Generative AI and Firm Productivity: Field Experiments in Online Retail[*]

Lu Fang[†] Zhe Yuan[‡] Kaifu Zhang[§]

Dante Donati[¶] Miklos Sarvary[‖]

October 10, 2025

Preliminary version, subject to change

Abstract

We quantify the impact of Generative Artificial Intelligence (GenAI) on firm productivity through a series of large-scale randomized field experiments involving millions of users and products at a leading cross-border online retail platform. Over six months in 2023-2024, GenAI-based enhancements were integrated into seven consumer-facing business workflows. We find that GenAI adoption significantly increases sales, with treatment effects ranging from 0% to 16.3%, depending on GenAI's marginal contribution relative to existing firm practices. Because inputs and prices were held constant across experimental arms, these gains map directly into total factor productivity improvements. Across the four GenAI applications with positive effects, the implied annual incremental value is approximately $5 per consumer—an economically meaningful impact given the retailer's scale and the early stage of GenAI adoption. The primary mechanism operates through higher conversion rates, consistent with GenAI reducing frictions in the marketplace and improving consumer experience. We also document substantial heterogeneity: smaller and newer sellers, as well as less experienced consumers, exhibit disproportionately larger gains. Our findings provide novel, large-scale causal evidence on the productivity effects of GenAI in online retail, highlighting both its immediate value and broader potential.

Keywords: Field Experiments, Generative AI, Productivity, Retail Platforms, Consumer Experience

JEL codes: C93, D24, L81, M31, O3

[*] The authors thank Andrey Simonov for valuable feedback, and participants at the TUM Workshop on Generative AI in Marketing, UCL School of Management seminar, Columbia Business School Seminar, the Business & Generative AI Conference 2025, and the School of Management Seminar at Zhejiang University, for helpful comments and discussions. The authors thank the anonymous e-commerce platform for its support of this project and are grateful for the valuable feedback provided by several industry colleagues. The paper only reflects the opinion of the authors, and not the opinion of any organizations. All rights reserved. All remaining errors are the responsibility of the authors. Lu Fang acknowledges financial support from the National Natural Science Foundation of China [Grants 72501258 and 72192803], the National Social Science Fund of China [Grant 22&ZD081]. Zhe Yuan acknowledges financial support from the National Key Research and Development Project [Grant 2024YFB3312900], the National Natural Science Foundation of China [Grants 72192803, 72141305 and 72203202], the Fundamental Research Funds for the Central Universities. Kaifu Zhang, Dante Donati, and Miklos Sarvary have no funding to report.

[†] Zhejiang University of Finance & Economics, fl_fanglu@zufe.edu.cn

[‡] Corresponding author, Zhejiang University, yyyuanzhe@gmail.com

[§] zhangkaifu314@gmail.com

[¶] Columbia University, dd3137@gsb.columbia.edu

[‖] Columbia University, miklos.sarvary@columbia.edu

"*The next industrial revolution has begun*," Nvidia Chief Executive Officer Jensen Huang said. *"AI will bring significant productivity gains to nearly every industry and help companies be more cost- and energy-efficient, while expanding revenue opportunities." Nvidia Stock Surges as Sales Forecast Delivers on AI Hopes*, Bloomberg, May 22, 2024.

1 Introduction

The rapid diffusion of Generative Artificial Intelligence (GenAI) tools has sparked growing interest in their potential to reshape productivity across sectors of the economy. Recent academic research has provided compelling evidence of GenAI's promise in various domains, including software development, customer support, education, and professional services (e.g., Brynjolfsson et al., 2025; Noy and Zhang, 2023; Peng et al., 2023; Eloundou et al., 2023). Yet, despite rapid adoption, there is little empirical evidence of measurable gains in aggregate or firm-level revenue-based productivity attributable to GenAI (Acemoglu, 2025). Similarly, investors and industry practitioners have raised concerns about whether massive AI investments will translate into sustained business returns.[1]

Identifying the firm-level productivity impact of GenAI poses three empirical challenges. First, constraints in technical expertise and the need for complementary investments may delay implementation and the realization of observable gains, even when long-run potential exists (Bonney et al., 2024).[2] Second, most existing applications of GenAI in firms remain at the pilot stage and focus on narrowly defined tasks—often at the worker level—making it difficult to detect productivity gains in aggregate firm-level data.[3] This narrow scope overlooks a central insight from the economics literature: value creation stems not from isolated tasks, but from interdependent routines—or workflows—which play a central role in driving firm productivity (e.g., Bloom and Van Reenen, 2007). Third, rigorous empirical analysis requires detailed revenue data and a setting that enables causal identification, both of which are rarely available.

This paper provides large-scale, real-world experimental evidence on the causal impact of GenAI on productivity at both the firm and workflow levels, using data from one of the world's largest cross-border online retail platforms. Over six months in 2023-2024, the platform integrated GenAI into seven consumer-facing business workflows—ranging from search query refinement to product description generation. In each workflow, GenAI augmented existing technologies with minimal or no displacement of labor and capital, ensuring that any observed changes in output reflect genuine productivity gains.[4] Each application was evaluated through randomized field experiments, with user groups ranging from tens of thousands to tens of millions. Leveraging granular consumer/product-level data, the experiments assess the short-term impact of GenAI on key performance outcomes such as sales (in dollar values) and conversion rates, allowing us to identify not only whether GenAI delivers measurable business outcomes and boosts productivity, but also where, how, and for whom those gains materialize.

[1] See, for example, recent articles by Sequoia Capital: www.sequoiacap.com/article/ais-600b-question; and The Economist: www.economist.com/leaders/2025/09/11/what-if-the-3trn-ai-investment-boom-goes-wrong.

[2] For example, gains from technological innovations often take time to materialize (Brynjolfsson and Hitt, 2003).

[3] For instance, recent studies have analyzed the effects of GenAI chatbots on workers' performance (Dell'Acqua et al., 2023; Otis et al., 2024), and earnings (Humlum and Vestergaard, 2025). See also: www.wsj.com/articles/companies-are-struggling-to-drive-a-return-on-ai-it-doesnt-have-to-be-that-way.

[4] In line with the classical Solow model (Solow, 1957), throughout the paper we interpret increases in output holding labor and capital constant as gains in total factor productivity.

We document three main findings. First, most GenAI deployments generate economically significant gains, though the effects vary across workflows—from no detectable impact to increases of up to 16.3% in sales, with the largest improvements observed in customer service and search applications. Because output rose while labor and capital inputs remained constant, these improvements map directly into total factor productivity (TFP) gains of comparable magnitude. Aggregating across the four GenAI applications with positive effects, we estimate an annual incremental value of approximately $5 per consumer. These impacts, observed both within and across workflows, are substantial given the retailer's scale and the early stage of GenAI adoption. Second, the productivity gains arise from GenAI's ability to reduce frictions in the online marketplace and increase purchase intention. Across workflows, we observe significantly higher conversion rates and no effect on average cart values, consistent with GenAI enhancing the consumer experience and inducing market expansion. Third, the effects are heterogeneous across seller and buyer segments: on both the demand and supply sides, less experienced and smaller buyers and sellers derive disproportionately larger gains from GenAI enhancements. Our setting represents markets across a diverse set of countries, languages, and cultures. This global scope makes the findings broadly generalizable to the online retail industry.

A key distinction of our study relative to existing GenAI literature is its focus on revenue-based outcomes. In contrast to prior work emphasizing input-side efficiency gains—such as improvements in worker performance (e.g., Brynjolfsson et al., 2025; Dell'Acqua et al., 2023; Peng et al., 2023)—we show that firm-level adoption of GenAI can enhance productivity through demand-side value creation. In our setting, the estimated gains stem entirely from higher sales: across most experiments, workflow costs remained unchanged and GenAI deployment did not alter the factors of production.[5] Our estimates therefore represent a conservative lower bound on the potential returns on investments in GenAI. The observed improvements reflect enhanced consumer experience through GenAI-driven reductions in market frictions, as evidenced by higher conversion rates. These results are consistent with theories emphasizing that in consumer-facing sectors, quality improvements—such as product innovation, personalization, and relevance—can raise productivity without lowering input costs (Syverson, 2011; Brynjolfsson and Hitt, 2003; De Loecker, 2011).

Moreover, prior studies of GenAI have primarily examined its impact on individual-level productivity in narrowly defined tasks conducted in laboratory settings. Such isolated applications in diverse contexts make it difficult to compare results due to differences in implementation quality and organizational practices (see Calvino et al., 2025 for a review). Moreover, productivity dynamics in lab settings may not capture the complexities of real-world firm adoption, where technical, organizational, and market factors interact in intricate ways (Brynjolfsson et al., 2025; Microsoft, 2024). In this study, we analyze the large-scale deployment of GenAI across multiple workflows—varying in business function and complexity—within a single firm. This setting allows us to assess GenAI's impact on firm-level productivity, a dimension largely unexplored in prior work. By holding implementation and organizational factors constant, we can attribute productivity differences across workflows to how effectively GenAI augments each application's baseline performance.

We partnered with a world-leading cross-border e-commerce platform to identify and quantify the sources of GenAI-driven productivity gains in online retail. The platform enables consumers worldwide to buy directly from manufacturers at competitive prices.[6] Between September 2023 and

[5] Section 3.3 provides details on the constant input structure of our experiments.

[6] The platform connects hundreds of thousands of predominantly small-business sellers with hundreds of millions

June 2024, the firm deployed GenAI solutions across seven consumer-facing business workflows: (1) Pre-sale Service Chatbot, (2) Search Query Refinement, (3) Product Description Generation, (4) Marketing Push Message Creation, (5) Google Advertising Title Optimization, (6) Chargeback Defense, and (7) Live Chat Translation. These correspond to three broader functional areas: (i) consumer and seller services (1, 6, 7); (ii) consumer–product matching (2, 3); and (iii) advertising and promotion (4, 5). Each workflow maps to a distinct stage of the customer journey, allowing us to assess GenAI's productivity impact across a wide range of retail operations.

Our setting involves multiple experiments applying GenAI to distinct business workflows. Several features enhance comparability across these experiments: all applications were developed and deployed by the same technical team—covering algorithm design, model fine-tuning, and online rollout—and operated within the same firm, under similar organizational and competitive conditions. Despite this common implementation environment, the effect of GenAI on workflow performance is likely to depend on the scope of its marginal contribution relative to baseline conditions, which differ across workflows. In each case, the control group reflects the firm's standard practices prior to GenAI adoption. For instance, in the Pre-Sale Service Chatbot, the control group received no customer support; in the Search Query Refinement, the search function in the control group relied on standard machine-learning-based search algorithms; and in most other workflows, the benchmark was human input.

Within each workflow, GenAI deployment was evaluated through a large-scale randomized field experiment, comparing the GenAI-enhanced workflow to a baseline version used as the control. Notably, baseline workflows often included automation or human input but did not incorporate GenAI technologies. The treatment condition differed solely in the integration of GenAI, while prices and costs remained constant across conditions. Randomization occurred at the level of consumers—and in one case, products—with minimal overlap (less than 1%) across experiments. For five of the seven experiments, we obtained detailed consumer/product-level transaction data, including expenditure, conversions, and clicks. For analysis, we aggregate these data to the consumer level (and to the product level in one experiment) and leverage consumer, seller, and product characteristics to study treatment effect heterogeneity. Our primary outcome is sales value, measured by total consumer expenditure, which—given fixed prices and workflow costs in our setting—serves as a proxy for revenue-based productivity. We also examine conversion rates, a standard performance metric in online retail that captures consumer experience.

Our results reveal that most GenAI deployments yield economically significant short-term productivity gains, though magnitudes differ across workflows. Among the five processes with detailed data, gains in sales range from no detectable effect in the advertising workflows to improvements of up to 16.3% in the Pre-Sale Chatbot application, consistent with prior research on GenAI's impact on individual tasks in lab settings (see, e.g., Peng et al., 2023). In the Search Query Refinement and Product Description workflows, the effects are smaller—generally 2-3%—yet still substantial for a platform of this scale and maturity. Back-of-the-envelope calculations based on the four deployments with detailed transaction data and positive effects—annualizing workflow-specific gains and assuming linear additivity—suggest that these GenAI applications generate an annual incremental value of approximately \$4.6–\$5 per consumer. These effects represent roughly 5.5–6% of the increase in per-user revenue observed in global e-commerce between 2023 and 2024. We also

of active buyers across more than 100 countries and regions. The platform supports around 20 languages, providing localized services that facilitate global communication and accessibility.

document notable improvements in workflows without granular data: a 15% higher success rate in Chargeback Defense and a 5.2% increase in consumer satisfaction from Live Chat Translation.

Taken together, these results show that GenAI generates sizable gains in targeted workflows and meaningful effects for a large and mature retailer, with further potential as adoption broadens and increasingly targets revenue-critical processes. For instance, while in 2023 the platform applied GenAI to only a handful of workflows, by 2024 it had expanded to more than 40 applications and by 2025 to over 60. At the same time, API calls to large language models increased twentyfold between 2024 and 2025, reflecting the rapid scaling of GenAI adoption across the platform. The long-run impact will ultimately depend on equilibrium forces—whether complementarities across workflows amplify these gains or industry-wide adoption offsets them through intensified competition.

We further shed light on the mechanisms underlying our results. We find that productivity improvements stem from enhanced consumer experience through GenAI-driven reductions in market frictions. Specifically, increased sales were strongly associated with higher conversion rates (extensive margins) and, where applicable, click-through rates, but not with higher average cart values (intensive margins). These conversion and engagement metrics increased by 1–22% across workflows, suggesting that GenAI applications contributed to productivity primarily through the creation of consumer value—most notably by improving the platform's user experience and expanding the market rather than increasing spending per buyer. In particular, better pre-sale chatbots and richer product descriptions reduced information asymmetries; GenAI-refined queries lowered search frictions, and automated push messages closed gaps in content personalization. While platforms already alleviate such frictions (Belleflamme and Peitz, 2021), our results show that GenAI adoption can further mitigate them.

Finally, we analyze heterogeneity in treatment effects across sellers, buyers, and products. If GenAI primarily reduces frictions on both the demand and supply sides, larger gains should arise among participants with lower baseline capabilities—smaller and less experienced sellers, buyers with limited platform engagement, and products in the long tail or in less concentrated categories. Consistent with this view, we find that productivity gains are more pronounced for smaller and less experienced sellers, in line with recent evidence that GenAI particularly benefits users with lower baseline skills (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023). On the demand side, less experienced consumers also benefit disproportionately. Overall, these patterns suggest that GenAI enhances the platform's ability to reduce frictions between buyers and sellers, narrowing outcome gaps across participants. By contrast, the effects across product groups are more context-dependent.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature. Section 3 describes the context, theoretical framework, empirical framework, and data. Section 4 presents the aggregate results, followed by Section 5, which explores the heterogeneity of treatment effects. Section 6 concludes. The Appendix provides additional details and results.

2 Contribution to the Literature

2.1 The Economic Impact of Generative AI

Recent advances in GenAI have attracted considerable attention for their potential economic and social implications. A growing body of research documents GenAI's ability to enhance individual productivity in simple and well-defined tasks, including mid-level writing (Noy and Zhang, 2023), software development (Peng et al., 2023; Cui et al., 2024), marketing copy generation (Dell'Acqua et al., 2023), and legal analysis (Choi et al., 2023). However, this literature largely focuses on individual-level impacts in laboratory settings, with limited empirical evidence of measurable productivity gains at the aggregate or firm level (Acemoglu, 2025; Calvino et al., 2025). In fact, the productivity dynamics observed in real-world firm adoption are likely more complex than those captured in lab environments due to technical, organizational, and market factors (Brynjolfsson et al., 2025; Microsoft, 2024).

More importantly, most prior studies assess the productivity potential of GenAI from a supply-side perspective, emphasizing labor savings or improvements in worker efficiency, typically measured by decreases in average task completion time or increases in the average number of completed tasks (e.g., Noy and Zhang, 2023; Dell'Acqua et al., 2023; Peng et al., 2023; Heller and Asam, 2024). Even the limited studies employing real-world firm data, such as Brynjolfsson et al. (2025) and Microsoft (2024), largely rely on similar supply-side metrics. By contrast, very few studies have examined GenAI's productivity effects through demand-side value creation, such as enhanced consumer experience and increased purchases.[7]

Furthermore, another important question concerns the asymmetric effects of GenAI across different user groups. Earlier waves of technological change, such as the adoption of computers and the Internet, were often described as "skill-biased" (Goldin and Katz, 2008), disproportionately favoring skilled users while leaving unskilled workers behind (Bresnahan et al., 2002; Autor et al., 2003; Bartel et al., 2007; Acemoglu and Restrepo, 2018). As of now, the heterogeneous impacts of GenAI on worker performance appear more context-dependent. On one hand, by lowering skill barriers, GenAI can promote inclusivity (Nguyen and Nadi, 2022; Eloundou et al., 2023; Chui et al., 2023), as it benefits users with lower levels of skills and expertise (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Peng et al., 2023; Hui et al., 2024). On the other hand, evidence from other studies points to the opposite outcome (Roldán-Monés, 2024; Otis et al., 2024).

Our paper aims to advance the understanding of how firm-level adoption of GenAI translates into tangible consumer value and measurable business outcomes in a real-world context. Particularly, we investigate how productivity gains emerge through enhanced consumer experiences while holding inputs constant. Leveraging field experiments in a leading global cross-border e-commerce platform, we offer a more comprehensive and nuanced view that complements and extends the predominantly supply-side focus of prior productivity research. In addition, using detailed data in the context of online retail platforms, we broaden the heterogeneity analysis of GenAI by exploring how its effects vary across seller and buyer groups with different levels of experience, addressing a notable gap in the existing literature.

2.2 Friction Reduction in Online Marketplaces

Our paper also contributes to the research on how technological innovations and market designs help reduce various forms of frictions in online marketplaces.

A key friction in online marketplaces is information asymmetry: buyers often cannot directly verify product quality or seller reliability prior to purchase (Jin and Kato, 2006; Tadelis, 2016).

[7] Chen and Chan (2024), Exner et al. (2025), Kapoor and Kumar (2025), and Hartmann et al. (2025) show that AI-generated ad copies or images in digital advertising can raise click-through rates, but they provide no evidence on actual purchase behavior.

