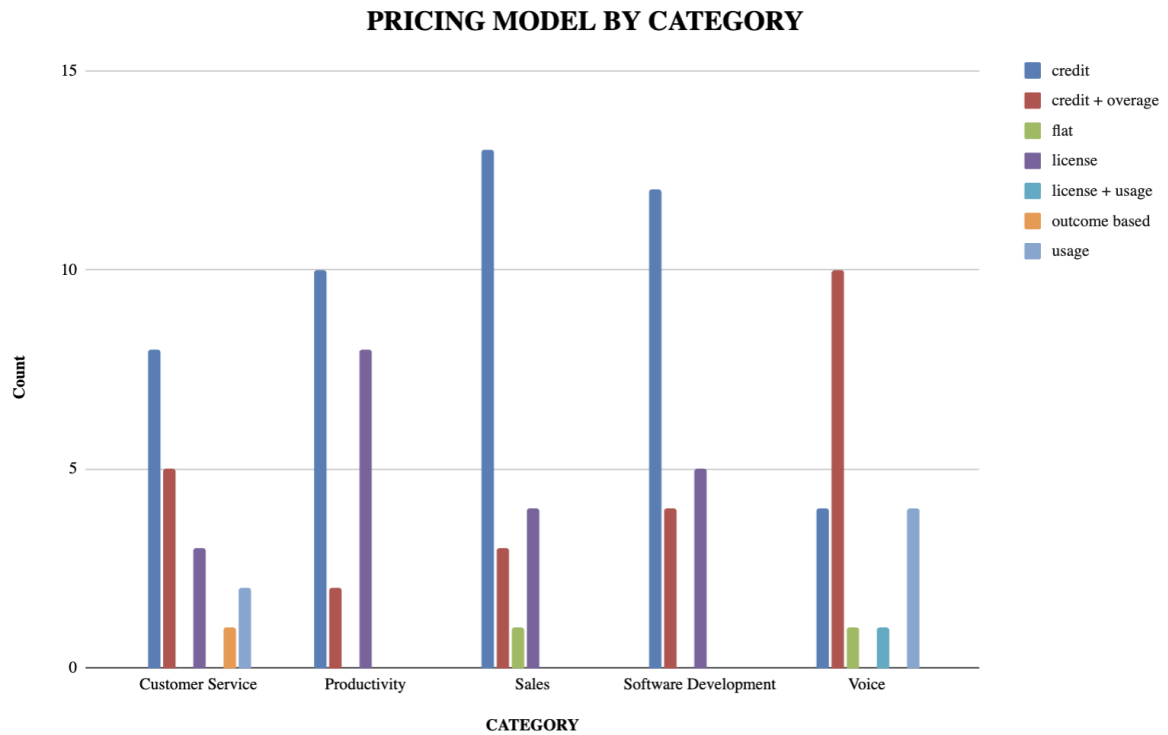
Appendix 12: Artisan Add (Carmichael-Jack, J., 2024)



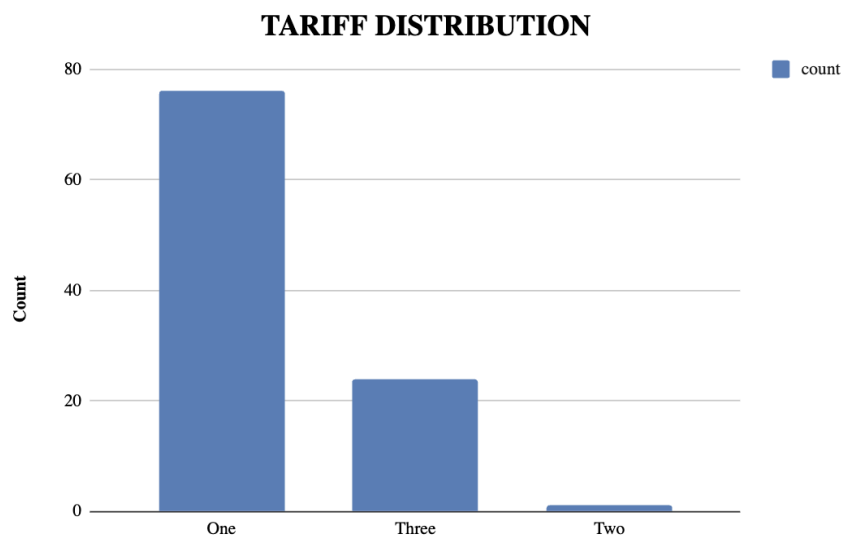
Appendix 13: Pricing Model Distribution (Authors own contribution)



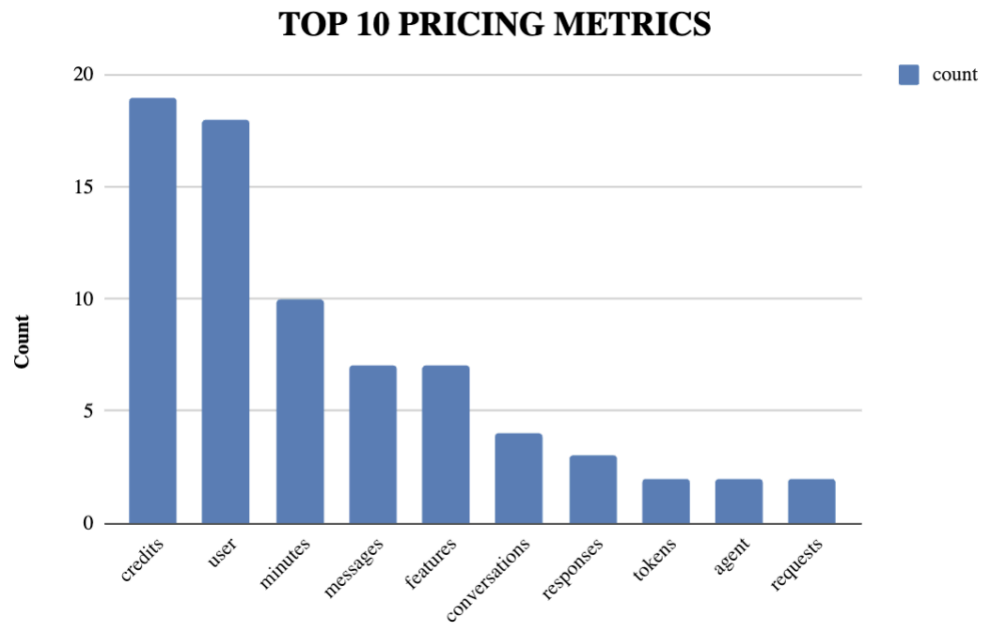
Appendix 14: Pricing Modality Distribution per Category (Authors own contribution)



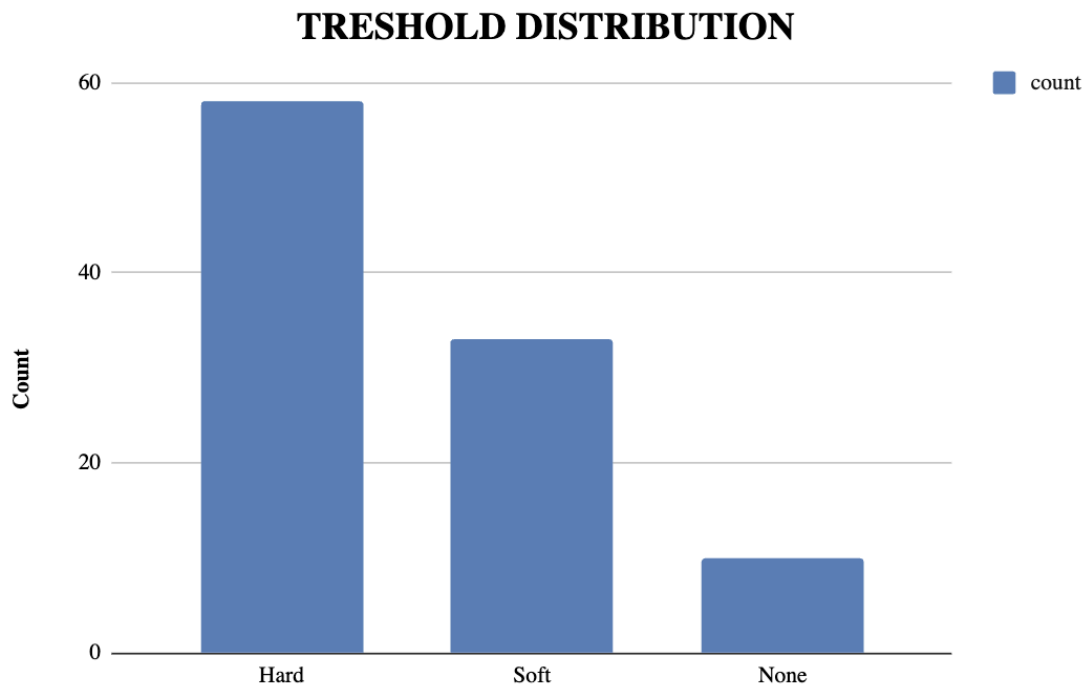
Appendix 15: Tariff Distribution (Authors own contribution)



Appendix 16: Top 10 Pricing Metric Distribution (Authors own contribution)



