red!50!black[1]

[1]We thank Manuela Collis, Hong Deng, Anindya Ghose, John Lalor, Shijie Lu, Alex Moehring, Christian Peukert, Mohammad Rahman, Yoonseock Son, Steve Tadelis, Sonny Tambe, Yun (Alicia) Wang, Joy Wu, Shuang Zheng and workshop and seminar participants at University of Notre Dame (Marketing), Carnegie Mellon University (Tepper School of Business), University of Pittsburgh, UC Irvine (Merage School of Business), Nanyang Business School, NYU AI in Strategic Management Workshop, NSF Convergence Workshop on Human-AI Frontier, NBER Digital Economics Spring Meeting, China India Insights Conference, Conference on Information Systems and Technology, Marketing Dynamics Conference, Workshop on Information Systems and Economics, Hi! PARIS Workshop on AI and Digital Economy, Online Platform Seminar Series, and Columbia/Wharton Management, Analytics, and Data Conference for helpful suggestions.

The Value of External Data for Digital Platforms: Evidence from a Field Experiment on Search Suggestions

Xiaoxia Lei
Shanghai Jiao Tong University[†]

Yixing Chen
University of Notre Dame[‡]

Ananya Sen
Carnegie Mellon University[§]

July 2023

Abstract

Firms increasingly leverage external data with an aim to unlock improvements in their offerings, but it is challenging to measure the value of external data. Collaborating with a technology company in China, we analyze a field experiment where we manipulated access to a leading search engine's application programming interface (API) to measure the causal impact of such data access on the click-through rate (CTR) of the focal company's search suggestions. We report three main findings: First, compared to the baseline with access to the API, API removal leads to a 4.6% decrease in the average CTR of search suggestions. Second, the negative effect due to API removal is more prevalent among heavy users, and it is driven by both mainstream and niche content. Third, the negative effect becomes less negative over time with the absolute magnitude in the longer term being half as much as what we would have obtained with a short-term experiment. We provide suggestive mechanism evidence of the longer-term effect: the focal company's reliance on the leading search engine's data limits the development of its algorithmic system based on its internal data. This research informs managers of whether and how the market leader's data affects a smaller player's product performance. It further sheds light on policies such as the Digital Markets Act that proposes data sharing by large digital platforms, as well as a recent debate on whether big data undermines market competition.

[†]Antai College of Economics and Management, Shanghai Jiao Tong University, dr.xiaoxia.lei@gmail.com.
[‡]Mendoza College of Business, University of Notre Dame, ychen43@nd.edu.
[§]Heinz College, Carnegie Mellon University, ananyase@andrew.cmu.edu.

1 Introduction

Data is considered the 'new oil' for digital platforms (Economist, 2017). Digital platforms collect large amounts of data, such as their users' digital footprints, and use it as a key input to develop algorithms that provide personalized recommendations (Sun et al., 2021). Increasingly, firms have tapped into 'external data', digital representation of acts, facts or information provided by external providers through data-sharing agreements, to unlock improvements in products and services.[1] A prevalent practice is that firms use Application Programming Interfaces (APIs) to gain access to external data from large players in the marketplace (Xue et al., 2019; Benzell et al., 2022). In the context of search, for example, Google custom search JSON API enables publishers and developers to retrieve response data from Google (e.g., metadata describing the requested search, search results).[2] Despite wide market prevalence and potential economic importance, there is limited evidence on the value of such external data because it is challenging to pin down its causal impact on the recipients' product performance. Measuring the causal impact of external data is important, not only from a managerial perspective, but also from a policy perspective. The Digital Markets Act that came into effect in November 2022 requires gatekeeper platforms that provide core services (e.g., search engines, web browsers) to open up for smaller players with data access. In particular, the Digital Markets Act mandates gatekeepers to provide smaller players with access to depersonalized search results to level the playing field.[3] Similarly, the guideline on building basic systems for data released by the State Council in China in December 2022 proposes a key measure on designing privacy-preserving data-sharing mechanisms to enable the growth of small and medium-sized companies.[4]

In this paper, we analyze a large-scale field experiment to measure the extent to which the *removal* of depersonalized candidates for search suggestions (external data) shared by a leading search engine (i.e., the market leader) affects the performance of our partner company's search

[1] See the survey report by Deloitte here https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html.
[2] https://developers.google.com/custom-search/v1/introduction.
[3] See details about the proposal of for a regulation on DMA: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en#documents.
[4] http://english.www.gov.cn/policies/policywatch/202212/21/content_WS63a264e0c6d0a757729e4a23.html.(English) http://www.gov.cn/zhengce/2022-12/19/content_5732695.htm.(Chinese)

suggestions (i.e., a smaller player; hereafter, "the company"). We further explore which and how users respond to the removal of such external data, as well as how users respond to the experimental variation in the longer term. The primary outcome of interest is click-through rate (CTR), a key performance indicator of search success, online advertising effectiveness, and online behavior at large. The company made its initial foray into the search market by launching search suggestions, an important initial application of generative artificial intelligence models (Serban et al., 2016; Kucharavy et al., 2023).[5] Similar to the Google autocomplete predictions,[6] search suggestions are generated by the underlying algorithm that uses multiple data sources to predict what users want to click in response to their queries. Providing users with relevant suggestions through query autocomplete is fundamental to user experience, so large players devote significant resources to the development of search suggestions (Agrawal et al., 2018). As an entrant, the company leveraged the market leader's autocomplete API to retrieve candidate suggestions as an external data input to improve its search suggestions. We partnered with the search product team to experiment with access to the market leader's autocomplete API while using the same algorithm to generate search suggestions, providing an ideal context to measure the value of the market leader's data.

