Managing Front-office and Back-office Effort Allocation: Unpacking the Effects of Workload on Service Performance

May 8, 2026

Abstract

Online customer service requires agents to balance customer-facing communication with less visible problem-solving work. This paper studies this service-effort allocation problem under workload pressure by measuring agents' front ratio, defined as the proportion of active service effort devoted to front-office interaction. Using granular operational data from a large e-commerce platform and a control-function approach to address endogeneity, we examine how workload reshapes this allocation and how the allocation affects service performance. We find that workload has a U-shaped effect on the front ratio: as workload increases from low to moderate levels, agents allocate relatively more effort to back-office problem solving, whereas under high workload they shift proportionally back toward front-office interaction. We further show that the front ratio significantly mediates the relationship between workload and service performance. Specifically, the front ratio has a U-shaped relationship with session duration and customer retrials, and an inverted-U-shaped relationship with customer ratings, indicating that service performance is optimized at an intermediate level of front-office effort. Textual analyses further explain this pattern. Guided by the service quality distinction between outcome quality and interaction quality, we show that moderate front-office effort improves both problem-centered communication and interaction quality, whereas excessive front-office emphasis crowds out substantive problem solving and eventually weakens communication quality. A counterfactual analyses suggest that increasing the front ratio by 0.53 standard deviations from the sample mean of 0.47 to 0.57 can reduce session duration by 9.81%, increase customer ratings by 10.73%, and reduce retrial rates by 13.04%. These findings extend workload and service operations research by identifying within-session effort allocation as a behavioral mechanism linking workload to efficiency, customer satisfaction, and durable resolution.

Keywords: *Online Customer Service; Workload; Effort Allocation; Front-Office/Back-Office; Behavioral Operations; Service Quality*

1 Introduction

Online customer service has become an increasingly important setting in service operations, particularly as digital platforms rely on live-chat systems to deliver timely and scalable customer support. In this setting, service agents do not merely "talk" to customers. They must also diagnose customer problems, retrieve

relevant information, coordinate with internal resources, and move cases toward substantive resolution. These activities create a fundamental service-effort allocation problem under workload pressure. On the one hand, agents need to devote effort to customer-facing communication, including clarification, explanation, responsiveness, and emotional management. On the other hand, they must allocate effort to less visible problem-solving activities that determine whether the customer's issue is actually resolved. This challenge corresponds to a classic front-office/back-office framework in service operations (Chase, 1978; Li and Zhang, 2000; Teboul, 2006). Whereas traditional theory often emphasizes decoupling these activities to improve productivity, online service agents must manage both types of work in real time. This raises a central managerial question: how does agents' within-session effort allocation affect service performance? This foundational challenge is further amplified in online customer service environments where multitasking across concurrent sessions is the norm (Kc, 2014), forcing agents to continuously navigate a complex allocation problem. Understanding this relationship is important not only for training human agents, but also for redesigning service strategies and developing AI tools that support the appropriate balance between customer-facing communication and substantive problem solving.

A significant body of operations research has examined how workload affects service performance. Prior studies show that workload can influence service speed, quality, multitasking, operational risk, and downstream customer behavior across healthcare, restaurants, banking, and customer contact centers (Kc and Terwiesch, 2009; Tan and Netessine, 2014; Goes et al., 2017; Xu et al., 2022). More recent work further shows that workers adapt to workload through specific behavioral responses, such as rushing, reduced diligence, task initiation, task selection, and changes in care intensity (Powell et al., 2012; Batt and Terwiesch, 2017; Ibanez et al., 2018; Kc et al., 2020; Soltani et al., 2022). However, this literature has not fully examined how workload changes the composition of effort within a service encounter. In online customer service, agents may respond to workload not only by working faster or slower, but also by reallocating effort between customer-facing communication and less visible problem-solving work. This within-session allocation decision remains underexplored, even though it may be the behavioral mechanism through which

workload affects session duration, customer ratings, and customer retrials.

In this paper, we open this behavioral black box by proposing a service-effort allocation framework in which workload affects service performance through agents' within-session allocation between front-office interaction and back-office problem solving. We operationalize this allocation using the front ratio, defined as the proportion of an agent's active service effort devoted to customer-facing interaction within a session. A higher front ratio indicates that the agent allocates relatively more effort to communication, clarification, and customer management, whereas a lower front ratio indicates relatively more effort devoted to diagnosis, information retrieval, coordination, and substantive resolution. This construct allows us to examine how workload changes the structure of service work, rather than only its speed or volume. Our conceptual model (Figure 1) guides this investigation, which we structure around three central research questions: First, how does workload affect agents' front ratio? Second, how does the front ratio mediate the relationship between workload and service performance? Third, what mechanisms explain why different levels of front-office effort lead to different service outcomes?

To answer these questions, we use granular operational data from a large Chinese e-commerce platform that provides customer support through live-chat service agents. Our data contain 4,683,611 session-level observations handled by 7,004 agents from October 25, 2023, to November 30, 2023. The data include timestamped service records, agent click-stream activities, workload measures, and multiple service outcomes, allowing us to observe both how agents allocate effort within each session and how each session performs. This setting is particularly useful for our research question because online service agents must repeatedly switch between customer-facing communication and background problem-solving tasks while handling concurrent customer sessions. To address potential endogeneity in workload and effort allocation, we adopt a control-function approach with instrumental variables.

Our analysis first shows that workload has a U-shaped relationship with the front ratio. As workload increases from low to moderate levels, agents allocate relatively more effort to back-office problem solving, causing the front ratio to decline. This pattern suggests that moderate workload may mobilize agents to

focus on diagnosis, information retrieval, and substantive resolution in order to move cases forward and avoid future rework. However, once workload becomes high, the relationship reverses: agents reduce back-office effort more sharply and protect visible customer-facing interaction, causing the front ratio to rise. This finding indicates that workload affects not only how much effort agents exert or how quickly they work, but also how they structure effort across different types of service activities within the session.

Second, we find that the front ratio significantly mediates the relationship between workload and service performance. Specifically, the front ratio has a U-shaped relationship with session duration, an inverted-U-shaped relationship with customer ratings, and a U-shaped relationship with customer retrials. These results indicate that neither too little nor too much front-office effort is optimal. When the front ratio is too low, insufficient customer-facing communication may create ambiguity, repeated clarification, and weak customer engagement. As the front ratio increases to a moderate level, agents can communicate more effectively with customers while still preserving sufficient back-office problem-solving effort, thereby improving efficiency, satisfaction, and resolution. However, when the front ratio becomes too high, excessive customer-facing effort may crowd out diagnosis, information retrieval, and substantive problem solving, increasing session duration, lowering ratings, and raising the probability of retrial. The mediation pathway is economically meaningful: the front-ratio mechanism accounts for 20.95% of the total absolute impact of workload on session duration, 36.07% on customer ratings, and 68.59% on retrial rate. Our counterfactual analysis further shows that increasing the front ratio by 0.53 standard deviations from the sample mean of 0.47 to 0.57 can reduce service duration by 9.81%, increase customer ratings by 10.73%, and reduce retrial rates by 13.04%.

To further understand why service performance is optimized at an intermediate front ratio, we use textual analysis to examine the content and quality of agent-customer conversations. The theoretical motivation comes from service quality research, which distinguishes between outcome quality and interaction quality (Grönroos, 1984; Brady and Cronin Jr, 2001). In our setting, outcome quality depends on whether the conversation helps diagnose the customer's problem and move the issue toward resolution, whereas interaction quality depends on whether the agent communicates clearly, manages the customer's emotions,

and maintains an effective service encounter. Guided by this distinction, we construct two sets of textual measures. First, using Latent Dirichlet Allocation (LDA), we measure problem-centered communication, which captures the extent to which a session focuses on diagnosis, issue clarification, and resolution-related content. This measure reflects the outcome-quality dimension of the conversation. Second, using a large language model trained on company evaluation data, we measure interaction quality through emotional management, communication, and overall approval, which reflect the customer-facing process of service delivery. The textual results show that a moderate front ratio improves both problem-centered communication and interaction quality. However, when the front ratio becomes too high, problem-centeredness declines and interaction quality eventually weakens, suggesting that additional customer-facing communication is no longer sufficiently supported by substantive problem-solving progress. These findings explain why excessive front-office emphasis can increase session duration, reduce customer ratings, and raise retrial rates despite appearing attentive during the focal interaction.

Based on the causal relationship between workload, effort allocation, and service performance, our findings identify the front-office/back-office balance as an actionable managerial lever. The results suggest that service platforms should not simply encourage agents to maximize customer-facing communication. Instead, managers should help agents maintain an intermediate front ratio that combines sufficient customer interaction with substantive problem-solving effort. This has implications for agent training, performance evaluation, and service process design. Training programs should help agents recognize when communication improves coordination and when it begins to substitute for problem solving. Performance metrics should balance immediate customer ratings with longer-term resolution outcomes such as retrials. In addition, AI tools for customer service should be designed not only to generate empathetic or responsive messages, but also to support diagnosis, information retrieval, and resolution-oriented communication. Thus, improving service performance requires managing the composition of agent effort, rather than merely increasing response speed or communication volume.

Our study makes three contributions to the literature. First, we contribute to the workload and behavioral

operations literature by identifying within-session effort allocation as a mechanism through which workload affects service performance. Prior research has shown that workload influences service speed, quality, multitasking, and operational risk; we extend this literature by showing that workload also changes how agents allocate effort between customer-facing interaction and problem-solving work. Second, we contribute to the service operations literature on front-office/back-office design. Whereas prior work often treats the front-office/back-office distinction as a structural design choice, we show that this boundary is also dynamically managed by agents within individual service sessions. Third, we contribute to the service quality literature by jointly examining efficiency, immediate customer satisfaction, and durable resolution. Our findings show that an intermediate front ratio can simultaneously reduce session duration, improve customer ratings, and reduce retrials, highlighting that service quality depends on balancing interaction quality with outcome quality rather than maximizing customer-facing effort alone.

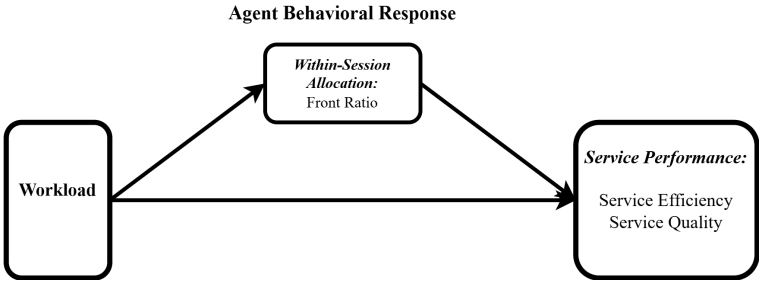


Figure 1: The conceptual framework

2 Related Literature

Our research contributes to three streams of literature: (i) customer contact operations, (ii) workload and behavioral adaptation in service operations, and (iii) front-office/back-office effort allocation and service quality.

The first stream of literature relevant to our work concerns customer contact operations. Prior work in this stream has studied how service systems should be designed and managed through staffing, routing, waiting-time control, and capacity allocation decisions (Gans et al., 2003; Aksin et al., 2007). More recent

work has extended this literature to online chat and digital contact centers, where agents often serve multiple customers simultaneously and where service time is shaped by both agent and customer behavior. Luo and Zhang (2013), Tezcan and Zhang (2014), and Long et al. (2024) examine staffing and routing in customer service chat systems. Goes et al. (2017) show that multitasking in online customer service can lead to longer delays, lower resolution rates, and lower customer satisfaction. Altman et al. (2020) study emotional load in online customer contact centers and show that negative customer emotions affect agent response delays, message length, and the number of messages required to complete service. On the customer side, Yu et al. (2017) show that delay announcements shape customer beliefs and waiting behavior, Hathaway et al. (2022) use individual-level customer history to predict abandonment and redialing behavior, and Ilk and Shang (2022) show that customer waiting affects customer response behavior during live-chat service. In general, this literature has greatly advanced our understanding of service-center operations and customer-agent interaction. However, limited work has examined how agents allocate effort within a service session between customer-facing interaction and background problem solving. Our study focuses on this within-session allocation decision and examines how it mediates the relationship between workload and service performance.

The second stream of relevant work studies workload and behavioral adaptation in service operations. A growing body of empirical operations research shows that workload can have complex and nonlinear effects on worker behavior and operational performance. Kc and Terwiesch (2009) show that hospital workers speed up under higher workload, with implications for patient safety. Tan and Netessine (2014) find an inverted-U-shaped relationship between workload and restaurant waitstaff's service speed and sales output. Kc (2014) shows that multitasking can initially improve performance but eventually harm quality, including revisit-related outcomes. Powell et al. (2012) find that high workload reduces physician diligence in paperwork. Batt and Terwiesch (2017) identify rushing and early task initiation as workload-related behavioral responses. Soltani et al. (2022) show that physician workload affects downstream care use through care intensity. Xu et al. (2022) find a U-shaped relationship between workload and operational risk, showing

that low and high workload can generate different types of failures. Together, these studies demonstrate that workload affects performance not only through congestion but also through behavioral adaptation. Nevertheless, prior work usually focuses on direct workload-performance effects or on specific behavioral responses such as speeding up, rushing, task initiation, multitasking, or reduced diligence. Less is known about how workload changes the composition of effort within a service encounter. We contribute to this literature by identifying front-office/back-office allocation as a behavioral mechanism through which workload affects service performance.

The third stream concerns front-office/back-office effort allocation and its consequences for service performance. The classic service operations literature emphasizes that customer-contact work and production-oriented work impose different operational requirements (Chase, 1978; Chase and Tansik, 1983; Safizadeh et al., 2003; Teboul, 2006; Zomerdijk and de Vries, 2007). Front-office work involves direct customer interaction, including clarification, explanation, responsiveness, and expectation management. Back-office work involves less visible production activities, including diagnosis, information retrieval, coordination, and substantive problem solving. Traditional service design often treats these two activities as structurally separable. In online customer service, however, this boundary is fluid because the same agent must repeatedly switch between customer-facing interaction and background problem-solving tasks during the same service session. Research on time allocation and discretionary task completion suggests that workers allocate scarce time and attention across competing activities and may adjust task depth, task sequence, and completion quality under pressure (Becker, 1965; Hopp et al., 2007, 2009; Coviello et al., 2014; Ibanez et al., 2018; Kc et al., 2020). In service systems, Roels (2014) highlights the co-productive nature of service delivery, and Legros et al. (2020) directly study front-office multitasking between service encounters and back-office tasks. Building on this literature, we conceptualize the front ratio as a session-level measure of how agents manage the boundary between visible customer-facing interaction and less visible problem-solving work.

This allocation decision matters because service performance depends on both the process of interaction and the outcome of resolution. Service quality research provides a useful framework for this distinction.

Grönroos (1984) distinguishes technical quality from functional quality, and Brady and Cronin Jr (2001) further conceptualize service quality as involving outcome quality, interaction quality, and the service environment. In our setting, back-office effort is closely related to outcome quality because it supports diagnosis, information retrieval, and durable problem resolution. Front-office effort is closely related to interaction quality because it supports clarification, responsiveness, emotional management, and perceived attentiveness. This distinction maps directly to our empirical outcomes. Session duration reflects the efficiency of the co-produced service process: insufficient front-office interaction may create ambiguity and repeated clarification, whereas excessive front-office interaction may delay substantive resolution by crowding out back-office work. Customer ratings capture immediate subjective evaluations of the service encounter and are therefore sensitive to interaction quality, including responsiveness, clarity, and perceived effort (Bitner et al., 1990; Tax et al., 1998; Maxham III and Netemeyer, 2002). Customer retrials capture a more durable quality outcome, because repeat contact indicates that the prior interaction did not provide sufficient resolution or closure (Cui et al., 2019; Hu et al., 2022). Therefore, ratings and retrials are related but distinct quality measures: ratings reflect how the service was experienced in the moment, whereas retrials reflect whether the issue was durably resolved.

In sum, prior research establishes that customer contact centers are behaviorally rich service systems, that workload changes worker behavior, and that service quality depends on both interaction and outcome dimensions. However, existing studies have not fully connected these ideas into a unified framework that explains how workload affects agents' within-session allocation between front-office and back-office effort, and how this allocation shapes efficiency, immediate satisfaction, and durable resolution. Our study addresses this gap by measuring the front ratio at the session level and examining it as an endogenous behavioral mechanism linking workload to service performance.

3 Hypothesis Development

Combining the online service context with the literature on workload, effort allocation, and service quality, we develop hypotheses for two linked relationships. First, we examine how workload affects agents' within-session allocation between front-office interaction and back-office problem solving. Second, we examine how this allocation affects service performance. Our central argument is that workload does not affect service performance only through congestion or average service speed. Rather, workload changes how agents allocate scarce time and attention across different types of service work, and this endogenous allocation shapes efficiency, immediate customer evaluation, and durable resolution.