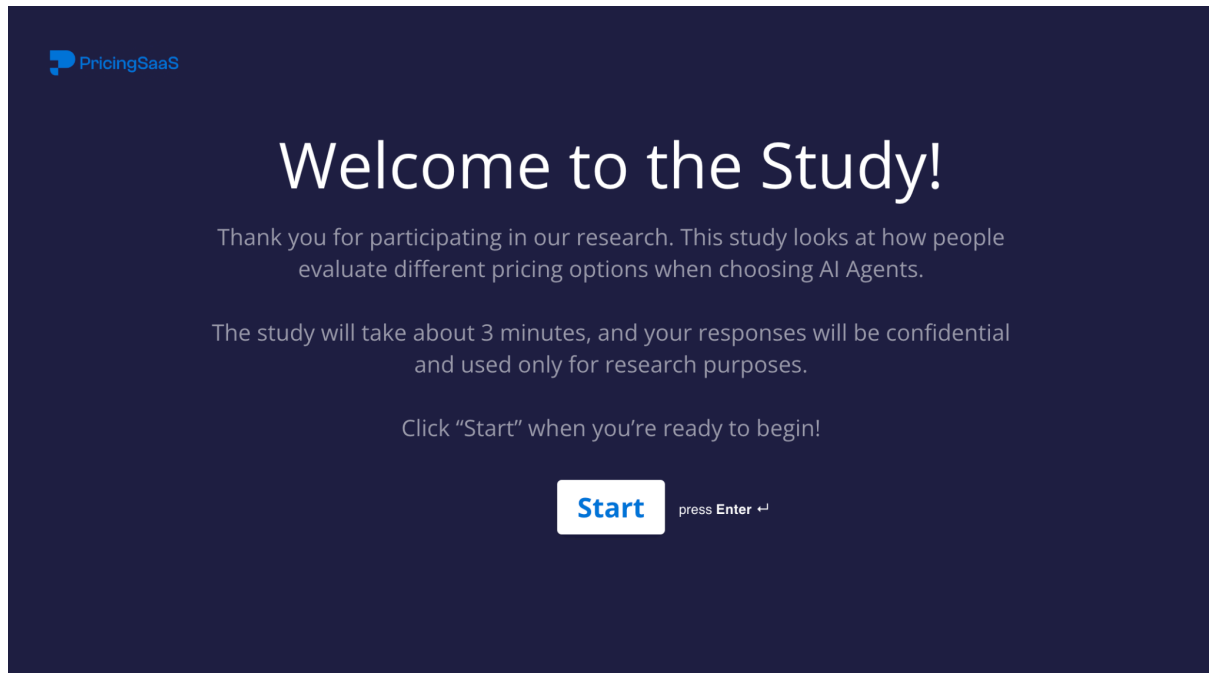
Appendix 17: Threshold Distribution (Authors own contribution)




Appendix 18: Free Plan Availability (Authors own contribution)



Appendix 19: Typeform Survey (Authors own contribution)



 PricingSaaS

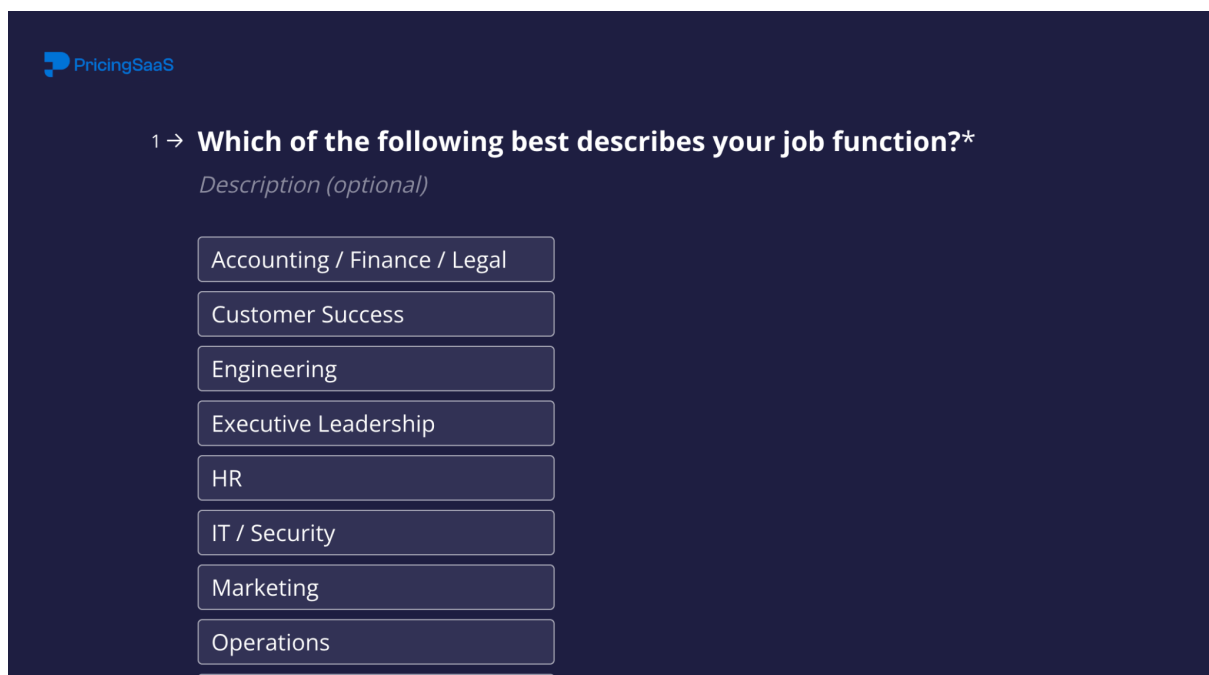
Welcome to the Study!


Thank you for participating in our research. This study looks at how people evaluate different pricing options when choosing AI Agents.

The study will take about 3 minutes, and your responses will be confidential and used only for research purposes.

Click "Start" when you're ready to begin!

Start press Enter ↵



 PricingSaaS

1 → **Which of the following best describes your job function?***

Description (optional)

- Accounting / Finance / Legal
- Customer Success
- Engineering
- Executive Leadership
- HR
- IT / Security
- Marketing
- Operations

2 → **Which of the following best describes your role?***

Description (optional)

Individual Contributor

Manager

Director

VP

C-Level

Other

[Add choice](#)

3 → **Which of the following best describes your company's category?***

Description (optional)

AI & Machine Learning

Business Operations

Cloud & DevOps

Collaboration & Productivity

CyberSecurity & Identity Management

Data & Analysis

Finance & Accounting

Marketing & Content

4 → **What best describes your company size?***

Description (optional)

Fewer than 25 employees

25-100

101-250

251-1,000

More than 1,000

[Add choice](#)

5 → **How would you describe your experience buying AI Agents?**

*

Description (optional)

I have never purchased an AI Agent

I am actively exploring AI Agents

I have purchased an AI Agent

[Add choice](#)

“ Meet Your New Support Teammate (An AI Agent)

Imagine having an AI Agent that can handle all of your Customer Support conversations for you — instantly, 24/7, with consistent quality.

Below, we'll show you a few different pricing models for this same AI support agent. Please review the options, and answer the follow-up questions as accurately as possible.

Description (optional)

Continue

press Enter ↵

6 → Please select the pricing model you would find MOST and LEAST preferable when purchasing an AI Agent? *

	Option A	Option B	Option C
Most preferred	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Least preferred	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

OK

Option A Pay-per-use Pay only for what you use. Cost scales with the number of conversations. Choose	Option B Credit - based Fixed monthly fee for a limited number of conversations. If you exceed the limit, you purchase more credits. Choose	Option C Flat Fee Pay a premium fee for unlimited use of the AI Agent. Choose
Option D Outcome - based Pay a premium per successful resolution. Choose	Option E License - based Fixed monthly fee for a limited number of conversations. If you exceed the limit, you pay an overage fee. Choose	

Powered by Typeform

7 → Which of the following do you find MOST and LEAST important when considering pricing models?*

	Cost Predictability	Cost Transparency	Fairness	Simplicity
MOST Important	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
LEAST important	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

OK

Powered by Typeform

8 → **Would you be more likely to sign up for an AI agent if it included a free plan with limited features?***

Yes

No

OK



Powered by [Typeform](#)

9 → **Can you shortly explain why chose that pricing model?**

Type your answer here...