The field experiment lasted 108 days and involved more than 2.3 million users among which each user was randomly assigned to one of two conditions. In the control condition (the status quo), the company's ranking algorithm ranks search-suggestion candidates retrieved from the market leader's autocomplete API along with its own search-suggestion candidates and generates a final list of search suggestions in response to user-submitted queries. These candidates retrieved from the market leader's autocomplete API is what we refer to as 'external data' in the paper.[7] The queries to the market leader's API are depersonalized because the company does not provide user-specific characteristics (location, browser, device type, etc.) in those API calls. In the treatment condition, we remove access to the API: i.e., the company's ranking algorithm does not have access to the market leader's search-suggestion candidates and only ranks its own candidates. This research

[5] Indeed, recent applications of generative artificial intelligence (e.g., ChatGPT) are referred to as "autocomplete for everything". For more details, see https://www.noahpinion.blog/p/generative-ai-autocomplete-for-everything.
[6] https://blog.google/products/search/how-google-autocomplete-works-search/.
[7] We use terms such as external data, the market leader's data, or data (input) from the market leader's API interchangeably throughout the paper.

design has two notable features. First, the manipulation is changing the supply of candidates at the ranking stage of the search-suggestion generating process, while holding the ranking algorithm and user interface constant. Therefore, the random assignment among millions of users enables us to precisely estimate the causal impact of access to the market leader's data. Second, maintaining a consistent treatment for 108 days enables us to estimate longer-term effects of access to the market leader's data. Third, this context captures prevalent data-sharing agreements in the market which are touted to have significant benefits for companies. Moreover, this is a policy-relevant setup aiming to quantify the impact of the provision of search-suggestion candidates as external data falling within the broader agenda of the Digital Markets Act, which states that smaller players in the search context should be provided with "access ... to ranking, query, click and view data".

We report three sets of findings. First, relative to the status quo, the removal of the market leader's API leads to a 4.6% ($\pm$0.3%) decline in average CTR, highlighting the value relevance of the market leader's data access. We benchmark the magnitude of the average treatment effect with existing studies in the literature and industry reports and find it economically meaningful.

Second, we document who are more (less) responsive to the removal of the market leader's data access and the conditions under which the market leader's data affects users. The negative impact of removing the market leader's data access is stronger for heavy users relative to light users, suggesting that such data access is more effective in helping in user retention by streamlining heavy users' experience than in ameliorating the cold-start problem for light users. Next, importantly, we find that the negative treatment effect is consistent for both mainstream and niche topics and that API removal reduces the coverage of topics searched. This result suggests that despite its depersonalized nature, the market leader's data is valuable by covering both mainstream content and long-tail information, which is hard to source and is crucial for product growth.

Third, the magnitude of treatment effect estimates varies significantly within the 16-week experimental period. It starts at about a 8.1%–9% decline in the first 2–3 weeks, and becomes less negative at 3.6%–4.5% in the last 3–4 weeks. These estimates suggest that, had we run a short-term experiment, we could have overestimated the value of the market leader's data access (the negative effect of API removal) by a factor of 2. Notably, the magnitude of treatment effects decreases over

the course of the experiment in absolute terms, even among new users who had no prior exposure to the search product. This mechanism evidence suggests that in the absence of the market leader's API, there has been a gradual improvement over time in the company's algorithmic recommendations because of the development of its internal data. This creates a trade-off between the short-run benefits of using the external data source and longer-term growth through enriching internal data. We further verify that this data pattern is not driven by differential attrition across two conditions because we do not find a significant treatment effect on the overall usage of search suggestions.

Our findings offer clear, practical implications for managers and policymakers. First, we inform managers of the potential economic implications and trade-offs of tapping into different external data levers. Our analysis provides empirical evidence suggestive of the potential impact of larger players removing access to their API, as was the case with Google's recent decision to restrict smaller players' access to its Autocomplete API.[8] Second, our findings inform policymakers in particular with regards to the Digital Markets Act. Our research provides novel causal evidence that the market leader's data does enhance a smaller player's product performance. We highlight a critical trade-off where the extensive reliance on external data can limit the organic development of the internal data, which can be used to improve the focal algorithm in the longer term. Our context is important from a policy perspective since regulation has focused on the need for large digital platforms/gatekeepers to share depersonalized search results with newer competitors. More generally, we provide empirical evidence that informs a recent debate on whether big data undermines market competition (Tucker, 2019; Crémer et al., 2019; Krämer et al., 2020). Lastly, more broadly, understanding the value of data sharing in search contexts is becoming increasingly important, especially with recent innovations of generative models and user-facing applications in search engine and web browsers (e.g., OpenAI large language model for Bing search engine and Edge browser).

Our paper contributes to three strands of academic literature. First, we contribute to the literature that aims to quantify the value (of different dimensions) of data for search engines and related online products. Yoganarasimhan (2020) analyzes how utilization of user-level data can help in personalization of search and quantifies significant returns to personal data. Chiou and Tucker (2017)

[8] https://developers.google.com/search/blog/2015/07/update-on-autocomplete-api.

analyze the efficacy of search recommendations when companies such as Yahoo! and Microsoft reduced the amount of individual-level data retention to a 90-day period and find no change in a user's CTR. Valavi et al. (2020) quantify the dynamic value of data by using platform data over time in a next word prediction task, similar to our search suggestions setting, with Reddit as their context. We augment this strand of research by analyzing how data access impacts search product performance using a large-scale, long-term field experiment. We demonstrate that depersonalized data inputs through third party API access could help search products grow successfully, at least in the short run, by catering to user preferences to both mainstream and long tail content.

Second, we contribute to a strand of literature that analyzes the impact of utilizing several forms of 'external data' on firm outcomes. These external sources can be varied. Beraja et al. (2022) show that access to government data, in the form of surveillance videos, lead to AI innovation in the facial recognition industry by private companies in China. Similarly, Nagaraj (2022) analyzes the private impact of public data in the Gold industry. Nagaraj (2022) finds that access to publicly available Landsat, a U.S. National Aeronautics and Space Administration satellite mapping program, lead to more gold discoveries especially by new entrants. In the case of online platforms, Neumann et al. (2019) highlight how third-party data from data brokers has become pervasive for online ad targeting and analyze the quality of such widespread third-party data. Chan et al. (2022) show that access to employer-verified employment data via Equifax benefits both auto loan borrowers and lenders. Wernerfelt et al. (2022) leverage a field experiment to show that a loss of off-platform cookie data increases the customer acquisition cost for advertisers on Meta. Given our experimental setup, we can cleanly identify effects along with heterogeneity within a policy-relevant search context of data-sharing through the API of the market leader. More generally, we focus on a widely used, source of external data through third-party APIs for online platforms, providing implications for other companies that aim to utilize such data for their product growth.

