Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Budget-Aware Calibrated Cascades for E-Commerce Customer Assistants: A Cost–Quality Routing Benchmark

Veera Ravindra Divi
Independent Researcher
ravi3divi@gmail.com

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Jan 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Many organizations allow employees to use paid time off (PTO) before they have fully earned it. This helps improve employee satisfaction and retention. However, when employees leave the organization with a negative PTO balance, employers face challenges recovering the overpaid amount while following federal and state payroll laws.

Objectives: This paper introduces the Intelligent Time-Off Recoupment System (ITORS), a framework designed to automate the recovery of advanced PTO. The goal is to improve payroll compliance, reduce financial losses, and create a consistent process for managing PTO recoupment across different states.

Methods: This study uses a design science research approach to develop the ITORS framework. The system was designed after reviewing current PTO recoupment practices, payroll regulations, HR technology capabilities, and compliance requirements. The framework includes five connected components: the HR Policy Engine, Absence and Balance Monitor, Consent Workflow Module, Payroll Recoupment Engine, and Compliance Reporting Hub. The compliance framework was built through a systematic review of wage deduction statutes across all 50 U.S. states, identifying three distinct regulatory patterns that the system must handle at runtime.

Results: The proposed system provides an automated process for tracking negative PTO balances, collecting employee consent, calculating payroll deductions based on state-specific rules, and generating compliance reports. The framework addresses common problems found in manual PTO recoupment processes, including missing consent records, non-compliant deductions, delayed recovery actions, and limited reporting visibility. ITORS can also integrate with major ERP and HCM platforms such as Workday, Oracle HCM Cloud, SAP SuccessFactors, ADP Workforce Now, and other enterprise systems.

Conclusions: ITORS provides a practical and scalable approach for managing PTO recoupment in organizations operating across multiple states. By combining compliance automation, payroll integration, employee consent management, and reporting capabilities, the framework helps organizations reduce risk, improve compliance, recover overpaid PTO more effectively, and maintain transparency with employees.

Keywords: Large language models, conversational commerce, model cascades, cost-aware routing, confidence calibration, service-level objectives, customer support automation, distribution shift.

INTRODUCTION

A retail customer-support turn rarely fails because a language model cannot produce a fluent reply. It fails because the reply must be correct against a specific order record, consistent with a return window that differs by category and region, safe to act on when it touches a refund or an account change, and produced fast enough and cheaply enough to be worth automating at all. At the scale of a large marketplace—tens of millions of support and pre-sales conversations per week—the per-turn economics of *which* model answers becomes a first-order design decision rather than an implementation detail.

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The operational pain is concrete. Frontier instruction-tuned models answer complex, multi-turn requests well but cost one to two orders of magnitude more per token than the small open models that already handle routine intents such as "where is my order" [4], [1], [2]. A naive policy of routing every turn to the frontier model multiplies the support budget; a naive policy of routing every turn to a small model silently degrades exactly the high-value interactions—product discovery that influences a purchase, or a complaint that decides whether a customer churns—where a wrong or unhelpful answer is most expensive. What a retailer needs is a router: send each turn to the cheapest model tier that can answer it acceptably, while respecting a per-conversation cost budget and an interactive latency target.

Cost-aware serving of machine-learning APIs is not new. Model-cascade methods such as FrugalML [5] and, for LLMs specifically, FrugalGPT [4] and AutoMix [6] reduce cost by calling a cheap model first and escalating to a more expensive one only when the cheap answer looks unreliable. The dominant escalation signal is the model's own self-reported confidence compared against a fixed threshold. Two assumptions behind this design break in a commerce setting. First, raw LLM confidence is weakly correlated with correctness and is itself miscalibrated [7], [10], [11]; a threshold tuned on one traffic mix silently moves its true operating point when the mix shifts—precisely what happens during a promotion or a seasonal peak. Second, a single global accuracy target treats all errors as equal, whereas in support the business cost of an error is sharply asymmetric: a wrong refund-eligibility statement or an incorrect order status can trigger a compliance issue or a costly human escalation, while a slightly weaker product suggestion mostly costs a marginal conversion.