Shift ⌘ + Enter ↵ to make a line break

Submit

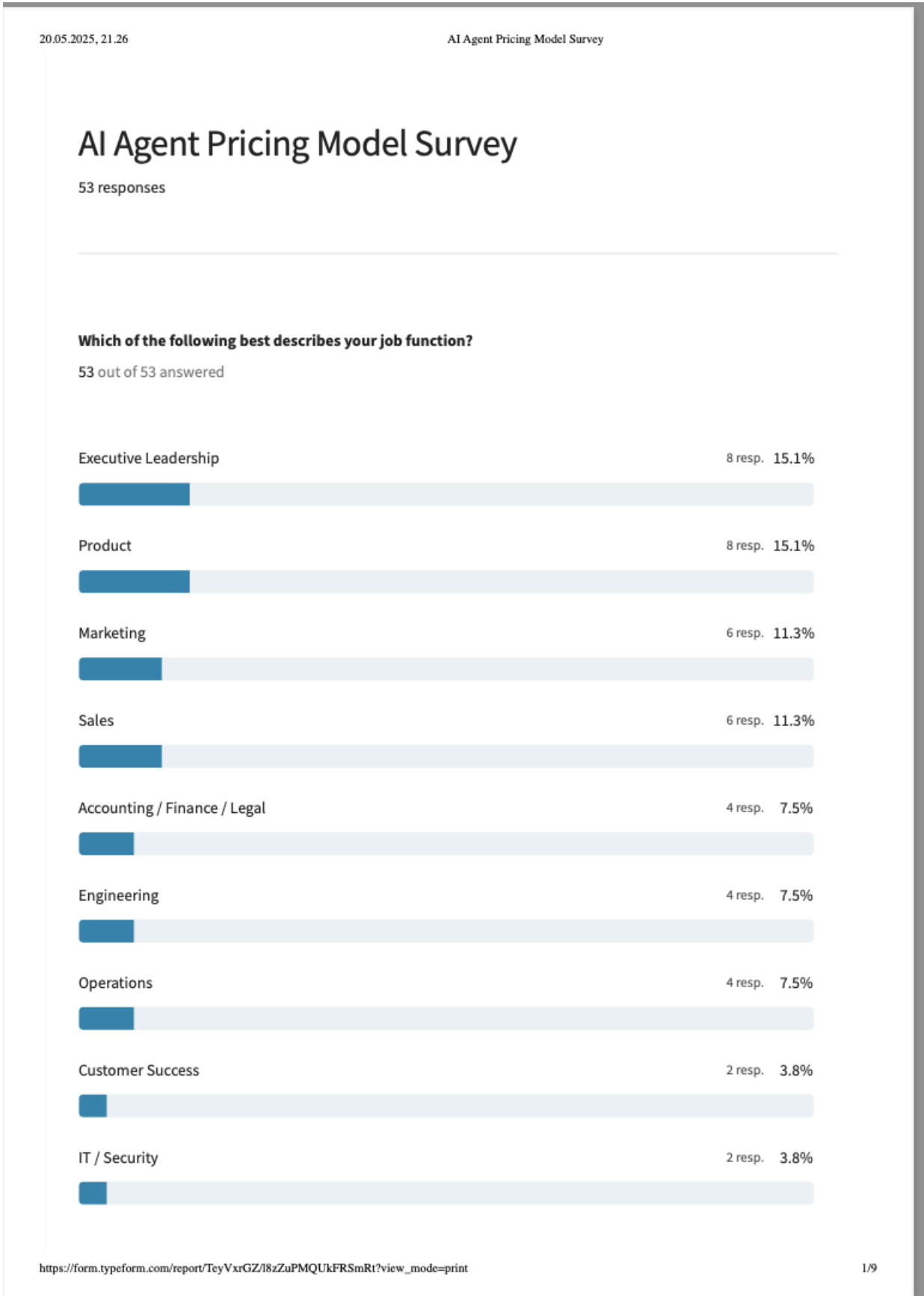
press Cmd ⌘ + Enter ↵

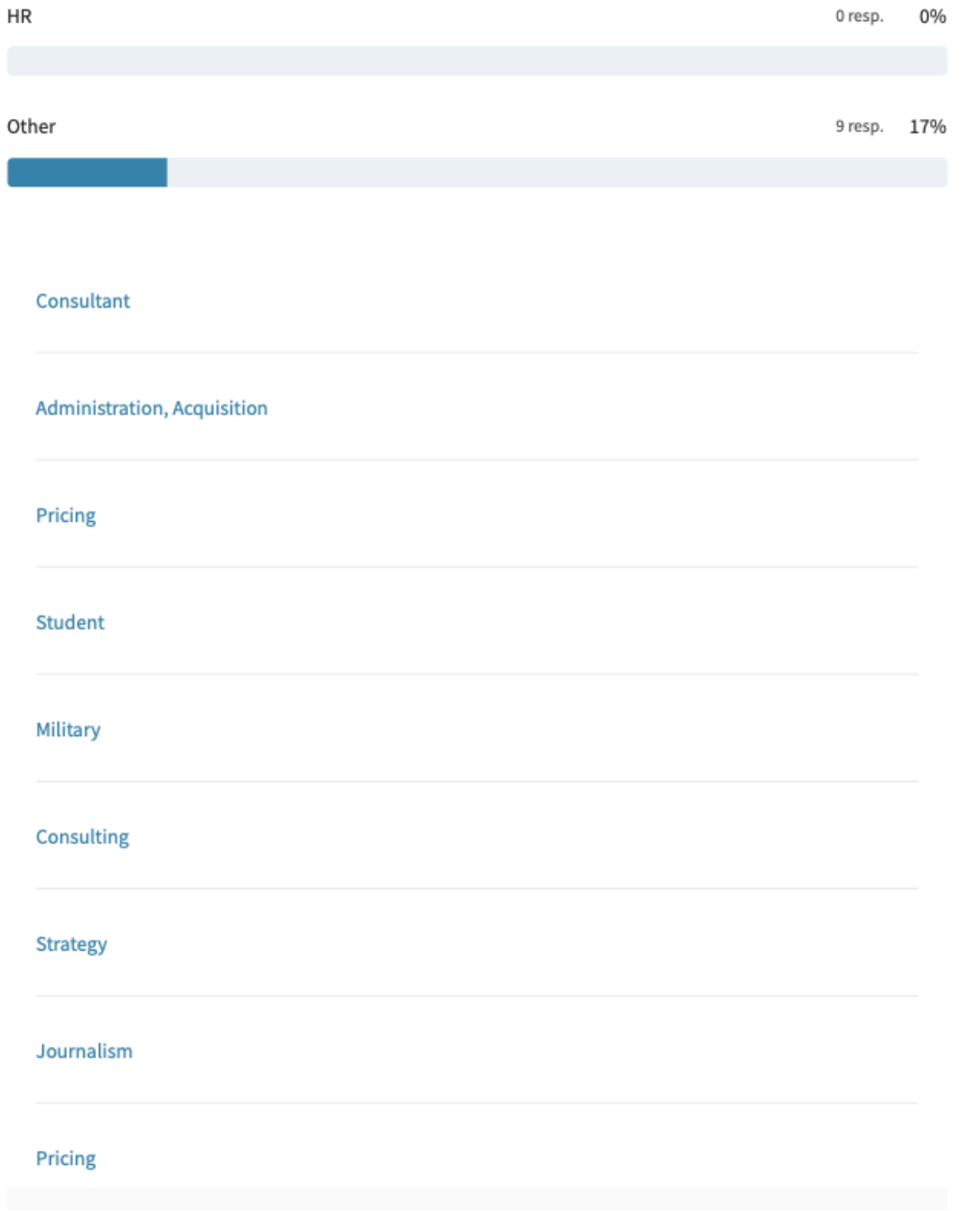


Powered by [Typeform](#)

Would you be more likely to sign up for an AI Agent if it included a free plan with limited features?

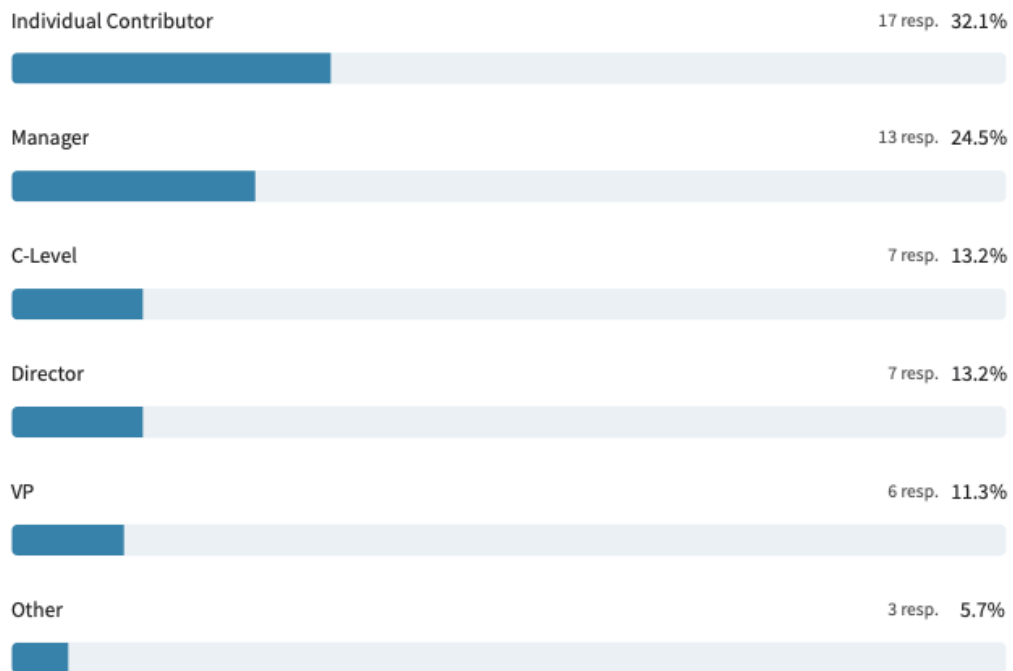
Appendix 20: Survey Results (Authors own contribution)





Which of the following best describes your role?

53 out of 53 answered



Student

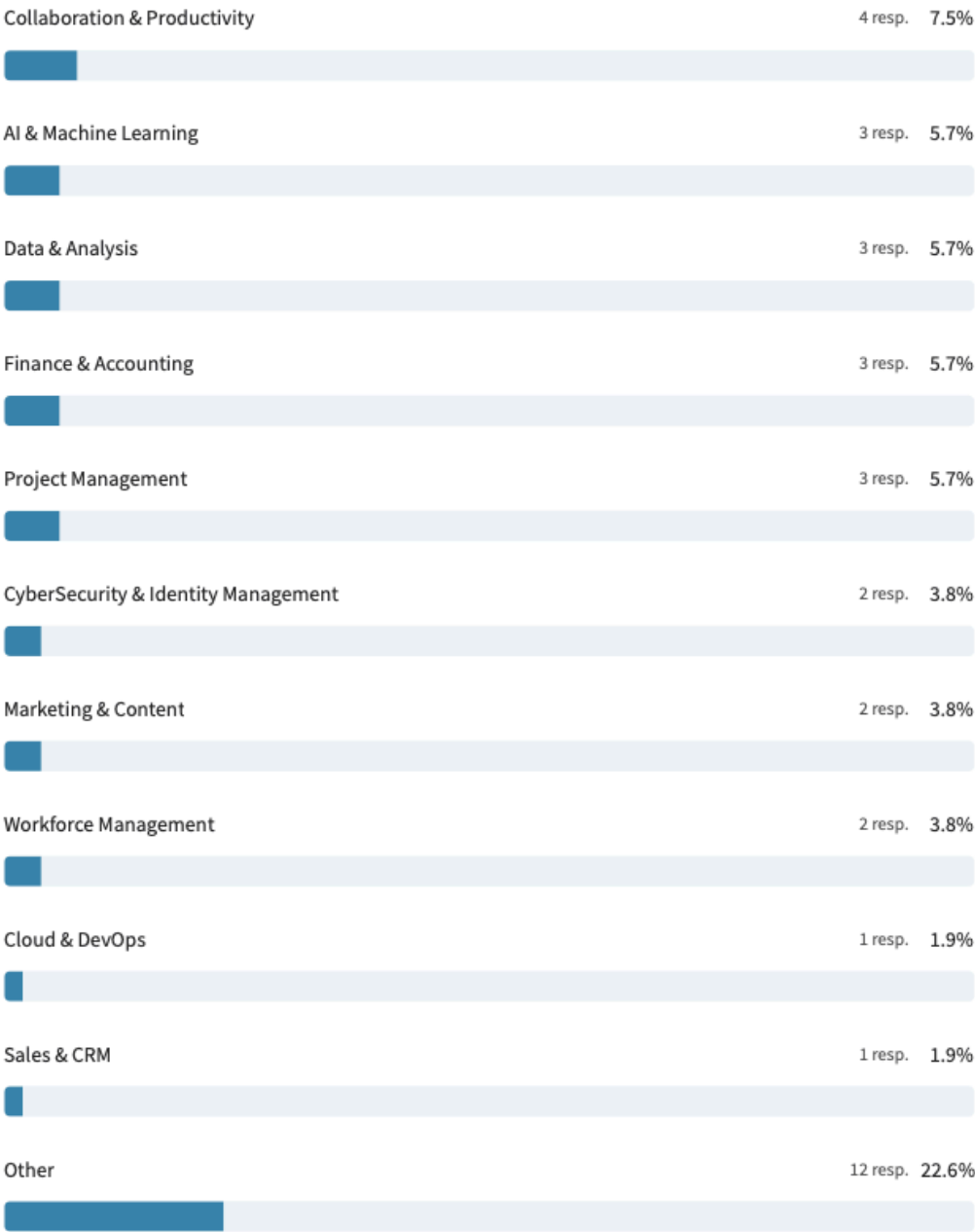
Undergraduate

Employee

Which of the following best describes your company's category?