Finally, more broadly, we contribute to the literature that analyzes different strategies for platform growth. Benzell et al. (2022) use aggregate data on online platforms to demonstrate how a firm can grow by opening itself up to third party complementors using APIs. Sun et al. (2021) simulate a privacy regulation through a field experiment on Alibaba to quantify the value of individual-level

data for users of the platform. They find significant negative effects on engagement and purchase when personal data is not used for product recommendations, especially for niche products. Relatedly, Claussen et al. (2019) quantify the value of personal data (or lack thereof) for an algorithm relative to human experts. Klein et al. (2022) demonstrate the need for large players to share user-generated data to improve the performance of competing search products. Relative to these papers, we take a middle path and analyze a privacy-preserving situation where depersonalized data from other (external) sources could be used as an input for algorithmic recommendations. Within this context, we highlight a trade-off where external data can help product performance initially but over-reliance in the longer term can inhibit focal algorithmic development based on internal data.

2 Empirical Setting

In this section, we describe the empirical setting with a focus on (1) search-suggestion generating process through the lens of the 'algorithmic funnel', (2) the role of the market leader's API in such a process, and (3) our choice of target metric that measures the performance of search suggestions.

We partnered with a large technology company in China that develops a web browser app. With more than 200 million monthly active users, the app offers its users various in-app products such as news feed, search engine, and video and eBooks streaming. The main search product is search suggestions, which was first launched in 2020. By design, search suggestions are predictive in nature and function as query autocomplete, which is similar to Google autocomplete feature offered by Google Chrome. Figure 1 provides an illustration of the app interface. When users enter a specific query term into the search bar (e.g., covid), they typically see a list of ten search suggestions in the form of phrases and/or sentences in response to the query (e.g., "What does a covid-19 Ag mean?"). Such a search suggestion setting is considered to be an early-stage application of generative artificial intelligence models. Generative models produce system responses that are autonomously generated word-by-word, which open up the possibility for realistic, flexible interactions (Serban et al., 2016).[9] A user can choose to click a matched suggestion, press the search button without adopting search

[9]Recent applications, such as generative pre-trained transformer (GPT) models, provide those flexible interactions.

suggestions, delete this query and enter a new one, or quit the session. Search suggestions and related contexts of predictive text are economically important with companies investing significant resources (Agrawal et al., 2018). Search suggestions bridge the gap between users' search intent and content consumption, helping streamline the search process.[10] Optimizing search suggestions is important from the perspective of our partner company because they intend to monetize this process. In particular, their medium-term aim is to earn revenue from each click that is made on search suggestions in the same way sponsored links generate revenue for a search engine.[11]

Figure 1: An Illustration of Search Suggestions



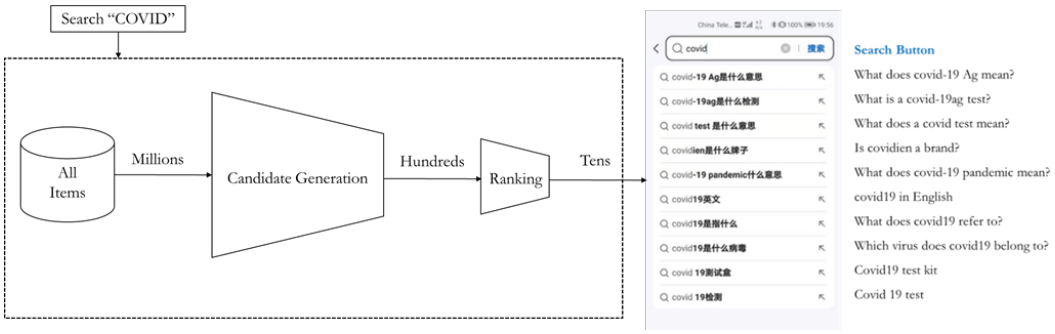
How does the company generate search suggestions in response to user-submitted queries? At a high level, the company uses a proprietary algorithmic funnel, a three-stage architecture similar to a recommender system (Covington et al., 2016): (1) item generation, (2) candidate generation, and (3) ranking. Figure 2 provides an illustration. At the *item generation* stage, the company first builds raw items based on (a) internal sources in the news feed (e.g., articles and videos) and query terms from users' active search histories and (b) external sources such as public trending news. Next, the company uses natural language processing techniques (e.g., keyword extraction,

[10] Such suggestions can also influence decisions of individuals in financial markets (Rubin and Rubin, 2021).

[11] Indeed, while our experiment was running, the team started deliberating about 'direct reach cards' (similar to Bing Snapshot Search and Google's Knowledge Graph). Direct reach cards would appear as a box with the instant information above search-suggestions results and lead the user straight to a website: i.e., such information represents factual summaries related to the queries (e.g., brand logo along with direct access to the the website or browsing and downloading the app). This is an important step in line with their goals of monetizing search suggestions.

text summarization) to transform raw items into phrases and sentences. As a result, this large item base contains millions of phrases and sentences that are filtered at the next stage. At the *candidate generation* stage, a subset of candidate items relevant to the submitted query are retrieved from the item base based on their popularity on the platform (e.g., common and trending searches). As is standard practice in such a context, this stage does not utilize personal data because of the need to filter through millions of data points (Mitra and Craswell, 2015). As a result, hundreds of candidate phrases and sentences are selected to enter the ranking stage. At the *ranking* stage, there are several steps of retrieving a larger number of features and the use of a pre-trained algorithm. Eventually, a ranking algorithm scores each candidate item according to its predicted click-through rate, which is a function of user (e.g., location, search histories) and query (e.g., topic, freshness) features. The highest scoring items are presented in a ranked order in the final list to users. The literature shows that sophisticated algorithms are utilized to handle a larger feature set at the ranking stage and rank fewer items. This contrasts with the candidate generation stage that uses generic models (e.g. logistic regression) and considers a larger set of items focusing on efficiently pruning duplicate and irrelevant items (Covington et al., 2016; Nandy et al., 2021).

