Marketing Science Institute Working Paper Series 2026

Report No. 26-112

Generative AI and Firm Productivity: Field Experiments in Online Retail

Lu Fang, Zhe Yuan, Kaifu Zhang, Dante Donati and Miklos Sarvary

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

Lu Fang† Zhe Yuan‡ Kaifu Zhang§

Dante Donati¶ Miklos Sarvary‖

February 10, 2026

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 sales 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 and improving consumer experience. Importantly, these effects are not associated with worse post-purchase outcomes, as product return rates and customer ratings do not deteriorate. Finally, we document substantial demand-side heterogeneity, with larger gains for less experienced consumers. 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, O33

*The authors thank Andrey Simonov for valuable feedback, and seminar and conference participants at the TSE Digital Economics Conference, AIML Conference, Operational Innovation Network Summit, University of Rochester, UC Davis, TUM Workshop on Generative AI in Marketing, UCL School of Management, Columbia Business School, Business & Generative AI Conference, and Zhejiang University School of Management for helpful comments and discussions. The authors thank the partner e-commerce platform for its support with this project and are grateful to its employees for patiently answering many of our questions. 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 72141305 and 72203202], the Fundamental Research Funds for the Central Universities. Kaifu Zhang, Dante Donati, and Miklos Sarvary have no funding to report. Lu Fang and Zhe Yuan served as consultants for the partner platform during the duration of the project. Kaifu Zhang is an employee of the partner platform. Dante Donati and Miklos Sarvary have no conflicts of interest to report.

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

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

§Partner Company, 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; Dell'Acqua 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 task-level focus overlooks a central insight from the economics literature: productivity gains arise from changes in interdependent routines, or workflows, rather than from isolated tasks (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 a six-month period 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

[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.

reflect genuine productivity gains.[4] Each application was evaluated through randomized field experiments, with sample sizes ranging from tens of thousands to tens of millions of users or products. Leveraging granular consumer- and product-level data, the experiments assess the short-term impact of GenAI on key performance outcomes such as sales (in dollar terms), conversion rates, and cart composition, allowing us to identify not only whether GenAI delivers measurable business value, but also where, how, and for whom these gains materialize.

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 sales 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, productivity gains arise primarily from GenAI's ability to reduce frictions along the customer journey and increase purchase probability. Across workflows, we observe significantly higher conversion rates but no effects on average cart values, product return rates, or customer ratings, consistent with GenAI improving the consumer experience and driving market expansion rather than altering spending intensity or misleading consumers. This mechanism is especially relevant in cross-country, multilingual online retail settings, where search and information frictions are common. Third, the effects of GenAI adoption are heterogeneous across market participants. On the demand side, less experienced consumers benefit disproportionately from GenAI-powered workflows. On the supply side, point estimates suggest larger gains for smaller and newer sellers, though these differences are estimated imprecisely and are often statistically insignificant.

A key distinction of our study relative to the 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 remain unchanged, and GenAI deployment does not alter the factors of production.[5] Our estimates therefore represent a conservative lower bound on the potential returns on investment in GenAI. The observed improvements reflect an 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 produc-

[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.

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

tivity 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, often conducted in laboratory or small-scale settings. Such isolated applications across 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). In addition, productivity dynamics observed 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 that remains 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 purchase directly from manufacturers at competitive prices.[6] Between September 2023 and 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 workflows span 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 corresponds 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 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 example, in the Pre-sale Service Chatbot workflow, the control group received no customer support; in Search Query Refinement, the control condition relied on standard machine-learning-based search algorithms; and in most other workflows, the benchmark was human input.

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

Within each workflow, the GenAI deployment was evaluated through a large-scale randomized field experiment that compared the GenAI-enhanced workflow in the treatment group to a baseline version used in the control group. Notably, baseline workflows often included automation or human input but did not incorporate GenAI technologies. The treatment condition differed solely through the integration of GenAI, while prices and inputs remained constant across conditions. Randomization occurred at the level of consumers—and, in one case, products—with minimal overlap (less than 1%) across experiments, enabling causal identification of the effects of GenAI on consumer behavior. For five of the seven experiments, we obtained detailed consumer- or 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 measure of consumer experience in online retail, as well as cart value, which captures changes in the behavior or preferences of consumers who make a purchase.

Our results reveal that most GenAI deployments yield economically significant short-term productivity gains, though the 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 Service Chatbot, 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, typically in the range of 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 correspond to roughly 5.5–6% of the per-user revenue growth observed in global e-commerce between 2023 and 2024. We also document notable improvements in workflows without granular data: a 15% increase in success rate for 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, mature retailer, with further potential as adoption broadens and targets revenue-critical processes. For example, 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. Meanwhile, API calls to proprietary GenAI tools 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, specifically 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 primarily from enhanced consumer experience through GenAI-driven reductions in market frictions along the customer journey. Specifically, increases in sales are strongly associated with higher conversion rates—and, where applicable, improvements in intermediate engagement metrics such as click-through rates—but not with changes in spending intensity among buyers. Across workflows, conversion rates increase by 1–22%, while average cart values and purchase frequency, conditional on consumers making a purchase, remain unchanged. This pattern indicates that GenAI contributes to productivity mainly through market expansion, rather than through changes in spending behavior or preferences among existing buyers.

In addition, we find no evidence that these gains come at the expense of consumer welfare. Product return rates and customer ratings do not deteriorate following GenAI adoption, and in some cases, returns decline or ratings improve. Together, these findings suggest that GenAI-driven friction reduction improves decision quality and consumer experience rather than inducing deceptive or impulsive purchases. In particular, GenAI-powered pre-sale chatbots and richer product descriptions reduce search costs and information asymmetries; GenAI-refined queries lower search frictions and improve matching; and automated push messages enhance content personalization. While online platforms already mitigate such frictions (Belleflamme and Peitz, 2021), our results show that GenAI adoption can further reduce them along multiple stages of the customer journey.

Finally, we analyze heterogeneity in treatment effects across buyers, sellers, and products. If GenAI primarily reduces frictions on both the demand and supply sides, one would expect larger gains among participants with lower baseline capabilities, such as less experienced buyers with limited platform engagement, smaller and newer sellers, and products in the long tail or in less concentrated categories. Consistent with this view, we find robust and statistically significant heterogeneity on the demand side: less experienced consumers benefit disproportionately more from GenAI-powered workflows. On the supply side, point estimates suggest larger gains for smaller and newer sellers, though these differences are estimated imprecisely and are often statistically insignificant. Overall, these patterns suggest that GenAI enhances outcomes for participants with lower baseline capabilities. By contrast, 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 and empirical frameworks, and data. Section 4 presents the main results, and Section 5 examines heterogeneity in 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

[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 the number of clicks or click-through rates, but they provide no evidence on actual purchase behavior.

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 GenAI 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 asymmetric information: buyers often cannot directly verify product quality or seller reliability prior to purchase (Jin and Kato, 2006; Tadelis, 2016). 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 and improving match value. 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 engagement 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.

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

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

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.

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).

[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.

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, for each consumer or product, the treatment status remained fixed throughout the experimental period. The key distinction between treatment and control was that the treatment group was exposed to workflows re-engineered with GenAI, whereas the control group continued under the platform's standard practices without GenAI integration.

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 44 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 of prioritizing human agents for post-sale rather than pre-sale support for self-sold products, given that pre-sale inquiries are generally less urgent. This setup also mirrors the constraints faced by many third-party sellers on the platform, particularly smaller-scale sellers, who lack the capacity to provide real-time multilingual customer service. By contrast, consumers in the treatment group were supported purely by GenAI-powered chatbots. GenAI is expected to reduce search costs and 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 because they are less commonly used on the platform and have historically been underserved by the

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	AI-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	Sellers & Platform

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

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 comprised five sub-experiments, each involving consumers who spoke English, Spanish, French, Portuguese, or Korean, which are among the most widely used languages on the platform. All sub-experiments ran for one week in December 2023, with staggered start dates. GenAI was employed to create multilingual, textual product descriptions for a predetermined product set of approximately 45,000 randomly selected platform self-sold products spanning a broad range of categories.[10] According to our partner company, self-sold products are primarily sourced from Chinese vendors, who typically provide image-based product presentations 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 to the experiment and evenly split into 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.[11] The treatment group is expected to face lower information asymmetries, as AI-generated descriptions are more complete, standardized, and structured, and are accessible in multiple languages.

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

[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] 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.

[11] 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 AI-generated descriptions only.

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 AI-generated messages, yielding nearly 2.7 million unique messages and thus far greater differentiation across individuals. The hypothesis in this experiment was that the very large number of AI-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 occurring 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 interest.[12] 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.[13] A key distinction of this experiment is that the GenAI model was not fine-tuned specifically for the advertising context within the e-commerce domain. As a result, the generated titles may fail to emphasize product attributes most relevant to consumer search and purchase decisions, leading us to adopt a more agnostic stance regarding the expected 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 were 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. We expect the treatment group to experience

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

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

higher resolutions, since GenAI can generate more tailored and context-specific responses than template-based staff.

Live Chat Translation The experiment was conducted over one month in October 2023 and involved approximately 0.2 million non-English-speaking consumers who contacted the platform's customer service for issues such as clarifying details of platform-level promotions or resolving disputes with sellers when no agreement was reached. Due to cost constraints, a large share of such requests is handled by customer service agents from the Philippines who provide assistance in English, as employing native agents for each market is roughly three times more expensive. During the experiment, non-English-speaking consumers who initiated inquiries to the platform's customer service were randomly split among treatment and control conditions. In the treatment group, consumers interacted with Filipino agents with real-time bidirectional GenAI translation support, while those in the control group engaged with Filipino agents without GenAI 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, consistent with the assumptions outlined in Section 3.2, labor inputs were held constant across treatment and control conditions in all experiments, and GenAI adoption involved little to no labor displacement, with the exception of one workflow. In the Pre-sale Service Chatbot and Search Query Refinement workflows, GenAI replaced tasks that were already automated, such as pre-programmed notifications or standard search algorithms, and therefore required no labor input. In other workflows, GenAI augmented rather than replaced existing labor. For example, AI-generated content complemented human inputs in Product Description, Marketing Push Message, and Google Advertising Title, while Live Chat Translation provided real-time support without altering agents' roles. The sole exception is Chargeback Defense, where GenAI replaced workers previously responsible for drafting dispute responses. However, the extent of displacement was negligible, affecting only three to five outsourced workers. Moreover, although energy and computing costs may have differed between treatment and control conditions, these costs were minimal relative to other operating expenses. Finally, GenAI adoption did not alter product prices.

3.4 Data and Estimation

We obtained comprehensive granular consumer- or product-level data for the first five of the seven experiments, allowing for in-depth analysis 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 experiments conducted at the consumer level, we record each consumer's treatment status and observe their complete set of activities, including the number of product views (Views), product clicks (Clicks), product orders (Orders), and total expenditure on those orders (Sales).[14] For comparability across workflows, our main analysis focuses on sales, measured as total consumer expenditure, which serves as our primary revenue-based measure of productivity. To shed light on the mechanisms underlying the sales effects, we additionally examine outcomes along the extensive and intensive margins. In particular, we analyze conversion rates, defined as a binary indicator for whether a consumer makes at least one purchase during the experimental period, and cart value, measured as average expenditure conditional on making a purchase. Conversion rates capture changes in purchase incidence and serve as a widely used proxy for consumer experience in online retail, while cart value reflects adjustments along the intensive margin. For the product-level experiment, we collect the same outcomes at the product level. Table 3 reports summary statistics for these key variables across the five experiments for which granular data are available.