To mitigate this challenge and facilitate consumer decision-making, platforms have traditionally relied on reputation and review systems to generate quality signals (Cabral and Hortaçsu, 2010; Donati, 2025; Fan et al., 2016; Wang et al., 2024). However, these feedback mechanisms suffer from well-documented limitations, including grade inflation (Nosko and Tadelis, 2015; Zervas et al., 2015), "cold start" problems (Bai et al., 2022), and score manipulation (Mayzlin et al., 2014; Luca and Zervas, 2016). Modern platforms are increasingly adopting advanced technologies and innovative market designs. For instance, AI models based on natural language processing are used to extract insights from textual reviews or to filter feedback for relevance, enabling more accurate inferences about product quality and consumer satisfaction (Milgrom and Tadelis, 2018; Li et al., 2020). Moreover, curated provision of off-site social media information during consumer search has been shown to assist consumer decision-making and enhance platform revenues (Ghose et al., 2019).

Consumers on digital platforms also face substantial frictions during online search, arising either from search costs, the effort and resources required to locate information, or search targetability, the effectiveness of search engines in retrieving the most relevant products. Such frictions can significantly influence market outcomes and market structures in online marketplaces (Ghose et al., 2014; Honka, 2014; Yang, 2013; Brynjolfsson and Smith, 2000; Brynjolfsson et al., 2011; Bar-Isaac et al., 2012). To mitigate search frictions, digital platforms have invested heavily in technology and design innovations aimed at encouraging consumer search behavior, improving match value, and enhancing consumer welfare. Prior research highlights several such advances in online search, including ranking algorithms (Dinerstein et al., 2018; Ursu, 2018; Yang et al., 2024), refinement tools like sorting and filtering (Chen and Yao, 2017; Fradkin, 2017), machine-learning-driven personalized search (Yoganarasimhan, 2020), category-refinement-based precision improvements (Zhou et al., 2025), and optimal search engine information layout (Gu and Wang, 2022).

In online marketplaces, advances in personalization and targeting technologies can also help reduce frictions by delivering products or content to users most likely to be interested. Bergemann and Bonatti (2011) develop theoretical models showing that improving advertisers' targeting ability increases the number of consumer-product matches, thus enhancing the overall social value of advertising. Empirically, targeted advertising has been found more effective than untargeted approaches. For instance, Goldfarb and Tucker (2011) observe that display ads matched to website context significantly raise purchase intent, while Blake et al. (2015) suggest that targeted ads are particularly valuable in lowering search friction when consumers would otherwise have difficulty discovering or learning about products. Extending these insights to recommendation systems, Sun et al. (2024) demonstrate that removing personalized recommendations discourages consumer search and purchasing, especially for small sellers and niche consumers.

Our paper builds on this literature by extending it to the emerging technological wave of generative AI. Leveraging seven large-scale randomized field experiments conducted across three major business functions on a leading cross-border online retail platform, we provide evidence on how GenAI can be utilized to reengineer multiple workflows, reduce different types of frictions, and ultimately motivate demand-side value creation through consumer experience enhancement. A comprehensive heterogeneity analysis further allows us to explore how GenAI-driven friction reduction disproportionately impacts distinct user groups.

3 Study Setting, Experimental Design and Data

3.1 Context

The seven field experiments analyzed in this paper were conducted over roughly six months, from September 2023 to June 2024. The company's GenAI initiatives, however, began earlier in 2023, with initial efforts devoted to model training, strategy formulation, and experimental preparation. The selection of workflows for GenAI reengineering was not systematic but instead reflected managerial judgment, with platform managers prioritizing those considered most promising in terms of technical feasibility, organizational costs, and potential productivity gains. The selected workflows cover several core modules of e-commerce operations, including customer service, consumer-product matching, advertising, and seller service. Table 1 provides a concise overview of these workflows, highlighting the associated business needs/objectives and the modifications implemented using GenAI. Generally, these deployments did not change labor or capital inputs, with only one minor exception discussed below.

Table 1: Business Workflows Re-engineered with Generative AI

	Functional Area	Business Workflow	Business Needs/Objectives	GenAI Capability	Description of GenAI Application
1	Customer Service	Pre-Sale Service Chatbot	Addressing each individual service request, providing unique, accurate, and content-rich answers.	AI agent	Deploying a GenAI-powered, 24/7 customer service chatbot that can respond to idiosyncratic consumer inquiries in all languages.
2	Consumer-product Matching	Search Query Refinement	Accurately decoding and translating the latent demands behind multilingual consumer search queries to improve consumer-product match.	Translation, content comprehension and generation	Using GenAI to improve consumers' demand expression by understanding, refining and translating their search queries, thus enhancing the matching accuracy of the search algorithm.
3	Consumer-product Matching	Product Description	Creating comprehensive, structured product descriptions tailored to diverse linguistic preferences and cultural norms (e.g. currently, nearly half of the self-sold products have no or limited description).	Content recognition, comprehension and generation	Using GenAI to produce comprehensive and structured textual descriptions for the product detail page's description module, adapted to each market.
4	Advertising	Marketing Push Message	Individual targeting of hundreds of millions of users with customized messages.	Content comprehension and generation	GenAI allows the generation of millions of messages, thereby enhancing the personalization of messages for precision marketing.
5	Advertising	Google Advertising Title	Creating product advertisement titles that closely match user interest and demands.	Content optimization and generation	Using GenAI to optimize product titles for Google ads for better user interest and engagement.
6	Seller Service	Chargeback Defense	Streamlining the complicated process in a cross-boarder context with language barriers and diverse regulations and customs (e.g. over half of chargeback disputes go unaddressed by sellers).	AI agent	Developing a GenAI-driven agent that offers a one-stop, automated solution for sellers to streamline the intricacies of chargeback defense.
7	Customer Service	Live Chat Translation	Delivering native-language customer services to a diverse, multilingual consumer base	Real-time translation	Integrating GenAI into the platform's core English customer service process to provide real-time translation for all languages.

3.2 Theoretical Framework

We model the impact of GenAI adoption on firm productivity through the lens of the standard Solow growth model (Solow, 1957). Assume that output is produced according to a Cobb–Douglas production function

$$Y = AK^{\alpha}L^{1-\alpha}, \qquad 0 < \alpha < 1, \tag{1}$$

where Y denotes output, K is the capital stock, L is labor input, and A is total factor productivity (TFP). This simple framework is particularly well-suited to our context of the online retail industry. Retail platforms operate at large scale, where marginal costs of digital operations are negligible, and productivity improvements often take the form of efficiency gains in existing processes (e.g., faster and higher-quality product page generation) rather than through large expansions of labor or capital. The Cobb–Douglas formulation, with TFP as a residual capturing efficiency, provides a natural and tractable way to interpret observed output changes in terms of underlying productivity shocks.

Specifically, differentiating (1) in logs yields the standard growth-accounting decomposition:

$$d\ln Y \;=\; d\ln A \;+\; \alpha\, d\ln K \;+\; (1-\alpha)\, d\ln L.$$

In this framework, changes in output can arise from (i) capital deepening, (ii) growth in labor input, or (iii) growth in TFP. Our focus is on GenAI adoption in business processes, where the technology primarily enhances the quality/efficiency of existing inputs rather than expanding them. In such settings, the additional gains from GenAI adoption can be interpreted as a shift in A, rather than as an increase in K or L.

Formally, if capital and labor inputs are held constant when the firm adopts GenAI, then

$$d\ln K = 0, \quad d\ln L = 0 \quad \Rightarrow \quad d\ln Y = d\ln A.$$

Under these conditions, any observed increase in output maps one-to-one into measured TFP growth. For clarity, this identification relies on the following assumptions:

1. **No capital deepening:** Although the platform trains and deploys its own GenAI models, these exhibit strong non-rivalry: once developed, they can be applied across millions of product listings at negligible marginal cost. The investments in model development and associated energy or computing costs are minimal relative to the scale of overall platform operations. Hence, GenAI use does not meaningfully expand the firm's measured capital stock.

2. **Fixed labor input:** GenAI is used primarily to automate tasks that were already automated, to augment tasks supported by existing labor inputs, or to perform tasks with negligible labor displacement. Accordingly, the number of workers and total hours remain constant during adoption.

3. **Constant prices:** Output prices are fixed, so revenue growth reflects real output growth rather than changes in prices or markups.

4. **Stable factor shares:** Input cost shares (α, $1-\alpha$) remain constant during the adoption period.

5. **Constant utilization:** Capital utilization and effective labor effort do not vary, so measured input quantities remain valid.

In Section 3.3, we explain why these assumptions are likely to hold in our context. Under such conditions, the potential gains from GenAI adoption can be interpreted as a pure productivity shock, represented by an upward shift in the A term of the production function. This interpretation is consistent with treatments of past general-purpose technologies (e.g., electrification or the internet), where adoption translated into improvements in TFP rather than capital accumulation.[8] Evidence from consumer-facing sectors further suggests that quality improvements—including reductions in information asymmetry, better matching, enhanced personalization and targeting, and product innovation—can raise revenue-based productivity without corresponding shifts in input units or costs (Syverson, 2011; Brynjolfsson and Hitt, 2003; De Loecker, 2011).

3.3 Empirical Framework

To test the productivity gains of the seven GenAI-improved business workflows, the firm conducted a series of large-scale, randomized field experiments. Six of these experiments were executed at the consumer level, with participating consumers randomly assigned to either treatment or control groups. The only exception was the Google Advertising Title, which was conducted at the product level, where a subset of products selected for Google ads was randomly divided into treatment or control groups. Consumer overlap across experiments was minimal (below 1%). In all cases, the key distinction between treatment and control was that users or products in the treatment group were exposed to workflows re-engineered with GenAI, whereas in the control group workflows remained unchanged and followed the platform's standard practices. The total size of the subject pool varied greatly between experiments, with the smallest experiment having 30 thousand subjects while the largest containing up to 13 million subjects. Most of the experiments featured an equal distribution between the treatment and control subjects, with each group comprising approximately half of the total sample. The exceptions are Pre-sale Service Chatbot and Live Chat Translation, where the treatment group consumers comprised two-thirds of the total sample. Below, we provide details on the experiments related to all seven business workflows, with a summary of key features presented in Table 2. Appendix A presents illustrative user interfaces and examples for each workflow.

Pre-sale Service Chatbot: The experiment, conducted over a two-month period from September to October 2023, included a random sample of 33 thousand consumers who initiated pre-sale customer service requests for the platform's self-sold products during the experimental period. These consumers were randomly divided into treatment and control groups. Consumers in the control group received the platform's automated response service, which delivered a pre-programmed standardized notification indicating that customer service was unavailable. This auto-response condition reflects the platform's standard operating practice. Due to resource considerations, the platform has historically prioritized allocating human agents to post-sale rather than pre-sale support for self-sold products, given that pre-sale inquiries are generally less urgent. By contrast, consumers in the treatment group were supported purely by GenAI-powered chatbots. GenAI is expected to reduce asymmetric information between buyers and sellers in the treatment group by providing richer, context-specific responses to consumer inquiries.

Search Query Refinement: The experiment comprised three sub-experiments, each targeting consumers using different languages: Arabic, Japanese, and Polish. These languages were chosen

[8] A natural question is whether GenAI adoption should be viewed as capital deepening rather than TFP growth. While, in principle, new investments in IT infrastructure (e.g., servers, GPUs, or proprietary model development) could expand the capital stock, in our context the platform already possessed the required infrastructure, and GenAI applications constituted incremental software upgrades operating on existing systems.

Table 2: Summary Descriptions of the Experiments

	Business Workflow	Time Frame	Sample Size	Control	Treatment	Data Availability	Product Sold By
1	Pre-sale Service Chatbot	Two months from Sep. to Oct. 2023	44,614 consumers	Pre-programmed auto response indicating no customer service	GenAI Agent	Yes	Platform
2	Search Query Refinement	Three nine-day sub-experiments from May. to Jun. 2024	1,849,382 consumers	Basic query translation with no semantic comprehension	GenAI-translated queries with semantic comprehension	Yes	Sellers & Platform
3	Product Description	Five one-week sub-experiments in Dec. 2023	4,772,937 consumers	Human-generated descriptions	GenAI-generated descriptions on top of those created by humans	Yes	Platform
4	Marketing Push Message	One day in Dec. 2023	13,715,528 consumers	Human-generated standardized messages	GenAI generated a large and diverse set of messages	Yes	Sellers & Platform
5	Google Advertising Title	Twelve days in Jan. 2024	1,244,016 products	Human-generated ad titles	GenAI-optimized ad titles	Yes	Sellers & Platform
6	Chargeback Defense	Two months from Oct. to Dec. 2023	About 30 thousand consumers	Human agent	GenAI agent	No	Sellers & Platform
7	Live Chat Translation	One month in Oct. 2023	About 0.2 million consumers	Filipino agent without translation assistance	Filipino agent with GenAI real-time translation assistance	No	Platform

[1] In column "Product Sold By", "Platform" denotes products procured and sold directly by the platform. "Sellers & Platform" indicates that the experiment included products sold both by the platform itself and by third-party sellers on the platform.

because they are less commonly used on the platform and have historically been underserved by the platform's traditional translation of search queries.[9] The sub-experiments were launched at different points between May and June 2024, each lasting nine days. During each period, a random subset of consumers conducting searches was assigned to the experiment. These consumers were then randomly divided into two groups, yielding a total sample of approximately 2 million consumers across all sub-experiments. In the control group, consumer search queries were subject only to basic translation without semantic comprehension. In the treatment group, GenAI was deployed to translate queries by comprehending their underlying intent and refining them to improve semantic accuracy and clarity. The enhancement is expected to reduce search friction in the treatment group, as it can improve consumers' demand expression and facilitate the search engine to present products more closely aligned with their needs.

Product Description: The experiment consisted of five sub-experiments, each encompassing consumers who spoke English, Spanish, French, Portuguese, or Korean. All sub-experiments ran for one week in December 2023, with staggered start dates. GenAI was employed to create multilingual product descriptions for a predetermined product set of approximately 45,000 randomly selected platform self-sold products spanning a broad range of categories. On this platform, product descriptions refer to the text content in the description module of product detail pages that summarizes key features and selling points. According to our partner, self-sold products are primarily sourced from Chinese vendors, who typically provide image-based introductions with limited Chinese text embedded in the images. While such image-based content aligns with Chinese consumer preferences, global consumers are more accustomed to text-based descriptive bullet points, such as the "About this item" section on Amazon. Consequently, nearly half of the self-sold products either lack textual descriptions or contain only minimal textual information. During each sub-experiment period, a random subset of consumers who clicked into the product detail pages of the selected products was assigned into the experiment and evenly split to treatment and control groups, resulting in a total of approximately 5 million participants. Control group consumers viewed the original human-generated descriptions, whereas treatment group consumers were shown the GenAI-created descriptions on top of the original, human-created descriptions.[10] The treatment group is expected to experience lower information asymmetries because GenAI descriptions are more complete, standardized, structured, and accessible.

Marketing Push Message: The experiment took place over the course of approximately one month in December 2023. A random subset of consumer who received push notifications on their mobile entered into the experiment and were then randomly assigned into either control or treatment groups. Given the large scale of this experiment, we restricted our analysis to the first day, which contained 13 million consumers. On our partner platform, push messages were traditionally created by staff, requiring 1-2 employees several hours each month to produce a few dozen messages. Given the platform's hundreds of millions of consumers, this limited volume meant that many consumers received identical content, constraining the potential for personalized marketing. Accordingly, in the control group of our experiment, consumers primarily received uniform, human-generated marketing content, totaling roughly 2,000 distinct messages. By contrast, in the treatment group, about 40% of consumers were randomly assigned to GenAI-generated messages, yielding nearly

[9] On our focal platform, the search algorithm initially translates multilingual queries into English to facilitate matching with product and seller information stored in English.

[10] For products without existing human-generated descriptions, control group consumers saw no description - as was historically the case for such products, and treatment group consumers saw the GenAI-generated descriptions only.

3 million unique messages and thus far greater differentiation across individuals. The hypothesis in this experiment was that the very large number of GenAI-generated messages would enable the platform to deliver more distinctive marketing content across consumers and achieve refined matching of consumers with messages, thereby leading to better responses.

Google Advertising Title: The experiment was conducted over twelve days in January 2024, with randomization occurred at the product level. The sample included 3.5 million products selected by the retail platform for advertising in the sponsored section of Google Shopping, representing a diverse set of categories. For Google ads, the quality of the advertisement title is critical: a well-crafted title not only increases product discoverability by aligning with user search keywords but also enhances user clicks by incorporating appealing terms that drive consumer conversion.[11] In our experiment, the control group retained the original product titles created by the sellers, while in the treatment group, titles used in the ads were optimized by GenAI based on seller titles.[12] One key difference between this experiment and the others is that the GenAI model was not specifically fine-tuned to the advertising domain. Therefore, the generated titles may fail to highlight product attributes most relevant for consumer search and purchase decisions, leading us to be more agnostic about the potential treatment effect in this setting.

Chargeback Defense: The experiment, conducted from late October to late December 2023, included over 30 thousand consumers. During this period, a random subset of consumers who initiated chargeback requests was assigned to the experiment and then randomly divided into two groups. Contesting chargeback disputes requires a broad skill set, including claim analysis, evidence collection, and persuasive defense writing, which is especially challenging in cross-border contexts characterized by language barriers and complex regulations and customs. As a result, more than half of chargeback disputes on the focal platform are left unaddressed by sellers. In the control group, consumer claims were initially addressed by sellers. If no action was taken, approximately 3-5 outsourced workers then intervened to resolve the claims. However, these employees could only handle a small fraction of cases elaborately, while most were processed using generalized templates that proved far less effective. In contrast, claims in the treatment group were initially managed by sellers and subsequently supported by GenAI agents. Note that this is the only experiment where costs were not identical between treatment and control conditions. Specifically, in the treatment group, the elimination of 3-5 outsourced employees reduced labor costs, but the effect was negligible. We expect the treatment group to experience higher resolutions, since GenAI can generate more tailored and context-specific responses than template-based staff.

Live Chat Translation: The experiment was conducted over a one-month period in October 2023 and covered about 0.2 million non-English-speaking consumers seeking assistance from the platform's customer service. Due to cost constraints, on our focal platform, a significant portion of requests from non-English-speaking consumers can only be addressed by customer service agents from the Philippines providing service in English, as employing native agents for every market is relatively 3 times more expensive compared to Filipino agents. During the experiment, non-English-speaking consumers were randomly split among treatment and control conditions. In the treatment group, consumers interacted with Filipino agents with real-time bidirectional GenAI

[11] Many e-commerce platforms maintain libraries of such buzzwords which, based on historical data, are known to boost consumer click-through and conversion rates.