Figure 2: An Overview of the Algorithmic Funnel



However, a key challenge of developing a new search product, such as search suggestions, is the lack of candidate items with respect to both quantity and quality. For example, at the initial development stage, the company's algorithmic funnel can generate only a short list of search suggestions that could match users' search intent. To address this problem, the company had started to leverage data via access to the autocomplete API of a leading search engine in China (i.e., the market

leader's API). The company's hypothesis is that because of its ability, due to their large user-base, to present relevant, timely results to of millions of users at scale, the market leader can (1) provide the company with access to more (precise) candidate items with respect to their popularity and freshness, and (2) expand the scope of the company's candidate items with respect to their topic coverage, both of which can be used as high-quality data inputs in the ranking stage to generate a list of search suggestions.

How does the market leader's API work? At a high level, the company and the market leader form a business-to-business contractual relationship where the company acquires a license to the market leader's API. To initiate the data request, the company provides the market leader's API with its user-submitted queries in real time, which are used by the market leader's API to return a set of candidate items. Importantly, these candidate items are depersonalized by design because the company provides the market leader's API with only query terms rather than users' personal data.[12] And then, at the ranking stage, the algorithm uses (1) candidates items from the market leader's API as an additional input along with (2) its own candidate items to generate a list of search suggestions in a ranked order. The company and market leader have a revenue-sharing agreement where the company pays service fees to the market leader at a undisclosed rate, the market leader compensates the company through the volume of search query terms (see Figure A1 in the Appendix).[13]

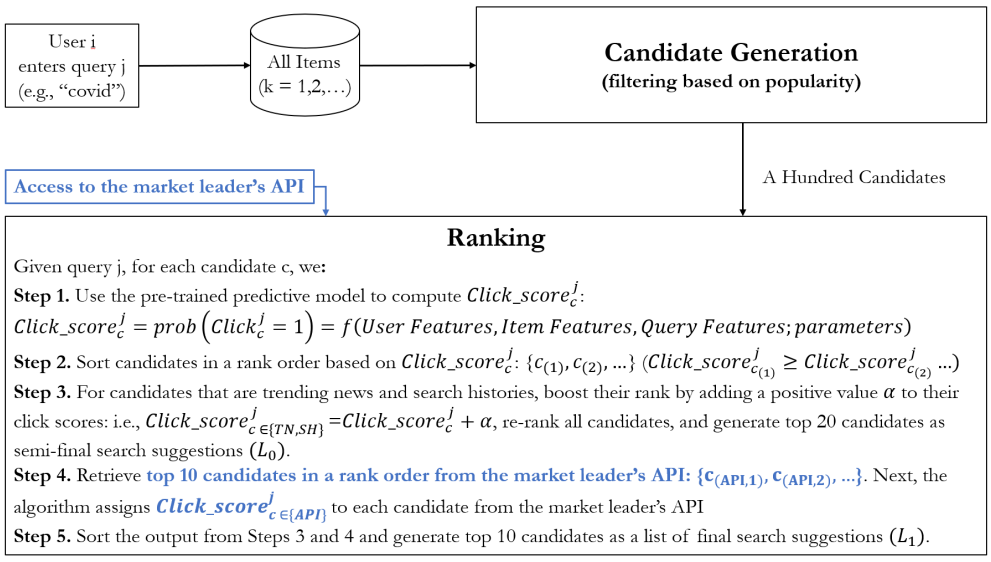
Concretely, building on Figure 2 and the discussion above, Figure 3 decomposes the entire process where the company's ranking algorithm meets the market leader's API. To fix ideas, consider the user i who enters the search query j into the search box (e.g, covid). First, the company's candidate generation model filters through millions of items and retrieves about 100 relevant candidate items from its item base (items indexed by k=1,2,3,...,K), associated with query j based on

[12] Our understanding is that the query terms are depersonalized due to privacy concerns, and there have been more discussions on regulating online platforms due to how they deal with personal data.

[13] It is pertinent to note that neither we as researchers, nor our partner company observes any details of the market leader's database. Hence it is a 'black box' for us. This elevates the need for our field experiment to understand the economic implications of such data inputs along various dimensions.

their popularity.[14] Second, the ranking algorithm scores each candidate item based on user (e.g., location, search histories), query (e.g., topic, freshness), item features, and interactions among such features and sort them according to their predicted click score $Click_Score_c^j$ (i.e., Steps 1–3 under **Ranking**). Third, given query j, another set of depersonalized candidate items is retrieved from the market leader's API in a ranked order ($c_{API(1)}$, $c_{API(2)}$...). Fourth, for each candidate item from the market leader's API, the ranking algorithm assigns a score $Click_Score_{c,API}^j$ (i.e., Step 4 under **Ranking**). Lastly, using standard tools for "crossing streams" in such settings,[15] the ranking algorithm re-ranks the mix of candidate items from two sources to generate a list of highest-scoring items as search suggestions (i.e., Step 5 under **Ranking**). Hence, access to the market leader's API changes the supply of candidate items entering into the ranking stage. Overall, the algorithmic architecture used by our partner company mirrors existing industry standards.

Figure 3: Search-Suggestion Generating Process: Decomposing the Algorithmic Funnel



[14] Indeed, retrieving semantically relevant items based on current and past popularity at the candidate-generation stage using simple models to filter irrelevant and duplicate items is a common principle across companies. For example, the Twitter also uses a similar architecture including logistic models to filter items at the candidate stage. See https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

[15] For details about Kafka tools used in such contexts to combine or "mix" output from different sources, see https://kafka.apache.org/28/documentation/streams/developer-guide/dsl-api.html#joining and https://www.confluent.io/blog/crossing-streams-joins-apache-kafka/.

Notably, we believe this setup reflects how one level of data-sharing agreements between companies could look in practice and provides a context to analyze the economic value of data sharing. First, at the time of the field experiment, the market leader's API was indeed available to other companies and developers in the market. Second, this setup mirrors how Google custom search API and autocomplete API provide developers with 'response data' that can be incorporated as an input into improving their product (Zaveri et al., 2017; Alrashed et al., 2020). Given its prevalence, external data retrieved from the market leader's API as input into the ranking stage makes it a key data-sharing context to analyze its economic value. Third, this context provides an example of how gatekeepers could provide smaller players or startups with "access on fair, reasonable and non-discriminatory terms to ranking, query, click and view data," if mandated within the framework of the Digital Markets Act. The details of the Digital Markets Act and similar regulations are still being finalized, so our analysis can be seen as providing information on the practicalities and potential benefits of data sharing.

