When Search Becomes a Conversation: The Impact of AI Integration on Demand in Online Marketplaces

Xi Nan[1], Ye Liu[2], Mingfeng Lin[2], Yong Tan[1]
[1]Foster School of Business, University of Washington, Seattle, Washington 98195, USA
[2]Scheller College of Business, Georgia Institute of Technology, Atlanta, Georgia 30308, USA

Abstract. As large language models (LLMs) become gateways to digital services, they may reshape how consumers discover sellers and how sellers compete. We study the economic impact of this shift by examining ChatGPT's February 2026 integration with OpenTable, an online restaurant reservation platform. Using daily reservation availability data from January 3 to April 18, 2026, we compare restaurants on OpenTable and Resy, a comparable non-integrated platform, in a difference-in-differences design. We find no significant effect on average demand for OpenTable restaurants, but substantial redistribution across restaurants. Restaurants with more cumulative reviews, typically a positive signal in human-driven discovery, experience demand losses relative to counterparts unaffected by the integration. In contrast, restaurants with textually richer reviews accessible to AI experience larger demand gains. Average ratings do not significantly moderate these effects. We further estimate restaurant-level heterogeneous treatment effects using a doubly robust estimator and find larger effects for restaurants with more distinctive characteristics and dining goals. Evidence on mechanisms reveals changes in both consumer search and restaurant recommendations: consumers express more free-form, intent-rich search criteria when interacting with AI, while AI surfaces restaurants differently from the platform's conventional recommendation algorithms. Our findings suggest that AI-mediated search can change the basis of competition among sellers by reshaping which information guides discovery and how demand is allocated.

Key words: AI integration; AI-mediated search; digital platforms; online marketplaces; recommendation systems; consumer search behavior; online reviews

1. Introduction

Conversational Artificial Intelligence (AI) has rapidly become a general-purpose interface for acquiring information, forming recommendations, and completing everyday tasks. Leading conversational AIs, such as ChatGPT, Claude, and Gemini, have reached hundreds of millions of weekly users, with ChatGPT alone near 900 million by early 2026.[1] Recent consumer surveys indicate that AI is increasingly becoming part of everyday search and decision-making across domains. For example, among consumers who had not yet used AI for shopping, 78% expressed interest in using it to search for product reviews and recommendations;[2] 89% of consumers globally reported wanting to use AI

[1] https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/.

[2] https://www.ibm.com/thought-leadership/institute-business-value/report/ai-retail

in future travel planning;[3] 60% of U.S. adults reported using AI for health or healthcare in the past three months[4] and 44% of Americans plan to use AI more to discover restaurants and book reservations in 2026.[5] On the supply side, these AI assistants are increasingly integrating with third-party applications. OpenAI's Apps framework, introduced in October 2025, allows users to interact and transact with outside services directly within ChatGPT.[6] In early 2026, Anthropic and Google also began developing similar application ecosystems around Claude and Gemini. As a result, such AIs are beginning to function not only as standalone assistant tools, but also as an access layer for digital services: users state their needs in natural language, and the AI searches, recommends, and routes users to integrated applications on their behalf.

As AI assistants become increasingly integrated with transaction platforms, we still know little about how these integrations reshape consumer behavior and the economic outcomes of sellers on those platforms. The change is not merely a new access channel. It could expand demand by reducing search frictions and bringing new consumers to the platform, or reallocate demand across sellers. First, consumer search behavior may change. Conversational search allows consumers to express richer and more contextual intent than keyword queries or categorical filters, which may alter both what information consumers act on and which businesses they ultimately discover. Second, the information-processing layer also changes. Unlike earlier integrations between operating systems and mobile apps, LLM integration does not simply route users to an external service; the model itself interprets consumer intent, retrieves relevant information, and generates recommendations. Because LLMs may weigh and retrieve information differently from consumers and from the platform's own ranking system, they may change which firms enter consumers' consideration sets. Together, these two forces may redistribute demand across firms. We therefore ask three questions. First, does AI integration expand aggregate demand on the integrated platform? Second, does it redistribute demand across firms listed on the platform, and if so, which firms gain or lose? Third, why does any redistribution arise?

We study these questions in the context of online restaurant reservation platforms, exploiting the integration between ChatGPT and OpenTable, a major online restaurant reservation platform. On February 16, 2026, OpenTable became directly accessible inside ChatGPT, so that users could discover and view reservation options through natural-language conversation rather than only through

[3] https://news.booking.com/bookingcom-releases-the-global-ai-sentiment-report/

[4] https://cdn.openai.com/pdf/2cb29276-68cd-4ec6-a5f4-c01c5e7a36e9/OpenAI-AI-as-a-Healthcare-Ally-Jan-2026.pdf

[5] https://www.opentable.com/c/top-restaurants/dining-trends/

[6] https://openai.com/index/introducing-apps-in-chatgpt/.

OpenTable's own interface. We use OpenTable restaurants as the treated group and restaurants on Resy, a comparable restaurant reservation platform that did not undergo a similar integration during our sample window, as the control group. We build a restaurant-day panel of reservation availability for OpenTable and Resy restaurants in New York City around the integration, and measure demand using the share of observed future reservation slots that are unavailable.

To understand which restaurants the AI channel surfaces, we focus on the variation in the information that already exists online about each restaurant, because this pre-existing online content is the raw material an LLM draws on when it interprets a query and forms a recommendation. We capture it with each restaurant's Google reviews. This pre-existing online information spans both the coarse cues that consumers commonly scan, such as a restaurant's review volume, average rating, and the full review text that an LLM can process in its entirety. This distinction maps directly onto our mechanism. If LLM-mediated discovery mainly amplifies conventional popularity or quality signals, restaurants with larger review stocks or higher ratings should benefit more. If instead the integration makes unstructured text more valuable, restaurants with more informative and more distinctive review content should gain more. From Google reviews we construct measures of a restaurant's information stock (review volume), its rating signal, and the textual informativeness of its reviews; and, going beyond these aggregates, we extract the characteristics diners describe and the goals they pursue and group them into embedding-based semantic clusters.

Our research design proceeds in steps that map to the three questions. We first estimate the platform-level effect with a difference-in-differences specification comparing OpenTable and Resy restaurants before and after the integration. We then estimate heterogeneous effects with triple-difference regressions that let the treatment effect vary with restaurant-level moderators, including each dimension of pre-existing online information described above. Finally, we examine evidence on consumer-side and system-side mechanisms separately. On the consumer side, we conduct three analyses. First, we analyze a corpus of realistic user prompts to characterize the preferences and contextual needs expressed in natural-language search. Second, we test whether longer reviews contain information about a broader range of restaurant characteristics and dining goals, and whether this information breadth moderates the treatment effect. Third, we use a doubly robust learner to estimate conditional average treatment effects on the treated (CATT) as a function of each restaurant's semantic profile (Lan et al. 2025), and examine whether these estimated effects are associated with the distinctiveness of its combination of characteristics and dining goals. On the system side, we compare ChatGPT's restaurant recommendations with OpenTable's own search results across three

prompt types: platform-native generic keywords, detailed semantic keywords, and realistic natural-language prompts. We examine both the overlap between the two systems' results and the review characteristics of the restaurants they recommend.

We find that the integration has no detectable effect on OpenTable's aggregate demand. The baseline difference-in-differences estimate is close to zero and statistically insignificant, and equivalence tests reject economically meaningful aggregate effects within our design. However, this null conceals a systematic redistribution that runs against the very cues humans rely on. Restaurants with larger information stocks experience estimated demand losses, whereas restaurants whose reviews are more informative, as proxied by average review length and within-review information breadth, experience estimated demand gains. Ratings do not significantly moderate the treatment effect. These findings indicate that the integration redistributes demand away from restaurants with larger accumulated review stocks and toward those with more informative review text.

To explore the information patterns associated with this redistribution, we construct cross-fitted pseudo-outcomes using a doubly robust learner and examine their association with restaurants' semantic review profiles within the treated sample. We represent each restaurant's review content as a composition across 86 clusters of characteristics and dining goals, and measure its distinctiveness relative to the semantic profiles of other restaurants. Greater semantic distinctiveness is positively associated with the pseudo-outcome, and this association persists after controlling for pre-integration restaurant characteristics. These profiles capture combinations of occasions, atmospheres, dishes, service needs, and constraints that consumers can express in natural language. This pattern is consistent with the possibility that AI-assisted search makes distinctive characteristics and dining goals more discoverable, particularly when they are difficult to express through conventional search categories and filters.

The prompt and recommendation analyses support this interpretation. Realistic user prompts contain multiple free-form search terms spanning several facets, with feature/vibe and contextual needs appearing especially often. These requests are difficult for OpenTable's conventional search interface to match: OpenTable returns at least one result for nearly all generic platform-native prompts, but much less often for detailed semantic prompts and realistic natural-language prompts. AI–platform recommendation overlap is correspondingly highest for generic prompts and substantially lower when prompts depart from OpenTable's maintained vocabulary. At the same time, restaurants surfaced by ChatGPT are more selected on review length than restaurants surfaced by OpenTable. Together, these facts indicate that redistribution arises from the interaction between richer user intent and the AI's ability to interpret machine-readable review content.

Taken together, these results show that embedding an AI into a transaction platform reorganizes rather than expands the market: discovery moves from category-based to intent-based search, and demand flows to restaurants with informative, distinctive review content rather than those distinguished primarily by cues such as higher review counts or higher ratings. The study contributes to a growing body of work on how AI and digital intermediation reshape consumer demand and competition by focusing on the AI–platform integration, and showing that it reallocates demand across competing firms even when it leaves the size of the market unchanged. It also contributes to research on AI-mediated search by estimating the market impact of AI integration on realized demand rather than only on search outputs or recommendations. The findings also carry direct managerial implications. For platforms, the value of an AI integration may lie less in growing total demand than in changing which listings surface, making intent matching and content structure central design problems. This also creates a broader strategic choice for platforms: whether to expose their inventory through external AI interfaces, embed AI-powered search within their own platform, or pursue both channels in parallel. For restaurants, and especially newer entrants, cultivating informative, distinctively positioned review content is a more effective path to being recommended than accumulating generic review volume or chasing higher ratings.

2. Related Literature

2.1. Operations of Online Restaurant Reservation Platforms

A large body of literature studies the operations of online restaurant reservation platforms. Within this literature, one stream examines how reservation platforms operate as digital intermediaries that shape market access, demand generation, and competition in capacity-constrained restaurant markets. Cheyre and Acquisti (2024) study restaurants' adoption of OpenTable and show that reservation platforms reshape restaurants' pricing and competitive incentives. Related work further shows that reservation platforms affect demand through platform curation and cross-channel discovery. Using OpenTable review data, Bang and Jang (2026) find that platform endorsements, such as OpenTable's curated restaurant lists, significantly affect consumers' perceived consumption values, with effects varying by the specificity of the endorsement cue. Lasmi et al. (2025) combine Instagram and OpenTable data and show that social-media attention spills over into OpenTable reservations, suggesting that reservation platforms are embedded in a broader digital discovery funnel. A second stream examines how information disclosed on restaurant reservation and discovery platforms shapes consumers' reservation decisions. Luca (2016) find that a one-star increase in Yelp ratings raises restaurant revenue by 5–9%, with the effect concentrated among independent restaurants, while Anderson and Magruder (2012) use prime-time reservation availability as a revealed measure

of demand and show that an additional half-star in Yelp ratings significantly increases the likelihood that a restaurant sells out. More recently, Hou et al. (2026a) show that displaying restaurant hygiene scores on Yelp unexpectedly reduces demand. Building on these studies, our paper examines AI-powered restaurant discovery as a new upstream layer of digital intermediation that can influence which restaurants consumers consider before making reservations.

Our study contributes to this literature by examining how a new form of digital intermediation, AI-powered restaurant discovery, affects demand in restaurant reservation markets. Unlike conventional reservation platforms, which primarily organize search and booking within a fixed platform interface (Roy et al. 2022), AIs can synthesize dispersed information, generate personalized recommendations, and rank restaurants conversationally before consumers arrive at a reservation platform. This upstream layer of intermediation may therefore influence which restaurants enter consumers' consideration sets, how attention is allocated across competing restaurants, and ultimately how reservation demand is distributed across restaurants.

2.2. The Effect of AI on User Behavior

A growing literature examines how AI is reshaping information acquisition and everyday decision-making. Recent evidence suggests that large language models (LLMs) are increasingly substituting for traditional search and other information intermediaries. Using clickstream data, Padilla et al. (2026) show that LLM adoption substantially reduces traditional search activity, especially for informational queries and visits to smaller or education-related websites. Relatedly, Aral et al. (2026) find that AI search can concentrate users' attention on a narrower set of sources, reducing source diversity and the visibility of long-tail content. As users increasingly obtain answers directly from AI systems, they may also contribute less to the online knowledge ecosystems on which traditional search relies, leading to declines in public knowledge sharing and user-generated contributions (del Rio-Chanona et al. 2024, Burtch et al. 2024). Beyond search, AI is now used across a wide range of everyday and professional tasks (Chatterji et al. 2025), with documented implications for individual creativity (Doshi and Hauser 2024) and labor market outcomes (Liu et al. 2023, Noy and Zhang 2023). Within this broader shift, an emerging stream studies how AI affects consumer demand. Prior work highlights several channels, including algorithmic recommendations that shape product discovery and consideration sets (Li et al. 2022, Lee and Hosanagar 2019), conversational agents that reduce search frictions and support purchase decisions (Hou et al. 2026b, Fang et al. 2025), AI-generated marketing content that changes consumer engagement and sales outcomes (Exner et al. 2025), and chatbots or virtual assistants that automate customer interactions and influence consumer choices (Luo et al. 2019).