[12] When promoting products on Google Shopping, the platform used the pricing and image information provided by sellers, and these factors remained unchanged across treatment and control groups in our experiment.

translation support, while those in the control group engaged with Filipino agents without the aid of translation assistance. We expect the treatment group to face lower communication frictions, as real-time GenAI translation reduces language barriers between consumers and agents, thereby improving service quality and potentially raising conversion rates.

Importantly, in line with the assumptions stated in Section 3.2, labor inputs were held constant between treatment and control conditions when adopting GenAI in each experiment. In particular, there was no or minimal labor displacement across the workflows. Three scenarios account for this outcome. First, in the Pre-Sale Service Chatbot and Search Query Refinement experiments, GenAI substituted tasks that were already automated, either through pre-programmed notification or standard search algorithms, and therefore required no labor input. Second, in cases where GenAI augmented rather than replaced existing labor, such as Product Description (where GenAI-generated description was presented ahead of the original human-generated one), Marketing Push Message (where some consumers were randomly assigned to GenAI-generated messages, while others continued receiving human-generated ones), Google Advertising Title (where GenAI optimized ad titles based on human-created product titles), and Live Chat Translation (where agents retained their roles but received real-time GenAI translation support), no reduction in labor inputs occurred. Third, the only exception was the Chargeback Defense experiment, where GenAI adoption technically replaced workers previously responsible for drafting defenses. However, the extent of displacement was negligible, affecting only 3–5 outsourced workers. Moreover, although energy/computing costs may have varied between treatment and control conditions, these were minimal and negligible compared to the cost of other operations. Finally, the adoption of GenAI did not alter product prices.

3.4 Data and Estimation

We obtained comprehensive granular consumer/product-level data for the first five of the seven experiments, allowing for in-depth measurement and exploration of productivity gains. For the remaining two experiments—Chargeback Defense and Live Chat Translation—the platform could not provide granular data. In these cases, we rely on analyses conducted by the platform's internal data science team. These estimates complement our direct observations, offering a broader perspective on the impact of GenAI across various business areas (see "Data Availability" in Table 2).

For the experiments operated at the consumer level, we recorded each consumer's treatment status. We gathered consumers' complete set of activities, including the number of product views (referred to as View), product clicks (Click), product orders (Order), and total expenditures on those orders (Sales).[13] For comparability across workflows, our benchmark analysis focuses on two primary outcome measures: (i) total consumer expenditure (Sales), which serves as a revenue-based measure of retail productivity; and (ii) the conversion rate, a binary indicator for whether a consumer made a purchase, representing a proxy for changes in consumer experience. Both sales and conversion rate are widely used industry metrics. For the product-level experiment, we collected analogous data at the product level. Table 3 presents summary statistics for the key variables in

[13] In the Search Query Refinement experiment, product views represent the number of products a consumer browses on the search results page, which displays a summarized collection of products immediately after a query search. In the Google Advertising Title experiment, product views refer to the number of views of advertised products within the Google Shopping tab. In the Search Query Refinement and Product Descriptions experiments, product clicks capture the number of times consumers clicked into product detail pages. In the Marketing Push Message experiment, product clicks reflect consumer clicks on push notifications, while in the Google Advertising Title experiment, they indicate clicks on advertised products.

Table 3: Summary Statistics of Main Outcomes

	Mean	Standard Dev.	Median	Min	Max
Pre-sale Service Chatbot					
Conversion Rate	0.068	0.253	0	0	1
Sales	1.86	9.749	0	0	517.34
Search Query Refinement					
View	313.36	615.02	125	1	105,883
Click	8.23	16.99	3	0	2024
Order	0.16	0.73	0	0	85
Conversion Rate	0.09	0.28	0	0	1
Sales	2.24	21.41	0	0	4,960
Product Description					
Click	1.98	2.06	1	1	173
Order	0.06	0.30	0	0	23
Conversion Rate	0.04	0.20	0	0	1
Sales	0.51	4.56	0	0	2,941
Marketing Push Message					
Click	0.058	0.234	0	0	1
Conversion Rate	0.0014	0.037	0	0	1
Order	0.0015	0.039	0	0	6
Sales	0.045	1.756	0	0	501
Google Advertising Title					
View	19.36	82.38	5	2	12,033
Click	0.22	1.69	0	0	627
Conversion Rate	0.004	0.069	0	0	6
Sales	0.129	2.97	0	0	322

[1] “View” refers to the number of product views. “Click” stands for the number of product clicks. “Order” is the number of product orders. “Sales” represents the total expenditure on product orders. “Conversion rate” measures consumers’ likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise. In Google Advertising Title experiment, the unit of observation is at the product level. The conversion rate is calculated as the number of purchases divided by the number of views for that product.

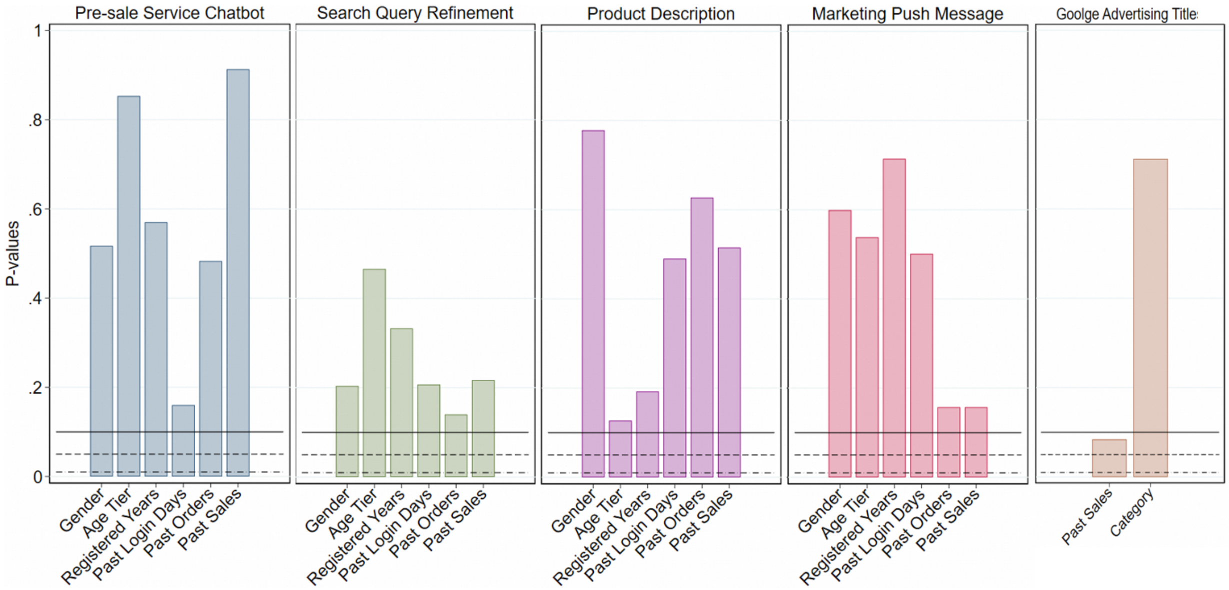
each of the five experiments where granular data are available.

To compare the mean purchase value and mean conversion rate between the treatment and control groups, we rely on the following general empirical specification, which we adapt as needed for each experiment:

$$y_i = \beta \times Treat_i + \alpha_{c(i)} + \epsilon_i, \tag{2}$$

where i denotes the randomized unit (consumer or product), and y_i is the outcome. $Treat_i$ is the treatment indicator, which equals one if the consumer or product belongs to the treatment group and zero otherwise. $\alpha_{c(i)}$ denotes the cohort fixed effects. Specifically, in the Pre-Sale Service Chatbot and Google Advertising Title experiments, consumers or products entered the experiments on different days, we therefore control for entry-day cohort fixed effects. In the Search Query Refinement and Product Description experiments, multiple sub-experiments were conducted across different languages at varying times, we thus include entry-day-by-language cohort fixed effects. For the Marketing Push Message experiment, the sample spans only a single day, so no cohort fixed effects are included. Details on the model specification for each experiment are provided in Appendix C.

Figure 1: P-Values for Covariate Balance Checks Across Experiments



[1] This figure presents the p-values for the covariate balance checks across all experiments. For the first four experiments, which are conducted at the consumer level, six consumers' demographic and behavioral variables are examined. For the single experiment conducted at the product level (Google Advertising Title), the focus is on two key measurements of products, which are the product's historical sales and the distribution of product categories.

[2] The bar presents the p-values. The solid line, dash line and dash-dot lines indicate p-values of 0.1, 0.05 and 0.01, respectively.

[3] "Gender" is predicted by the platform based on consumers' past shopping behaviors. "Age Tier" is a 7-point scale from 1 (youngest) to 7 (oldest). "Registered Years" indicates the duration from the year of consumer registration to the year of experiment. "Past Login Days" represents the number of days consumers have logged into the platform in the 30 days prior to the experiment. "Past Orders" is the number of product orders in the 30 days prior to the experiment. "Past Sales" represents the total expenditure on product orders in the 30 days prior to the experiment. "Category" is the category associated with a product.

We estimate Equation (2) via OLS, adjusting the standard errors for heteroskedasticity. Under random assignment, β recovers the average treatment effect of GenAI adoption, expressed as the absolute lift in outcomes. We also report results in percent lift, rescaling β by the control group mean. For sales, we use levels to address concerns regarding log transformations with zero outcomes (Chen and Roth, 2024). For conversions (a binary outcome), we primarily estimate a linear probability model and confirm that the findings are robust to logit specifications. We also estimate the model using pre-experiment covariates as controls, and the results remain consistent. Consumer overlap across experiments was minimal (less than 1%), and our findings are robust to excluding overlapping observations. This design allows treatment effects to be solely attributed to individual workflows, though it does not capture potential complementarities across GenAI applications.

To enrich our analysis, we obtained pre-experiment seller data for products included in the experiments (e.g. seller size measured by annual sales, seller operational years, and the number of sub-accounts linked to a seller's online store). These data enable us to examine seller-level heterogeneity in the three experiments involving products sold by both third-party sellers and the platform (see column "Product Sold By" in Table 2). However, this analysis is not applicable to the Pre-sale Service Chatbot and Product Description experiments, as the products in these cases were platform self-sold products—sold exclusively by a limited number of platform-operated sellers—resulting in insufficient variation in seller characteristics for meaningful heterogeneity analysis. Additionally, we collect product characteristics, such as the concentration ratio of its associated category, price, and annual sales quantity, to explore product-level heterogeneity.

We also augmented the data with consumer demographics and pre-experiment shopping history (years of registration, activity level, purchase volume, etc.), which supported both the analysis of buyer-level heterogeneity and the verification of random group assignment in the experiments.[14] As confirmed by the covariate balance checks reported in Figure 1, we found no systematic significant pre-experiment differences across consumers between the control and treatment groups. Thus, the randomization process was effective at allocating comparable consumers/products into the two groups. Details on the covariate balance checks are presented in Appendix B.

4 Main Results

Table 4 reports average effects across the GenAI-driven business workflows. The table presents, from left to right: (i) the impact of GenAI re-engineering on sales, our measure of productivity; and (ii) the mechanism underlying productivity gains, captured by the conversion rate as a proxy for consumer experience. Columns (1) and (3) report the estimated average treatment effects (ATE, absolute lift), while columns (2) and (4) present percentage changes relative to the control group (relative lift). Additional results are provided in Appendix C.

4.1 Productivity Impact by Workflow

Table 4 shows overall productivity improvements from most GenAI deployments, alongside substantial heterogeneity across workflows. The largest effect arises in the Pre-sale Service Chatbot, where GenAI increases sales by 16.3% ($p < 0.01$) relative to the control group (Column 2). In this setting, consumers in the treatment group were assisted by a GenAI-powered chatbot, while those in the control group received an automated message indicating that no support was available. This auto-response condition reflects the platform's standard practice of allocating human agents primarily to urgent post-sale inquiries, while offering no assistance for most pre-sale inquiries on self-sold products due to limited resources. However, it is plausible that consumers in the control group became frustrated by the lack of assistance, potentially reducing their likelihood of purchase and leading to an overestimation of our treatment effects. For this reason, the platform conducted further experiments on the GenAI-powered chatbot.

Appendix Table C1 reports additional comparisons, including evaluations against and in combination with human agents. The results show that the GenAI chatbot delivers service quality comparable to human customer support (Columns 3–4). Integrating GenAI with human agents further increases sales: when comparing the no pre-sale service condition with the treatment that combines GenAI assistance and human escalation when needed, sales rise by 25% (Columns 5–6), indicating strong complementarities between GenAI and human labor. More importantly, comparing consumers who receive GenAI-assisted service with human escalation to those served exclusively by human agents shows that the former spend 11.5% more (Columns 7–8). We interpret this latter comparison as a conservative lower bound on GenAI's productivity impact in pre-sale customer support.

[14] According to our research agreement with the partner platform, all consumers in our data are anonymous to ensure consumer privacy. We identify consumers by hashed IDs instead of knowing their actual names.

Table 4: Summary of Average Treatment Effects of GenAI Adoption Across Workflows

		Productivity Impact (Sales, $)		Mechanisms (Conversion Rate)	
		(1)	(2)	(3)	(4)
	Business Workflow	Coefficient	% Change	Coefficient	% Change
1	**Pre-sale Service Chatbot**	0.274*** (0.0995)	16.3%	0.0131*** (0.00256)	21.7%
2	**Search Query Refinement**	0.0648** (0.0314)	2.93%	0.00101** (0.00041)	1.15%
3	**Product Description**	0.0104** (0.00417)	2.05%	0.000554** (0.000187)	1.27%
4	**Marketing Push Message**	0.000402 (0.000812)	1.6%	0.000048** (0.0000218)	3.0%
5	**Google Advertising Title**	-0.00602 (0.00534)	-4.5%	-0.000137 (0.000124)	-3.3%
6	**Chargeback Defense**†	15% defense success rate increase			
7	**Live Chat Translation**†	5.2% consumer satisfaction increase			

[1] "Sales" represents the total expenditure on product orders, in USD."Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[2] Columns (1) and (3) report the estimated coefficients for sales and conversion rate, respectively, with standard errors in brackets. Columns (2) and (4) report % Change for sales and conversion rate, respectively. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

[3] † For these experiments data are not available: we report findings estimated by the platform's internal data science team.

The remaining four workflows with detailed data reported in Table 4 exhibit smaller effects on sales, ranging from a negative and statistically insignificant change to gains of up to 3%. Specifically, the Search Query Refinement application increased sales by 2.93% ($p < 0.05$), while the Product Description generated a 2.05% ($p < 0.05$) gain—still substantial effects for a platform of this scale and maturity. The Marketing Push Message workflow shows a positive yet not statistically significant improvement in sales (1.6%). This likely reflects the combination of a very low baseline conversion rate (only 0.14% of consumers make a purchase) and high variance in expenditures among converters, suggesting that broader implementation could provide sufficient power to detect treatment effects.[15] It is also important to note that only a subset of the treatment group was exposed to GenAI-generated messages in this experiment, which may attenuate the observed treatment effects under partial exposure. By contrast, the Google Advertising Title workflow exhibits an insignificant negative effect. This pattern is consistent with the lack of fine-tuning of the GenAI model to the advertising context, which led it to omit commonly used commercial keywords. An alternative explanation is a divergent delivery bias in ad distribution: Google's advertising algorithm may have recognized and deprioritized AI-generated titles, reducing their visibility. Together, these findings underscore the importance of domain-specific fine-tuning or retraining of foundation models when applying GenAI to industry-specific tasks requiring specialized knowledge (Deloitte, 2023).

For the last two business processes studied (Chargeback Defense and Live Chat Translation), we did not obtain granular transaction data and instead relied on the company's internal metrics and

[15] Note that this case exhibits significant conversion rate increase in Table 4 as well as significant order increase in Appendix Table C5.

analysis. While not directly comparable to the effects on sales, these measures indicate substantial improvements in two important outcomes: a 15% increase in defense success rates for Chargeback Defense and a 5.2% increase in consumer satisfaction attributable to Live Chat Translation.[16]

These results provide new evidence on the potential of GenAI to enhance revenue-based productivity in online retail, thereby contributing to the broader debate on the economic consequences of GenAI adoption (Peng et al., 2023; Acemoglu, 2025). A key distinction of our study is its focus on revenue outcomes rather than input-side efficiency gains, which makes our estimates not directly comparable to those from studies emphasizing labor productivity or task efficiency. Among the seven deployments we examine, five deliver measurable performance gains, showing that GenAI can generate substantial improvements in firm outcomes under real-world operating conditions. Overall, the evidence points to a positive but heterogeneous impact of GenAI along the customer journey.

The variation we observe across workflows is unlikely to reflect differences in implementation quality—which was comparable across applications—but instead arises from differences in the scope of GenAI's marginal contribution relative to baseline conditions. In each case, the control group reflects the firm's standard practices prior to GenAI adoption. For example, in the Pre-Sale Service Chatbot, the control group received no customer support; in Search Query Refinement, it relied on standard machine-learning–based search algorithms; and in most other workflows, the benchmark was human input. The results therefore point to genuine heterogeneity in where GenAI is most effective: customer-support applications, such as Pre-Sale Service Chatbots, deliver the largest improvements; search and product-discovery tasks yield smaller gains; and advertising-related applications show no significant effects.[17] Together, this highlights both the role of baseline conditions in shaping treatment effects and the comparative effectiveness of GenAI across functional areas.

4.2 Mechanism

Most of the existing literature attributes the productivity-enhancing potential of GenAI to supply-side mechanisms, such as labor savings or efficiency improvements (e.g., reductions in the time required to complete a task). Our context and findings highlight an additional channel: GenAI can unlock productivity gains by enriching the demand-side consumer experience. We measure this mechanism using conversion rates (i.e., the likelihood that consumers complete a purchase), which serves as a widely recognized industry standard for consumer satisfaction in the e-commerce context. Across the various workflows, we document significant increases in conversion rates ranging from 1% to 22% (Column 4 of Table 4), which in turn translate into higher output as measured by sales. Notably, we do not find any significant effects along the intensive margin. Table 5 shows that the average cart value among consumers who made at least one purchase (or among products purchased at least once) remains unchanged following GenAI adoption. In our setting, GenAI primarily drove market expansion by converting a larger share of consumers, rather than inducing existing buyers to purchase higher-priced or larger quantities of products.

