# Reasoning AI, Consumer Search, and Purchase: Evidence from an Online Platform Field Experiment

May 15, 2026

**Abstract**

We study how more deliberative conversational assistance reshapes consumer search and purchase behavior on digital platforms. We examine this question in the context of reasoning AI, using a large-scale field experiment on a major online travel platform, where more than 510,000 users were randomly assigned to an AI assistant powered by either a reasoning model (DeepSeek-R1) or a comparable non-reasoning model (DeepSeek-V3). Access to the reasoning assistant reduced hotel bookings by 2.5%. The effect operates through the search funnel: users initiated fewer search queries, browsed fewer hotel options, and clicked fewer hotels for deeper inspection. Additional evidence suggests that this narrowing is more consistent with the assistant's more informative replies reducing consumers' perceived need for further search than with a time-cost-based explanation. At the same time, the reasoning assistant increased subsequent engagement with the chat feature. Our findings show that more deliberative conversational assistance can narrow consumer search, creating a trade-off between assistant engagement and search-driven purchases.

**Keywords:** AI assistant, Large Language Model, reasoning model, e-commerce, consumer search
**Word Count:** 6,591

# 1 Introduction

Generative AI is changing how consumers search for products on digital platforms. A growing number of online marketplaces now embed AI assistants directly into the shopping experience, allowing consumers to describe their needs in natural language and receive conversational guidance alongside traditional search. Rather than relying only on keywords and listings, consumers can now use embedded assistants to ask questions, receive recommendations, and navigate back to listings or product pages. These assistants therefore create a new layer of information acquisition before consumers continue searching or purchase.

For platforms, a central design choice is how deliberative this conversational assistance should be. This is not simply a technical model-selection problem. As large platforms introduce reasoning-capable embedded assistants, such as Amazon's Rufus and Expedia's Romie, they increasingly face the operational decision of how much reasoning, explanation, and synthesis these assistants should provide within the search and purchase process. Recent advances in reasoning-capable large language models make this choice especially salient.[1] Compared with non-reasoning LLMs, reasoning models are designed to deliberate before answering (e.g. Guo et al., 2025). In consumer-facing assistants, this capability can translate into more structured and justificatory responses: the assistant may synthesize information, weigh trade-offs, and provide recommendations that appear more complete. Reasoning models therefore offer a managerially important setting for studying the effects of making conversational assistance more deliberative.

In order to understand how deliberative conversational assistance affects consumer search and purchase, we draw on emerging multi-stage models of consumer search, which depart

[1]Platform examples include Amazon's Rufus and Expedia's Romie. Amazon describes Rufus as a generative AI-powered conversational shopping assistant and notes that its development expanded into use cases requiring "advanced reasoning" and "multi-step reasoning."; see `https://aws.amazon.com/blogs/machine-learning/how-rufus-scales-conversational-shopping-experiences-to-millions-of-amazon-customers-with-amazon-bedrock`. Expedia Group describes Romie as a generative AI-powered travel assistant and notes that "on-the-fly reasoning" helps travelers refine their preferences; see `https://medium.com/expedia-group-tech/elevating-travel-experiences-with-ai-acdb2cf2ec13`. More broadly, major model providers have introduced reasoning, thinking, or extended-thinking capabilities, including OpenAI, Anthropic, Google, and DeepSeek.

from one-stage search by allowing consumers to acquire information through multiple sequential actions, such as discovering or initiating additional search and then inspecting alternatives more deeply (e.g. Gibbard, 2022; Greminger, 2022). Once search is viewed as a sequence of information-acquisition margins, the effect of making assistance more deliberative becomes ambiguous. By synthesizing information and explaining trade-offs, a more deliberative assistant may make consumers' subsequent search more productive, helping them identify relevant options and rule out poor matches more efficiently. But the same features may also compress the search process by leading consumers to narrow consideration sets or stop exploring earlier, even when additional search could still be valuable. Our empirical analysis therefore examines how making conversational assistance more deliberative changes behavior across the search funnel and affects purchase. If search activity declines, a further question is why. One possibility is that the reasoning assistant discourages additional search by increasing the burden of the interaction; another is that its longer and more information-rich replies lower the perceived marginal value of further search.

To study this question, we exploit a large-scale field experiment on a major online travel platform centered on its AI assistant, Ctrip Wendao. The assistant is embedded in the platform's standard search and recommendation pages, allowing it to shape the consumer search process at multiple points. In particular, it can recommend specific hotels through clickable links and suggest search queries that take users directly to hotel listings, thereby influencing both search initiation and browsing. Our experiment exploits a model upgrade in this assistant. Users in the treatment group were assigned to an assistant powered by DeepSeek-R1, a reasoning model, while users in the control group interacted with the existing DeepSeek-V3 assistant, a comparable non-reasoning model. Relative to the control assistant, the reasoning assistant provides more deliberative conversational assistance, with responses that are more synthesized, structured, and justificatory. This design allows us to trace effects across the full funnel, from chat interaction to search initiation, browsing, clicks, and purchase.

The experiment ran for eight days, from April 3 to April 10, 2025. During this period, logged-in users who initiated a session with the AI assistant were randomly assigned at the user level, with equal probability, to either the treatment or control group; assignment remained fixed for each user throughout the experiment. This randomization resulted in a sample of over 510,000 users. For each user, we collected a comprehensive dataset including demographics, full chat transcripts with response-level timestamps, and complete purchase records across all product categories. In addition, for the hotel category, we observe detailed data on search queries, browsing activity on listing pages, and clickstream behavior.

The experiment yields a surprising top-line result. Rather than boosting sales, assignment to the reasoning assistant reduced hotel bookings by 2.5%, with no discernible effect on flights or other products. Users assigned to the reasoning assistant initiated fewer search queries, browsed fewer hotel listings, and clicked fewer hotel options, while conditional click-through and conversion remained largely unchanged. These patterns suggest that the decline in hotel bookings reflects compression of the search process rather than weaker conversion conditional on search. The category pattern is consistent with this search-funnel interpretation: we find no discernible effect on flight or other-product (e.g., attraction tickets) orders, where users typically face a narrower feasible set and a more standardized comparison process than in hotel search, leaving less scope for reduced search, browsing, and clicking to affect purchase.

We next examine why search activity declines. The reasoning assistant is indeed slower: users wait longer for AI replies and spend more time in the chat interface. However, treated users do not spend less total time searching in the hotel channel, and accounting for chat-time and waiting-time exposure does little to attenuate the effects on search queries, browsing, or clicks. By contrast, accounting for differences in reply informativeness substantially attenuates these effects, and the assistant's replies become more information-rich under treatment. While these analyses are not a direct test of users' beliefs, the patterns suggest that the reduction in search is not primarily a mechanical consequence of slower responses or time displacement. Instead, they are more consistent with longer and more informative replies

lowering the perceived marginal value of additional search.

Although the reasoning assistant reduced short-run hotel sales, it increased engagement with the chat feature. Users assigned to the treatment group were more likely to send a message and to return to the feature within 30 days. Engagement also persisted well beyond the experiment: over the three months after the experiment, treated users made 11% more chat requests.

Several checks help rule out alternative interpretations of the main order effect. The reduction in hotel bookings does not appear to reflect purchase postponement, differential cancellation, or an offsetting shift toward higher-value bookings; nor are the main funnel results sensitive to alternative specifications or sample definitions. Moreover, the effect is also present among users with no prior experience using the platform's AI assistant, which makes it less likely that the results are driven simply by incumbent users reacting negatively to a model change. The effects vary across consumer segments: the negative effect on hotel purchases is more pronounced among female users and older, more experienced consumers, while search effects are broader but not identical across funnel stages.

To the best of our knowledge, this is the first large-scale field experiment to compare a reasoning-based platform assistant with a comparable non-reasoning assistant in a consumer search setting. The paper shows that making conversational assistance more deliberative can affect not only final purchase, but also the sequence of search actions through which consumers reach purchase decisions. In our setting, the reasoning assistant compresses the hotel search funnel without improving conditional conversion, suggesting that richer and more structured assistance can narrow search rather than simply make search more productive. For managers, the results point to a clear trade-off: more deliberative AI may increase engagement with the assistant while reducing search exploration and short-run sales.

# 2 Literature Review

Our work joins a fast-growing stream of research exploring the impact of generative AI on marketing. Much of this early work focuses on using LLMs for market research (Brand et al., 2023; Gui and Toubia, 2023; Dominguez-Olmedo et al., 2024; Goli and Singh, 2024; Li et al., 2024). Other scholars have begun to explore generative AI's role across the entire customer journey, examining its use in personalized advertising (Kapoor and Kumar, 2025), content recommendation (Ye et al., 2025), AI-powered search summaries (Zheng et al., 2025), and after-sales service (Ni et al., 2025; Zhang and Narayandas, 2026). Recent work by Padilla et al. (2025) shows that LLM adoption reduces online search and traffic to smaller websites, suggesting that LLMs can substitute for traditional search tools. We extend this literature by using a field experiment comparing reasoning and non-reasoning shopping assistants to show that making conversational assistance more deliberative reduces search initiation, browsing, and deeper inspection, with consequences for purchase.

AI shopping assistants existed long before the emergence of modern LLMs, and studies of these earlier, simpler assistants generally find positive effects on consumer demand (Luo et al., 2019; Sun et al., 2025). Related descriptive work in the same setting shows that platform-embedded AI chat often appears alongside traditional search and is used disproportionately for exploratory, hard-to-keyword tasks (Yan et al., 2026). Lin et al. (2025) use a lab experiment to study a prototype LLM assistant and find that it increased both shopping time and purchase likelihood. We extend this literature by moving to a large-scale field setting and showing that more deliberative assistance can increase engagement with the assistant while compressing search activity and reducing short-run purchases.

We also contribute to the consumer search literature by studying how platform-embedded conversational assistance affects sequential search behavior. Broadly, our paper relates to work on online consumer search and pre-purchase search behavior (e.g., Honka et al., 2019, 2024; Ursu et al., 2025). More specifically, our framework builds on recent multi-stage search models that separate information acquisition into sequential decisions, such as whether

to discover or initiate additional search and whether to inspect alternatives more deeply (Greminger, 2022; Gibbard, 2022). Our results also relate to work on platform-guided search, including AI-generated summaries, recommendations, and platform steering (Nocke and Rey, 2024; Zheng et al., 2025; Yuan et al., 2025; Zhou et al., 2025). We show that making conversational assistance more deliberative can reduce search initiation, browsing, and deeper inspection without improving conditional conversion.