3.1 The Effect of Workload on Agent Within-Session Allocation

In online customer service, agents must perform both front-office and back-office activities within the same session. Front-office work includes direct customer communication, clarification, emotional management, and expectation alignment. Back-office work includes diagnosis, information retrieval, internal coordination, and substantive problem solving. The classic front-office/back-office framework emphasizes that these two types of work impose different operational demands on service systems (Chase, 1978; Chase and Tansik, 1983; Safizadeh et al., 2003; Zomerdijk and de Vries, 2007). In our setting, however, the boundary between them is not fixed by organizational design. Instead, the same agent continuously allocates effort between customer-facing interaction and background problem solving during the service session. We capture this allocation using the front ratio, defined as the proportion of active service effort allocated to front-office interaction.

Workload may affect this front ratio through two opposing mechanisms. The first mechanism is a problem-solving mobilization effect. Research on attention and effort suggests that moderate workload can increase arousal, focus, and active engagement, while very low workload may leave workers less engaged (Kahneman, 1973; Hockey, 1997; Speier et al., 1999; Wickens, 2002). Operations studies similarly show that workers often respond to rising workload by speeding up, reallocating effort, or becoming more focused

on task completion (Kc and Terwiesch, 2009; Tan and Netessine, 2014; Berry Jaeker and Tucker, 2017; Xu et al., 2022). In our setting, when workload increases from a low level, agents may become more motivated to move cases toward resolution and avoid downstream congestion. Because unresolved cases can create repeated work and future service burden, agents may allocate relatively more effort to back-office problem solving, such as diagnosis and information retrieval. This mechanism predicts that workload initially reduces the front ratio.

The second mechanism is a capacity-compression effect. When workload becomes high, agents face binding limits on time and attention. Theories of time allocation and discretionary task completion suggest that workers under pressure adjust not only how much effort they exert, but also which tasks they prioritize and how deeply they complete them (Becker, 1965; Hopp et al., 2007, 2009; Coviello et al., 2014). Empirical research shows that workers under workload may reduce diligence, rush tasks, alter task initiation, or select tasks differently (Powell et al., 2012; Batt and Terwiesch, 2017; Ibanez et al., 2018; Kc et al., 2020; Soltani et al., 2022). In our setting, back-office work is less visible to customers and often requires concentrated cognitive effort, whereas front-office responsiveness is immediately observable and helps manage customer expectations. Under heavy workload, agents may therefore protect visible customer-facing interaction while compressing less visible back-office problem solving. This mechanism predicts that workload increases the front ratio when workload is already high.

We argue that the problem-solving mobilization effect dominates when workload is low to moderate, whereas the capacity-compression effect dominates when workload is high. Thus, as workload increases from a low level, agents allocate relatively more effort to back-office problem solving and the front ratio decreases. Once workload becomes sufficiently high, back-office work is compressed more sharply than front-office interaction, causing the front ratio to rise.

Hypothesis 1 (H1): *The relationship between workload and an agent's front ratio is U-shaped.*

3.2 The Effect of Within-Session Allocation on Service Performance

Having hypothesized how workload affects the front ratio, we next examine how the front ratio affects service performance. Service performance in our setting has three dimensions: session duration, customer rating, and customer retrial. These outcomes capture different aspects of service delivery. Session duration reflects the efficiency of the service process. Customer rating captures the customer's immediate evaluation of the service encounter. Customer retrial captures whether the issue was durably resolved or whether the customer returned with the same or related problem. We argue that the front ratio affects these outcomes through two opposing forces: a coordination-enhancing effect and a problem-solving crowd-out effect.

3.2.1 Front Ratio and Session Duration

The coordination-enhancing effect suggests that increasing front-office effort from a low level can reduce session duration. Service encounters are dyadic interactions in which agents and customers exchange information and adjust their behavior during the process (Solomon et al., 1985). Co-productive service theory similarly emphasizes that service outcomes depend on the allocation of tasks and information between the provider and the customer (Roels, 2014). In online service, customers often need to describe the issue, provide missing information, understand the agent's instructions, and accept the next steps. If front-office effort is too low, agents may provide limited clarification, weak guidance, or overly generic responses. This can create ambiguity, repeated clarification, and rework. Communication theory suggests that richer and more adaptive communication helps reduce equivocality and establish common ground (Daft and Lengel, 1986; Weitz et al., 1986; Bitner et al., 1990; Clark and Brennan, 1991). Therefore, increasing the front ratio from a low level can improve coordination between the agent and the customer and shorten the session.

However, the problem-solving crowd-out effect suggests that excessive front-office effort can increase session duration. Agents have limited time and attention. When more effort is devoted to customer-facing communication, less effort is available for diagnosis, information retrieval, and internal coordination (Ocasio, 1997; Monsell, 2003; Leroy, 2009). Prior work on customer contact and front-office/back-office mul-

titasking highlights this tension between interaction and production work in service systems (Chase, 1978; Roels, 2014; Legros et al., 2020). When the front ratio becomes too high, the interaction may remain active, but it may not advance efficiently toward resolution. Thus, the session can become longer because communication is no longer sufficiently supported by substantive problem solving.

We argue that the coordination-enhancing effect dominates when the front ratio is low, whereas the problem-solving crowd-out effect dominates when the front ratio is high. Therefore, session duration should first decrease and then increase as the front ratio rises.

Hypothesis 2 (H2): *The relationship between front ratio and session duration is U-shaped.*

3.2.2 Front Ratio and Service Quality

We next consider how the front ratio affects service quality. Service quality research distinguishes between what customers receive and how the service is delivered. Grönroos (1984) separates technical quality from functional quality, and Brady and Cronin Jr (2001) conceptualize service quality as involving outcome quality, interaction quality, and the service environment. This distinction maps closely to our setting. Back-office effort supports outcome quality because it enables diagnosis, information retrieval, and substantive resolution. Front-office effort supports interaction quality because it enables responsiveness, clarification, emotional management, and perceived attentiveness. Effective service therefore requires both interaction quality and outcome quality.

Customer rating captures the customer's immediate evaluation of the service encounter. When the front ratio is low, customers may perceive the agent as inattentive, unclear, or insufficiently supportive, even if the agent is working on the problem in the background. Service encounter research shows that employee responsiveness, explanations, and interpersonal treatment strongly shape customer evaluations (Bitner et al., 1990; Tax et al., 1998; Maxham III and Netemeyer, 2002). Visible effort can also increase perceived value when customers believe that the effort reflects real service work (Buell and Norton, 2011; Buell et al., 2017). Thus, increasing front-office effort from a low level should improve ratings by strengthening interaction

quality.

However, this benefit should not increase without limit. When the front ratio becomes too high, communication may no longer be supported by sufficient substantive progress. The too-much-of-a-good-thing perspective suggests that beneficial behaviors can become harmful beyond an optimal point when they displace complementary activities (Pierce and Aguinis, 2013). In our setting, excessive front-office emphasis may reduce the back-office problem solving needed to make communication credible. As a result, customers may eventually evaluate the service less favorably because the interaction appears responsive but does not sufficiently move the issue toward resolution. Therefore, customer rating should be maximized at an intermediate level of front-office effort.

Hypothesis 3 (H3): *The relationship between front ratio and customer rating is inverted U-shaped.*

Customer retrial captures a more durable form of service quality. Unlike customer rating, which reflects the customer's immediate subjective evaluation, retrial indicates whether the focal interaction provided enough resolution and closure to prevent the customer from returning with the same or related issue. Prior research on call centers and service systems treats retrial and repeat contact as important indicators of unresolved service needs and downstream system burden (Oliva and Sterman, 2001; Cui et al., 2019; Hu et al., 2022).

When the front ratio is low, insufficient customer-facing communication may leave the issue poorly explained, the next steps unclear, or the customer unconvinced that the problem has been resolved. This lack of shared understanding can increase uncertainty and prompt repeat contact. Low front-office effort may also weaken emotional support and reduce the customer's willingness to accept the current handling of the case (Tax et al., 1998; Maxham III and Netemeyer, 2002). Thus, increasing the front ratio from a low level should reduce retrials by improving explanation, expectation alignment, and customer closure.

However, when the front ratio becomes too high, retrials may rise again because excessive front-office effort crowds out the back-office work required for durable resolution. Under limited attention, greater effort devoted to explanation and interaction maintenance necessarily reduces the attention available for diagnosis, information retrieval, and substantive problem solving (Ocasio, 1997). In this range, the interaction

may appear active and responsive, yet still fail to resolve the customer's underlying issue. As a result, customers may contact the firm again not because communication was absent, but because the interaction was insufficiently resolution-oriented. Therefore, retrial should be minimized at an intermediate front ratio.

Hypothesis 4 (H4): *The relationship between the front ratio and the customer retrial rate is U-shaped.*

4 Empirical Setting and Data

4.1 Empirical Setting

We collaborate with the after-sales service department of a major Chinese online retailer. This department specializes in answering customer inquiries and resolving seller disputes, which include return and refund, account information, membership, shipping, and complaints. The retailer receives roughly 1 million daily customer support requests. To manage this volume cost-effectively, the department employs a two-tier service model in its contact center: the Automated Tier, where initial inquiries are routed to a chatbot that resolves routine issues using a predefined knowledge base, and the Human-Agent Tier, where unresolved cases (approximately 300,000 customers daily) are escalated to human agents.

The department employs a blended workforce strategy, integrating gig workers to supplement its traditional, full-time agents. This approach is adopted because blending the workforce can lead to significant cost reductions and a more balanced service level (Dong and Ibrahim, 2020). The full-timers are salaried workers. They work on site and are given higher authority to handle requests that the gig agents cannot handle.

The gig agents work remotely and communicate with customers via live-chat. Before working, they undergo a one-month online training program covering all service types. Their compensation follows a piece-rate structure, consisting of a payment per session and a performance bonus tied to high customer ratings. On average, agents complete ten sessions per one-hour shift, yielding approximately 21 RMB (equivalent to approximately three USD) in earnings. In addition, these agents are categorized into five skill levels, with level one being new hires, and levels two to five based on prior months' performance in ascending skill

levels. Higher-level agents are assigned more customer cases simultaneously (i.e., higher multi-tasking capacity). Specifically, Level one agents handle one session, Level two agents up to two sessions, Levels three and four agents up to three sessions, and Level five agents up to four sessions concurrently. This increased capacity enables higher-level agents to serve more customers and thereby potentially increase their earnings.

We focus on gig workers in this study because the front-office versus back-office effort allocation is likely to be particularly salient in this population. Gig agents are compensated on a short-term, piece-rate basis, making their effort allocation decisions more responsive to workload conditions. This responsiveness not only provides clearer empirical variation to examine our proposed mechanism, but also makes the findings directly relevant to the growing share of fissured, on-demand labor in service operations (Weil, 2014; Benjaafar and Hu, 2020).

Agents use a proprietary tool to provide service. Figure 2 shows a simplified version of an agent interface. This tool offers three main functions: 1) displaying the current sessions the agent is managing, 2) communicating with customers, and 3) integrating various support tools (e.g., knowledge repository, customer history tracking).

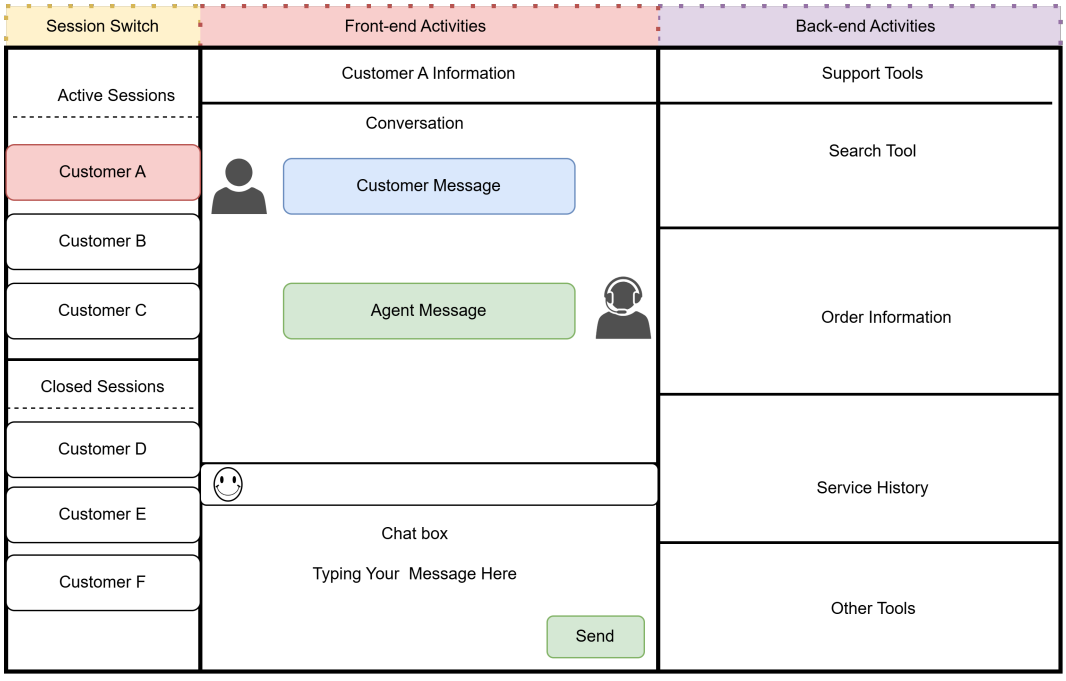


Figure 2: Illustration of Service Tool Interface

The service process of a typical live-chat session is shown in Figure 3. Following the framework established by Dubé-Rioux and Schmitt (1989), we divide the service process into three stages from the customers' perspective: pre-service, main service, and close-down stages. During the pre-service stage, a utilization-based routing algorithm assigns customers to agents with the lowest utilization (defined as the ratio of active customers to an agent's aforementioned maximum multitasking capacity). This utilization-based routing algorithm is unchanged during our observation period.

During the main service stage, agents aim to comprehend and resolve the customer's issues. Although Figure 3 depicts one message per party (either agent or customer) at a time, multiple messages may be sent by the same party during one interaction round. We define an interaction round as a consecutive sequence of messages initiated by one party (either the agent or the customer) and followed by a response from the other party. When agents manage multiple sessions simultaneously (i.e., multitasking), they must switch between sessions to provide service, as noted as "switch duration" in Figure 3.

Finally, the close-down stage includes the period after the last message is sent by either the customer or agent until the session is terminated. In practice, this stage serves as a buffer time to detect customers' silent abandonment in service so that agents can close abandoned session and release service capacity (Castellanos et al., 2025).

After completing the service, customers are invited to rate their experience on a scale from one to five, with higher scores indicating greater satisfaction. This rating serves as our measure of short-term service quality because it reflects customers' instantaneous satisfaction. However, the rating response rate is approximately 16.40%, which raises concerns about potential selection bias in this metric. In other words, probably only the most satisfied and most unsatisfied customers may rate their experience (Hu et al., 2017; Schoenmueller et al., 2020). To address this potential bias, we employ a Heckman selection model to address these factors as a robustness check in Appendix E. Furthermore, some customers (35% of the customers in our sample) revisit the system to resolve the same issue within a week. This behavior is referred to as *retrials* in the literature (Hu et al., 2022). Retrials not only increase the platform's operational costs but

also undermine the customer's service experience in the long run. We therefore consider this behavior as a long-term service quality measure.

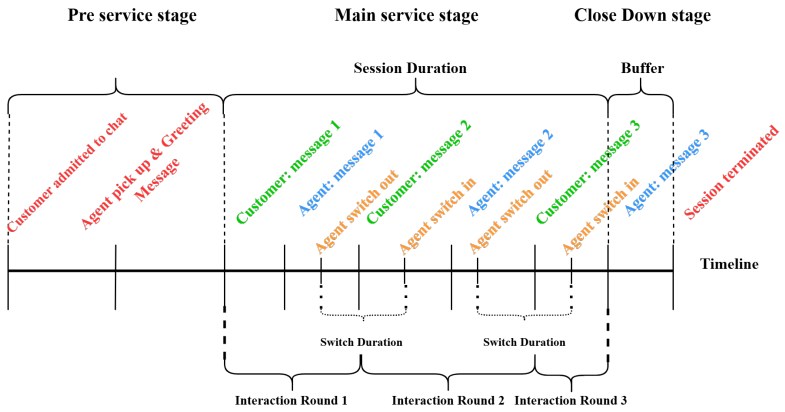


Figure 3: Illustration of Service Process

4.2 Data

Our data span from October 25, 2023 to November 30, 2023. A critical feature of our empirical setting is that the agents in our sample belong to a specialized team dedicated exclusively to serving VIP customers. By restricting our analysis to this homogeneous customer tier, we effectively hold customer status constant by design. This approach isolates the effects of workload and agent behavior, eliminating potential confounding variation that could arise from unobserved differences in service protocols, routing priorities, or resolution policies between VIP and non-VIP segments.