Our study contributes to this literature by examining a distinct and increasingly important channel: the integration of AI into a digital platform. We analyze how users change their search behavior when LLM assistance is embedded directly into the platform interface. We find that AI integration enables users to issue more complex and expressive prompts that combine multiple facets of their preferences and semantic requirements, many of which are difficult to articulate through conventional search interfaces. In doing so, our study shows that platform-integrated AI does not merely reduce search costs, but also expands the set of product attributes and preference dimensions that consumers can operationalize during search.

2.3. AI and Search

In recent years, an increasing number of researchers have turned their attention to how AI systems search, recommend, and shape consumer choices. On the computational side, recent work shows that LLM-based recommenders use natural-language interaction to elicit preferences (Huang et al. 2025), construct prompt-based user profiles (Chien et al. 2026), recover multiple preference dimensions from interaction histories (Wang et al. 2026), and plan recommendations under objectives such as relevance, diversity, satisfaction, and engagement (Raj et al. 2026). Related work on AI search emphasizes task-oriented retrieval: agents must retrieve evidence useful for downstream decisions (Liu et al. 2026), filter distracting context (Xue et al. 2026), and account for source authority (Lee et al. 2026). A parallel information-systems and marketing stream studies the resulting consumer and market outcomes. Allouah et al. (2026) show that AI shopping agents produce model-specific market shares and respond to product position, reviews, ratings, and platform endorsements; Ursu et al. (2026) show that AI-assisted product search presents products in a selective "best-for" format that affects choice quality; and Kaiser and Schulze (2026) show that ChatGPT referrals remain small but form a distinct e-commerce intermediation channel. Together, this literature suggests that AI-mediated search and recommendation can reshape not only how consumers acquire information, but also how attention and demand are allocated across firms.

Our study contributes to this emerging literature by moving from AI search outputs and agent behavior to realized market outcomes on a transaction platform. Where prior work primarily provides descriptive evidence on how AI agents rank, recommend, cite, or refer products and sources, we estimate the causal impact of platform-level AI integration on actual consumer demand across competing firms.

3. Research Context, Data, and Variables

3.1. Research Context

In October 2025, OpenAI launched its Apps feature, allowing users to access products, services, tools, and content directly from within the chat.[7] This integration makes the discovery and consumption of platform offerings more seamless. Rather than returning text results without such integration, when an integration is enabled for an App, the chat returns direct links that follow a web design similar to the platform's own. Figure 1 illustrates an example before and after the integration. Recognizing the large volume of traffic from AI chat, many platforms have adopted this SDK, such as Booking.com, Canva, Expedia, Spotify, and Zillow.



(a) Before integration (b) After integration

Figure 1 ChatGPT interface before and after the OpenAI Apps integration. Following the integration, users can directly access platform-specific content and booking interfaces within the chat environment rather than relying solely on text-based recommendations.

To investigate the impact of AI integration on digital platforms, we focus on two online restaurant reservation platforms: OpenTable, which is the largest online restaurant reservation platform and integrated with ChatGPT on February 16, 2026,[8] and Resy, the second-largest online restaurant reservation platform which did not undergo a similar integration during that period.[9] Following prior studies that exploit platform-level design changes as quasi-natural experiments (Jiang et al. 2023, Wang et al. 2022), we identify the effect of AI integration using the platform-level integration with ChatGPT. Even though Resy's overall market share is smaller than OpenTable's, it holds a

[7] https://openai.com/index/introducing-apps-in-chatgpt/

[8] The integration was announced on October 26, 2025. For details, see https://www.linkedin.com/posts/sebastiansanjuan_futureofdining-opentable-ai-share-7407849762848325632-g-Ti

[9] Resy started the integration with Claude on April 23, 2026. https://blog.resy.com/newsroom/resy-restaurants-claude-anthropic

comparable share in major cities, such as New York City.[10] Therefore, we focus our analysis on the online restaurant reservation market in New York City. This focus on a single city also reduces geographic heterogeneity and is consistent with prior literature on restaurant reservation platforms (Cheyre and Acquisti 2024).

3.2. Data

3.2.1. Restaurant Reservation Platforms For each platform, we first identify the restaurants through a city-level search, then supplement the restaurant list with the records from the New York City Department of Health and Mental Hygiene (DOHMH) registry[11]. In total, we identify 1,484 restaurants bookable on OpenTable (treatment group), and 1,882 restaurants on Resy (control group). We drop the 313 restaurants that multi-home on both platforms to keep the treatment and control groups cleanly separated, resulting in 1,171 restaurants in the treatment group and 1,569 in the control group.

Next, we collect daily restaurant reservation availability from January 3 to April 18, 2026, through web scraping. Following Anderson and Magruder (2012), we use reservation availability as a proxy for restaurant demand. We exclude the period from February 2 to February 14 to avoid demand spikes associated with holidays, resulting in 31 days of pre-treatment and 60 days of post-treatment observations. Each day, we collect reservation availability twice, at 12:00 PM ET and 9:00 PM ET, and aggregate the two snapshots to construct our daily measure. In each web-scraping session, we record the availability of every 15-minute time slot between 5:00 PM and 9:00 PM for a party of two over the subsequent seven days. For example, at 12:00 PM ET and 9:00 PM ET on January 3, 2026, we set the party size to two and record whether each time slot, including 5:00, 5:15, 5:30, ..., and 9:00 PM, is available for reservations on January 4, January 5, ..., through January 10. We then construct our outcome measure, the demanded slot proportion, Y_{it}, for restaurant i on date t, by computing the share of unavailable time slots, measured at 15-minute intervals, among all observed slots over the following seven days. With this construction, a higher value of Y_{it} indicates that more time slots are unavailable, reflecting higher demand.

Although the demanded slot proportion provides a comparable measure of reservation availability across restaurants, it may introduce measurement error. First, we do not observe the exact inventory initially offered for each reservation time. Treating each time slot as either available or unavailable therefore abstracts from differences in underlying inventory: an unavailable slot may reflect strong

[10] https://ny.eater.com/2023/2/1/23475584/how-resy-won-reservation-wars-opentable

[11] https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/about_data

demand, limited initial capacity, or both. This concern is partially mitigated by restaurant fixed effects, as restaurants' baseline reservation capacity is likely to be relatively stable over time. Thus, time-invariant differences in inventory across restaurants are absorbed by the fixed effects. Second, the measure cannot fully capture demand at either boundary. When the demanded slot proportion equals one, all observed slots are unavailable, and we cannot observe demand beyond full booking. When it equals zero, all observed slots remain available, even though some reservations may already have been made. To assess sensitivity to these boundary observations, we exclude restaurants with extreme pre-integration mean demand and report the results in Section 5.2. Third, the measure may be sensitive to the temporal granularity of reservation slots. As a robustness check, we repeat the analysis after aggregating the demanded slot proportion to 30-minute intervals, with results reported in Section 5.3. Fourth, reservations may be made more than a week in advance, so availability observed after the integration may partly reflect bookings made before it, particularly in the days immediately following the integration. To assess sensitivity to this timing mismatch, we estimate a donut-hole difference-in-differences specification that excludes observations around the integration date and report the results in Appendix B.

We define $Treat_i$ as an indicator equal to one for restaurants bookable through OpenTable and zero for restaurants on Resy. $Post_t$ is an indicator equal to one for dates after the ChatGPT–OpenTable integration on February 16, 2026, and zero otherwise. Our panel comprises 235,156 restaurant-day observations. Panel A of Table 1 reports summary statistics for the outcome and design variables, and Appendix A reports the corresponding summary statistics separately for OpenTable and Resy restaurants.

3.2.2. Google Reviews To capture heterogeneous effects across restaurants, particularly given differences in existing information online, we collect detailed Google reviews of each restaurant. Google Reviews is one of the largest and most widely used online review sources for restaurants,[12] allowing us to measure the pre-existing online information of these restaurants in a consistent way across platforms. The use of Google Reviews, rather than reviews within each platform, allows us to measure pre-existing online information of restaurants in both the treatment and control groups consistently. Appendix F examines whether pre-integration differences between Google and platform-displayed ratings and review counts moderate the treatment effect. Neither interaction is statistically significant, which provides additional support for using Google reviews to measure restaurants' pre-existing online information consistently across platforms. Specifically, for each restaurant, we collect

[12] https://www.reviewtrackers.com/blog/restaurant-review-sites/

all existing Google reviews till the end of 2025. We choose this cutoff to ensure that the reviews are generated before our study period and incorporated into the model's knowledge. For each review, we collect the review text, author information, including whether the author is labeled as a local guide, the review timestamp, and the star rating.

Based on the reviews of each restaurant, we first measure three dimensions of pre-existing online information.

- **Information stock**: This dimension captures the amount of information accumulated about a restaurant before the integration. Our main measure is $Google_Review_Count_i$, the cumulative number of Google reviews, which proxies for the size of the public information stock available to consumers and AI.[13]
- **Rating**: This dimension captures the overall evaluative signal available about a restaurant before the integration. We measure rating using the average Google review rating, $Rating_i$,[14] which provides a salient quality cue commonly used by consumers, recommendation systems, and ranking algorithms.
- **Textual informativeness**: This dimension captures the amount of textual information contained in existing reviews. For the heterogeneity specifications below, we proxy for textual informativeness using the average length of restaurant i's reviews that contain text. Specifically, $Avg_Review_Length_i$ is defined as the average review length, in characters,[15] across all of restaurant i's non-empty Google reviews.

To gain insights beyond these aggregate measures and examine the role of semantic content, we construct two vector representations ("embeddings") for each restaurant based on the textual content of its Google reviews. We focus on two aspects of the semantic meaning of review content, motivated by two common patterns in how consumers interact with AI when describing their needs for restaurant reservations. First, consumers may provide granular descriptions of desired restaurant attributes, such as cuisine, ambience, service, price level, or location. Second, consumers may describe their needs in a more contextual and goal-oriented way, such as seeking a restaurant for a date night, a business dinner, a family gathering, or a casual meal with friends. Accordingly, we construct a characteristics-based embedding that describes the attribute-based characteristics of a restaurant, and a goal-based

[13] In Appendix C, we also examine two alternative proxies: restaurant tenure, $Tenure_i$, measured as the number of months since the earliest observed Google review, and $Local_Guide_Review_Count_i$, the number of reviews contributed by local guides. These appendix analyses show that the information-stock pattern is robust to alternative measures.

[14] The average rating calculated from Google reviews is exactly the same as the rating displayed on Google.

[15] The number of characters and the number of words are highly correlated ($r = 0.9979$, $p < 0.01$).

embedding that describes the context- or goal-based occasions for which the restaurant may be suitable. We describe the steps of constructing both embeddings in detail below.

- **Step 1. Construct a corpus of reviews.** For each restaurant, we retain up to its 400 most recent reviews posted in 2025. For restaurants with fewer than 400 reviews in 2025, we instead collect the 50 most recent reviews available regardless of posting date.
- **Step 2. Extract a set of characteristics and goals.** We randomly sample 15% of the reviews collected in Step 1 and use gpt-5-mini (Singh et al. 2025) to extract candidate restaurant characteristics and consumer goals. Each prompt contains up to 100 reviews. After processing all sampled reviews, we manually review, consolidate, and refine the extracted labels by removing categories that are excessively rare, overly granular, or too broad. This procedure yields 243 characteristic labels and 506 goal labels.
- **Step 3. Cluster characteristics and goals.** To reduce the dimensionality of the characteristic- and goal-based embeddings, we first embed each characteristic and goal label using text-embedding-3-small[16]. We then perform agglomerative clustering with cosine distance and average linkage. Following a data-driven approach, we select distance thresholds corresponding to the largest jump in the merge-height curve (0.599 for characteristics and 0.569 for goals), resulting in 29 characteristic clusters and 57 goal clusters.
- **Step 4. Label reviews.** We apply the finalized characteristic and goal embeddings to all retained reviews, recording the presence of each raw label at the review level. These labels are subsequently mapped to their corresponding semantic clusters, allowing each review to be represented both by detailed semantic labels and by cluster measures.
- **Step 5. Aggregate to the restaurant level and create the restaurant-level embeddings.** Finally, we aggregate the review-level labels and clusters to construct restaurant-level measures of semantic content. These measures serve as the high-dimensional moderators in the heterogeneous treatment effect analyses presented in Section 6.1.3.

We report summary statistics for the measures constructed from Google Reviews in Panel B of Table 1.

4. Empirical Analysis

4.1. Main Results

We first ask whether OpenTable restaurants experienced an overall change in demand relative to Resy restaurants following ChatGPT's integration with OpenTable. We estimate the following two-way fixed effects difference-in-differences (DiD) (Bertrand et al. 2004) specification:

$$Y_{it} = \beta_0 + \beta_1 \left(Treat_i \times Post_t \right) + \alpha_i + \delta_t + \varepsilon_{it}, \tag{1}$$

[16] https://developers.openai.com/api/docs/guides/embeddings

Table 1 Summary Statistics for the Final Clean Estimation Sample (Pooled OpenTable and Resy)

Variable	N	Mean	SD	Min	P25	Median	P75	Max
Panel A. Outcome and design variables (restaurant–day level)								
Y_{it} (demanded slot proportion)	235,156	0.294	0.320	0.000	0.008	0.160	0.479	1.000
$Treat_i$	235,156	0.409	0.492	0.000	0.000	0.000	1.000	1.000
$Post_t$	235,156	0.646	0.478	0.000	0.000	1.000	1.000	1.000
Panel B. Pre-existing online information (restaurant level)								
$Google_Review_Count_i$	2,740	952.679	1,394.146	4.000	211.000	544.000	1,119.750	24,324.000
$Rating_i$ (Google)	2,740	4.480	0.249	2.400	4.300	4.500	4.600	5.000
$Avg_Review_Length_i$	2,740	231.647	75.242	66.800	178.680	214.809	267.482	888.556
$Labels_per_Review_i$	2,738	3.861	0.983	0.000	3.240	3.740	4.386	8.571
$Clusters_per_Review_i$	2,738	3.085	0.625	0.000	2.710	3.060	3.440	6.222

where Y_{it} denotes the demand for restaurant i on day t. Restaurant fixed effects α_i absorb time-invariant differences across restaurants, including cuisine type, location, baseline popularity, reservation inventory, persistent platform-level differences, and other pre-existing differences among restaurants. Date fixed effects δ_t absorb common shocks affecting all restaurants on a given date, such as day-of-week patterns, holidays, seasonality, and citywide demand fluctuations. The coefficient of interest, β_1, captures the differential change in demand for OpenTable restaurants relative to Resy restaurants after the integration, that is, the average treatment effect on the treated (ATT). Because time-varying platform observables may themselves be affected by the ChatGPT–OpenTable integration, including them as controls could block part of the mechanism or introduce post-treatment bias (Tafti and Shmueli 2020, Cinelli et al. 2024), a concern commonly referred to as the bad-control problem (Angrist and Pischke 2009, ch. 3, pp. 47–49). Following established practice in two-way fixed-effects designs (Braghieri et al. 2022), our primary specification therefore does not condition on potentially endogenous time-varying controls.[17]

Table 2 The Impact of LLM–Platform Integration on Overall Restaurant Demand

	(1)
Treat × Post	−0.0003
	(0.0037)
Restaurant FE	Yes
Date FE	Yes
Observations	235,156
R^2	0.9290

Notes: Difference-in-differences estimate of Equation (1) with restaurant and date fixed effects. Standard errors clustered at the restaurant level in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

[17] Appendix E reports robustness specifications that add platform-displayed rating and platform review count as controls; the results remain qualitatively consistent.