This evidence indicates that the observed productivity gains arise from GenAI's capacity to reduce market frictions—both by enabling new services and by improving existing ones—thereby

[16] In Chargeback Defense, further gains not captured in our calculations may also arise from cost reductions, as the GenAI-enhanced workflow eliminated the need for manual intervention.

[17] In the Marketing Push Message workflow, experimental design features may also influence measured effects, as only a subset of the treated population was exposed to GenAI, which likely attenuated measured impacts.

Table 5: Impact of GenAI Adoption on Average Cart Value (intensive margins)

		Cart Value ($)	
		(1)	(2)
	Business Workflow	Coefficient	% Change
1	**Pre-sale Service Chatbot**	-0.859 (1.203)	-3.1%
2	**Search Query Refinement**	0.376 (0.334)	1.49%
3	**Product Description**	0.0942 (0.0807)	0.81%
4	**Marketing Push Message**	0.024 (0.473)	0.15%
5	**Google Advertising Title**	-0.784 (0.992)	-2.3%

[1] The table presents the treatment effects of GenAI adoption on "Cart Value". For workflows 1-4, "Cart Value" refers to the expenditure per consumer, conditional on the consumer making a purchase. For workflow 5, "Cart Value" refers to the expenditure per product, conditional on the product being purchased.

[2] Columns (1) report the estimated coefficient, with standard errors in brackets. Columns (2) report % Change, which is calculated dividing the treatment effect by the control group mean. *** p<0.01, ** p<0.05, * p<0.1.

enhancing the shopping experience and influencing purchasing behavior. For example, GenAI mitigates information asymmetries by providing relevant and timely assistance through pre-sale chatbots (a 21.7% increase in conversion rate, $p < 0.01$) and by generating comprehensive and structured product descriptions (a 1.27% increase, $p < 0.05$). It reduces search frictions and improves match quality by enhancing the translation and semantic comprehension of consumer queries (a 1.15% increase, $p < 0.05$). In addition, GenAI enables personalization of marketing content by enabling large-scale generation of customized marketing messages across a broad product portfolio (a 3% increase, $p < 0.05$). Finally, evidence from Chargeback Defense and Live Chat Translation points to improvements consistent with enhanced user experience, including higher success rates and greater consumer satisfaction.

In Appendix C, we provide additional results on user activities where data are available. For example, GenAI-refined search queries led to a 2.02% ($p < 0.01$) increase in click-through rates—the ratio of product clicks to views—indicating that consumers found the displayed products more appealing and were more likely to seek additional information after viewing the summarized search results. GenAI-generated product descriptions increased the number of orders placed by 1.08% ($p < 0.05$), while GenAI-created marketing messages raised clicks by 3% ($p < 0.05$) and orders by 2.8% ($p < 0.10$). These additional outcomes provide consistent evidence that GenAI enhances intermediate engagement measures—such as clicks, orders, and click-through rates—reflecting a smoother and more informative shopping experience that, in turn, underpins the higher conversion and sales gains documented above.

4.3 Aggregate Productivity Gains Across Workflows

In this section, we aggregate productivity gains across workflows to estimate the AI-driven incremental value per consumer observed in our experiments. Specifically, we focus on the four workflows with positive treatment effects in sales, excluding the Google Advertising Title experiment, which serves as a testable bad case and can be readily improved in future iterations.

Table 6 summarizes the key variables used in our calculations. Column (1) reports the AI-driven incremental value per consumer for each experimental workflow, reflecting the estimated average

Table 6: Aggregate Productivity Gains Across Workflows

	Business Workflow	(1) Incremental Value Per Consumer ($)	(2) Time Multiplier	(3) Annualized Incremental Value Per Consumer ($)
1	**Pre-sale Service Chatbot**	0.274 (upper-bound)	6.0	1.6
		0.218 (lower-bound)	6.0	1.3
2	**Search Query Refinement**	0.0648	40.6	2.6
3	**Product Description**	0.0104	52.1	0.5
4	**Marketing Push Message**	0.0004	365.0	0.1
	Total (linear additivity)			**5.0** (upper-bound)
				4.6 (lower-bound)

Column (1) reports the absolute lifts in sales (treatment effects of GenAI) for each workflow from Table 4. Column (2) shows the factor used to annualize the workflow-specific estimates. Column (3) reports the annualized values.

treatment effects (the absolute sales lift per consumer) reported in Column (1) of Table 4. Because each experiment had a different duration, Column (2) presents the time multiplier used to extrapolate these effects to an annual horizon. For instance, the Pre-Sale Service Chatbot experiment spans two months, yielding a time multiplier of six (i.e., 12/2). For each worklow, we obtain the annualized AI-driven incremental value per consumer in Column (3) by multiplying the estimated effect in Column (1) by the corresponding time multiplier in Column (2). This calculation assumes that the treatment effects observed during the experimental period remain constant over time, abstracting from potential amplification (e.g., greater engagement and purchases as GenAI-generated content enhances consumer satisfaction and platform loyalty) or attenuation (e.g., consumer dissatisfaction and product returns due to potential mismatches between AI-generated content and actual product characteristics).

Finally, we aggregate the annualized estimates across the four workflows to obtain the total annual incremental value per consumer attributable to GenAI (see "Total" in Column (3)). This aggregation assumes that effects across workflows are linearly additive and abstracts from potential cross workflow interactions, such as cannibalization among touchpoints or synergies. Based on the four GenAI applications with positive sales effects, we estimate an annual incremental value of about $5 per consumer, which decreases to $4.6 when applying the lower-bound estimate from the Pre-Sale Service Chatbot experiment. These effects represent roughly 5.5-6% of the increase in revenue per user observed in global e-commerce between 2023 and 2024 (Statista, 2024), highlighting the economic significance of these gains relative to broader industry trends.

It is worth noting that these estimates capture only a partial and time-dependent view of the firm's efforts to scale up GenAI across workflows. While in 2023 the platform applied GenAI to only a handful of workflows, by 2024 it was deployed in more than 40 applications and by 2025 in over 60. This rapid expansion is also reflected in the growth of API calls to large language models: in mid-2024, AI-related API requests averaged over 50 million per day, rising to more than 1 billion per day by mid-2025—a twentyfold increase. Hence, our point estimates should be interpreted with caution and in light of these rapid adoption trends, which nonetheless indicate that the firm anticipates substantial value from GenAI.

Taken together, these results reveal sizable gains in targeted workflows and measurable contributions to overall platform sales, with substantial potential as GenAI applications diffuse across use cases and models are further refined for domain-specific tasks. This pattern aligns with Acemoglu (2025), who emphasize that the aggregate productivity effects of new general-purpose technologies materialize gradually as complementary investments and organizational adaptations accumulate.

5 Heterogeneous Treatment Effects

To further shed light on the mechanism behind our results, we examine heterogeneity in treatment effects across sellers, buyers, and products. If GenAI primarily reduces frictions on both the demand and supply sides, one would expect relatively larger gains among participants with lower baseline capabilities—namely, smaller and less experienced sellers, buyers with limited platform engagement, and products in the long tail of sales or in less concentrated categories. We classify each dimension into "high" and "low" groups based on pre-experiment characteristics that capture scale, experience, or market position. For sellers, the classification is based on size and tenure on the platform. For buyers, we use measures of online shopping experience and purchasing intensity. For products, we rely on indicators of category concentration, sales volume, and relative price. These groupings allow us to assess whether GenAI adoption disproportionately benefits smaller or less experienced sellers, less sophisticated buyers, and lower-volume or lower-priced products, thereby revealing how the technology shapes outcomes across different segments of the marketplace. Note that, depending on the context and data availability, some heterogeneity analyses cannot be implemented for certain workflows. We find consistent results across seller and buyer characteristics: disadvantaged groups, such as small sellers and inexperienced buyers, experience greater gains from GenAI-powered workflows. In contrast, the effects across product groups are more context-dependent.

5.1 Heterogeneous Effects Across Sellers

The platform hosts a highly diverse population of sellers, varying substantially in firm characteristics. We focus on sellers' size and sophistication to investigate whether the productivity potential of GenAI vary between big versus small sellers. We classify sellers into high and low groups based on three pre-experiment metrics: annual sales value, years of operation on the platform (Operation Years), and the number of sub-accounts linked to the seller's online store (# of Sub-Accounts). For each metric, sellers in the low group are generally small sellers, defined as meeting the following pre-experiment criteria: (i) accounting for the bottom 50% cumulative share of total sales when sellers are ranked by annual sales; (ii) having operated on the platform for fewer than five years;[18] or (iii) maintaining fewer than three sub-accounts for their online store.[19] Table 7 presents the summarized results on seller heterogeneity, reporting only the percent change and significance levels for brevity. Additional results and details can be found in Appendix D.

Search Query Refinement. Table 7 shows that small sellers experience significant gains in both sales and conversion rates (Columns 2, 4, and 6), whereas effects for larger sellers tend to be smaller

[18] The platform closely monitors seller tenure, with five years of continuous operation serving as an important performance indicator.

[19] According to the advice of the platform's internal staff, small online stores are typically run by individuals or operate as mom-and-pop businesses, often with no more than two sub-accounts. Stores that have more sub-accounts generally employ additional staff to assist with tasks related to store operations, suggesting that these are more likely to be larger sellers.

Table 7: Summary of Heterogeneous Treatment Effects Across Sellers

			(1)	(2)	(3)	(4)	(5)	(6)
	Business	Dependent	Annual Sales		Operation Years		# of Sub-Accounts	
	Workflow	Variable	High	Low	High	Low	High	Low
1	Search Query	Sales	2.18%	3.68%**	2.28%	3.19%**	0.97%	3.48%**
	Refinement	Conversion	0.21%	1.69%***	1.24%	1.01%*	0.88%	1.20%**
2	Marketing	Sales	1.9%	2.1%	6.2%	0.7%	7.5%	-0.8%
	Push Message	Conversion	3.3%	3.2%*	1.9%	3.8%**	-0.1%	5.3%**
3	Google	Sales	-5.1%	-4.4%	-5.4%	-4.4%	-5.9%	-4.5%
	Ad Titles	Conversion	-6.5%	-0.3%	-2.3%	-3.8%	-4.7%	-3.2%

[1] We classify sellers into high and low groups based on three different pre-experiment metrics. Low-group sellers are generally small sellers, as defined by meeting the following pre-experiment criteria: (1) accounting for the bottom 50% cumulative share of total sales when sellers are ranked by annual sales; (2) having operated on the platform for fewer than five years; or (3) maintaining fewer than three sub-accounts for their online store.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] We report the % change. % change is calculated by dividing the treatment effect by the control group average. The asterisks notation indicates the significance of treatment effect. *** p<0.01, ** p<0.05, * p<0.1.

[4] The heterogeneity analysis of different seller types was not applicable for experiments of Pre-sale Service Chatbot and Product Description as these two were conducted on platform self-sold products, which involved only a few platform-operated sellers, thereby lacking variation in seller characteristics for meaningful examination of seller heterogeneity.

and statistically insignificant (Columns 1, 3, and 5). GenAI enhances the search algorithm's ability to translate queries by understanding consumer intent and refining queries to improve semantic accuracy and clarity. This refinement allows the search engine to retrieve products more relevant to consumer needs, thereby reducing search frictions, improving matching efficiency, and ultimately increasing purchases for small sellers. These findings are consistent with prior work showing that lowering search costs or improving search quality fosters long-tail dynamics and expands market share for smaller and niche sellers (Brynjolfsson et al., 2011; Bar-Isaac et al., 2012; Yang, 2013; Zhou et al., 2025).

Marketing Push Message. A similar pattern, but only with respect to conversion rates, is observed in the platform's advertising function, where only small sellers show significant gains. A likely explanation is that human drafters, constrained by time and attention, tend to reference or replicate the product characteristics of top sellers when composing marketing content. By contrast, GenAI enables the large-scale generation of customized marketing messages across a broad product portfolio, enhancing content differentiation for smaller sellers' offerings and stimulating consumer purchases. These findings align with Sun et al. (2024), who show that personalized recommendation systems disproportionately benefit smaller and less established sellers.

Google Advertising Title. In this experiment, heterogeneous treatment effects across seller types are statistically insignificant for both high and low groups.

Other Workflows. We cannot assess seller heterogeneity in the Pre-sale Service Chatbot and Product Description experiments, as both were conducted exclusively on platform self-sold products with only a handful of platform-operated sellers, leaving insufficient variation for meaningful analysis. In the case of Product Description, as well as for experiments where granular data are unavailable (e.g., Chargeback Defense), the platform's internal data science team provides descriptive evidence from third-party sellers to offer anecdotal insights into potential heterogeneity from GenAI. Specifically, third-party sellers are divided into quartiles based on pre-experiment annual sales. The share of products with no or limited descriptions is 1.2 times higher among bottom-quartile (small) sellers relative to top-quartile (large) sellers, while the share of unaddressed chargeback disputes is 1.7 times higher. These figures suggest that small sellers are particularly constrained by shortages of talent and resources, lacking the capacity to perform or outsource functions such as customer service, product content generation, or compliance management. Taken together, these patterns provide anecdotal support for the view that small sellers stand to benefit disproportionately from GenAI.

Overall, our findings indicate that the productivity gains from GenAI are disproportionately larger for small sellers, characterized by lower transaction volumes, shorter operational histories, and fewer sub-accounts. This pattern is consistent with prior research on GenAI adoption by individual workers, which shows that low-skilled and less experienced labor—often disadvantaged by earlier technological advances—derive greater benefits from GenAI (Brynjolfsson et al., 2025; Hui et al., 2024; Chen and Chan, 2024; Noy and Zhang, 2023; Dell'Acqua et al., 2023; Choi et al., 2023; Peng et al., 2023). Our results extend this insight to the domain of entrepreneurship on online retail platforms.

5.2 Heterogeneous Effects Across Consumers

Beyond seller heterogeneity, little research has examined the impact of GenAI on different types of consumers. Our analysis focuses on the differential effects of the technology on experienced versus inexperienced consumers. We classify consumers into high and low groups based on three pre-experiment indicators of online shopping experience: years since registration on the platform (Registered Years), number of login days during the 30 days prior to the experiment (Past Login Days), and total expenditure during the same period (Past Sales). For each indicator, consumers in the low group are defined as relatively inexperienced if they meet the following criteria: (i) a registration duration below the platform-wide median; (ii) login days below the platform-wide median; or (iii) total purchases below the platform-wide median. Our results show significant heterogeneity in GenAI's effects across these consumer segments. Summary results are presented in Table 8, with additional details reported in Appendix E.

Pre-sale Service Chatbot. Table 8 shows that although both consumer groups benefit from the intervention, the gains are relatively larger for inexperienced consumers—namely, those with shorter registration histories, fewer prior login days, and lower past spending (Column 2, 4 and 6). One explanation is that experienced online shoppers are generally able to locate and evaluate product information independently, even when customer support is limited to pre-programmed auto responses (as in the control condition). By contrast, less experienced users rely more heavily on customer service for product and seller information to inform their decisions. As a result, GenAI-based customer support enables these consumers to navigate the platform more effectively and make better purchase choices.

Search Query Refinement. In this experiment, increases in sales and conversion rates were significant and pronounced only for less experienced consumers, while the effects for more experienced consumers were not statistically significant. This pattern likely reflects the challenges faced by inexperienced consumers, who generally possess weaker information-search skills and less fa-

Table 8: Summary of Heterogeneous Treatment Effects Across Consumers

			(1)	(2)	(3)	(4)	(5)	(6)
	Business	Dependent	Registered Years		Past Login Days		Past Sales	
	Workflow	Variable	High	Low	High	Low	High	Low
1	Pre-sale	Sales	13.7%**	22.4%**	15.0%**	18.5%**	8.6%	25.9%***
	Serv Chatbot	Conversion	19.8%***	26.1%***	17.1%***	29.6%***	17.6%***	25.4%***
2	Search Query	Sales	1.27%	5.01%**	0.63%	8.16%***	-0.56%	7.46%***
	Refinement	Conversion	-0.09%	2.79%***	0.43%	2.32%***	0.44%	1.93%***
3	Product	Sales	1.09%	3.06%***	-1.02%	6.24%***	0.03%	5.63%***
	Description	Conversion	1.10%*	1.39%**	0.65%	2.06%***	1.05%*	1.59%**
4	Marketing	Sales	0.6%	7.2%	2.3%	1.7%	-6.1%	22.9%***
	Push Message	Conversion	1.9%	5.5%**	0.5%	5.5%***	-2.3%	15.4%***

[1] We classify consumers into high and low groups and capture their online shopping experience using three pre-experiment indicators. Consumers in the low group are considered relatively inexperienced, as determined if they meet the following criteria: (1) a registration duration below the median among all consumers on the platform; (2) a number of 30-day login days below the platform-wide median; or (3) a 30-day total purchase amount below the platform-wide median.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] We report the % change. % change is calculated by dividing the treatment effect by the control group average. The asterisks notation indicates the significance of treatment effect. *** p<0.01, ** p<0.05, * p<0.1.

[4] The heterogeneity analysis of different consumer types is not applicable for the workflow of Google Advertising Title because the platform cannot access detailed consumer demographic data from Google.

miliarity with the online search environment, making it more difficult for them to articulate their needs effectively through query-based searches.

Product Description. Augmenting product descriptions with GenAI-generated content led to substantial increases in consumer purchases, particularly among inexperienced consumers, as reflected in significant sales gains for this group. A plausible explanation is that inexperienced consumers often lack domain knowledge and access to alternative information sources relative to experienced users, and therefore rely more heavily on detailed descriptions to guide their decisions. These findings suggest that GenAI can help bridge informational gaps and improve the shopping experience for less experienced consumers.

Marketing Push Message. In this experiment, the results remain consistent with the previous findings. When GenAI is used to enhance the personalization and differentiation of marketing messages across a broad product portfolio, the perceived relevance of the content increases, making it more likely to capture the attention of less experienced consumers and support their purchase decisions.

Google Advertising Title. This experiment is not suitable for consumer heterogeneity analysis, due to the fact that the it was conducted off-site at product level, preventing the platform from accessing detailed consumer demographic data from Google.

To sum up, we find that inexperienced consumers—those with shorter registration histories, lower login frequencies, and lower spending levels—benefit more from GenAI than their more experienced counterparts. This result is consistent with prior evidence that GenAI can support vulnerable groups such as low-skilled workers (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Peng et al., 2023), and it extends this insight to the consumer domain by showing that GenAI disproportionately benefits less sophisticated consumers. It also aligns with existing research indicating that enhanced e-commerce technologies—such as improved information provision, search refinement, and personalized recommendation systems—deliver greater benefits to consumers who are older, newer to the platform, or have lower purchasing power (Fang et al., 2024; Sun et al., 2024), while extending these patterns to the context of GenAI.