# 3 Field Experiment

## 3.1 Institutional Details

Our study takes place on Ctrip, China's largest online travel agency (OTA). As a dominant market leader, the platform serves over 165 million users and captures more than half of China's domestic travel market. Similar to Expedia or Booking.com, Ctrip offers a comprehensive suite of travel services, including hotels and flights, in a single app. Over 90% of the platform's traffic comes from its mobile app, which is the focus of our study. Below, we first outline the baseline hotel booking flow and then describe where the AI assistant can intervene.

The primary channel for hotel bookings follows a standard search-to-purchase funnel (Appendix Figure A.1). From the home screen (Panel A), users initiate a search (Panel B) by selecting a city, dates, and other criteria. They are then taken to a listings page (Panel C) displaying available hotels with key details such as ratings and price, from which they can click through to a hotel's detail page to complete the booking. While consumers can also land on a detail page through other entry points, such as content pages, front-page recommendations, or the AI assistant, about 80% of hotel orders in our setting originate from user clicks on the listings page. We therefore focus our analysis on search and conversion behavior within this hotel funnel.[2]

[2]We further show that the main treatment effect on hotel orders is driven by conversions from hotel

Our experiment centers on "Ctrip Wendao," the platform's generative AI assistant. Launched in July 2023, the assistant is designed to support the travel planning and booking process from initial planning to final purchase. As shown in Figure 1, users access Wendao through an icon on the app's home screen, which opens the chat interface that we vary across treatment and control. In both conditions, users can minimize the chat window to return to the standard search flow, browse hotel listings, and click on products, and they can move back and forth between chat and search during the same session. This design is not unique to Ctrip; Amazon's Rufus and Expedia's Romie similarly let users move between chat and conventional search or browsing (Appendix Figure A.2).

The assistant can intervene in this funnel in three ways. First, it can link users directly to a hotel's detail page when a specific property is mentioned in chat through a clickable hyperlink. Second, it can move users into the listings view by offering suggested search queries appended to replies that open results without the user typing keywords or pre-setting filters. Third, it can shape search more indirectly by influencing which filters or keywords users choose and how they navigate within the listings, such as how broadly or deeply they explore.

Our experiment leverages a major version upgrade of Ctrip Wendao in April 2025, which transitioned the underlying model from DeepSeek-V3, a non-reasoning LLM, to DeepSeek-R1, a reasoning model.[3] Both models share the same underlying architecture and can answer standard travel-related queries such as finding hotels near a landmark, showing available flights, or checking amenities. The main difference lies in their training objectives. DeepSeek-V3, the control model, was tuned to provide quick, functional answers to routine requests, whereas DeepSeek-R1, the treatment model, was trained through a multi-stage process that emphasizes step-by-step reasoning and logical consistency. As a result, R1 tends to generate more deliberative and structured responses, for example by clarifying constraints, checking distances or times, justifying recommendations, and producing travel plans tailored to

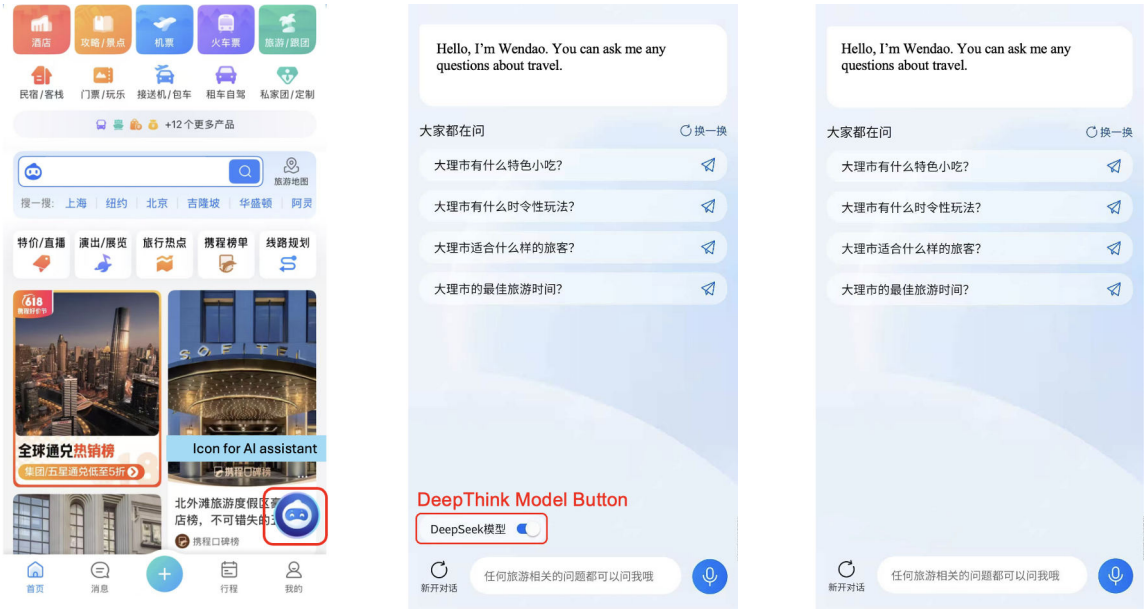
listings pages.

[3]DeepSeek-V3 was released on December 26, 2024, while DeepSeek-R1 was released on January 20, 2025.

individual users (Guo et al., 2025). Sample responses from the two models are shown in Appendix Table A.1.

The experiment ran from April 3 to April 10, 2025. Treatment assignment was at the user level. When a logged-in user clicked the Wendao icon for the first time during the experimental window, the platform randomly assigned that user with equal probability to either the treatment (reasoning model) or control (non-reasoning model) condition. This assignment was fixed for the duration of the experiment, so any subsequent sessions by the same user remained in the same condition. In the treatment group, users were assigned to the reasoning model by default, but could use a button to switch back to the non-reasoning model (Figure 1, Panel B) or simply exit without interacting with the assistant. We therefore estimate intent-to-treat effects throughout. Although repeated use was allowed, less than 0.4% of users in our final sample initiated more than one chat session during the eight-day experimental window.

Figure 1: Experiment Design




Panel A: Start of Experiment Panel B: Treated Group Panel C: Control Group

*Note:* The experimental interface. Panel A shows the icon used to access the AI assistant from the app's home screen. Panel B displays the interface for the treatment group, featuring a button that allows users to opt out of the DeepSeek-R1 reasoning model (enabled by default). Panel C shows the control group's interface, which uses the baseline DeepSeek-V3 model.

## 3.2 Data

Our initial dataset includes 514,402 users who entered the assistant interface during the experiment. Per the experimental design, these users were randomly assigned and approximately evenly split between the treatment group (257,486 users) and the control group (256,916 users). For each user, we observe account information, including demographics, platform tenure, device type, and membership status. We then collect behavioral outcomes during the experimental window, including complete transaction records across hotel, flight, and other travel product bookings, full chat transcripts with the AI assistant, response-level timestamps, and measures of time spent in chat and search. Detailed search-funnel data are available only for hotels, for which we observe search queries, browsing activity on list-

ing pages, and clickstream behavior. Finally, we observe post-experiment activity for up to three months after the experiment ends, which allows us to examine the persistence of chat engagement and hotel-order outcomes beyond the initial eight-day window.

To arrive at our final regression sample, we begin with the 514,402 users in the full experimental sample and remove 16,837 users (3.27%). We apply two trimming rules intended to retain ordinary individual user behavior. First, we remove users above the 99th percentile of search, browsing, or click activity within the hotel channel. These extreme observations are more likely to reflect bot-like or otherwise abnormal activity than typical consumer search behavior; for example, the 99th-percentile cutoff for total search queries corresponds to 137 searches within the eight-day outcome window. Second, we remove users above the 99.9th percentile of booking activity. We use a less aggressive cutoff for bookings because order outcomes are much sparser and have much lower means than search, browsing, and click outcomes; the 99.9th-percentile rule targets the small number of users whose booking volumes are inconsistent with ordinary individual travel planning. Based on internal discussions with Ctrip, users in this extreme upper tail of order volume are more likely to be non-individual accounts, such as business accounts or travel agents booking on behalf of others. The final sample contains 497,565 users (summary statistics reported in Table 1). Appendix Table A.2 shows that the main results remain qualitatively similar under looser and tighter cutoff rules.

As a validity check, we confirm that the randomization was successful. Panel A in Appendix Table A.3 shows that the treatment and control groups were well-balanced on key demographics and pre-experiment behavior. Panel B further confirms that the trimming rule removes roughly the same share of users from the control and treatment groups, suggesting that trimming does not materially change the distribution of treatment assignment in the final sample.

Table 1: Summary Statistics for Key Variables

| | Control (N = 248,581) | | Treated (N = 248,984) | |
|---|---|---|---|---|
| Variable | Mean | SD | Mean | SD |
| Panel A: User Demographics | | | | |
| Female | 0.545 | 0.498 | 0.547 | 0.498 |
| Age Group | 2.731 | 0.843 | 2.730 | 0.842 |
| New User | 0.487 | 0.500 | 0.487 | 0.500 |
| Star User | 0.266 | 0.442 | 0.267 | 0.443 |
| Device: iOS | 0.350 | 0.477 | 0.350 | 0.477 |
| VIP User | 0.098 | 0.297 | 0.098 | 0.297 |
| Panel B: Order Behavior | | | | |
| Total Hotel Orders | 0.158 | 0.531 | 0.154 | 0.521 |
| Total Flight Orders | 0.182 | 0.681 | 0.180 | 0.678 |
| Total Other Orders | 0.088 | 0.407 | 0.089 | 0.412 |
| Panel C: Search, Browse, and Click Behavior | | | | |
| Total Search Time (min) | 49.917 | 236.530 | 49.188 | 232.753 |
| Total Search Queries | 13.251 | 21.272 | 13.143 | 21.182 |
| Total Browse | 161.749 | 273.282 | 160.238 | 271.513 |
| Total Clicks | 3.362 | 8.587 | 3.275 | 8.404 |
| Panel D: Chat Behavior and Retention | | | | |
| Pre-experiment Chat User | 0.197 | 0.398 | 0.198 | 0.399 |
| Used Chat | 0.066 | 0.249 | 0.071 | 0.258 |
| Total Chat Time (min) | 0.129 | 2.796 | 0.152 | 3.074 |
| Retained (30 Days) | 0.004 | 0.066 | 0.005 | 0.070 |
| Total Chat Requests (Months 1–3) | 0.126 | 1.336 | 0.140 | 2.371 |
| Total Orders (Months 1–3) | 2.724 | 6.561 | 2.699 | 5.871 |

*Note:* All variables are measured at the user level. This table reports summary statistics separately for the control and treated groups. The full sample includes 248,581 control users and 248,984 treated users. In total, 34,288 users (6.89%) used the chat at least once, including 16,500 control users and 17,788 treated users. Only 1,937 users had more than one chat session, accounting for 0.39% of all users and 5.65% of chat users.

