International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Multi-Service AI Orchestration in SaaS

Rudransh Shahi[1], Pawan Chaturvedi[2], Rohit Srivastava[3]

[1,2,3]BBDNIIT

Abstract

Contemporary AI Software-as-a-Service platforms exhibit fragmentation across multiple providers including OpenAI, Anthropic, and Google Cloud AI, compelling users to navigate heterogeneous authentication systems, billing structures, and service quality considerations. This research presents MULTI SERVICE AI ORCHESTRATION, a novel framework integrating Cost-Spectrum Contrastive Routing with adaptive Machine Learning as a Service composition and multi-stakeholder explainability mechanisms tailored for subscription-based platforms. Through comprehensive evaluation utilizing 1,247 authentic user queries spanning 147 active participants over a twelve-week deployment period, our findings demonstrate cost reduction of 23.1 percent while sustaining 97.1 percent quality across text, image, and document processing services. The framework achieves 92 percent fairness equity across subscription tiers, with free-tier users experiencing 92 percent of paid-tier quality. Additionally, we observe 80.4 percent explanation robustness accompanied by 33 percent improvement in user trust metrics. The synergistic integration of routing and composition mechanisms yields 34.8 percent total cost reduction, enabling simultaneous optimization of efficiency and equity, addressing a critical gap in existing multi-service orchestration literature.

Keywords: AI service orchestration, cost-aware routing, fairness, explainability, SaaS platforms.

INTRODUCTION

The proliferation of heterogeneous artificial intelligence services has created substantial operational fragmentation within the technology landscape. Users frequently alternate between platforms, consuming three to five minutes per composite task while managing distinct authentication credentials and navigating service-specific pricing models ranging from 0.0005 to 0.03 dollars per API call. This fragmented ecosystem introduces several unaddressed challenges encompassing cost optimization without budget-aware routing mechanisms, quality maintenance under service degradation conditions, transparency requirements mandated by regulatory frameworks including GDPR Article 22 and the EU AI Act, and fairness considerations across subscription tiers.

The contemporary AI service landscape presents fundamental questions regarding resource allocation efficiency and equitable access. Organizations deploying Software-as-a-Service platforms must balance competing objectives including operational cost minimization, service quality assurance, user satisfaction across diverse subscription levels, and regulatory compliance regarding algorithmic decision-making transparency.

This research introduces MULTI-SERVICE AI ORCHESTRTAION, a comprehensive framework designed to address these multifaceted challenges through integrated mechanisms for intelligent service routing, adaptive composition, and transparent decision-making. Our approach builds upon recent advances in cost-aware routing and service composition while introducing novel extensions addressing

subscription-based fairness constraints and multi-stakeholder explainability requirements. The primary contributions of this work include subscription-aware Cost-Spectrum Contrastive Routing extension incorporating tier-specific cost weighting parameters enabling cost-fairness trade-offs absent in original formulations. We introduce multi-modal fingerprinting enabling unified routing across text, image, and document modalities with minimal generalization degradation. Our fairness-constrained composition approach adapts existing adaptive Machine Learning as a Service frameworks with explicit fairness constraints achieving substantial tier-based equity. We develop multi- stakeholder explainability mechanisms providing context- dependent explanations achieving robust performance across diverse user populations. Finally, we present production validation spanning 1,247 authentic queries from 147 users over twelve weeks demonstrating synergistic cost reduction.

RELATED WORK

Existing research in cost-aware routing has demonstrated significant potential for reducing operational expenses in large language model deployments. Recent work by Shirkavand and colleagues achieves 28 percent cost reduction for pay-per-call language model services through cost- spectrum InfoNCE loss mechanisms. However, subscription models incorporating tiered access structures and fixed budget constraints remain unaddressed. No prior work examines cost-aware Fairness in machine learning has generated extensive literature addressing demographic parity and equalized odds. However, subscription-tier fairness operationalization represents a novel contribution. Existing fairness research does not define fairness metrics for resource allocation across user cohorts with heterogeneous budget constraints, representing a significant gap in subscription-based service platforms.

SYSTEM ARCHITECTURE

The MULTI-SERVICE AI ORCHESTRATION

framework comprises five primary components that work in concert to enable intelligent, fair, and transparent multi-service orchestration. The frontend layer implements a React-based user interface that facilitates task submission and results visualization. The routing engine utilizes Cost-Spectrum Contrastive Routing mechanisms for service selection with sub-five millisecond latency requirements. The composition engine implements adaptive Machine Learning as a Service with integrated fairness constraints. The explainability module provides multi-stakeholder transparency through context-aware explanation generation. The data layer employs PostgreSQL databases to store all 1,247 queries, comprehensive decision logs, and complete audit trails for regulatory compliance and system improvement.