Our data include five parts: (1) session characteristics, (2) agent activity logs, (3) chat transcripts, (4) agent demographics, and (5) customer demographics. More specifically, session characteristics contain data such as agent and customer IDs, session ID, customer issue type, customer ratings, and customer retrials (the system codes 1 if the customer revisits the system within 7 days for the same request). Agent activity logs capture timestamped agents' activities, including switching between concurrent sessions, typing responses, sending messages, searching Standard Operating Procedures (SOPs), and reviewing customer service histories (see Appendix B for examples). We consider those activities involving real-time interacting with the customers as the front-office activities (Teboul, 2006), and other activities like researching the topic, reviewing histories, seeking solutions, and documentation as the back-office activities. The granularity of

the timestamped data, a key feature of our study, enables us to observe the time or effort an agent allocates to the front-office and back-office activities, respectively. Chat transcripts record the full-text chat history between the agent and the customer. Such data allow us to conduct textual analysis to delineate our proposed mechanisms in Section 6. Agent demographics capture agents' gender, age, tenure and previous service performance, while customer demographics reflect customers' service history, rating rate, and tenure on the platform.

After consolidating these five parts, our full dataset contains 4,683,611 session observations from 7,004 agents. We removed the sessions with fewer than three interaction rounds (approximately 15% of the data), because such chats provide insufficient conversational data to estimate sentiment effects reliably, and they are disproportionately comprised of abandonments, mis-routes, or single-question transactions that fall outside our focal construct of substantive service interactions. Nevertheless, we still used the full unfiltered data to conduct robustness checks, which showed consistent results (Section 7.2). Finally, we winsorized the data at the 1st and 99th percentiles to mitigate the effect of outliers. Our final data set consists of 3,965,028 sessions and 7,004 agents.

We believe our data set is valuable to study our research questions for the following reasons. Compared to data used in previous empirical studies on workload, the granularity of our data set enables us to unpack the impact of workload on agents' effort allocation decisions and further investigate the underlying mechanisms linking workload to service performance. The availability of chat transcripts also allows us to apply textual analysis, offering deeper insights into how effort allocation decisions manifest in service patterns and providing practical managerial implications. Lastly, the large sample size ensures sufficient statistical power for our analyses.

4.3 Measures and Key Variables

4.3.1 Service Performance Measures

We examine three key service performance outcomes which reflect service efficiency, short-term and long-term service qualities, respectively.

1. We define a service efficiency measure, $SessionDuration_{ij}$, as the total time in seconds from agent j's first intervention in a session i until this session is terminated.

2. Variable $CustomerRating_{ij}$ is the customer rating of session i handled by agent j. It is rated on a scale between 1 (lowest) and 5 (highest). This metric captures the customer's immediate, subjective evaluation of the service experience.

3. Variable $IfRetrial_{ij}$, is coded as 1 if the customer returns to the contact center for the same request within a week after session i handled by agent j. To ensure the robustness of this measure and avoid false positives (e.g., a customer calling back for a different order), we construct the variable using a rigorous, dual-source validation process. First, we utilize the platform's automated system to link sessions associated with the same unique Order ID. Second, we cross-reference this with agent-reported data, where agents manually categorize the specific request type after the session. A subsequent session is classified as a retrial only if it matches the original session on both the Order ID and the specific request category. This multi-source verification ensures that our measure captures genuine unresolved issues rather than unrelated subsequent interactions. We consider this retrial behavior as the long-term quality measure.

4.3.2 Independent and Mediating Variables

1. We operationalize our independent variable, $Workload_{ij}$ as the time-averaged number of sessions that agent j handles simultaneously together with the focal session i (Tan and Netessine, 2014). We also include the quadratic term to test our hypothesis about the non-linear effect of workload. For ease of interpretation purposes, both the linear and quadratic workload terms are standardized by subtracting their respective means and then dividing by their standard deviations.

2. We construct our mediating variable $FrontRatio_{ij}$ as the ratio between the time spent on front-office activities and the sum of time spent on front- and back-office activities by agent j during session i. In particular, we use agent's click-stream data and categorize all in-session activities, excluding the

switch actions, as either front-office or back-office. Following the established distinction between front-office and back-office tasks (Teboul, 2006), we classify all agent activities involving real-time customer communication as front-office. These include sending messages, delivering Standard Operating Procedures (SOPs), and typing responses, as the latter is visible to the customer. Conversely, all other in-session activities, are classified as back-office, which includes tasks such as reviewing customer history, searching the knowledge base, and documenting the interaction. Variable $FrontRatio_{ij}$ reflects the agent's relative effort allocation between the front- and back-office activities. Similar to *Workload*, we standardize this variable in our analysis for consistent interpretations. Note that the sum of time spent on the front- and back-office activities is not necessarily equal to the total session time (i.e., $SessionDuration$) because an agent may swap out during the focal session to work on other sessions simultaneously. We use time as a proxy for effort provision because longer engagement typically reflects greater cognitive or motivational investment, especially in self-paced or complex tasks (Stoeber et al., 2010; Alós-Ferrer and Buckenmaier, 2021). In other words, more challenging or effortful tasks generally require more time to complete. Admittedly, invested time is not the only proxy for effort provision. We use the count of corresponding front-office or back-office actions as an alternative effort measure as a robustness check and find similar results in Section 7.3.

4.3.3 Control Variables

1. We control for $Fatigue_{ij}$, measured as the cumulative number of one-hour shifts the agent has worked on a given day up to the focal session to account for potential fatigue effects (Bavafa and Jónasson, 2024).

2. We also account for customer-driven factors that influence an agent's behavior (Goes et al., 2017). In particular, we include variable $AvgCustomerMessageLength_{ij}$, the average Chinese character count of the customer's messages because longer messages should require an agent to allocate more front-office effort to craft a comprehensive response. In addition, we adjust for the average time between an agent

completes a message and the customer's response in seconds (i.e., the $AvgCustomerResponseDelay_{ij}$). A customer's slow response may prompt an agent to switch to other sessions or otherwise reallocate more effort to back-office activities. Crucially, we also include $IsFirst_{ij}$, a dummy variable indicating if the session is the first contact for a request. This controls for the distinct service dynamics and customer expectations associated with initial versus follow-up interactions.

3. We finally include a set of fixed effects to account for unobserved heterogeneity. In particular, agent fixed effects account for agents' idiosyncratic tendencies in allocation preference. Problem-type fixed effects control for heterogeneity across nine distinct problem categories. These categories, which are predefined by the platform and identified by the routing algorithm using chatbot data, include: purchasing, promotions, after-sales, account management, membership, shipping, complaints, logistics, and risk-related issues. Lastly, date-hour fixed effects adjust for temporal variations, such as fluctuating traffic volume or other time-specific unobserved factors.

Table 1 presents the descriptive statistics for the key variables. On average, an agent handles a 2.45 concurrent sessions, with a standard deviation of 0.94, indicating substantial variation in the workload pressure agents face. The average FrontRatio is 0.47, suggesting that agents, on average, dedicate slightly less than half of their active time to direct front-office interaction, with the remainder spent on back-office work. In fact, the average front-office time spent is 146.96 seconds, while the average back-office time spent is 170.3 seconds per session. Regarding service performance, the average *SessionDuration* is approximately 516.59 seconds (8.6 minutes) and we adopt the commonly used approach of taking the log transformation for this measure in our regression analysis because of the large standard deviation (429.70) (Kc and Terwiesch, 2009; Staats and Gino, 2012). While the average *CustomerRating* is 3.05 on a five-point scale, our long-term quality measure shows that 35% of all sessions result in a customer *Retrial* within one week. This high failure rate aligns with our sample composition, where 74% of interactions are initial inquiries and the remaining 26% are repeat contacts. To put this figure in context, industry benchmarks for First Call Resolution (FCR)—the inverse of the retrial rate—consider a rate of 70-79% to be strong performance (SQM

Group, 2025). Our observed 35% retrial rate (equivalent to a 65% FCR) is therefore substantially below this industry standard, highlighting the significant practical and financial importance of understanding the behavioral drivers of effective first-contact resolution. Appendix F.1 plots the distributions of Workload and FrontRatio. The figures show substantial variation in both workload pressure and agents' within-session effort allocation, which provides useful empirical support for our analysis of how workload reshapes the front-office/back-office balance.

Table 1: Definition and Descriptive Statistics (Session Granularity)

Variable	Description	Min	Max	Mean	SD
SessionDuration	Total service time (in seconds) for focal session.	59.00	2236.00	516.59	429.70
CustomerRating	The score (1-5) that customer rated for focal session.	1.00	5.00	3.05	1.87
$IfRetrial$	A binary variable indicating whether the customer revisits within one week after focal session for same request.	0.00	1.00	0.35	0.48
Workload	The time-averaged number of sessions when focal session is ongoing.	1.00	5.89	2.45	0.94
$FrontRatio$	The proportion of effort that the agent allocates to front-office tasks in the focal session.	0.09	0.90	0.47	0.19
$FrontOfficeEffort$	The total time agent spends on front-office tasks in focal session.	10.00	763.00	146.96	139.75
$BackOfficeEffort$	The total time agent spends on back-office tasks in focal session.	11.00	866.00	170.30	163.49
Fatigue	The total number of one-hour shifts agent worked up to focal session in that day.	1.00	8.00	3.610	2.11
AvgCustomerMessageLength	The average Chinese character count of customers' messages in focal session.	2.50	27.75	8.77	4.45
AvgCustomerResponseDelay	The average time between the agent completes a message and the customer's response in seconds in focal session.	3.28	196.50	35.20	34.62
IsFirst	A binary variable indicating whether the session is the first contact for the request.	0.00	1.00	0.74	0.44

Table 2: Correlation Matrix for Session-level Variables

	1	2	3	4	5	6	7	8	9
Log(SessionDuration)	1.000								
CustomerRating	-0.368***	1.000							
$IfRetrial$	0.070***	-0.152***	1.000						
Workload	-0.050***	0.007***	-0.017***	1.000					
$FrontRatio$	-0.045***	-0.050***	-0.041***	-0.047***	1.000				
Fatigue	-0.016***	-0.002	0.004*	0.118***	-0.011***	1.000			
IsFirst	-0.094***	0.184***	-0.707***	0.007***	0.027***	-0.006***	1.000		
Log(AvgCustomerMessageLength)	0.253***	-0.123***	-0.011***	-0.019***	0.111***	-0.013***	0.015***	1.000	
Log(AvgCustomerResponseDelay)	0.598***	-0.147***	0.019***	0.021***	-0.192***	-0.011***	-0.030***	0.193***	1.000

Note: * is for $p < 0.1$, ** is for $p < 0.01$, *** is for $p < 0.001$

5 Empirical Analysis

We follow the causal pathways outlined in our conceptual framework and conduct our empirical analysis in two stages. First, we examine the direct causal effect of workload on the agent's within-session effort allocation between front- and back-office activities. We then conduct a mediation analysis to examine how effort allocation channels the effect of workload onto our three service performance outcomes: efficiency, short-term, and long-term service quality, respectively.

5.1 The Effect of Workload on Within-Session Effort Allocation

5.1.1 Estimation Strategy

We specify a fixed-effects model at the granular session level to examine the relationship between workload and effort allocation:

$$FrontRatio_{ij} = \beta_0 + \beta_1 Workload_{ij} + \beta_2 Workload_{ij}^2 + \beta X_{ij} + \mu_{ij} \quad (1)$$

Vector X_{ij} includes the control variables explained in Subsection 4.3.3. Although our fixed-effects model controls for both observed and unobserved heterogeneity at session level, the model may still be prone to endogeneity. Unobserved factors can influence both workload and effort allocation, causing omitted variable bias. For example, unobserved agent's motivation, or problem complexity can affect both workload and how agents allocate their effort.

To alleviate the endogeneity issue mentioned above, we employ the instrumental variable (IV) approach (Kennedy, 2008). A valid IV must meet two conditions: (1) Relevance: It must be correlated with the endogenous variables. (2) Exclusion Restriction: It should only affect the dependent variable through the endogenous variables (Wooldridge, 2010). In our estimation, we use two types of IVs.

We introduce a Hausman-type instrument (Hausman, 1994). In particular, we construct *HWorkload* as the average workload of other agents (i.e., excluding the focal agent j) with the *same* skill level (ensuring comparable service ability) working in the *same* shift (ensuring comparable traffic patterns), and include its

square term. This instrument is relevant because all agents working in the same shift face common demand shocks. It should also satisfy the exclusion restriction because an individual agent's allocation decision is unlikely to be directly affected by the specific workloads of their coworkers, beyond the indirect effect through their own workload. We show the first-stage results in Appendix A. *HWorkload* has positive and significant coefficients of 0.935 and 0.593 in the first stage estimations. The Cragg-Donald Wald F statistics (24.4 and 24.5) also support that these IVs are strong. All instrumental variables are standardized to match the normalization of the endogenous workload variables.

With these two types of IVs, we employ a control-function (CF) approach for estimation, which is often preferred for nonlinear causal effect models (Petrin and Train, 2010). Here are the model specifications of our CF estimation approach:

$$Workload_{ij} = \gamma_0 + \gamma_1 HWorkload_{ij} + \gamma_2 HWorkload_{ij}^2 + \gamma X_{ij} + \varepsilon_{ij} \tag{2}$$

$$Workload_{ij}^2 = \alpha_0 + \alpha_1 HWorkload_{ij} + \alpha_2 HWorkload_{ij}^2 + \alpha X_{ij} + e_{ij} \tag{3}$$

$$FrontRatio_{ij} = \beta_0 + \beta_1 Workload_{ij} + \beta_2 Workload_{ij}^2 + \beta X_{ij} + \hat{v}_{ij} + \mu_{ij} \tag{4}$$

In the first-stage (Equations 2 and 3), we regress the endogenous variables (*Workload* and *Workload*^{2}) on the instruments (i.e., *HWorkload*, and *HWorkload*^{2}) and control variables (X_{ij}). From these regressions, we obtain the estimated residuals of the first-stage regressions, denoted as $\hat{v}_{ij}$ ($\hat{v}_{ij} = (\hat{\varepsilon}_{ij}, \hat{e}_{ij})$). In the second stage, we estimate Equation 4, including the first stage residuals ($\hat{v}_{ij}$) as additional control variables. These residuals control for the portion of the workload variables that is correlated with the unobserved error term μ_{ij}, allowing for an unbiased estimation of the causal coefficients β_1 and β_2. In our analysis, we cluster the standard errors at agent level to account for potential intra-agent correlation over time.

5.1.2 Estimation Results

Table 3 presents the estimation results. Both OLS (column 2) and CF (column 4) results consistently suggest a U-shaped relationship between workload and effort allocation to the front-office activities, supporting H1. In particular, the coefficients of the linear term of workload are negative and statistically significant (β_1 : -0.109, -0.221), while the coefficients of the quadratic term are positive and statistically significant (β_2 : 0.086, 0.212). In addition, a partial F-test confirms that the quadratic model provides a significantly better fit than a linear specification (455.2, $p < 0.001$), further supporting the U-shaped relationship between workload and effort allocation.

Interpreting the coefficients of the CF estimation, we plot the non-linear relationship, in Figure 4. The turning point of this U-shaped curve, $-\beta_1/(2\beta_2)$, occurs at approximately 0.521 standard deviations (SD = 0.94) above the mean workload (mean = 2.45), which is equivalent to approximately 2.94 concurrent sessions. As workload increases from one session to this critical point, agents increasingly prioritize back-office work, when they can batch process customers' requests in the back-office. In fact, an increase in workload from one to 2.94 concurrent sessions corresponds to a 27.3% decrease in the proportion of time spent on front-office interactions ((0.459 – 0.631) / 0.631 $\approx$ 27.3%). This pattern is consistent with the idea that under moderate pressure, agents appear to use this additional engagement to protect core diagnostic and resolution-oriented work, consistent with research on discretionary task completion and workload-induced task prioritization. However, beyond this 2.94-session threshold, the trend inverts. As workload increases to five sessions, the front ratio increases by 42% ((0.652-0.459) / 0.459 = 42.0%). This shift suggests agents face tighter cognitive and temporal constraints, making it more difficult to sustain substantive back-office processing across multiple concurrent sessions. In this range, front-office communication becomes more operationally salient because it is visible to customers and helps manage waiting, expectations, and perceived responsiveness. Thus, the upward portion of the curve should not be interpreted as agents necessarily increasing absolute front-office effort. Rather, it reflects a relative shift in effort allocation: back-office work becomes more compressed than front-office interaction, causing the front ratio to rise. In other words,

agents prioritize front-office tasks under extremely high workload probably because communicating with customers is positively related to customer satisfaction ratings (we show this result in Table 4) and is generally not as time-consuming as back-end tasks.