To compare mean sales, conversion rates, and cart values between the treatment and control groups, we use the following general empirical specification, adapted 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.

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 prob-

[14]In Search Query Refinement, 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 Google Advertising Title, product views refer to the number of views of advertised products within the Google Shopping tab. In Search Query Refinement and Product Descriptions, product clicks capture the number of times consumers clicked into product detail pages. In Marketing Push Message, product clicks reflect consumer clicks on push notifications, while in Google Advertising Title, they indicate clicks on advertised products.

ability 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.

Table 3: Summary Statistics of Main Outcomes

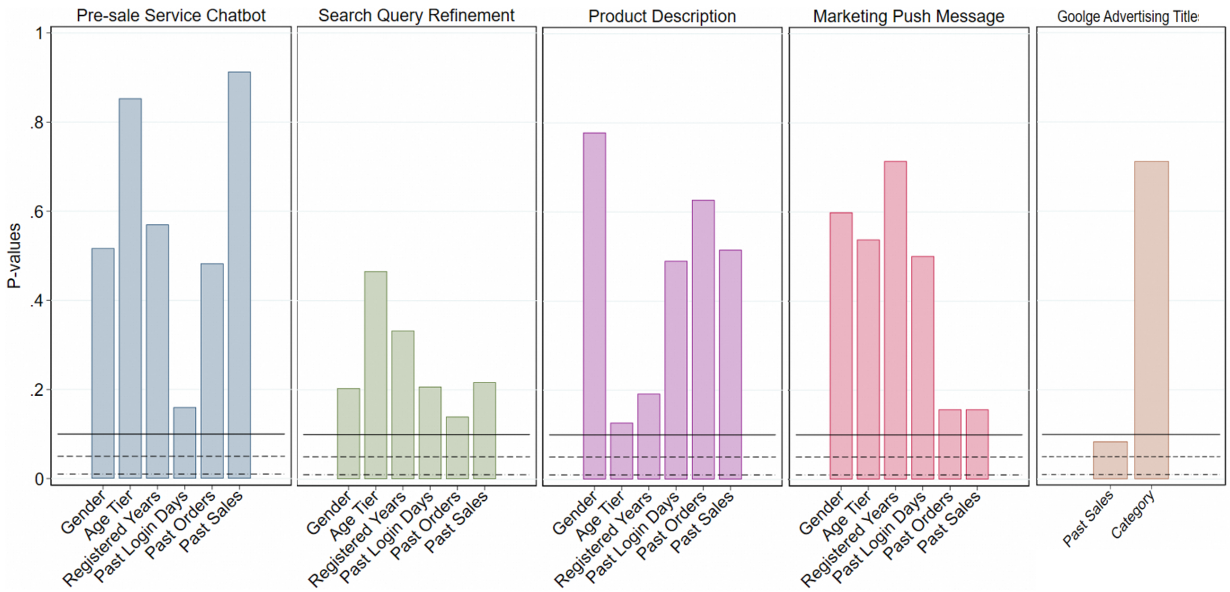
	Mean	Standard Dev.	Median	Min	Max
Pre-sale Service Chatbot					
Sales	1.86	9.75	0	0	517
Conversion Rate	0.07	0.25	0	0	1
Cart Value	26.71	26.21	18.61	1.11	517
Search Query Refinement					
Views	313.36	615.02	125	1	105,883
Clicks	8.23	16.99	3	0	2,024
Orders	0.16	0.73	0	0	85
Sales	2.24	21.41	0	0	4,960
Conversion Rate	0.09	0.28	0	0	1
Cart Value	25.39	67.83	10.42	0.01	4,960
Product Description					
Clicks	1.98	2.06	1	1	173
Orders	0.06	0.30	0	0	23
Sales	0.51	4.56	0	0	2,942
Conversion Rate	0.04	0.21	0	0	1
Cart Value	11.62	18.51	7.06	0.59	2,942
Marketing Push Message					
Clicks	0.017	0.13	0	0	1
Orders	0.0018	0.05	0	0	6
Sales	0.025	1.51	0	0	501
Conversion Rate	0.0016	0.04	0	0	1
Cart Value	15.52	33.97	7.81	0.01	501
Google Advertising Title					
Views	19.36	82.38	5	2	12,033
Clicks	0.22	1.69	0	0	627
Sales	0.13	2.97	0	0	322
Conversion Rate	0.004	0.07	0	0	6
Cart Value	41.35	72.16	22.06	0.54	322

[1] "Views" refers to the number of product views. "Clicks" denotes the number of product clicks. "Orders" is the number of product orders. "Sales" represents the total expenditure on product orders. "Conversion rate" measures consumers' likelihood of making a purchase. 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" is the expenditure per consumer, conditional on the consumer making a purchase.

[1] In the Google Advertising Title experiment, the unit of observation is the product. The conversion rate is calculated as the number of purchases divided by the number of views for a given product, while cart value refers to sales per product, conditional on the product being purchased.

To enrich our analysis, we obtained pre-experiment seller data for products included in the experiments (e.g. seller size measured by annual sales, 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). This analysis is not applicable to the Pre-sale Service

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.

Chatbot and Product Description experiments, as both involve platform self-sold products sold 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 level of the 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.[15] 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 and products into the two groups. Details on the covariate balance checks are presented in Appendix B.

[15] 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.

4 Main Results

4.1 Productivity Impact by Workflow

We begin by examining the impact of GenAI on productivity across business workflows. Table 4 reports the estimated effects on sales, our primary productivity measure. Column (1) presents the average treatment effect (ATE), capturing the absolute change in sales, while column (2) reports the corresponding percentage change relative to the control group. The results presented here do not account for differences in the frequency of consumer interactions with the workflows or in the duration of the corresponding experiments. For example, search affects nearly all consumers on a daily basis, whereas chatbot interactions occur less frequently. In Section 4.3, we provide adjusted measures that account for these differences. Additional results and robustness analyses are reported in Appendix C.

Table 4: Average Treatment Effects of GenAI Adoption on Sales Across Workflows

	Productivity Impact (Sales, $)			
	(1)	(2)	(3)	(4)
Business Workflow	Coefficient	% Change	Obs	Unit of Obs
Pre-sale Service Chatbot	0.274*** (0.0995)	16.3%	44,614	Consumer
Search Query Refinement	0.0648** (0.0314)	2.9%	1,849,382	Consumer
Product Description	0.0104** (0.00417)	2.1%	4,772,937	Consumer
Marketing Push Message	0.000402 (0.000812)	1.6%	13,715,528	Consumer
Google Advertising Title	-0.00602 (0.00534)	-4.5%	1,244,016	Product
Chargeback Defense†	15% defense success rate increase			
Live Chat Translation†	5.2% consumer satisfaction increase			

[1] "Sales" represents the total expenditure on product orders, in USD.
[2] Columns (1) reports the estimated coefficients, with standard errors in parentheses. Columns (2) reports % Change, calculated as the treatment effect divided by the control group mean. Columns (3) reports the number of observations. *** 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 table shows that most GenAI deployments are associated with productivity gains, with substantial heterogeneity across workflows. The largest effect is observed in the pre-sale service chatbot workflow, where GenAI increases sales by 16.3% relative to the control group ($p < 0.01$; Column 2). In this setting, consumers in the treatment group interact with a GenAI-powered chatbot, whereas consumers in the control group receive an automated message indicating that real-time support is unavailable. This control condition reflects the platform's standard operating practice: human agents are prioritized for post-sale inquiries, leaving a substantial share of pre-sale inquiries for self-sold products without live assistance due to capacity constraints. This setting also mirrors the operating reality faced by many third-party sellers on the platform, particularly smaller-scale sellers who lack the capacity to provide real-time, multilingual customer support. Taken together, these features suggest that GenAI enhances productivity by performing tasks that are difficult for human agents to fulfill due to talent shortages or resource constraints.

One potential concern is that consumers in the control group may have been frustrated by the absence of assistance, mechanically reducing their likelihood of purchase and thus inflating the estimated treatment effects. To address this concern, we conducted additional experiments on the GenAI-powered chatbot. Appendix Table C1 reports a series of comparisons, including benchmarks against human agents and hybrid configurations that combine GenAI assistance with human escalation when needed. The results indicate that the GenAI chatbot delivers service quality comparable to human customer support (Column 2). Moreover, integrating GenAI with human agents yields substantially larger effects: relative to the no pre-sale service condition, sales increase by 25% when GenAI assistance is combined with human escalation ($p < 0.01$; Columns 3), pointing to strong complementarities between GenAI and human labor. Most 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 ($p < 0.1$; Columns 4). We interpret this comparison as a conservative lower bound on GenAI's productivity impact in pre-sale customer support.

The remaining four workflows with detailed data reported in Table 4 exhibit more modest effects on sales, ranging from a negative and statistically insignificant estimate to gains of up to 3%. In particular, the Search Query Refinement application increases sales by 2.9% ($p < 0.05$), while automated Product Description generation raises sales by 2.1% ($p < 0.05$). Although smaller in magnitude, these effects are economically meaningful for a platform of this scale and maturity, especially given the near-universal exposure of consumers to search and product descriptions.

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 0.16% of consumers make a purchase) and high variance in expenditures among converters, suggesting that broader implementation could provide sufficient power to detect treatment effects. Notably, however, this case exhibits significant increases in conversion rate and number of orders, which we discuss in detail later. It is also important to note that, for risk-mitigation purposes, only a subset of the treatment group was exposed to AI-generated messages in this experiment, which may attenuate the observed treatment effects under partial exposure.

By contrast, the Google Advertising Title workflow exhibits a statistically insignificant negative effect. This pattern is consistent with the GenAI model not being fine-tuned to the advertising context within e-commerce, leading it to omit commercially salient keywords commonly used in ad titles. As a result, these ads may receive lower quality scores by Google, reducing their probability of winning auctions at a given bid. In addition, Google's advertising algorithm may identify and deprioritize AI-generated titles, further lowering their visibility. Consistent with this interpretation, Table C8 in the Appendix shows that ads with AI-generated titles receive fewer views and clicks than those with human-generated titles. Taken together, these results underscore the importance of domain-specific fine-tuning or retraining of foundation models when deploying GenAI in industry-specific tasks that require specialized knowledge (Deloitte, 2023).

For the final two business processes studied—Chargeback Defense and Live Chat Translation—we do not observe granular transaction-level data and instead rely on the platform's internal metrics and analyses, which do not report statistical significance. While these outcomes are not directly comparable to the sales effects discussed above, they nevertheless point to substantial improvements in two operational dimensions. Specifically, Chargeback Defense is associated with a 15% increase in defense success rates, and Live Chat Translation increases consumer satisfaction by 5.2%. Both effects point directionally to successful implementation of the GenAI applications.[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 input-side task efficiency. Among the seven deployments we examine, most 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 in effects across workflows is unlikely to reflect differences in implementation quality, which was comparable across applications. Instead, it appears to arise from differences in the marginal contribution of GenAI relative to baseline conditions. In each case, the control group reflects the firm's standard operating practices prior to GenAI adoption. For example, in the Pre-Sale Service Chatbot workflow, the control group received no customer support; in Search Query Refinement, it relied on standard machine-learning-based search algorithms with lexical translation into English; 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, generate the largest improvements; search and product-discovery tasks yield more modest gains; and advertising-related applications exhibit no statistically significant effects.[17] Taken together, these findings highlight both the role of baseline

[16]In the case of Chargeback Defense, additional gains not captured in our calculations may arise from cost reductions, as the GenAI-enabled workflow eliminates the need for manual intervention.