How do we evaluate the performance of search suggestions? Our conversations with the search product team revealed that the target metric is click-through rate (CTR), a widely-used key performance indicator of search success in online markets. CTR is important from the company's perspective because customer satisfaction with the search product depends on the search engine's ability to serve relevant results (Yoganarasimhan, 2020). In the context of search suggestions, CTR is measured as the ratio of number of clicks to search suggestions in a list to the number of exposures. Specifically, when a user starts typing a keyword into the search bar, this user is exposed to a list of search suggestions in this session (i.e., one exposure). If a user clicks to any search suggestion in this list, such an action will be counted as one click. To ensure that our results are not sensitive to the variation in exposures, we conduct robustness checks using alternative measures such as the number of clicks and the probability of any click. We can measure CTR for each user on a daily basis or over a longer period (e.g., week). To capture the overall user activities, rather than a user's specific search session, we focus on the *aggregate* CTR for each user: the ratio of the total number of clicks to total number of exposures over the same period of time.

However, measuring the value of the market leader's data (candidate items) on target metrics,

such as CTR, is challenging without exogenous variation on this dimension. To circumvent this challenge, we conducted a large-scale field experiment where we manipulated access to the market leader's API while holding all other aspects (e.g., algorithms, user interface) equal. Therefore, we can cleanly identify the impact of access to the market leader's data on search-suggestion CTR.

3 Field Experiment

3.1 Experimental Design

The randomization was implemented at the user level: when users started typing keywords into the search bar in a given time point during the experiment, they were randomly assigned to one of two conditions. Once a user was assigned to a condition, this user stayed in the same condition until the experiment ended. We maintained a consistent treatment assignment for 108 days (about 16 weeks) from May 2021 to September 2021.[16]

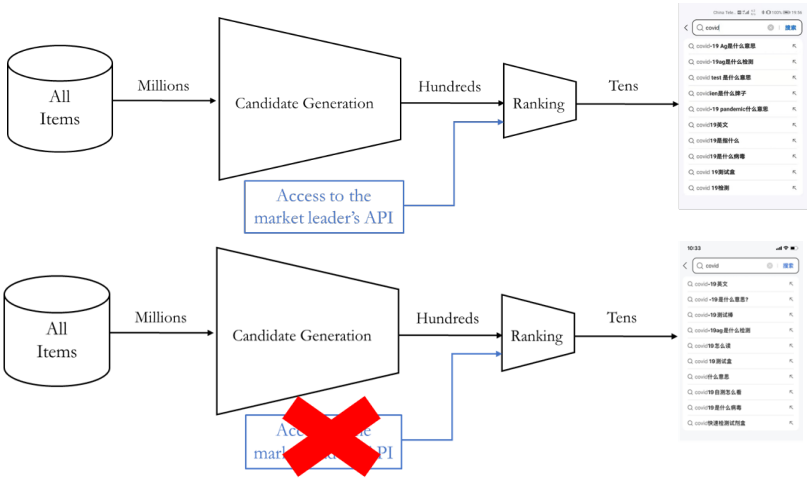
Control (N=1,194,619): In response to user-submitted queries, search suggestions are generated by the company's proprietary algorithm funnel, including item generation, candidate generation, and ranking. In the control condition, at the ranking stage, the ranking algorithm scores candidate items from two data sources: including (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. Next, the ranking algorithm generates the highest scoring items in a ranked list. As a result, users in this condition see a sizable proportion of search suggestions supplied by the market leader even though the number of search suggestions from the market leader might differ across users depending on queries, demographics, and search histories.

Treatment (N=1,195,625): In response to user-submitted queries, search suggestions are generated by the same proprietary algorithm funnel, which does not rely on the market leader's API at the ranking stage. The ranking algorithm (the same as the one used in the control condition) scores only candidate items retrieved from the candidate generation stage and generates the highest scoring items in a ranked list (i.e., there is no Step 4 in Figure 3). As a result, users in this condition

[16] The company ensured that users in this experiment did not overlap with the users in any other experiment conducted simultaneously on the platform. Like other major technology companies, the company uses the standard design of overlapping experiments to run experiments simultaneously and efficiently (e.g., Figure 2b in Tang et al. (2010)).

never see search suggestions supplied by the market leader. Figure 4 visualizes our design. Notably, our treatment does not cause any change in the user interface: i.e., there is no disclosure to the user of whether a search suggestion comes from the market leader's API.[17]

Figure 4: Experimental Design



3.2 Data

The primary data set is at the user-day level where we observe a unique user identifier, treatment status, number of exposures to search suggestions, number of clicks to search suggestions, login status, and search button usage. In addition, we observe pre-experimental characteristics, such as demographics (e.g., gender, city of residence); mobile operating system (e.g., Android); and activity level, (e.g., active days in the past 30 days). The main analysis is conducted at the user level which is the unit of randomization. The dependent variable, aggregate CTR, is computed for each user as the ratio of the total number of clicks over the entire experiment to total number of exposures over the entire experiment. We conduct several checks to ensure that the random assignment is successful. Table 1 shows that the mean difference in observables across conditions is neither economically nor statistically significant. Figures A2 and A3 show that users were equally

[17] It is also pertinent to note that any updates to the training data during the experimental period remains separate for treatment and control conditions.

likely to be assigned to either condition over the course of our experiment, regardless of whether they are new users in any given day (e.g., if a user had not used the search bar since the beginning of the experiment and did so for the first time on day t, she is considered a new user on day t).

Table 1: Randomization Checks

User Characteristics	Control	Treatment	p. value
Male	0.503	0.504	0.672
	(0.000)	(0.000)	
Larger Cities	0.505	0.505	0.425
	(0.000)	(0.000)	
Smaller Cities	0.434	0.435	0.349
	(0.000)	(0.000)	
Mobile Operating System: Apple iOS	0.107	0.107	0.736
	(0.000)	(0.000)	
Mobile Operating System: Android	0.837	0.837	0.794
	(0.000)	(0.000)	
Active days in the past 30 days (search activities)	100	99.826	0.426
	(0.005)	(0.005)	
Query views in the past 30 days (search activities)	100	99.383	0.126
	(0.046)	(0.045)	

Notes: This table shows the balance between users in the treated relative to control groups along several observable dimensions. The first two columns provide the mean of the variables with standard error in parentheses. Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, values reported in the last two rows were normalized so that the variable means in the control condition are 100.