53 out of 53 answered





Transport

Product & Pricing Strategy

Machinery construction

Not employed

Security

All the above

IT Service Provider

Distribución

Publishing

Shipping

Education

Telematics

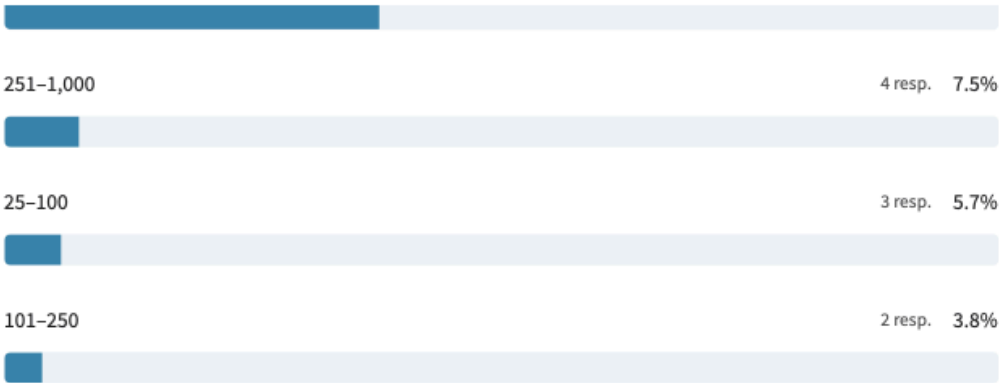
What best describes your company size?

53 out of 53 answered

More than 1,000 24 resp. 45.3%

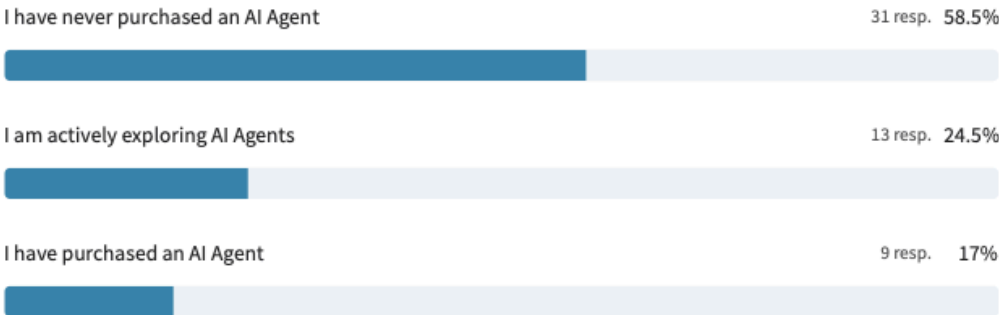


Fewer than 25 employees 20 resp. 37.7%



How would you describe your experience buying AI Agents?

53 out of 53 answered



Would you be more likely to sign up for an AI agent if it included a free plan with limited features?

53 out of 53 answered



No

7 resp. 13.2%

**Can you shortly explain why chose that pricing model?**

35 out of 53 answered

Fixed costs - unlimited access

Pricing mode needs to adapt to entities business conditions and day-to-day business activities. Pricing model needs to be simple and transparent.

Regarding evolution of the model observing close up the results

Pricing model is crucial for implementing AI in the company. My desire for controlling expenditure and seeing results would lead towards a pay per use solution in the upstart, and based on the first months of usage/results and evaluation lead to the best fit pricing model from there.

Free is better than nothing and you can test if working and can expect something more for pay model

Because I would have the chance to test qithout a big investment or comitment

Because it aligns both with the value proposition of the product and the needs of the target customer segment. This model offers a clear and simple benefit: customers pay a higher, fixed fee in exchange for unrestricted use of the product.

provides access with controls to not let costs spiral (which PAYG or outcome does not)

In this narrow use case (which is not typical of AI agents) outcomes are clearly defined and predictable.

Because it's less work and makes the most sense for me. I don't want to book again while I'm working and maybe my flat is gone

Fixed Price to explore the tool

My CS budget is fixed, I need the replacement to be fixed as well

Knowing how much the service will cost is very important to me

Because you can see how it works before committing to it

Barrier to entry zero. I try it out. If I like it, then I'll consider buying it. Not all AIs are created equal.

Mostly the chat bots are useless and dont take losd off support team, so it makes sense to pay for successful outcomes when user is happy and saved the support cost

Know your cost

User friendly

I wouldn't need to commit from the start to a pricing model, without being able to test the platform and its success beforehand.

My usage varies, so I don't want to pay more than my usage

You don't get attached to anything.

Risk Assessment. Will be able to test the Environment before committing

I like the predictability of a flat fee.

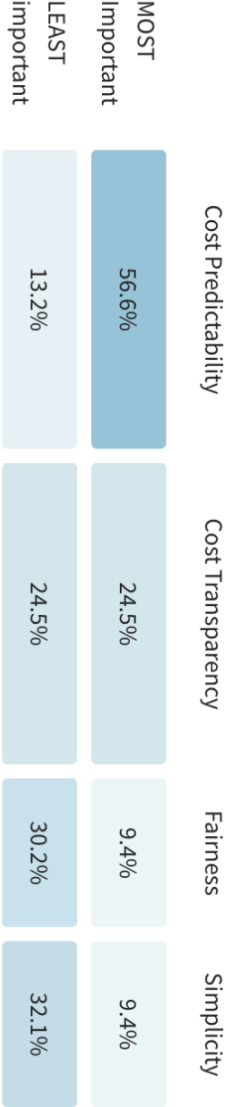
Important to try it out and evaluate before committing

Predictability

Powered by Typeform

Which of the following do you find MOST and LEAST important when considering pricing models?

53 out of 53 answered



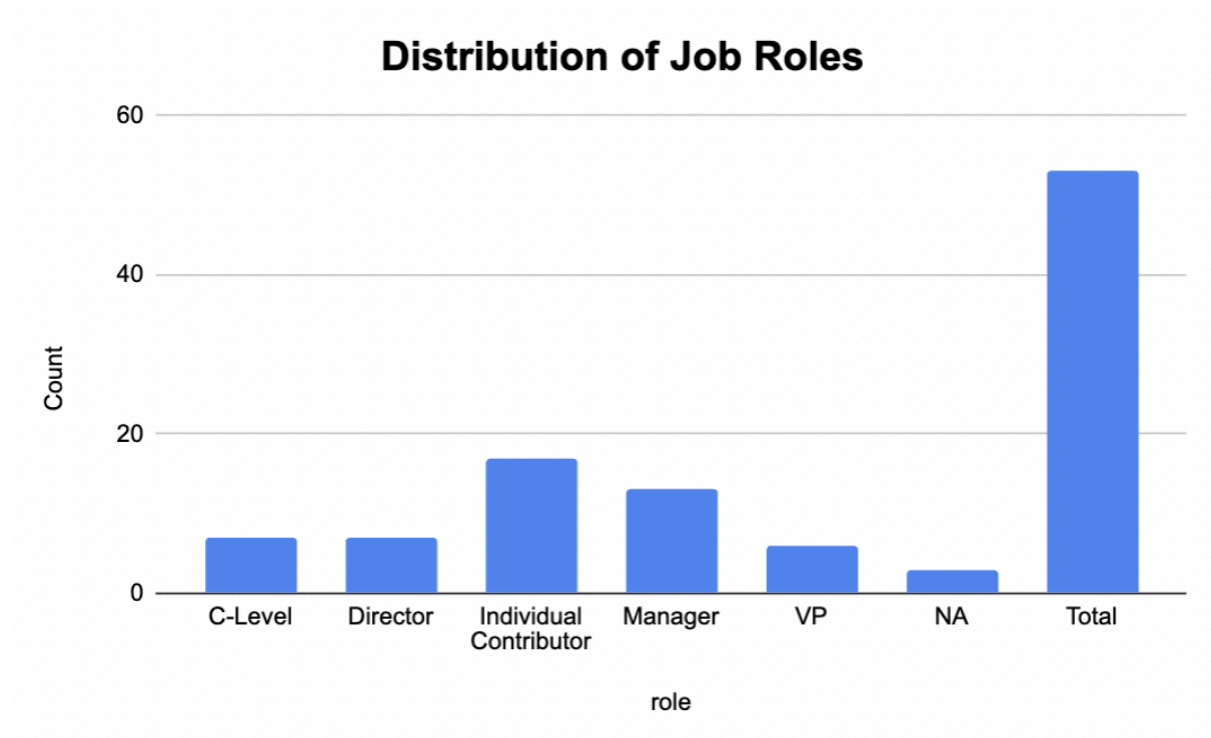


Please select the pricing model you would find MOST and LEAST preferable when purchasing an AI Agent?