[17]In the Marketing Push Message workflow, features of the experimental design may also affect measured impacts,

conditions in shaping treatment effects and the differential effectiveness of GenAI across functional areas.

4.2 Mechanisms

Much of the existing literature attributes the productivity-enhancing potential of GenAI to input-side mechanisms, such as labor savings or efficiency improvements that reduce the time required to complete tasks. In this section, we examine an additional channel operating on the demand side: GenAI can increase productivity by improving the consumer experience during the shopping process.

We investigate two primary mechanisms behind the observed impact on sales. The first is market expansion, captured by an increase in the probability that a consumer completes a purchase at the extensive margin. We measure this mechanism using conversion rates, a widely used industry metric that reflects consumers' revealed satisfaction and perceived friction during the ex ante search and evaluation process in e-commerce. The second mechanism is a change in the behavior or preferences of consumers who make a purchase, reflected in adjustments to spending intensity and order composition. In this case, GenAI may influence consumers' willingness to pay, alter purchasing patterns, or shift demand toward products at different price points. We assess this channel using cart value and purchase frequency, conditional on consumers making a purchase.

Table 5: Average Treatment Effects of GenAI Adoption on Conversion Rate and Cart Value

	Conversion Rate			Cart Value ($)		
	(1)	(2)	(3)	(4)	(5)	(6)
Business Workflow	Coefficient	% Change	Obs	Coefficient	% Change	Obs
Pre-sale Service Chatbot	0.0131***	21.7%	44,614	-1.264	-4.5%	3,076
	(0.00256)			(1.036)		
Search Query Refinement	0.00101**	1.1%	1,849,382	0.370	1.5%	163,381
	(0.000411)			(0.334)		
Product Description	0.000554***	1.3%	4,772,937	0.0944	0.8%	210,155
	(0.000187)			(0.0805)		
Marketing Push Message	0.000048**	3.0%	13,715,528	-0.206	-1.3%	22,425
	(0.0000218)			(0.454)		
Google Advertising Title	-0.000137	-3.3%	1,244,016	-0.784	-2.3%	4,811
	(0.000124)			(0.992)		

[1] "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. 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 sales per product, conditional on the product being purchased.

[2] Columns (1) and (4) report the estimated coefficients, with standard errors in parentheses. Columns (2) and (5) report % Change, calculated as the treatment effect divided by the control group mean. Columns (3) and (6) report the number of observations. *** p<0.01, ** p<0.05, * p<0.1.

as only a subset of treated users was exposed to GenAI, likely attenuating estimated effects.

Table 5 reports the results on conversion rates and cart value, with additional evidence on purchase frequency provided in Appendix C. Across workflows, we document economically and statistically significant increases in conversion rates ranging from 1% to 22% (Column 2), which translate directly into higher output as measured by sales. By contrast, we find no evidence of effects along the intensive margin. Columns 4 and 5 show that cart value, as measured by the average order value among consumers who make at least one purchase (or among products purchased at least once, in the case of Google Advertising Title), remains unchanged following GenAI adoption. We likewise find no significant changes in purchase frequency, defined as the number of orders placed conditional on making at least one purchase (Appendix C). Taken together, these results indicate that GenAI primarily drives productivity gains through market expansion rather than by altering spending patterns among existing buyers.

The evidence on increased conversion rates in Table 5 suggests that GenAI has the potential to reduce market frictions, both by enabling new services and by improving existing ones, thereby enhancing the shopping experience and increasing the probability of purchase. For example, GenAI can mitigate search costs and information asymmetries by providing relevant and timely assistance through a pre-sale chatbot (a 21.7% increase in conversion rates, $p < 0.01$) and by generating more comprehensive and structured product descriptions (a 1.3% increase, $p < 0.01$). It can also reduce search frictions and improve match quality by enhancing the translation and semantic understanding of consumer queries (a 1.1% increase, $p < 0.05$). In addition, GenAI enables personalization of marketing content at scale by generating customized messages across a broad product portfolio (a 3% increase, $p < 0.05$). Finally, evidence from Chargeback Defense and Live Chat Translation reported in Table 4 points to improvements consistent with enhanced experiences for sellers and consumers, respectively, including higher dispute success rates and greater customer satisfaction.

In Appendix C, we report additional evidence on how GenAI reduces frictions along the customer journey, where such data are available. For example, Table C4 shows that GenAI-refined search queries increase the likelihood of a product click by 0.3% ($p < 0.01$; Column 1), consistent with improved search performance in combination with the previously documented increase in purchase conversion. Columns 2 and 3 further indicate that treated consumers view fewer products prior to clicking or purchasing, suggesting a reduction in search intensity. These patterns align with prior literature showing that improvements in search quality increase conversion while reducing search effort (Yang, 2013; Zhou et al., 2025). In addition, we observe a statistically significant 2.0% increase in the click-through rate ($p < 0.01$; Column 4), defined as the ratio of product clicks to product views, while the click-to-order conversion rate (Column 5), defined as the ratio of orders to clicks, remains statistically insignificant. Together, these findings suggest that GenAI-refined queries primarily affect the composition of products retrieved at the search stage, rather than the information displayed on product detail pages, reinforcing the proposed query-refinement mechanism.

Table C5 shows that AI-generated product descriptions increase the number of orders placed by 1.1% ($p < 0.05$; Column 2). Moreover, when we stratify products by the length of their original, human-generated descriptions, we find substantial heterogeneity in effects. Products with no or insufficient textual descriptions (fewer than 50 words) experience a 6.1% increase in sales following augmentation with AI-generated content ($p < 0.05$), whereas products with sufficiently detailed descriptions (more than 50 words) exhibit no significant increase (Columns 1 and 2 of Table C6).[18] This pattern indicates that the descriptions augmented by GenAI are most effective for products with limited baseline information.

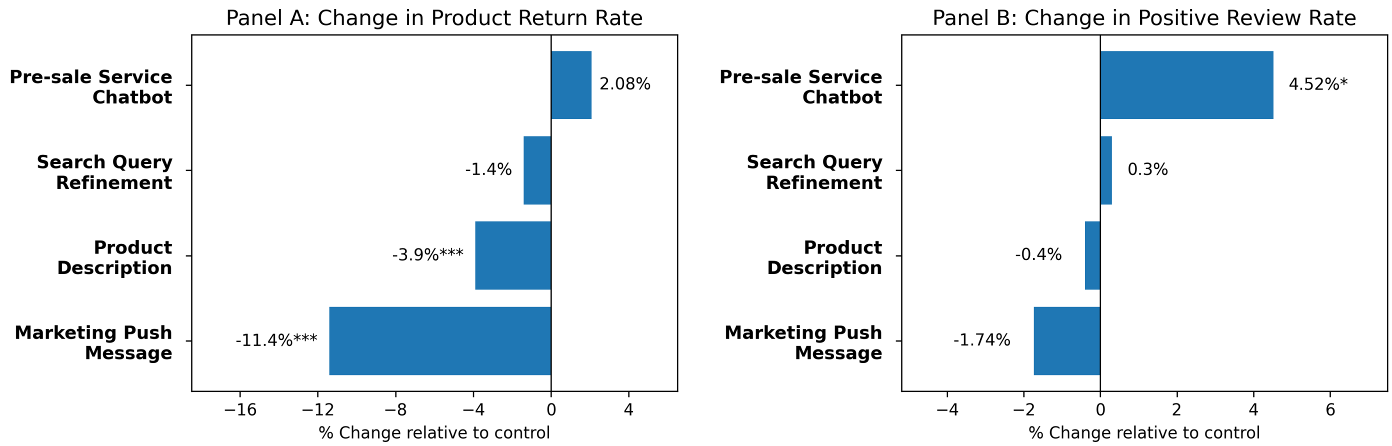
In the marketing push workflow, GenAI enables large-scale variation in customer-facing content. The treatment group includes approximately 2.7 million unique message variants, compared with only roughly 2,000 variants in the control group, which relies exclusively on human-generated content. Consistent with the expanded scope for personalization, AI-generated marketing messages increase clicks by 3.1% ($p < 0.05$) and orders by 2.8% ($p < 0.10$; Columns 2 and 3 of Table C7). Taken together, these results provide consistent evidence that GenAI enhances intermediate engagement outcomes, such as clicks, orders, and click-through rates, reflecting a smoother and more informative shopping experience that underpins the higher conversion and sales gains documented above.

A natural concern with deploying GenAI in customer-facing processes is that it may introduce hallucinations or distortions that mislead consumers—for example, by inducing impulse purchases or manipulation—thereby reducing ex post consumer satisfaction. In our case, increases in revealed purchase behavior may not map cleanly into welfare gains, as higher sales could reflect distorted choices rather than improved consumer surplus. To assess this possibility, we examine two post-purchase outcomes that capture consumers' evaluations of their purchase decisions. The first is the product return rate, defined as the share of orders that are returned. The second is the positive review rate, defined as the share of rated orders that receive four- or five-star ratings on a five-point scale.

As shown in Figure 2 (full regression results are provided in Table C9), we find no evidence that the observed sales gains come at the expense of ex post consumer satisfaction. Product return rates and customer ratings do not deteriorate following GenAI adoption. Notably, return rates decline significantly for Product Descriptions and Marketing Push Messages (by 3.9% and 11.4%, respectively; $p < 0.01$), consistent with improvements in both ex ante decision quality and ex post consumer satisfaction in these workflows. We also observe a positive effect on customer ratings for the Pre-sale Service Chatbot (a 4.5% increase; $p < 0.10$). In all other cases, the estimated effects are statistically insignificant, providing no evidence that GenAI induces deceptive practices that

[18]Based on internal research and expert surveys conducted by our partner platform, descriptions containing fewer than 50 words are classified as providing insufficient or minimal textual information.

Figure 2: Average Treatment Effects of GenAI Adoption on Product Returns and Reviews



[1] "Return Rate" is defined as the share of orders that are returned. "Positive Review Rate" is defined as the share of rated orders that receive four- or five-star ratings on a five-point scale. The estimation of Return Rate is conditional on consumers with orders, while the estimation of Positive Review Rate is conditional on consumers provide ratings. Such data are not available for the Google Advertising Title workflow.

[2] Full regression results are reported in Table C9. *** p<0.01, ** p<0.05, * p<0.1.

mislead consumer decisions.

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, explicitly accounting for differences in the frequency of consumer interactions across workflows. 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 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.[19] For instance, the Pre-sale Service Chatbot experiment spans two months, yielding a time multiplier of six (i.e., 12/2). For each workflow, 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 AI-generated content

[19] We assume that each experiment's duration reflects the typical interval between treatment opportunities for a representative consumer on the platform. For example, a representative consumer is assumed to interact with the Pre-sale Service Chatbot about once every two months, whereas they could receive a Marketing Push Message daily.

Table 6: Aggregate Productivity Gains Across Workflows

Business Workflow	(1) Incremental Value Per Consumer ($)	(2) Time Multiplier	(3) Annualized Incremental Value Per Consumer ($)
Pre-sale Service Chatbot	0.274 (upper-bound)	6.0	1.64
	0.218 (lower-bound)	6.0	1.31
Search Query Refinement	0.0648	40.6	2.63
Product Description	0.0104	52.1	0.54
Marketing Push Message	0.000402	365.0	0.15
Total (linear additivity)			**4.96** (upper-bound)
			4.63 (lower-bound)

[1] 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 for each workflow.