Of course, the ratio reflects the relative effort allocation between the two types of activities, analyzing the absolute time spent on each activity provides deeper insight into the underlying behavioral mechanisms. We further analyzed the absolute time agents spend on front- and back-office tasks. The results, presented in Appendix C, provide direct evidence for the behavioral mechanisms underlying the U-shaped front-ratio pattern. Specifically, workload has an inverted U-shaped relationship with total active engagement time (the sum of front- and back-office effort), with the turning point at approximately 2.39 concurrent sessions. Workload also has inverted U-shaped relationships with both front-office effort and back-office effort, but the turning points differ: front-office effort peaks earlier, at approximately 1.77 concurrent sessions, whereas back-office effort peaks later, at approximately 2.62 concurrent sessions. The front ratio itself reaches its minimum later, at approximately 2.94 concurrent sessions. These decomposition results clarify why the front ratio is U-shaped even though the underlying absolute effort measures are inverted U-shaped. As workload rises from low to moderate levels, agents increase overall active engagement, and back-office effort grows or remains protected relative to front-office effort. This lowers the front ratio. As workload rises further, total active engagement begins to decline, and both types of effort are compressed. However, back-office effort eventually contracts more sharply relative to front-office interaction, causing the front ratio to turn upward. In this sense, the U-shaped relationship between workload and front ratio reflects two linked dynamics: an inverted U-shaped pattern in total active engagement and a workload-dependent shift in the composition of effort between front-office and back-office tasks.

The results for the control variables are consistent with our expectations. We find that higher agent fatigue is associated with a lower front ratio, suggesting that tired agents tend to retreat from more interactive, emotion-consuming front-office tasks. Direct customer interaction is typically considered energy-consuming emotional labor (Altman et al., 2020). Additionally, customer behavior directly affects how

agents allocate time. When customers write longer messages, agents spend more time on the front end to craft detailed responses. In contrast, when customers take longer to reply, agents shift to back-office work, reducing their front ratio. Crucially, the nature of the session significantly affects service strategy. For first-contact sessions ($IsFirst = 1$), agents maintain a higher front ratio, likely because initial interactions require more information gathering and expectation alignment. Conversely, in retrial sessions ($IsFirst = 0$), agents can rely more heavily on prior service history and back-office records, reducing the need for real-time customer-facing communication. Together, these patterns reinforce our interpretation that agents' front ratio reflects an endogenous effort-allocation response to both workload pressure and customer-side interaction demands.

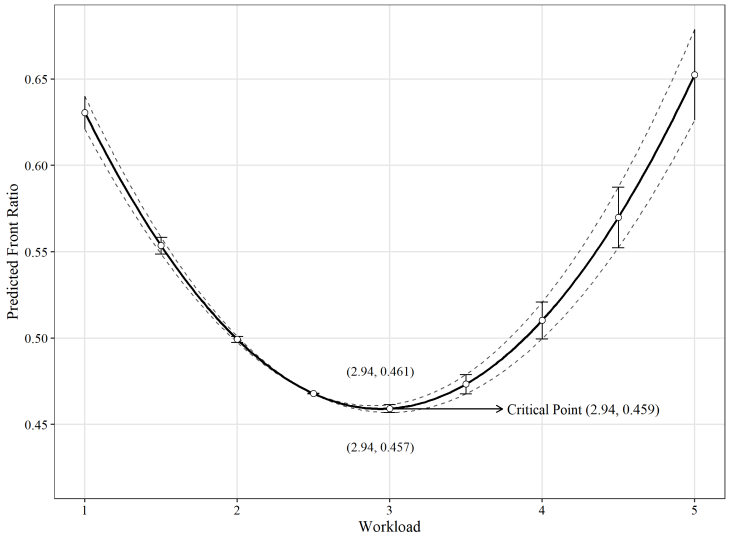


Figure 4: Predicted Front Ratio by Workload

Table 3: Impact of Workload on Allocation Decision

Dependent Variable	*FrontRatio*			
	(1)	(2)	(3)	(4)
Workload	-0.024***	-0.109***	0.007***	-0.221***
	(0.001)	(0.003)	(0.002)	(0.009)
$Workload^2$		0.086***		0.212***
		(0.003)		(0.009)
Fatigue	-0.002***	-0.002***	-0.002***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)
IsFirst	0.031***	0.030***	0.031***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)
Log(*AvgCustomerMessageLength*	0.291***	0.291***	0.292***	0.292***
	(0.002)	(0.002)	(0.002)	(0.002)
Log(*AvgCustomerResponseDelay*)	-0.251***	-0.250***	-0.251***	-0.249***
	(0.001)	(0.001)	(0.001)	(0.001)
$ResWorkload1(\hat{\varepsilon}_{ij})$			-0.040***	0.129***
			(0.002)	(0.010)
$ResWorkload2(\hat{e}_{ij})$				-0.150***
				(0.009)
Agent Effects	Yes	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes	Yes
N	3,965,028	3,965,028	3,965,028	3,965,028
$adj.R^2$	0.200	0.200	0.200	0.200

Notes. Robust standard errors clustered at the agent level in parentheses.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.2 The Mediating Role of the Front Ratio on Service Performance

5.2.1 Estimation Strategy

Having established how workload influences agents' effort allocation between front- and back-office activities, we next examine how this allocation mediates the relationship between workload and service performance. Our objective is to estimate the direct and indirect service performance effects of workload via front ratio. However, the primary independent variable (*Workload*) and the mediator (*FrontRatio*) are subject to endogeneity. Workload has been found to be endogenous to service performance because of omitted variable bias and reverse causality in previous work (e.g., Tan and Netessine (2014)). The front ratio is also

likely to be endogenous. For example, service quality expectations can influence effort allocation, creating a simultaneity bias. In addition, unobserved factors such as agent motivation can affect both effort allocation and service quality, causing an omitted variable bias. To address these endogeneity concerns, we adopt a multi-stage control-function (CF) approach. This strategy requires valid instrumental variables (IVs) for both endogenous variables.

For the endogenous *Workload* terms, we use the Hausman-type of instrument (*HWorkload*) explained in Subsection 5.1.1. Similar to estimating the front ratio impact, the instruments should satisfy the relevance condition. In addition they should satisfy the exclusion restriction condition because other agents' workload should not affect the focal agent's service performance beyond their workload (Xu et al., 2022).

For the endogenous *FrontRatio* terms, we follow established approaches (Hausman, 1994; Soltani et al., 2022) and construct a new Hausman-type IV based on session similarity. *HFrontRatioPD*, is defined as the average front ratio of other sessions handled by the same agent, for the same problem type, and on the same day. These instruments satisfy the relevance condition because agents often adopt consistent allocation strategies in similar situations (e.g., for a given workload or on a given day). They should also satisfy the exclusion restriction because the front ratio of other sessions is unlikely to affect the service performance of the focal session except through its influence on the focal session's front ratio. As shown in Appendix A, the first-stage regressions corroborate that these instruments are highly relevant and not weak ($0.074, p < 0.001$,and $0.040, p < 0.001$, respectively), with Cragg-Donald Wald F-statistics well exceeding the conventional threshold of 10 (80.0 for SessionDuration, 18.1 for CustomerRating, and 109.9 for IfRetrial).

We then adapt the CF approach for mediation (following the principles in Hayes and Andrew (2009)). The estimation proceeds in three stages:

- **Stage1:** We replicate Equations 2 and 3 in Section 5.1.1 to obtain the residuals of workload, $\hat{v}_{ij}$.
- **Stage2:** We then regress the endogenous *FrontRatio* variables on the instrument (*HFrontRatioPD*), the *Workload* variables, all exogenous controls. We obtain the second set of residuals, $\hat{\mu}_{ij}$ ($\hat{\vartheta}_{ij}$,$\hat{\varsigma}_{ij}$).

$$FrontRatio_{ij} = \alpha_0 + \alpha_1 Workload_{ij} + \alpha_2 Workload_{ij}^2 + \gamma_1 HFrontRatioPD_{ij} + \gamma_2 HFrontRatioPD_{ij}^2 + \alpha X_{ij} + \hat{\vartheta}_{ij} \quad (5)$$

$$FrontRatio_{ij}^2 = \beta_0 + \beta_1 Workload_{ij} + \beta_2 Workload_{ij}^2 + \rho_1 HFrontRatioPD_{ij} + \rho_2 HFrontRatioPD_{ij}^2 + \beta X_{ij} + \hat{\varsigma}_{ij} \quad (6)$$

- **Stage3:** We finally regress service performance variables Y_{ij} on workload variables, front ratio variables, the residual vector $\hat{v}_{ij}$ obtained from stage 1, the residual vector $\hat{\mu}_{ij}$ obtained from stage 2, and the control variables. We estimate the model using Ordinary Least Squares (OLS) for our continuous outcomes (*SessionDuration* and *CustomerRating*). For our binary outcome, *IfRetrial*, we employ a linear probability model.

$$Y_{ij} = \gamma_0 + \gamma_1 Workload_{ij} + \gamma_2 Workload_{ij}^2 + \gamma_3 FrontRatio_{ij} + \gamma_4 FrontRatio_{ij}^2 + \gamma X_{ij} + \hat{v}_{ij} + \hat{\mu}_{ij} + \varepsilon_{ij} \quad (7)$$

In Equation 7, the coefficients of the *Workload* terms (γ_1, γ_2) capture the direct effect, while the indirect effects are identified by the product of the first-stage coefficients (β_1,β_2) from Section 5.2.1 and the corresponding second-stage coefficients (γ_3, γ_4). This multi-stage CF approach allows us to obtain unbiased estimates of the direct and indirect effects.

5.2.2 Mediation Results

Table 4 presents the mediation analysis results. In the session duration results (Column 1), the coefficients of workload and its squared term are 0.122 and -0.116 ($p < 0.001$), suggesting an inverted-U-shaped direct relationship between workload and service duration. This pattern is consistent with prior findings, where Tan and Netessine (2014) find an inverted-U-shaped relationship between workload and meal duration in a restaurant. More importantly, the mediating variable *FrontRatio* has a U-shaped relationship with session duration because the coefficients are significant and equal to -0.317 and 0.230. The turning point occurs approximately $(0.317/(2 \times 0.230) \approx 0.69$ standard deviations (0.19) above the mean (0.47), which is approximately equal to 0.60. In other words, allocating either too little or too much effort to front-office

(customer-facing) tasks leads to longer service duration, with the service efficiency being maximized (i.e., duration minimized) when the front ratio is at an intermediate level at approximately 0.60. At very low front ratios, agents underinvest in customer-facing communication, which creates severe information frictions. Without sufficient interaction to clarify ambiguous issues and establish a shared understanding, the service process is hindered by misunderstandings, repeated clarification loops, and rework, ultimately prolonging the session. As the front ratio increases toward the optimal moderate level, interaction efficiency improves; agents can accurately diagnose customer needs and facilitate smooth co-production, thereby accelerating resolution. However, when the front ratio becomes extremely high, this efficiency is lost. Disproportionate effort dedicated to front-office communication crowds out the essential back-office tasks, such as information retrieval and coordination, required for substantive progress. Consequently, the interaction becomes less problem-centered and more repetitive, stalling actual resolution and lengthening the overall service duration. Because both workload and front ratio enter the model with linear and quadratic terms, the mediated effect is not summarized by a single product of coefficients. We therefore compute the indirect effect using the marginal-effect decomposition described below. Specifically, at each workload level W, we decompose the marginal effect of workload on session duration into a direct effect and an indirect effect operating through the front ratio. The indirect effect is computed via the chain rule as $(dM/dW) \times (dY/dM)$, where (dM/dW) is the marginal effect of workload on front ratio and (dY/dM) is the marginal effect of front ratio on the session duration. To capture the substantive magnitude of this mediation pathway, we report the indirect share of absolute marginal impact, defined as the average absolute indirect effect divided by the sum of the average absolute direct and indirect effects. We find that this indirect pathway accounts for 20.95% of the total absolute impact of workload on session duration. This demonstrates that a managerially significant portion of the relationship between workload and service speed is driven by the agent's endogenous behavioral response.

Column 2 shows customer rating results. Similar to service duration results, the direct effect of workload on customer satisfaction is also inverted-U-shaped (coefficients are 0.079 and -0.175) with the critical point

at 0.22 standard deviations above the mean. This relationship supports previous work which often shows an inverted-U-shaped relationship between workload and service quality (Tan and Netessine, 2014; Kc, 2014). The coefficient of our mediating variable *FrontRatio* is significant and positive (1.047), while the coefficient of its quadratic term turns out to be negative and significant (-0.808). These results support Hypothesis 3 and suggest an inverted-U shape between front ratio and customer ratings, an immediate-term service quality measure. The turning point is approximately 0.65 standard deviations above the sample mean or equivalent to 0.59. This indicates that customer ratings improve as front-office effort increases from a low level, but decline once front-office effort becomes excessive. This finding is consistent with service quality theory, which distinguishes between interaction quality and outcome quality (Grönroos, 1984; Brady and Cronin Jr, 2001). Moderate front-office effort improves ratings because customers are more likely to perceive the agent as responsive, attentive, and clear in communication. However, customer-facing effort is valuable only when it is supported by credible substantive progress. When front-office effort becomes excessive, it may crowd out back-office problem solving and weaken the perceived usefulness of the interaction. In that range, additional communication may no longer signal effective service effort, but instead become less informative or less credible. The absolute average indirect effect calculation shows that the front-ratio pathway accounts for 36.07% of the total absolute impact of workload on customer ratings. Of course, customer ratings are susceptible to self-selection bias due to systematic motivational differences between customers who choose to provide a rating and those who do not. The rating response rate in our data set is approximately 16.40%. Hence, we conduct a robustness check using a Heckman selection model in Appendix E and find consistent results.

Finally, Column 3 shows the long-term service quality results measured in *IfRetrial.* The coefficient of *Workload* and its squared term are 0.011 and -0.008 with critical point at 0.69 standard deviations above the mean or approximately to 3.10 concurrent sessions.. These results suggest that workload has a direct inverted-U-shaped relationship with retrial rates, consistent with prior work where Goes et al. (2017) find that when multitasking levels increase from 1 to 3 in a call center could lower problem resolution rates. In

addition, the coefficients of the mediating variables *FrontRatio* and $FrontRatio^2$ are significant and equal to -0.172 and 0.162, respectively, representing probability-scale changes under the LPM and implying a U-shaped relationship, supporting Hypothesis 4. Its critical point is approximately 0.53 standard deviations above the sample mean or equivalent to 0.57. This result highlights that immediate customer satisfaction and durable resolution are related but distinct service outcomes. When the front ratio is too low, customers may not receive sufficient clarification, explanation, or emotional closure, making the interaction less satisfying but also more likely to return with the same or related issue (Tax et al., 1998; Maxham III and Netemeyer, 2002; Hu et al., 2022). As the front ratio increases from a low level, better communication and expectation alignment reduce uncertainty and improve closure. However, when the front ratio becomes too high, excessive front-office emphasis can displace the back-office work required for actual resolution. In that case, customers may still need to return even if the focal interaction appears attentive. This interpretation is consistent with research on retrials and repeat contact as downstream consequences of unresolved service needs. The absolute average indirect effect calculation shows that the front-ratio pathway accounts for 68.59% of the total absolute impact of workload on retrial rate. We further unpack this mechanism using chat-text evidence in Section 6.

To recap, we delineate the direct effects of workload and the indirect effects mediated by agents' strategic effort allocation on service performance. This behavioral pathway is managerially significant across all outcomes: on average, accounting for 20.95% of the effect on session duration, 36.07% on customer ratings, and 68.59% on the retrial rate. The nonlinear marginal effects underlying these decomposition results are visualized in Appendix F.2. These findings indicate that the front-office/back-office balance is a meaningful operational lever rather than a passive byproduct of workload. Our estimates suggest that managers can adjust agents' effort allocation to improve service performance. For example, increasing the front ratio by 0.53 standard deviations from the sample mean (0.47) to 0.57 can achieve a Pareto optimal balance among service duration, customer ratings, and retrial rate. With this adjustment, service duration decreases by 9.81% (equivalent to a 50.75-second reduction); customer ratings increase by 10.73% (equivalent to a 0.33

points improvement); retrial rate is minimized and decreases by 13.04% (equivalent to a 4.57 percentage-point reduction).

Finally, the control variables provide additional evidence consistent with our interpretation. The coefficients on *IsFirst* show that, compared with retrial sessions, first-contact sessions are shorter, receive higher ratings, and are much less likely to result in another retrial. This underscores the operational importance of first-contact resolution and validates our focus on retrial as a durable service-quality outcome. As detailed in Section 7, these core findings are robust to a wide range of alternative specifications and measures.