# 4 Results

## 4.1 Main Effects on the Hotel Search Funnel

We estimate intent-to-treat effects of assignment to the reasoning assistant on user-level outcomes. Random assignment ensures that these estimates identify the intent-to-treat effect of assignment to the reasoning-assistant condition relative to the non-reasoning-assistant condition. Because the main search-funnel outcomes are nonnegative counts with many zeros and substantial over-dispersion, we use negative binomial (NB) models as our default specification and report coefficients as $e^{\beta} - 1$, which are interpretable as percent changes in the expected outcome relative to the control group. For outcomes measured as rates or durations, we use OLS specifications, as indicated in each table.

Table 2 reports the intent-to-treat effects of assignment to the reasoning AI assistant on the four main outcomes along the hotel search funnel. The results reveal a coherent pattern. Users assigned to the reasoning assistant initiated 0.8% fewer search queries, browsed 0.9% fewer hotel listings, and clicked 2.6% fewer hotels for deeper inspection. This compression of search is accompanied by a 2.5% decline in hotel orders. All four effects are statistically significant at conventional levels. These results suggest that access to the reasoning assistant narrows exploration within the hotel channel, with the decline in purchases emerging alongside clear reductions in the search behaviors that typically precede booking.

To understand whether the purchase decline is more consistent with reduced entry into the search funnel or weaker performance conditional on reaching later stages, we examine two sets of diagnostic analyses. First, Appendix Table A.4 reports sequential regressions that add post-treatment search, browse, and click outcomes as controls. These specifications should not be interpreted as randomized estimates of direct effects after conditioning on mediators. Rather, they provide a decomposition of how the hotel-order coefficient changes as search-funnel outcomes are added. The negative treatment effect on hotel orders falls from 2.5% to 1.1% and becomes statistically insignificant once we control sequentially for search

queries, browse, and clicks, consistent with the order decline operating through reduced search activities.

Second, Appendix Table A.5 reports conditional funnel outcomes among users who reach later stages of the funnel. Because these samples are selected on post-treatment behavior, we interpret the results descriptively rather than as standalone causal treatment effects. With that caveat, the estimates provide little evidence that the reasoning assistant worsens outcomes conditional on reaching later stages: average browse per search, click-through rates, and conversion rates remain statistically unchanged.

We next examine whether the effect operates through direct AI recommendations inside the chat interface. Appendix Table A.6 shows that the reasoning assistant provides users with more hotel suggestions, but does not increase orders on AI-mentioned hotels. Appendix Table A.7 further shows that the negative order effect is concentrated in hotel orders originating from hotel listing pages rather than other channels. Together, these results suggest that the decline in hotel purchases arises primarily through users' search and browsing behavior on the main hotel funnel, rather than through direct purchases inside the chat interface.

Table 2: ITT Effect of Reasoning AI Assistant on Search and Purchase Outcomes

| | Total Search Queries | Total Browse | Total Clicks | Total Hotel Orders |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Reasoning AI Assistant | $-0.008^{***}$ | $-0.009^{***}$ | $-0.026^{***}$ | $-0.025^{***}$ |
| | (0.003) | (0.003) | (0.007) | (0.009) |
| Model | NB | NB | NB | NB |
| Reported effect | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* This table reports the effect of assignment to the reasoning AI assistant on user search and purchase outcomes in the hotel channel during the 8-day experimental window. Columns 1–3 report intent-to-treat effects on total search queries, total browse, and total clicks. Column 4 reports the negative binomial estimate for total hotel orders. All coefficients are reported as $\exp(\beta) - 1$ and are interpreted as percentage changes in the expected count of the dependent variable. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The category asymmetry is also informative. While the reasoning assistant significantly

reduces hotel orders, we find no corresponding effect on flight orders or on orders for other travel products (Appendix Tables A.8 and A.9). This pattern is consistent with the mechanism operating most strongly when purchase depends on continued browsing and inspection. Hotel choice typically requires consumers to compare many differentiated alternatives along attributes such as location, neighborhood fit, amenities, ratings, reviews, price, and room characteristics, many of which are difficult to evaluate from a single query. Flights, by contrast, typically involve a much smaller feasible option set than hotels for a given trip, with alternatives compared along a smaller set of common attributes such as departure time, duration, stops, airline, and price. There is therefore less scope for the assistant to reduce browsing and product-level inspection in a way that affects final purchase.

The hotel-order decline also does not appear to reflect intertemporal substitution, cancellation behavior, or a shift toward fewer but higher-value purchases. As shown in Appendix Figure A.3, the treatment effect does not show a rebound during the entire experimental window, which is inconsistent with a pure postponement story in which treated users merely purchase later. This interpretation is also consistent with the post-experiment order patterns discussed in Section 4.3, where the short-run hotel-order effect attenuates over time rather than being offset by a subsequent positive rebound. This is reinforced by the institutional fact that booking activity in our setting is heavily front-loaded, with over 77.6% of purchasers booking within the first three days of their visit. More broadly, Appendix Table A.10 shows that the decline is not driven by differential cancellation patterns: the treatment has no significant effect on canceled orders, but reduces completed orders. Nor does the decline appear to be offset by a shift toward higher-value purchases: Appendix Table A.11 shows more negative effects in the middle order-amount tiers, and the treatment also reduces users' total weighted order amount.

Our main findings are robust across alternative specifications and sample definitions. The negative effect on hotel orders remains when demographic controls are added (Appendix Table A.12) and is similar under OLS and Poisson specifications (Appendix Table A.13). The

search, browse, and click results are likewise robust in Poisson models (Appendix Table A.14). In addition, Appendix Table A.2 shows that the main funnel results remain qualitatively similar under both looser and tighter outlier cutoff rules. Appendix Table A.15 further indicates that the browse and click estimates are more highly powered, while the search and hotel-order estimates remain precise enough to detect economically meaningful effects.

## 4.2 Timing, Reply Structure, and Mechanism Evidence

The reasoning assistant compresses the search funnel, but why? We examine two broad accounts. Under a cost-based mechanism, the reasoning assistant may discourage continued search by increasing waiting time or time spent in chat. We capture this account using chat-time exposure, which includes total time spent in the chat interface and waiting-time exposure from a user's request to the first token of the assistant's final answer. Under a belief-based or perceived-completeness mechanism, the assistant's more informative replies may lower the perceived marginal value of additional search by making users feel that they already have a sufficiently informative shortlist. We capture this account using content exposure, measured by the amount of information entropy contained in the assistant's reasoning text and final answer.[4] We then use mediation-style attenuation specifications to ask whether the treatment effects on search-funnel outcomes are more strongly attenuated after conditioning on chat-time exposure or on content exposure.

We begin by showing that assignment to the reasoning assistant significantly changes the proposed mediators themselves. As shown in Appendix Table A.16, the treatment increases both chat-time exposure and content exposure: users wait longer, spend more time in the chat interface, and receive longer and more information-rich assistant responses. We then

[4]The results are very similar when we instead measure content exposure using the length of the assistant's reply text. We use entropy in the main specification because it captures information volume beyond the mechanical length of the response. For users who entered the assistant interface but did not submit a request and therefore did not receive an AI response, we set chat-time and content exposure to zero. These variables should therefore be interpreted as realized exposure measures under assignment, and the corresponding estimates as intent-to-treat effects rather than conditional comparisons among the selected set of users who submitted requests.

ask which of these changes appears most closely related to the decline in search activity.

Table 3: Mechanism Evidence: Conditioning on Chat-Time and Content Exposure

| Panel A | Total Search Queries | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Reasoning AI Assistant | $-0.008^{***}$ | $-0.009^{***}$ | $-0.004$ |
| | (0.003) | (0.003) | (0.003) |
| Panel B | Total Browse | | |
| | (4) | (5) | (6) |
| Reasoning AI Assistant | $-0.009^{***}$ | $-0.010^{***}$ | $-0.004$ |
| | (0.003) | (0.003) | (0.003) |
| Panel C | Total Clicks | | |
| | (7) | (8) | (9) |
| Reasoning AI Assistant | $-0.026^{***}$ | $-0.028^{***}$ | $-0.018$ |
| | (0.007) | (0.007) | (0.011) |
| Model | NB | NB | NB |
| Chat-time controls | No | Yes | No |
| Content controls | No | No | Yes |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 497,565 | 497,565 | 497,565 |

*Note:* This table reports mechanism evidence for search, browse, and click outcomes. For each outcome, the first column reports the baseline ITT effect, the second adds chat-time exposure controls, and the third adds content exposure controls. Chat-time controls include two separate measures in seconds: total chat-time exposure from entry to exit and waiting-time exposure from user request to the first token of the final answer, excluding reasoning text. Content controls include two separate information volume measures: total entropy in the AI reasoning text and total entropy in the final answer text, where entropy measures the amount of distinct content contained in the text. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 3 implements this mediation-style attenuation exercise. Adding chat-time exposure controls leaves the treatment effects on total search, browse, and clicks largely unchanged. By contrast, adding content-exposure controls substantially attenuates the treatment effects, which become statistically insignificant across all three funnel outcomes. This pattern is difficult to reconcile with a purely cost-based account driven by waiting time or chat-time displacement alone. Instead, it is more consistent with a belief-based or perceived-completeness mechanism, in which the reasoning assistant's more information-rich replies lower users' per-

ceived marginal value of further search.

Table 4: ITT Effects on User Time Allocation

| | Total Search Time (min) (1) | Total Chat Time (min) (2) |
|---|---|---|
| Reasoning AI Assistant | −0.729 | 0.023*** |
| | (0.665) | (0.008) |
| Baseline | 49.553 | 0.141 |
| Model | OLS | OLS |
| User Demographics | No | No |
| Observations | 497,565 | 497,565 |

*Note:* This table reports ITT effects of the reasoning AI assistant on user-level time allocation on the platform. Column 1 reports total search time (minutes) within the hotel channel, and Column 2 reports total time (minutes) spent on the chat interface. Both outcomes are measured in minutes and estimated using OLS. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Other evidence points in the same direction. Table 4 shows that treated users spend more time in chat, but do not spend less total time searching within the hotel channel, which is inconsistent with a simple time-displacement effect in which users run out of time for browsing. In addition, Appendix Table A.17 shows that the assistant's replies become more attribute-rich and structured under treatment, consistent with the idea that they provide more synthesized and seemingly complete recommendations.