The user subscription tiers define three distinct service levels. Free tier users experience a cost weight parameter of 2.0 with a monthly allocation of 50 API calls. Professional tier users receive cost weight parameter of 1.0 with monthly allocation of 2,000 API calls. Enterprise tier users benefit from cost weight parameter of 0.3 with unlimited API call allocation. These tier-specific parameters enable differentiated service delivery while maintaining fairness constraints.

METHODOLOGY

1. Core Innovation: Tier-Aware Cost Weighting

The original cost-spectrum contrastive routing formulation selects services through cosine similarity maximization between query embeddings and service descriptors, which is penalized by service costs

through a uniform cost sensitivity parameter. Our tier-aware adaptation modifies this approach by introducing subscription-specific cost weighting.

The routing decision maximizes the objective function combining embedding similarity with tier-adjusted cost penalty. The cost weight parameter for individual users equals the base cost sensitivity multiplied by tier-specific factors. Free-tier users receive higher cost sensitivity prioritizing economical service selection. Enterprise users experience reduced cost sensitivity emphasizing quality optimization. We empirically calibrated these parameters through iterative A/B testing minimizing refund requests and support tickets while maximizing user satisfaction metrics.

2. Multi-Modal Fingerprinting

Our approach extends fingerprinting mechanisms across three distinct modalities. Text services including GPT-4 and Claude utilize logit probability distributions with 256-dimensional representations. Image generation services including DALL-E employ latent space embeddings with 512dimensional representations. Document processing services implement feature extraction producing 256-dimensional representations. All fingerprint vectors undergo normalization to unit hypersphere, enabling unified metric space comparisons across heterogeneous service types.

3. Fairness-Constrained Composition

We operationalize fairness through tier-based quality equity metrics. Fairness for each subscription tier equals the ratio of free-tier quality to paid tier quality expressed as percentage, with target threshold of 85 percent or higher. Modified composition rules incorporate Rule 6 addressing Cost Equity Score ensuring free-tier budget compliance. The fairness- adjusted composition score sums weighted rule contributions minus fairness penalty proportional to maximum fairness impact, where fairness penalty weight parameter equals 0.15 controlling trade-off intensity.

4. Explanation Robustness

For each routing or composition decision, we generate ten paraphrased query variants through T5 sequence-to-sequence models. Feature importance rankings computed for original and paraphrased queries undergo Spearman rank correlation analysis. Robustness equals the mean correlation across all paraphrased variants. We establish target threshold of
0.75 or higher, with 14 of 300 explanations
representing 4.7 percent flagged below 0.70 threshold for manual expert review and refinement.

RESULTS

1. Cost-Accuracy Trade-off

TABLE I. COST AND QUALITY COMPARISON

Method	Cost/Query	Quality	Fairness
Baseline	$0.0502	95.2%	68%
CSCR-Standard	$0.0358	92.1%	72%
Multi-Service AI orchestration	$0.0386	97.1%	92%

Our tier-aware approach achieves 23.1 percent cost reduction maintaining 97.1 percent quality preservation. This represents 5.6 percentage point cost reduction sacrifice compared to standard Cost-Spectrum Contrastive Routing achieving 28.7 percent reduction, but yields 20 percentage point fairness improvement. Tier-specific analysis reveals free-tier users experience 42.4 percent cost reduction,

professional tier users achieve 18.0 percent reduction, and enterprise tier users realize 0.8 percent reduction. Average routing latency measures 2.3 milliseconds with 95th percentile at 4.8 milliseconds, successfully meeting sub-five millisecond target requirement.

2. Fairness Across Tiers

Service-type specific fairness analysis demonstrates consistent equity across modalities. Text services achieve 91 percent fairness with free-tier users receiving 82.1 percent quality and paid-tier users receiving 92.4 percent quality. Image generation services attain 93 percent fairness with free-tier quality at 78.2 percent and paid-tier quality at 84.1 percent. Document processing services reach 92 percent fairness with free-tier quality at 77.1 percent and paid-tier quality at 84.2 percent. The cost of fairness implementation manifests as 2.8 percent aggregate platform cost increase yielding 20 percentage point fairness improvement. Free-tier users experience 24 percentage point service quality improvement relative to baseline configuration without fairness constraints.

3. Multi-Modal Generalization

Cross-modality performance evaluation through Area Under Deferral Curve metrics demonstrates acceptable generalization. Text-to-image transfer achieves 0.512 AUDC representing 0.016 degradation compared to image-trained models. Text-to-document transfer reaches 0.438 AUDC representing 0.014 degradation compared to document-trained models. Unified multi-modal routing achieves 0.498 AUDC across all modalities with 2 to 3 percent performance loss deemed acceptable for unified infrastructure benefits.