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 across touchpoints or expansion through synergies. Based on the four GenAI applications with positive sales effects, we estimate an annual incremental value of up to $4.96 per consumer, which decreases to $4.63 when using the lower-bound estimate from the Pre-sale Service Chatbot experiment (human agents vs. GenAI assistance equipped with human escalation). These effects represent roughly 5.5-6% of the per-user revenue growth 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 the platform's GenAI tools: in mid-2024, AI-related API requests averaged over 50 million per day, rising to more than 1 billion per day by mid-2025, representing 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

A central question in online retail is which platform participants benefit most from GenAI-driven improvements. To address this, we examine heterogeneity in treatment effects across buyers, sellers, 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, less experienced buyers with limited platform engagement, smaller and newer sellers, and products in the long tail of the sales distribution or in less concentrated categories.

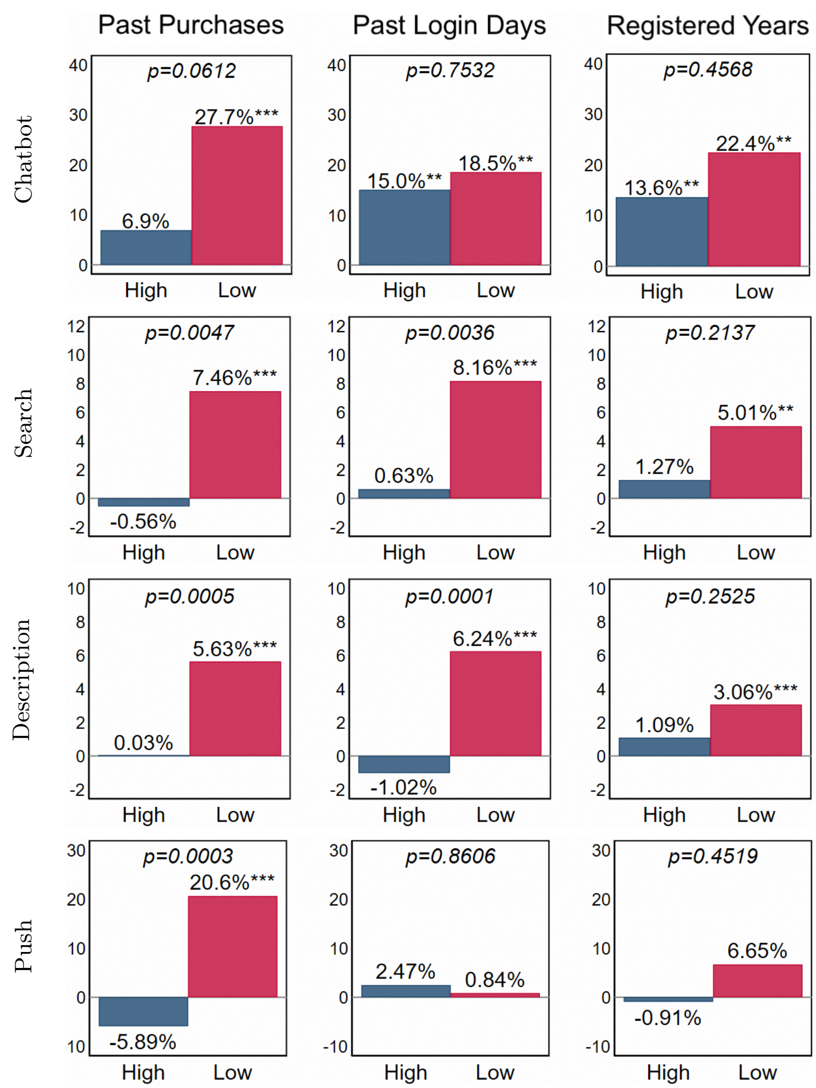
For each dimension, we classify observations into "high" and "low" groups based on pre-experiment characteristics that capture experience, scale, or market position. For buyers, we use measures of prior online shopping experience and purchasing intensity. For sellers, we classify firms based on size and tenure on the platform. 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 less sophisticated buyers, smaller or newer sellers, and products in the long tail, thereby shedding light on how GenAI reshapes outcomes across segments of the marketplace.

We find consistent evidence of heterogeneity on the demand side: buyers with lower prior spending and less platform experience benefit more from GenAI-powered workflows. On the supply side, the estimated effects point to larger gains for smaller and newer sellers, although these estimates are less precise. Finally, heterogeneity across product groups is more context-dependent. We note that, due to data limitations and workflow-specific features, some heterogeneity analyses cannot be implemented for all applications.

5.1 Heterogeneous Effects Across Consumers

We classify consumers into high- and low-experience groups based on three pre-experiment indicators of online shopping experience: total expenditure over the period preceding the experiment (Past Purchases), number of login days in the 30 days prior to the experiment (Past Login Days), and years since registration on the platform (Registered Years). For each indicator, consumers in the low group are defined as relatively inexperienced if they fall below the median of the corresponding distribution. We estimate treatment effects separately for high- and low-experience consumers and test for differences in percentage treatment effects across groups using Wald tests implemented via seemingly unrelated estimation.

Figure 3: Heterogeneous Treatment Effects on Sales Across Consumers



Notes: The bars indicate the % change in treatment effects for the high (blue) and low (red) groups. The p-values are from seemingly unrelated estimations testing the equality of % treatment effects across the two groups.

Results on sales are summarized in Figure 3, with regression results and additional evidence on conversion rates and cart values reported in Appendix D.[20] Overall, we find statistically significant heterogeneity in GenAI's effects across consumer segments. We begin by analyzing heterogeneity by past purchases, our preferred proxy for consumers' prior experience on the platform. Across workflows, GenAI-driven gains are significant and systematically larger for consumers with fewer past purchases, while the effects for consumers with more purchase experience are smaller and insignificant. Further analysis confirms that the differences between these two consumer groups are statistically significant. We find qualitatively similar patterns when stratifying consumers by past login activity and platform tenure, although the differences along these dimensions are generally

[20]The Google Advertising Title experiment is not suitable for consumer-level heterogeneity analysis, as it was conducted off-platform at the product level, preventing access to detailed consumer characteristics.

less pronounced.

Our results indicate that the technology disproportionately benefits less experienced and less sophisticated users, aligning with prior evidence that GenAI can disproportionately benefit vulnerable groups, such as low-skilled workers (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Peng et al., 2023), and extend this insight to the consumer domain. The results are also consistent with earlier work demonstrating that improvements in e-commerce technologies, such as enhanced information provision, search refinement, and personalization, generate greater benefits for consumers who are newer to the platform or have lower purchasing power (Sun et al., 2024; Fang et al., 2024).

A plausible explanation for the observed buyer heterogeneity is that less experienced consumers face greater informational and search frictions throughout the purchase journey. With limited domain knowledge and fewer alternative information sources, they rely more heavily on customer service and product descriptions to acquire relevant information. They also exhibit weaker search skills and less familiarity with online search environments, making GenAI-enhanced search particularly valuable for articulating their needs and identifying relevant products. In addition, GenAI-driven message personalization increases the perceived relevance of marketing content, which is especially effective in supporting the purchase decisions of less experienced consumers.

5.2 Heterogeneous Effects Across Sellers

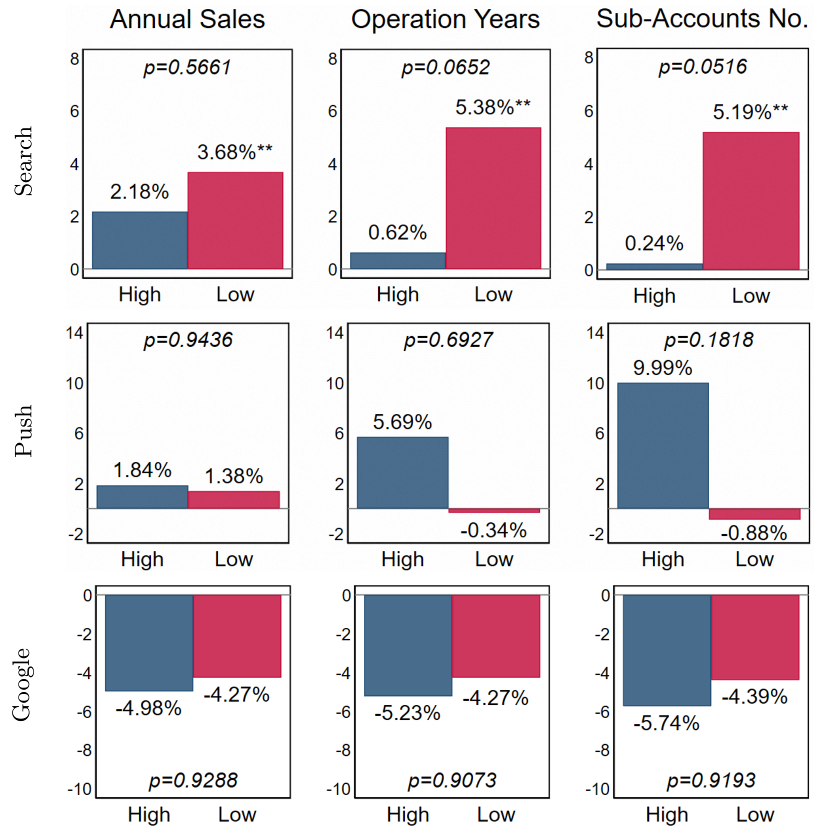
We examine heterogeneity in GenAI's effects across sellers by focusing on differences in seller size and sophistication. Sellers are classified into high and low groups based on three pre-experiment characteristics: annual sales volume (Annual Sales), years of operation on the platform (Operation Years), and the number of sub-accounts associated with the seller's store (# of Sub-Accounts). Sellers in the low group correspond to smaller and less established firms, defined by lower sales, shorter platform tenure, or fewer sub-accounts.[21] We then decompose consumer purchases into transactions involving high- and low-group sellers and estimate treatment effects separately for each group. Differences in percentage treatment effects are formally tested using Wald tests implemented via seemingly unrelated estimation.

Summary results are reported in Figure 4, with regression results and additional details on conversion rates and cart values provided in Appendix E. Overall, we find limited statistical evidence of differential effects across seller types. While point estimates often suggest larger gains for smaller and less established sellers, most differences between high- and low-group sellers are statistically insignificant, reflecting imprecise estimates and limited power. This pattern holds across

[21] Sellers are classified into the low group if they (i) account for the bottom 50% of cumulative annual sales, (ii) have operated on the platform for fewer than five years (the platform treats five years as a key performance benchmark), or (iii) maintain fewer than three sub-accounts (following platform guidance, small stores are typically individually operated or mom-and-pop businesses, whereas stores with more sub-accounts generally employ additional staff). Platform self-operated sellers are classified as large.

the workflows for which seller-level heterogeneity can be assessed.[22]

Figure 4: Heterogeneous Treatment Effects on Sales Across Sellers



Notes: The bars indicate the % change in treatment effects for the high (blue) and low (red) groups. The p-values are from seemingly unrelated estimations testing the equality of % treatment effects across the two groups.