Table 4: Mediating Effects of *FrontRatio* on Service Performance

Dependent Variable	$Log(SessionDuration)$	$CustomerRating$	$IfRetrial$
	(1)	(2)	(3)
$FrontRatio$	-0.317***	1.047**	-0.172***
	(0.053)	(0.379)	(0.028)
$FrontRatio^2$	0.230***	-0.808*	0.162***
	(0.054)	(0.380)	(0.029)
$Workload$	0.122***	0.078*	0.011***
	(0.006)	(0.036)	(0.003)
$Workload^2$	-0.116***	-0.174***	-0.008***
	(0.005)	(0.033)	(0.002)
$Fatigue$	-0.002***	-0.001	0.000
	(0.000)	(0.002)	(0.000)
$IsFirst$	-0.064***	0.651***	-0.762***
	(0.001)	(0.005)	(0.000)
$Log(AvgCustomerMessageLength)$	0.223***	-0.491***	0.007***
	(0.002)	(0.016)	(0.001)
$Log(AvgCustomerResponseDelay)$	0.542***	-0.137***	-0.008***
	(0.002)	(0.014)	(0.001)
$ResWorkload1$	-0.146***	-0.251***	-0.009***
	(0.006)	(0.037)	(0.003)
$ResWorkload2$	0.060***	0.403***	0.006*
	(0.005)	(0.034)	(0.002)
$ResFrontRatio1$	0.696***	-1.617***	0.164***
	(0.053)	(0.380)	(0.028)
$ResFrontRatio2$	-0.603***	1.193**	-0.159***
	(0.053)	(0.381)	(0.029)
Agent Effects	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes
N	3,965,028	650,660	3,965,028
$adj.R^2$	0.468	0.164	0.537

Notes. Robust standard errors clustered at the agent level in parentheses.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6 Unpacking the Mechanism: Textual Analysis

Our mediating results show that service duration and retrial are minimized, and customer ratings are maximized, at an intermediate level of front-office effort. To explain this pattern, we analyze the content of agents' chat interactions (3.9 million session transcripts). Textual analysis allows us to examine whether

these nonlinear effects arise because front ratio changes both the substantive content and the interaction quality of service conversations. Our textual measures are motivated by the service quality framework of Brady and Cronin Jr (2001), which emphasizes that customers' service quality perceptions are shaped by multiple dimensions, especially outcome quality and interaction quality. In our online chat setting, outcome quality is reflected in whether the conversation remains focused on diagnosing and resolving the customer's problem, whereas interaction quality is reflected in whether the agent manages emotions and communicates clearly. We use two complementary methods: a Latent Dirichlet Allocation (LDA) model to identify the thematic content of agent communication, and a proprietary Large Language Model (LLM) to score the nuanced quality of the interactions. The details about the methods are provided in Appendix D.

6.1 LDA Analysis: The Shift of Service Outcome Quality

We first examine service content using problem-centeredness. We define problem-centeredness as the extent to which the conversation focuses on diagnosis, information retrieval, solution explanation, and substantive problem resolution. Theoretically, this measure captures the outcome-quality dimension in the service quality framework (Brady and Cronin Jr, 2001). It is also closely related to communication theories emphasizing that effective service requires reducing ambiguity and establishing shared understanding (Daft and Lengel, 1986; Clark and Brennan, 1991). If front-office communication is productive, it should help the agent and customer clarify the problem and move the issue toward resolution.

To construct this measure, we apply LDA to identify latent topics within the agent conversation corpus (Blei et al., 2003; Liu and Toubia, 2018; Zhong and Schweidel, 2020). We then prompt a large language model (DeepSeek-R1) to classify each machine-generated topic as either "problem-solving oriented" (focused on logistics, refunds, product specs) or "relational/comforting" (focused on apologies, empathy, pleasantries) based on its associated keywords. We use the LLM model instead of hiring people to classify because we want to code the chats systematically and reduce subjectivity bias. For example, an LDA-generated topic that contains "shipping fee, door-to-door, pickup, freight cost, usage, vip, benefits, monthly,

enjoy, claim, return, member, returns and exchanges, offset, failed" is classified as "problem-solving oriented" by the LLM, while a topic that includes "sorry, caused, experience, really, platform, shopping, unpleasant, inconvenience, understanding, indeed, feeling, feedback, please don't be upset, wish, truly" is considered "comforting oriented". Based on this classification, we then compute the proportion of "problem-solving oriented" topics per session weighted by sentence relevance to construct variable *Problem-Centered Ratio*$_1$ to measure how strongly the service concentrates on the problem. As a robustness check, we alternatively ask the LLM to rate the topics on the scale between 1 (least problem-solving oriented) and 5 (most problem-solving oriented) and calculate the average score weighted by topics relevance per session to be *Problem-Centered Ratio*$_2$. The details can be found in Appendix D.

After constructing these two measures, we repeat Models 5 through Equation 7 to understand the impact of front ratio on agents' tendency of service focus. The results, presented in Table 5, reveal an inverted-U shaped relationship between an agent's front ratio and the problem-centeredness measures. These results suggest that moderate front-office effort improves the substantive focus of the service interaction. When front ratio increases from a low level, agents communicate more with customers, clarify the issue, and keep the conversation more focused on problem resolution. However, once front ratio becomes too high, problem-centeredness declines. This pattern is consistent with the limited-attention logic: excessive front-office effort can crowd out back-office diagnosis, information retrieval, and substantive problem solving (Ocasio, 1997). Thus, front-office effort improves outcome-related communication up to a point, but beyond that point additional front-office emphasis becomes less connected to actual resolution.

Table 5: The Shift of Communication Content

Dependent Variable	$Problem\text{-}Centered\ Ratio_1$	$Problem\text{-}Centered\ Ratio_2$
	(1)	(2)
$FrontRatio$	0.224***	0.162***
	(0.020)	(0.014)
$FrontRatio^2$	-0.242***	-0.169***
	(0.020)	(0.014)
$Workload$	-0.014***	-0.019***
	(0.002)	(0.001)
$Workload^2$	0.010***	0.015***
	(0.002)	(0.001)
$Fatigue$	0.001***	0.000***
	(0.000)	(0.000)
$IsFirst$	0.043***	0.042***
	(0.000)	(0.000)
$Log(AvgCustomerMessageLength)$	-0.035***	-0.025***
	(0.001)	(0.001)
$Log(AvgCustomerResponseDelay)$	-0.007***	0.000
	(0.001)	(0.000)
$ResWorkload1$	0.016***	0.021***
	(0.002)	(0.001)
$ResWorkload2$	-0.009***	-0.014***
	(0.002)	(0.001)
$ResFrontRatio1$	-0.246***	-0.177***
	(0.020)	(0.014)
$ResFrontRatio2$	0.250***	0.177***
	(0.020)	(0.014)
Agent Effects	Yes	Yes
Problem Effects	Yes	Yes
Day Effects	Yes	Yes
Hour Effects	Yes	Yes
N	3,965,028	3,965,028
$adj.R^2$	0.189	0.171

Notes. Robust standard errors clustered at the agent level in parentheses.
† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6.2 LLM Analysis: The Shift in Interaction Quality

Building on the LDA analysis, which reveals how the front ratio affects the outcome quality, we next employ large language model (LLM) analysis to gain additional insight into the impact on interaction quality. Our model was trained on a large corpus of human evaluations conducted by the company to emulate customer judgment when assessing service conversations across the following dimensions:

- Emotion Management: an agent's effectiveness in transforming a customer's emotion from negative to positive.
- Communication: the agent's ability to understand the customer and convey information with clarity.
- Overall Approval: a general assessment of the customer's approval of the agent.

We used this LLM to score a random sample of approximately 600,000 sessions from our data because of computational limitations. More details about our LLM approach can be found in Appendix D.

Table 6 shows the LLM results. Across all three measures, we find that FrontRatio again has an inverted U-shaped relationship with interaction quality. These findings indicate that moderate front-office effort improves interaction quality. When front ratio is low, agents may provide limited clarification, less adaptive communication, or overly generic responses. From the customer side, weaker communication may reduce engagement or make it harder to understand what information is needed and what progress is being made. This can generate misunderstanding, repeated clarification, and rework, thereby prolonging the session. However, when front ratio becomes too high, additional communication may no longer be supported by substantive progress. Communication can become repetitive, less informative, or less credible, weakening Emotional Management, Communication, and Overall Approval.

Taken together, the LDA and LLM results explain the mediation findings in Section 5.2. Moderate front-office effort improves both outcome quality and interaction quality: conversations become more problem-centered, emotionally better managed, clearer, and more positively evaluated. Too little front-office effort weakens dyadic service interaction and customer engagement, which helps explain longer session duration. Too much front-office effort crowds out back-office problem solving, reducing problem-centeredness and eventually weakening communication quality, which helps explain declining ratings and higher retrial rates. These textual results therefore show why service performance is optimized at an intermediate front ratio.

Table 6: LLM Results

Dependent Variable	*EmotionManagement*	*Communication*	*OverallApproval*
	(1)	(2)	(3)
$FrontRatio$	0.703*	0.718**	0.759*
	(0.320)	(0.244)	(0.333)
$FrontRatio^2$	-0.581†	–0.629*	-0.641†
	(0.320)	(0.244)	(0.333)
$Workload$	0.183***	0.247***	0.283***
	(0.029)	(0.022)	(0.030)
$Workload^2$	-0.328***	-0.354***	-0.440***
	(0.026)	(0.020)	(0.028)
$Fatigue$	-0.004**	-0.003**	-0.004**
	(0.001)	(0.001)	(0.001)
$IsFirst$	0.501***	0.335***	0.551***
	(0.005)	(0.004)	(0.005)
$Log(AvgCustomerMessageLength)$	-0.349***	-0.236***	-0.409***
	(0.012)	(0.010)	(0.014)
$Log(AvgCustomerResponseDelay)$	-0.079***	-0.078***	-0.097***
	(0.012)	(0.009)	(0.010)
$ResWorkload1$	-0.215***	-0.266***	-0.308***
	(0.030)	(0.023)	(0.031)
$ResWorkload2$	0.403***	0.407***	0.510***
	(0.028)	(0.021)	(0.029)
$ResFrontRatio1$	-1.039**	-1.030***	-1.143***
	(0.320)	(0.244)	(0.333)
$ResFrontRatio2$	0.793*	0.844***	0.874**
	(0.320)	(0.245)	(0.334)
Agent Effects	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes
N	594,156	594,156	594,156
$adj.R^2$	0.138	0.148	0.162

Notes. Robust standard errors clustered at the agent level in parentheses.
† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

7 Robustness Checks

We conduct a comprehensive set of robustness checks that test the sensitivity of our results to alternative model specifications, sample definitions, and variable measures. As detailed below, our core results, the non-linear effects of workload on agent effort allocation and the subsequent mediating effects of the front ratio, remain consistent across all tests.

7.1 Alternative Model Specifications

In our main analysis, we use quadratic terms to capture non-monotonic responses. As a robustness check, we follow prior literature (Xu et al., 2022; Soltani et al., 2022) to conduct spline regressions. We use two knots

to divide workload and *FrontRatio* into three equally sized intervals and estimate the effects of workload on agents' effort allocation (measured in *FrontRatio*) within each interval and the effects of *FrontRatio* on service outcomes, respectively. The results, shown in Tables 7a and 7b. The slope estimates are qualitatively congruent with the results presented in Tables 3 and 4. However, for the customer rating and retrial rate, only the coefficients of the first spline are significant while the coefficient of the left two spline knots are not significantly different from zero.

The spline estimates are qualitatively consistent with our main findings. In Table 7a, the effect of Workload on FrontRatio is negative in the low-workload interval and positive in the middle- and high-workload intervals, consistent with the U-shaped relationship reported in Table 3. In Table 7b, the slope estimates for FrontRatio also follow the patterns implied by our mediation results: the relationship with session duration is broadly U-shaped, whereas the relationships with customer rating and retrial rate move in the directions expected from the inverted-U-shaped and U-shaped patterns, respectively. However, similar to the spline-robustness evidence in Soltani et al. (2022), not all local spline coefficients are statistically significant, especially in the later intervals for customer rating and retrial rate. We therefore interpret the spline regressions as supportive evidence on the shape of the relationships rather than as the primary statistical test of non-monotonicity. The formal evidence for the nonlinear relationships comes from the quadratic specifications and the Lind and Mehlum tests reported below.

Table 7: Alternative Spline Regression

(a) Spline Regressions of *Workload*

Dependent Variable	*FrontRatio*
	(1)
*Workload*0% – 33.33%	-0.046***
	(0.007)
*Workload*33.33% – 66.67%	0.040***
	(0.004)
*Workload*66.67% – 100%	0.065***
	(0.005)
N	3,965,028
$adj.R^2$	0.202

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

(b) Spline Regressions of *Front Ratio*

Dependent Variable	*Log*(*SessionDuration*)	*CustomerRating*	*IfRetrial*
	(1)	(2)	(3)
*FrontRatio*0% – 33.33%	-0.086***	0.279***	-0.013***
	(0.008)	(0.051)	(0.004)
*FrontRatio*33.33% – 66.67%	-0.014***	-0.025	0.002
	(0.005)	(0.029)	(0.002)
*FrontRatio*66.67% – 100%	0.017***	-0.038	0.001
	(0.005)	(0.031)	(0.002)
N	3,965,028	650,660	3,965,028
$adj.R^2$	0.469	0.173	0.538

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Because spline regressions estimate separate local slopes and do not by themselves provide a formal test

of U-shaped or inverted-U-shaped relationships, we further employ the Lind and Mehlum test to validate the non-linear relationships (Lind and Mehlum, 2010). The test was used in previous literature (e.g, Tan and Netessine 2014; Xu et al. 2022) to examine U-shaped (or inverted U-shaped) relationships in linear regression models. We further conduct it on both Probit & Logit specifications for our binary outcome, *IfRetrial*, which yield consistent signs and significance patterns in Table 8b. The results presented in Table 8a confirm the validity of the specified non-linear relationships for both main results and mediation results at least at the 95% confidence level. As can be seen, the coefficients of the slopes are all significant. All the slope coefficients have the expected signs, consistent with the quadratic specification results in our main analysis.

Table 8: Lind Testing and Alternative Models

(a) Lind and Mehlum U-shape Testing

	Main Model (*Workload*)		Mediation Model (*SessionDuration*)		Mediation Model (*CustomerRating*)		Mediation Model (*IfRetrial*)	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Interval	-1.545	3.676	-2.507	2.809	-2.507	2.809	-2.507	2.809
Slope	-0.870	1.324	-1.420	0.935	4.884	-3.324	-0.935	0.670
t-value	-23.380	23.141	-4.616	3.938	2.251	-1.988	-5.737	5.574
$P > \|t\|$	0.000	0.000	0.000	0.000	0.012	0.023	0.000	0.000

(b) The Alternative Models

	IfRetrial	
	Probit	Logit
FrontRatio	-1.114***	-2.318***
	(0.150)	(0.280)
$FrontRatio^2$	1.060***	2.211***
	(0.151)	(0.282)
Workload	0.063***	0.111***
	(0.014)	(0.026)
$Workload^2$	-0.048***	-0.086***
	(0.013)	(0.025)
Controls	✓	✓
FixedEffects	✓	✓
N	3,965,028	3,965,028

Notes. Robust standard errors clustered at the agent level.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

7.2 Alternative Sample

As mentioned in Section 4.2, our main analysis excluded sessions with fewer than three interaction rounds. We use the full sample of 4,683,611 sessions without winsorization to assess the robustness of our findings. The results, presented in Table 9, remain consistent with our findings. They suggest a U-shaped relationship between workload and front ratio, a U-shaped relationship between front ratio and session duration, an inverted-U shaped relationship between front ratio and customer rating, and a U-shaped relationship between

front ratio and customers' retrial rate.

Table 9: Alternative Sample Robustness Results

Dependent Variable	$FrontRatio$	$Log(SessionDuration)$	$CustomerRating$	$IfRetrial$
	(1)	(2)	(3)	(4)
$Workload$	–0.132***	0.062***	-0.040†	0.008***
	(0.006)	(0.004)	(0.022)	(0.001)
$Workload^2$	0.127***	-0.061***	-0.055**	-0.006***
	(0.005)	(0.004)	(0.020)	(0.001)
$FrontRatio$		-0.246***	0.897*	-0.165***
		(0.056)	(0.364)	(0.026)
$FrontRatio^2$		0.164**	-0.647†	0.155***
		(0.056)	(0.366)	(0.026)
$Fatigue$	-0.003***	-0.002***	0.000	0.000
	(0.000)	(0.000)	(0.002)	(0.000)
$IsFirst$	0.039***	-0.054***	0.666***	-0.758***
	(0.001)	(0.001)	(0.006)	(0.000)
$Log(AvgCustomerMessageLength)$	0.243***	0.188***	-0.416***	0.003***
	(0.001)	(0.002)	(0.014)	(0.001)
$Log(AvgCustomerResponseDelay)$	-0.191***	0.542***	-0.150***	-0.006***
	(0.001)	(0.002)	(0.010)	(0.001)
$ResWorkload1$	0.046***	-0.141***	-0.033	-0.007***
	(0.006)	(0.004)	(0.023)	(0.002)
$ResWorkload2$	-0.073***	0.050***	0.189***	0.005***
	(0.005)	(0.004)	(0.020)	(0.001)
$ResFrontRatio1$		0.690***	-1.440***	0.154***
		(0.056)	(0.364)	(0.026)
$ResFrontRatio2$		-0.585***	1.008**	-0.151***
		(0.056)	(0.366)	(0.026)
Agent Effects	Yes	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes	Yes
N	4,683,611	4,683,611	687,249	4,683,611
$adj.R^2$	0.178	0.449	0.159	0.534

Notes. Robust standard errors clustered at the agent level in parentheses.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

7.3 Alternative Measure of Front Ratio

As described in Section 4, our main analysis uses time-based measures for agents' effort allocation. As a robustness check, we employ count-based measures. Using the same categorization of within-session

activities, we first count the number of front-office actions and back-office actions within each session. We then define the count-based front ratio, $FrontRatio_2$, as the proportion of total within-session actions that are categorized as front-office. We also standardize variable $FrontRatio_2$ for consistency. Table 10 shows the results of the alternative measure of front ratio. They are consistent with our main findings. Workload has a U-shaped relationship with front ratio with the critical point at 0.49 standard deviations above the mean. In addition, front ratio has a U-shaped relationship with session duration, with the critical point being 0.66 standard deviations above the mean. It is also exhibiting an inverted-U shaped relationship associated with customer ratings, with the critical point being 0.64 standard deviations above the mean. Finally, front ratio has a U-shaped relationship with retrial rates, and the critical point is 0.52 standard deviations above the mean.