4 Empirical Analyses

4.1 Empirical Framework

We use a potential outcome framework to describe our model. For illustration purposes, we consider an experiment with N users who are randomly assigned to one of two conditions: i.e., treatment (e.g., API removal) or control condition. For a set of independent and identically distributed users $i = 1,...,n$, we observe the outcome of interest Y_i (e.g., CTR aggregated over 108 days); treatment assignment T_i; and a vector of user characteristics Z_i (e.g., demographics, past search activities). For each user i, there are two potential outcomes: if a user is assigned to the treatment condition, we observe the outcome $Y_i = Y_{i1}$, and if the user is assigned to the control condition, we observe

$Y_i = Y_{i0}$. In theory, the average treatment effect (ATE) is $E[Y_{i1} - Y_{i0}]$, can be used to assess whether the treatment causes changes in Y_i. Alternatively, we can estimate the following regression:

$$Y_i = \alpha + \beta \times T_i + \epsilon_i, \tag{1}$$

where β captures the causal effect of the removal of the market leader's API on the outcome of interest (e.g., CTR aggregated over 108 days). Because of the successful randomization, we do not expect the controls Z_i to affect the estimate of β. Therefore, we estimate the regressions without control variables as the baseline results and use heteroskedasticity-robust standard errors.

Another important consideration is that the regression coefficient β itself does not provide a sense of the effect magnitude. Therefore, we report the estimates as *lift* to facilitate the interpretation of estimates as the magnitude and comparison across experiments (Gordon et al., 2022): the incremental CTR among treated users relative to control users as a percentage of CTR among control users. ($\frac{\overline{Y_1} - \overline{Y_0}}{\overline{Y_0}}$ or $\frac{\hat{\beta}}{\hat{\alpha}}$). A negative (positive) value of the estimated lift indicates a decrease (an increase) in CTR among treated users relative to control users. Importantly, the lift is a ratio of two random variables, making it a random variable. Therefore, we use the Delta method to derive approximations for the mean and variance estimates of the lift (Casella and Berger, 2002; Deng et al., 2018).

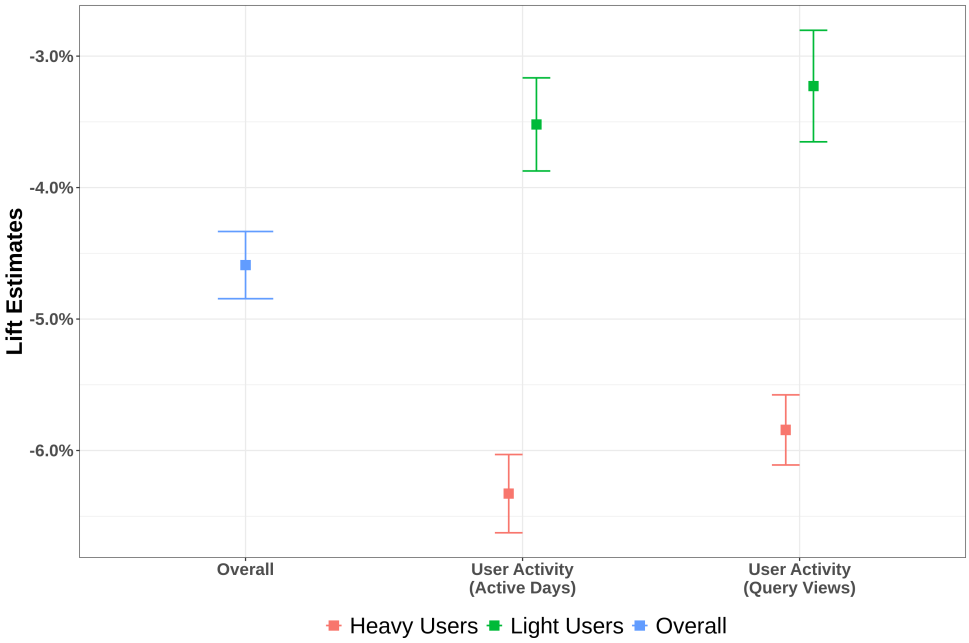
4.2 Baseline Results

Our starting point is estimating the average treatment effect. There is a significant and negative average treatment effect as seen in Figure 5: API removal leads to a decrease in CTR by 4.6% ($\pm$0.3%) in the treatment condition relative to the control condition (i.e., the average lift is –4.6% with the 95% confidence interval in parentheses). The large sample size allows us to reach considerable precision, such that the width of the 95% confidence interval is about 10% of the absolute value of the point estimate. The magnitude is economically significant, especially in the search context.[18] We can also assess the magnitude relative to other studies in the literature to

[18] See more details here Kohavi et al. (2013) and comments by Ronny Kohavi, former VP of Analytics and Experimentation at Bing, here https://www.facebook.com/watch/?v=2368597899925946.

see if its comparable and economically meaningful. The average effect of 4.6% is comparable to Berman and Israeli (2022) showing that descriptive analytics increases revenue between 4%-10% and Brynjolfsson and McElheran (2016) showing that data driven decision making increases firm performance by about 3%. Bar-Gill et al. (2021) shows that providing access to a data dashboard leads to an increase in sales by 3%. Finally, Kim (2019) demonstrates that providing access to competitor information leads companies to increase sales by 8%. Our average estimate, as well as its evolution across the four months of the experiment falls within the range of the estimates mentioned above. Indeed, these studies analyze the impact of additional data or information provided to a firm or seller which is in line with the framework of the current study.

Figure 5: Average Treatment Effect and Heterogeneous Treatment Effects by User Characteristics



Notes: Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users (e.g., the average treatment effect is –4.6%). The estimates are based on OLS regressions as in equation (1). The overall estimates are based on the full sample while estimates are based on sub-samples for heterogeneity results. Error bars represent 95% confidence intervals.