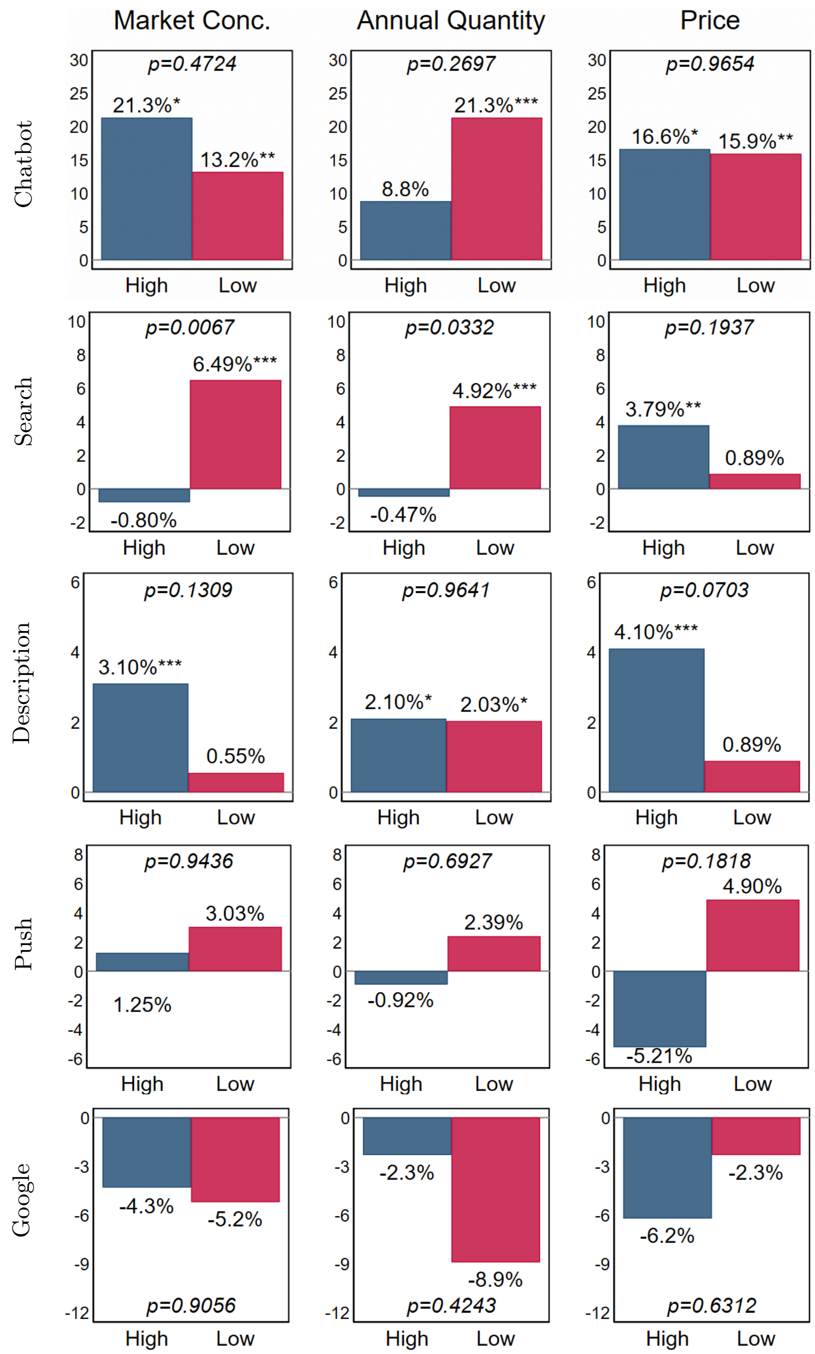
5.3 Heterogeneous Effects Across Products

Finally, we study heterogeneity in GenAI's effects across products by classifying them into high and low groups along three pre-experiment dimensions. First, category market concentration, measured by the sales share of the top 1% of products within each category, captures the degree of product standardization and demand homogeneity. Second, annual sales quantity within category distinguishes head from tail products. Third, product price within category proxies for purchase stakes and potential information asymmetries. For each dimension, products in the low group correspond to less concentrated categories, lower-sales (tail) products, or lower-priced items.[23] We decompose consumer purchases into transactions involving high- and low-group products and estimate

[22]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.

[23]Products are classified into the low group if they (i) belong to categories with concentration below the platform average, (ii) account for the bottom 50% of cumulative sales quantity within their category, or (iii) are priced below the category median.

Figure 5: Heterogeneous Treatment Effects on Sales Across Products



Notes: The bars indicate the % change in treatment effects for the high (blue) and low (red) groups. The p-values are from seemingly unrelated estimations testing the equality of % treatment effects across the two groups.

treatment effects separately by group, testing for differences in percentage treatment effects using seemingly unrelated estimation.

Summary results are reported in Figure 5, with regression results and additional details provided in Appendix F. Overall, we find no uniform or sharply defined pattern of product-level heterogeneity across workflows. While several point estimates suggest larger GenAI-driven gains for products in less concentrated categories, tail products or high-priced items—depending on the specific application—most differences between high- and low-group products are statistically insignificant. The direction of the effects is often consistent with GenAI reducing search and information frictions in settings with greater product differentiation, limited sales history or higher decision stakes, but these patterns are not robust across all workflows. Overall, the evidence points to limited and context-dependent product-level heterogeneity in the effects of GenAI.

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 experimental 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 while holding inputs constant, as reflected in increased sales across several business workflows. Although the magnitude of these gains varies widely, ranging from negligible effects to double-digit increases, the evidence shows that GenAI can generate economically meaningful value when deployed in consumer-facing processes. Second, these improvements arise primarily from enhanced consumer experience through GenAI-driven reductions in marketplace frictions, rather than from cost savings on the input side. Across workflows, we observe higher conversion rates but no changes in spending intensity or adverse post-purchase outcomes. By enriching pre-sale communication, refining search queries, generating richer product descriptions, and personalizing marketing messages, GenAI improves matching efficiency and mitigates information asymmetries along the customer journey. Third, the benefits of GenAI adoption are heterogeneous across market participants. On the demand side, less experienced consumers derive disproportionately larger gains. On the supply side, point estimates suggest larger benefits for smaller and newer sellers, though these effects are less precisely estimated. Heterogeneity across products is more context-dependent. Taken together, these patterns highlight GenAI's potential to improve outcomes for participants with lower baseline capabilities.

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.[24] Back-of-the-envelope calculations—annualizing

[24] One exception is the Product Description experiment, where observed consumer outcomes pertain only to the

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 correspond to 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 its proprietary GenAI tools increased between 2024 and 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. Therefore, our analysis reflects only the immediate, short-run effects of GenAI. The impacts of sustained GenAI use may differ as sellers and consumers adapt their behavior over time and as platforms refine model deployment. Relatedly, we lack data to assess longer-term outcomes such as consumer retention and repeat purchasing. For example, while our results show no deterioration in post-purchase outcomes, as measured by return rates and customer ratings, they do not speak to longer-run consumer responses, such as changes in trust, loyalty, or engagement with the platform. 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, including 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 labor 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 within this environment. As GenAI adoption spreads and competing platforms deploy similar technologies, these relative advantages may diminish. However, if the mechanism we highlight—higher willingness to purchase due to 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.

products included in the experiment and do not capture behavior toward competing products.

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 by creating demand-side value, both within individual workflows and across the platform, through reductions in market frictions and improvements in the consumer shopping experience. These gains are economically meaningful given the scale and maturity of the partner platform. Over the longer run, GenAI's impact may expand as firms move beyond early use cases, capture cost-reduction opportunities, and adapt organizational structures to integrate the technology more effectively. Continued advances in computational speed, accuracy, and domain coverage are likely to further amplify these effects. At the aggregate level, however, widespread adoption raises open questions about equilibrium effects, competitive dynamics, and the persistence of observed gains. Our study provides a first step by offering causal evidence on how GenAI can reshape core retail workflows to improve meaningful business outcomes, while highlighting important directions for future research on general equilibrium effects, 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, and Zhe Yuan (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*.

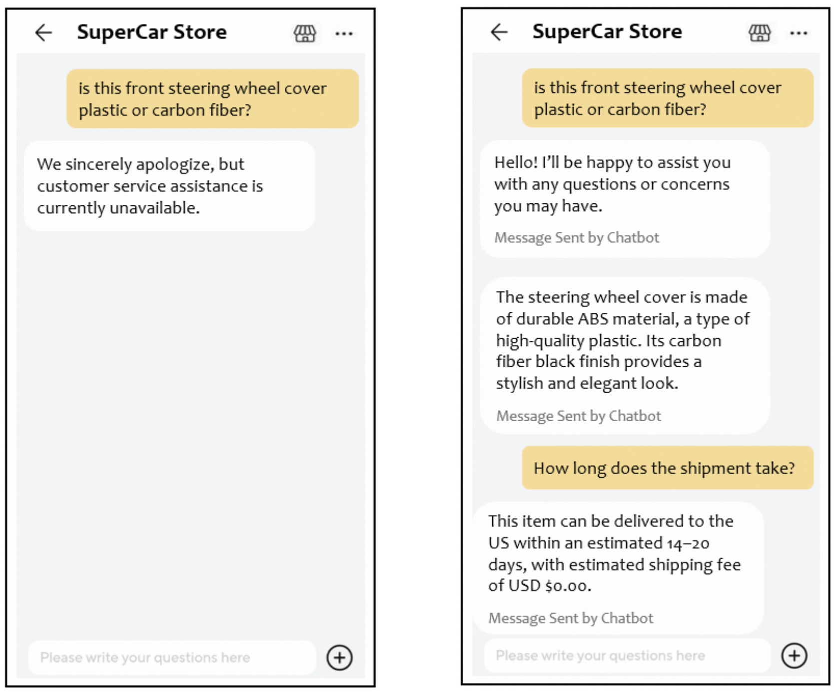
Zhou, Wei, Mingfeng Lin, Mo Xiao, and Lu Fang (2025). "Higher Precision is Not Always Better: Search Algorithm and Consumer Engagement". In: *Management Science* 71.7, pp. 6204–6226.

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



(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, as well as 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 de-

Figure A2: Illustration of Search Query Refinement

(a) Search Results of Consumer Query

(b) Search Results of GenAI-Refined Query

scriptions are layered on top of those written by humans. The AI-generated content provides more comprehensive and structured bullet-point-style descriptions that highlight product features, benefits, and typical use cases.

Figure A3: Illustration of Product Description

Description

Waterproof; 3D Floating Blade System; Intelligent LED Display; Rechargeable & Cordless Design; Long-Lasting Battery Life; Universal Voltage

(a) Human-Created Description

Description

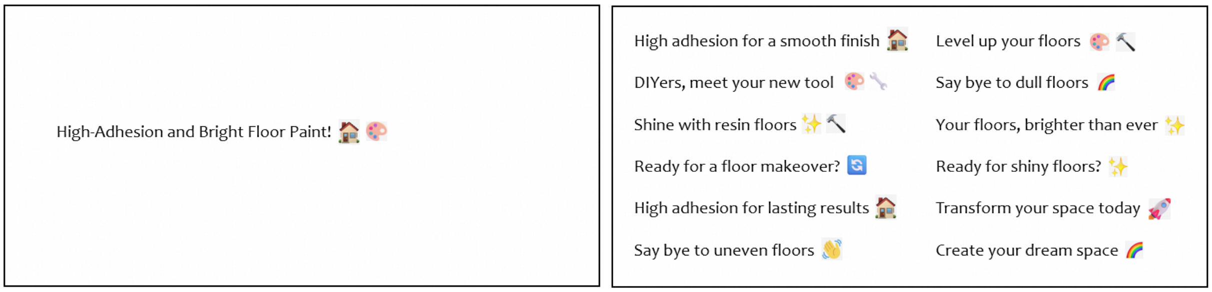
1. **Waterproof & Whole Body Washable:** Directly use and easily rinse under running water, ensuring effortless cleaning and long-lasting performance.
2. **3D Floating Blade System:** Adapts to facial contours for a close, smooth, and accurate shave, reaching every hard-to-reach area.
3. **Intelligent 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; 3D Floating Blade System; Intelligent LED Display; Rechargeable & Cordless Design; Long-Lasting Battery Life; Universal Voltage

(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.