These findings help connect the mechanism evidence back to our conceptual framework. More deliberative replies could compress search activity either by increasing the burden of interacting with the assistant or by making its recommendations appear sufficiently complete that users perceive less need for further exploration. The evidence points more strongly to the second mechanism. Treated users spend more time in chat but not less time searching, and timing controls do little to attenuate the funnel effects. By contrast, the assistant's replies become longer, more attribute-rich, and more structured, and content-exposure controls substantially attenuate the effects on search-funnel outcomes. Thus, the reasoning assistant

appears to narrow search less by mechanically displacing search time and more by reducing the perceived marginal value of additional exploration.[5]

## 4.3 Engagement and Persistence

The purchase decline does not mean that users dislike the reasoning assistant. In fact, assignment to the reasoning assistant increases engagement with the chat feature during and after the experiment. Table 5 reports the effects of assignment to the reasoning assistant on chat engagement. During the experimental window, treated users are significantly more likely to initiate a conversation by sending a message. This stronger engagement persists beyond the experiment: as shown in Appendix Table A.18, treated users generate 10.7%, 15.2%, and 8.0% more chat requests in months 1, 2, and 3 after the experiment, respectively, and 11.0% more in cumulative requests over months 1–3 as shown in Table 5. After the experiment, the reasoning assistant became broadly available on the platform, so these post-experiment estimates should be interpreted as persistence from initial exposure rather than continued differential access to the two models. Because travel app use and AI-chat use are episodic, this initial-exposure effect can persist for several months even as the experimental contrast gradually attenuates. The same appendix table also shows that the negative effect on hotel orders attenuates over time, remaining negative in the first two post-experiment months but becoming small and statistically insignificant thereafter. These results reveal a managerial trade-off: the reasoning assistant narrows search activity and reduces hotel purchases in the short run, while also increasing subsequent engagement with the chat feature.

[5]The idea that explanations from AI systems can increase user trust is well-supported in the literature (e.g. Bauer et al., 2023). This interpretation is also consistent with recent work showing that when consumers perceive platform recommendations to be more precise, they may reduce search effort and explore fewer alternatives (Nocke and Rey, 2024; Zhou et al., 2025).

Table 5: Effect of Reasoning AI Assistant on AI Chat Sessions

| | During Experiment | Post-experiment | |
|---|---|---|---|
| | Used Chat | Retained (30 Days) | Total Requests (Months 1–3) |
| | (1) | (2) | (3) |
| Reasoning AI Assistant | 0.079*** | 0.082** | 0.110*** |
| | (0.011) | (0.040) | (0.017) |
| User Demographics | No | No | No |
| Model | Logit | Logit | NB |
| Observations | 497,565 | 497,565 | 497,565 |

*Note:* Column 1 reports the effect on chat use during the experiment. Column 2 reports whether users returned to the chat feature within 30 days after their last request. Column 3 reports the effect on cumulative total chat requests over months 1–3 after the experiment. Columns 1 and 2 are estimated using logistic regression. Column 3 is estimated using a negative binomial (NB) model, with coefficients reported as $\exp(\beta)-1$. Standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## 4.4 Heterogeneity

Appendix Table A.19 reports exploratory heterogeneity analyses by gender, age, device type, and platform tenure. Given the more limited power of subgroup analyses, we interpret these results cautiously. The clearest heterogeneity appears at the purchase stage: the hotel-order effect is more negative for female users, users aged 35 and above, Android users, and experienced users, while the corresponding estimates for male users, younger users, iOS users, and new users are smaller and statistically indistinguishable from zero. By contrast, the search-funnel effects are broader but less uniform across segments, with declines appearing on different margins for different groups.

We conduct two additional exploratory checks. First, to assess whether the stronger hotel-order effect among experienced users reflects habit formation or resistance to change, we examine heterogeneity by whether users had ever used Wendao before their experiment entry time (Appendix Table A.20). This is a more direct measure of prior familiarity with the assistant than platform tenure: as shown in Table 1, about 19.7% of users had used Wendao

before the experiment. The results do not support a simple change-aversion account. Users with prior Wendao experience do not show a uniformly stronger negative response across the funnel, and the hotel-order decline is statistically significant among users without prior Wendao experience. Thus, the main findings are not driven solely by incumbent assistant users reacting to a model change. Second, causal-forest estimates in Appendix C and Figure A.4 show that individual-level CATEs for hotel orders are centered below zero, suggesting that the average negative effect is not driven by a small set of extreme users.

Taken together, these analyses suggest that responses to the reasoning assistant vary across users, especially at the purchase stage. However, because the subgroup patterns are not identical across funnel margins and the estimates are less powered than the main effects, we view them as suggestive evidence of heterogeneous behavioral responses rather than as a basis for a fully specified personalization rule. At a broader design level, the results nevertheless suggest that platforms may need to tune the depth and structure of AI assistance to the user and decision context, for example by offering more concise or expandable responses when recommendations that appear sufficiently complete risk curtailing further exploration.

# 5 Conclusion

This paper studies how more deliberative conversational assistance reshapes consumer search and purchase behavior on digital platforms. Using a large-scale field experiment on a major online travel platform, we show that assignment to a reasoning-based assistant compresses earlier stages of the hotel search funnel: treated users initiate fewer search queries, browse fewer hotel listings, and click fewer hotels for deeper inspection, resulting in lower short-run hotel bookings. At the same time, the reasoning assistant increases engagement with the chat feature during the experiment and in post-experiment usage.

These results suggest that more deliberative AI does not automatically improve commercial outcomes. In our setting, richer and more structured replies appear to reduce consumers'

perceived need for further search, narrowing exploration even as engagement with the assistant rises. More broadly, the paper highlights how AI embedded in the search funnel can reallocate attention across multiple stages of search and thereby affect eventual purchases.

Our findings also point to several important open questions for future research. First, our experiment identifies the effect of deploying a reasoning-based assistant relative to a comparable non-reasoning assistant in this particular platform setting, rather than isolating which specific features of more deliberative assistance drive the observed behavioral changes. A natural next step is to unpack more cleanly which elements of reasoning-style responses, such as explanation, structure, length, or perceived completeness, matter most, and how their effects depend on interface and platform design. Second, the strongest evidence in our study comes from the hotel channel, where browsing and inspection are central to choice. More generally, the portability of these findings remains an open question: the magnitude of the effect may differ across product categories and platform settings, especially where search is more structured or less differentiated. Third, although we examine post-experiment behavior for up to three months, these results do not substitute for a longer experiment with continued randomized exposure and therefore cannot fully identify long-run learning, habit formation, or novelty effects. Future research should therefore examine when more deliberative assistance complements rather than substitutes for consumer search, how these effects evolve under sustained exposure, and what they imply for consumer welfare and platform design.

# Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

# References


Bauer, K., von Zahn, M., Hinz, O. 2023. Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. Information Systems Research. 34 (4), 1582–1602.

Brand, J., Israeli, A., Ngwe, D. 2023. Using llms for market research. Harvard Business School Marketing Unit Working Paper. (No. 23-062).

Dominguez-Olmedo, R., Hardt, M., Mendler-Dünner, C. 2024. Questioning the survey responses of large language models. Advances in Neural Information Processing Systems. 37, 45850–45878.

Gibbard, P. 2022. A model of search with two stages of information acquisition and additive learning. Management Science. 68 (2), 1212–1217.

Goli, A., Singh, A. 2024. Frontiers: Can large language models capture human preferences? Marketing Science. 43 (4), 709–722.

Greminger, R. P. 2022. Optimal search and discovery. Management Science. 68 (5), 3904–3924.

Gui, G., Toubia, O. 2023. The challenge of using LLMs to simulate human behavior: A causal inference perspective. arXiv preprint arXiv:2312.15524.

Guo, D., Yang, D., Zhang, H. et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature. 645 (8081), 633–638.

Honka, E., Hortaçsu, A., Wildenbeest, M. 2019. Empirical search and consideration sets. In Handbook of the Economics of Marketing. 1, 193–257, Elsevier.

Honka, E., Seiler, S., Ursu, R. 2024. Consumer search: What can we learn from pre-purchase data? Journal of Retailing. 100 (1), 114–129.

Kapoor, A., Kumar, M. 2025. Frontiers: Generative ai and personalized video advertisements. Marketing Science. 44 (4), 733–747.

Li, P., Castelo, N., Katona, Z., Sarvary, M. 2024. Frontiers: Determining the validity of large language models for automated perceptual analysis. Marketing Science. 43 (2), 254–266.

Lin, S., Shi, Z. J., Sun, X. 2025. Towards intelligent shopping assistant: How can llm-based chatbots transform online shopping process? Available at SSRN 5088975.

Luo, X., Tong, S., Fang, Z., Qu, Z. 2019. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. Marketing Science. 38 (6), 937–947.

Ni, X., Wang, Y., Feng, T., Lu, L. X., Wang, Y., Zhou, C. 2025. Generative AI in action: Field experimental evidence on worker performance in e-commerce customer service operations. Available at SSRN 5012601.

Nocke, V., Rey, P. 2024. Consumer search, steering, and choice overload. Journal of Political Economy. 132 (5), 1684–1739.

Padilla, N., Lam, H. T., Lambrecht, A., Hollenbeck, B. 2025. The impact of LLM adoption on online user behavior. Available at SSRN 5393256.

Sun, C., Shi, Z., Liu, X., Ghose, A. 2025. The effect of voice ai on digital commerce. Information Systems Research. 36 (2), 1147–1166.

Ursu, R., Seiler, S., Honka, E. 2025. The sequential search model: A framework for empirical research. Quantitative Marketing and Economics. 23 (1), 165–213.

Yan, S., Zhong, H., Zhong, Z., Zhou, W. 2026. Shopping with a platform ai assistant: Who adopts, when in the journey, and what for. arXiv preprint arXiv:2603.24947.

Ye, Z., Yoganarasimhan, H., Zheng, Y. 2025. Lola: LLM-assisted online learning algorithm for content experiments. Marketing Science.

Yuan, Z., Chen, A. Y., Wang, Y., Sun, T. 2025. How recommendation affects customer search: A field experiment. Information Systems Research. 36 (1), 84–106.

Zhang, S., Narayandas, D. 2026. Engaging customers with ai in online chats: Evidence from a randomized field experiment. Management Science. 72 (1), 73–95.

Zheng, S., Zhu, Y., Ye, X., Shen, L. 2025. Generative search: Evidence from a large-scale field experiment. Available at SSRN 5297194.

Zhou, W., Lin, M., Xiao, M., Fang, L. 2025. Higher precision is not always better: Search algorithm and consumer engagement. Management Science. 71 (7), 6204–6226.