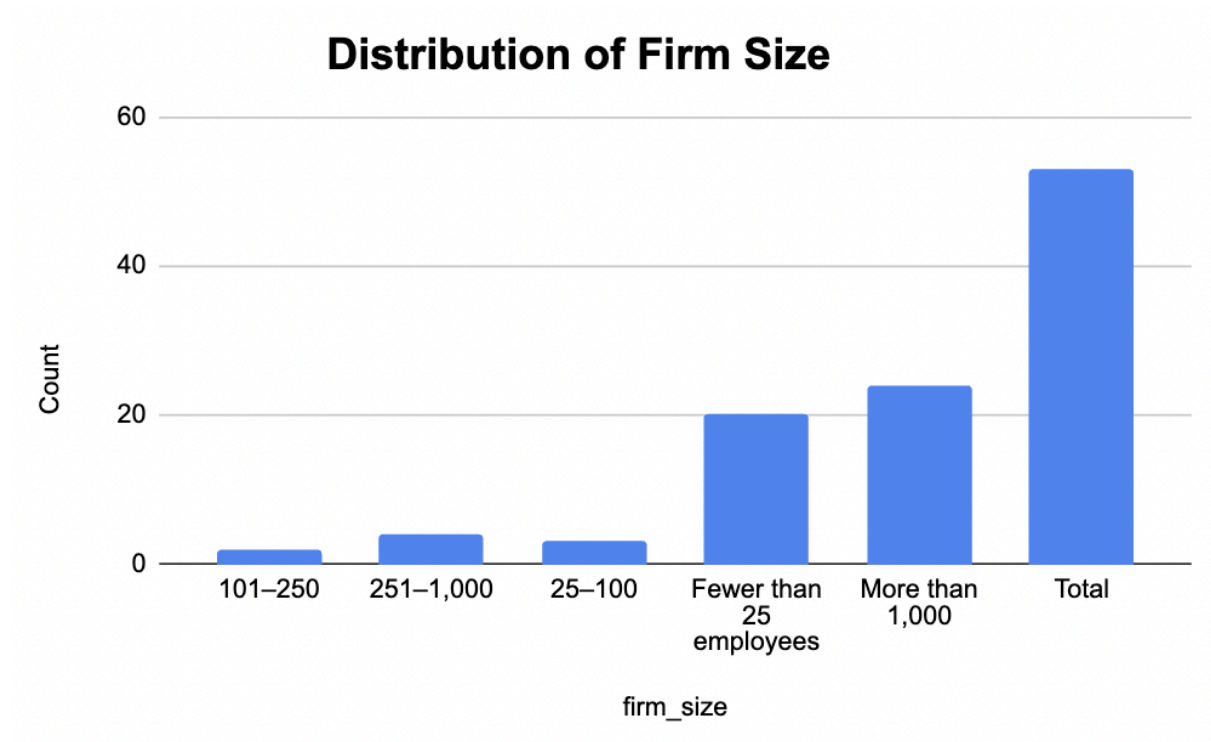
53 out of 53 answered

	Option A	Option B	Option C	Option D	Option E
Most preferred	5.7%	22.6%	43.4%	24.5%	3.8%
Least preferred	37.7%	9.4%	1.7%	9.4%	26.4%

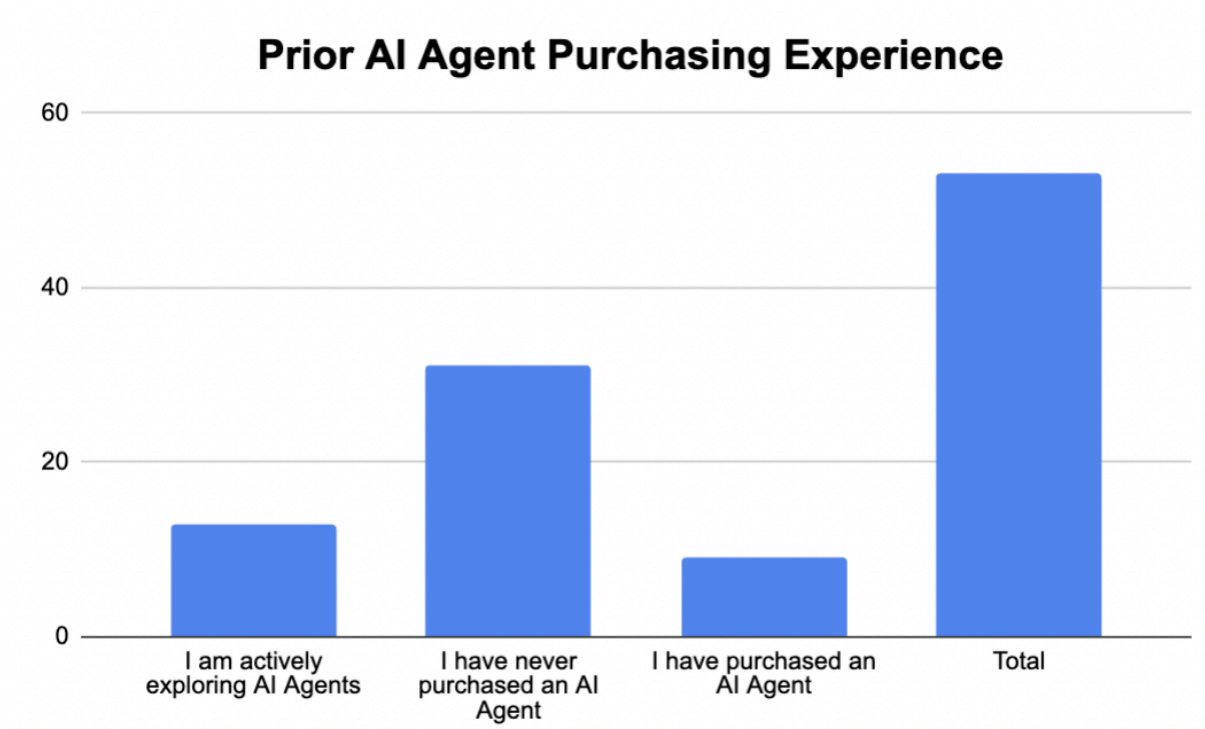
Appendix 21: Distribution of Job Roles (Author's own contribution)



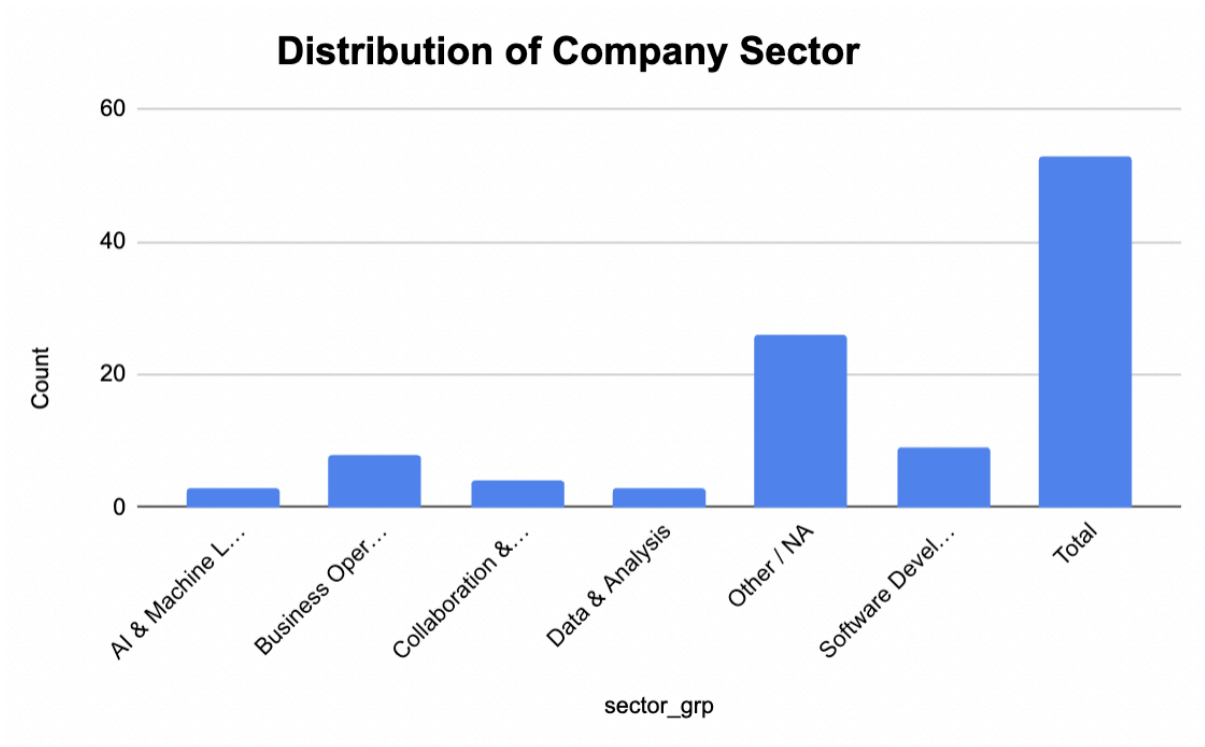
Appendix 22: Distribution of Firm Size (Author's own contribution)