Table 2 reports the baseline specification. We do not find a statistically significant effect of the integration on aggregate restaurant demand: the coefficient on $Treat \times Post$ is close to zero and statistically insignificant. To distinguish this null finding from a merely imprecise estimate, Section 5.1 reports an equivalence test showing that the aggregate effect is statistically equivalent to zero within economically meaningful bounds. We also examine the parallel-trends assumption using an event-study specification. As shown in Figure 2, the pre-integration coefficients are close to zero and statistically insignificant, supporting the validity of the difference-in-differences design.

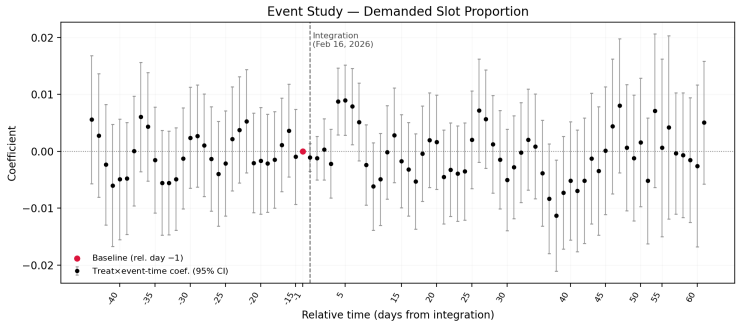


Figure 2 Event study of the integration's effect on demand

Notes: This figure reports the estimated event-time coefficients from the following specification: $Y_{it} = \beta_0 + \sum_{k \neq -1} \beta_k \left(Treat_i \times \mathbf{1}\{t - t_0 = k\} \right) + \alpha_i + \delta_t + \varepsilon_{it}$, where t_0 denotes the February 16, 2026 integration date. The omitted baseline is $k = -1$, the day before the integration. The figure reports 95% confidence intervals based on restaurant-clustered standard errors. The pre-integration coefficients are flat and statistically insignificant, supporting the parallel-trends assumption behind our difference-in-differences estimates.

A likely explanation of the insignificant effect is that the integration primarily reshapes search and discovery among OpenTable's existing users, rather than inducing users to switch across reservation platforms, at least over a relatively short time horizon such as the two-month post-integration period. This interpretation is consistent with prior literature showing that switching costs across platforms can be substantial (Chen and Hitt 2002, Klemperer 1987).

4.2. Heterogeneity Across Review Cues

We next examine heterogeneity in the effects of AI integration by testing whether the treatment effect varies across restaurants with different attributes. To do so, we estimate triple-difference regressions that allow the treatment effect to vary with a restaurant-level moderator Z_i, specified as in Equation (2).

$$Y_{it} = \beta_0 + \beta_1 \left(Treat_i \times Post_t \right) + \beta_2 \left(Post_t \times Z_i \right) + \beta_3 \left(Treat_i \times Post_t \times Z_i \right) + \alpha_i + \delta_t + \varepsilon_{it}. \tag{2}$$

The coefficient β_3 measures how the treatment effect varies with Z_i. Here, we examine three dimensions of pre-existing online information: information stock, measured by $\log(1 + Google_Review_Count_i)$; rating signal, measured by $Rating_i$; and textual informativeness, proxied by $\log(Avg_Review_Length_i)$. Review count and rating provide salient discoverability cues that consumers can readily observe and act on. Average review length serves as a proxy for the amount of textual information available for an AI model to extract, summarize, and match to user intent. The results using either continuous moderators or binary indicators constructed by discretizing each moderator are reported in Table 3.

Table 3 Heterogeneous Effects by Review Cues and Textual Content

	Continuous Moderators			Median Splits		
	(1) Google Review Count	(2) Rating	(3) Review Length	(4) Google Review Count	(5) Rating	(6) Review Length
Treat × Post	0.0373*	0.0227	−0.1611**	0.0095	−0.0025	−0.0078
	(0.0192)	(0.0653)	(0.0706)	(0.0058)	(0.0048)	(0.0051)
× log(1 + *Google_Review_Count*$_i$)	−0.0061**					
	(0.0029)					
× *Rating*$_i$		−0.0051				
		(0.0146)				
× log(*Avg_Review_Length*$_i$)			0.0298**			
			(0.0131)			
× *High Google Review Count*$_i$				−0.0186**		
				(0.0075)		
× *High Rating*$_i$					0.0057	
					(0.0077)	
× *High Review Length*$_i$						0.0158**
						(0.0074)
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	235,156	235,156	235,156	235,156	235,156	235,156
R^2	0.9291	0.9290	0.9291	0.9291	0.9290	0.9291

Notes: Triple-difference estimates of Equation (2). Columns (1)–(3) use continuous moderators: the log of one plus Google review count, Google rating, and the log of average review length, respectively. Columns (4)–(6) replace each moderator with an indicator equal to one if restaurant i lies above the cross-restaurant median of that moderator and zero otherwise. Each column is a separate regression. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

From Equation 2, the ATT at a given moderator level z is $\beta_1 + \beta_3 z$. Thus, a negative β_3 means that the relative post-integration effect for OpenTable declines as Z_i increases, while a positive β_3 means that it rises with Z_i; whether the implied effect is positive or negative at a given z depends on the sign of $\beta_1 + \beta_3 z$. Column (1) shows that the estimated treatment effect decreases with accumulated review volume, implying demand gains for restaurants with sufficiently small review stocks and demand losses for those with sufficiently large review stocks. In this specification, the estimated effect is

positive for restaurants with fewer than roughly 443 Google reviews and negative for restaurants with more than roughly 443 Google reviews. It is also consistent with the idea that AI creates more value when consumers have less pre-existing structured information to rely on. The same pattern holds for the two other information-stock cues reported in Appendix C: accumulated tenure and credible local-guide reviews, both of which capture information sources on which consumers traditionally relied. The rating signal, by contrast, does not moderate the effect, as shown in column (2), despite the established role of ratings in demand (Luca 2016, Chevalier and Mayzlin 2006).

The textual informativeness moderator behaves differently. Column (3) of Table 3 shows that $\log(Avg_Review_Length_i)$ carries a positive triple-difference coefficient of 0.0298 ($p < 0.05$). In this specification, the estimated effect is negative for restaurants with average review length below roughly 222 characters and positive for restaurants with average review length above roughly 222 characters. Thus, restaurants with relatively short reviews experience relative losses, whereas those with sufficiently long reviews experience relative gains. Unlike ratings and information-stock measures, which are highly visible cues that consumers can easily process, detailed review text contains more information that is more difficult for humans to fully digest. Because LLMs are well suited to extracting and synthesizing information from unstructured text, the integration may increase the value of textual content while reducing reliance on simpler cues. We examine this mechanism in greater detail in Section 6.1.

Because several moderators enter the continuous specifications in logs and have long right tails, we also check whether the results depend on using the full cardinal variation in these variables. The specifications in the last three columns of Table 3 therefore replace each moderator Z_i with an indicator for whether restaurant i lies above the cross-restaurant median of that moderator, so that β_3 compares the treatment effect for restaurants in the upper half of the distribution with those in the lower half. The estimates preserve the main qualitative pattern from Table 3: restaurants above the median in Google review volume have a treatment effect 0.0186 lower than those at or below the median ($p < 0.05$), whereas restaurants above the median in review length have a treatment effect 0.0158 higher than those at or below the median ($p < 0.05$). The rating split remains statistically insignificant. Appendix D jointly includes review-count quintile indicators and average review length. The estimates show the same pattern.

5. Robustness Tests

We conduct several robustness tests to better support our outcome variable and our interpretation of the effects.

5.1. Equivalence Test for the Null Aggregate Effect

First, because a statistically insignificant estimate does not by itself establish a null effect, we report an equivalence test using the two one-sided tests procedure. The null hypothesis is that the aggregate effect is outside a prespecified equivalence region, $H_0 : \beta \leq -\Delta$ or $\beta \geq \Delta$, while the alternative is that the effect lies inside that region, $H_1 : -\Delta < \beta < \Delta$. The tightest rejectable equivalence margin is 0.0064, corresponding to 2.16% of mean demand. Following Balafoutas et al. (2026), we set Δ to the design's minimum detectable effect size at 80% power and $\alpha = 0.05$; under this bound, the estimate is statistically equivalent to zero (Table 4). For scale, Cohen's conventional small-effect benchmark of 0.20 standard deviations would imply a much wider bound of about 0.064, so our rejected margin is far tighter than a conventional small effect. Thus, the aggregate result is not only imprecisely estimated around zero; it is bounded tightly enough to rule out economically meaningful aggregate demand expansion in our design.

Table 4 Equivalence (TOST) tests for the aggregate DiD effect on demand.

	Baseline
$\hat{\beta}$ (Treat×Post)	−0.00027
Standard error	0.00370
p-value (H_0: $\beta = 0$)	0.942
90% confidence interval	[−0.0064, 0.0058]
Δ_{min} (tightest rejectable, abs.)	0.0064
as % of mean demand	2.16%
in standard-deviation units	0.020
Δ = MDES (power 0.8; Balafoutas et al. 2026)	0.0104
TOST p-value	0.003
equivalent at $\alpha = 0.05$?	Yes

5.2. Robustness to Censoring the Demand Measure

Because our booking-demand measure is bounded, censoring may arise at both ends. When all booking slots are available, observed availability does not necessarily imply zero latent demand. Similarly, when all slots are unavailable, the measure may fail to reveal the maximum latent booking demand. We therefore examine whether censoring at the lower or upper bound of the demand measure drives the main results by trimming restaurants with extreme pre-period mean demand. The first rule drops restaurants with $\bar{Y}_i^{pre} \notin [0.01, 0.99]$; the second drops restaurants in the bottom and top 5 percent of the pre-period mean-demand distribution. Table 5 and Table 6 report the results. The baseline DiD estimate remains small and statistically insignificant, while the heterogeneity patterns are broadly preserved.

Table 5 Censoring Robustness: Baseline Demand Effects

	(1) Keep $\bar{Y}_i^{pre} \in [0.01, 0.99]$	(2) Drop top and bottom 5%
Treat × Post	−0.0003	0.0033
	(0.0047)	(0.0038)
Restaurant FE	Yes	Yes
Date FE	Yes	Yes
Observations	180,649	217,145
R^2	0.8955	0.9103

Notes: Difference-in-differences estimates of Equation (1) re-estimated after trimming restaurants with extreme pre-period mean demand $\bar{Y}_i^{pre}$. Column (1) retains restaurants with $\bar{Y}_i^{pre} \in [0.01, 0.99]$; column (2) drops the top and bottom 5% of restaurants ranked by $\bar{Y}_i^{pre}$. The untrimmed baseline is reported in Table 2. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 6 Censoring Robustness: Heterogeneous Effects by Review Cues

	Keep $\bar{Y}_i^{pre} \in [0.01, 0.99]$			Drop top and bottom 5%		
	(1) Review Count	(2) Rating	(3) Review Length	(4) Review Count	(5) Rating	(6) Review Length
Treat × Post	0.0380	0.0211	−0.1930**	0.0464**	0.0108	−0.1583**
	(0.0244)	(0.0825)	(0.0913)	(0.0203)	(0.0683)	(0.0748)
× log(1 + *Google_Review_Count*$_i$)	−0.0062*			−0.0070**		
	(0.0038)			(0.0031)		
× *Rating*$_i$		−0.0048			−0.0017	
		(0.0183)			(0.0152)	
× log(*Avg_Review_Length*$_i$)			0.0357**			0.0300**
			(0.0169)			(0.0140)
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	180,649	180,649	180,649	217,145	217,145	217,145
R^2	0.8956	0.8955	0.8956	0.9103	0.9103	0.9103

Notes: Triple-difference estimates of Equation (2) re-estimated under each censoring rule; the untrimmed baseline is reported in Table 3. Columns (1)–(3) retain restaurants with pre-period mean demand $\bar{Y}_i^{pre} \in [0.01, 0.99]$; columns (4)–(6) drop the top and bottom 5% of restaurants ranked by $\bar{Y}_i^{pre}$. Each column is a separate regression whose moderator Z_i corresponds to the column title; Review Count refers to Google review count. Rows beginning with × report interactions of the indicated moderator with Treat × Post. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

5.3. Reservation-Slot Aggregation

Second, we examine whether the results depend on measuring reservation demand at the finest 15-minute grid. We rebuild the same demand measure on a coarser 30-minute grid, counting a half-hour bin as available if either of its two constituent 15-minute slots is open. The aggregate DiD effect remains economically small and statistically insignificant and the heterogeneity pattern is preserved as

shown in Table 7: the information-stock interaction remains negative, the rating interaction remains null, and the review-length interaction remains positive.