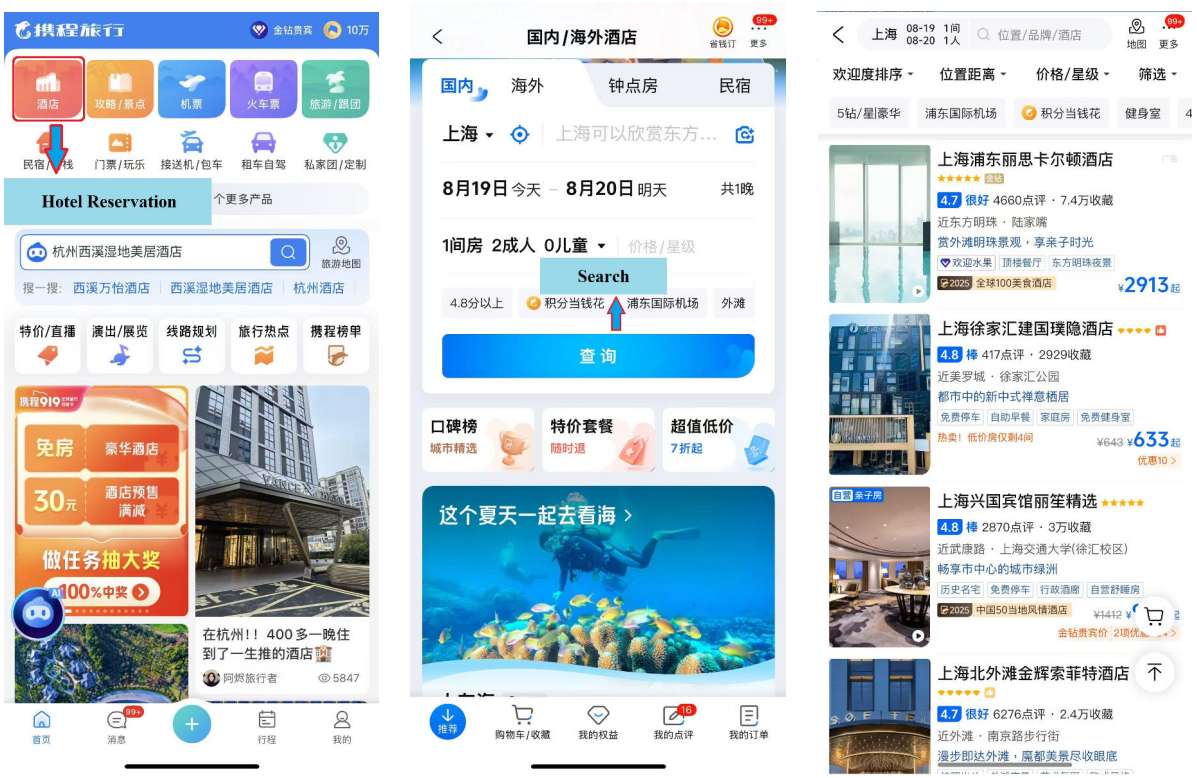
# Online Appendix to "Reasoning AI, Consumer Search, and Purchase: Evidence from an Online Platform Field Experiment"

**Abstract**

This online appendix accompanies the paper "Reasoning AI, Consumer Search, and Purchase: Evidence from an Online Platform Field Experiment" and provides supplementary material. Section A provides additional background on the Ctrip mobile application. Section B presents further empirical results. Section C details the causal forest approach used to estimate the conditional average treatment effect (CATE) in the heterogeneity analysis reported in the main text.

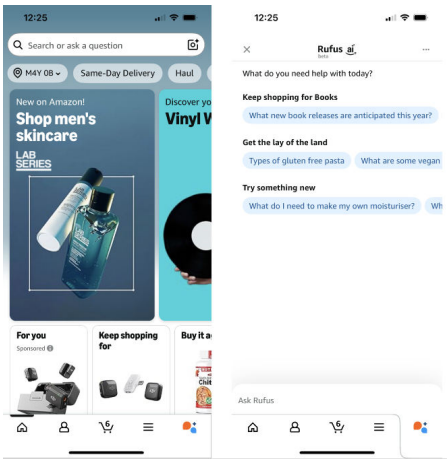
# A Additional Background on the Ctrip App

Figure A.1: Consumer Shopping Process

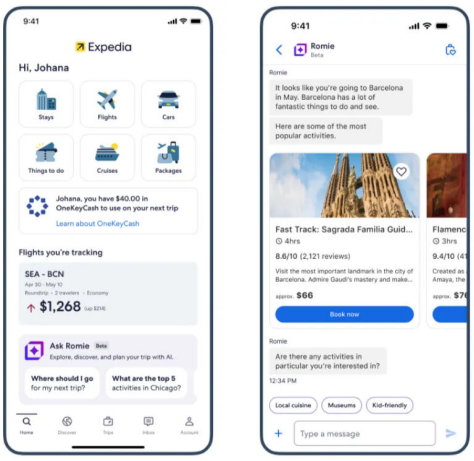



Panel A: Entry Page Panel B: Search Page Panel C: Listing Page

*Note:* This figure illustrates the consumer shopping process on the Ctrip mobile app using hotel reservations as an example. Panel A shows the entry page, where users can either tap the hotel booking button or enter a search query. Panel B displays the search results page, and Panel C presents the product listing page, where available hotels are shown with basic information such as name, rating, location, and price.




(a) Shopping Assistant "Rufus" – Amazon




(b) Travel Assistant "Romie" – Expedia

Figure A.2: Examples: AI Assistants on Large Digital Platforms

# B Additional Results

Table A.1: Illustrative Example of Reasoning vs. Non-Reasoning AI Hotel Recommendations

| **Model** | **Conversation Example** |
|---|---|
| Reasoning | [10:03:12] **Ask**: Can you recommend some New York hotels near Times Square that are great for families?<br>[10:03:15] **Reply**: The user wants family-friendly picks in the Times Square area. I filter for Midtown properties within an easy walk of Times Square or its subway hub, prioritizing properties that offer family suites, connecting rooms, and kid-friendly amenities.<br>[10:03:24] **Here is my recommendation**:<br>1. Tempo by Hilton New York Times Square: on Times Square (steps to the subway); modern rooms with family options (connecting rooms, free cribs/infant beds, mini-fridges); fitness center; around USD 229–290/night.<br>2. Hotel Edison Times Square: 2–4 min walk to 49 St/Times Sq stations; Signature family suites (1 king bed and 2 full-size beds, sleep up to 6); two on-site restaurants with Grab&Go option; fitness center; around USD 170–240/night.<br>3. YOTEL New York Times Square: 10–12 min walk to Times Square; family-friendly amenities (suites with multiple beds, smart TVs, interesting robot luggage handler "Yobot" in the lobby); fitness center; around USD 160–210/night. |
| Non-reasoning | [10:03:12] **Ask**: Can you recommend some New York hotels near Times Square that are great for families?<br>[10:03:14] **Reply**:<br>1. Tempo by Hilton New York Times Square: family options; gym; around USD 229–290/night.<br>2. Hotel Edison Times Square: family rooms; dining; around USD 170–240/night.<br>3. YOTEL New York Times Square: modern family rooms; gym; around USD 160–210/night. |

Table A.2: Robustness Check Alternative Sample Cutoff Thresholds

| | Total Search Queries | Total Browse | Total Clicks | Total Hotel Orders |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Cutoff Threshold 1: Top 2% for Search/Browse/Clicks and Top 0.2% for Orders* | | | | |
| Reasoning AI Assistant | $-0.0064^{**}$ | $-0.0069^{**}$ | $-0.0167^{**}$ | $-0.0214^{**}$ |
| | (0.0030) | (0.0032) | (0.0069) | (0.0095) |
| Model | NB | NB | NB | NB |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 487,726 | 487,726 | 487,726 | 487,726 |
| *Cutoff Threshold 2: Top 1.5% for Search/Browse/Clicks and Top 0.15% for Orders* | | | | |
| Reasoning AI Assistant | $-0.0075^{**}$ | $-0.0084^{***}$ | $-0.0214^{***}$ | $-0.0175^{*}$ |
| | (0.0031) | (0.0032) | (0.0069) | (0.0094) |
| Model | NB | NB | NB | NB |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 491,655 | 491,655 | 491,655 | 491,655 |
| *Cutoff Threshold 3: Top 0.5% for Search/Browse/Clicks and Top 0.05% for Orders* | | | | |
| Reasoning AI Assistant | $-0.0074^{**}$ | $-0.0080^{**}$ | $-0.0207^{***}$ | $-0.0210^{**}$ |
| | (0.0032) | (0.0034) | (0.0068) | (0.0095) |
| Model | NB | NB | NB | NB |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 498,856 | 498,856 | 498,856 | 498,856 |
| *Cutoff Threshold 4: Top 0.1% for Search/Browse/Clicks and Top 0.025% for Orders* | | | | |
| Reasoning AI Assistant | $-0.0081^{**}$ | $-0.0079^{**}$ | $-0.0218^{***}$ | $-0.0215^{**}$ |
| | (0.0033) | (0.0034) | (0.0069) | (0.0094) |
| Model | NB | NB | NB | NB |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 501,334 | 501,334 | 501,334 | 501,334 |

*Note:* This table reports robustness checks using different sample cutoff thresholds. Each panel re-estimates the main specification under a different cutoff rule. All columns are estimated using negative binomial (NB) models, and coefficients are reported as $\exp(\beta)-1$. The specific cutoff values for each user-level outcomes are as follows. Threshold 1: Total Search Queries (109), Total Browse (1,419), Total Clicks (49), and Total Hotel Orders (5). Threshold 2: Total Search Queries (121), Total Browse (1,578), Total Clicks (59), and Total Hotel Orders (6). Threshold 3: Total Search Queries (160), Total Browse (2,098), Total Clicks (61), and Total Hotel Orders (8). Threshold 4: Total Search Queries (177), Total Browse (2,317), Total Clicks (65), and Total Hotel Orders (9). Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.3: Randomization and Balance Check

| Panel A: Randomization Check | | | | |
|---|---|---|---|---|
| Variable | Control | Treated | Difference | P-value |
| N | 248,581 | 248,984 | | |
| *User Demographics* | | | | |
| Female | 0.545 | 0.547 | 0.002 | 0.149 |
| Device: iOS | 0.350 | 0.350 | 0.000 | 0.581 |
| New User | 0.487 | 0.487 | 0.000 | 0.720 |
| VIP User | 0.098 | 0.098 | 0.000 | 0.941 |
| Star User | 0.266 | 0.267 | 0.001 | 0.413 |
| Age Above 35 | 0.622 | 0.622 | 0.000 | 0.911 |
| Pre-exp Chat User | 0.197 | 0.198 | 0.001 | 0.238 |
| *Pre-experiment User Behavior* | | | | |
| Total Search Queries | 11.454 | 11.385 | −0.069 | 0.328 |
| Total Browse | 139.151 | 138.255 | −0.896 | 0.318 |
| Total Clicks | 3.216 | 3.233 | 0.017 | 0.598 |
| Unique Hotels Clicked | 2.317 | 2.323 | 0.006 | 0.800 |
| Click-through-rate (%) | 0.088 | 0.088 | 0.000 | 0.847 |
| Click Conversion Rate (%) | 5.307 | 5.249 | −0.058 | 0.483 |
| Panel B: Trimming Rates (%) by Treatment Groups | | | | |
| Outcome | Control | Treated | Difference | P-value |
| If removed | 3.240 | 3.300 | 0.060 | 0.245 |

*Note:* Panel A reports group means and two-sample t-tests comparing baseline characteristics between the control and treatment groups in our final sample. The pre-experiment period is defined as the interval from April 3, 2025 to each user's A/B test entry time. Panel B reports $t$-tests comparing the share of users removed under our main trimming rule (top 1% of users in total search queries, browse, and clicks, and the top 0.1% in total orders) between the control and treatment groups, which gives us a final regression sample of 497,565 users.