Table 10: Results for Service Performance with Respect to Alternative FrontRatio Measure

Dependent Variable	$FrontRatio_2$	$Log(SessionDuration)$	$CustomerRating$	$IfRetrial$
	(1)	(2)	(3)	(4)
$Workload$	-0.201***	0.119***	0.082*	0.011***
	(0.010)	(0.006)	(0.036)	(0.003)
$Workload^2$	0.205***	-0.108***	-0.191***	-0.008***
	(0.009)	(0.005)	(0.033)	(0.002)
$FrontRatio_2$		-0.367***	1.217**	-0.187***
		(0.058)	(0.408)	(0.031)
$FrontRatio_2^2$		0.276***	-0.957*	0.180***
		(0.059)	(0.417)	(0.031)
$Fatigue$	0.003***	-0.001***	-0.003†	0.000
	(0.000)	(0.000)	(0.002)	(0.000)
$IsFirst$	0.114***	-0.051***	0.610***	-0.758***
	(0.002)	(0.002)	(0.011)	(0.001)
$Log(AvgCustomerMessageLength)$	0.255***	0.219***	-0.489***	0.005***
	(0.002)	(0.002)	(0.013)	(0.001)
$Log(AvgCustomerResponseDelay)$	-0.183***	0.550***	-0.159***	-0.004***
	(0.001)	(0.002)	(0.009)	(0.001)
$ResWorkload1$	0.163***	-0.138***	-0.272***	-0.009***
	(0.010)	(0.006)	(0.037)	(0.003)
$ResWorkload2$	-0.163***	0.050***	0.430**	0.007**
	(0.009)	(0.005)	(0.035)	(0.002)
$ResFrontRatio_2 1$		0.394***	-1.011*	0.196***
		(0.058)	(0.408)	(0.031)
$ResFrontRatio_2 2$		-0.390***	0.787†	-0.191***
		(0.059)	(0.417)	(0.031)
Agent Effects	Yes	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes	Yes
N	3,965,028	3,965,028	650,660	3,965,028
$adj.R^2$	0.331	0.467	0.155	0.537

Notes. Robust standard errors clustered at the agent level in parentheses.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

8 Conclusion Remarks

In this paper, we use detailed operational data from a large e-commerce platform to study how workload affects agents' within-session allocation between front-office interaction and back-office problem solving, and how this allocation subsequently shapes service performance. We measure agents' allocation using the

front ratio, defined as the proportion of active service effort devoted to customer-facing interaction within a session. Adopting a control-function approach to address potential endogeneity, we find that workload has a U-shaped relationship with the front ratio. When workload is low to moderate, increasing workload leads agents to allocate relatively more effort to back-office problem solving, reducing the front ratio. However, when workload becomes high, further increases in workload compress back-office effort more sharply and shift agents proportionally back toward front-office interaction. This finding shows that workload changes not only the level of service pressure agents face, but also the structure of effort they allocate within the service encounter.

We further show that this effort allocation decision is consequential for service performance. The front ratio significantly mediates the relationship between workload and three key outcomes: session duration, customer ratings, and customer retrials. Specifically, the front ratio has a U-shaped relationship with session duration, an inverted-U-shaped relationship with customer ratings, and a U-shaped relationship with retrial rate. These findings suggest that service performance is optimized at an intermediate level of front-office effort. Too little front-office effort may weaken clarification, guidance, and customer engagement, thereby prolonging the session and increasing the likelihood of repeat contact. Too much front-office effort, however, may crowd out diagnosis, information retrieval, and substantive problem solving, which can again lengthen the session, reduce customer ratings, and increase retrials. Quantitatively, the front-ratio pathway accounts for 20.95% of the total absolute impact of workload on session duration, 36.07% on customer ratings, and 68.59% on retrial rate. Counterfactual analysis further suggests that increasing the front ratio by 0.53 standard deviations from the sample mean of 0.47 to 0.57 can reduce session duration by 9.81%, increase customer ratings by 10.73%, and reduce retrial rates by 13.04%.

Textual analyses provide evidence for the mechanisms underlying these nonlinear performance effects. Guided by the service quality distinction between outcome quality and interaction quality, we examine both the problem-centeredness and interaction quality of agent-customer conversations. Using LDA-based measures, we find that a moderate front ratio increases problem-centered communication, suggesting that

customer-facing interaction is most effective when it helps clarify the issue and move the case toward resolution. Using LLM-based measures trained on company evaluation data, we further find that a moderate front ratio improves emotional management, communication, and overall approval. However, when the front ratio becomes too high, problem-centeredness declines and interaction quality eventually weakens. These findings suggest that excessive front-office emphasis can make communication less supported by substantive problem-solving progress, thereby explaining why service performance deteriorates beyond the optimal front ratio.

From a managerial perspective, our findings suggest that service platforms should manage the composition of agent effort rather than simply emphasizing response speed or communication volume. The optimal policy is not to maximize customer-facing interaction, but to maintain a balance between front-office communication and back-office problem solving. This has implications for training, performance evaluation, service process design, and AI-enabled support. Training programs should help agents recognize when communication improves coordination and when it begins to substitute for substantive resolution. Performance evaluation systems should balance immediate customer ratings with longer-term outcomes such as retrials. Service processes should also be designed to protect agents' capacity for diagnosis, information retrieval, and coordination under high workload. Finally, AI tools in customer service should not only help agents generate responsive or empathetic messages, but also support problem diagnosis, information retrieval, and resolution-oriented communication.

Our study makes three contributions to the literature. First, we contribute to workload and behavioral operations research by identifying within-session effort allocation as a mechanism through which workload affects service performance. Prior research has shown that workload influences service speed, quality, multitasking, and operational risk; we extend this literature by showing that workload also changes how agents allocate effort between customer-facing interaction and problem-solving work. Second, we contribute to the service operations literature on front-office/back-office design. Whereas prior work often treats the front-office/back-office distinction as a structural design choice, we show that this boundary is also dynamically

managed by agents within individual service sessions. Third, we contribute to service quality research by jointly examining efficiency, immediate customer satisfaction, and durable resolution. Our findings show that service performance is optimized at an intermediate front ratio, highlighting that high-quality service requires balancing interaction quality with outcome quality rather than maximizing customer-facing effort alone.

Finally, it is important to understand the limitations of our work and establish future research directions. First, our analysis measures the front ratio as a session-level average. Future research with more granular interaction data could examine how agents dynamically adjust front-office and back-office effort within a single session. Second, our workload measure is based on the number of concurrent sessions, which may not fully capture variation in customer problem complexity or emotional intensity. Future research could develop complexity-weighted workload measures that incorporate problem type, customer sentiment, or expected resolution difficulty. Third, our study focuses on live-chat service in a large e-commerce platform. Future work could examine whether similar effort-allocation mechanisms arise in other service contexts, such as technical support, telemedicine, and financial services. Finally, although our textual analyses provide evidence on the communication mechanisms underlying the performance effects, future research could combine textual measures with field interventions to more directly test how training, routing, or AI-enabled support tools change agents' effort allocation and service outcomes.

References

Aksin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Alós-Ferrer, Carlos, Johannes Buckenmaier. 2021. Cognitive sophistication and deliberation times. *Experimental Economics* **24**(2) 558–592.

Altman, Dan, Galit Bracha Yom-Tov, Marcelo A. Olivares, Shelly Ashtar, Anat Rafaeli. 2020. Do customer

emotions affect agent speed? an empirical study of emotional load in online customer contact centers. *Manufacturing & Service Operations Management* **23** 854–875.

Armona, Luis, Greg Lewis, Georgios Zervas. 2024. Learning product characteristics and consumer preferences from search data. *Marketing Science* .

Batt, Robert J, Christian Terwiesch. 2017. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.

Bavafa, Hessam, Jónas Oddur Jónasson. 2024. The distributional impact of fatigue on performance. *Management Science* **70**(5) 3319–3337.

Becker, G.S. 1965. A theory of the allocation of time. *The Economic Journal* **75**(299) 493–517.

Benjaafar, Saif, Ming Hu. 2020. Operations management in the age of the sharing economy: What is old and what is new? *Manufacturing & Service Operations Management* **22**(1) 93–101.

Berry Jaeker, Jillian A, Anita L Tucker. 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* **63**(4) 1042–1062.

Bitner, Mary Jo, Bernard H Booms, Mary Stanfield Tetreault. 1990. The service encounter: diagnosing favorable and unfavorable incidents. *Journal of Marketing* **54**(1) 71–84.

Blei, David M, Andrew Y Ng, Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**(Jan) 993–1022.

Brady, Michael K, J Joseph Cronin Jr. 2001. Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing* **65**(3) 34–49.

Buell, Ryan W, Tami Kim, Chia-Jung Tsay. 2017. Creating reciprocal value through operational transparency. *Management Science* **63**(6) 1673–1695.

Buell, Ryan W, Michael I Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Science* **57**(9) 1564–1579.

Castellanos, Antonio, Galit B Yom-Tov, Yair Goldberg, Jaeyoung Park. 2025. Silent abandonment in text-based contact centers: Identifying, quantifying, and mitigating its operational impacts. *arXiv preprint arXiv:2501.08869* .

Chase, R B. 1978. Where does the customer fit in a service operation?? *Harvard Business Review* **56**(6) 137–142.

Chase, Richard B, David A Tansik. 1983. The customer contact model for organization design. *Management Science* **29**(9) 1037–1050.

Clark, Herbert H, Susan E Brennan. 1991. Grounding in communication. .

Coviello, Decio, Andrea Ichino, Nicola Persico. 2014. Time allocation and task juggling. *American Economic Review* **104**(2) 609–623.

Cui, Shiliang, Xuanming Su, Senthil Veeraraghavan. 2019. A model of rational retrials in queues. *Operations Research* **67**(6) 1699–1718.

Daft, Richard L, Robert H Lengel. 1986. Organizational information requirements, media richness and structural design. *Management Science* **32**(5) 554–571.

de Kok, Ties. 2025. Chatgpt for textual analysis? how to use generative llms in accounting research. *Management Science* **71**(9) 7888–7906.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. URL `https://arxiv.org/abs/2501.12948`.

Dong, Jing, Rouba Ibrahim. 2020. Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research* **68**(4) 1238–1264.

Dubé-Rioux, Laurette, Bernd H. Schmitt. 1989. Consumers' reactions to waiting: When delays affect the perception of service quality. *ACR North American Advances* .

Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Goes, Paulo, Noyan Ilk, Mingfeng Lin, J. Leon Zhao. 2017. When more is less: Field evidence on unintended consequences of multitasking. *Ewing Marion Kauffman Foundation Research Paper Series* .

Grönroos, Christian. 1984. A service quality model and its marketing implications. *European Journal of Marketing* **18**(4) 36–44.

Hathaway, Brett Alan, Seyed Morteza Emadi, Vinayak Deshpande. 2022. Personalized priority policies in call centers using past customer interaction information. *Management Science* **68**(4) 2806–2823.

Hausman, Jerry A. 1994. *Valuation of new goods under perfect and imperfect competition*. National Bureau of Economic Research Cambridge, Mass., USA.

Hayes, F. Andrew. 2009. Beyond baron and kenny: Statistical mediation analysis in the new millennium. *Communication Monographs* **76**(4) 408–420.

Heckman, James. 1974. Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society* 679–694.

Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of economic and social measurement, volume 5, number 4*. NBER, 475–492.

Hinton, Geoffrey, Oriol Vinyals, Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Hockey, G Robert J. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology* **45**(1-3) 73–93.

Hopp, Wallace J, Seyed MR Iravani, Fang Liu. 2009. Managing white-collar work: An operations-oriented survey. *Production and operations management* **18**(1) 1–32.

Hopp, Wallace J, Seyed MR Iravani, Gigi Y Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.

Hu, Kejia, Gad Allon, Achal Bassamboo. 2022. Understanding customer retrials in call centers: Preferences for service quality and service speed. *Manufacturing & Service Operations Management* **24**(2) 1002–1020.

Hu, Nan, Paul A. Pavlou, Jie Jennifer Zhang. 2017. On self-selection biases in online product reviews. *MIS Q.* **41** 449–471. URL `https://api.semanticscholar.org/CorpusID:37225725`.

Ibanez, Maria R, Jonathan R Clark, Robert S Huckman, Bradley R Staats. 2018. Discretionary task ordering: Queue management in radiological services. *Management Science* **64**(9) 4389–4407.

Ilk, N., Guangzhi Shang. 2022. The impact of waiting on customer-instigated service time: Field evidence from a live-chat contact center. *Journal of Operations Management* .

Kahneman, Daniel. 1973. Attention and effort .

Kc, Diwas S, Bradley R Staats, Maryam Kouchaki, Francesca Gino. 2020. Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science* **66**(10) 4397–4416.

Kc, D.S. 2014. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

Kc, D.S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

Kennedy, Peter. 2008. *A guide to econometrics*. John Wiley & Sons.

Legros, Benjamin, Oualid Jouini, O Zeynep Akşin, Ger Koole. 2020. Front-office multitasking between service encounters and back-office tasks. *European Journal of Operational Research* **287**(3) 946–963.

Leroy, Sophie. 2009. Why is it so hard to do my work? the challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes* **109**(2) 168–181.

Li, L., H. Zhang. 2000. The multistage service facility start-up and capacity model. *Operations Research* **48**(3) 490–497.

Lind, Jo Thori, Halvor Mehlum. 2010. With or without u? the appropriate test for a u-shaped relationship. *Oxford bulletin of economics and statistics* **72**(1) 109–118.

Liu, Jia, Olivier Toubia. 2018. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science* **37**(6) 930–952.

Long, Zhenghua, Tolga Tezcan, Jiheng Zhang. 2024. Routing and staffing in customer service chat systems with generally distributed service and patience times. *Manufacturing & Service Operations Management* **26**(5) 1674–1691.

Luo, Jun, Jiheng Zhang. 2013. Staffing and control of instant messaging contact centers. *Operations Research* **61** 328–343.

Maxham III, James G, Richard G Netemeyer. 2002. A longitudinal study of complaining customers' evaluations of multiple service failures and recovery efforts. *Journal of Marketing* **66**(4) 57–71.

Monsell, Stephen. 2003. Task switching. *Trends in Cognitive Sciences* **7**(3) 134–140.

Netzer, Oded, James M Lattin, Vikram Srinivasan. 2008. A hidden markov model of customer relationship dynamics. *Marketing Science* **27**(2) 185–204.

Niu, Yimeng, Jing Wu, Shenyang Jiang, Zhibin Jiang. 2025. The bullwhip effect in servitized manufacturers. *Management Science* **71**(1) 1–20.

Ocasio, William. 1997. Towards an attention-based view of the firm. *Strategic Management Journal* **18**(S1) 187–206.

Oliva, Rogelio, John D Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7) 894–914.

Petrin, Amil, Kenneth Train. 2010. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research* **47**(1).

Pierce, Jonathon R, Herman Aguinis. 2013. The too-much-of-a-good-thing effect in management. *Journal of Management* **39**(2) 313–338.

Powell, Adam, Sergei Savin, Nicos Savva. 2012. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management* **14**(4) 512–528.

Ravn, Morten, Stephanie Schmitt-Grohé, Martin Uribe. 2006. Deep habits. *The Review of Economic Studies* **73**(1) 195–218.

Röder, Michael, Andreas Both, Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

Roels, Guillaume. 2014. Optimal design of coproductive services: Interaction and work allocation. *Manufacturing & Service Operations Management* **16**(4) 578–594.

Safizadeh, M Hossein, Joy M Field, Larry P Ritzman. 2003. An empirical analysis of financial services processes with a front-office or back-office orientation. *Journal of Operations Management* **21**(5) 557–576.

Schoenmueller, Verena, Oded Netzer, Florian Stahl. 2020. The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research* **57**(5) 853–877.

Siano, Federico. 2025. The news in earnings announcement disclosures: Capturing word context using llm methods. *Management Science* .