5.3 Heterogeneous Effects Across Products

The platform hosts hundreds of millions of products spanning diverse categories. To study heterogeneity across products, we classify them into high and low groups based on three pre-experiment metrics. First, category market concentration, measured by the sales share of the top 1% of products (ranked by annual sales) within each category. This metric reflects the extent of product and demand differentiation, with higher concentration indicating more standardized offerings and relatively homogeneous consumer preferences (e.g., laptops typically exhibit higher concentration than dresses, where product features and consumer tastes are more varied). Products in the low group belong to categories with concentration below the platform average. Second, annual sales quantity, defined within category. Products in the low group—referred to as tail products—comprise the bottom 50% cumulative share of units sold when ranked by annual sales quantity within a category. Third, product price, also defined within category to account for category-specific pricing variation. Low-priced products are those priced below the median of their respective category. Summary results are reported in Table 9, with additional details in Appendix F.

Pre-sale Service Chatbot. GenAI-enabled pre-sale service chatbot exhibits larger effects on sales and conversion rates for tail products with low sales volume (Column 4). A plausible explanation is that such products often suffer from poor visibility and information gaps that heighten consumer hesitation (Fan et al., 2016). GenAI chatbot mitigates these frictions by clarifying product details and providing richer information before the purchase, particularly for niche or long-tail items with weak marketing signals. Heterogeneity across prices is more nuanced, with the intervention benefiting both high- and low-priced products (Columns 5 and 6). For high-priced items, the chatbot likely supports conversion by helping consumers justify larger expenditures through detailed explanations of value, quality assurance, and comparative advantages. For low-priced goods, the chatbot reduces decision frictions by streamlining information provision and simplifying the purchase process. Similarly, heterogeneity across categories defined by market concentration is also nuanced (Columns 1 and 2), indicating that the GenAI-augmented chatbot benefits products in both highly concentrated and less concentrated categories.

Search Query Refinement. For this experiment, our results show significant increases in sales and conversion rates only for products in low-concentration categories (Column 2), tail products (Column 4), and high-priced segments (Column 5). These patterns are consistent with the idea that GenAI reduces search friction by enhancing the comprehension and articulation of consumer demand in query-based search engines. This effect is particularly valuable in categories with high product and demand differentiation, because in such settings, consumers often struggle to express their needs precisely, yet accurate and content-clear queries are critical for retrieving well-matched products. For example, searching for a laptop is relatively straightforward, as the category is concentrated around key brands and standardized specifications, whereas searching for a dress is more complex given the wide variation in designer, color, texture, style, and other subjective

Table 9: Summary of Heterogeneous Treatment Effects Across Products

			(1)	(2)	(3)	(4)	(5)	(6)
	Business	Dependent	Market Concentration		Annual Quantity		Price	
	Workflow	Variable	High	Low	High	Low	High	Low
1	Pre-sale	Sales	21.3%*	13.2%*	8.8%	21.3%***	16.6%*	15.9%*
	Serv Chatbot	Conversion	26.6%***	18.5%***	10.3%	30.3%***	17.4%*	24.5%***
2	Search Query	Sales	-0.80%	6.49%***	-0.47%	4.92%***	3.79%**	0.89%
	Refinement	Conversion	0.45%	1.85%***	0.00%	2.07%***	1.19%*	0.94%
3	Product	Sales	3.10%***	0.55%	2.10%*	2.03%*	4.10%***	0.89%
	Description	Conversion	1.38%**	0.81%	2.07%***	0.69%	3.32%***	0.63%
4	Marketing	Sales	4.4%	-1.5%	-0.5%	3.0%	0.6%	2.7%
	Push Message	Conversion	2.4%	3.5%*	0.6%	5.2%***	8.9%***	-0.2%
5	Google	Sales	-4.3%	-5.2%	-2.3%	-8.9%	-6.2%	-2.3%
	Ad Titles	Conversion	-4.1%	-2.1%	-2.9%	-4.4%	-4.2%	-2.8%

[1] We classify products into high and low groups based on three pre-experiment metrics. (1) Category market concentration, measured by the sales share of the top 1% of products (ranked by annual sales) within each category. Products in the low group belong to categories with concentration levels below the platform average. (2) Annual sales quantity, also defined within category. Products in the low group—referred to as tail products—are those comprising the bottom 50% cumulative share of total units sold when ranking products by annual sales quantity within each category. (3) Product price, defined within each category to control for category-level pricing variation. Low-priced products are those priced below the median of their respective category.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] We report the % change. % change is calculated by dividing the treatment effect by the control group average. The asterisks notation indicates the significance of treatment effect. *** p<0.01, ** p<0.05, * p<0.1.

attributes. In addition, the reduction in frictions can also increase the visibility of long-tail products and strengthening purchase determination of high-priced items.

Product Description. We find significant gains from GenAI-generated product descriptions in high-concentration categories (Column 1), in high-priced products (Column 5), and across both head and tail products (Columns 3 and 4). These patterns are consistent with several mechanisms. In high-concentration categories such as laptops, product attributes are more standardized and therefore easier to communicate through text-based descriptions. For high-priced items, more comprehensive descriptions may help mitigate information asymmetry, which is particularly salient in costly purchase decisions. The results for head versus tail products may partly reflect the experimental context: the intervention was implemented on platform self-sold products, where Chinese vendors typically provide image-based content that often misaligns with cross-border consumer preferences. Consequently, the head–tail classification may not accurately reflect the quality of text-based description content.

Marketing Push Message. Our results show no significant effects of GenAI deployment on sales across product segments, consistent with the overall analysis. However, we find significant conversion gains in categories with low market concentration (Column 2), in tail products (Column 4), and in high-priced products (Column 5). These patterns align with intuitive mechanisms. In categories such as apparel and fashion, where products are highly differentiated and consumer choices hinge on subtle differences in style, quality, and branding, GenAI-powered marketing—especially when personalized—can more effectively capture attention. Tail products, which typically lack historical data or reputational signals, benefit disproportionately from enhanced marketing outreach. For high-priced items, tailored messaging helps emphasize value, quality assurance, and differentiation, making premium offerings more persuasive.

Google Advertising Title. All heterogeneity results are statistically insignificant in this experiment.

Overall, our analysis of product heterogeneity shows that the effects of GenAI applications are context-dependent. Search Query Refinement and Marketing Push Message generate larger improvements in low-concentration categories, where product differentiation is greater and consumer preferences more diverse. By contrast, Pre-Sale Service Chatbot and Product Description yield stronger gains in highly concentrated categories. Segmenting products by annual sales and price further reveals more consistent patterns: across most workflows, GenAI tends to deliver larger benefits for tail products and for high-priced items.

6 Discussion and Conclusions

The rapid advances in GenAI have generated widespread expectations among investors and business leaders, driving unprecedented investment in infrastructure and applications. Yet doubts remain regarding the extent to which GenAI can generate substantial productivity improvements at scale. This paper offers some of the first large-scale, real-world evidence on GenAI adoption in online retail, shedding light on how firm-level deployment translates into tangible consumer value and measurable business outcomes.

Our findings yield three main insights. First, GenAI can deliver measurable productivity improvements with comparable inputs, as reflected in increased sales across several business workflows. While the magnitude of these gains varies widely—from negligible effects to double-digit increases—the evidence demonstrates that GenAI can generate substantial economic value when deployed in consumer-facing processes. Second, these improvements arise primarily from enhanced consumer experience and satisfaction through the reduction of frictions in the marketplace, rather than through cost savings on the input side. Across workflows, we observe higher conversion rates: by enriching pre-sale communication, refining search queries, generating richer product descriptions, and personalizing marketing messages, GenAI improves matching efficiency and mitigates information asymmetries. Third, the benefits of GenAI adoption are heterogeneous. Smaller sellers, less experienced consumers, tail products, and high-priced items derive disproportionately larger gains, highlighting GenAI's role in bridging capability gaps across different segments of the marketplace.

Because most experiments were randomized at the consumer level and overlap across experiments was minimal (less than 1%), the observed effects capture incremental demand (i.e., market expansion) rather than substitution across products. Back-of-the-envelope calculations—annualizing workflow-specific gains and assuming linear additivity—suggest that the four GenAI applications with positive sales effects generate an annual incremental value of approximately \$4.6–\$5 per consumer. These effects represent roughly 5.5–6% of global per-user e-commerce revenue growth in 2023–2024. Taken together, these figures show that even a few GenAI applications already yield substantial gains for a large, mature retailer, with the potential for much larger effects as adoption broadens and increasingly targets revenue-critical workflows. By 2025, the partner platform had deployed GenAI in more than 60 workflows, with usage rising twentyfold as API calls to large language models increased from 2024 to 2025.

At the same time, our study has several limitations that inform the interpretation of the results and point to avenues for future research. First, the adoption horizon in our experiments was short—spanning several weeks to months—so we capture only the immediate, short-run effects of GenAI. Impacts of sustained GenAI use may differ as sellers and consumers adapt their behavior and as platforms refine model deployment. Relatedly, we lack data to assess long-term consumer responses, such as product returns and retention. For instance, while GenAI may initially boost purchases, consumers may ultimately become dissatisfied if the AI-generated content fails to align with actual product characteristics. Second, our analysis is limited to seven workflows that were selected by the platform based on managerial assessments of technical feasibility, organizational costs, and expected productivity improvements, rather than representing the full spectrum of business processes where the technology could be deployed. Other business processes, including logistics, inventory management, or dynamic pricing, remain unexplored and could yield distinct productivity effects. Third, while our estimates map directly into total factor productivity gains under the assumption of constant inputs, we cannot rule out future changes in labor and capital inputs. Many of the studied processes—such as customer service, content creation, and Chargeback Defense—are currently staffed or supported by human labor. Over time, GenAI adoption could displace or augment these functions, yielding additional cost-side efficiency improvements not captured in our current analysis.

Another limitation concerns external validity in general equilibrium. Our experiments were conducted on a single, albeit large, global retail platform. The effects we document partly reflect relative improvements in user experience and satisfaction within this environment. If GenAI adoption becomes widespread and competing platforms deploy similar technologies, these relative advantages may diminish. However, if the mechanism we highlight—enhanced consumer experience through reduced frictions—holds more broadly, there may still be scope for market expansion even in a more competitive environment. In addition, our experiments abstract from strategic responses by competitors —such as changes in pricing or advertising strategy—, which could either amplify or attenuate realized productivity gains.

Taken together, our results highlight both the current scope and the future potential of GenAI adoption in online retail. In the short run, GenAI delivers measurable productivity gains both within workflows and across the platform by reducing market frictions and enhancing the consumer shopping experience, constituting a substantial contribution given the scale and maturity of our focal platform. Over the longer run, its transformative impact could grow if firms expand adoption beyond early-use cases, capture cost-reduction opportunities, and adapt organizational structures to integrate GenAI more effectively. Continued advances in computational speed, accuracy, and domain coverage would further amplify this potential. At the aggregate level, widespread industry adoption raises open questions about equilibrium effects, competitive dynamics, and the durability of observed gains. Our study offers a first step by providing causal evidence on how GenAI can reengineer core retail workflows to deliver meaningful business outcomes, while pointing to important avenues for future research on general equilibrium impacts, cost-side adjustments, and long-term productivity growth.

References

Acemoglu, Daron (2025). “The simple macroeconomics of AI”. In: *Economic Policy* 40.121, pp. 13–58.

Acemoglu, Daron and Pascual Restrepo (2018). “Low-skill and high-skill automation”. In: *Journal of Human Capital* 12.2, pp. 204–232.

Autor, David H., Frank Levy, and Richard J. Murnane (2003). “The skill content of recent technological change: An empirical exploration”. In: *The Quarterly Journal of Economics* 118.4, pp. 1279–1333.

Bai, Jie, Maggie X Chen, Jin Liu, Xiaosheng Mu, and Daniel Yi Xu (2022). *Stand Out from the Millions: Market Congestion and Information Friction on Global E-Commerce Platforms.*

Bar-Isaac, Heski, Guillermo Caruana, and Vicente Cuñat (2012). “Search, Design, and Market Structure”. In: *American Economic Review* 102.2, pp. 1140–1160.

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw (2007). “How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills”. In: *The quarterly journal of Economics* 122.4, pp. 1721–1758.

Belleflamme, Paul and Martin Peitz (2021). *The economics of platforms.* Cambridge University Press.

Bergemann, Dirk and Alessandro Bonatti (2011). “Targeting in Advertising Markets: Implications for Offline versus Online Media”. In: *RAND Journal of Economics* 42.3, pp. 417–443.

Blake, Thomas, Chris Nosko, and Steven Tadelis (2015). “Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment”. In: *Econometrica* 83.1, pp. 155–174.

Bloom, Nicholas and John Van Reenen (2007). “Measuring and explaining management practices across firms and countries”. In: *The quarterly journal of Economics* 122.4, pp. 1351–1408.

Bonney, Kathryn, Cory Breaux, Cathy Buffington, Emin Dinlersoz, Lucia S Foster, Nathan Goldschlag, John C Haltiwanger, Zachary Kroff, and Keith Savage (2024). *Tracking firm use of AI in real time: A snapshot from the Business Trends and Outlook Survey.* Tech. rep. National Bureau of Economic Research.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt (2002). “Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence”. In: *The Quarterly Journal of Economics* 117.1, pp. 339–376.

Brynjolfsson, Erik and Lorin M Hitt (2003). “Computing productivity: Firm-level evidence”. In: *Review of economics and statistics* 85.4, pp. 793–808.

Brynjolfsson, Erik, Yu Hu, and Duncan Simester (2011). “Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales”. In: *Management Science* 57.8, 1373–1386.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond (2025). “Generative AI at Work”. In: *The Quarterly Journal of Economics.*

Brynjolfsson, Erik and Michael D. Smith (2000). “Frictionless Commerce? A Comparison of Internet and Conventional Retailers”. In: *Management Science* 46.4, pp. 563–585.

Cabral, Luis and Ali Hortaçsu (2010). “The Dynamics of Seller Reputation: Evidence from eBay”. In: *Journal of Industrial Economics* 58.1, pp. 54–78.

Calvino, Flavio, Jelmer Reijerink, and Lea Samek (2025). “The effects of generative AI on productivity, innovation and entrepreneurship”. In: *OECD Artificial Intelligence Papers 39.*

Chen, Jiafeng and Jonathan Roth (2024). “Logs with zeros? Some problems and solutions”. In: *The Quarterly Journal of Economics* 139.2, pp. 891–936.

Chen, Y. and S. Yao (2017). “Sequential search with refinement: Model and application with click-stream data”. In: *Management Science* 63.12, pp. 4345–4365.

Chen, Zenan and Jason Chan (2024). "Large language model in creative work: The role of collaboration modality and user expertise". In: *Management Science* 70.12, pp. 9101–9117.

Choi, Jonathan H., Amy Monahan, and Daniel Schwarcz (2023). "Lawyering in the age of artificial intelligence". In: *SSRN* 4626276.

Chui, Michael, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmel (2023). *The Economic Potential of Generative AI: The Next Productivity Frontier*. McKinsey & Company. Accessed: 2025-10-07. URL: `https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier`.

Cui, Zheyuan (Kevin), Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz (2024). "The effects of generative ai on high skilled work: Evidence from three field experiments with software developers". In: *SSRN* 4945566.

De Loecker, Jan (2011). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity". In: *Econometrica* 79.5, pp. 1407–1451. DOI: `10.3982/ECTA7610`.

Dell'Acqua, Fabrizio, Saran Rajendran, Edward McFowland III, Lisa Krayer, Ethan Mollick, François Candelon, Hila Lifshitz-Assaf, Karim R. Lakhani, and Katherine C. Kellogg (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality". In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper*.

Deloitte (2023). "A new rontier in artificial intelligence: Implications of Generative AI for businesses". In: *Deloitte AI Institute*.

Dinerstein, Michael, Liran Einav, Jonathan Levin, and Neel Sundaresan (2018). "Consumer price search and platform design in internet commerce". In: *American Economic Review* 108.7, pp. 1820–1859.

Donati, Dante (2025). "The End of Tourist Traps: The Impact of Review Platforms on Quality Upgrading". In: *Marketing Science*.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock (2023). "Gpts are gpts: An early look at the labor market impact potential of large language models". In: *arXiv preprint arXiv:2303.10130*.

Exner, Yannick, Jochen Hartmann, Oded Netzer, and Shunyuan Zhang (2025). "AI in disguise-How AI-generated ads' visual cues shape consumer perception and performance". In: *Available at SSRN*.

Fan, Ying, Jiandong Ju, and Mo Xiao (2016). "Reputation premium and reputation management: Evidence from the largest e-commerce platform in China". In: *International Journal of Industrial Organization* 46, pp. 63–76.

Fang, Lu, Yanyou Chen, Chiara Farronato, Zhe Yuan, and Yitong Wang (2024). "Platform Information Provision and Consumer Search: A Field Experiment". In: *NBER No. w32099*.

Fradkin, Andrey (2017). "Search, matching, and the role of digital marketplace design in enabling trade: Evidence from Airbnb". In: *Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb (March 21, 2017)*.

Ghose, A., P. G. Ipeirotis, and B. Li (2019). "Modeling consumer footprints on search engines: An interplay with social media". In: *Management Science* 65.3, pp. 1363–1385.

Ghose, Anindya, Avi Goldfarb, and Sang Pil Han (2014). "How Is the Mobile Internet Different? Search Costs and Local Activities". In: *Information Systems Research* 24.3, pp. 613–631.

Goldfarb, Avi and Catherine Tucker (2011). "Online Display Advertising: Targeting and Obtrusiveness". In: *Marketing Science* 30.3, pp. 389–404.

Goldin, Claudia and Lawrence F. Katz (2008). *The race between education and technology.* Harvard Univ. Press.

Gu, C. and Y. Wang (2022). “Consumer online search with partially revealed information”. In: *Management Science* 68.6, pp. 4215–423.

Hartmann, Jochen, Yannick Exner, and Samuel Domdey (2025). “The power of generative marketing: Can generative AI create superhuman visual marketing content?” In: *International Journal of Research in Marketing* 42.1, pp. 13–31.