4. Cold-Start Learning

TABLE II. LEARNING CURVE ANALYSIS

Labeled Examples	Accuracy	Confidence
10	63%	Low
100	**86%**	**Good**
500	93%	Excellent

Cold-start performance analysis indicates 100 labeled examples provide sufficient training data for production deployment, exceeding 85 percent confidence threshold established for acceptable routing accuracy.

5. Degradation Detection

Service degradation detection mechanisms provide 15.2 Hour average lead time before user submitted problem reports, with 95 percent confidence interval spanning 12 to 18 hours. Degradation severity influences detection latency. Ten percent quality degradation receives detection 3.2 hours before eventual user reports at 22.1 hours. Twenty-five percent quality degradation achieves detection 0.7 hours before user reports at 6.8 hours. Overall prevention rate reaches 75 percent with two issues reported by users compared to eight issues detected proactively by system monitoring.

6. Composition Recommendation

Composition modification recommendation accuracy achieves 85.7 percent with 18 of 21 accepted recommendations yielding measurable improvement. Quality improvement from accepted recommendations averages 8.3 percent with 95 percent confidence interval spanning 6.1 to 10.5 percent. User acceptance rates vary by recommendation mode with 95 percent acceptance for automatic implementation with user deferral option and 42 percent acceptance for manual recommendation requiring explicit approval. Explanation quality correlates with acceptance rates at

0.67 correlation coefficient, suggesting transparency enhances user trust in system recommendations.

7. Explanation Robustness

TABLE III. STAKEHOLDER-SPECIFIC ROBUSTNESS

Stakeholder Type-	Robustness
Student	78.1%
Teacher	80.4%
Admin	82.3%

Multi-stakeholder explanation robustness demonstrates consistent performance across user populations. Overall robustness measures 80.4 percent successfully exceeding 75 percent target threshold across all stakeholder categories.

8. User Trust Impact

Controlled A/B study with 94 participants using counterbalanced design reveals significant trust improvements attributable to explanation provision. Control group without explanations exhibits minimal trust change from 5.2 to 5.4 out of 10, representing statistically non-significant 0.2 point increase. Treatment group receiving explanations demonstrates substantial trust improvement from 5.1 to 6.8 out of 10, representing statistically significant

1.7 point increase with p-value below 0.001. Explanation provision produces beneficial calibration effects. User overconfidence decreases from 58 percent to 31 percent with p-value below 0.05. Verification behavior increases from 12 percent to 34 percent with p-value below 0.01, suggesting explanations promote appropriate skepticism rather than blind acceptance of system recommendations.

9. Synergy Analysis

TABLE IV. COMPONENT INTEGRATION BENEFITS

Configuration	Cost	Quality	Fairness
Baseline	$0.0502	95.2%	68%
CSCR only	$0.0358	92.1%	72%
MLaaS only	$0.0482	96.1%	68%
Combined	**$0.0327**	**94.1%**	**92%**

Integration synergy analysis reveals super-additive benefits from combined routing and composition mechanisms. Expected additive benefit predicts 0.0164 dollar reduction representing 32.7 percent improvement. Observed performance achieves 0.0175 dollar reduction representing 34.8 percent improvement, yielding 1.1 percent synergy bonus beyond independent component contributions. This synergy emerges because composition mechanisms identify underperforming services typically offering lower costs, enabling routing engine to substitute alternatives without incurring fairness penalties. The integrated approach optimizes across both service selection and service pool composition simultaneously.

System Overhead Per-query latency for routing decisions without composition averages 2.2 milliseconds, comprising 1.2 milliseconds for embedding generation, 0.8 milliseconds for k-nearest neighbor search, and 0.2 milliseconds for tier adaptation computation. Composition assessment executing hourly as background process requires 41.6 milliseconds encompassing service assessment, rule evaluation, and contextual multi-armed bandit optimization, with results cached for immediate routing access.

DISCUSSION

Our findings reveal several key insights regarding multi-service AI orchestration. First, tier-aware routing achieves 92 percent fairness through explicit mechanism design rather than emergent optimization properties. Cost-blind routing approaches including standard Cost- Spectrum Contrastive Routing improve fairness marginally from 68 percent to 72 percent, validating necessity of fairness-specific engineering rather than relying on implicit optimization behaviors.

Second, multi-modal fingerprinting generalizes across modalities with 2 to 3 percent Area Under Deferral Curve degradation, representing acceptable trade-off for unified routing infrastructure benefits. Separate modality- specific models would increase system complexity substantially while providing minimal performance gains. Third, composition adaptation mechanisms provide 15.2 hour detection lead time enabling proactive system evolution. This contrasts sharply with reactive approaches responding only after user complaints accumulate. The 75 percent prevention rate demonstrates substantial value in anticipatory service quality management.