Solomon, Michael R, Carol Surprenant, John A Czepiel, Evelyn G Gutman. 1985. A role theory perspective on dyadic interactions: the service encounter. *Journal of Marketing* **49**(1) 99–111.

Soltani, Mohamad, Robert J Batt, Hessam Bavafa, Brian W Patterson. 2022. Does what happens in the ed stay in the ed? the effects of emergency department physician workload on post-ed care use. *Manufacturing & Service Operations Management* **24**(6) 3079–3098.

Speier, Cheri, Joseph S Valacich, Iris Vessey. 1999. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences* **30**(2) 337–360.

SQM Group. 2025. First Call Resolution (FCR): A Comprehensive Guide. URL `https://www.sqmgroup.com/resources/library/blog/fcr-metric-operating-philosophy`.

Staats, Bradley R, Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management Science* **58**(6) 1141–1159.

Stoeber, Joachim, Daryl Chesterman, Terri-Anne Tarn. 2010. Perfectionism and task performance: Time on task mediates the perfectionistic strivings–performance relationship. *Personality and Individual Differences* **48**(4) 458–462.

Tan, F.T., S. Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

Tax, Stephen S, Stephen W Brown, Murali Chandrashekaran. 1998. Customer evaluations of service complaint experiences: implications for relationship marketing. *Journal of Marketing* **62**(2) 60–76.

Team, Qwen. 2024. Qwen2.5: A party of foundation models. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Teboul, James. 2006. *Service is front stage: positioning services for value advantage*. Springer.

Tezcan, Tolga, Jiheng Zhang. 2014. Routing and staffing in customer service chat systems with impatient customers. *Operations Research* **62** 943–956.

Weil, David. 2014. The fissured workplace: Why work became so bad for so many and what can be done to improve it. *The fissured workplace*. Harvard University Press.

Weitz, Barton A, Harish Sujan, Mita Sujan. 1986. Knowledge, motivation, and adaptive behavior: A framework for improving selling effectiveness. *Journal of Marketing* **50**(4) 174–191.

Wickens, Christopher D. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* **3**(2) 159–177.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Xu, Y., F.T. Tan, S. Netessine. 2022. The impact of workload on operational risk: Evidence from a commercial bank. *Management Science* **68**(4) 2668–2693.

Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* .

Yoganarasimhan, Hema, Irina Iakovetskaia. 2024. From feeds to inboxes: A comparative study of polarization in facebook and email news sharing. *Management Science* **70**(9) 6461–6472.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.

Zhang, Shunyuan, Das Narayandas. 2025. Engaging customers with ai in online chats: evidence from a randomized field experiment. *Management Science* .

Zhong, Ning, David A Schweidel. 2020. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science* **39**(4) 827–846.

Zomerdijk, Leonieke G, Jan de Vries. 2007. Structuring front office and back office work in service delivery systems: an empirical study of three design decisions. *International Journal of Operations & Production Management* **27**(1) 108–131.

Online Appendix

Appendix A: Validation of Instrumental Variables

Appendix A reports the first-stage regressions used in our control-function estimation. Panel (a) validates the relevance of the Hausman-type instruments for workload and its squared term. Panel (b) reports the first-stage regressions for the front-ratio instruments used in the mediation analysis.

Table 11: Validation of Instrumental Variables

(a) First-stage regression of *Workload*

Dependent Variable	*Workload*	$Workload^2$
	(1)	(2)
HWorkload	0.935***	0.593***
	(0.008)	(0.011)
$HWorkload^2$	-0.281***	0.084***
	(0.007)	(0.010)
Fatigue	-0.012***	-0.008***
	(0.001)	(0.001)
IsFirst	-0.006***	-0.003***
	(0.001)	(0.001)
Log(*AvgCustomerMessageLength*)	-0.030***	-0.032***
	(0.001)	(0.001)
Log(*AvgCustomerResponseDelay*)	-0.003***	-0.011***
	(0.001)	(0.001)
Agent Effects	Yes	Yes
Problem Effects	Yes	Yes
Day Effects	Yes	Yes
Hour Effects	Yes	Yes
N	3,965,028	3,965,028
$adj.R^2$	0.539	0.524

Notes. Robust standard errors clustered at the agent level in parentheses.
† *p < 0.1,* p < 0.05, ** p < 0.01, *** p < 0.001.*

(b) First-stage regression of *FrontRatio*

Dependent Variable	*FrontRatio*	$FrontRatio^2$
	(1)	(2)
HFrontRatioPD	0.074***	0.040***
	(0.004)	(0.004)
$HFrontRatioPD^2$	-0.005	0.029***
	(0.004)	(0.004)
Workload	-0.105***	-0.116***
	(0.003)	(0.003)
$Workload^2$	0.083***	0.092***
	(0.003)	(0.003)
Fatigue	-0.002***	-0.002***
	(0.000)	(0.000)
IsFirst	0.030***	0.028***
	(0.001)	(0.001)
Log(*AvgCustomerMessageLength*)	0.291***	0.265***
	(0.002)	(0.002)
Log(*AvgCustomerResponseDelay*)	-0.250***	-0.226***
	(0.001)	(0.001)
Agent Effects	Yes	Yes
Problem Effects	Yes	Yes
Day Effects	Yes	Yes
Hour Effects	Yes	Yes
N	3,965,028	3,965,028
$adj.R^2$	0.203	0.187

Notes. Robust standard errors clustered at the agent level in parentheses.
† *p < 0.1,* p < 0.05, ** p < 0.01, *** p < 0.001.*

Appendix B: Example of Agent Activities in Service

This appendix provides an illustrative example of how we classify agents' click-stream activities and construct the front-ratio measure. The service system records each agent action with a timestamp and an action name. We classify actions involving real-time customer communication, such as typing in the chat box,

sending messages, and sending solutions, as front-office activities. We classify actions related to diagnosis, information retrieval, service-history review, solution search, and documentation as back-office activities. Switch actions, such as switching into or out of the focal session, are used to identify whether the agent is actively engaged in the session, but they are not classified as either front-office or back-office activities.Switch actions and inactive buffer time are excluded when computing the front ratio.

For each activity, we calculate its duration as the elapsed time from the timestamp of the current action to the timestamp of the next action. The front ratio for a session is then calculated as the total duration of front-office activities divided by the total duration of front-office and back-office activities. Time associated with switching across sessions or inactive buffer periods is excluded from this denominator. Table 12 provides a simplified example of the action sequence and classification.

Table 12: An Illustrative of Agent Actions in Service

Index	Timestamp	Action Name	Activity Type	
0	2023-10-30 9:26:10	Switch In	Switch	
1	2023-10-30 9:26:34	Send Message	Front-office	
2	2023-10-30 9:26:40	Send Message	Front-office	
3	2023-10-30 9:26:45	Search Information	Back-office	
4	2023-10-30 9:26:53	Typing in chat-box	Front-office	
5	2023-10-30 9:26:58	Send Message	Front-office	
6	2023-10-30 9:27:02	Track Service History	Back-office	
7	2023-10-30 9:27:12	Check Order Information	Back-office	
8	2023-10-30 9:27:15	Typing in chat-box	Front-office	
9	2023-10-30 9:27:19	Send Message	Front-office	
10	2023-10-30 9:27:27	Track Service History	Back-office	
11	2023-10-30 9:27:43	Search Solutions	Back-office	
12	2023-10-30 9:27:50	Typing in chat-box	Front-office	
13	2023-10-30 9:27:57	Send Message	Front-office	
14	2023-10-30 9:27:59	Click Search Results	Back-office	Session Duration
15	2023-10-30 9:28:05	Check Solutions	Back-office	
16	2023-10-30 9:28:22	Send Solutions	Front-office	
17	2023-10-30 9:28:24	Typing in chat-box	Front-office	
18	2023-10-30 9:28:28	Send Message	Front-office	
19	2023-10-30 9:28:37	Switch Out	Switch	
20	2023-10-30 9:29:49	Switch In	Switch	
21	2023-10-30 9:29:56	Typing in chat-box	Front-office	
22	2023-10-30 9:30:08	Send Message	Front-office	
23	2023-10-30 9:30:19	Search Knowledge	Back-office	
24	2023-10-30 9:30:29	Send Knowledge	Front-office	
25	2023-10-30 9:32:10	Switch Out	Switch	
26	2023-10-30 9:32:47	Switch In	Switch	
27	2023-10-30 9:32:59	Typing in chat-box	Front-office	
28	2023-10-30 9:33:18	Send Message	Front-office	
29	2023-10-30 9:33:36	Switch Out	Switch	
30	2023-10-30 9:35:22	Switch In	Switch	Buffer Time
31	2023-10-30 9:35:25	Closed	Close	

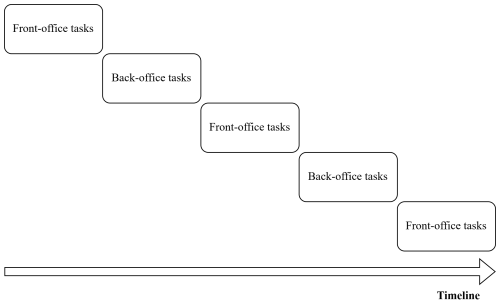


Figure 5: Illustration of Agent Activities in Service

Appendix C: Additional Analysis of Workload

Appendix C decomposes the workload-front-ratio relationship by examining how workload affects agents' absolute effort levels. Because the front ratio is a relative measure, a change in the front ratio may reflect changes in front-office effort, back-office effort, or both. We therefore estimate the effect of workload on three absolute effort measures: total active engagement time, front-office effort, and back-office effort. Total active engagement time is defined as the sum of front-office and back-office effort within the focal session, excluding switch actions and inactive buffer time. The results show that workload has inverted-U-shaped relationships with all three effort measures. These patterns indicate that agents initially increase active engagement as workload rises, but eventually compress effort under high workload. Importantly, front-office and back-office effort peak at different workload levels, which explains why the front ratio itself follows a U-shaped pattern.

Table 13: Decomposition of Workload Effects on Absolute Effort

Dependent Variable	$Log(ActiveEngagementTime)$	$Log(FrontOfficeEffort)$	$Log(BackOfficeEffort)$
	(1)	(2)	(3)
Workload	-0.018**	-0.098***	0.085***
	(0.006)	(0.007)	(0.007)
$Workload^2$	-0.148***	-0.068***	-0.240***
	(0.005)	(0.006)	(0.007)
Fatigue	-0.004***	-0.006***	-0.003***
	(0.000)	(0.000)	(0.000)
IsFirst	-0.073***	-0.059***	-0.085***
	(0.001)	(0.001)	(0.001)
$Log(AvgCustomerMessageLength)$	0.225***	0.370***	0.113***
	(0.001)	(0.001)	(0.001)
$Log(AvgCustomerResponseDelay)$	0.369***	0.240***	0.462***
	(0.001)	(0.001)	(0.001)
*ResWorkload*1	-0.204***	-0.160***	-0.262***
	(0.006)	(0.007)	(0.008)
*ResWorkload*2	0.197***	0.139***	0.258***
	(0.006)	(0.007)	(0.007)
Agent Effects	Yes	Yes	Yes
Problem Effects	Yes	Yes	Yes
Day Effects	Yes	Yes	Yes
Hour Effects	Yes	Yes	Yes
N	3,965,028	3,965,028	3,965,028
$adj.R^2$	0.327	0.251	0.320

Notes. Robust standard errors clustered at the agent level in parentheses.
† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

These decomposition results clarify the mechanism behind the U-shaped front-ratio pattern documented in Section 5.1.2. At low-to-moderate workload levels, back-office effort is maintained or increases relative to front-office effort, causing the front ratio to decline. At high workload levels, both types of effort decline, but back-office effort is compressed more sharply than front-office interaction, causing the front ratio to rise. Thus, the upward part of the front-ratio curve should be interpreted as a relative shift in effort allocation rather than a simple increase in absolute front-office effort.

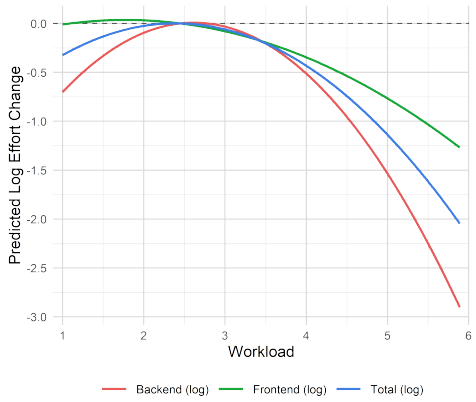


Figure 6: Workload Effects on Active Engagement, Front-Office Effort, and Back-Office Effort

Appendix D: Textual Analysis Methodology

This appendix describes the textual analysis methods used in Section 6. The analysis contains two complementary components. First, we use LDA to construct measures of problem-centered communication, which capture the outcome-quality dimension of the service conversation. Second, we use a proprietary LLM to score interaction quality, including emotional management, communication, and overall approval. Together, these measures allow us to examine whether the front ratio affects both the substantive content and the interaction quality of agent-customer conversations.

Part I: LDA-Based Measure of Problem-Centered Communication

This appendix details the methodology used for the Latent Dirichlet Allocation (LDA) analysis presented in Section 6.1. The process involved four main stages: data pre-processing, hyper-parameter tuning, topic interpretation using a Large Language Model (LLM), and the construction of our final measures.

1. Data Pre-processing

Before estimating the LDA model, we pre-processed the conversation data from our final sample of 3,965,028 sessions using standard text analysis procedures: (i) Focus on Agent Utterances: To capture service patterns from the agent's perspective, we first filtered the dataset to include only messages sent by agents. (ii) Word Segmentation: We utilized Jieba, an open-source Python package specifically designed for Chinese text processing, to segment the conversation transcripts into individual words. (iii) Stopword Removal: We employed a standard library of common stop-words and augmented it with a manually curated list of high-frequency, low-information words specific to our dataset (e.g., common greetings). (iv) Vocabulary Pruning: We removed words that appeared in fewer than five conversations or in more than 50% of the entire corpus to eliminate very rare or overly common terms. After these pre-processing steps, we trained our LDA model on the full set of 3,965,028 conversations.

2. Hyper-parameter Tuning

The key hyper-parameter for an LDA model is the number of topics (K). To identify the optimal K for our dataset, we used the topic coherence score, a measure that reflects the interpretability of the generated topics and correlates well with human judgment (Röder et al., 2015). We trained separate LDA models for K ranging from 5 to 60 and plotted the resulting coherence scores (Figure 7). The score peaks in the $K \in [20, 25]$ range, suggesting the optimal number of topics for our corpus. We ultimately selected $K = 25$, as this value yielded a high coherence score (approx. 0.6) and provided a sufficient number of topics to capture meaningful thematic variations in agent communication. The final model was fitted using

the Gensim package in Python with 100 passes to ensure convergence.

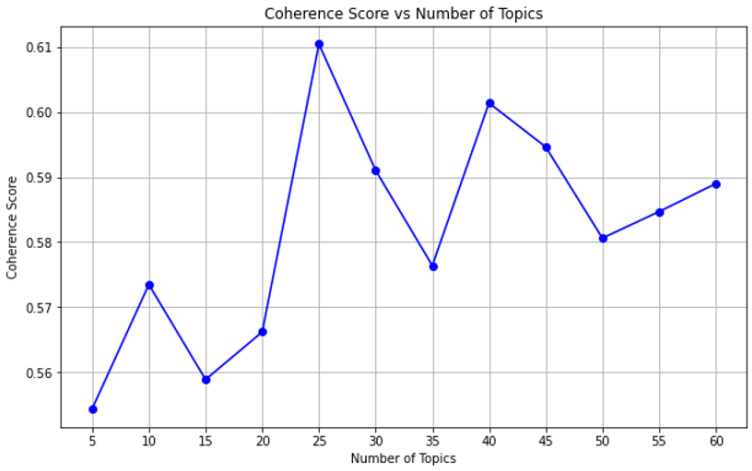


Figure 7: Coherence score for different number of topics

3. LLM-based Topic Interpretation and Measure Construction

A key challenge in LDA is the subjective interpretation of the generated topics. To address this in a systematic way, we leveraged a Large Language Model (LLM) to characterize the 25 topics identified by our model. We utilized DeepSeek-R1, an open-source LLM known for its strong performance in understanding Chinese context and executing reasoning tasks (DeepSeek-AI, 2025). We designed two distinct prompting approaches to create our final measures, using the default parameters of the LLM.

Approach 1: Topic Categorization To classify each topic, we provided the LLM with the following prompt:

User Prompt: "You are a customer service expert. Based on the following keywords for different chat topics identified by an LDA model, please classify each topic into one of two categories: (1) primarily focused on 'problem-solving' or (2) primarily focused on 'customer-comforting'. Please return the classification for each topic. [Insert LDA keywords for each of the 25 topics here]."

Illustrative Model Response: "Understood. Based on my expertise and the provided keywords, I will classify each topic as either 'problem-solving' or 'customer-comforting'."