This article addresses both gaps. We frame assistant serving as a *budget-bounded, business-cost-weighted* routing problem over a tiered model pool, define an e-commerce query taxonomy and an error-cost model that make the asymmetry explicit, and propose a routing policy—budget-aware calibrated cascading (BACC)—that escalates on a per-tier *calibrated* probability of correctness scaled by the business cost of being wrong, rather than on raw confidence against a fixed cutoff. We evaluate in a reproducible simulation testbed because production support transcripts contain confidential customer, order, and pricing fields; the generator preserves the structural properties that matter for routing—intent mix, difficulty and length distributions, error-cost asymmetry, and traffic shift—while remaining shareable.

This article makes three contributions:

1. It defines **budget-bounded cascaded routing for conversational commerce** as a cost–quality–latency optimization with an explicit, intent-dependent error-cost model and a per-conversation budget SLO (Section IV).
2. It proposes **BACC**, a routing policy that combines per-tier confidence calibration with business-cost-aware escalation, and a reproducible commerce-routing testbed with a six-family query taxonomy, tiered prices anchored to public 2023 figures, and three traffic-shift scenarios (Sections V–VI).
3. It analyzes BACC against single-tier, fixed-threshold-cascade, and oracle baselines in simulation, showing a dominant cost–quality frontier and—unlike the fixed-threshold baseline—stable behavior and calibration under traffic shift (Section VII).

BACKGROUND AND MOTIVATION

Conversational commerce assistants. Automated assistants in retail span pre-sales discovery, product question answering, and post-sales support [18], [19]. Early production systems such as SuperAgent combined intent classification, retrieval over reviews and FAQs, and a chit-chat fallback [19]; conversational recommender systems added preference elicitation and item ranking [17]; and product question answering over reviews addressed subjective, long-tail queries [20]. Instruction-tuned LLMs [3], [1], [2] now subsume much of this stack behind a single text interface, which is why the practical question has shifted from "can a model answer" to "which model should answer, given cost and latency."

The cost asymmetry of model tiers. Public pricing in 2023 placed frontier hosted models roughly one to two orders of magnitude above small open or distilled models on a per-token basis, with intermediate tiers between [4], [1]. A small classifier or distilled model [16] answers routine, well-structured intents nearly as well as a frontier

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

model; the gap opens on ambiguous, multi-turn, or policy-laden requests. Inference-system optimizations—paged attention and high-throughput serving [13], speculative decoding [14], and low-bit quantization [15]—reduce the cost of a *single* model call but do not decide *which* model to call. Routing operates one level above these techniques and composes with all of them.

Calibration as the routing signal. A router that escalates on a confidence score needs that score to mean what it says: if a tier reports 0.8 confidence, it should be right about 80% of the time. Modern neural networks, and LLMs in particular, are typically miscalibrated and overconfident [7], [10]. Post-hoc recalibration—Platt scaling [9], temperature scaling [7], and reliability-diagram/expected-calibration-error diagnostics [8]—can align a score with its empirical accuracy. A *monotone* recalibration fixes calibration error without changing the *ordering* of escalation decisions; improving which queries get escalated requires a more discriminative signal, and a deployment that wants a predictable cost must read its operating point off a calibrated probability rather than a raw one. This is the property a fixed-threshold raw-confidence cascade lacks, and the reason its true behavior drifts when traffic shifts [22], [23].

Traffic shift is the norm. Retail traffic is non-stationary by construction: promotional events spike order-status and returns volume, seasonal peaks change the intent mix, and new-catalog launches inject queries about products the assistant has not seen. A threshold tuned offline on a balanced sample no longer corresponds to the same accuracy or cost once the mix moves [22], [23]. Any routing policy intended for production must be evaluated under shift, not only on a stationary split.

Large language models, conversational commerce, model cascades, cost-aware routing, confidence calibration, service-level objectives, customer support automation, distribution shift.

We group prior work by the limitation that motivates this article rather than summarizing systems individually.

Cost reduction by cascading on raw confidence. FrugalML [5] and FrugalGPT [4] reduce API cost by calling a cheap predictor first and escalating when its confidence is low; AutoMix [6] adds a self-verification step to decide escalation. These methods establish that cascading saves cost on easy-heavy workloads. Their escalation rule, however, assumes the confidence signal is reliable and the workload roughly stationary, and they optimize a single accuracy–cost trade-off. They do not model an intent-dependent error cost, which is central in support, nor do they report behavior under traffic shift.