Table A.4: Serial Analysis: Search-to-Order Funnel

| | Total Search | Total Browse | Total Clicks | Total Hotel Orders | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Reasoning AI Assistant | $-0.008^{**}$ | $-0.005^{***}$ | $-0.007$ | $-0.025^{**}$ | $-0.011$ |
| | (0.005) | (0.0002) | (0.007) | (0.010) | (0.009) |
| Total Search | | $0.030^{***}$ | $0.038^{***}$ | | $0.012^{***}$ |
| | | (0.00000) | (0.0003) | | (0.0004) |
| Total Browse | | | $0.003^{***}$ | | $-0.0003^{***}$ |
| | | | (0.00003) | | (0.00003) |
| Total Clicks | | | | | $0.072^{***}$ |
| | | | | | (0.0004) |
| Model | NB | NB | NB | NB | NB |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* This table reports sequential regressions for outcomes along the search-to-purchase funnel. Column 1 reports the treatment effect of assignment to the reasoning AI assistant on total search. Column 2 reports the treatment effect on total browse controlling for total search. Column 3 reports the treatment effect on total clicks controlling for total search and total browse, to examine if the click effect operates through the search funnel. Columns 4 and 5 report the treatment effect on total hotel orders, first without controls and then controlling for total search, total browse, and total clicks, to assess whether the order effect is mediated by upstream search and clicks. All columns are estimated using negative binomial (NB) models, and coefficients are reported as $\exp(\beta)-1$. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.5: Effect of the Reasoning AI Assistant on Search Intensity and Outcomes

| | Avg. Browse per Search | CTR (%) | Conversion (%) |
|---|---|---|---|
| | (1) | (2) | (3) |
| Reasoning AI Assistant | 0.003 | −0.0002 | −0.0002 |
| | (0.027) | (0.0002) | (0.001) |
| Baseline | 9.992 | 0.031 | 0.058 |
| Observations | 338,444 | 338,444 | 172,640 |

*Note:* This table reports the effect of the reasoning AI assistant on search intensity and downstream outcomes within the hotel channel. Column 1 reports the effect on average browse per search. Columns 2 and 3 report effects on click-through rate (CTR) and conversion rate, respectively. All columns are estimated by OLS. Columns 1 and 2 use the subsample of users with at least one browse, while column 3 uses the subsample of users with at least one click. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.6: Effect of Reasoning AI Assistant: Options Offered and User Outcomes

| Panel A: Options Offered by AI Assistant (User Level) | | | | |
|---|---|---|---|---|
| | Has Suggested Hotel | | log (1+ Total Hotels Suggested) | |
| | All Users | Chat Users | All Users | Chat Users |
| | (1) | (2) | (3) | (4) |
| Reasoning AI Assistant | 0.004*** | 0.038*** | 0.005*** | 0.056*** |
| | (0.0004) | (0.005) | (0.0004) | (0.006) |
| User Demographics | No | No | No | No |
| Observations | 497,565 | 34,288 | 497,565 | 34,288 |
| **Panel B: Outcomes for AI-recommended Hotel (Hotel-request Level)** | | | | |
| | Is Clicked | | Is Ordered | |
| | (5) | | (6) | |
| Reasoning AI Assistant | −0.022*** | | −0.001 | |
| | (0.003) | | (0.001) | |
| Observations | 20,923 | | 20,923 | |
| **Panels C & D: Matched Click/Order Outcomes with AI Chat (User Level)** | | | | |
| Panel C: Clicks | | | | |
| | Mentioned | Not Mentioned | log (1+ Total Mentioned) | log (1+ Total Not Mentioned) |
| | (7) | (8) | (9) | (10) |
| Reasoning AI Assistant | −0.00004 | −0.004*** | −0.00005 | −0.009*** |
| | (0.0001) | (0.001) | (0.0001) | (0.003) |
| User Demographics | No | No | No | No |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |
| Panel D: Orders | | | | |
| | Mentioned | Not Mentioned | log (1+ Total Mentioned) | log (1+ Total Not Mentioned) |
| | (11) | (12) | (13) | (14) |
| Reasoning AI Assistant | 0.00004 | −0.001 | 0.00003 | −0.001 |
| | (0.00004) | (0.001) | (0.00003) | (0.001) |
| User Demographics | No | No | No | No |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* Panel A reports the effects of assignment to the reasoning AI Assistant on hotel recommendations received by users. Columns 1 and 2 show the effect on the probability of receiving at least one hotel suggestion, for all users and for chat users, respectively. Columns 3 and 4 report the effect on the number of hotel options, transformed as log(1 + Total Hotels Suggested), for all users and chat users. Panel B uses hotel-request observations and reports the effect of the reasoning AI Assistant on whether a hotel mentioned by the AI is clicked or ordered. Panels C and D report user-level outcomes for hotels mentioned and not mentioned by the AI. Panel C reports click outcomes, including both extensive-margin and log-count measures. Panel D reports order outcomes, including both extensive-margin and log-count measures. Robust standard errors are in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.7: Effect of Reasoning AI Assistant on Hotel Orders by Channel

| | Total Hotel Orders | |
|---|---|---|
| | Hotel Listing Pages | Other Channels |
| | (1) | (2) |
| Reasoning AI Assistant | $-0.0027^{**}$ | 0.0004 |
| | (0.0012) | (0.0006) |
| User Demographics | No | No |
| Observations | 497,565 | 497,565 |

*Note:* The table reports a decomposition of the effect of the reasoning AI assistant on hotel orders by order channel. Column 1 reports the treatment effect on the number of hotel orders originating from hotel listing pages, while Column 2 reports the treatment effect on hotel orders from other channels. In the final sample, users place 63,646 hotel orders in total, of which 51,631 (81.12%) originate from hotel listing pages and 12,015 (18.88%) originate from other channels. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.8: Effect of Reasoning AI Assistant on Flight Orders

| Panel A | All Users | | |
|---|---|---|---|
| | Total Flight Orders | | |
| | (1) | (2) | (3) |
| Reasoning AI Assistant | −0.002 | −0.852 | −0.852 |
| | (0.002) | (1.139) | (0.661) |
| User Demographics | No | No | No |
| Model | OLS | NB | POS |
| Observations | 497,565 | 497,565 | 497,565 |
| **Panel B** | **Users with Any Flight Order** | | |
| | Total Flight Orders | | |
| | (4) | (5) | (6) |
| Reasoning AI Assistant | −0.003 | −0.171 | −0.171 |
| | (0.012) | (0.666) | (0.666) |
| User Demographics | No | No | No |
| Model | OLS | NB | POS |
| Observations | 47,055 | 47,055 | 47,055 |

*Note:* Panel A reports estimates for all users, and Panel B reports estimates for users with any flight order. Total flight orders is the dependent variable in all columns. Columns 1 and 4 report OLS estimates. Column 2 and column 5 report negative binomial (NB) estimates. Column 3 and column 6 report Poisson (POS) estimates. For the NB and POS regressions, coefficients are transformed as $\exp(\beta)-1$. Robust standard errors are in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.9: Effect of Reasoning AI Assistant on Other Orders

| Panel A | All Users | | |
|---|---|---|---|
| | Total Other Orders | | |
| | (1) | (2) | (3) |
| Reasoning AI Assistant | 0.001 | 1.007 | 1.007 |
| | (0.001) | (1.352) | (0.962) |
| User Demographics | No | No | No |
| Model | OLS | NB | POS |
| Observations | 497,565 | 497,565 | 497,565 |
| Panel B | Users with Any Other Order | | |
| | Total Other Orders | | |
| | (4) | (5) | (6) |
| Reasoning AI Assistant | 0.009 | 0.623 | 0.623 |
| | (0.010) | (0.958) | (0.958) |
| User Demographics | No | No | No |
| Model | OLS | NB | POS |
| Observations | 29,837 | 29,837 | 29,837 |

*Note:* Panel A reports estimates for all users, and Panel B reports estimates for users with any other order. Total other orders is the dependent variable in all columns. Columns 1 and 4 report OLS estimates. Column 2 and column 5 report negative binomial (NB) estimates. Column 3 and column 6 report Poisson (POS) estimates. For the NB and POS regressions, coefficients are transformed as $\exp(\beta)-1$. Robust standard errors are in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.10: Effect of Reasoning AI Assistant on Hotel Orders by Order Status

| | Total Orders | |
|---|---|---|
| | Cancelled | Completed |
| | (1) | (2) |
| Reasoning AI Assistant | −0.012 | −0.025*** |
| | (0.016) | (0.009) |
| Model | NB | NB |
| User Demographics | No | No |
| Reported Effect | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ |
| Observations | 497,565 | 497,565 |