Based on the LLM's binary classification for each topic, we then constructed our first measure, $ProblemCenteredRatio_1$. For each session, this ratio represents the proportion of the conversation's total topic relevance that is attributed to topics classified as "problem-solving oriented".

$$ProblemCenteredRatio_1 = \sum ProblemSolvingTopicRelevance \,/\, \sum TopicRelevance$$

where *Topic Relevance* is the relevance value for each topic obtained from LDA and *ProblemSolving Topic Relevance* is the relevance for topics categorized as "problem-solving oriented".

Approach 2: Topic Rating To obtain an alternative measure as a robustness check, we prompted the LLM to rate each topic on a continuous scale:

User Prompt: "You are a customer service expert. Based on the following keywords for different chat topics identified by an LDA model, please rate each topic on a scale of 1 to 5 according to how problem-solving oriented it is. A rating of 1 means 'least problem-solving oriented', and a rating of 5 means 'highly problem-solving oriented'. Please return the rating for each topic. [Insert LDA keywords for each of the 25 topics here]."

Illustrative Model Response: "Certainly. I will rate each topic on a scale of 1 to 5, indicating its degree of problem-solving orientation."

Based on the LLM's 1-to-5 rating for each topic, we constructed our second measure, $ProblemCenteredRatio_2$. For each session, this score represents the overall problem-solving orientation of the conversation, calculated as the relevance-weighted average of the problem-solving rating across all topics present in that session.

$$ProblemCenteredRatio_2 = \sum ProblemSolvingScore \cdot TopicRelevance \,/\, \sum TopicRelevance$$

where *Topic Relevance* is the relevance value for each topic obtained from LDA.

Table 14 and 15 present the 25 topics, their top keywords, and the resulting classification and rating

provided by the LLM for each approach. The binary classification and the 1-to-5 rating capture related but distinct judgments. A topic may be classified as problem-oriented in the binary approach while receiving a relatively low problem-solving score when its keywords contain both procedural/problem-related and relational elements.

Topic ID	Keywords	Topic Category	Problem Solving Score
0	coupon, usage, refund, expiration date, threshold, return, successful, check, taobao, expired, choose, information, distribute, rules,failed	problem-oriented	5
1	one-time, shop, coupon, pre-sale, deposit, only, remove, products,together, stackable,submit order, same, partial, friendly, final payment	problem-oriented	5
2	claim, account, benefits, taobao, link, click, member center, member, VIP, phone number, quark, Netease, Music, access, page	problem-oriented	5
3	report, brand, event, context, message, task, live-stream, video, rules, participate, comment, eligible, disclose, image	problem-oriented	5
4	shipping fee, door-to-door, pickup, freight cost, usage, vip, benefits, monthly, enjoy, claim, return, member, returns and exchanges, offset, failed	problem-oriented	5
5	feedback, automatic, renew, taobao, 2023, page, system, activate,money-saving, loyalty score, vip, taopiaopiao, event, access, various	problem-oriented	5
6	check, click, success, alipay, page, payment, details, operation, taobao, confirm, access, proceed, receipt, information, notification	problem-oriented	5
7	a moment, appreciate you effort, minute, 12, awaiting, check, verify, in progress, screenshot, check, assist, sorry to bother you, not have, immediately, maybe	comforting-oriented	3
8	sorry, caused, experience, really, platform, shopping, unpleasant, inconvenience, understanding, indeed, feeling, feedback, please don't be upset, wish, truly	comforting-oriented	1
9	agent, contact, service, thank you, wish you, cainiao, online, pleasant, life, support, goodbye, time, taobao, understanding, tmall	comforting-oriented	2
10	return, refund, seller, platform, request, reason, product, support, verify, provide, address, click, i want, express, intervention	problem-oriented	5
11	agent, hour, awaiting, exclusive, feedback, escalate, 24, verify, message, refund, patience, submit, member, platform, esteemed	problem-oriented	4
12	points, exchange, event, tmall, page, center, access, expired, 31, click here, shopping coupon, claim, benefits, member center, available	problem-oriented	5

Table 14: LDA topics

Topic ID	Keywords	Topic Category	Problem Solving Score
13	request, refund, seller, platform, after-sales, initiate, intervention, entrance, reject, exchange, click, notification, timely, follow-up, contact	problem-oriented	5
14	product, we suggest that, order, try, time, platform, purchase, exists, unable to, transaction, notification, verify, account, behavior, current	problem-oriented	5
15	shipping fee, cover, evidence, submit, upload, request, follow up, the agent will, seller, promise, product, freight cost, guarantee, seller, member	problem-oriented	5
16	merchant, contact, negotiate, feedback, unable, refund, we suggest you, 24, not have, promise, cancel, communicate, transaction, reply,a moment	problem-oriented	5
17	vip, benefits, shopping, enjoy, activate, life, supermarket, membership, daily, comprehensive, member center, global, access, youku, tmall	problem-oriented	4
18	subsidy, ten billion, product, order, purchase, the same, unable, customer, discount, address, partial, enjoy, channel, notification, case	problem-oriented	5
19	benefits, claim, member, youku, vip, period, expatriation date, expire, anew, current, already, Mango TV, enjoy, link, renew	problem-oriented	5
20	taobao, event, double 11, double 12, time, check, homepage, 18, search, access, engage, during, 30, second floor, interaction	problem-oriented	5
21	member, benefits, vip, request, cancel membership, continue, acknowledge, activate, enjoy, usage, amount, submit, unable, whether or not, certainly	problem-oriented	4
22	ship, complaint, seller, initiate, platform, check, i want to, compensation, click, time, proceed, request, merchant, review, progress	problem-oriented	2
23	product, usage, discount, payment, shall prevail, consumption, amount, calculate, tag, featuring, whether or not, page, 23:59:59, coins, support	problem-oriented	2
24	switch, plan,member, current, benefits, vip, 10, change,period, comprehensive, support, expiration date, expires, claim, price difference	problem-oriented	2

Table 15: LDA topics

Part II: Proprietary LLM-Based Measures of Interaction Quality

This section describes the proprietary LLM used to generate the interaction-quality measures in Section 6.2. Unlike the LDA-based measures, which capture the thematic content of agent communication, the proprietary LLM evaluates the quality of the service interaction from the customer's perspective. The model scores each conversation on three dimensions: Emotion Management, Communication, and Overall Approval.

1. Model Development and Validation

Capitalizing on recent advancements in LLMs (Yoganarasimhan and Iakovetskaia, 2024; Niu et al., 2025; de Kok, 2025; Siano, 2025; Zhang and Narayandas, 2025), the focal platform developed a proprietary model to assess service quality with greater nuance than traditional metrics. The development process involved a multi-stage, state-of-the-art approach: (i) Base Model: The initiative began with a powerful open-source foundation model, Qwen2-7B (Yang et al., 2024; Team, 2024). (ii) Evaluation Framework Design: A multi-dimensional framework for evaluating service quality was established, comprising three key metrics: Emotion Management (the agent's effectiveness in transforming a customer's emotion from negative to positive), Communication (the agent's ability to understand the customer and convey information with clarity), and Overall Approval (a general assessment of the customer's approval of the agent). (iii) "Teacher" Data Generation: A dataset of 10,000 service sessions was annotated on these three dimensions by two then-state-of-the-art "teacher" LLMs (GPT-4 and Qwen Max), providing a robust set of initial labels. (iv) Fine-tuning via Knowledge Distillation: The base model (Qwen2-7B) was then fine-tuned on this labeled data using knowledge distillation techniques (Hinton et al., 2015), effectively transferring the nuanced judgment capabilities of the larger models to the more efficient "student" model. (v) Human Alignment and Validation: Finally, the fine-tuned model's performance was validated and aligned with human judgment through a "chat arena" process involving experienced human agents. The resulting proprietary LLM achieved a precision and recall of approximately 90% when compared against human ratings on these dimensions according to the platform's internal report. This proprietary LLM has been available for internal use since September 2024.

2. Application for Research Data Generation

Ideally, we would use this validated proprietary LLM to measure all of our chat sessions. However, that analysis would consume a significant amount of computational resources and time. Given these computational constraints, we applied this LLM to a large, random sub-sample of our data. The process was as

follows: (i) We first randomly sampled 600,000 sessions from our original dataset. (ii) We then used the proprietary LLM to score each of these sampled sessions on the three predefined dimensions on a scale from 1 (lowest) to 10 (highest). This process yielded a final dataset of 594,156 sessions with complete LLM-generated scores.

The exact internal prompt is confidential. However, its structure and core instructions mirror the illustrative example below, which defines "Good" and "Poor" performance for each dimension before asking the model to evaluate a given dialogue.

Prompt: "You are a Consumer Experience Officer. Your task is to analyze the provided conversation transcript and evaluate the overall service experience based on the following quality standards, from a consumer's perspective. Please provide a score from 1 to 10 for each dimension (1 being the poorest performance, 10 being the best).

1. Emotion Management:

Poor : The service made me very angry; the agent did not ease my negative emotions or even exacerbated them.

Good: My emotional state changed from negative to positive during or after the interaction with the agent.

2. Communication:

Poor: The agent did not understand my questions/requests, answered irrelevantly, repeatedly asked for clarification, used difficult-to-understand language, or was mechanical/rigid (e.g., repeatedly sending policies without addressing the issue).

Good: Communication was smooth; I felt understood without needing excessive explanation from my side.

3. Overall Approval:

Poor: I do not approve of this agent and doubt their work quality/ability.

Good: I approve of this agent and might even express gratitude.

Dialogue Record:

{dialog_text}

Please evaluate the dialogue based on these standards and provide your scores."

Illustrative Model Response: "Certainly. I will analyze the dialogue and return the ratings for each dimension."

Appendix E: Heckman Selection Model for Customer Ratings

Customer ratings are observed only when customers choose to provide post-service feedback. Because the rating response rate in our data is approximately 16.40%, the observed rating sample may be subject to self-selection bias. To examine whether our customer-rating results are robust to this concern, we estimate a Heckman selection model (Heckman, 1974, 1976). This analysis complements the control-function approach used in the main mediation model: the control-function approach addresses endogeneity in workload and front ratio, whereas the Heckman correction addresses nonrandom rating response.

In the first stage, we estimate a Probit model for whether customer i provides a rating after the session handled by agent j. We then calculate the inverse Mills ratio (IMR), denoted by λ_{ij}, and include it in the second-stage customer-rating regression. The IMR controls for the nonrandom selection process through which customer ratings are observed.

We model the rating decision as a function of two types of variables: the customer's historical behavior and the current service experience. Historical behavior includes the number of prior chat sessions and the customer's historical rating response rate Ravn et al. (2006); Netzer et al. (2008); Armona et al. (2024). Current service experience includes session duration, average agent response delay, average customer response delay, average agent message length, average customer message length, total interaction rounds, and agent fatigue, consistent with prior literature (Goes et al., 2017). We also control for customer demographic characteristics, including VIP status and months since registration.

In particular, we employ the following first-stage selection Probit model:

$$IfRate_{ij} = \begin{cases} 1 & IfRate^*_{ij} > 0 \\ 0 & otherwise \end{cases} \tag{8}$$

$$IfRate^*_{ij} = \alpha_0 + \beta_1 H_{ij} + \beta_2 S_{ij} + \gamma X_{ij} + \kappa_i + \theta_j + \varepsilon_{ij} \tag{9}$$

where H_{ij} is a vector of variables representing the customer's historical behavior, including the number of previous chat sessions they had with the customer service department on the platform and their historical rating response rate. S_{ij} is a vector capturing the customer's current service experience, including session duration, average agent response delay, average user response delay, average agent message length (in words), average customer message length (in words), total interaction rounds, and agent fatigue level. X_{ij} is a vector of customer demographic variables, such as VIP status and months since registration. In the second stage, we re-estimate the model of customer rating by including the IMR as an additional control variable. This inclusion aims to correct for the selection bias.

For identification, the selection equation (9) should include at least one variable that affects the probability of observing a rating (i.e., relevance condition) but does not directly affect the rating score conditional on the current service experience and controls (i.e., exclusion restriction condition) (Heckman, 1976). We use the customer's historical rating response rate as the exclusion variable. The logic is that a customer's past tendency to provide ratings should strongly predict whether they rate the current session, but conditional on the current service experience, agent fixed effects, problem fixed effects, and other controls, it should not directly determine the satisfaction score for the current interaction. This exclusion restriction is an identifying assumption, and the Heckman analysis should therefore be interpreted as a robustness check.

Table 16 reports the results. The coefficient of the IMR (λ) is statistically significant, indicating nonrandom selection into the observed rating sample. More importantly, similar to our main results, the coefficient for the front ratio remains positive and significant, while the quadratic term remains negative and significant.

Thus, the inverted-U-shaped relationship between front ratio and customer ratings is robust to correcting for rating-selection bias.

Table 16: Heckman Selection Correction for Customer Ratings

Dependent Variable	*CustomerRating*	
	Main CF Model (1)	CF Model+Heckman Correction (2)
FrontRatio	1.047**	1.079**
	(0.379)	(0.378)
$FrontRatio^2$	-0.808*	-0.832*
	(0.380)	(0.379)
Workload	0.078*	0.081*
	(0.036)	(0.036)
$Workload^2$	-0.174***	-0.181***
	(0.033)	(0.033)
λ		2.128***
		(0.038)
Fatigue	-0.001	0.000
	(0.002)	(0.002)
IsFirst	0.651***	0.611***
	(0.005)	(0.005)
Log(*AvgCustomerMessageLength*)	-0.491***	-0.441***
	(0.016)	(0.016)
Log(*AvgCustomerResponseDelay*)	-0.137***	-0.157***
	(0.014)	(0.014)
*ResWorkload*1	-0.251***	-0.264***
	(0.037)	(0.037)
*ResWorkload*2	0.403***	0.405***
	(0.034)	(0.034)
*ResFrontRatio*1	-1.617***	-1.553***
	(0.380)	(0.378)
*ResFrontRatio*2	1.193**	1.136**
	(0.381)	(0.379)
Agent Effects	Yes	Yes
Problem Effects	Yes	Yes
Day Effects	Yes	Yes
Hour Effects	Yes	Yes
N	650,660	650,660
$Adj.R^2$	0.164	0.170

Notes. Robust standard errors clustered at the agent level in parentheses.

† $p < 0.1$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix F: Distributions and Nonlinear Mediation Decomposition

Appendix F provides visual evidence supporting the empirical analysis. Figure 8 shows the distributions of the two main explanatory variables, workload and front ratio. Figure 9 visualizes the nonlinear mediation decomposition used in Section 5.2.2 by separating the marginal effect of workload into a direct component and an indirect component operating through the front ratio.

F.1 Distributions of Main Variables

Figure 8 presents the distributions of the two main variables in our analysis: workload and front ratio. Workload is measured as the time-averaged number of concurrent sessions handled by the focal agent during the focal session. Front ratio is measured as the proportion of active service effort allocated to front-office customer interaction. These distributions show substantial variation in both workload pressure and within-session effort allocation, supporting the empirical analysis of how workload affects agents' front-office/back-office balance.

(a) Distribution of Workload

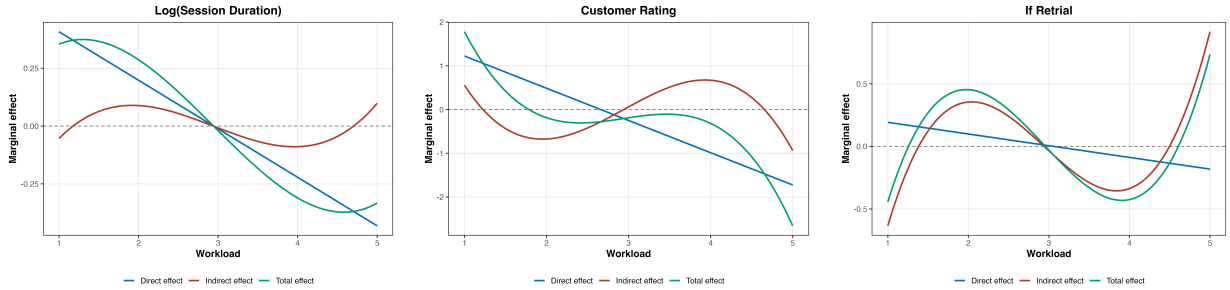
(b) Distribution of Front Ratio

Figure 8: Distributions of *Workload* and *FrontRatio*

F.2 Visualization of Direct and Indirect Effects

Figure 9 visualizes the marginal-effect decomposition used in the mediation analysis. Because workload and front ratio enter the empirical models non-linearly, the mediated effect varies across workload levels. We therefore decompose the marginal effect of workload on each service outcome into a direct component

and an indirect component operating through the front ratio. The three panels show this decomposition for session duration, customer rating, and retrial rate, respectively.



(a) Session Duration (b) Customer Rating (c) Retrial Rate

Figure 9: Decomposition of Workload Effects into Direct and Front-Ratio-Mediated Components