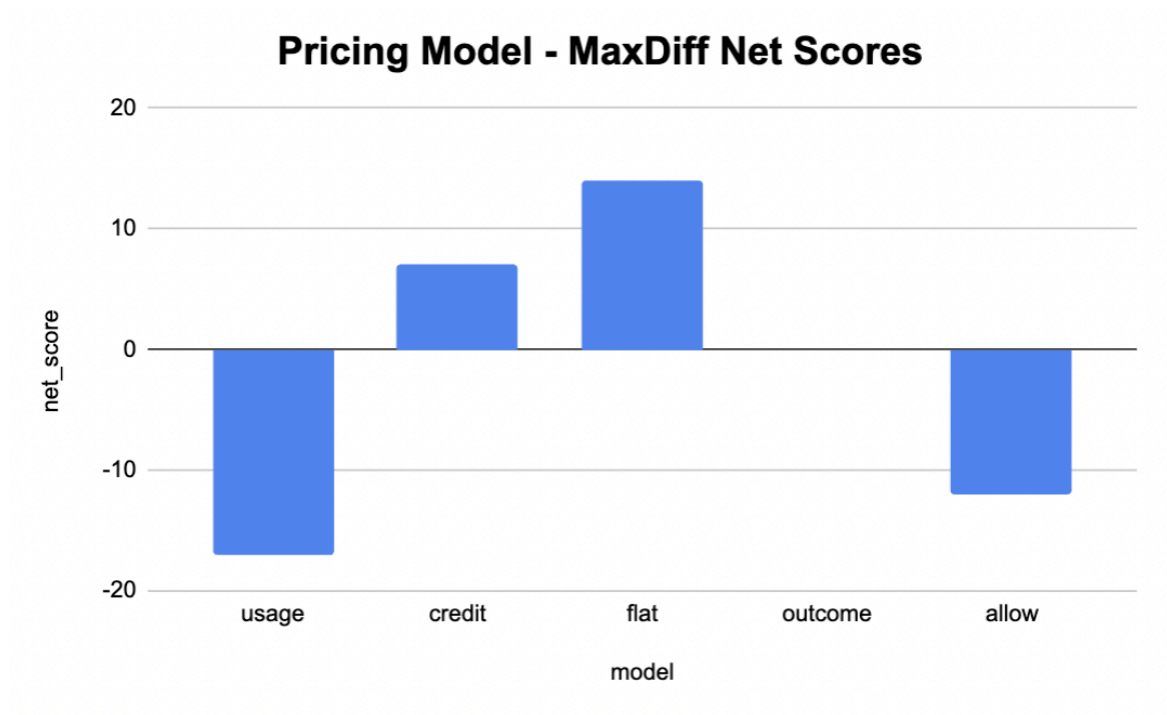
Appendix 23: Distribution Prior AI Agent Purchasing Experience (Author's own contribution)



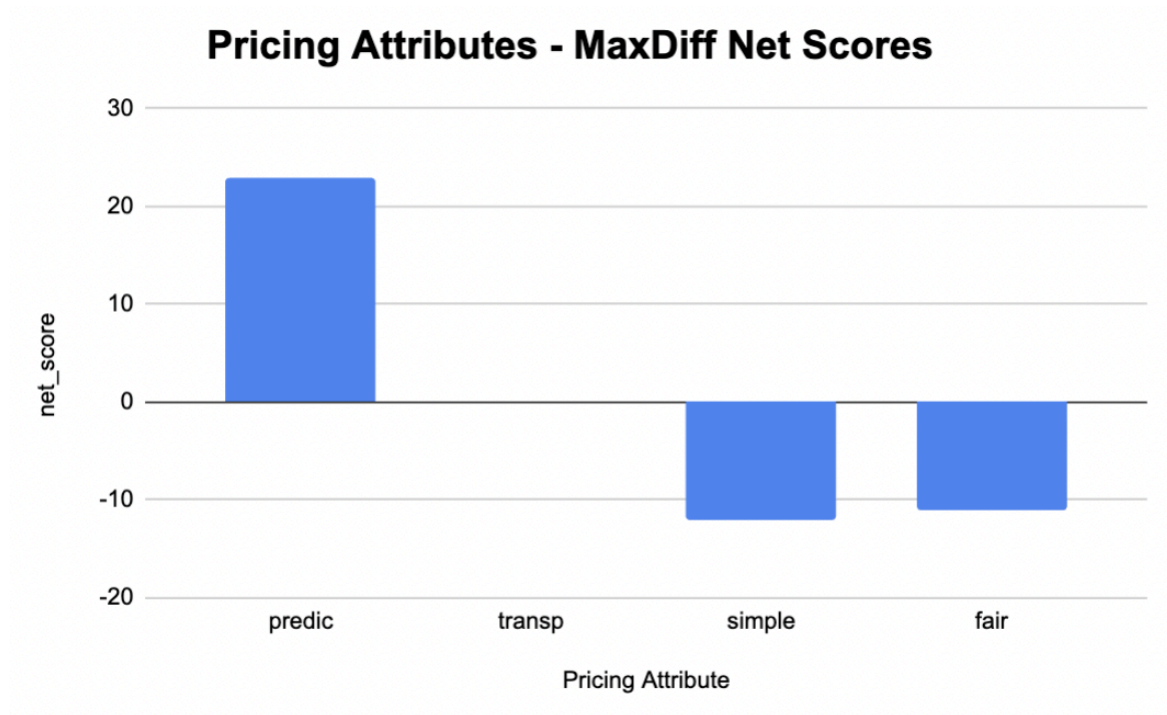
Appendix 24: Distribution of Company Sectors (Author's own contribution)



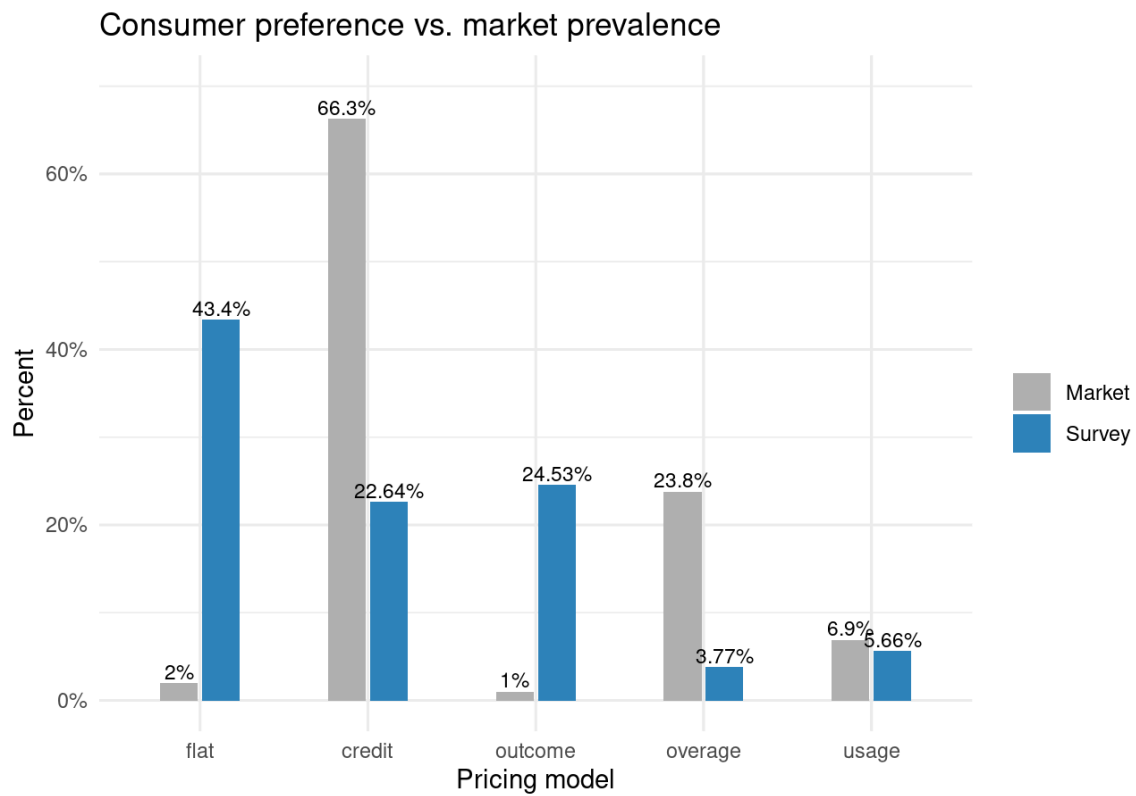
Appendix 25: Pricing Model - MaxDiff Net Scores (Author's own contribution)



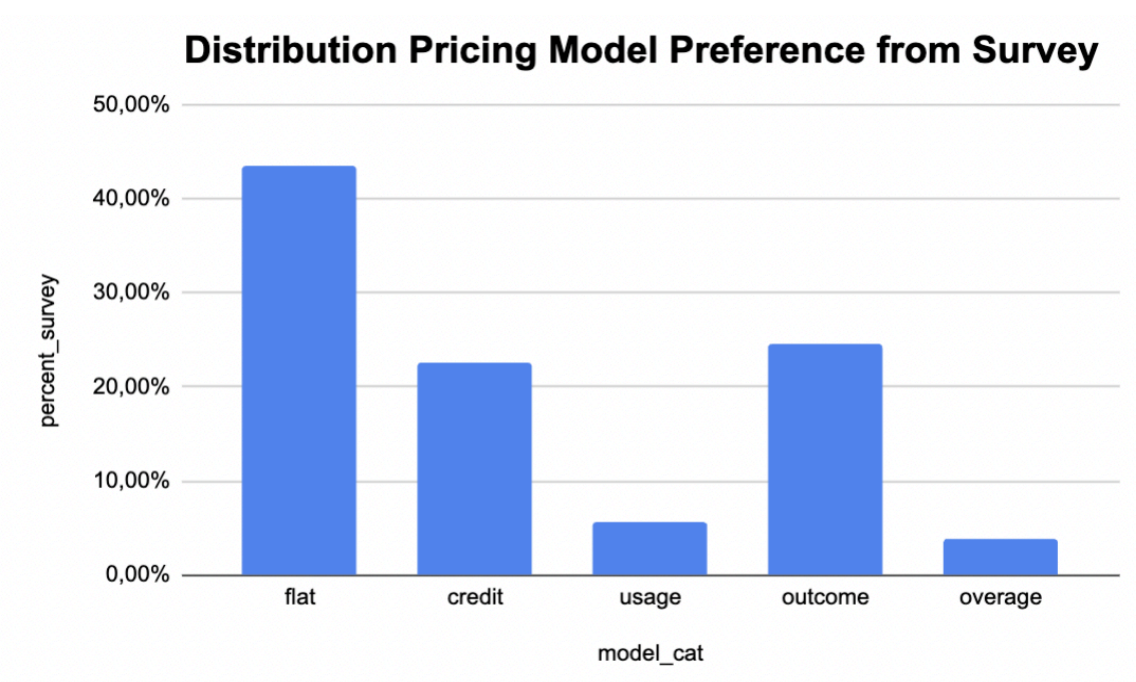
Appendix 26: Pricing Attributes - MaxDiff Net Scores (Author's own contribution)