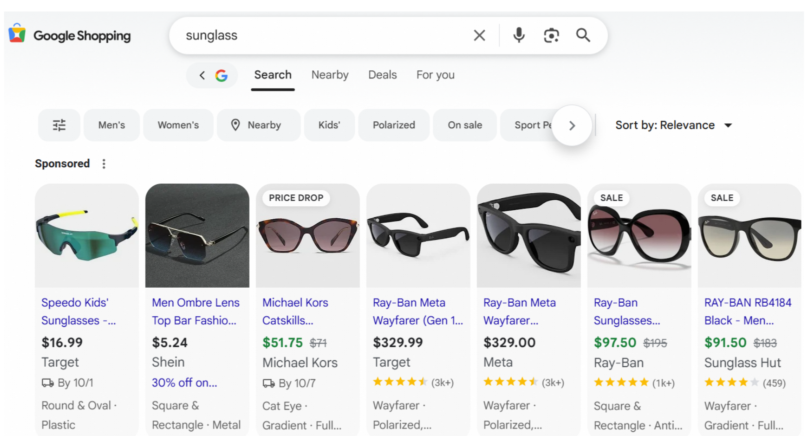
Figure A4: Illustration of Marketing Push Message



(a) Human-Created Push Message (b) GenAI-Created Push Messages

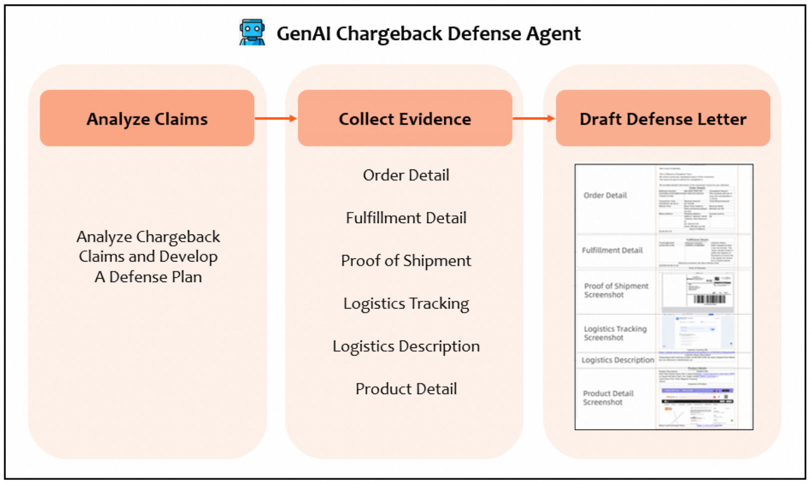
Google Advertising Title In this case, GenAI is applied to optimize human-generated product titles for Google advertising. Since the model was not fine-tuned with e-commerce domain knowledge, the performance of GenAI-optimized titles is lower than that of human-generated titles. For example, the original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



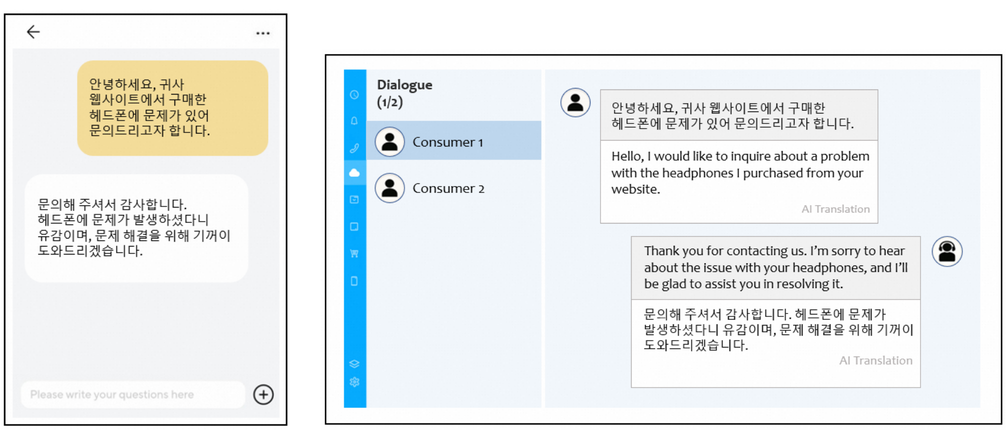
Chargeback Defense Figure A6 illustrates the process by which the chargeback defense agent interprets and analyzes consumer claims, gathers relevant evidence—including transaction records, product details, and shipping information—and drafts persuasive defense letters. By automating this complex workflow,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

Figure A6: Illustration of Chargeback Defense



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 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



(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 Randomization Checks

	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	622,883	

[1] Mean (Std. Dev.) are shown with all values normalized by the corresponding variable's control group mean. For the definitions of each variable, 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 the consumer. The Google Advertising Title experiment is conducted at the product level and thus the unit of observation is the 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 reduce search costs, support decision-making, improve customer 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".

Table C1 presents the impact of the GenAI chatbot on sales, while Table C2 reports its effects on conversion rates and cart value. In both tables, Columns (1)–(2) report effects when the treatment is GenAI Reply, and Columns (3)–(4) report effects for GenAI+Human Reply. Within each set, Columns (1) and (3) use No Service as the control, while Columns (2) and (4) use Human Reply as the control.

Focusing on the comparison between the No Service control and the GenAI Reply treatment (Column 1), we continue to find sizable productivity improvements: sales increase by 16.3% and conversion rate rises by 21.7% (both significant at the 1% level). Using Human Reply as the control (Column 2), the GenAI Reply treatment shows no statistically significant differences in either sales or conversion, suggesting that the GenAI chatbot matches the quality of human service but does not outperform it. When the No Service control is compared with the GenAI+Human Reply treatment (Column 3), the gains are even larger—sales improve by 25.0% and conversion by 29.0%—indicating complementarities between GenAI and human agents. By contrast, relative to the Human Reply control (Column 4), the GenAI+Human Reply treatment yields a marginally significant 11.5% increase in sales, with no statistically significant change in conversion (4.8%), implying that the hybrid approach can enhance revenue even when benchmarked against human agents. Finally, Table C2 shows that all estimated effects on cart value are statistically insignificant.

Table C1: Effects of Pre-sale Service Chatbot on Sales Across Experiments

Treatment: Control:	GenAI reply No Service (1)	GenAI reply Human Reply (2)	GenAI+Human reply No Service (3)	GenAI+Human reply Human Reply (4)
Treat	0.274***	0.0701	0.422***	0.218*
	(0.0995)	(0.0992)	(0.115)	(0.1145)
%Change	16.3%	3.7%	25.0%	11.5%
Observations	44,614	44,736	30,345	30,467
R-squared	0.000	0.000	0.000	0.000

[1] "Sales" represents the total expenditure on product orders.
[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group mean. *** p<0.01, ** p<0.05, * p<0.1.

Table C2: Conversion Rate and Cart Value in Pre-sale Service Chatbot Experiments

Treatment: Control:	GenAI reply No Service (1)	GenAI reply Human Reply (2)	GenAI+Human reply No Service (3)	GenAI+Human reply Human Reply (4)
Extensive margin: Conversion Rate				
Treat	0.0131***	-0.000768	0.0175***	0.00358
	(0.00256)	(0.00261)	(0.00295)	(0.00301)
%Change	21.7%	-1.0%	29.0%	4.8%
Observations	44,614	44,736	30,345	30,467
R-squared	0.001	0.000	0.001	0.000
Intensive margin: Cart Value (conditional on purchase)				
Treat	-1.264	1.220	-0.859	1.624
	(1.036)	(0.929)	(1.203)	(1.078)
%Change	-4.5%	4.8%	-3.1%	6.4%
Observations	3,076	3,300	2,092	2,316
R-squared	0.000	0.001	0.000	0.001

[1] The first block reports the *Conversion Rate* (binary indicator equal to 1 if consumer makes at least one order during the experimental period). The second block reports *Cart Value*, 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 mean. *** p<0.01, ** p<0.05, * p<0.1.

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 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). We find no significant impact on cart value and purchase frequency (Columns 6 and 7).

We also explore further mechanism analysis of the Search Query Refinement application. In Table C4, we find that the likelihood of a product click increases by 0.3% ($p < 0.01$; Column 1), consistent with improved search performance in combination with the documented purchase conversion increase in our main effect. Columns 2 and 3 further indicate that treated consumers view fewer products prior to clicking or purchasing, suggesting a reduction in search intensity. In addition, we observe a statistically significant 2.0% increase in the click-through rate ($p < 0.01$; Column 4), defined as the ratio of product clicks to product views, suggesting that consumers found the exposed products more appealing and chose to seek additional details after viewing the summarized search results. The click-to-order conversion rate (Column 5), defined as the ratio of orders to clicks, remains statistically insignificant. This pattern echoes 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.

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

Table C3: Main Effect of Search Query Refinement

	(1) Views	(2) Clicks	(3) Orders	(4) Sales	(5) Conversion Rate	(6) Cart Value	(7) Purchase Frequency
Treat	-0.549	0.0901***	0.00154	0.0648**	0.00101**	0.370	-0.00315
	(0.887)	(0.0245)	(0.00107)	(0.0314)	(0.000411)	(0.334)	(0.00842)
%Change	-0.18%	1.10%	0.94%	2.93%	1.15%	1.47%	-0.17%
Observations	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	163,381	163,381
R-squared	0.038	0.041	0.019	0.004	0.030	0.010	0.004

[1] "Views" stands for the number of product views in the summarized search results pages. "Clicks" stands for the number of product clicks into product detail pages. "Orders" 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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase. "Purchase Frequency" refers to the number of orders 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 mean. *** p<0.01, ** p<0.05, * p<0.1.

Table C4: More Exploration of Search Query Refinement

	(1) Is Click	(2) Views Conditional on Click	(3) Views Conditional Order	(4) Click-through Rate	(5) Click-to-Order Conversion Rate
Treat	0.00267***	-1.728*	-10.34*	0.000767***	-0.000088
	(0.00057)	(1.034)	(5.362)	(0.000079)	(0.000182)
%Change	0.33%	-0.46%	1.48%	2.02%	-0.35%
Observations	1,849,382	1,508,873	163,381	1,849,382	1,508,873
R-squared	0.013	0.039	0.069	0.003	0.014

[1] "Is Click" measures consumers' likelihood of making a click. It is a binary indicator for click, which equals 1 if a consumer makes at least one click during the experimental period, and 0 otherwise. "Views Conditional on Click" refers to the product views per consumer, conditional on making a click. "Views Conditional on Order" refers to the product views 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, measuring 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, measuring 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 mean. *** p<0.01, ** p<0.05, * p<0.1.

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, $\alpha_{cl(i)}$.

Column 1 of Table C5 shows no statistically significant differences in product clicks between the control and treatment groups. However, conditional on consumers clicking through to product detail pages—where the descriptions are displayed—treated consumers place 1.08% more orders and spend 2.05% more on those orders (Columns 2 and 3). This improvement is also reflected in a 1.27% increase in the conversion rate (Column 4), indicating a higher likelihood of purchase following exposure to AI-generated descriptions. By contrast, we detect no statistically significant changes in cart value or purchase frequency. Taken together, these results suggest that augmenting human-generated product descriptions with AI-generated content primarily affects purchase incidence rather than spending intensity, leading to higher overall sales.