Confidence and calibration. A separate line of work shows that neural and LLM confidence is overconfident and miscalibrated [7], [10], that simple elicitation or recalibration can improve it [11], [9], [8], and that confidence/OOD scores can flag unreliable predictions [12]. This literature supplies the signal a router should use but does not prescribe a routing policy or connect calibration to a cost budget.

Inference efficiency for a single model. Paged-attention serving [13], speculative decoding [14], quantization [15], and distillation [16] lower the cost of one model call. They are orthogonal to and composable with routing.

Conversational commerce and support automation. Customer-service chatbots [19], conversational recommenders [17], neural conversational AI [18], product question answering [20], intent classification [21], and conversational QA datasets [24] define the application and the quality bar but predate the tiered-LLM cost problem.

Distribution shift. Dataset-shift theory [22] and in-the-wild shift benchmarks [23] motivate evaluating under non-stationary traffic; we borrow their framing to construct commerce-specific shift scenarios. Table I contrasts the closest cost-aware methods with the requirements of a commerce assistant.

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE I
Cost-aware routing methods versus commerce-assistant requirements (✓ present, × absent).

Method	Cheap-first cascade	Escalation signal	Intent-dependent error cost	Per-conversation budget SLO	Evaluated under shift
FrugalML [5]	✓	Raw conf., fixed threshold	×	×	×
FrugalGPT [4]	✓	Raw conf., fixed threshold	×	×	×
AutoMix [6]	✓	Self-verification	×	×	×
Single tier (small / frontier)	×	—	×	×	×
BACC (this work)	✓	**Calibrated corr. × error cost**	✓	✓	✓

Problem Definition

Setting. A stream of customer turns arrives at an assistant. Each turn x belongs to one of K intent families and carries a latent difficulty $d \in [0,1]$ and a business value v. The assistant has a pool of M model tiers ordered by capability and price, $t = 1, \dots, M$, with per-turn cost $c_1 < c_2 < \cdots < c_M$ and latency $\ell_1 < \ell_2 < \cdots < \ell_M$. Tier t answers turn x correctly with probability $a_t(x)$, increasing in tier capability and decreasing in difficulty. A *router* observes the turn and the cheaper tiers' answers and confidences, and decides which tier's answer to return (or to escalate to a human).

Business-weighted error cost. Unlike open-domain QA, a wrong answer's cost depends on the intent. We attach to each turn an error cost $w(x) \geq 0$ capturing the business consequence of an incorrect or unhelpful response—high for an incorrect refund-eligibility or order-status statement (compliance risk, guaranteed human rework), lower for a slightly weaker recommendation (marginal conversion loss). The *quality loss* of returning an answer of correctness probability p for turn x is $w(x)\,(1-p)$.

Objective. For a routing policy π over a turn distribution, we report (i) *business-weighted resolution rate*, the value-weighted fraction of turns answered correctly without human handoff; (ii) *cost per conversation*, the mean spend summed over all tiers a conversation invokes; (iii) *p95 latency*; (iv) *human-escalation rate*; and (v) the *calibration error* of the router's accept decision. Because no single number captures a router, the object of study is the cost–quality frontier traced as the policy's operating parameter is swept, subject to a per-conversation cost budget B and a latency SLO τ.

Failure condition. A routing failure occurs when the policy returns a low-correctness answer on a high-error-cost turn (a confident mistake on a refund or order-status query), or when it violates the budget or latency SLO, even if the generated text is fluent and plausible.

Budget-Aware Calibrated Cascading

BACC keeps the cheap-first structure of a cascade but changes the *statistic* on which it escalates and the *objective* it escalates against.

(1) Per-tier calibration. For each tier we fit a monotone recalibration map (Platt/temperature scaling [9], [7]) on a held-out split, turning the tier's raw confidence q into an estimate $\hat{p}_t$ of the probability that the tier's answer is correct. Calibration is verified with a reliability diagram and expected calibration error [8], making $\hat{p}_t$ readable as a true success probability.