Table 7 Baseline and Heterogeneous Effects Under 30-Minute Slot Aggregation

	(1) Baseline	(2) Google Review Count	(3) Rating	(4) Review Length
Treat × Post	−0.0015	0.0416**	0.0005	−0.1528**
	(0.0037)	(0.0194)	(0.0686)	(0.0721)
$\times \log(1 + Google_Review_Count_i)$		−0.0070**		
		(0.0030)		
$\times Rating_i$			−0.0004	
			(0.0153)	
$\times \log(Avg_Review_Length_i)$				0.0280**
				(0.0134)
Restaurant FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Clusters	2,740	2,740	2,740	2,740
Observations	235,156	235,156	235,156	235,156
R^2	0.9208			

Notes: Column (1) reports the baseline difference-in-differences estimate of Equation (1); columns (2)–(4) report triple-difference estimates of Equation (2). The demand outcome is aggregated on a 30-minute grid, treating a half-hour interval as available if either constituent 15-minute slot is available. The corresponding 15-minute estimates are reported in Tables 2 and 3. Each column is a separate regression. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

5.4. Placebo Test

Third, we conduct placebo integration-date tests by assigning fake integration dates within periods that do not contain the true integration cutoff. The pre-period placebo uses January 13, 2026 within the January 3–February 1, 2026 window, and the post-period placebo uses March 27, 2026 within the February 16–April 18, 2026 window. None of the placebo heterogeneity estimates (Table 8) is statistically significant at conventional levels.

6. Mechanisms

To shed light on the mechanisms underlying these heterogeneous effects, we focus on two primary explanations: changes in consumer search behavior and changes in AI- or platform-generated restaurant recommendations.

6.1. Consumer-side Mechanisms

To uncover the mechanisms underlying consumer behavior, we proceed in three steps. First, we analyze the prompts users use when searching for restaurant reservations to understand how this search behavior differs from traditional search. Second, we examine how restaurants with longer

Table 8 Placebo integration dates: review-cue heterogeneity

Panel A. Pre-period placebo date: January 13, 2026	(1) Google Review Count	(2) Rating	(3) Review Length
Treat × $\text{Post}^{placebo}$	0.0051	0.0407	0.0129
	(0.0142)	(0.0496)	(0.0531)
$\times \log(1 + Google_Review_Count_i)$	−0.0008		
	(0.0022)		
$\times Rating_i$		−0.0092	
		(0.0112)	
$\times \log(Avg_Review_Length_i)$			−0.0023
			(0.0100)
Restaurant FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	80,465	80,465	80,465
R^2	0.9596	0.9596	0.9597
Panel B. Post-period placebo date: March 27, 2026	(1) Google Review Count	(2) Rating	(3) Review Length
Treat × $\text{Post}^{placebo}$	0.0294	−0.0084	−0.0988
	(0.0183)	(0.0585)	(0.0640)
$\times \log(1 + Google_Review_Count_i)$	−0.0046		
	(0.0028)		
$\times Rating_i$		0.0019	
		(0.0130)	
$\times \log(Avg_Review_Length_i)$			0.0184
			(0.0119)
Restaurant FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	150,763	150,763	150,763
R^2	0.9466	0.9466	0.9466

Notes: Triple-difference placebo estimates of Equation (2). Panel A uses only the pre-integration window and assigns a placebo cutoff on January 13, 2026; Panel B uses only the post-integration window and assigns a placebo cutoff on March 27, 2026. Each column is a separate regression whose moderator Z_i is named in the header. Restaurant and date fixed effects; standard errors clustered at the restaurant level in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

reviews differ from other restaurants and link these differences to restaurant demand. Third, we characterize high-dimensional heterogeneous treatment effects across restaurants to identify which specific types of information drive the benefits of richer review content.

6.1.1. Real User Prompt Analysis We begin by characterizing the kinds of requests users pose to an AI assistant before comparing recommendation outputs. We construct a corpus of 1,000 realistic New York City dinner-recommendation prompts by asking three frontier models, GPT-5.1,[18] Claude

[18] https://developers.openai.com/api/docs/models/gpt-5.1

Sonnet 5,[19] and Gemini 3.5 Flash[20] to generate roughly equal shares of user-level requests (334, 333 and 333 respectively). The generation follows a snowball design. In each round, we ask for realistic requests whose intent differs from the prompt types already covered, so successive rounds expand the space of user intents rather than repeating it. Appendix G compares facet coverage across the three generating models. Appendix K presents ten prompts randomly drawn from the 100-prompt subset used in the recommendation experiment.

To analyze the prompts and compare it with traditional search behavior, we pass every prompt to GPT-5.1 and ask it to translate the request into the OpenTable search query that a user would actually execute: a short free-text query together with extracted keywords, each tagged by facet, including cuisine, dish, neighborhood, borough, landmark, price, feature or vibe, occasion, party size, or time.

The average prompt maps to 4.87 extracted keywords spanning 3.37 aspects, meaning that users routinely combine several contextual constraints within a single prompt; Figure 3 shows the full distribution of the number of keywords per prompt. Figure 4 reports how often each aspect appears: feature or vibe cues dominate, appearing in 86.0% of prompts. This suggests that users rely heavily on detailed attributes that are hard to include in OpenTable's search queries, especially open-vocabulary contextual features such as romantic, quiet, cozy, trendy, and date-night appeal. Appendix H lists the most frequent keywords within each of the ten aspects, for example, cozy and quiet for feature/vibe, and date night and birthday for occasion.

6.1.2. Are longer reviews more informative? We then examine whether longer reviews contain more information about the restaurant characteristics and dining goals expressed in natural-language prompts. A positive association would support the use of average review length as a proxy for textual informativeness, so the demand result above reflects genuine informational content rather than mere verbosity. Guided by the prompt analysis, we focus on two dimensions of restaurant information: detailed restaurant characteristics and the consumer goals or occasions that a restaurant may serve, both of which feature prominently in users' natural-language requests.

We test this at two levels. At the review level, we regress each review's information breadth on $\log(1 + \mathit{Review_Length}_{ij})$ with restaurant fixed effects, where breadth is defined as the number of distinct original characteristic or goal labels or distinct clusters contained in review j of restaurant i. At the restaurant level, we regress average breadth per review on $\log(\mathit{Filtered_Avg_Review_Length}_i)$, where $\mathit{Filtered_Avg_Review_Length}_i$ is the average review length over the subset of reviews retained

[19] https://www.anthropic.com/news/claude-sonnet-5

[20] https://deepmind.google/models/model-cards/gemini-3-5-flash/

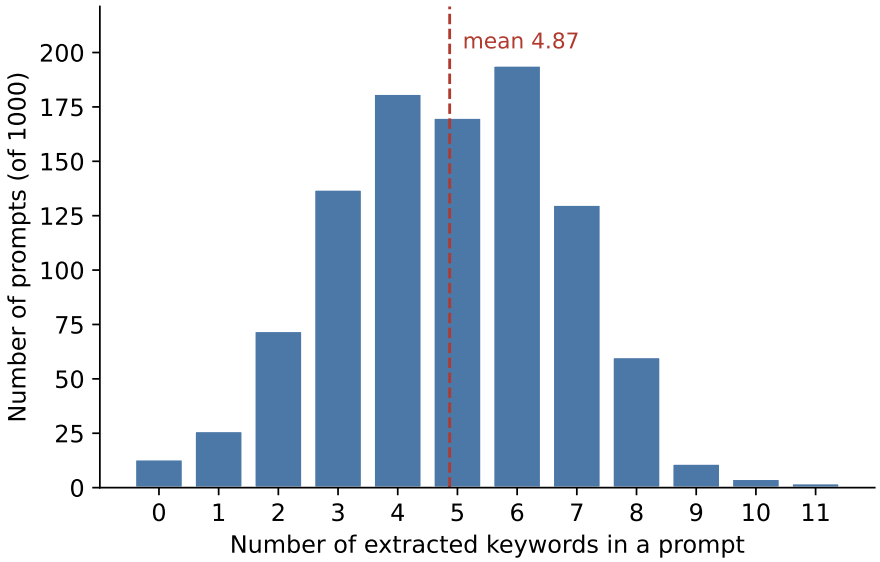


Figure 3 **Distribution of the number of extracted keywords per prompt across the 1,000 prompts (mean 4.87). Users typically combine several contextual constraints in a single request rather than issuing a single tag.**

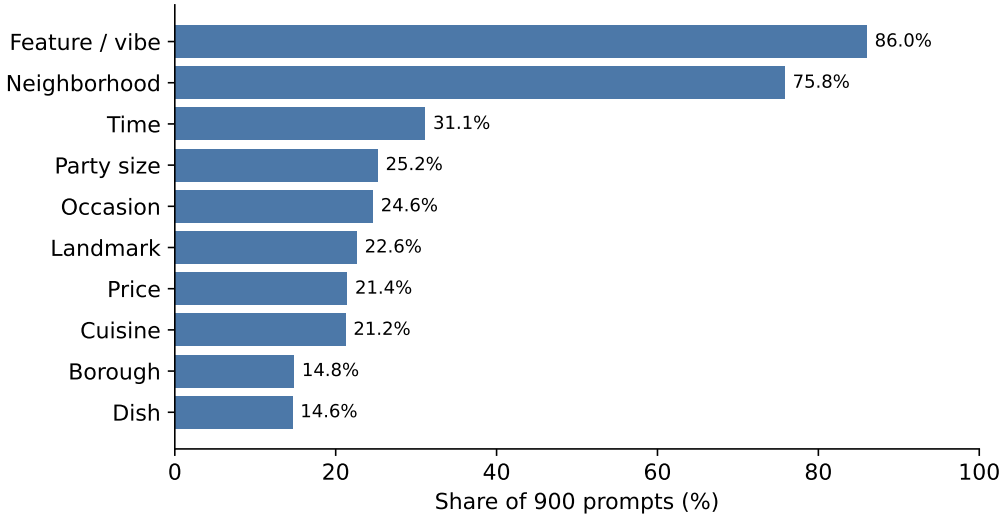


Figure 4 **Share of the 1,000 prompts that contain each search facet. Feature/vibe and neighborhood cues are by far the most common; explicit cuisine and dish terms are comparatively rare.**

for characteristic and goal extraction, controlling for the log total review count. Table 9 reports both sets of results. Panels A and B show that, in both the review-level and restaurant-level specifications, longer reviews are associated with more distinct characteristics and goals, whether measured at the original-label or the cluster level.

Table 9 Information Breadth and Review Length

	Original Char.	Original Goal	Char. Cluster	Goal Cluster
Panel A. Review-level (restaurant FE)				
$\log(1 + Review_Length_{ij})$	1.045^{***}	0.328^{***}	0.600^{***}	0.292^{***}
	(0.008)	(0.004)	(0.004)	(0.003)
Restaurant FE	Yes	Yes	Yes	Yes
Clustered SE	Restaurant	Restaurant	Restaurant	Restaurant
Observations	321,041	321,041	321,041	321,041
Within R^2	0.402	0.132	0.308	0.135
Panel B. Restaurant-level cross section				
$\log(Filtered_Avg_Review_Length_i)$	0.784^{***}	0.195^{***}	0.449^{***}	0.175^{***}
	(0.045)	(0.016)	(0.021)	(0.014)
Observations	2,740	2,740	2,740	2,740
R^2	0.172	0.138	0.181	0.164

Notes: The dependent variable is the information-breadth measure shown in each column. Standard errors are clustered at the restaurant level in Panel A and are HC1 robust in Panel B, both reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

To tie this validation directly to demand, we construct two *within-review breadth* measures that capture how much distinct, LLM-usable semantic content the average review carries. For each review used to extract characteristics and goals, we count the distinct semantic items it mentions and then average this count across restaurant i's most recent reviews used in that extraction stage. $Labels_per_Review_i$ is the average number of distinct characteristic-plus-goal *labels* per review, while $Clusters_per_Review_i$ is the average number of distinct characteristic-plus-goal embedding *clusters* per review, after collapsing near-synonyms. We then re-estimate the triple-difference specification in Equation (2), using each breadth measure as the restaurant-level moderator (Table 10).

Table 10 Heterogeneous Effects of LLM Integration by Within-Review Information Breadth

	(1) $Labels_per_Review_i$	(2) $Clusters_per_Review_i$
Treat × Post	-0.0384^{**}	-0.0464^{**}
	(0.0176)	(0.0234)
$\times Labels_per_Review_i$	0.0099^{**}	
	(0.0045)	
$\times Clusters_per_Review_i$		0.0150^{**}
		(0.0076)
Restaurant FE	Yes	Yes
Date FE	Yes	Yes
Observations	235,041	235,041
R^2	0.9290	0.9290

Notes: Standard errors clustered at the restaurant level in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Both within-review breadth moderators are positive and statistically significant (Table 10). In column (1), one additional characteristic-or-goal label per review is associated with a 0.99 percentage-point larger post-integration demand change for OpenTable restaurants relative to Resy restaurants. In column (2), one additional semantic cluster per review is associated with a 1.50 percentage-point larger relative demand change. Combined with the negative baseline Treat × Post coefficients, these positive interaction coefficients imply estimated demand losses for restaurants with low within-review information breadth and estimated demand gains for those with sufficiently high breadth. The estimated treatment effect crosses zero at approximately 3.9 distinct characteristic-or-goal labels per review in column (1) and 3.1 distinct semantic clusters per review in column (2). These results indicate that the AI channel rewards review text not merely for being longer, but for containing enough distinct information for the model to parse, summarize, and match to user intent.