Moreover, we stratify products based on the length of their original, human-generated descriptions, distinguishing between those with no or insufficient text (fewer than 50 words) and those with sufficiently detailed descriptions (more than 50 words). Products in the former group experience a 6.1% increase in sales following augmentation with AI-generated content ($p < 0.05$), whereas products in the latter group show no significant effect (Columns 1 and 2 of Table C6), indicating that richer descriptions facilitate consumer decision-making.[25][26]

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.

[25] Based on internal research and expert surveys conducted by our partner platform, descriptions containing fewer than 50 words are classified as providing insufficient textual information.

[26] The analysis is restricted to English-speaking consumers, as description-length data are available only for English-language content.

Table C5: Main Effect of Product Description

	(1) Clicks	(2) Orders	(3) Sales	(4) Conversion Rate	(5) Cart Value	(6) Purchase Frequency
Treat	0.00233	0.000600**	0.0104**	0.000554***	0.0944	-0.00186
	(0.00183)	(0.000278)	(0.00417)	(0.000187)	(0.0805)	(0.00320)
%Change	0.12%	1.08%	2.05%	1.27%	0.82%	-0.15%
Observations	4,772,937	4,772,937	4,772,937	4,772,937	210,155	210,155
R-squared	0.055	0.008	0.002	0.008	0.010	0.021

[1] "Clicks" stands for the number of product clicks into product detail pages. "Orders" 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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase. "Purchase Frequency" refers to the number of orders 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 mean. *** p<0.01, ** p<0.05, * p<0.1.

Table C6: Product Description: Heterogeneity by Length of Original, Human-generated Description

	(1) Sales ≤50 words	(2) Sales >50 words	(3) Conversion Rate ≤50 words	(4) Conversion Rate >50 words	(5) Cart Value ≤50 words	(6) Cart Value >50 words
Treat	0.0115**	0.0001	0.00047**	0.00044*	0.367	-0.188
	(0.0050)	(0.0064)	(0.00022)	(0.00026)	(0.235)	(0.198)
%Change	6.50%	0.03%	2.43%	1.60%	3.99%	-1.50%
P-value	0.0536		0.5490		0.0653	
Observations	1,628,409	1,628,409	1,628,409	1,628,409	31,759	45,500
R-squared	0.000	0.001	0.003	0.005	0.001	0.004

[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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

[2] "≤50 words" refers to descriptions containing fewer than 50 words, while ">50 words" refers to descriptions containing more than 50 words.

[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group mean. The p-values are from seemingly unrelated estimations testing the equality of % treatment effects across the two groups. *** p<0.01, ** p<0.05, * p<0.1.

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.

Table C7 shows that the use of AI-generated marketing messages leads to increases in both consumer engagement and purchases. Specifically, clicks increase by 3.1%, orders by 2.8%, and total purchase amount by 1.6% (Columns 1–3). In addition, the probability that a consumer makes a purchase rises by 3%, while cart value and purchase frequency remain statistically unchanged (Columns 4–6). Taken together, these results suggest that GenAI enhances the effectiveness of marketing content primarily by expanding consumer participation, consistent with its ability to unlock personalization at scale in settings where human-generated content is constrained by limited resources.

Table C7: Main Effect of Marketing Push

	(1) Clicks	(2) Orders	(3) Sales	(4) Conversion Rate	(5) Cart Value	(6) Purchase Frequency
Treat	0.000529**	0.00005*	0.000402	0.000048***	-0.206	-0.013
	(0.00007)	(0.00003)	(0.000816)	(0.0000218)	(0.453)	(0.0085)
%Change	3.1%	2.8%	1.6%	3.0%	-1.32%	-0.12%
Observations	13,715,528	13,715,528	13,715,528	13,715,528	22,425	22,425
R-squared	0.000	0.000	0.000	0.000	0.000	0.000

[1] "Clicks" stands for the number of clicks on the marketing messages. "Orders" 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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase. "Purchase Frequency" refers to the number of orders 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 mean. *** 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 $\alpha_{c(i)}$.

Table C8 presents the main findings. Columns 1 and 2 indicated a 7.6% decrease in ad views and a 10.2% decrease in ad clicks for the treatment group, respectively. Columns 3, 4, and 5 reported a 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 C8: Main Effect of Google Advertising

	(1) Views	(2) Clicks	(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	4,811
R-squared	0.000	0.000	0.000	0.000	0.000

[1] "Views" represents the number of times the product advertisement is viewed on Google. "Clicks" 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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per product, conditional on the product being purchased.

[2] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group mean. *** 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 chargeback 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 inquiries 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.

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.

Post-purchase Outcomes In Table C9, we provide evidence on two important post-purchase measures of consumer satisfaction with the products, namely product return rates and customer ratings. We do not find evidence that GenAI applications deteriorate such outcomes.

Table C9: Average Treatment Effects of GenAI Adoption on Product Returns and Reviews

	Return Rate			Positive Review Rate		
	(1)	(2)	(3)	(4)	(5)	(6)
Business Workflow	Coefficient	% Change	Obs	Coefficient	% Change	Obs
Pre-sale Service Chatbot	0.0208 (0.0582)	2.08%	3,076	0.0452* (0.0247)	4.52%	1,008
Search Query Refinement	-0.0141 (0.0152)	-1.4%	163,381	0.0031 (0.0040)	0.3%	26,209
Product Description	-0.0386*** (0.0142)	-3.9%	210,155	-0.0042 (0.0034)	-0.4%	31,916
Marketing Push Message	-0.114*** (0.0394)	-11.4%	22,425	-0.0174 (0.0122)	-1.74%	2,853

[1] "Return Rate" is defined as the share of orders that are returned. "Positive Review Rate" is defined as the share of rated orders that receive four- or five-star ratings on a five-point scale. The estimation of Return Rate is conditional on consumers with orders, while the estimation of Positive Review Rate is conditional on consumers provide ratings. Such data are not available for the Google Advertising Title workflow.

[2] Columns (1) and (4) report the estimated coefficients, with standard errors in parentheses. Columns (2) and (5) report % Change, calculated as the treatment effect divided by the control group mean. Columns (3) and (6) report the number of observations. *** p<0.01, ** p<0.05, * p<0.1.

Table D1: Consumer HTE for Pre-sale Service Chatbot

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Purchases	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.268**	0.283**	0.318**	0.230**	0.160	0.343***
	(0.134)	(0.135)	(0.156)	(0.114)	(0.180)	(0.105)
%Change	13.6%	22.4%	15.0%	18.5%	6.86%	27.7%
P-value	0.4568		0.7532		0.0612	
Panel B: Conversion Rate						
Treat	0.0134***	0.0128***	0.0128***	0.0135***	0.0129***	0.0130***
	(0.00340)	(0.00368)	(0.00386)	(0.00321)	(0.00434)	(0.00300)
%Change	19.7%	26.1%	17.1%	29.6%	16.5%	27.0%
P-value	0.4579		0.1218		0.1976	
Panel C: Cart Value						
Treat	-1.469	-0.744	-0.516	-2.331	-2.470	0.157
	(1.275)	(1.770)	(1.387)	(1.525)	(1.568)	(1.320)
% Change	-5.07%	-2.88%	-1.82%	-8.56%	-8.27%	0.61%
P-value	0.7787		0.3702		0.2710	
Observation (Panel A&B)	26,984	17,630	22,570	22,044	18,608	26,006
Observation (Panel C)	2,066	1,010	1,877	1,199	1,609	1,467

[1] We classify consumers into high and low groups and capture their online shopping experience based on registration duration, 30-day pre-experiment login days and 30-day pre-experiment purchase. For each indicator, consumers in the low group are defined as relatively inexperienced if they fall below the median of the corresponding distribution.
[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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.
[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group mean. *** p<0.01, ** p<0.05, * p<0.1.

D Details on Consumer Treatment Effect Heterogeneity

Pre-sale Service Chatbot In this experiment, we focus on the comparison between consumers who are served only by the GenAI chatbot and those who are served by an auto-response indicating service unavailability. Table D1 reports the heterogeneous treatment effects across consumer groups. Our analysis reveals that the gains in sales are more pronounced among newer consumers with shorter registration histories (Column 2), less active consumers with fewer login days (Column 4), and consumers with less past spending (Column 6). However, differences across consumer groups are generally not statistically significant, with the exception of sales across past purchases, for which the difference is significant at the 10% level.

Search Query Refinement Table D2 reports the heterogeneous treatment effect of refining search query with GenAI on different consumer groups. For sales, the benefits are significant and larger for inexperienced consumers (Columns 2, 4, and 6), while the effects for more experienced consumers are smaller and insignificant (Columns 1, 3, and 5). The differences between these two

Table D2: Consumer HTE for Search Query Refinement

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Purchases	
	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)	(0.0566)	(0.0303)	(0.0757)	(0.0291)
%Change	1.27%	5.01%	0.63%	8.16%	-0.56%	7.46%
P-value	0.2137		0.0036		0.0047	
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%
P-value	0.0023		0.0421		0.1122	
Panel C: Cart Value						
Treat	0.329	0.440	-0.015	1.126**	-0.355	1.224***
	(0.431)	(0.526)	(0.452)	(0.460)	(0.484)	(0.450)
%Change	1.34%	1.69%	-0.05%	5.36%	-1.35%	5.13%
P-value	0.8965		0.0363		0.0139	
Observation (Panel A&B)	813,317	1,036,065	890,044	959,338	596,721	1,252,661
Observation (Panel C)	93,362	70,019	104,531	58,850	88,135	75,246

[1] Refer to Table D1 for detailed notes.

consumer groups are statistically significant when consumers are classified by past purchases and past login days. When focusing on conversion rates, we also find statistically significant differences between consumer groups defined by years since registration. 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 applying GenAI to the semantic translation consumer queries.

Product Description In Table D3, we observe a similar pattern to search. There is a significant and more pronounced sales increase for inexperienced consumers (Columns 2, 4, and 6). The differences between high- and low-group consumers are statistically significant for consumers classified by past purchases and past login days. Augmenting product descriptions with AI-generated content substantially enhances the sufficiency and clarity of product information, thereby motivating consumer purchase decisions, particularly among less experienced consumers.

Marketing Push Message In Table D3, we observe a pattern similar to that in Search Query Refinement. Sales increases are larger and statistically significant for less experienced consumers (Columns 2, 4, and 6). Differences between high- and low-experience consumers are statistically significant when consumers are classified by past purchases and past login days. Augmenting product descriptions with AI-generated content improves the sufficiency and clarity of product

Table D3: Consumer HTE for Product Description

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Purchases	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	0.00671	0.0130***	-0.00606	0.0263***	0.000224	0.0206***
	(0.00761)	(0.00455)	(0.00655)	(0.00522)	(0.00718)	(0.00423)
%Change	1.09%	3.06%	-1.02%	6.24%	0.03%	5.63%
P-value	0.2525		0.0001		0.0005	
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%
P-value	0.7441		0.1194		0.5235	
Panel C: Cart Value						
Treat	0.003	0.184*	-0.180	0.438***	-0.127	0.447***
	(0.126)	(0.098)	(0.112)	(0.115)	(0.115)	(0.098)
%Change	0.03%	1.62%	-1.50%	3.96%	-1.06%	4.09%
P-value	0.2496		0.0001		0.0001	
Observation (Panel A&B)	2,035,278	2,737,659	2,322,437	2,450,500	2,386,336	2,386,601
Observation (Panel C)	107,098	103,057	115,708	94,447	129,462	80,693

[1] Refer to Table D1 for detailed notes.

information, thereby increasing the likelihood of purchase, particularly among less experienced consumers.