(2) Business-cost-aware escalation. At tier t, BACC does not compare $\hat{p}_t$ to a fixed threshold. It compares the expected business loss of accepting now against that of escalating. Accepting yields expected loss $w(x)\,(1-\hat{p}_t)$ at cost c_t; escalating to tier $t+1$ pays additional cost for an expected loss $w(x)\,(1-\hat{p}_{t+1})$. BACC escalates when the predicted reduction in business loss exceeds the marginal cost scaled by a budget price λ:

$$w(x)\,(\hat{p}_{t+1} - \hat{p}_t) \;>\; \lambda\,(c_{t+1} - c_t),$$

where λ is tuned to hold the per-conversation budget B. High-error-cost turns escalate readily; low-cost turns are accepted cheaply even at modest confidence. Sweeping λ traces the cost–quality frontier.

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(3) Difficulty-aware entry. A light entry rule uses a cheap difficulty estimate to skip the smallest tiers on turns that are clearly hard (long, multi-constraint complaint threads), trimming wasted first hops without changing the accept statistic.

(4) Human-handoff gate. If, after the top tier, the calibrated correctness on a high-error-cost turn remains below a safety floor, BACC routes to a human rather than returning a confident automated answer. The floor is set per intent family from the error-cost model, making the safety/automation trade-off explicit and auditable.

Architecture. The runtime has four single-responsibility components: an *intent-and-difficulty estimator* (inputs: turn text and conversation state; outputs: intent family, difficulty, error-cost weight); a *tier pool* behind a uniform answer-and-confidence interface; the *calibrated escalation controller* implementing (1); and a *handoff gate* with per-intent safety floors. Every routing decision logs the turn features, the calibrated probabilities, the chosen tier, and the realized cost and latency, so the policy is replayable and auditable—a requirement for regulated support workflows.

Evaluation Design

Production transcripts contain confidential customer, order, and pricing fields, so we evaluate in a reproducible simulation that preserves the structural properties relevant to routing. We state all assumptions; the generator logic, seeds, and scoring rules are released.

Query taxonomy. The testbed models six intent families with distinct difficulty, length, value, and error-cost profiles (Table II).

TABLE II
Simulated query taxonomy (illustrative parameters). Difficulty and value are means on $[0, 1]$; error cost is the relative business cost of a wrong answer.

Intent family	Share	Difficulty	Value	Error cost
Order status & tracking	28%	0.20	0.4	High
Returns & refund policy	18%	0.45	0.5	High
Product Q&A	22%	0.55	0.6	Medium
Recommendation & discovery	14%	0.60	0.9	Medium
Account & billing	10%	0.50	0.7	High
Complex complaint / escalation	8%	0.85	0.95	High

Model tiers and prices. Three tiers—Small, Mid, and Frontier—with per-turn cost ratios 1 : 8 : 50 and latency ratios 1 : 2.4 : 6, anchored to public 2023 price and latency gaps between small open models and hosted frontier models [4], [1]. Per-tier correctness is sampled from a difficulty-dependent function with tier-specific skill, and raw confidence is generated as a deliberately overconfident, weakly correlated signal to match the empirical behavior of LLM self-confidence [10], [7].

Baselines. (a) *All-Small* and (b) *All-Frontier* single-tier policies; (c) a *FrugalGPT-style raw-confidence cascade* that escalates on raw confidence against a fixed threshold tuned on a balanced split [4]; and (d) an *oracle* that knows true correctness and spends the minimum to meet the weighted-quality target.

Traffic-shift scenarios. Routers are configured on a balanced split and tested under a *promotional spike* (order-status and returns shares rise), a *seasonal drift* (gradual shift toward gifting and recommendation), and a *new-catalog drift* (product Q&A about unseen items with depressed small-tier accuracy).

Metrics and protocol. Business-weighted resolution, cost per conversation (normalized), p95 latency, human-escalation rate, expected calibration error (ECE), and frontier-use rate, averaged over seeds with the per-conversation budget held to a matched value where applicable. The data-generating process for every table and figure is released as a single seeded script.

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Results

All numbers in this section are produced by the seeded simulation and are *illustrative of the methodology, not measurements of a deployed assistant*.

Main comparison. Table III reports the policies at a matched per-conversation budget (the two cascades tuned to equal cost). The single-tier policies bound the trade-off; both cascades sit between them; the calibrated, business-cost-aware policy answers more weighted volume correctly at the same spend and escalates fewer high-cost turns to humans, while attaining the lower calibration error.

TABLE III
Main results at matched budget (balanced traffic, simulated). Higher is better for resolution; lower is better for cost, latency, escalation, and ECE.