6.1.3. High-dimensional heterogeneous treatment effect We next examine how the integration's effect on demand varies with the restaurant characteristics and dining goals described in text reviews. We use a doubly robust meta learner to capture the heterogeneity of treatment effects across restaurants' semantic profiles, represented by a high-dimensional vector of characteristics and dining goals. We then focus on the *distinctiveness* of these semantic profiles at the restaurant level. The intuition is that when a restaurant exhibits relatively uncommon characteristics or dining goals, detailed review text can make these features more visible and allow the AI integration to better match the restaurant to users' specific requests in ways that are difficult to achieve through traditional search. We therefore examine whether restaurants with more distinctive semantic profiles experience greater estimated treatment effects.

To characterize each restaurant's semantic profile, we construct an 86-dimensional vector $W_i = (q_{i1}, \ldots, q_{i86})$, covering 29 characteristic clusters and 57 goal clusters. Let c_{if} denote the number of restaurant i's recent reviews mentioning cluster f, counting each cluster at most once per review. For each restaurant, we define

$$q_{if} = \frac{c_{if}}{\sum_{k=1}^{86} c_{ik}}.$$

These shares capture the relative composition of semantic attributes mentioned in each restaurant's reviews.

Building on the framework for CATE estimation in Kennedy (2023), we construct doubly robust pseudo-outcomes using restaurant-level changes in average demanded-slot proportion between the pre- and post-integration periods, following the adaptation to the difference-in-differences setting

described in Lan et al. (2025). Let D_i indicate treatment and let $\Delta Y_i = \bar{Y}_{i,\text{post}} - \bar{Y}_{i,\text{pre}}$ denote the change in restaurant i's average demanded-slot proportion. Define

$$g_d(w) = \mathbb{E}[\Delta Y_i \mid D_i = d, W_i = w], \qquad p(w) = \Pr(D_i = 1 \mid W_i = w), \tag{3}$$

where $g_d(w)$ describes expected demand changes within treatment group d and $p(w)$ is the propensity score. Under no anticipation, overlap, and parallel trends conditional on W_i, the difference $g_1(w) - g_0(w)$ identifies the average treatment effect for treated restaurants with semantic profile w. Then we combine estimates of these functions to construct the doubly robust pseudo-outcome[21]

$$\widehat{U}_i = \widehat{g}_1(W_i) - \widehat{g}_0(W_i) + \frac{D_i\{\Delta Y_i - \widehat{g}_1(W_i)\}}{\widehat{p}(W_i)} - \frac{(1 - D_i)\{\Delta Y_i - \widehat{g}_0(W_i)\}}{1 - \widehat{p}(W_i)}. \tag{4}$$

The first two terms give the predicted difference in demand changes between treated and control restaurants with the same semantic profile. The remaining terms adjust this difference using propensity-weighted prediction errors. We implement this procedure using 5-fold cross-fitting. The treatment-assignment model ($\hat{p}$) uses L_2-penalized logistic regression. We estimate the outcome regressions ($\hat{g}_1$, $\hat{g}_0$) separately for treated and control restaurants using XGBoost.[22]

We next examine the association between the HTE and the extent to which each restaurant's semantic composition differs from that of other restaurants. Following Corritore et al. (2020) and Ahn and Greve (2025), we measure semantic distinctiveness using Jensen–Shannon distance (Lin 1991):

$$\text{Distinctiveness}_i = \sqrt{\tfrac{1}{2}\sum_f q_{if} log(\frac{2q_{if}}{q_{if} + q_{-if}}) + \tfrac{1}{2}\sum_f q_{-if} log(\frac{2q_{-if}}{q_{if} + q_{-if}})}, \tag{5}$$

where $q_{-if} = \frac{\sum_{j \neq i} q_{jf}}{\sum_{f'} \sum_{j \neq i} q_{jf'}}$ is the share of semantic category f in the reference distribution formed by averaging the composition vectors of all other restaurants and renormalizing to sum to one. A higher value of Distinctiveness_i indicates a more unusual combination of restaurant characteristics and dining goals. Restaurants without extracted content receive a distinctiveness value of zero by convention; their composition vectors are set to zero when constructing the reference distribution.

We standardize Distinctiveness_i using its mean and standard deviation among the 1,171 treated restaurants, denoting the resulting measure by $\widetilde{\text{Distinctiveness}}_i$. Using the 1,171 treated restaurants, we estimate the following regression with and without controls:

$$\widehat{U}_i = \alpha + \beta \widetilde{\text{Distinctiveness}}_i + \gamma' C_i + \varepsilon_i. \tag{6}$$

[21] The pseudo-outcome captures the CATE and the noise of each individual observation.

[22] Model hyperparameters are selected using three-fold cross-validation within each outer training fold, choosing the configuration that best predicts pre-to-post changes in demand as measured by mean squared error.

In the specification with controls, C_i includes pre-integration platform rating, price-category indicators, 12 cuisine indicators, and restaurant tenure. The specification without controls omits $\gamma' C_i$. In both specifications, semantic distinctiveness is positively and significantly associated with the estimated treatment effect. A one-standard-deviation increase in distinctiveness is associated with an approximately 2.7–2.8 percentage-point greater treatment effect, and the estimate is highly similar after adding restaurant controls. This pattern is consistent with the idea that the AI integration generates greater demand benefits for restaurants whose review text is more distinctive.

Table 11 **Semantic Distinctiveness and Treatment Effects: XGBoost**

Dependent variable: $100\widehat{U}_i$	(1)	(2)
$\widetilde{\text{Distinctiveness}}_i$	2.678**	2.842**
	(1.208)	(1.246)
Controls	No	Yes
Observations	1,171	1,171
R^2	0.0042	0.0202

Notes: The sample includes treated restaurants. Distinctiveness is measured by Jensen–Shannon distance and standardized within the treated sample. Controls include pre-period platform rating, price-category indicators, 12 cuisine indicators, and restaurant tenure. XGBoost outcome-model hyperparameters are selected using three-fold cross-validation within each outer training fold. The propensity model uses L2-penalized logistic regression with $C = 0.05$. Standard errors appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

6.2. AI- or Platform-side Mechanisms

To understand how system-side differences affect demand, we compare how restaurants recommended by AI differ from those surfaced by OpenTable under similar search criteria. Specifically, we construct three sets of search queries that vary in complexity and specificity. First, we use single-keyword searches based on generic terms drawn from OpenTable's publicly available cuisine and dining-topic labels (e.g., Sushi, Romantic, and Good for Groups).[23] These queries are intended to mimic the relatively simple search behavior commonly supported by traditional restaurant platforms. Second, we use single-keyword searches based on more detailed attributes drawn from our broader semantic inventory of 86 clusters, including 57 goal-oriented clusters and 29 restaurant-characteristic clusters. Third, we construct multi-keyword searches using combinations of attributes derived from real user prompts, which more closely mimic the richer and more complex preferences users express when interacting with AI assistants. We randomly sample 50, 50, and 100 queries from the three search sets, respectively.

[23] https://www.opentable.com/s/dinerschoice?metroid=8

For each query, we compare the top 10 results, or all available results when fewer than 10 are returned, from OpenTable with recommendations generated by a web-search-grounded OpenAI model, GPT-5.6 Luna.[24] We prompt the model to recommend up to 10 New York City restaurants that are bookable on OpenTable for the same party size, date, and time used in the corresponding OpenTable search. Each query is evaluated in five replicates on both OpenTable and OpenAI. Across replicates, we hold the core search request fixed and vary only the future booking slot, allowing us to account for both time-varying reservation inventory and stochasticity in AI-generated recommendations.

Table 12 summarizes the overlap between the AI assistant's recommendations and the platform's search results at the prompt-replicate level. Each observation corresponds to one replicate of a prompt, with five replicates per prompt.[25]

Table 12 AI–Platform Recommendation Overlap by Prompt Type

		Distinct restaurants returned		Mean per observation			
Family	N	AI	Platform	AI	Platform	AI ∩ Platform	Jaccard
Generic	250	825	428	9.96	9.31	1.70	0.106
Detailed	250	652	201	9.99	3.56	0.23	0.015
Real-user	500	1,160	132	10.00	1.35	0.11	0.006

Notes: *Generic* prompts use platform cuisine and dining-topic keywords; *Detailed* prompts use review-derived clusters; and *Real-user* prompts are free text. N counts prompt–replicate observations. Each prompt has five replicates that differ only in future booking slot. *Distinct restaurants returned* is the type-level union across all prompts and replicates. Under *Mean per observation*, *AI* and *Platform* are the mean numbers of distinct restaurants returned per prompt–replicate observation, *AI* ∩ *Platform* is their mean intersection, and *Jaccard* is the mean observation-level ratio $|A \cap P|/|A \cup P|$ (not a ratio of column means). Observations with no platform result are retained with Jaccard zero. Appendix J reports results pooling the five replicates within each prompt and conditioning on nonempty platform results.

First, we find that the overlap between ChatGPT and OpenTable is low even for very generic search queries. The mean Jaccard index is only 0.106, indicating that, on average, only 10.6% of the unique restaurants appearing in the combined recommendation sets are shared by both platforms. Second, overlap declines as queries become more specific: the mean Jaccard indices are 0.106, 0.015, and 0.006 for generic, detailed, and real-user prompts, respectively. Pairwise two-sided Mann–Whitney U tests indicate statistically significant differences in the distributions of overlap across all three prompt types (Holm-adjusted $p < 0.001$ for each comparison). Third, the particularly low overlap for real-user prompts is not driven solely by differences in the restaurants recommended by the two platforms; it also reflects OpenTable's frequent failure to return any restaurant at all. OpenTable returns at

[24] https://developers.openai.com/api/docs/models/gpt-5.6-luna

[25] The prompt-level analysis yields similar results, which we report in Appendix J.

least one restaurant for 98% of generic prompts and 66% of detailed prompts, but for only 16% of real-user prompts. These results could arise from either limitations in OpenTable's search algorithm or the AI model returning recommendations that are not sufficiently grounded in restaurants that actually satisfy the search criteria; however, both explanations point to systematic differences in how the two systems translate user queries into restaurant recommendations.

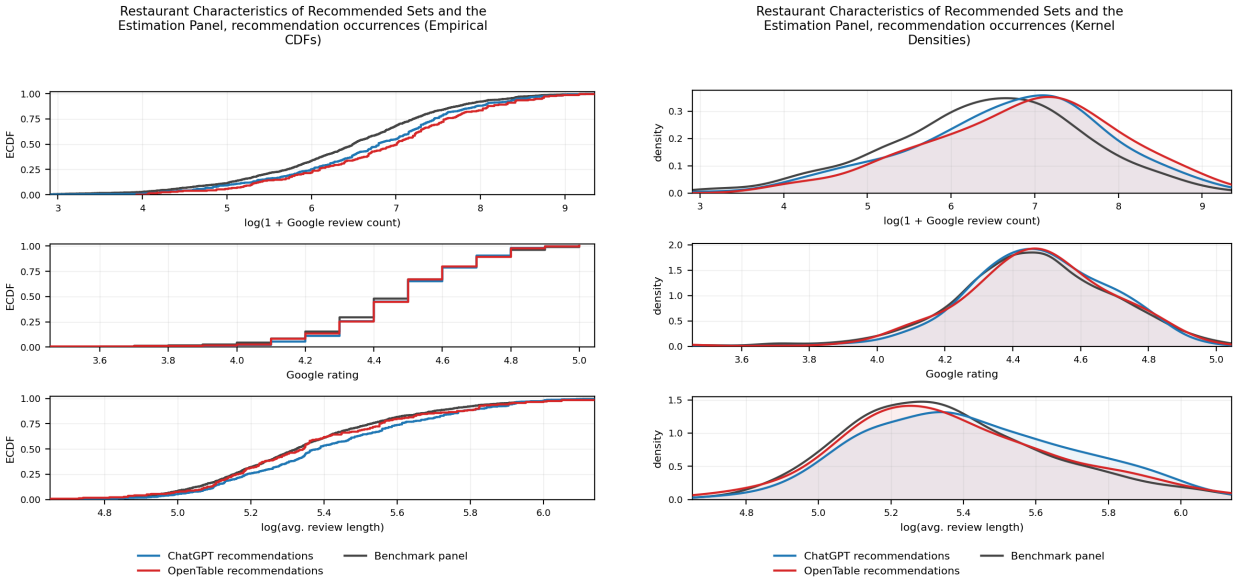
We next compare the characteristics of the restaurants surfaced by the two systems and focus on the restaurants in our analysis sample. In total, ChatGPT recommended 734 distinct restaurants and OpenTable returned 333 distinct restaurants. We plot the distributions of Google review count, Google rating, and average review length for restaurants recommended by OpenTable, restaurants recommended by ChatGPT, and restaurants in our OpenTable estimation panel, as shown in Figure 5.[26]

First, in terms of the number of Google reviews, both ChatGPT and OpenTable surface restaurants with more reviews than the average restaurant. The distribution of the total number of Google reviews for restaurants recommended by OpenTable lies to the right of that for restaurants recommended by ChatGPT, indicating first-order stochastic dominance. This pattern is consistent with the earlier heterogeneity results: restaurants with more reviews are not disproportionately favored in the AI channel. Second, for the ratings, the overall distribution is similar across restaurants recommended by AI, restaurants recommended by platforms as well as the overall distribution of all restaurants. Third, regarding review length, the distribution for restaurants recommended by ChatGPT lies to the right of both the estimation panel and the OpenTable recommendations, whereas the OpenTable distribution remains relatively close to that of the estimation panel. This pattern is again consistent with the earlier heterogeneity results, which suggest that restaurants with longer reviews are relatively more favored in the AI channel. This evidence together suggests that the systematic differences between platform recommendations and AI recommendations also contribute to the heterogeneous effects of AI integration on restaurant demand. Appendix I provides complementary evidence by repeating the distributional comparison for generic prompts and reporting tests of differences in mean and median review characteristics between the two systems' recommendations.