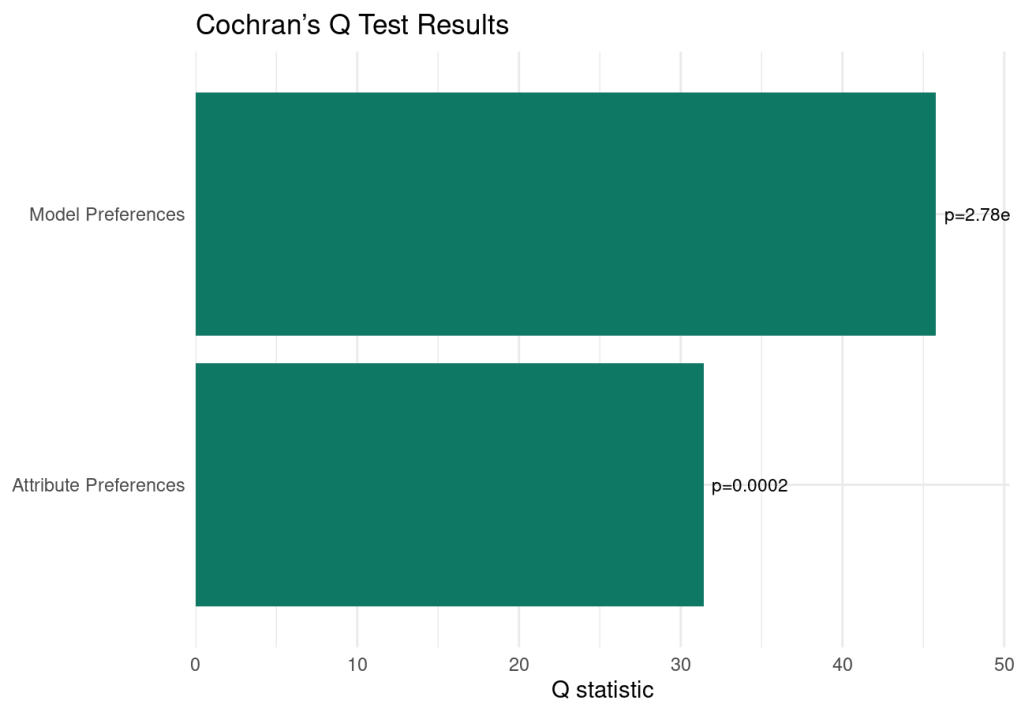
Appendix 27: Consumer preference vs market prevalence (Created in R)



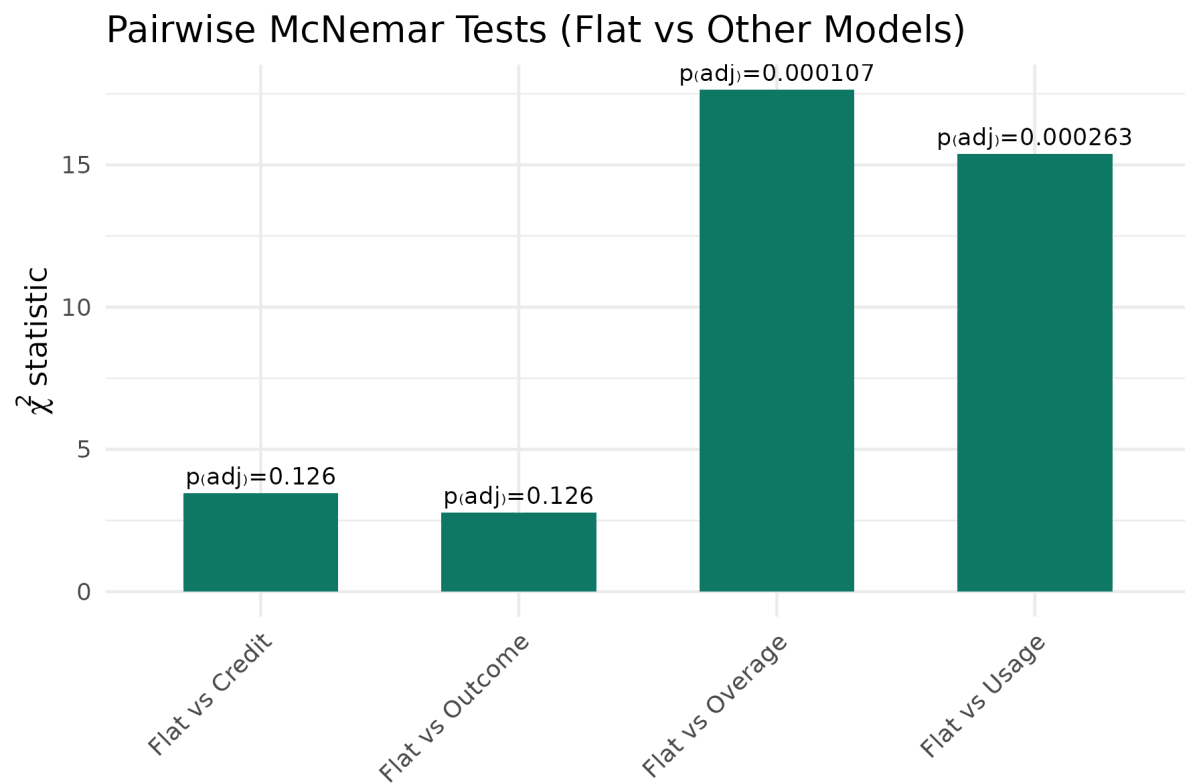
Appendix 28: Distribution Pricing Model Preference From Survey (Author's own contribution)



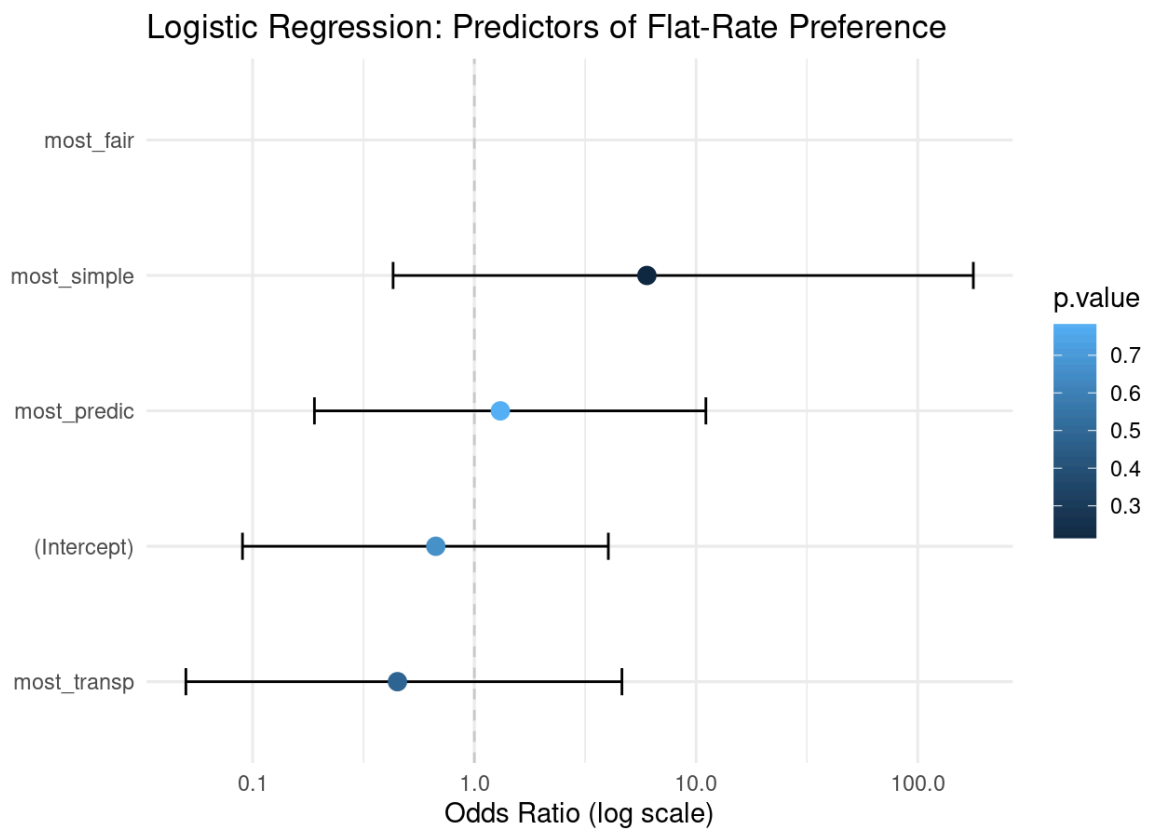
Appendix 29: Cochran's Q Test (Created in R)



Appendix 30: Pairwise McNemar Test (Flat vs Other Models) (Created in R)

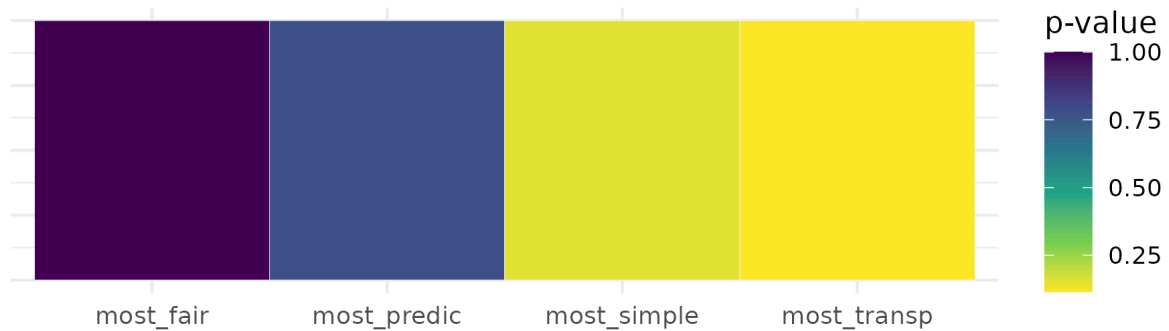


Appendix 31: Binary logistic Regression (Flat vs Other) (Created in R)