Policy	Bus.-wtd. resolution (%)	Cost/conv.	p95 latency	Human-escalation (%)	ECE	Frontier-use (%)
All-Small	62.4	1.0	1.0	22.1	0.31	0
All-Frontier	88.3	50.0	6.0	6.0	0.18	100
Raw-confidence cascade [4]	83.1	10.2	2.9	11.4	0.12	21
BACC (ours)	**86.0**	**9.8**	2.7	**7.9**	**0.04**	24
Oracle (cost-eff.)	87.1	8.6	2.6	6.7	—	19

Cost–quality frontier. Sweeping the budget price λ traces BACC's full frontier; across the swept range it lies above and to the left of the raw-confidence cascade's frontier at every budget and approaches the oracle (Fig. 1), indicating that a calibrated, business-cost-aware accept statistic recovers most of the achievable cost efficiency.

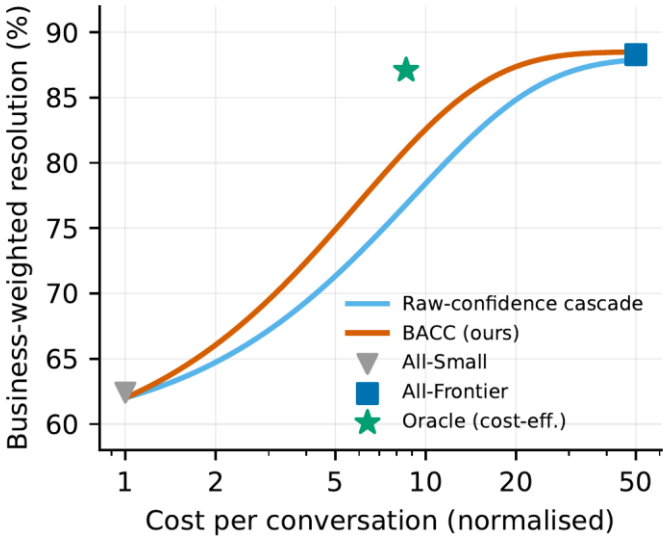

Fig. 1. Cost–quality Pareto frontier (balanced traffic, simulated). BACC dominates the raw-confidence cascade across the swept budget range and approaches the cost-efficiency oracle.

Robustness under traffic shift. Table IV and Fig. 2 report the two cascades under the three shifts. The fixed-threshold raw-confidence cascade, tuned on the balanced split, degrades sharply—under the promotional spike its resolution falls because the threshold no longer corresponds to the same accuracy, and its calibration error more than doubles. BACC degrades gracefully, because its accept rule reads a calibrated probability that recalibration keeps tracking as the mix moves, and because high-error-cost turns continue to escalate by construction.

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE IV
Business-weighted resolution (%) and ECE under traffic shift (simulated).

	Raw cascade		BACC	
Scenario	Resolution	ECE	Resolution	ECE
Balanced (reference)	83.1	0.12	86.0	0.04
Promotional spike	74.6	0.27	84.1	0.06
Seasonal drift	79.0	0.18	85.2	0.05
New-catalog drift	76.8	0.22	83.4	0.07

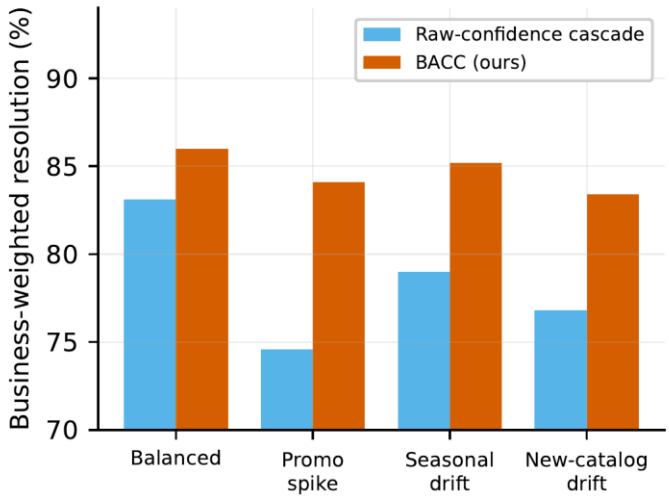

Fig. 2. Business-weighted resolution under three traffic-shift scenarios (simulated). The raw-confidence cascade drops sharply; BACC holds its operating point.