To dig into *who* is more (less) responsive to the removal of external data, we explore heterogeneous treatment effects along user characteristics. Specifically, we examine whether users with

different activity levels are more or less responsive to the market leader's data as additional candidate items into the ranking algorithm. A priori, it is unclear whether the additional data will help solve the cold start problem for light, less active users (i.e., customer acquisition) or provide a more streamlined experience for heavy, more active users (i.e., customer retention), or both. To examine this, we use the number of search-related active days in the past month prior to the experiment to differentiate irregular users from heavy users. We construct an indicator variable, heavy user, which is equal to 1 when a user's number of search-relate active days is strictly above the median and zero otherwise. Figure 5 shows that the magnitude (absolute value) of the negative effect of API removal among heavy users is 6.3% (±0.3%), which is significantly larger than that among light users, 3.5% (±0.4%). This pattern is highly consistent when we use an alternate measure: i.e., the number of query views in the past month prior to the experiment (1 if strictly above the median and zero otherwise).

These results highlight that relative to light users, users with heavier search activities in the pre-experimental period are more responsive to the removal of the market leader's data. A plausible explanation is that because heavy users experience quality recommendations more often, they may set a higher expectation of product quality and are more sensitive to changes in the quality of search suggestions due to the absence of the market leader's data. In this regard, our results complement Sun et al. (2021), where the authors find that the use of personal data in the recommendation benefits light users more when data volume and customer resilience coexist in their context.

4.3 Treatment Effects by Query Type

To dig into *how* and *why* external data affects user clicks, we explore how treatment effects vary by query type. First, we examine whether the effect of API removal varies by the clarity of users' search intent. When users enter a query term with an explicit search intent (e.g., Covid), the company's natural language processing tool is able to classify the topics of such terms into 31 first-level categories (e.g., Health) and 167 second-level categories (e.g., Health-Disease). When users enter query terms with implicit search intent, the company's natural language processing tool classifies such terms as null or other. Therefore, we construct an indicator variable, explicit search

intent, which is equal to 1 if the content category of a user's query can be explicitly labeled by the tool (e.g., Covid: Health-Disease) and 0 if it is classified as null or other.

For this analysis, we use log files from users who clicked at least once on a search suggestion across the experimental period. We compute the CTR among query terms with explicit search intent and CTR among those with implicit search intent for each user each day, aggregate them across the entire experiment at the user level, and use them as two separate dependent variables. Table 2, Columns (1) and (2) show that the negative treatment effect due to API removal is much higher when users' search intent is implicit. In other words, the market leader's data is more valuable when users' search intent is implicit. A plausible explanation is that search suggestions based on real searches by all users on the market leader's platform are more likely to capture users' interests when they are less explicit about their queries.

Table 2: Treatment Effects by Query Type

	(1)	(2)	(3)	(4)	(5)
	Explicit Intent	Implicit Intent	75%	25%	First Level
Variables	CTR Explicit	CTR Implicit	CTR Popular	CTR Niche	Unique Cat.
API Removal	-0.0071***	-0.1085***	-0.0158***	-0.0101**	-0.0032***
	(0.0015)	(0.0036)	(0.0017)	(0.0032)	(0.0005)
Observations	1,631,260	1,631,260	1,415,958	1,415,958	1,415,958

* $p<0.05$, ** $p<0.01$, *** $p<0.001$. Robust standard errors in parentheses. The sub-sample estimates are based on OLS regressions as specified in Equation (1).

Second, we examine whether the effect of API removal varies by query popularity conditional on explicit search intent. We conduct this analysis with search records of users that clicked at least once and those with explicit content categories (i.e., content categories are neither null nor other).[19] We define the content categories that generate 75% of the clicks in the control condition as popular (mainstream) content (e.g., Education-K-12, Books-Novel, Health-Disease) and the remaining 25% as niche content (e.g., Education-Study Abroad, Music-Music Radio, Government Affairs-Nonprofit Organization).[20] Columns (3) and (4) suggest that the negative treatment effect is driven by both

[19] We verified that the user demographics and usage activities are balanced among users in the treatment and control conditions in this subsample (see Table A1).

[20] These results are robust to alternate thresholds, such as 90-10 split (see Table A2 in the Appendix).

popular and niche topics. This suggests that the proprietary data from the market leader, despite being depersonalized, helps identify both mainstream and long-tail topics. This result dispels the concern that because the queries sent to the market leader's API are depersonalized, candidate suggestions generated by the market leader's API would do a better job at capturing popular content, rather than niche content.

To dig deeper into this result, we examine whether the market leader's data affects what users search with respect to content categories. To do so, we use the total number of unique content categories across the entire experiment as the dependent variable. Column (5) of Table 2 shows that API removal reduces the number of content categories searched, suggesting that the market leader's API is capable of expanding the breadth of topics that the users can search for. This result is further supported when we look at finer, second-level content categories, as seen in Column (6) of Table A2 in the Appendix. The fact that the API provides a wider set of topics and has a larger user base allows it to overcome, at least to some extent, the fact that the queries do not contain any personal information. This has implications for managers and policy makers that are looking to (privacy-preserving) data sharing regulations across companies.

4.4 Long(er)-Term Effects

A novel aspect of our design is that a consistent 16-week treatment assignment allows us to estimate longer-term effects. Following Huang et al. (2018), we estimate a regression for each week as if that week were the final one. This gives us 16 separate regression estimates. Figure 6 shows how the absolute values of the lift vary over time, with the solid (dotted) line representing the point estimates (the 95% confidence intervals). The estimates start at about a 8.1%–9%, with absolute magnitudes decreasing to 3.6%–4.5% in the last few weeks. Thus, had we run a short-term experiment, we could have overestimated the value of the market leader's data by a factor of 2.