Heller, David and Dominik Asam (2024). “Generative AI and Firm-level Productivity: Evidence from Startup Funding Dynamics”. In: *Available at SSRN 4877505.*

Honka, Elisabeth (2014). “Quantifying Search and Switching Costs in the U.S. Auto Insurance Industry”. In: *RAND Journal of Economics* 45.4, pp. 847–884.

Hui, Xiang, Oren Reshef, and Luofeng Zhou (2024). “The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market”. In: *Organization Science* 35.6, pp. 1977–1989.

Humlum, Anders and Emilie Vestergaard (2025). *Large Language Models, Small Labor Market Effects.* Tech. rep. National Bureau of Economic Research.

Jin, Ginger Zhe and Andrew Kato (2006). “Price, Quality, and Reputation: Evidence from an Online Field Experiment”. In: *RAND Journal of Economics* 37.4, pp. 983–1005.

Kapoor, Anuj and Madhav Kumar (2025). “Frontiers: Generative AI and personalized video advertisements.” In: *Marketing Science.*

Li, Lingfang, Steven Tadelis, and Xiaolan Zhou (2020). “Buying reputation as a signal of quality: Evidence from an online marketplace”. In: *The RAND Journal of Economics* 51.4, pp. 965–988.

Luca, Michael and Georgios Zervas (2016). “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud”. In: *Management Science* 62.12, pp. 3412–3427.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014). “Promotional Reviews: An Empirical Investigation of Online Review Manipulation”. In: *American Economic Review* 104.8, pp. 2421–2455.

Microsoft (2024). “Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research”. In.

Milgrom, Paul and Steven Tadelis (2018). “How Artificial Intelligence and Machine Learning Can Impact Market Design”. In: *NBER Working Paper No. 24282.*

Nguyen, Nhan and Sarah Nadi (2022). “An empirical evaluation of GitHub copilot’s code suggestions”. In: *Proceedings of the 19th International Conference on Mining Software Repositories.*

Nosko, Chris and Steven Tadelis (2015). “The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment”. In: *NBER Working Paper No. 20830.*

Noy, Shakked and Whitney Zhang (2023). “Experimental evidence on the productivity effects of generative artificial intelligence”. In: *Science* 381, pp. 187–92.

Otis, Nicholas G., Rowan Clarke, Sol‘ene Delecourt, David Holtz, and Rembrand Koning (2024). “The uneven impact of generative AI on entrepreneurial performance”. In: *SSRN* 4671369.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer (2023). “The impact of ai on developer productivity: Evidence from github copilot”. In: *arXiv preprint arXiv:2302.06590.*

Roldán-Monés, Toni (2024). “When GenAI increases inequality: evidence from a university debating competition”. In.

Solow, Robert M (1957). “Technical change and the aggregate production function”. In: *The review of Economics and Statistics* 39.3, pp. 312–320.

Statista (2024). *eCommerce: Market Data & Analysis.* Statista report. Accessed: 2025-10-08. URL: https://www.statista.com/study/42335/ecommerce-report/.

Sun, Tianshu, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu (2024). "The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy". In: *Management Science* 70.4, pp. 2645–2660.

Syverson, Chad (2011). "What Determines Productivity?" In: *Journal of Economic Literature* 49.2, pp. 326–365. DOI: 10.1257/jel.49.2.326.

Tadelis, Steven (2016). "Reputation and feedback systems in online platform markets". In: *Annual review of economics* 8.1, pp. 321–340.

Ursu, Raluca M. (2018). "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions". In: *Marketing Science* 37.4, pp. 530–552.

Wang, H., B. Williams, K. Xie, and W. Chen (2024). "Quality Differentiation and Matching Performance in Peer-to-Peer Markets: Evidence from Airbnb Plus". In: *Management Science* 70.7, pp. 4260–4282.

Yang, Huanxing (2013). "Targeted search and the long tail effect". In: *The RAND Journal of Economics* 44.4, pp. 733–756.

Yang, J., N. S. Sahni, H. S. Nair, and X Xiong (2024). "Advertising as information for ranking e-commerce search listings". In: *Marketing science* 43.2, pp. 360–37.

Yoganarasimhan, Hema (2020). "Search Personalization Using Machine Learning". In: *Management Science* 66.3, pp. 1045–1070.

Zervas, Georgios, Davide Proserpio, and John Byers (2015). "A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average". In: *Working Paper*.

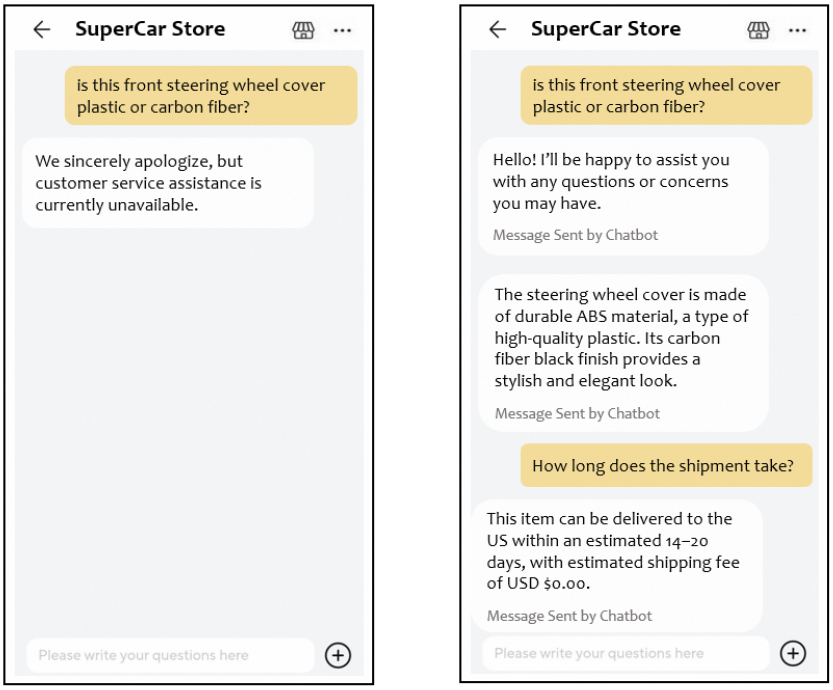
Zhou, Wei, Mingfeng Lin, Mo Xiao, and Lu Fang (2025). "Higher Precision is Not Always Better: Search Algorithm and Consumer Engagement". In: *Management Science* 71.7, pp. 6204–6226.

Appendix

A Illustrative User Interfaces and Examples of Each Experiment

Pre-sale Service Chatbot. In Figure A1, Panel (a) depicts the control condition, where a pre-programmed auto response indicates that no service is available, while Panel (b) presents the treatment condition, featuring support from a GenAI-driven chatbot. The chatbot acts as a virtual sales assistant, available 24/7 to provide instant responses to pre-sale inquiries, covering product features, pricing, availability, and delivery options across multiple languages.

Figure A1: Illustration of Pre-sale Service Chatbot



is this front steering wheel cover plastic or carbon fiber?

We sincerely apologize, but customer service assistance is currently unavailable.

is this front steering wheel cover plastic or carbon fiber?

Hello! I'll be happy to assist you with any questions or concerns you may have.

Message Sent by Chatbot

The steering wheel cover is made of durable ABS material, a type of high-quality plastic. Its carbon fiber black finish provides a stylish and elegant look.

Message Sent by Chatbot

How long does the shipment take?

This item can be delivered to the US within an estimated 14–20 days, with estimated shipping fee of USD $0.00.

Message Sent by Chatbot

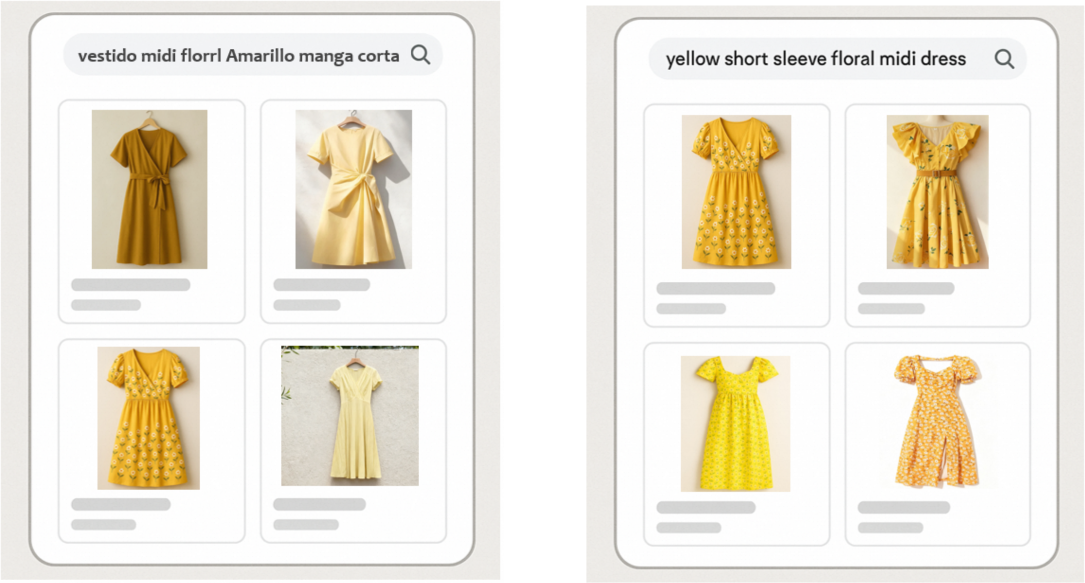
(a) Auto-Response (No Service)

(b) GenAI-Driven Chatbot

Search Query Refinement. In Figure A2, Panel (a) displays the consumer's original search query in spanish along with the corresponding results, while Panel (b) presents the structured English query translated and refined by GenAI, along with the corresponding search results retrieved. The GenAI-powered query refinement can improve the expression of consumer demand and facilitate the matching efficiency of search engines.

Product Description. In Figure A3, Panel (a) illustrates the control condition with human-generated descriptions, while Panel (b) shows the treatment condition, where GenAI-created descriptions are layered on top of those written by humans. The GenAI-generated content provides more comprehensive and structured bullet-point-style descriptions that highlight product features, benefits, and typical use cases.

Figure A2: Illustration of Search Query Refinement



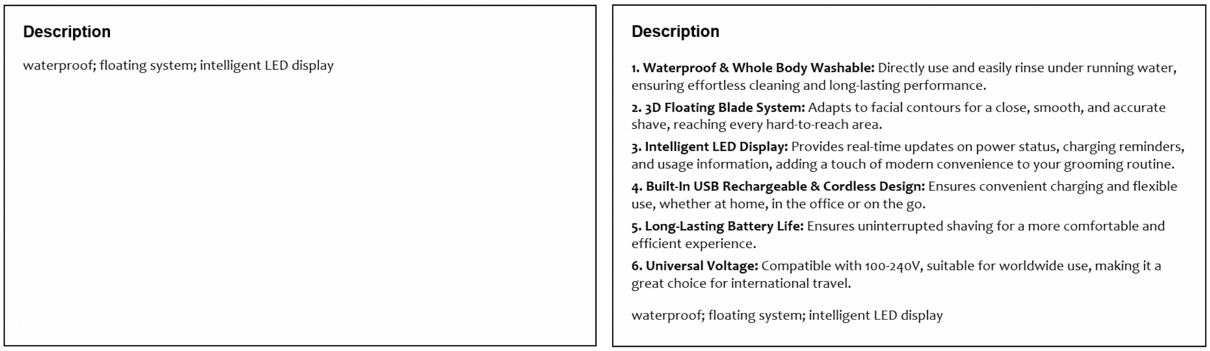
vestido midi florrl Amarillo manga corta

yellow short sleeve floral midi dress

(a) Search Results of Consumer Query

(b) Search Results of GenAI-Refined Query

Figure A3: Illustration of Product Description



Description

waterproof; floating system; intelligent LED display

Description

1. **Waterproof & Whole Body Washable:** Directly use and easily rinse under running water, ensuring effortless cleaning and long-lasting performance.
2. **3D Floating Blade System:** Adapts to facial contours for a close, smooth, and accurate shave, reaching every hard-to-reach area.
3. **Intelligent LED Display:** Provides real-time updates on power status, charging reminders, and usage information, adding a touch of modern convenience to your grooming routine.
4. **Built-In USB Rechargeable & Cordless Design:** Ensures convenient charging and flexible use, whether at home, in the office or on the go.
5. **Long-Lasting Battery Life:** Ensures uninterrupted shaving for a more comfortable and efficient experience.
6. **Universal Voltage:** Compatible with 100-240V, suitable for worldwide use, making it a great choice for international travel.

waterproof; floating system; intelligent LED display

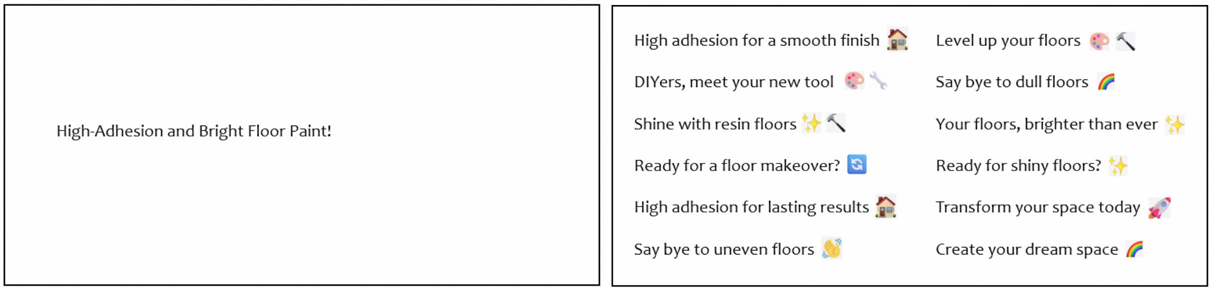
(a) Human-Created Description

(b) GenAI-Created Description Ahead of Original Human Input

Marketing Push Message. In Figure A4, Panel (a) illustrates a human-generated marketing push message, whereas Panel (b) displays multiple messages produced by GenAI for the same product. Generative AI enables large-scale creation of diverse marketing content, increasing the likelihood that consumers encounter differentiated messages and thereby allowing platforms to leverage the benefits of personalized marketing.

Google Advertising Title. In this case, GenAI is applied to optimize human-generated product titles for Google advertising. Since the model was not fine-tuned with e-commerce domain knowledge, the performance of GenAI-optimized titles is lower than that of human-generated titles. For example, the original human-generated title for a pair of sunglasses is: “2024 New Arrival Polarized

Figure A4: Illustration of Marketing Push Message



High-Adhesion and Bright Floor Paint!

High adhesion for a smooth finish 🏠
DIYers, meet your new tool 🎨🔧
Shine with resin floors ✨🔨
Ready for a floor makeover? 🔄
High adhesion for lasting results 🏠
Say bye to uneven floors 👋

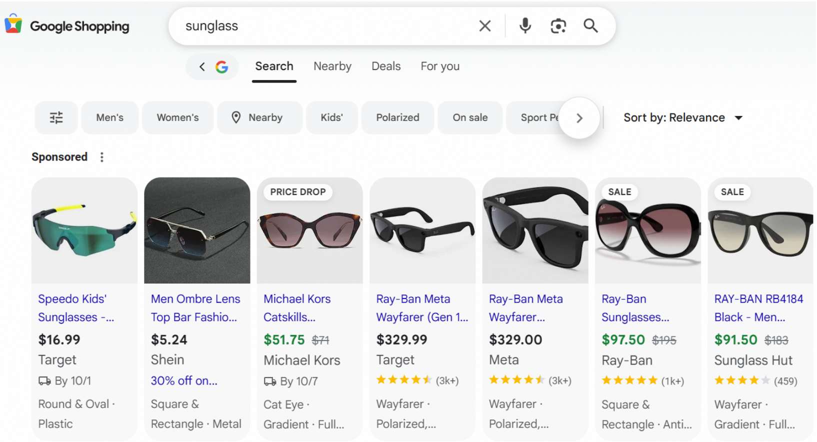
Level up your floors 🎨🔨
Say bye to dull floors 🌈
Your floors, brighter than ever ✨
Ready for shiny floors? ✨
Transform your space today 🚀
Create your dream space 🌈

(a) Human-Created Push Message

(b) GenAI-Created Push Messages

Pitboss 2 Sunglasses Men Cycling Eyewear Goggles Bicycle Glasses". The GenAI-optimized version is: "Men's Polarized Pitboss 2 Sunglasses - Polycarbonate Frame for Cycling, Sports, Bike Goggles Bicycle". On Google Shopping, the first few words of a product title are the most prominent, as shown in Figure A5. Thus, by removing popular keywords such as "New Arrival," GenAI may reduce consumer attention.

Figure A5: Illustration of Google Shopping Interface



Speedo Kids' Sunglasses -... $16.99 Target By 10/1 Round & Oval · Plastic

Men Ombre Lens Top Bar Fashio... $5.24 Shein 30% off on... Square & Rectangle · Metal

Michael Kors Catskills... $51.75 ~~$71~~ Michael Kors By 10/7 Cat Eye · Gradient · Full...

Ray-Ban Meta Wayfarer (Gen 1... $329.99 Target ★★★★★ (3k+) Wayfarer · Polarized,...

Ray-Ban Meta Wayfarer... $329.00 Meta ★★★★★ (3k+) Wayfarer · Polarized,...

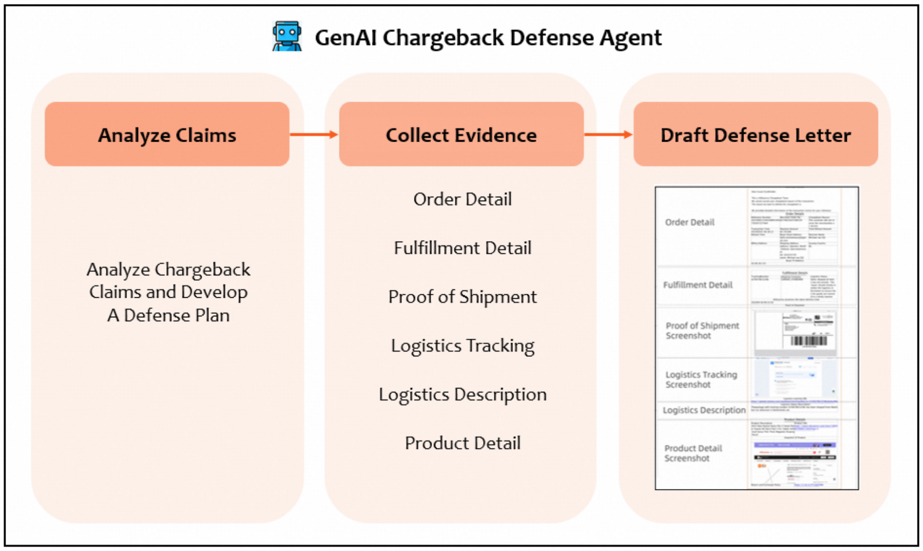
Ray-Ban Sunglasses... $97.50 ~~$195~~ Ray-Ban ★★★★★ (1k+) Square & Rectangle · Anti...