Component ablation. Removing each BACC component in turn (Table V) shows each earns its place. Reverting from the calibrated accept statistic to raw confidence is the most damaging—resolution drops and ECE balloons—confirming that calibration, not the cascade plumbing, carries the result. Removing the business-cost weighting mostly hurts the high-error-cost intents, which is where misrouting is most expensive; removing difficulty-aware entry costs a little cost efficiency by wasting first hops on clearly hard threads. Fig. 3 shows the accept decision is near-calibrated for BACC and systematically overconfident for the raw cascade.

TABLE V
Iso-budget ablation under mixed/shifted traffic (simulated).

Variant	Bus.-wtd. resolution (%)	ECE	High-cost-intent error (%)
BACC (full)	84.7	0.05	6.1
– calibration (raw conf.)	77.9	0.26	12.8
– business-cost weighting	82.0	0.06	11.3
– difficulty-aware entry	83.1	0.05	6.4

Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

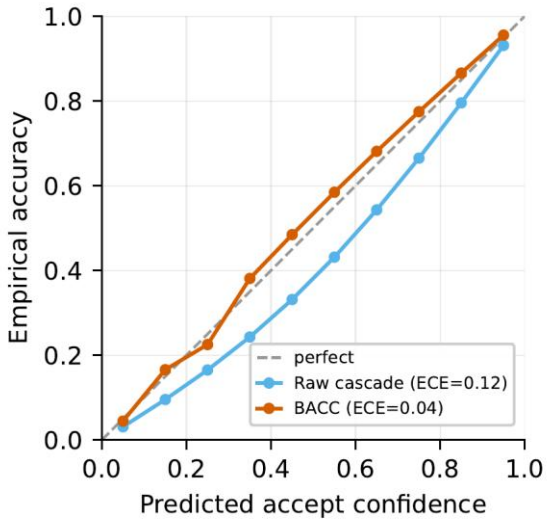

Fig. 3. Reliability diagram of the accept decision (simulated). BACC lies near the diagonal; the raw-confidence cascade is systematically overconfident.

DISCUSSION

The central message is that, for a commerce assistant, the deferral *statistic* governs cost efficiency more than the cascade structure. Raw self-confidence is a weak, drifting signal; a calibrated, business-cost-aware accept rule keeps the same cheap-first plumbing but makes the cost predictable and the high-error-cost turns safe. This matters operationally because a support organization commits to a per-conversation budget and an automation-versus-escalation policy in advance, and can only do so if the router's reported confidence means what it says.

The approach buys this predictability at a price. Per-tier calibration requires a labelled held-out split per intent family and periodic re-fitting as traffic moves, adding a monitoring and data-governance burden. The business-cost weighting requires the organization to quantify the cost of an error per intent—a modelling choice that should be owned by the business, not hidden in a threshold. The handoff gate improves safety on high-cost turns but raises human-escalation cost where the safety floor is conservative; the floor is therefore a deliberate, auditable lever. Routing also composes with single-model efficiency techniques: cheaper frontier inference [13], [14], [15] changes the tier price ratios the testbed uses but not the routing decision.

Threats to Validity and Limitations

Construct (simulation). Every number derives from a seeded generator, not live traffic or live model calls. We chose this because production transcripts carry confidential data, and because simulation gives ground-truth correctness, exact control over the mix and error costs, and a reproducible shift sweep. We mitigate construct risk by setting only mechanism parameters—per-tier skill, price and latency ratios, confidence reliability, and the error-cost model—and letting outcomes emerge; by modelling self-confidence as the weak, overconfident signal it empirically is [10], [7]; and by anchoring price ratios to public 2023 figures [4], [1]. Absolute values should be read as relative comparisons among policies. Instantiating the testbed on a live tiered pool with a learned estimator and human-labelled correctness is the primary line of future work.

Internal. Per-tier accuracy functions, confidence reliabilities, and the error-cost weights are parameters. The ablation and shift study vary the regime to show the qualitative conclusions—calibration drives the frontier; fixed thresholds drift; business-cost weighting protects high-cost intents—are insensitive to their exact values.

External. The taxonomy uses six intent families, three tiers, and three shift scenarios; real catalogs, locales, and regulatory regimes introduce further structure. The framework—calibrated accept statistic plus business-cost weighting under a budget—transfers unchanged, but the parameters would need domain calibration.