7. Conclusion

This paper studies how integrating a large language model into a transaction platform affects consumer demand. Using the February 2026 ChatGPT–OpenTable integration as a natural experiment,

[26] Restaurants recommended multiple times are counted multiple times in these distributions. This weighting preserves the frequency with which each restaurant is surfaced by a system and therefore reflects the characteristics of the recommendations users are more likely to encounter.



(a) Empirical CDFs (b) Kernel densities

Figure 5 Recommended Restaurants Against the Estimation Panel

Notes: The figures compare restaurants surfaced by ChatGPT and OpenTable with the OpenTable restaurants in the estimation panel. The three restaurant characteristics are Google review count, Google rating, and average review length, transformed as in Table 3. The density panel uses weighted Gaussian kernel densities with Scott's-rule bandwidth.

we find no detectable expansion in aggregate demand for OpenTable restaurants relative to comparable Resy restaurants. The aggregate null, however, masks systematic redistribution across restaurants. Restaurants with larger accumulated review stocks experience estimated demand losses following the integration, while those with more informative review text experience estimated demand gains. Review ratings do not meaningfully moderate the treatment effect. Using a doubly robust meta-learner, we find that restaurants whose reviews describe more distinctive combinations of characteristics and dining goals have larger estimated treatment effects. Our mechanism analyses provide complementary evidence: realistic user prompts illustrate the contextual preferences that conversational search can accommodate, while the recommendation comparison shows that the AI system surfaces restaurants differently from OpenTable's conventional search interface. The integration therefore reorganizes demand by changing both the discovery and matching processes, rather than simply adding traffic to the integrated platform.

The study contributes to the literature in several ways. First, it contributes to research on online restaurant reservation platforms by examining a new form of digital intermediation: integration with AI assistants. We show that such integration can shift demand across restaurants even when aggregate demand does not increase. Second, the paper contributes to research on consumer behavior in the

AI era by showing how AI-mediated interfaces change the way consumers search for information. In particular, users can express more detailed, contextual, and intent-rich requests than conventional platform search interfaces typically support, with consequences for how demand is allocated. Third, it contributes to the literature on consumer search and recommendation in differentiated markets. Prior research has emphasized that digital platforms reduce search costs but may also concentrate attention on a relatively small set of salient or popular products. We show that LLM-mediated search operates differently: review counts and ratings become less predictive of demand gains, while rich review text and semantic attributes become more valuable. By matching unstructured product information to contextual consumer intent, LLM-mediated search may expand the effective set of products consumers can discover and consider, potentially increasing effective product variety.

These findings have practical implications for restaurants, platforms, and governance. For restaurants, especially newer or less established ones, AI-mediated discovery may create an opening to surface differentiated attributes without first accumulating very large review stocks. Restaurants can improve discoverability by encouraging informative reviews, making distinctive features and use cases legible online, and ensuring that machine-readable descriptions consistently communicate what kinds of diners and occasions the restaurant serves. For platforms, the results highlight a strategic choice about integration. A platform can expose inventory to an external AI assistant, build its own AI-mediated search layer, or do both. External integration may expand reach and meet users where they search, but it can also shift control over ranking, presentation, and consumer data to the AI intermediary. Internal integration preserves more control over marketplace objectives such as availability, diversity, fairness, and seller exposure. Finally, because AI-mediated discovery increases the value of machine-readable content, it may also increase incentives for strategic content manipulation. Platforms and regulators may need to pay closer attention to review authenticity, content manipulation, disclosure, and the accountability of recommendation systems that increasingly shape market access.

Several limitations point to future research. First, our evidence captures short-run effects following a major integration. Future work could study the new equilibrium that emerges after restaurants, platforms, and consumers adapt: ratings may change, review production may respond, and restaurants may strategically alter how they describe themselves online. Second, we do not directly observe individual user behavior inside ChatGPT or OpenTable. Experiments or clickstream data could measure the full search path, including what users ask, which recommendations they inspect, how often they click through from the AI assistant to the platform, what information they attend to, and which recommendation explanations change booking decisions. Third, our empirical setting is

the restaurant industry, where experiential attributes and textual reviews are unusually important. Future research could examine whether similar mechanisms operate in other AI-mediated markets, including travel, local services, retail product discovery, and high-consideration purchases in which AI assistants increasingly guide buying decisions.

References

Ahn Y, Greve HR (2025) Cultural spawning: Founders bringing organizational cultures to their startup. *Organization Science* 36(1):411–428.

Allouah A, Besbes O, Figueroa JD, Kanoria Y, Kumar A (2026) What is your AI agent buying? evaluation, biases, model dependence, & emerging implications of agentic e-commerce. *Proceedings of the ACM Web Conference 2026*, 8697–8700.

Anderson M, Magruder J (2012) Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122(563):957–989.

Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton, NJ: Princeton University Press).

Aral S, Li H, Zuo R (2026) The rise of AI search: Implications for information markets and human judgement at scale. *arXiv preprint arXiv:2602.13415* .

Balafoutas L, Fornwagner H, Hauser E, Hauser O (2026) Gender-inclusive language and economic decision-making. *Management Science* 72(9):8042–8056.

Bang D, Jang SS (2026) Platform endorsements as drivers of consumer behavior: A quasi-experimental study using restaurant review data. *International Journal of Hospitality Management* 133:104502.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119(1):249–275.

Braghieri L, Levy R, Makarin A (2022) Social media and mental health. *American Economic Review* 112(11):3660–3693.

Burtch G, Lee D, Chen Z (2024) The consequences of generative AI for online knowledge communities. *Scientific Reports* 14(1):10413.

Chatterji A, Cunningham T, Deming DJ, Hitzig Z, Ong C, Shan CY, Wadman K (2025) How people use ChatGPT. Technical Report 34255, National Bureau of Economic Research.

Chen PY, Hitt LM (2002) Measuring switching costs and the determinants of customer retention in Internet-enabled businesses: A study of the online brokerage industry. *Information Systems Research* 13(3):255–274.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43(3):345–354.

Cheyre C, Acquisti A (2024) Online intermediation in legacy industries: Evidence from the adoption of restaurant reservation platforms. *Available at SSRN 4721874* .

Chien CY, Yen AZ, Huang HH, Chen HH (2026) User profiling for specification-sensitive recommendations with large language model prompting. *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, 609–618.

Cinelli C, Forney A, Pearl J (2024) A crash course in good and bad controls. *Sociological Methods & Research* 53(3):1071–1104.

Corritore M, Goldberg A, Srivastava SB (2020) Duality in diversity: How intrapersonal and interpersonal cultural heterogeneity relate to firm performance. *Administrative Science Quarterly* 65(2):359–394.

del Rio-Chanona RM, Laurentsyeva N, Wachs J (2024) Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus* 3(9):pgae400.

Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10(28):eadn5290.

Exner Y, Hartmann J, Ding Z, Zhang S, Netzer O (2025) AI in disguise—quasi-experimental analysis of a large-scale deployment of AI-generated ads. *Available at SSRN 5096969* .

Fang L, Yuan Z, Zhang K, Donati D, Sarvary M (2025) Generative AI and sales productivity: Field experiments in online retail. *arXiv preprint arXiv:2510.12049* .

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3):291–313.

Hou J, Jiang DL, Ma X, Gopal A (2026a) The costs of ambiguous information disclosure: On the unintended consequences of providing restaurant hygiene scores on platforms in the United States. *Information Systems Research* 37(1):81–102.

Hou J, Yang S, Xiong G, Pavlou PA (2026b) Enhancing AI-assisted purchase decisions: The role of the sense of autonomy. *MIS Quarterly* 50(2):589–614.

Huang X, Lian J, Lei Y, Yao J, Lian D, Xie X (2025) Recommender AI agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems* 43(4):1–33.

Jiang L, Ravichandran T, Kuruzovich J (2023) Review moderation transparency and online reviews: Evidence from a natural experiment. *MIS Quarterly* 47(4):1693–1708.

Kaiser M, Schulze C (2026) Frontiers: ChatGPT referrals to e-commerce websites: How do LLMs compare against traditional channels? *Marketing Science* 45(4):699–715.

Kennedy EH (2023) Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2):3008–3049.

Klemperer P (1987) Markets with consumer switching costs. *The Quarterly Journal of Economics* 102(2):375–394.

Lan H, Chang H, Dillon E, Syrgkanis V (2025) A meta-learner for heterogeneous effects in difference-in-differences. *arXiv preprint arXiv:2502.04699* .

Lasmi H, Lee CH, Ceran Y (2025) Popularity brings better sales or vice versa: Evidence from Instagram and OpenTable. *Global Business Review* 26(4):923–967.

Lee D, Hosanagar K (2019) How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research* 30(1):239–259.

Lee S, Back J, Jeon D, Kwon S, Kim M, Kang I, Lee J (2026) From relevance to authority: Authority-aware generative retrieval in web search engines. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics: Industry Track*, 796–811.

Li X, Grahl J, Hinz O (2022) How do recommender systems lead to consumer purchases? a causal mediation analysis of a field experiment. *Information Systems Research* 33(2):620–637.

Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

Liu J, Xu X, Nan X, Li Y, Tan Y (2023) "Generate" the future of work through AI: Empirical evidence from online labor markets. *arXiv preprint arXiv:2308.05201* .

Liu W, Ma X, Zhu Y, Li Y, Shi D, Yin D, Dou Z (2026) Agentic-R: Learning to retrieve for agentic search. *Findings of the Association for Computational Linguistics: ACL 2026*, 15987–16005.

Luca M (2016) Reviews, reputation, and revenue: The case of Yelp.com. *Harvard Business School NOM Unit Working Paper* (12-016).

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.

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104(8):2421–2455.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

Padilla N, Lam HT, Lambrecht A, Hollenbeck B (2026) The impact of LLM adoption on online user behavior. *Available at SSRN 5393256* .

Raj S, Jha AV, Anand M, Saha S (2026) HARPO: Hierarchical agentic reasoning for user-aligned conversational recommendation. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 35580–35599.

Roy D, Spiliotopoulou E, de Vries J (2022) Restaurant analytics: Emerging practice and research opportunities. *Production and Operations Management* 31(10):3687–3709.

Singh A, Fry A, Perelman A, Tart A, Ganesh A, El-Kishky A, McLaughlin A, Low A, Ostrow A, Ananthram A, et al. (2025) OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267* .

Tafti A, Shmueli G (2020) Beyond overall treatment effects: Leveraging covariates in randomized experiments guided by causal structure. *Information Systems Research* 31(4):1183–1199.

Ursu R, Rao A, Embrey J (2026) The help me search era: Do we choose "better" with AI? *Available at SSRN 7208681* .

Wang Y, Qin MS, Luo X, Kou Y (2022) Frontiers: How support for Black Lives Matter impacts consumer responses on social media. *Marketing Science* 41(6):1029–1044.

Wang Z, Du Y, Wei T, Chua H, Bi J, Zhang J, Sun Z (2026) Adaptive and representative multi-interest modeling for recommendation with large language model. *Findings of the Association for Computational Linguistics: ACL 2026*, 2484–2496.

Xue B, Wu B, Qiao S, Wang S, Wang R, Du Y, Wang H, Pan JZ, Yilmaz E, Wong KF, Lipani A (2026) Mitigating context interference for reliable and efficient search agents. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3541–3558.

Online Appendix to "When Search Becomes a Conversation: The Impact of AI Integration on Demand in Online Marketplaces"

Appendix A: Summary Statistics by Platform

This appendix reports summary statistics separately for OpenTable and Resy restaurants. We present two sets of variables for each platform. The first set is measured at the restaurant–day level and includes the outcome variable and post-period indicator. The second set is measured at the restaurant level and captures the review-cue and textual-content measures used in the main heterogeneity analyses, including Google review volume, Google rating, average review length, and semantic breadth extracted from review text.

Table A1 Summary Statistics by Platform

Variable	N	Mean	SD	Min	P25	Median	P75	Max
Panel A. OpenTable restaurants								
Outcome, treatment, and time-varying controls (restaurant–day level)								
Y_{it} (demanded slot proportion)	96,275	0.214	0.287	0.000	0.000	0.092	0.311	1.000
$Post_t$	96,275	0.630	0.483	0.000	0.000	1.000	1.000	1.000
$Rating_{it}$ (platform)	93,394	4.583	0.277	1.000	4.500	4.600	4.800	5.000
$\log(1+Reviews_{it})$ (platform)	95,019	5.313	1.951	0.000	4.007	5.407	6.711	9.911
Pre-existing online information (restaurant level)								
$Google_Review_Count_i$	1,171	1141.991	1593.942	4.000	302.000	659.000	1341.500	24324.000
$Rating_i$ (Google)	1,171	4.455	0.239	3.200	4.300	4.500	4.600	5.000
$Avg_Review_Length_i$	1,171	220.244	67.394	66.800	173.155	205.331	251.588	546.662
$Labels_per_Review_i$	1,171	3.930	0.997	0.000	3.254	3.788	4.520	8.571
$Clusters_per_Review_i$	1,171	3.108	0.612	0.000	2.711	3.080	3.444	5.714
Panel B. Resy restaurants								
Outcome, treatment, and time-varying controls (restaurant–day level)								
Y_{it} (demanded slot proportion)	138,881	0.349	0.331	0.000	0.042	0.252	0.584	1.000
$Post_t$	138,881	0.657	0.475	0.000	0.000	1.000	1.000	1.000
$Rating_{it}$ (platform)	138,881	4.611	0.240	2.400	4.500	4.700	4.800	5.000
$\log(1+Reviews_{it})$ (platform)	138,791	6.416	1.741	0.693	5.260	6.504	7.697	11.019
Pre-existing online information (restaurant level)								
$Google_Review_Count_i$	1,569	811.389	1205.269	5.000	174.000	451.000	985.000	19100.000
$Rating_i$ (Google)	1,569	4.498	0.255	2.400	4.400	4.500	4.700	5.000
$Avg_Review_Length_i$	1,569	240.157	79.565	82.843	185.524	223.784	276.080	888.556
$Labels_per_Review_i$	1,569	3.804	0.979	0.000	3.220	3.714	4.287	8.190
$Clusters_per_Review_i$	1,569	3.064	0.643	0.000	2.703	3.054	3.426	6.222

Notes: This table reports the same variables as Table 1, but separately for treated OpenTable restaurants and control Resy restaurants. Within each platform panel, the first block is at the restaurant–day level and the second block is at the restaurant level. Tenure and local-guide review volume are reported in Appendix C. Restaurants from which no characteristic or goal label could be extracted are assigned zeros for the review-breadth variables $Labels_per_Review_i$ and $Clusters_per_Review_i$.