RAY-BAN RB4184 Black - Men... $91.50 ~~$183~~ Sunglass Hut ★★★★☆ (459) Wayfarer · Gradient · Full...

Chargeback Defense. Figure A6 illustrates the process by which the chargeback defense agent interprets and analyzes consumer claims, gathers relevant evidence—including transaction records, product details, and shipping information—and drafts persuasive defense letters. By automating this complex workflow, GenAI enables sellers to respond to chargeback claims more quickly and consistently, thereby improving win rates while reducing compliance burdens and financial losses.

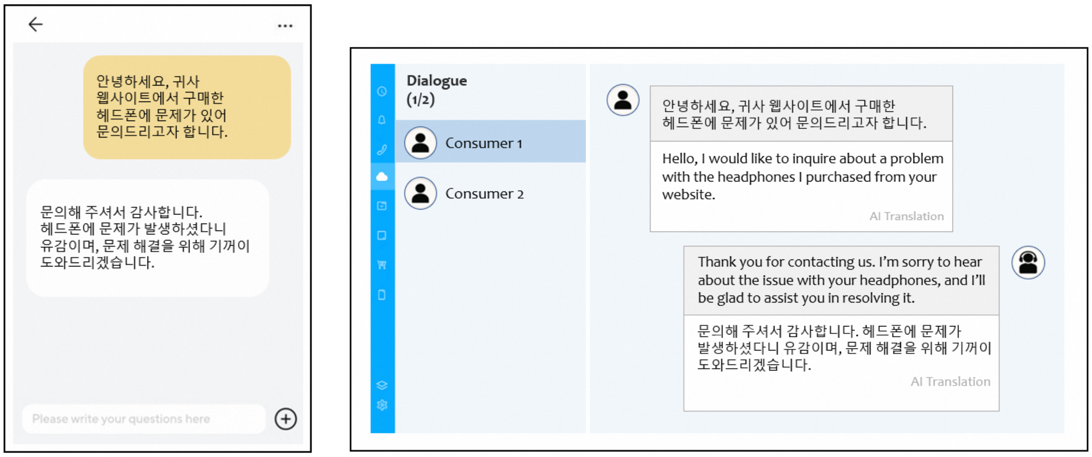
Live Chat Translation. Figure A7 depicts the live chat translation system. Panel (a) shows the consumer interface, where a Korean consumer submits a query. Panel (b) presents the service agent interface, where a Filipino agent receives the query translated in real time from Korean to English by GenAI. The system supports bidirectional translation: the agent's English response is simultaneously translated into Korean, allowing the consumer to receive the reply in their native

Figure A6: Illustration of Chargeback Defense



language. This functionality enables English-speaking Filipino agents to communicate seamlessly with consumers across multiple languages on the platform.

Figure A7: Illustration of Live Chat Translation



안녕하세요, 귀사
웹사이트에서 구매한
헤드폰에 문제가 있어
문의드리고자 합니다.

문의해 주셔서 감사합니다.
헤드폰에 문제가 발생하셨다니
유감이며, 문제 해결을 위해 기꺼이
도와드리겠습니다.

안녕하세요, 귀사 웹사이트에서 구매한
헤드폰에 문제가 있어 문의드리고자 합니다.

Hello, I would like to inquire about a problem with the headphones I purchased from your website.

AI Translation

Thank you for contacting us. I’m sorry to hear about the issue with your headphones, and I’ll be glad to assist you in resolving it.

문의해 주셔서 감사합니다. 헤드폰에 문제가
발생하셨다니 유감이며, 문제 해결을 위해 기꺼이
도와드리겠습니다.

AI Translation

(a) Consumer Interface

(b) Customer Service Agent Interface

B Covariance Balance Checks for the Field Experiments

In this section, we present detailed covariance balance checks for each of the five experiments for which granular data are available.

Table B1: Covariate Balance and Randonmization Check

	Control	Treatment	*p-value* (C=T)
Pre-sale Service Chatbot			
Gender	1.000 (0.584)	1.005 (0.438)	0.517
Age Tier	1.000 (0.267)	1.001 (0.187)	0.853
Registered Years	1.000 (0.937)	0.994 (0.703)	0.570
Past Login Days	1.000 (0.080)	1.002 (0.080)	0.160
Past Orders	1.000 (3.108)	0.977 (1.959)	0.483
Past Sales	1.000 (3.875)	1.005 (2.759)	0.913
N. of Consumers	15,457	29,157	
Search Query Refinement			
Gender	1.000 (0.799)	1.001 (0.799)	0.204
Age Tier	1.000 (0.289)	1.000 (0.289)	0.466
Registered Years	1.000 (0.860)	1.001 (0.861)	0.333
Past Login Days	1.000 (1.051)	1.002 (1.054)	0.207
Past Orders	1.000 (3.214)	1.007 (3.468)	0.140
Past Sales	1.000 (6.850)	1.014 (8.206)	0.217
N. of Consumers	929,188	920,194	
Product Description			
Gender	1.000 (0.912)	1.000 (0.912)	0.778
Age Tier	1.000 (0.283)	1.000 (0.283)	0.127
Registered Years	1.000 (0.942)	1.001 (0.942)	0.192
Past Login Days	1.000 (0.942)	0.999 (0.944)	0.490
Past Orders	1.000 (2.437)	1.001 (2.516)	0.627
Past Sales	1.000 (3.613)	0.998 (3.613)	0.515
N. of Consumers	2,392,803	2,380,134	
Marketing Push Message			
Gender	1.000 (0.894)	1.000 (0.894)	0.599
Age Tier	1.000 (0.275)	1.000 (0.277)	0.538
Registered Years	1.000 (1.009)	1.000 (1.008)	0.714
Past Login Days	1.000 (2.386)	0.999 (2.371)	0.501
Past Orders	1.000 (3.304)	1.003 (3.062)	0.157
Past Sales	1.000 (6.115)	1.004 (5.278)	0.157
N. of Consumers	6,869,558	6,845,970	
Google Advertising Title			
Past Sales	1.000 (2.261)	0.993 (2.238)	0.084
Industry ID	1.000 (0.414)	1.000 (0.416)	0.712
N. of Products	621,133	623,001	

[1] Mean (Std. Dev.) are shown with all values normalized by the corresponding variable's mean in the control group. For the definitions of each variable, please refer to the notes in Figure 1.

[2] The first four experiments are conducted at the consumer level and thus the unit of observation is consumer. The Google Advertising Title experiment is conducted at the product level and thus the unit of observation is product.

C Main Results: Model, Estimation, and Additional Outcomes

Pre-sale Service Chatbot. Pre-sale inquiries regarding product and seller information (e.g., product attributes, promotions, and logistics) play a critical role in shaping consumer purchase decisions. To support decision-making, improve consumer service, and enhance the overall consumer experience, the platform introduced a GenAI-powered chatbot capable of delivering accurate, content-rich responses tailored to a diverse consumer base and available around the clock.

We conduct our analysis using the following regression model:

$$y_i = \beta \times Treat_i + \alpha_{c(i)} + \epsilon_i, \tag{3}$$

where i indicates the consumer, y_i stands for a consumer's outcome (e.g., conversion rate or sales), $Treat_i$ is an indicator for whether the consumer is assigned to the treatment group. Since consumers entered the experiments on different days, we control for their entry-day cohort fixed effects using $\alpha_{c(i)}$.

In addition to the main experiment comparing an auto-response indicating service unavailability ("No Service") with a GenAI chatbot service ("GenAI Reply"), we also studied three supplementary experiments: (1) "No Service" versus "GenAI+Human Reply", where consumers initially interacted with a GenAI chatbot and unresolved issues were escalated to human agents; (2) "Human Reply", where consumers were exclusively served by human agents, versus "GenAI Reply"; (3) "Human Reply" versus "GenAI+Human Reply".

In Table C1, Columns (1)–(4) report effects when the treatment is GenAI Reply, and Columns (5)–(8) report effects for GenAI+Human Reply. Within each set, Columns (1)–(2) and (5)–(6) use No Serice as the control, while Columns (3)–(4) and (7)–(8) use Human Reply as the control. Focusing on the comparison between the No Service control and the GenAI Reply treatment (Cols. 1–2), we continue to find sizable productivity improvements: the conversion rate rises by 21.7% and sales increase by 16.3% (both significant at the 1% level). When the No Service control is compared with the GenAI+Human Reply treatment (Cols. 5–6), the gains are even larger—conversion improves by 29.0% and sales by 25.0%—indicating complementarities between GenAI and human agents. Using Human Reply as the control, the GenAI Reply treatment shows no statistically significant differences in either conversion or sales (Cols. 3–4), suggesting that the GenAI chatbot matches the quality of human service but does not outperform it. By contrast, relative to the Human Reply control, the GenAI+Human Reply treatment yields a small, statistically insignificant increase in conversion (4.8%) and a marginally significant 11.5% increase in sales (Cols. 7–8), implying that the hybrid approach can enhance revenue even when benchmarked against human agents. Table C2 reports the Cart Value in the Pre-sale Service Chatbot experiment. We find that all Cart Values are statistically insignificant.

Search Query Refinement. Consumers arrive at e-commerce platforms with diverse needs. The search engine serves as the primary channel to facilitate consumers' discovery of desired products, allowing them to express preferences through search queries. Our focal platform seeks to accurately decode the latent demands behind consumers' multilingual queries, translate the queries, and retrieve products that align with their underlying needs. The effectiveness of this process is crucial in determining match quality, which in turn impacts consumer purchase decisions and platform revenues. GenAI is well-positioned to improve the search algorithm's capabilities in translating

Table C1: Main Effect of Pre-sale Service Chatbot

Treatment:	GenAI reply		GenAI reply		GenAI+Human reply		GenAI+Human reply	
Control:	No Service		Human Reply		No Service		Human Reply	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Conv Rate	Sales	Conv Rate	Sales	Conv Rate	Sales	Conv Rate	Sales
Treat	0.0131***	0.274***	-0.000768	0.0701	0.0175***	0.422***	0.00358	0.218*
	(0.00256)	(0.0995)	(0.00261)	(0.0992)	(0.00295)	(0.115)	(0.00301)	(0.1145)
%Change	21.7%	16.3%	-1.0%	3.7%	29.0%	25.0%	4.8%	11.5%
Observations	44,614	44,614	44,736	44,736	30,345	30,345	30,467	30,467
R-squared	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000

[1] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Table C2: Cart Value in Pre-sale Service Chatbot Experiment

Treatment:	GenAI reply	GenAI reply	GenAI+Human reply	GenAI+Human reply
Control:	No Service	Human Reply	No Service	Human Reply
	(1)	(2)	(3)	(4)
Treat	-0.859	1.624	-1.264	1.220
	(1.203)	(1.078)	(1.036)	(0.929)
%Change	-3.1%	6.4%	-4.5%	4.8%
Observations	2,092	2,316	3,076	3,300
R-squared	0.000	0.001	0.000	0.001

[1] The dependent variable is the "Cart Value", which refers to the expenditure per consumer, conditional on making a purchase. All analyses are restricted to those who make a purchase.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

consumer queries based on semantic understanding and refinement.

We conduct our analysis using the following regression model:

$$y_i = \beta \times Treat_i + \alpha_{cl(i)} + \epsilon_i, \tag{4}$$

where i denotes a consumer, y_i stands for a consumer's outcome variables, $Treat_i$ is an indicator of a consumer's treatment status. Consumers are classified into various cohorts based on their language groups and their first day of entry in the experiment. As multiple sub-experiments were conducted across consumers in different languages at varying dates, we include entry-day-by-language cohort fixed effects, $\alpha_{cl(i)}$.

The results are summarized in Table C3. Column 1 indicated no significant differences in product views between the two groups. However, treatment group consumers generated 1.10% more clicks (Column 2), spent 2.93% more (Column 4), and were 1.15% more likely to make a purchase (Column 5). These effects were primarily driven by a significant 2.02% rise in click-through rate (Column 6)—the ratio of product clicks to views—suggesting that consumers found the exposed products more appealing and chose to seek additional details after viewing the summarized search results. Furthermore, the click-to-order conversion rate (Column 7)—the ratio of product orders to clicks—remained insignificant, echoing the fact that query refinement influenced only the composition of products retrieved immediately after a query search, not the information displayed on product detail pages. Overall, GenAI-facilitated query refinement enhanced the search algorithm's ability to satisfy consumers' demand, resulting in more effective matching and improved consumer purchases.

Table C3: Main Effect of Search Query Refinement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
							Click-	Click-to-Order
					Conversion	Cart	through	Conversion
	View	Click	Order	Sales	Rate	Value	Rate	Rate
Treat	-0.549	0.0901***	0.0015	0.0648**	0.0010**	0.376	0.0008***	-0.0001
	(0.887)	(0.0245)	(0.0011)	(0.0314)	(0.0004)	(0.334)	(0.0001)	(0.0002)
%Change	-0.18%	1.10%	0.94%	2.93%	1.15%	1.49%	2.02%	-0.35%
Observations	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	163,471	1,849,382	1,508,873
R-squared	0.038	0.041	0.019	0.004	0.030	0.010	0.003	0.014

[1] "View" stands for the number of product views in the initial search results pages. "Click" stands for the number of product clicks into product detail pages. "Order" is the number of product orders. "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase. "Click-through Rate" is the ratio of the number of product clicks to the number of product views. It measures the degree of conversion from views to clicks. "Click-to-Order Conversion Rate" stands for the ratio of the number of product orders to the number of product clicks. It measures the degree of conversion from clicks to purchases.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Product Description. Well-crafted product descriptions are essential for informing consumers about product features, benefits, and uses, thereby reducing information asymmetry, facilitating consumer decision-making and driving platform sales. Despite its importance, our studied platform shows that nearly half of the self-sold products either lack a textual description or contain only a minimal description. GenAI's strengths in content recognition, comprehension, and generation can offer an effective solution to create comprehensive and structured product descriptions for a global audience.

We conduct our analysis using the following regression model:

$$y_i = \beta \times Treat_i + \alpha_{cl(i)} + \epsilon_i, \tag{5}$$

where i denotes a consumer, y_i stands for a consumer's outcome variables, $Treat_i$ is an indicator for whether the consumer is assigned to the treatment group. Because multiple sub-experiments were implemented across language groups on different start dates, we include entry-day-by-language cohort fixed effects, $\alpha_{cl(i)}$.

In Column 1 of Table C4, we found no significant differences in product clicks between the control and treatment groups. However, once consumers clicked and accessed the product detail pages where descriptions were displayed, the treatment group consumers tended to place 1.08% more orders and spend 2.05% more on those orders (Columns 2 and 3). This improvement was also evidenced by the 1.27% increase in the conversion rate in Column 4, indicating that GenAI-generated descriptions promoted a higher likelihood of purchasing. Consequently, augmenting human-generated product descriptions with GenAI-generated content effectively motivated consumers' purchase decisions, leading to higher sales for the platform.

Table C4: Main Effect of Product Description

	(1) Click	(2) Order	(3) Sales	(4) Conversion Rate	(5) Cart Value
Treat	0.0023	0.0006**	0.0104**	0.0006***	0.0942
	(0.0018)	(0.0003)	(0.0042)	(0.0002)	(0.0807)
%Change	0.12%	1.08%	2.05%	1.27%	0.81%
Observations	4,772,937	4,772,937	4,772,937	4,772,937	209,371
R-squared	0.055	0.008	0.002	0.008	0.010

[1] "Click" stands for the number of product clicks into product detail pages. "Order" is the number of product orders. "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Marketing Push Message. The platform leverages marketing push notifications, direct messages sent to consumers via their platform app, to draw traffic and boost transactions. Because creating a large volume of diverse and targeted marketing messages manually was challenging, the number of marketing messages is far smaller than the hundreds of millions of consumers, often resulting in many consumers receiving identical content. With the introduction of GenAI, the platform can produce millions of distinct messages, enabling highly personalized marketing strategies through push notifications.

We conduct our analysis using the following regression model:

$$y_i = \alpha + \beta \times Treat_i + \epsilon_i \tag{6}$$

where i denotes a consumer, y_i stands for a consumer's outcome variables, $Treat_i$ is an indicator for whether the consumer is assigned to the treatment group.

In Table C5, Column 1 reveals that GenAI-generated content accounts for 40% of the marketing push message in the treatment group. With the use of GenAI-generated messages, there are observable rises in consumer clicks and purchases, with the number of clicks increasing by 3.0%, the number of orders increasing by 2.8%, and the purchase amount growing by 1.6% (Columns 2 to 4). Additionally, the likelihood of a consumer making a purchase increased by 3.1% in Column 5. The Cart Value is not statistically significant (Column 6). The results highlight the benefits of using GenAI in unlocking the personalization potential of marketing content, particularly in environments where human-generated content alone may be limited by resource constraints.

Table C5: Main Effect of Marketing Push

	(1)	(2)	(3)	(4)	(5)	(6)
	Is AI Task	Click	Order	Sales	Conversion Rate	Cart Value
Treatment	0.394***	0.000048**	0.00005*	0.000402	0.000529***	0.024
	(0.000187)	(0.00002)	(0.00002)	(0.000812)	(0.00007)	(0.473)
%Change		3.0%	2.8%	1.6%	3.1%	0.15%
Observations	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528	11,030
R-squared	0.245	0.000	0.000	0.000	0.000	0.000

[1] "Is AI Task" is a binary variable which equals one if the message is generated by AI. "Click" stands for the number of clicks on the marketing messages. "Order" is the number of product orders. "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Google Advertising Titles The platform buys advertising slots in the sponsored section of Google Shopping to promote its products on Google and attract traffic to its site. To maximize purchases derived from Google advertisements, a critical operational decision for the platform is how to design product titles to increase both product discoverability and the likelihood of user clicks. The platform leveraged GenAI to refine titles based on the original seller-created titles, optimizing them for better visibility and engagement.