Fourth, explanation provision calibrates user trust appropriately rather than promoting unconditional acceptance. Verification behavior increases to 34 percent while overconfidence decreases to 31 percent, suggesting users engage more critically with system recommendations when provided transparent rationales. This represents desirable outcome for high-stakes decision contexts.

Fifth, synergy between routing and composition components yields 1.1 percent bonus beyond additive contributions. Composition identifies underperforming low-cost services enabling routing substitution, while routing provides signals enabling composition to prioritize high-impact optimization opportunities. This bidirectional information flow creates multiplicative rather than additive benefits.

1. Limitations

Several limitations constrain interpretation and generalization of our findings. The twelve-week evaluation period may not capture multi-month or seasonal service degradation patterns common in production environments.

Our service pool comprises eight distinct services, while realistic platforms provision 50 to 100 or more services introducing substantially greater complexity.

User population characteristics exhibit demographic concentration with 87 percent technical background and mean age 24.3 years, limiting conclusions regarding diverse user populations. Fairness operationalization addresses tier-based equity exclusively without examining demographic parity across protected attributes including gender, race, or socioeconomic status.

2. Generalizability

Despite these limitations, our approach generalizes to several related domains. Federated learning systems face analogous client selection challenges balancing computational cost, communication overhead, and fairness across heterogeneous devices. Edge computing deployments must route computation across devices exhibiting diverse capabilities, energy constraints, and availability patterns. Healthcare diagnostic systems require service triaging across multiple diagnostic modalities optimizing accuracy, cost, and wait times. Digital marketplaces need vendor selection mechanisms ensuring seller fairness while optimizing buyer outcomes. Each domain exhibits structural similarities to multi- service AI orchestration including heterogeneous service providers, diverse quality-cost trade-offs, and competing stakeholder interests requiring explicit fairness mechanisms.

CONCLUSION

This research addresses the critical gap of cost- aware, fair, and transparent multi-service orchestration within subscription-based Software-as- a Service platforms. Our tier-aware Cost-Spectrum Contrastive Routing adaptation combined with fairness-constrained composition and multistakeholder explainability mechanisms jointly achieve 23.1 percent cost reduction while maintaining 92 percent fairness equity and 80.4 percent explanation robustness across diverse stakeholder populations.

The observed 34.8 percent synergistic cost reduction demonstrates that efficiency optimization and equity assurance represent complementary rather than conflicting objectives when appropriate mechanisms coordinate routing and composition decisions. This challenges conventional wisdom suggesting inevitable trade-offs between operational efficiency and fairness considerations.

Future research directions include multi-year longitudinal evaluation capturing seasonal patterns and long-term service evolution, scaling to larger service pools approaching realistic platform complexity, incorporating demographic diversity in user populations enabling comprehensive fairness analysis across multiple protected attributes, and exploring cross-domain generalization to federated learning, edge computing, healthcare,
and marketplace applications.

The MULTI-SERVICE AI ORCHESTRATION

framework provides practitioners with actionable mechanisms for deploying fair, efficient, and transparent multi- service AI platforms while offering researchers novel formulations of subscription-tier fairness, multi- modal routing, and explanation robustness suitable for continued investigation and refinement.

REFERENCES

1. A. Kanneganti, S. Ramanathan, and P. Garg, "Adaptive composition of machine learning services for internet-of-things applications," in Proc. IEEE Int. Conf. Cloud Comput. (CLOUD), 2025, pp. 234–241.
2. R. Shirkavand, M. Alimadadi, and K. Zare, "Cost- aware contrastive routing for distributed language model deployment," in Proc. ACM Conf. Mach. Learn. Syst. (MLSys), 2025, pp. 456–468.
3. S. Vascotto, L. Marinelli, and P. De Stefano, "Robustness of ensemble explanation methods for machine learning systems," in Proc. European Conf. Artif. Intell. (ECAI), 2025, pp. 112–124.
4. C. Fairfield, "Fairness in machine learning systems: Principles and practices," ACM Comput. Surv., vol. 52, no. 4, pp. 78–102, 2024.
5. M. P. Johnson and T. K. Richardson, "Multi- stakeholder explainability for AI systems," in Proc. AAAI Conf. Artif. Intell. (AAAI), 2024, pp. 345–352.
6. D. Herrera, G. Laurent, and S. Chen, "Service composition under budget constraints," in Proc. IEEE Int. Conf. Serv. Comput. (SCC), 2024, pp. 89–96.
7. B. Morrison and R. Thompson, "Explainability evaluation frameworks for AI orchestration," IEEE Trans. Softw. Eng., vol. 50, no. 2, pp. 210– 228, 2024.
8. "Towards Explainable Vulnerability Detection with Large Language Models" (Oct 2025)
9. "Advanced Smart Contract Vulnerability Detection via LLM- Powered Multi-Agent Systems" (Oct 2025)
10. Journal of Artificial Intelligence Research (JAIR) (2024)