Appendix B: Washout and Donut Difference-in-Differences Around the Integration Date

Because our main outcome is a forward-looking measure of reservation availability rather than contemporaneous booking flows, observations immediately after the integration date may still reflect reservations booked before February 16, 2026. Consistent with this concern, OpenTable reports that 56% of diners make holiday reservations at least one week in advance, implying that reservation availability may adjust only gradually following the integration.[27] To assess whether this timing blur attenuates the treatment estimates, we re-estimate the paper's baseline DiD and heterogeneity DDD specifications after deleting the first L post-integration days from the sample while keeping the true treatment date fixed. We report the two shortest washouts, $L = 7$ and $L = 14$, which are the most direct tests of whether the early post period is contaminated by pre-integration bookings.

Table B1 shows that the aggregate Treat×Post coefficient remains economically small and statistically indistinguishable from zero under both washouts. Table B2 shows that the main review-cue heterogeneity patterns also survive: the review-volume interaction remains negative, the rating interaction remains null, and the review-length interaction remains positive and statistically significant. These estimates therefore suggest that blurred treatment timing does not drive the paper's substantive conclusions.

Table B1 Washout robustness of the baseline DiD: deleting the first L post-integration days.

	(1) $L = 7$	(2) $L = 14$
Treat × Post	−0.0009	−0.0009
	(0.0039)	(0.0040)
Restaurant FE	Yes	Yes
Date FE	Yes	Yes
Observations	215,878	198,998
R^2	0.9277	0.9274

Notes: Difference-in-differences estimates of Equation (1) after deleting the first L post-integration days, with L given by the column header. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

[27] https://www.opentable.com/restaurant-solutions/valentines-day-trends

Table B2 Washout robustness of the heterogeneity results: deleting the first L post-integration days.

	(1) Google Review Count	(2) Rating	(3) Review Length
Panel A. $L = 7$			
Treat × Post	0.0373^{*}	0.0154	-0.1618^{**}
	(0.0198)	(0.0685)	(0.0731)
$\times \log(1 + Google_Review_Count_i)$	-0.0062^{**}		
	(0.0031)		
$\times Rating_i$		−0.0037	
		(0.0153)	
$\times \log(Avg_Review_Length_i)$			0.0298^{**}
			(0.0136)
Restaurant FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	215,878	215,878	215,878
R^2	0.9278	0.9277	0.9278
Panel B. $L = 14$			
Treat × Post	0.0360^{*}	0.0052	-0.1616^{**}
	(0.0203)	(0.0706)	(0.0747)
$\times \log(1 + Google_Review_Count_i)$	-0.0060^{*}		
	(0.0031)		
$\times Rating_i$		−0.0014	
		(0.0158)	
$\times \log(Avg_Review_Length_i)$			0.0298^{**}
			(0.0139)
Restaurant FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	198,998	198,998	198,998
R^2	0.9274	0.9274	0.9274

Notes: Triple-difference estimates of Equation (2) after deleting the first L post-integration days, with L given by the panel header. Each column is a separate regression whose moderator Z_i corresponds to the column title. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Appendix C: Heterogeneity by Tenure and Local-Guide Review Volume

Table 3 reports the review-volume, rating and review-length moderators. This appendix reports the two additional information-stock moderators of Equation (2): $\log(1+Tenure_i)$, the log of the restaurant's age in months as of the day before integration, and $\log(1+Local_Guide_Review_Count_i)$, the log count of reviews written by Google Local Guides, whom the literature regards as the most credible reviewers (Mayzlin et al. 2014, Forman et al. 2008). Table C1 reports their summary statistics for the pooled sample and separately by platform. Together with total review volume these are progressively narrower measures of the same underlying construct, from accumulated tenure to total reviews to local-guide review volume.

Table C1 Summary Statistics for Tenure and Local-Guide Review Volume

Sample	Variable	N	Mean	SD	Min	P25	Median	P75	Max
Pooled	$Tenure_i$ (months)	2,740	100.972	91.990	1.938	28.868	75.378	152.932	421.518
Pooled	$Local_Guide_Review_Count_i$	2,740	610.677	922.893	0.000	118.750	323.500	762.250	19,153.000
OpenTable	$Tenure_i$ (months)	1,171	115.546	95.212	1.938	37.714	96.945	168.676	421.518
OpenTable	$Local_Guide_Review_Count_i$	1,171	722.892	1,086.357	1.000	162.500	407.000	870.000	19,153.000
Resy	$Tenure_i$ (months)	1,569	90.095	87.979	1.938	24.507	58.804	134.198	421.518
Resy	$Local_Guide_Review_Count_i$	1,569	526.927	768.621	0.000	90.000	282.000	667.000	11,097.000

Notes: This table reports restaurant-level summary statistics for the auxiliary information-stock measures used in this appendix. Tenure is measured in months since the earliest observed Google review. Local-guide review volume is the number of Google reviews contributed by identity-disclosed Local Guides.

Table C2 shows that both additional measures of information stock have similar moderating effects to Column (1) of Table 3 of our main paper. Column (1) implies that OpenTable restaurants with less than 64 months of tenure experience a positive relative post-integration effect, whereas those with longer tenure experience a negative one. Column (2) shows that restaurants with fewer than 276 local-guide reviews gain after the integration, while those above this threshold lose. Thus, both alternative information-stock measures imply a redistribution of demand: restaurants below the respective thresholds experience estimated demand gains, whereas those above them experience estimated demand losses.

Table C2 Heterogeneous Effects by Tenure and Local-Guide Review Volume

	(1) Tenure	(2) Local Guide
Treat × Post	0.0313^{**}	0.0315^{*}
	(0.0154)	(0.0165)
$\times \log(1 + Tenure_i)$	-0.0075^{**}	
	(0.0034)	
$\times \log(1 + Local_Guide_Review_Count_i)$		-0.0056^{**}
		(0.0027)
Restaurant FE	Yes	Yes
Date FE	Yes	Yes
Controls	No	No
Observations	235,156	235,156
R^2	0.9291	0.9291

Notes: Triple-difference estimates of Equation (2). Each column is a separate regression whose moderator Z_i corresponds to the column title. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Appendix D: Joint Moderator Specifications

Table D1 extends Table 3 by entering review count and review length jointly. Specifically, to estimate the joint moderating effect in a more granular way, we replace the continuous review-count moderator with a set of count-bin indicators. Let B_{ik} denote an indicator equal to one if restaurant i falls in review-count bin k, with the lowest-count bin omitted. We estimate

$$\begin{aligned} Y_{it} = {} & \alpha_i + \delta_t + \beta_0(Treat_i \times Post_t) \\ & + \sum_{k=2}^{K} \gamma_k(Post_t \times B_{ik}) + \sum_{k=2}^{K} \beta_k(Treat_i \times Post_t \times B_{ik}) \\ & + \gamma_l(Post_t \times \log(Avg_Review_Length_i)) + \beta_l(Treat_i \times Post_t \times \log(Avg_Review_Length_i)) + \varepsilon_{it}. \end{aligned} \tag{7}$$

The coefficient β_k compares the post-integration treatment effect for restaurants in count bin k with that for restaurants in the omitted lowest-count bin, while β_l captures how review length further moderates the treatment effect.

The estimates are qualitatively similar to our main results. Holding average review length fixed, restaurants in the middle and upper-middle review-count bins have lower estimated treatment effects than those in the lowest-count bin. This pattern is consistent with the continuous specification, in which the estimated treatment effect becomes negative once accumulated review volume exceeds the estimated cutoff. The pattern is not strictly monotonic: the most reviewed restaurants experience a smaller negative differential than restaurants in the middle bins, which is suggestive of a non-monotonic relationship between review volume and the demand effect. At the same time, review length positively moderates the treatment effect, indicating that restaurants with more informative textual review content are relatively more likely to benefit from AI-mediated discovery.

Table D1 Joint Heterogeneity with Review-Count Quintile Bins and Review Length

	(1) Quintile Bins
Panel A. Count-bin definitions	
Omitted bin	4–168
Bin 2	169–390
Bin 3	392–717
Bin 4	718–1,332
Bin 5	1,333–24,324
Restaurants per bin	548
Panel B. Regression estimates	
Treat × Post	−0.1153
	(0.0807)
Treat × Post × Bin 2	−0.0187
	(0.0120)
Treat × Post × Bin 3	-0.0221^{*}
	(0.0134)
Treat × Post × Bin 4	-0.0246^{**}
	(0.0125)
Treat × Post × Bin 5	−0.0115
	(0.0127)
Treat × Post × log($Avg_Review_Length_i$)	0.0242^{*}
	(0.0146)
Restaurant FE	Yes
Date FE	Yes
Observations	235,156
R^2	0.9291

Notes: This table reports estimates of Equation (7). Review count is discretized into quintile bins, each containing 548 restaurants. The omitted bin is the lowest review-count quintile. The coefficient on Treat × Post is therefore the estimated treatment effect for restaurants in the omitted bin. The coefficients on Treat × Post × Bin k measure the differential effect for restaurants in bin k relative to the omitted bin. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Appendix E: Specifications with Time-Varying Platform Controls

This appendix reports the specifications that add time-varying platform controls: the platform-displayed rating and the log cumulative platform review count. These variables may themselves respond to platform activity, so the main text reports the cleaner fixed-effects specifications without them and uses the estimates here as robustness checks.

Table E1 Summary Statistics for Platform Controls

Variable	N	Mean	SD	Min	P25	Median	P75	Max
$Rating_{it}$ (platform)	232,275	4.600	0.256	1.000	4.500	4.600	4.800	5.000
$\log(1 + Reviews_{it})$ (platform)	233,810	5.968	1.908	0.000	4.718	6.087	7.378	11.019

Notes: Platform controls are measured at the restaurant–day level. The control specifications use observations for which both controls are observed, yielding 230,929 restaurant–day observations and 2,731 restaurants.

Table E2 Aggregate DiD with Time-Varying Platform Controls

	(1)
Treat × Post	−0.0023
	(0.0036)
Restaurant FE	Yes
Date FE	Yes
Platform controls	Yes
Observations	230,929
R^2	0.9319

Notes: Difference-in-differences estimate of Equation (1), adding platform-displayed rating and $\log(1 + Platform_Review_Count_{it})$. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table E3 Binary-Split DDD with Time-Varying Platform Controls

	(1) Google Review Count	(2) Rating	(3) Review Length
Treat × *Post*	0.0053	−0.0040	−0.0108**
	(0.0056)	(0.0047)	(0.0049)
×*High Google Review Count*$_i$	−0.0143*		
	(0.0073)		
×*High Rating*$_i$		0.0044	
		(0.0075)	
×*High Review Length*$_i$			0.0182**
			(0.0072)
Restaurant FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Platform controls	Yes	Yes	Yes
Observations	230,929	230,929	230,929
R^2	0.9320	0.9320	0.9320

Notes: Triple-difference estimates of Equation (2), where each continuous moderator is replaced by an above-median indicator. All columns add platform-displayed rating and $\log(1+Platform_Review_Count_{it})$. Standard errors clustered at the restaurant level are reported in parentheses. $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.10$.

Appendix F: Comparing Google and Platform Information

The moderators in Table 3 are measured using restaurants' Google information. We use Google information because it is available for both OpenTable and Resy restaurants and, importantly, because it gives us access to review text on both platforms. This appendix shows that our results are not driven by systematic differences between Google-side information and the information displayed on the reservation platforms. We define

$$Rating_Gap_i = Google_Rating_i - Platform_Rating_i, \tag{8}$$

$$Log_Count_Gap_i = \log(1 + Google_Reviews_i) - \log(1 + Platform_Reviews_i). \tag{9}$$

The platform-side variables are measured before the integration, so each gap is a restaurant-level, time-invariant characteristic, as required by Equation (2). Table F1 enters each gap as a continuous moderator. Neither the rating gap nor the review-count gap significantly moderates the treatment effect, suggesting that the main heterogeneity results are not driven by discrepancies between Google and platform-displayed information.

Table F1 Heterogeneous Effects by Google–Platform Information Gaps

	(1) Rating Gap	(2) Review-Count Gap
$Treat \times Post$	−0.0004	−0.0017
	(0.0046)	(0.0042)
$\times Rating_Gap_i$	−0.0014	
	(0.0184)	
$\times Log_Count_Gap_i$		0.0021
		(0.0028)
Restaurant FE	Yes	Yes
Date FE	Yes	Yes
Controls	No	No
Observations	235,156	235,156
R^2	0.9290	0.9290

Notes: Triple-difference estimates of Equation (2). The moderator in column (1) is the difference between a restaurant's Google rating and its pre-integration platform-displayed rating. The moderator in column (2) is the difference between $\log(1 + Google_Reviews_i)$ and $\log(1 + Platform_Reviews_i)$, so it can be interpreted as a log review-count ratio. The platform-side variables are measured before the integration, making both gaps restaurant-level, time-invariant characteristics. Standard errors clustered at the restaurant level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix G: Cross-Model Comparison of User Prompts

The prompt corpus in Section 6.1.1 is generated in equal shares by three AI model families, which allows us to ask whether the demand facets a diner is imagined to provide depend on which assistant generates the prompt. This comparison is substantively relevant because OpenTable's assistant initially ran on OpenAI's model and later also integrated with Gemini, while Resy integrated with Claude; all three models therefore represent plausible interfaces through which users could make restaurant-recommendation requests. For each prompt, we use the same ten facets as in Section 6.1.1, tagged by the same GPT-5.1 extractor for every prompt, and record whether the prompt mentions each facet, together with the total number of facets mentioned.