We conduct our analysis employing the following regression model:

$$y_i = \beta \times Treat_i + \alpha_{c(i)} + \epsilon_i, \tag{7}$$

where i denotes a product, y_i stands for the outcome variables for a product, $Treat_i$ is an indicator for whether the product is assigned to the treatment group. We control for product entry-day cohort fixed effect α_{ci}.

Table C6 presents the main findings. Columns 1 and 2 indicated a 7.6% decrease in ad views and an 10.2% decrease in ad clicks for the treatment group, respectively. Columns 3, 4 and 5 reported non-significant reduction in sales, conversion rate, and cart value, respectively. As we discussed in Section 4.1, the null effect on sales can be attributed to the lack of fine-tuning using e-commerce domain knowledge when setting up the GenAI model.

Table C6: Main Effect of Google Advertising

	(1) View	(2) Click	(3) Sales	(4) Conversion Rate	(5) Cart Value
treat	-1.547***	-0.0247***	-0.00602	-0.000137	-0.784
	(0.148)	(0.00304)	(0.00534)	(0.000124)	(0.992)
%Change	-7.6%	-10.2%	-4.5%	-3.3%	-2.3%
Observations	1,244,016	1,244,016	1,244,016	1,244,016	2,437
R-squared	0.000	0.000	0.000	0.000	0.000

[1] "View" represents the number of times the product advertisement is viewed on Google. "Click" refers to the number of times the product advertisement is clicked on Google. "Sales" represents the total expenditure on product orders. "Conversion Rate" measures users' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Chargeback Defense Online sellers often struggle to defend against chargebacks due to reasons such as non-receipt of goods. Chargebacks can lead to significant financial losses and jeopardize the long-term sustainability of sellers' businesses on the platform. The whole process of contesting chargebacks disputes includes analyzing claims, collecting necessary evidence (e.g., order details, fulfillment records, proof of shipment and logistics tracking through ways like interfacing with diverse APIs), and crafting compelling chargeback defense letters. The rapid advancement of GenAI enabled the platform to develop a chargeback defense agent that offers a one-stop solution to streamlining the intricacies of chargeback disputes.

As the data was not available for us to review, we report findings estimated by the platform's internal data science team. Their estimates indicate that the adoption of the GenAI agent helps increase sellers' success rate of chargeback defense by 15%.

Live Chat Translation E-commerce platforms must provide robust consumer services for consumers seeking consultation or negotiation with the platform, such as addressing inquiries about the platform's promotional details and resolving disputes when consensus with sellers is not reached. For our focal platform, delivering native-language consumer services to a diverse, multilingual consumer base incurs significant costs. Thus, a significant portion of non-English consumer inquires were supported by Filipino consumer service agents in English. A straightforward application of GenAI allows the platform to equip low-cost Filipino agents with robust real-time translation support, aiming to provide more effective communication between consumers and agents.

Table D1: Seller HTE for Search Query Refinement

	(1)	(2)	(3)	(4)	(5)	(6)
	Annual Sales		Operation Years		# of Sub-Accounts	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.0241	0.0407**	0.0143	0.0505**	0.00477	0.0600**
	(0.0243)	(0.0180)	(0.0165)	(0.0251)	(0.0154)	(0.0261)
%Change	2.18%	3.68%	2.28%	3.19%	0.97%	3.48%
Panel B: Conversion Rate						
Treat	0.00008	0.00105***	0.000369	0.000715*	0.000156	0.000936**
	(0.000291)	(0.000352)	(0.000249)	(0.000374)	(0.000194)	(0.000390)
%Change	0.21%	1.69%	1.24%	1.01%	0.88%	1.20%
Observation	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382

[1] We classify sellers into high and low groups based on three different proxies. Low-group sellers are generally small sellers, as defined by meeting the following pre-experiment criteria: (1) accounting for the bottom 50% cumulative share of total sales when sellers are ranked by annual sales; (2) having operated on the platform for fewer than five years; or (3) maintaining fewer than three sub-accounts for their online store.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Similarly, due to the unavailability of raw data, the effect estimate for this process is from the platform's internal data science team that we couldn't verify. During the experiment, consumers were queried whether they were satisfied or not with the service via a survey question immediately after service completion. As a result, there was a 5.2% increase in consumer satisfaction, suggesting that GenAI helped Filipino agents to better serve non-English-speaking consumers.

D Details on the Heterogeneous Treatment Effects Across Seller

Search Query Refinement. Table D1 reports the heterogeneous impact on high- versus low-type of sellers classified based on three metrics: annual past sales, tenure on the platform, and number of sub-accounts. Small sellers with lower transaction volumes (Column 2), shorter operational histories (Column 4), and fewer sub-accounts (Column 6) experience a significant increase in both sales (Panel A) and conversion rates (Panel B) from treated consumers. In contrast, expenditures and conversion rates on big, tenured and scaled sellers show no significant change. Thus, the GenAI-powered search query refinement generates more values among small sellers.

Marketing Push Message. Table D2 exhibited the heterogeneous impact of Marekting Push Message. Similarly, we observe that small sellers experience a significant increase in conversion rates from treated consumers (Columns 2, 4 and 6 in Panel B), while the conversion rates on big sellers show no significant changes between treatment and control groups. Thus, GenAI helps the platform to more effectively leverage content personalization in its marketing activities, particularly benefiting small sellers.

Google Advertising Titles. Table D3 shows that the heterogeneous treatment effects are statistically insignificant across both high- and low-seller groups, for sales as well as conversion rate.

Table D2: Seller HTE for Marketing Push Message

	(1)	(2)	(3)	(4)	(5)	(6)
	Annual Seller Sales		Operation Years		# of Sub-Accounts	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.000217	0.000278	0.000354	0.000142	0.000635	-0.000140
	(0.000594)	(0.000559)	(0.000520)	(0.000628)	(0.000419)	(0.000700)
%Change	1.9%	2.1%	6.2%	0.7%	7.5%	-0.8%
Panel B: Conversion Rate						
Treat	0.0000241	0.0000278*	0.000088	0.000043**	-0.000058	0.000052**
	(0.0000147)	(0.0000161)	(0.0000134)	(0.0000172)	(0.0000117)	(0.0000185)
%Change	3.3%	3.2%	1.9%	3.8%	-0.1%	5.3%
Observation	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528

[1] Please refer to Table D1 for detailed notes.

Table D3: Seller HTE for Google Advertising Titles

	(1)	(2)	(3)	(4)	(5)	(6)
	Annual Seller Sales		Operation Years		# of Sub-Accounts	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	-0.00605	-0.00610	-0.00656	-0.00578	-0.00824	-0.00574
	(0.00711)	(0.00797)	(0.00935)	(0.00650)	(0.0181)	(0.00556)
% Change	-5.1%	-4.4%	-5.4%	-4.4%	-5.9%	-4.5%
Panel B: Conversion Rate						
Treat	-0.000261	-0.0000139	-0.000083	-0.000163	-0.000163	-0.000134
	(0.000176)	(0.000175)	(0.000196)	(0.000157)	(0.000326)	(0.000134)
% Change	-6.5%	-0.3%	-2.3%	-3.8%	-4.7%	-3.2%
Observations	622,083	621,933	397,339	846,677	140,005	1,104,011

[1] Please refer to Table D1 for detailed notes.

Table E1: Consumer HTE for Pre-sale Service Chatbot

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Sales	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.268**	0.283**	0.318**	0.230**	0.194	0.314***
	(0.134)	(0.135)	(0.156)	(0.114)	(0.166)	(0.109)
%Change	13.7%	22.4%	15.0%	18.5%	8.6%	25.9%
Panel B: Conversion Rate						
Treat	0.0134***	0.0128***	0.0128***	0.0135***	0.0134***	0.0120***
	(0.00340)	(0.00368)	(0.00386)	(0.00321)	(0.00407)	(0.00310)
%Change	19.8%	26.1%	17.1%	29.6%	17.6%	25.4%
Observations	26,984	17,612	22,570	22,026	20,934	23,662

[1] We classify consumers into high and low groups and capture their online shopping experience using three metrics. Consumers in the low group are considered relatively inexperienced, as determined by meeting the following pre-experiment criteria: (1) a registration duration below the median among all consumer on the platform; (2) a number of 30-day login days below the platform-wide median; or (3) a 30-day total purchase amount below the platform-wide median.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

E Details on the Heterogeneous Treatment Effects Across Consumers

Pre-sale Service Chatbot. In this experiment, we focus on the comparison between consumers who are served only by GenAI chatbot with those who are served by an auto-response indicating service unavailability. Table E1 reports the heterogeneous treatment effects across consumer groups. Our analysis reveals that the increases of both sales (Panel A) and converstion rate (Panel B) are more pronounced among newer consumers with shorter registration histories (Column 2), less active consumers with fewer logins days (Column 4), and lighter consumers with less past spending (Column 6).

Search Query Refinement. Table E2 indicates the heterogeneous treatment effect of refining search query with GenAI on different consumer groups. For both sales (Panel A) and conversion rate (Panel B), the benefits are only significant for inexperienced consumers (Columns 2, 4 and 6). Since consumers with limited online experience often struggle to effectively articulate their needs through query-based searches, they tend to benefit more from enhanced match quality achieved by utilizing GenAI to translate and refine consumer queries.

Product Description. In Table E3, we observe similar pattern. There is significant and more pronounced sales increases for inexperienced consumers (Columns 2, 4 and 6). Augmenting product descriptions with GenAI-generated content substantially enhanced the sufficiency and clarity of product information, thereby motivating consumer purchase decisions, particularly among less experienced consumers.

Marketing Push Message. Table E4 presents the buyer-level heterogeneous treatment effects of the Marketing Push Message experiment. For sales (Panel A), we find a significant increase among

Table E2: Consumer HTE for Search Query Refinement

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Sales	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.0360	0.0867**	0.0203	0.104***	-0.0216	0.106***
	(0.0526)	(0.0380)	(0.0565)	(0.0302)	(0.0757)	(0.0291)
%Change	1.27%	5.01%	0.63%	8.16%	-0.56%	7.46%
Panel B: Conversion Rate						
Treat	-0.000104	0.00186***	0.000500	0.00141***	0.000651	0.00115***
	(0.000697)	(0.000487)	(0.000675)	(0.000482)	(0.000907)	(0.000421)
%Change	-0.09%	2.79%	0.43%	2.32%	0.44%	1.93%
Observation	813,317	1,036,065	890,044	959,338	596,721	1,252,661

[1] Please refer to Table E1 for detailed notes.

Table E3: Consumer HTE for Product Description

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Sales	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.00671	0.0130***	-0.00606	0.0263***	0.000224	0.0206***
	(0.00762)	(0.00455)	(0.00655)	(0.00523)	(0.00718)	(0.00423)
%Change	1.09%	3.06%	-1.02%	6.24%	0.03%	5.63%
Panel B: Conversion Rate						
Treat	0.000575*	0.000520**	0.000325	0.000785***	0.000568*	0.000533**
	(0.000312)	(0.000229)	(0.000284)	(0.000245)	(0.000292)	(0.000233)
%Change	1.10%	1.39%	0.65%	2.06%	1.05%	1.59%
Observation	2,035,278	2,737,659	2,322,437	2,450,500	2,386,336	2,386,601

[1] Please refer to Table E1 for detailed notes.

Table E4: Consumer HTE for Marketing Push Message

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Sales	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.000181	0.00138	0.000655	0.000385	-0.00261	0.00264***
	(0.00115)	(0.00112)	(0.00136)	(0.00100)	(0.00161)	(0.000713)
% Change	0.6%	7.2%	2.3%	1.7%	-6.1%	22.9%
Panel B: Conversion Rate						
Treat	0.0000336	0.0000754**	0.000008	0.0000843***	-0.0000613	0.000121***
	(0.0000301)	(0.0000312)	(0.0000345)	(0.0000281)	(0.0000415)	(0.0000210)
% Change	1.9%	5.5%	0.5%	5.5%	-2.3%	15.4%
Observations	7,959,821	5,755,707	5,796,905	7,918,623	6,085,062	7,630,466

[1] Please refer to Table E1 for detailed notes.

consumers with lower spending in the 30-day pre-experiment window (Column 6). For conversion rate (Panel B), the analysis indicates that only inexperienced consumers derive significant benefits from GenAI-driven marketing push activities (Columns 2, 4, and 6).

F Details on the Heterogeneous Treatment Effects Across Products

Pre-sale Service Chatbot Table F1 reports the product-level heterogeneous treatment effects of the Pre-sale Service Chatbot experiment. We detect tail products (Column 4) benefiting more from GenAI adoption. The results for market concentration (Columns 1 and 2) and price levels (Columns 5 and 6) tend to be nuanced.

Search Query Refinement. In Table F2, we find that the gains are only significant for products in low-concentration categories (Column 2), products with fewer sales volume (Column 4), and products with high price level (Column 5).

Product Description. In Table F3, for market concentration and price level, we only observe significant gains for high-concentration categories (Column 1) and high-priced products (Column 5). The conversion rate increase is only significant for head products while the sales increase is significant for both head and tail products (Columns 3-4).

Marketing Push Message. Table F4 presents that tail products (Column 4) and high-priced products (Column 5) receive a significant increase in conversion rate (Panel B). In terms of the market concentration, products in low-concentration categories experience a significant increase in conversion rate from treated consumers (Column 2).

Google Advertising Titles. In Table F5, Panels A and B report the heterogeneity in Sales and conversion rate, respectively, while all results are insignificant.

Table F1: Product HTE for Pre-sale Service Chatbot

	(1)	(2)	(3)	(4)	(5)	(6)
	Market Concentration		Annual Quantity		Price	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.136**	0.138*	0.0597	0.214***	0.137*	0.137**
	(0.0532)	(0.0821)	(0.0564)	(0.0800)	(0.0782)	(0.0589)
% Change	21.3%	13.2%	8.8%	21.3%	16.6%	15.9%
Panel B: Conversion Rate						
Treat	0.00647***	0.00667***	0.00265	0.0105***	0.00404**	0.00910***
	(0.00166)	(0.00196)	(0.00163)	(0.00199)	(0.00158)	(0.00202)
% Change	26.6%	18.5%	10.3%	30.3%	17.4%	24.5%
Observation	44,614	44,614	44,614	44,614	44,614	44,614

[1] We classify products into high and low groups based on three key pre-experiment dimensions. (1) Category market concentration, measured by the sales share of the top 1% of products (ranked by annual sales) within each category. Products in the low group belong to categories with concentration levels below the platform average. (2) Annual sales quantity, also defined within category. Products in the low group—referred to as tail products—are those comprising the bottom 50% cumulative share of total units sold when ranking products by annual sales quantity within each category. (3) Product price, defined within each category to control for category-level pricing variation. Low-priced products are those priced below the median of their respective category.

[2] "Sales" represents the total expenditure on product orders. "Conversion Rate" measures consumers' likelihood of making purchases. It is a binary indicator for purchase, which equals 1 if a consumer makes at least one order during the experiment period, and 0 otherwise.

[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group average. *** p<0.01, ** p<0.05, * p<0.1.

Table F2: Product HTE for Search Query Refinement

	(1)	(2)	(3)	(4)	(5)	(6)
	Market Concentration		Annual Quantity		Price	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	-0.00862	0.0734***	-0.00381	0.0686***	0.0590**	0.00586
	(0.0174)	(0.0253)	(0.0156)	(0.0258)	(0.0296)	(0.00864)
% Change	-0.80%	6.49%	-0.47%	4.92%	3.79%	0.89%
Panel B: Conversion Rate						
Treat	0.000247	0.000823***	0.0000005	0.00121***	0.000526*	0.000548
	(0.000332)	(0.000302)	(0.000305)	(0.000343)	(0.000301)	(0.000342)
% Change	0.45%	1.85%	0.00%	2.07%	1.19%	0.94%
Observations	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382

[1] Please refer to Table F1 for detailed notes.

Table F3: Product HTE for Product Description

	(1)	(2)	(3)	(4)	(5)	(6)
	Market Concentration		Annual Quantity		Price	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.00932***	0.00113	0.00497*	0.00547*	0.00760***	0.00285
	(0.00299)	(0.00284)	(0.00281)	(0.00299)	(0.00291)	(0.00286)
% Change	3.10%	0.55%	2.10%	2.03%	4.10%	0.89%
Panel B: Conversion Rate						
Treat	0.000384**	0.000141	0.000456***	0.000166	0.000393***	0.000209
	(0.000151)	(0.000120)	(0.000134)	(0.000141)	(0.00010)	(0.000164)
% Change	1.38%	0.81%	2.07%	0.69%	3.32%	0.63%
Observations	4,772,937	4,772,937	4,772,937	4,772,937	4,772,937	4,772,937

[1] Please refer to Table F1 for detailed notes.

Table F4: Product HTE for Marketing Push Message

	(1)	(2)	(3)	(4)	(5)	(6)
	Market Concentration		Annual Quantity		Price	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.000562	-0.00019	-0.0000513	0.000428	0.000092	0.000285
	(0.000529)	(0.00062)	(0.000479)	(0.000660)	(0.000741)	(0.000340)
% Change	4.4%	-1.5%	-0.5%	3.0%	0.6%	2.7%
Panel B: Conversion Rate						
Treat	0.000023	0.00002*	0.00000488	0.0000408***	0.0000478***	-0.000002
	(0.0000169)	(0.000014)	(0.0000155)	(0.0000153)	(0.0000128)	(0.0000177)
% Change	2.4%	3.5%	0.6%	5.2%	8.9%	-0.2%
Observations	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528

[1] Please refer to Table F1 for detailed notes.

Table F5: Product HTE for Google Advertising Titles

	(1)	(2)	(3)	(4)	(5)	(6)
	Market Concentration		Annual Quantity		Price	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	-0.00564	-0.00651	-0.00333	-0.00973	-0.00831	-0.00282
	(0.00709)	(0.00811)	(0.00740)	(0.00759)	(0.00758)	(0.00727)
% Change	-4.3%	-5.2%	-2.3%	-8.9%	-6.2%	-2.3%
Panel B: Conversion Rate						
Treat	-0.000178	-0.0000793	-0.000136	-0.000142	-0.000144	-0.000140
	(0.000174)	(0.000173)	(0.000179)	(0.000163)	(0.000149)	(0.000212)
% Change	-4.1%	-2.1%	-2.9%	-4.4%	-4.2%	-2.8%
Observations	714,642	529,374	714,642	529,374	714,872	529,144

[1] Please refer to Table F1 for detailed notes.