E Details on Seller Treatment Effect Heterogeneity

Search Query Refinement Table E1 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 and larger increase in sales from treated consumers. By contrast, sales on larger, tenured, and scaled sellers show no significant change. The differences between these two seller groups are statistically significant when sellers are divided by operation years and the number of sub-accounts. When using conversion rates, we also find statistically significant differences between seller groups defined by annual sales. Thus, the GenAI-powered search query refinement generates more values among small sellers.

Marketing Push Message In Table E2, the differences between high- and low-group sellers are all statistically insignificant.

Table D4: Consumer HTE for Marketing Push Message

	(1)	(2)	(3)	(4)	(5)	(6)
	Registered Years		Past Login Days		Past Purchases	
	High	Low	High	Low	High	Low
Panel A: Sales						
Treat	-0.000269	0.00128	0.000693	0.000195	-0.00250	0.00238***
	(0.00115)	(0.00112)	(0.00136)	(0.00100)	(0.00161)	(0.000712)
% Change	-0.91%	6.65%	2.47%	0.84%	-5.89%	20.6%
P-value	0.4519		0.8606		0.0003	
Panel B: Conversion Rate						
Treat	0.0000314	0.0000681**	0.0000122	0.0000737***	-0.0000590	0.000111***
	(0.0000301)	(0.0000312)	(0.0000345)	(0.0000281)	(0.0000416)	(0.0000210)
% Change	1.76%	4.97%	0.71%	4.82%	-2.22%	14.2%
P-value	0.4373		0.2809		0.0000	
Panel C: Cart Value						
Treat	-0.432	0.225	0.285	-0.573	-0.6	0.828
	(0.578)	(0.731)	(0.714)	(0.583)	(0.558)	(0.761)
% Change	-2.62%	1.60%	1.75%	-3.8%	-3.76%	5.62%
P-value	0.4956		0.3551		0.1286	
Observations (Panel A&B)	7,959,821	5,755,707	5,796,905	7,918,623	6,085,062	7,630,466
Observations (Panel C)	14,351	8,074	10,026	12,399	16,025	6,400

[1] Refer to Table D1 for detailed notes.

Table E1: 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**	0.0070	0.0578**	0.00238	0.0624**
	(0.0243)	(0.0180)	(0.0182)	(0.0237)	(0.0174)	(0.0246)
%Change	2.18%	3.68%	0.62%	5.38%	0.24%	5.19%
P-value	0.5661		0.0652		0.0516	
Panel B: Conversion Rate						
Treat	0.000085	0.00105***	0.000444	0.000673**	0.000242	0.00100***
	(0.000291)	(0.000352)	(0.000343)	(0.000306)	(0.000317)	(0.000330)
%Change	0.21%	1.69%	0.76%	1.47%	0.49%	1.85%
P-value	0.0573		0.3493		0.0738	
Panel C: Cart Value						
Treat	0.550	0.238	-0.042	0.773	-0.054	0.613
	(0.554)	(0.267)	(0.287)	(0.486)	(0.325)	(0.426)
%Change	2.06%	1.34%	-0.22%	3.29%	-0.26%	2.75%
P-value	0.7711		0.1484		0.2069	
Observation (Panel A&B)	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382
Observation (Panel C)	76,536	115,804	108,790	85,195	91,774	100,643

[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; (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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

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

Table E2: 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	0.000216	0.000186	0.000788	-0.000386	0.000573	-0.000171
	(0.000594)	(0.000559)	(0.000562)	(0.000591)	(0.000418)	(0.000700)
% Change	1.84%	1.38%	5.69%	-0.341%	9.99%	-0.88%
P-value	0.9436		0.6927		0.1818	
Panel B: Conversion Rate						
Treat	0.0000240	0.0000236	0.0000248	0.0000228*	-0.0000004	0.000048**
	(0.0000147)	(0.0000161)	(0.0000175)	(0.0000131)	(0.0000117)	(0.0000184)
% Change	3.27%	2.69%	2.39%	3.97%	-0.09%	4.20%
P-value	0.8299		0.1826		0.1485	
Panel C: Cart Value						
Treat	-0.221	-0.195	0.430	-1.403	1.237	-0.829
	(0.734)	(0.564)	(0.485)	(0.918)	(0.835)	(0.540)
% Change	-1.38%	-1.27%	3.22%	-7.1%	10.1%	-4.87%
P-value	0.9852		0.0817		0.0467	
Observation (Panel A&B)	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528
Observation (Panel C)	10,212	12,213	14,405	8,020	6,410	16,015

[1] Refer to Table E1 for detailed notes.

Google Advertising Titles Table E3 shows that the differences between high- and low-group sellers are all statistically insignificant.

F Details on Product Treatment Effect Heterogeneity

Pre-sale Service Chatbot Table F1 reports the product-level heterogeneous treatment effects of the Pre-sale Service Chatbot experiment. When focusing on conversion rates, we detect statistically significant differences between product groups classified by annual sales quantity, indicating that the GenAI-driven chatbot service is more effective at converting consumers for tail products.

Search Query Refinement In Table F2, we find that for sales, the gains are larger and statistically significant only for products in low-concentration categories (Column 2), products with fewer sales volume (Column 4), and products with high price level (Column 5). The differences between these two products groups are statistically significant when products are classified by market concentration and annual quantity, indicating that GenAI reduces search frictions in settings with greater product differentiation or limited sales history.

Product Description Table F3 shows that, for sales, statistically significant differences across product groups arise only when products are classified by price level. This pattern is consistent with

Table E3: 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	-4.98%	-4.27%	-5.23%	-4.27%	-5.74%	-4.39%
P-value	0.9288		0.9073		0.9193	
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	-6.33%	-0.33%	-2.25%	-3.73%	-4.61%	-3.18%
P-value	0.3549		0.8248		0.8823	
Panel C: Cart Value						
Treat	-0.0528	-1.588	-0.990	-0.697	-0.960	-0.749
	(1.372)	(1.425)	(1.896)	(1.162)	(3.811)	(1.016)
% Change	-0.16%	-4.41%	-2.84%	-2.09%	-2.28%	-2.28%
P-value	0.4433		0.8971		0.9997	
Observations (Panel A&B)	622,083	621,933	397,339	846,677	140,005	1,104,011
Observation (Panel C)	2,343	2,468	1,411	3,400	468	4,343

[1] Refer to Table E1 for detailed notes.

GenAI being more effective at reducing information asymmetries in settings with higher decision stakes.

Marketing Push Message Table F4 shows that, in most cases, we do not observe statistically significant differences between the two product groups.

Google Advertising Titles Table F5 shows no statistically significant differences between the two product groups.

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	0.138*	0.136**	0.0597	0.214***	0.137*	0.137**
	(0.0821)	(0.0532)	(0.0564)	(0.0800)	(0.0782)	(0.0589)
% Change	21.3%	13.2%	8.77%	21.3%	16.6%	15.9%
P-value	0.4724		0.2697		0.9654	
Panel B: Conversion Rate						
Treat	0.00667***	0.00647***	0.00265	0.0105***	0.00404**	0.00910***
	(0.00196)	(0.00166)	(0.00163)	(0.00199)	(0.00158)	(0.00202)
% Change	26.6%	18.5%	10.3%	30.3%	17.4%	24.5%
P-value	0.3444		0.0182		0.4139	
Panel C: Cart Value						
Treat	-1.3	-1.108	-0.366	-1.990	-0.258	-1.584*
	(1.551)	(1.182)	(1.370)	(1.472)	(2.182)	(0.926)
% Change	-4.22%	-4.47%	-1.38%	-6.86%	-0.725%	6.84%
P-value	0.9704		0.4498		0.3981	
Observation (Panel A&B)	44,614	44,614	44,614	44,614	44,614	44,614
Observation (Panel C)	1,802	1,274	1,223	1,853	1,151	1,925

[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 the category. Products in the low group—referred to as tail products—are those comprising the bottom 50% cumulative share of total 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 experimental period, and 0 otherwise. "Cart Value" refers to the expenditure per consumer, conditional on making a purchase.

[3] Standard errors are in parentheses. % Change is calculated by dividing the treatment effect by the control group mean. *** 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%
P-value	0.0067		0.0332		0.1937	
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%
P-value	0.0888		0.0064		0.7493	
Panel C: Cart Value						
Treat	-0.292	1.106**	-0.098	0.587	0.717	-0.032
	(0.291)	(0.534)	(0.319)	(0.411)	(0.615)	(0.132)
%Change	-1.49%	4.35%	-0.55%	2.46%	2.04%	-0.28%
P-value	0.0208		0.2052		0.2583	
Observations (Panel A&B)	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382	1,849,382
Observation (Panel C)	101,975	82,912	84,799	109,265	82,272	108,485

[1] 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%
P-value	0.1309		0.9641		0.0703	
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)	(0.000100)	(0.000164)
% Change	1.38%	0.81%	2.07%	0.69%	3.32%	0.63%
P-value	0.5014		0.0845		0.0047	
Panel C: Cart Value						
Treat	0.186**	-0.019	0.018	0.138	0.092	0.032
	(0.089)	(0.141)	(0.108)	(0.104)	(0.204)	(0.071)
%Change	1.72%	-0.16%	0.16%	1.24%	0.59%	0.33%
P-value	0.1895		0.4258		0.8605	
Observations (Panel A&B)	4,772,937	4,772,937	4,772,937	4,772,937	4,772,937	4,772,937
Observation (Panel C)	133,779	83,687	106,252	115,810	57,458	159,492

[1] 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	0.000255	0.000146	-0.0000554	0.000457	-0.000429	0.000831
	(0.000742)	(0.000339)	(0.000349)	(0.000737)	(0.000511)	(0.000636)
% Change	1.25%	3.03%	-0.92%	2.39%	-5.21%	4.90%
P-value	0.9436		0.6927		0.1818	
Panel B: Conversion Rate						
Treat	0.0000263	0.0000213**	0.0000115	0.0000361*	0.000013	0.000034*
	(0.0000198)	(0.00001)	(0.0000108)	(0.000019)	(0.00001)	(0.000019)
% Change	1.98%	7.60%	2.90%	2.97%	4.06%	2.68%
P-value	0.8299		0.1826		0.1485	
Panel C: Cart Value						
Treat	-0.108	-0.730	-0.566	-0.0896	-2.292*	0.285
	(0.504)	(1.038)	(0.772)	(0.546)	(1.381)	(0.444)
% Change	-0.71%	-4.24%	-3.71%	-0.57%	8.92%	2.17%
P-value	0.6095		0.6106		0.0825	
Observations (Panel A&B)	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528	13,715,528
Observations (Panel C)	18,432	3,993	5,498	16,927	4,483	17,942

[1] 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%
P-value	0.9056		0.4243		0.6312	
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%	-4.3%	-4.1%	-2.8%
P-value	0.7661		0.8272		0.8407	
Panel C: Cart Value						
Treat	-0.819	-0.723	-0.389	-1.477	-1.546	0.0417
	(1.258)	(1.608)	(1.221)	(1.700)	(1.567)	(1.159)
% Change	-2.48%	-2.07%	-1.18%	-4.18%	-3.69%	0.16%
P-value	0.9439		0.6157		0.4577	
Observations (Panel A&B)	714,642	529,374	714,642	529,374	714,872	529,144
Observations (Panel C)	2,906	1,905	3,132	1,679	2,347	2,464

[1] Refer to Table F1 for detailed notes.