Table G1 reports, for each facet, the share of prompts from each model that mention it. GPT-5.1 and Gemini exhibit broadly similar patterns. Feature/vibe is mentioned in more than 90% of prompts for both models, and neighborhood is also highly prevalent, appearing in 91.3% of GPT-5.1 prompts and 85.3% of Gemini prompts. The shares mentioning time, occasion, landmark, and borough are also relatively similar across the two AI models. Some differences emerge for more specific demand facets: Gemini mentions cuisine and dish more frequently than GPT-5.1, whereas GPT-5.1 mentions price and party size more frequently than Gemini. Overall, the average number of facets per prompt is nearly identical for GPT-5.1 and Gemini (3.68 versus 3.60). Claude shows a somewhat different pattern, most notably for location-related information. Only 50.8% of Claude prompts mention a neighborhood, compared with 85.3% for Gemini and 91.3% for GPT-5.1. Claude also mentions feature/vibe, time, and landmark less frequently than the other two models, while occasion is mentioned more frequently. By contrast, borough, cuisine, price, dish, and party size remain broadly comparable in magnitude across at least two of the three models. Consistent with these differences, Claude prompts contain fewer facets on average (2.78) than GPT-5.1 (3.68) and Gemini (3.60). The observed differences across model families may suggest systematic variation in both user behaviors in using different models and how different models formulate prompts, which is a limitation of the paper that we could not address. Overall, however, a consistent pattern across the three models is that users frequently express restaurant demand through dimensions such as feature/vibe and occasion, which are difficult to specify through conventional search interfaces.

Table G1 User Behavior by Demand Facet, by AI Family

	Share of prompts (%)		
Facet	GPT-5.1	Gemini	Claude
Feature / vibe	90.1	91.3	76.6
Neighborhood	91.3	85.3	50.8
Time	35.0	38.7	19.5
Occasion	15.6	20.4	37.8
Landmark	33.8	22.8	11.1
Borough	13.8	13.8	16.8
Price	29.6	13.5	15.3
Cuisine	14.7	33.9	15.0
Dish	5.4	27.9	10.5
Party size	38.3	12.6	24.6
Mean # facets	3.68	3.60	2.78
Prompts	334	333	333

Notes: Each share column reports the percentage of prompts generated by the corresponding assistant (334 for GPT-5.1, 333 for Gemini, and 333 for Claude) that mention the named facet. All prompts are classified using the same ten-facet GPT-5.1 tagging procedure described in Section 6.1.1. "Mean # facets" is the average number of facets mentioned per prompt.

Appendix H: Most Frequently Used Keywords in Real-User Prompts by Facet

Table H1 reports the number of distinct keywords in the facet and the most frequently mentioned terms for each of the ten search facets that emerge from the 1,000-prompt corpus (Section 6.1.1). Open-vocabulary facets—feature/vibe, time, occasion—are highly granular, spanning dozens to hundreds of distinct terms, whereas structured facets such as price and borough draw on a small closed set. This is what makes natural-language requests hard for a filter-and-tag interface: most of the descriptive content lives in facets whose vocabulary is long-tailed and free-form.

Table H1 Most Frequent Keywords by Facet (1,000 Realistic Prompts)

Facet	# distinct	Most frequent keywords (# prompts)
Feature / vibe	661	cozy (59), quiet (58), good for groups (52), casual (50), bar seating (45), family-friendly (44), not touristy (33), lively (32)
Neighborhood	73	east village (65), lower east side (62), upper west side (55), midtown west (51), soho (46), west village (46), williamsburg (43), chelsea (38)
Time	125	tonight (64), friday (13), saturday (12), open late (12), thursday (11), 6:00 pm (10), 8:00 pm (9), friday night (8)
Occasion	78	date night (36), birthday (27), pre-theater (22), business dinner (17), celebration (15), anniversary (12), date (10), special occasion (10)
Party size	46	2 (67), 4 (34), 3 (26), 8 (19), 5 (16), 6 (11), 10 (8), 12 (5)
Landmark	65	madison square garden (23), penn station (20), times square (18), the met (18), lincoln center (18), barclays center (13), grand central (8), 9/11 memorial (7)
Price	4	$$ (110), $$$ (46), $$$$ (27), $ (22)
Cuisine	55	italian (33), american (20), seafood (16), french (16), greek (16), mediterranean (16), steakhouse (15), korean bbq (12)
Dish	50	pasta (34), ramen (25), noodles (13), dumplings (11), burger (10), steak (9), tacos (9), hot pot (7)
Borough	12	manhattan (47), brooklyn (32), queens (29), new york city (19), lower manhattan (9), downtown manhattan (7), bronx (4), anywhere in the city (3)

Notes: "# distinct" is the number of distinct keywords tagged to the facet across the 1,000 prompts; the last column lists the eight most frequent keywords, with the number of prompts mentioning each in parentheses. Price is normalized to $–$$$$.

Appendix I: Comparison of Restaurant Characteristics in ChatGPT and OpenTable Recommendations

I.1. Review Characteristics of Restaurants Recommended by ChatGPT and OpenTable

To complement the graphical evidence in Figure 5, Table I1 formally tests for differences in the means and medians of review characteristics between restaurants recommended by ChatGPT and those recommended by OpenTable. We pool recommendations across the 200 questions, five replicates, and three prompt types, restricting the comparison to restaurants in the estimation panel. Restaurants are weighted by how often each system recommends them. The table reports means and medians for review volume, ratings, and average review length, together with tests of differences between the two systems.

OpenTable recommends restaurants with higher review volume: the mean and median of log review count are both higher than for ChatGPT's recommendations. Because the outcome is in logs, the gaps in Table I1 are larger than the raw numbers suggest. The difference in means of 0.156 log points implies that the restaurants OpenTable recommends have roughly 17% more Google reviews ($e^{0.156} - 1 \approx 0.17$), and the difference in medians of 0.211 log points implies about 24% more at the median—approximately 1,100 versus 890 reviews. By contrast, ChatGPT recommends restaurants with longer reviews, with higher values of both the mean and median of log average review length; the differences of 0.046 and 0.042 log points correspond to reviews that are about 4–5% longer on average. All four differences are statistically significant at the 1% level. Ratings are very similar across the two systems: both medians are 4.500, and the mean difference of 0.009 points on a five-point scale (roughly 0.2% of the mean) is significant only at the 10% level. These comparisons suggest that ChatGPT's recommendations differ from OpenTable's primarily in review volume and, to a lesser extent, review length, with no difference in ratings.

Table I1 Restaurants Recommended by ChatGPT versus OpenTable

	Mean			Median		
	ChatGPT	OpenTable	Welch p	ChatGPT	OpenTable	Median p
$\log(1 + \textit{Google_Review_Count})$	6.705	6.861	$< 0.001^{***}$	6.796	7.007	$< 0.001^{***}$
Rating	4.475	4.466	0.092^{*}	4.500	4.500	0.162
$\log(\textit{Avg_Review_Length})$	5.417	5.371	$< 0.001^{***}$	5.385	5.343	$< 0.001^{***}$

Notes: The two columns are the restaurants ChatGPT returned and the restaurants OpenTable returned, pooling the 200 questions, their five replicates and the three prompt types, and restricted to the 1,171-restaurant estimation panel. A restaurant contributes one observation for each prompt–replicate in which that system returned it, giving 5,510 observations on the ChatGPT side and 2,220 on the OpenTable side, from 734 and 333 distinct restaurants. The 276 restaurants both systems returned appear on both sides, so the two samples are not independent; the positive correlation this induces means the reported standard errors overstate rather than understate the true ones. "Welch p" is the two-sided unequal-variance t-test on the difference in means; "Median p" is Mood's median test. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Appendix J: Additional Evidence on AI–Platform Recommendation Overlap

This appendix supplements Table 12 with two analyses. First, we pool recommendations across the five replicates of each prompt to assess whether the overlap patterns persist when recommendations are aggregated across booking slots. Second, we restrict the comparison to prompts or prompt–replicate observations for which OpenTable returns at least one restaurant, separating the platform's ability to return results from the similarity of the two systems' recommendations when it does.

Table J1 reports overlap after pooling each system's recommendations across the five replicates within each prompt. The ordering across prompt types is unchanged: mean Jaccard overlap is highest for generic prompts (0.147), followed by detailed prompts (0.018) and realistic user prompts (0.008).

Table J1 AI–Platform Recommendation Overlap by Prompt Type: Prompt-Level Aggregation

		Distinct restaurants returned		Mean per prompt			
Family	N	AI	Platform	AI	Platform	AI ∩ Platform	Jaccard
Generic	50	825	428	26.42	13.36	4.40	0.147
Detailed	50	652	201	33.26	4.72	0.64	0.018
Real-user	100	1,160	132	28.67	1.69	0.28	0.008

Notes: Prompt types are defined in Table 12. N counts prompts. Each prompt has five replicates that differ only in future booking slot. Each system's recommendations are pooled across these replicates before computing the prompt-level statistics. *Distinct restaurants returned* is the type-level union across all prompts and replicates. Under *Mean per prompt*, *AI* and *Platform* are the mean sizes of the two pooled recommendation sets, *AI* ∩ *Platform* is their mean intersection, and *Jaccard* is the mean prompt-level ratio $|A \cap P|/|A \cup P|$ (not a ratio of column means). Prompts with no platform result in any replicate are retained with Jaccard zero. Table J2 reports results conditional on a nonempty platform result.

The unconditional comparisons assign zero overlap to queries for which OpenTable returns no restaurants. Table J2 reports overlap conditional on a nonempty platform result, alongside the number of units satisfying this condition. OpenTable returns at least one restaurant for 49 of 50 generic prompts, 33 of 50 detailed prompts, and 16 of 100 realistic user prompts. Generic prompts continue to exhibit the highest overlap. For realistic user prompts, mean Jaccard overlap rises from 0.008 to 0.050 at the prompt level and from 0.006 to 0.040 at the prompt–replicate level, exceeding the corresponding values for detailed prompts (0.027 and 0.022). Thus, the particularly low unconditional overlap for realistic user prompts partly reflects the platform's frequent failure to return results, although overlap remains low even when it does.

Table J2 Recommendation Overlap, Units Where the Platform Returned at Least One Restaurant

Family	N	Returns a result	AI	Platform	AI ∩ Platform	Jaccard
Panel A. Prompt level (N = prompts; five replicates pooled on each side)						
Generic	50	49	26.43	13.63	4.49	0.150
Detailed	50	33	33.36	7.15	0.97	0.027
Real-user	100	16	29.81	10.56	1.75	0.050
Panel B. Observation level (N = prompt–replicate rows)						
Generic	250	245	9.96	9.50	1.74	0.108
Detailed	250	165	9.98	5.39	0.35	0.022
Real-user	500	80	9.99	8.44	0.70	0.040

Notes: Columns are defined as in Table 12, with one addition. "Returns a result" is the number of units, out of the N in that row, for which the platform's own search returned at least one restaurant—in Panel A, a prompt counts if the platform returned a restaurant in at least one of its five replicates. *AI*, *Platform*, *AI* ∩ *Platform*, and *Jaccard* are computed over those units only, so each row's denominator is its "Returns a result" count rather than N. Panel A of Table 12 reports the same statistic computed over all N units.

Appendix K: Examples of Dining Requests in Realistic User Prompts

Table K1 presents ten prompts drawn at random without replacement from the 100 realistic user prompts used in the recommendation comparison in Section 6.2. These model-generated prompts illustrate the kinds of dining requests the corpus is designed to capture. They express not only cuisine and location preferences but also the purpose of a meal, desired atmosphere, dietary needs, budget, and practical constraints. For example, requests for a skyline-view date, a first dinner with a partner's parents, or a quick meal before a show combine restaurant attributes with the context in which the meal will take place.

Table K1 A Random Sample of Realistic User Dining Requests

No.	Prompt
1	want a date night with an actual skyline view, splurge-worthy, brooklyn or manhattan doesn't matter
2	it's pouring rain and freezing. where can my husband and i get cozy comforting food (maybe ramen or hot pot) in chinatown tonight where we won't have to stand outside in line forever?
3	meeting my boyfriend's parents for the first time tomorrow night, need somewhere nice but not intimidating, midtown east area
4	Im seeing a show at 8pm tonight near times square and need a quick solo dinner before. where can i get a really good meal sitting at the bar without a reservation around 6:30?
5	want a proper old school italian red sauce joint, like the kind with checkered tablecloths, does that still exist in nyc
6	where can i get a good steak in the financial district that wont bankrupt me lol
7	need a spot in downtown Brooklyn or Dumbo for dinner with my boss and her husband, kind of classy but not white-tablecloth stiff, she doesn't eat red meat
8	vegetarian but my friend is a picky eater who basically eats pasta and pizza only, need dinner that works for both of us in the east village
9	Need a place for 8 people in Midtown East for my cousin's graduation dinner. A couple of teens, a couple of grandparents, so menu has to be pretty broad. Ideally they take reservations for a group and don't rush us out.
10	My friend who is a major foodie is visiting from LA. I need to take her somewhere cool with great natural wine and sharing plates, maybe Greenpoint or Williamsburg. Somewhere with a really good aesthetic.

Notes: The ten prompts are sampled without replacement from the 100 distinct realistic user prompts in the recommendation experiment, using random seed 20261002. These prompts originate from the model-generated corpus described in Section 6.1.1; they are not observed user queries. Original wording is preserved before adaptation to the experimental booking slots. Sampling does not condition on either system's recommendations or their overlap.