Scope. The study does not claim BACC eliminates wrong answers. It limits where confident automated answers are returned on high-error-cost turns, by requiring a calibrated correctness estimate above an auditable, per-intent safety floor before accepting.

Practical Implications

For teams building LLM-backed commerce assistants, four practices follow. First, *route on a calibrated probability, not a raw confidence*: fit and monitor a per-tier recalibration map and read the operating point off it so the budget is predictable. Second, *make the error cost explicit per intent*: a returns-policy or order-status mistake should escalate far more readily than a marginal recommendation, and that asymmetry belongs in the routing objective. Third, *evaluate under traffic shift before launch*: a router that looks optimal on a balanced sample can silently overspend or leak errors during a promotion. Fourth, *gate high-cost turns to humans on an auditable floor*, and log every routing decision for replay in regulated workflows.

CONCLUSION

We framed serving an e-commerce customer assistant as a budget-bounded, business-cost-weighted routing problem over a tiered model pool, and proposed budget-aware calibrated cascading: escalate on a per-tier calibrated probability of correctness scaled by the intent-dependent business cost of an error, under a per-conversation budget and latency SLO. In a reproducible simulation testbed with a six-family commerce query taxonomy, tiered prices anchored to public 2023 figures, and three traffic-shift scenarios, the calibrated, business-cost-aware policy traced a more favourable cost–quality frontier than a FrugalGPT-style raw-confidence cascade and, unlike the fixed-threshold baseline, held its operating point and calibration under shift. The reported numbers are simulated and illustrative; the immediate next step is to instantiate the testbed on a live tiered pool with human-labelled correctness, turning the framework's relative comparisons into measured operating curves for a deployed assistant.

Acknowledgment

The views and opinions expressed in this work are solely those of the author and do not represent, reflect, or imply the views, positions, policies, or opinions of any employer or affiliated organization. The content is offered for informational purposes only. Neither the author's employer nor any affiliated organization makes representations or warranties concerning the accuracy or completeness of the material, and none endorses or accepts responsibility for any statements, conclusions, or content included herein.

REFERENCES

[1] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.
[2] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023.
[3] L. Ouyang *et al.*, "Training Language Models to Follow Instructions with Human Feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[4] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance," arXiv:2305.05176, 2023.
[5] L. Chen, M. Zaharia, and J. Zou, "FrugalML: How to Use ML Prediction APIs More Accurately and Cheaply," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
[6] Madaan *et al.*, "AutoMix: Automatically Mixing Language Models," arXiv:2310.12963, 2023.
[7] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.
[8] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," in *Proc. AAAI Conf. Artif. Intell.*, 2015.
[9] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, 1999.
[10] S. Kadavath *et al.*, "Language Models (Mostly) Know What They Know," arXiv:2207.05221, 2022.
[11] K. Tian *et al.*, "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2023.

[12] Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
[13] W. Kwon *et al.*, "Efficient Memory Management for Large Language Model Serving with PagedAttention," in *Proc. ACM Symp. Oper. Syst. Princ. (SOSP)*, 2023.
[14] Y. Leviathan, M. Kalman, and Y. Matias, "Fast Inference from Transformers via Speculative Decoding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.
[15] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in *NIPS Deep Learning Workshop*, 2015.
[17] Jannach, A. Manzoor, W. Cai, and L. Chen, "A Survey on Conversational Recommender Systems," *ACM Comput. Surv.*, vol. 54, no. 5, 2021.
[18] J. Gao, M. Galley, and L. Li, "Neural Approaches to Conversational AI," *Found. Trends Inf. Retr.*, vol. 13, no. 2–3, 2019.
[19] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "SuperAgent: A Customer Service Chatbot for E-commerce Websites," in *Proc. ACL System Demonstrations*, 2017.
[20] J. McAuley and A. Yang, "Addressing Complex and Subjective Product-Related Queries with Customer Reviews," in *Proc. Int. World Wide Web Conf. (WWW)*, 2016.
[21] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," arXiv:1902.10909, 2019.
[22] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (eds.), *Dataset Shift in Machine Learning*, MIT Press, 2009.
[23] P. W. Koh *et al.*, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
[24] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge," *Trans. Assoc. Comput. Linguist. (TACL)*, vol. 7, 2019.