*Note:* The table reports the effect of the Reasoning AI Assistant on cancelled and completed hotel orders using negative binomial (NB) regression. Coefficients are reported as $\exp(\beta) - 1$ for interpretation as proportional changes, with delta method robust standard errors in parentheses. Both columns use the full sample of users. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.11: Effect of Reasoning AI Assistant on Total Hotel Orders by Order Amount Tier

| | Total Orders by Total Amount Tier | | | | Total Weighted Order Amount |
|---|---|---|---|---|---|
| | Bottom 20% | 20–50% | 50–80% | Top 20% | Log Transformed |
| | (1) | (2) | (3) | (4) | (5) |
| Reasoning AI Assistant | 0.0001 | $-0.001^{*}$ | $-0.002^{**}$ | $-0.001$ | $-0.007^{**}$ |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.003) |
| Model | NB | NB | NB | NB | OLS |
| User Demographics | No | No | No | No | No |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* Columns 1–4 report the effects of the reasoning AI assistant on the total number of hotel orders within each total amount tier. These tiers are defined by a scaled index for the order's total dollar amount from 0 to 100: $\leq 11$ for the bottom 20%, $(11, 22.5]$ for 20–50%, $(22.5, 52.8]$ for 50–80%, and $> 52.8$ for the top 20%. Columns 1–4 are estimated using negative binomial (NB) models, and coefficients are reported as $\exp(\beta)-1$. To construct the total order weighted value in Column 5 for each user, orders are weighted by the median amount index of the corresponding tier: 5.5, 16.75, 37.65, and 76.4, respectively. Column 5 reports the effect on the log-transformed total weighted order amount using OLS. Robust standard errors are in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.12: Effect of Reasoning AI Assistant on Hotel Orders (User Demographics)

| | All Users | | |
|---|---|---|---|
| | Total Hotel Orders | | |
| | (1) | (2) | (3) |
| Reasoning AI Assistant | $-0.004^{**}$ | $-0.023^{**}$ | $-0.023^{**}$ |
| | (0.002) | (0.009) | (0.009) |
| User Demographics | Yes | Yes | Yes |
| Model | OLS | NB | POS |
| Observations | 446,194 | 446,194 | 446,194 |

*Note:* This table reports the estimated effect of exposure to the reasoning AI Assistant on hotel orders with user demographic controls using users with complete user demographic information. Columns 1 report OLS estimates, Column 2 report negative binomial (NB) estimates, and Column 3 report Poisson (POS) estimates. For the NB and POS regressions, coefficients are transformed as $\exp(\beta)-1$. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.13: Alternative Models: Effect of the Reasoning AI Assistant on Hotel Orders

| | All Users | | Users with Any Hotel Order | |
|---|---|---|---|---|
| | OLS | Poisson | OLS | Poisson |
| | (1) | (2) | (3) | (4) |
| Reasoning AI Assistant | $-0.004^{***}$ | $-0.025^{***}$ | $-0.013^{*}$ | $-0.009^{*}$ |
| | (0.001) | (0.009) | (0.007) | (0.005) |
| User Demographics | No | No | No | No |
| Model | OLS | POS | OLS | POS |
| Reported effect | $\beta$ | $\exp(\beta) - 1$ | $\beta$ | $\exp(\beta) - 1$ |
| Observations | 497,565 | 497,565 | 53,568 | 53,568 |

*Note:* This table reports alternative model specifications for the effect of the reasoning AI assistant on total hotel orders. Left panel reports estimates for all users, and right panel reports estimates for users with any hotel order. Columns 1 and 3 report OLS estimates, with coefficients reported as $\beta$. Columns 2 and 4 report Poisson estimates, with coefficients reported as $\exp(\beta) - 1$. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.14: Poisson Model: Effect of Reasoning AI Assistant on Search, Browse, and Click Outcomes

| | Total Search Queries | Total Browse | Total Clicks |
|---|---|---|---|
| | (1) | (2) | (3) |
| Reasoning AI Assistant | $-0.008^{***}$ | $-0.009^{***}$ | $-0.026^{***}$ |
| | (0.003) | (0.003) | (0.007) |
| Model | POS | POS | POS |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 497,565 | 497,565 | 497,565 |

*Note:* This table reports the estimated effect of assignment to the reasoning AI Assistant on total search queries, total browse, and total clicks. All columns are estimated using Poisson (POS) regressions. Coefficients are transformed as $\exp(\beta)-1$. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.15: Power Test and Minimum Detectable Effect (MDE)

| Statistic | Total Search Queries | Total Browse | Total Clicks | Total Hotel Orders |
|---|---|---|---|---|
| Estimated effect | -0.008 | -0.009 | -0.026 | -0.025 |
| Robust SE | 0.003 | 0.003 | 0.007 | 0.009 |
| P-value | 0.0088 | 0.0045 | 0.0003 | 0.0091 |
| 95% CI | [-0.014, -0.002] | [-0.016, -0.003] | [-0.040, -0.012] | [-0.043, -0.006] |
| Power (%) | 74.50 | 81.12 | 94.91 | 74.13 |
| MDE 80% | -0.009 | -0.009 | -0.020 | -0.027 |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* This table reports the power test and minimum detectable effect (MDE) for the negative binomial estimates for the main user outcomes: total search queries, total browse, total clicks, and total hotel orders. Estimated effect is reported as $\exp(\beta) - 1$, and the 95% confidence interval is constructed on the same scale. Robust SE reports the exponent-adjusted standard error used in the power calculation. MDE 80% is the minimum detectable effect corresponding to 80% power under the reported standard error.

Table A.16: ITT Effects on Chat-time and Content Exposure Mediators

| | Chat Duration | | AI Reply Info | |
|---|---|---|---|---|
| | Waiting Time for AI Answer | Entry to Exit | Reasoning | Answer |
| Panel A: All Users | | | | |
| | (1) | (2) | (3) | (4) |
| Reasoning AI Assistant | 0.099*** | 0.023*** | 0.041*** | 0.010*** |
| | (0.002) | (0.003) | (0.0005) | (0.001) |
| User Demographics | No | No | No | No |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |
| Panel B: Chat Users | | | | |
| | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | 1.270*** | 0.096*** | 0.566*** | 0.006 |
| | (0.017) | (0.018) | (0.004) | (0.008) |
| User Demographics | No | No | No | No |
| Observations | 34,288 | 34,288 | 34,288 | 34,288 |

*Note:* This table reports the estimated effect of exposure to the reasoning AI Assistant on chat timing and information content. Panel A uses all users, and Panel B is restricted to chat users. All columns report OLS estimates using log(1+x) transformed dependent variables. Waiting time for AI Answer measures the time (seconds) from a user's request to the first token of the AI's final answer. Entry to Exit measures total chat duration (seconds). Information content is measured by total information entropy. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.17: ITT on User-Level Chat Attribute Use

| Panel A | Total User Requests | | | | | | |
|---|---|---|---|---|---|---|---|
| | Any Attr. | Private Info | Transport | Time | Place | Person | Hotel List |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Reasoning AI Assistant | 0.003 | 0.003 | 0.001* | 0.003*** | 0.006*** | 0.0003 | 0.001 |
| | (0.002) | (0.002) | (0.0004) | (0.0005) | (0.001) | (0.0003) | (0.0005) |
| Baseline | 0.104 | 0.104 | 0.007 | 0.009 | 0.062 | 0.006 | 0.011 |
| Panel B | Total AI Responses | | | | | | |
| | Any Attr. | Incomplete | Transport | Price | Place | Rating | Room Type |
| | (8) | (9) | (10) | (11) | (12) | (13) | (14) |
| Reasoning AI Assistant | 0.005*** | 0.004*** | 0.012*** | 0.016*** | 0.005*** | 0.001*** | 0.003*** |
| | (0.002) | (0.0003) | (0.001) | (0.001) | (0.002) | (0.0002) | (0.0003) |
| Baseline | 0.078 | 0.010 | 0.038 | 0.029 | 0.075 | 0.003 | 0.005 |
| Model | OLS | OLS | OLS | OLS | OLS | OLS | OLS |
| User Demographics | No | No | No | No | No | No | No |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* This table reports ITT effects of the reasoning AI assistant on the content of user requests and AI replies. Panel A measures the total number of user requests that contain specific attributes, including any attribute, private user information, transportation, time, place, person, or hotel-list requests. Panel B measures the total number of AI replies that contain any attribute, incomplete content, transport, place, price, rating, or room type information. All columns are estimated using OLS. Standard errors are in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.18: Effect of Reasoning AI Assistant on Post-experiment User Activity

| Panel A | Total Chat Requests | | | |
|---|---|---|---|---|
| | Month 1 | Month 2 | Month 3 | Month 1–3 |
| | (1) | (2) | (3) | (4) |
| Reasoning AI Assistant | 0.107*** | 0.152*** | 0.080*** | 0.110*** |
| | (0.025) | (0.030) | (0.026) | (0.017) |
| Panel B | Total Hotel Orders | | | |
| | Month 1 | Month 2 | Month 3 | Month 1–3 |
| | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | −0.013* | −0.014* | −0.001 | −0.009 |
| | (0.007) | (0.008) | (0.008) | (0.006) |
| Model | NB | NB | NB | NB |
| User Demographics | No | No | No | No |
| Reported effect | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ | $\exp(\beta)-1$ |
| Observations | 497,565 | 497,565 | 497,565 | 497,565 |

*Note:* This table reports the effects of the reasoning AI assistant on post-experiment user activity. Panel A reports effects on total chat requests in months 1, 2, and 3 after the experiment, as well as the cumulative total over months 1–3. Panel B reports the corresponding effects on total hotel orders. All specifications are estimated using negative binomial (NB) models, and coefficients are reported as $\exp(\beta)-1$. Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.19: Heterogeneous Treatment Effects of the Reasoning AI Assistant by User Demographics

| | Gender | | Age Group | | Device Type | | User Type | |
|---|---|---|---|---|---|---|---|---|
| | Female | Male | Age $< 35$ | Age $\geq 35$ | iOS | Android | Old User | New User |
| Panel A | | | | Total Search | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | $-0.014^{***}$ | 0.001 | $-0.009^{*}$ | $-0.006$ | $-0.009^{*}$ | $-0.006$ | $-0.001$ | $-0.015^{***}$ |
| | (0.004) | (0.005) | (0.005) | (0.004) | (0.005) | (0.004) | (0.005) | (0.005) |
| Observations | 237,740 | 195,513 | 166,880 | 266,373 | 157,717 | 275,536 | 216,627 | 216,626 |
| Panel B | | | | Total Browse | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | $-0.016^{***}$ | 0.0001 | $-0.009$ | $-0.009^{**}$ | $-0.010^{*}$ | $-0.009^{*}$ | $-0.006$ | $-0.012^{**}$ |
| | (0.004) | (0.006) | (0.006) | (0.004) | (0.005) | (0.004) | (0.005) | (0.005) |
| Observations | 237,740 | 195,513 | 166,880 | 266,373 | 157,717 | 275,536 | 216,627 | 216,626 |
| Panel C | | | | Total Clicks | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | $-0.036^{***}$ | $-0.020^{*}$ | $-0.008$ | $-0.045^{***}$ | $-0.022^{*}$ | $-0.035^{***}$ | $-0.027^{***}$ | $-0.032^{***}$ |
| | (0.009) | (0.012) | (0.012) | (0.009) | (0.012) | (0.009) | (0.010) | (0.010) |
| Observations | 237,740 | 195,513 | 166,880 | 266,373 | 157,717 | 275,536 | 216,627 | 216,626 |
| Panel D | | | | Total Hotel Orders | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Reasoning AI Assistant | $-0.031^{**}$ | $-0.018$ | 0.015 | $-0.047^{***}$ | $-0.013$ | $-0.031^{***}$ | $-0.038^{***}$ | $-0.012$ |
| | (0.013) | (0.014) | (0.016) | (0.012) | (0.016) | (0.012) | (0.013) | (0.014) |
| Observations | 237,740 | 195,513 | 166,880 | 266,373 | 157,717 | 275,536 | 216,627 | 216,626 |
| Model | NB | NB | NB | NB | NB | NB | NB | NB |
| Reported Coefficient | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ |

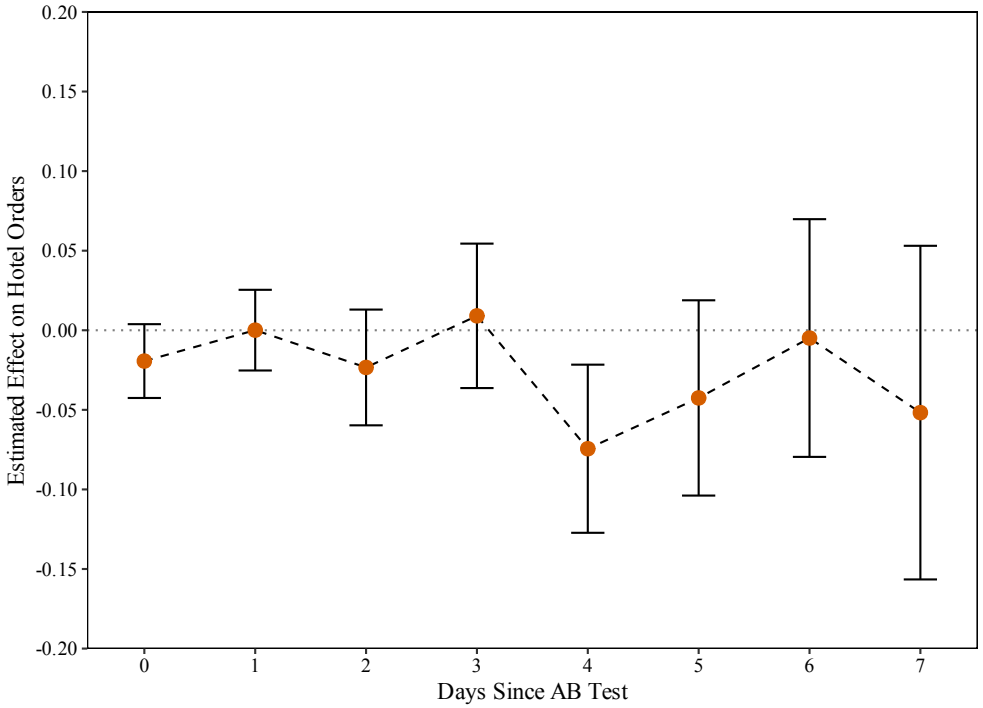
*Note:* The table reports subgroup estimates of the treatment effect of the reasoning AI Assistant by gender, age group, device type, and user type. Panel A reports total search, Panel B reports total browse, Panel C reports total clicks, and Panel D reports total hotel orders. Coefficients in all panels are reported as $\exp(\beta) - 1$. Old and new users are defined by a 50% sample split based on registration date, using September 8, 2023 as the cutoff. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A.20: Effect of Reasoning AI Assistant by Prior AI Chat Usage

| | Used Chat | Total Search Queries | Total Browse | Total Clicks | Total Orders (A/B Period) | Total Orders (1 Month Post) |
|---|---|---|---|---|---|---|
| Panel A: Used Chat Before | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Reasoning AI | 0.114*** | −0.017*** | −0.014** | −0.034** | −0.029 | −0.025 |
| | (0.027) | (0.006) | (0.006) | (0.014) | (0.020) | (0.019) |
| Observations | 98,279 | 98,279 | 98,279 | 98,279 | 98,279 | 98,279 |
| Panel B: Never Used Chat Before | | | | | | |
| | (7) | (8) | (9) | (10) | (11) | (12) |
| Reasoning AI | 0.072*** | −0.006 | −0.008** | −0.024*** | −0.024** | −0.009 |
| | (0.012) | (0.004) | (0.004) | (0.008) | (0.011) | (0.008) |
| Model | Logit | NB | NB | NB | NB | NB |
| Reported | $\beta$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ | $\exp(\beta) - 1$ |
| Observations | 399,286 | 399,286 | 399,286 | 399,286 | 399,286 | 399,286 |

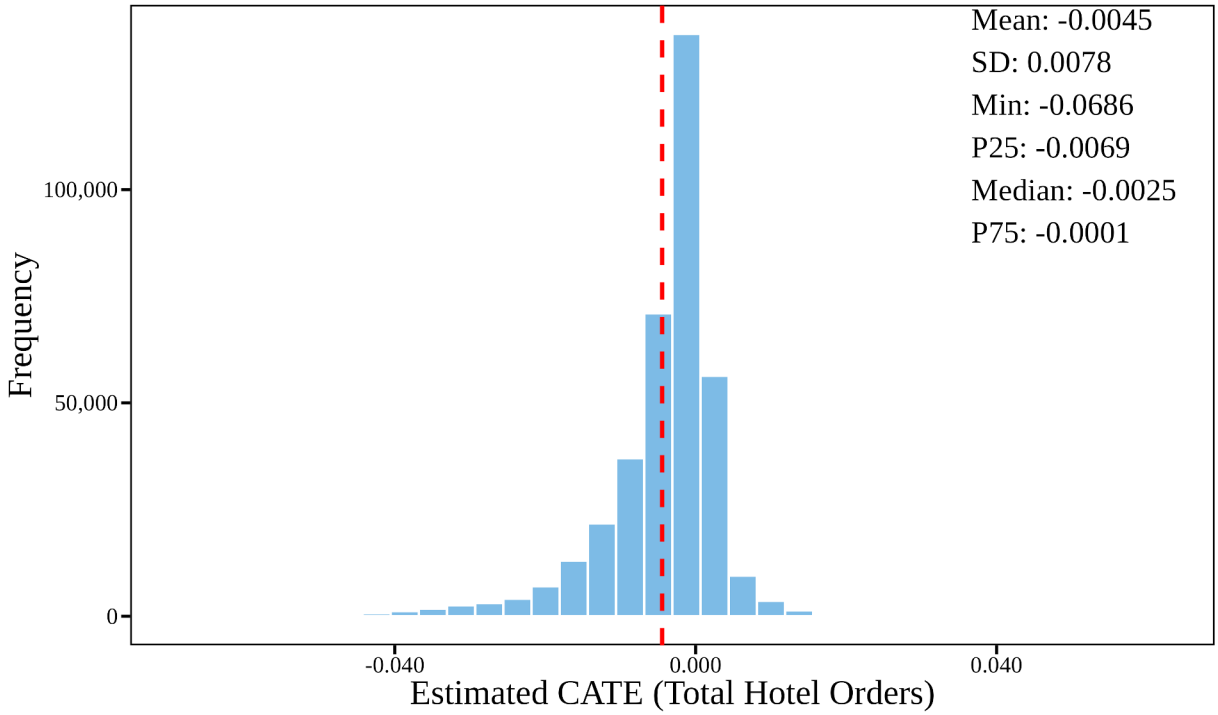
*Note:* This table reports the effect of assignment to the reasoning AI assistant on chat use, search, browse, clicks, and hotel orders among users with different prior AI chat experience. Panel A reports estimates for users who used AI chat before their experiment entry time (N=98,279, 19.75%), and Panel B reports estimates for users who never used AI chat before their entry (N=399,286, 80.25%). Within both groups, treatment and control users are balanced on user demographics and pre-experiment behavior. Columns 1 and 7 report logit coefficients. All other columns report transformed coefficients as $\exp(\beta) - 1$. Robust standard errors are reported in parentheses.

Figure A.3: Dynamic Effects on Total Hotel Orders




*Note:* This figure plots the estimated treatment effect for each day in the experiment, controlling for user demographics. The vertical bars indicate 95% confidence intervals based on robust standard errors.

Figure A.4: Distribution of Individual-Level CATE Estimates for Hotel Orders




*Note*: This figure shows the distribution of Conditional Average Treatment Effect (CATE) estimates from causal forests at the individual level. We estimate CATEs separately for hotel orders using post-treatment user-level data. The x-axis plots estimated CATE values, while the y-axis indicates the number of users within each bin.

# C Estimation of CATEs through Causal Forests

To examine the heterogeneity in treatment effects, we estimate the Conditional Average Treatment Effect (CATE) using the causal forest method (Wager and Athey, 2018; Nie and Wager, 2021). Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for consumer $i$ under treatment and control, with the observed outcome defined as

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0),$$

where $W_i \in \{0, 1\}$ is the treatment indicator for whether the reasoning model is assigned, and $X_i$ are observed covariates capturing consumer characteristics. The CATE is defined as

$$\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i].$$

Causal forests extend random forests to estimate $\tau(X_i)$ by combining three key steps:

First, "local centering" is applied to partial out selection effects. Nuisance models $\hat{m}(X_i) = \mathbb{E}[Y_i \mid X_i]$ and $\hat{e}(X_i) = \mathbb{P}(W_i = 1 \mid X_i)$ are estimated to obtain residuals:

$$\tilde{Y}_i = Y_i - \hat{m}(X_i), \quad \tilde{W}_i = W_i - \hat{e}(X_i).$$

Second, a forest of "honest" causal trees is grown, in which separate subsamples are used for splitting and estimation to avoid overfitting. Each tree partitions the covariate space to maximize heterogeneity in $\tau(X)$.

Finally, treatment effects are estimated using Robinson's transformation within an adaptive kernel defined by the forest weights $\alpha_i(x)$. For any target $x$, the estimated CATE is

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)\, \tilde{W}_i\, \tilde{Y}_i}{\sum_{i=1}^{n} \alpha_i(x)\, \tilde{W}_i^2}.$$

Intuitively, the forest assigns higher weights to consumers $i$ who fall in the same leaves as

the target $x$ across many trees, thereby forming a localized neighborhood in high-dimensional space. This approach yields asymptotically unbiased and normally distributed estimates of $\tau(X)$, even in the presence of high-dimensional covariates and complex treatment heterogeneity.

# References

Nie, X., Wager, S. 2021. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 108 (2), 299–319.

Wager, S., Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association. 113 (523), 1228–1242.