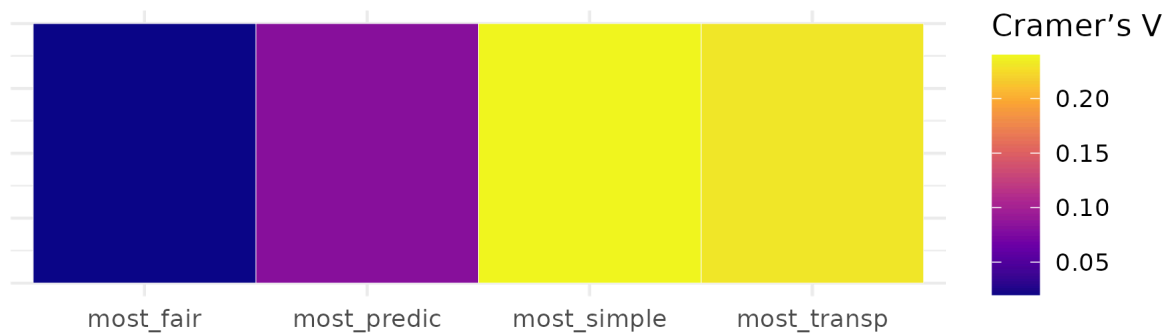
Appendix 32: Fisher's Test - Heat Map (Created in R)

Fisher's Exact Test p-values



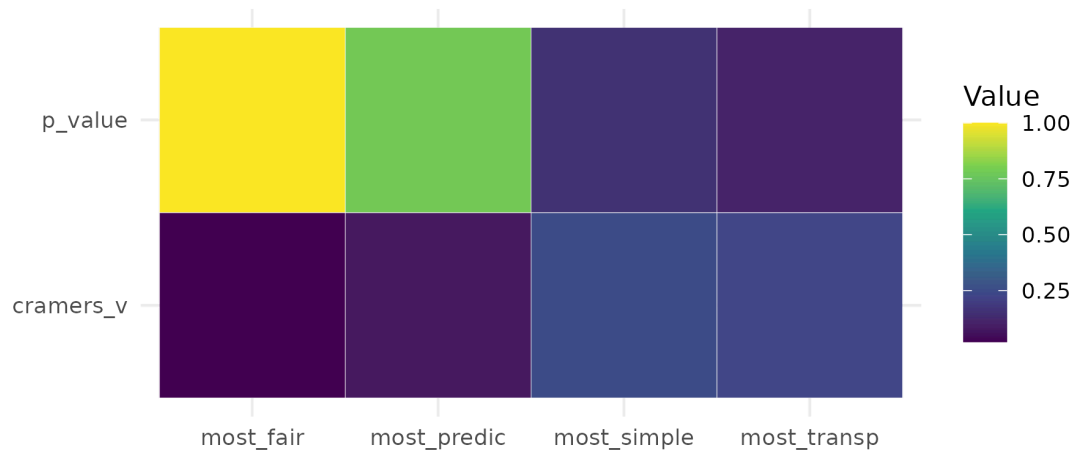
Appendix 33: Cramer's V - Heat Map (Created in R)

Cramer's V Effect Sizes



Appendix 34: Fisher's Test & Cramers V - Heat Map (Created in R)

Fisher p-values and Cramer's V by Attribute



Appendix 35: Open-ended explanations for pricing model choice

Fixed costs - unlimited access
Pricing mode needs to adapt to entities business conditions and day-to-day business activities.
Pricing model needs to be simple and transparent.
Regarding evolution of the model observing close up the results
Free is better than nothing and you can test if working and can expect something more for pay model
Pricing model is crucial for implementing AI in the company. My desire for controlling expenditure and seeing results would lead towards a pay per use solution in the upstart, and based on the first months of usage/results and evaluation lead to the best fit pricing model from there.
Because I would have the chance to test qithout a big investment or comitment
Because it aligns both with the value proposition of the product and the needs of the target customer segment. This model offers a clear and simple benefit: customers pay a higher, fixed fee in exchange for unrestricted use of the product.
provides access with controls to not let costs spiral (which PAYG or outcome does not)
In this narrow use case (which is not typical of AI agents) outcomes are clearly defined and predictable.
Because it's less work and makes the most sense for me. I don't want to book again while I'm working and maybe my flat is gone
Fixed Price to explore the tool
My CS budget is fixed, I need the replacement to be fixed as well
Knowing how much the service will cost is very important to me
Because you can see how it works before committing to it
Barrier to entry zero. I try it out. If I like it, then I'll consider buying it. Not all AIs are created equal.
Mostly the chat bots are useless and dont take losd off support team, so it makes sense to pay for successful outcomes when user is happy and saved the support cost
Know your cost
User friendly
I wouldn't need to commit from the start to a pricing model, without being able to test the platform and its success beforehand.
My usage varies, so I don't want to pay more than my usage
You don't get attached to anything.
Risk Assessment. Will be able to test the Environment before committing
I like the predictability of a flat fee.
Important to try it out and evaluate before committing
Predictability
I'm skeptical of if the AI can do what it promises. I would therefore prefer to not pay for unsuccessful

conversations
I don't want a surprise bill and need to be able to budget. However, I would prefer a pack of successful resolutions included in the plan and the ability to purchase more packs when I exceed rather than an abstracted credit model.
Budgeting is key for a company of my size. Having unexpected invoices due to heavy usage is not an option
Try and buy is effective in AI
It's the most simple model and it needs a minimum attention to be controlled
Flexibility
Because I use three different AI products
Most experience with
I like that I only pay for successful outcomes
I like it

Appendix 36: Philosophy of Science

Philosophy of Science

This chapter explains the philosophical foundations that guide every subsequent design decision in this thesis. It positions the study within a post-positivist paradigm and shows how that worldview informs the deductive quantitative methodology and the criteria used to assess rigour and ethics.

Research Paradigm: Post-Positivism

Post-positivism is a modern extension of classical positivism that acknowledges an objective reality while recognising that it can only ever be approached imperfectly through probability-based inquiry. The paradigm is characterised by four core features (Creswell, 2014):

1. **Critical realism:** reality exists independently of the researcher, but observations are theory-laden and fallible.
2. **Probabilistic knowledge:** claims are never proven, only corroborated or falsified with varying degrees of confidence.
3. **Deductive logic:** hypotheses derived from existing theory are subjected to empirical tests.
4. **Rigorous control and statistical inference:** bias is mitigated through design features such as randomisation, standardised instruments and inferential statistics.

These attributes are applied to this study: to test theory-driven hypotheses about which pricing models consumers prefer for AI agents and whether a flat-rate bias applies in that new context (see chapter 1.6). A post-positivist stance legitimises the use of an experiment, a structured market audit and statistical modelling to make cautious causal inferences.

Ontological Assumptions

The study adopts critical realist ontology: there is a single external reality in which consumer preferences exist regardless of whether they are observed. However, observations are always partial and mediated by measurement error. The objective of research is therefore to approximate, rather than capture exhaustively, those real preferences.

Epistemological Assumptions

Epistemologically, knowledge is regarded as conjectural and falsifiable. Reliable insights arise when hypotheses are subjected to systematic empirical scrutiny using probability theory. In this thesis, preference distributions across five pricing models are compared with Cochran's Q and follow-up McNemar tests, while binary logistic regression explores predictors of flat-rate choice. Statistical significance ($\alpha = .05$) is treated as conditional support, not proof, of a claim.

Axiology and Researcher Reflexivity

Post-positivism aspires to objectivity, yet recognises that the researcher's prior beliefs and disciplinary background influence the study. Bias was mitigated through:

- neutral phrasing and randomised stimulus order;
- co-development of the instrument with external domain experts;
- third-party audit of ambiguous coding decisions in the market review.

These steps are detailed in chapter 5.4, but they are rooted in the belief that transparency and critical self-reflection increase trustworthiness.

Paradigm-to-Methodological Coherence

The logical chain from worldview to concrete procedures are summarised as follows:

Philosophical layer	This study's stance	Methodological manifestation
Ontology	Critical realism – preferences exist, observable imperfectly	Treat preferences as latent but measurable through carefully designed choice tasks
Epistemology	Falsification & statistical estimation	Hypotheses H ₁ (1-3) tested with inferential statistics; confidence intervals quantify uncertainty
Methodological approach	Deductive, quantitative	Within-subjects experiment + structured market audit

Methods	Standardised online survey; Randomisation, MaxDiff scaling, secondary database coding logistic regression
---------	---

The within-subjects design makes each respondent their own control, enhancing internal validity. Combining primary (survey) and secondary (market audit) data enables triangulation, meeting the post-positivist goal of converging evidence from multiple sources.

Rigour and Quality Criteria

This summarises how methodological quality standards (Creswell, 2014; Greener, 2008) are operationalised.

Internal validity	Randomised presentation order; uniform scenario; fatigue-minimising 3-minute survey
External validity	Purposive but expert sample of business professionals; market audit cross-checks survey results
Reliability	Standardised Typeform instrument; piloting; double coding of market data
Objectivity	Pre-set inclusion rules; external audit of coding; transparent reporting of analytical decisions

Ethical Considerations

Ethical integrity, though sometimes treated separately from philosophy of science, is integral to axiology. The study followed five principles: informed consent, anonymity, right to withdraw, data minimisation and secure storage. Because no personal identifiers were collected and the scenario was hypothetical, risk to participants was minimal.

Chapter Summary

This chapter established a post-positivist foundation for the thesis. It articulated a critical realist ontology, a probabilistic epistemology, and an objective but reflexive axiology. From these premises flow a deductive quantitative strategy, specifically a within-subjects experiment triangulated with a market audit and a set of rigour criteria emphasising validity, reliability and objectivity. Subsequent chapters build on this philosophical footing to detail methods, present results and discuss implications.