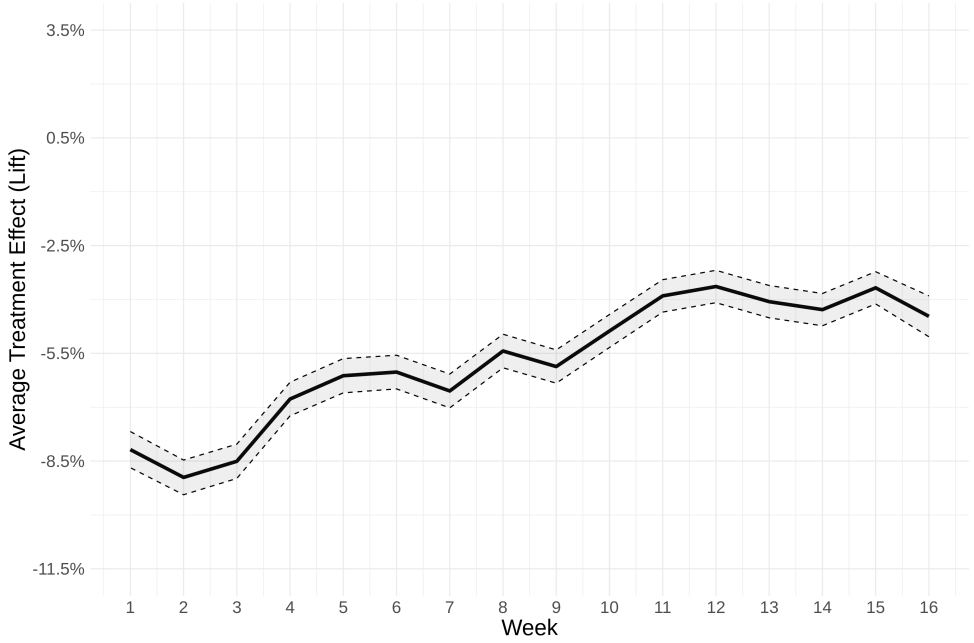
There are a couple of possible explanations for the treatment effect becoming less negative over time. First, the decline in absolute magnitude could be driven by the improvement in the company's algorithmic recommendations in the treatment condition relative to the control condition. Specifically, after removing the market leader's candidate items, the company's algorithm could be

trained on additional internal data generated through user activities related to internal candidate items. A key point here is that: whereas the focal company's algorithm has offline pipelines to avoid transforming internal candidate items into features real-time, it only has access to the output from the market leader's API and does not have pre-computing features for API items that can be further used to train the algorithm due to data sparsity (Sarwar, 2001).[21] Moreover, it is unclear whether such API-related data sharing agreements allow for using the API output for training the focal algorithm. Indeed, a recent example is that of OpenAI API usage agreement which "prohibits using output form the API to develop a competing product" and "prohibits reverse engineering the source code, model parameters and algorithm".[22] Similar concerns of reverse engineering and the use of information to build competing was one of the reasons why Google restricted access to its autocomplete API. Together, as a result, in the absence of the market leader's API access, the relative improvement in the company's algorithmic systems, based on its internal data, may compensate for the initial decrease in the quality of search suggestions due to API removal.

Second, there are two (competing) behavioral factors that could impact users over time. On the one hand, there could be a novelty effect that would explain the upward trend over time for users. For example, experienced users, who are used to a certain level of product performance, might have a strong initial negative reaction to the changes in quality. Over time, though, they might get used to the decline in product quality, which could explain the upward trend in the treatment effect. Yet we believe that this is less plausible because our treatment is not a new feature experimentation and does not involve the disclosure of API items to the user. On the other hand, there could be a competing behavioral hypothesis which would predict the opposite trend. In particular, because experienced users are sensitive to the quality change, they could learn that search suggestions have degraded over time and then stopped clicking on them. If this were the dominant mechanism, treatment effects among experienced users should become more negative over time. Importantly, we posit that both of these behavioral explanations should be absent (or much weaker) among new users in the treatment condition because they do not have prior experience with the product.

[21] For details, see https://mlops.community/why-real-time-data-pipelines-are-so-hard/ and Stoica et al. (2017).
[22] See the terms of agreement here https://openai.com/policies/terms-of-use.

Figure 6: Longer-Term Treatment Effects

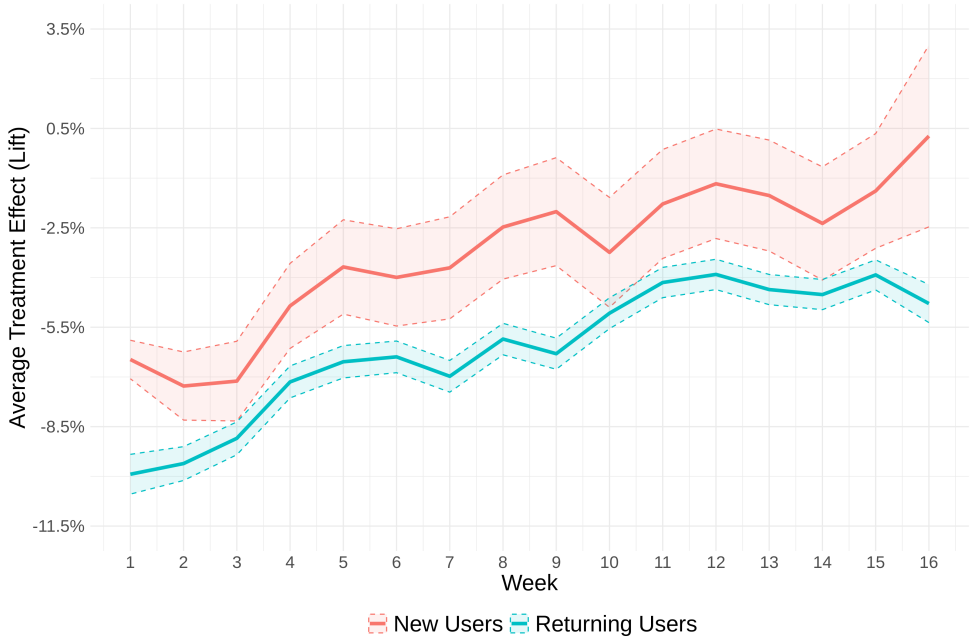


Notes: Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users. The figure plots weekly lift estimates based on linear regression for the entire sample. Error bands represent 95% confidence intervals.

To shed light on the underlying mechanism, we plot treatment effect estimates over time across two subgroups: new users vs. returning (experienced) users. We define new users on a weekly basis: During the 16-week experimental period, new users at week t are those who have never used the search bar since the start of the experiment and used the search bar for the first time at week t. Our hypothesis is that new users had few interactions with the product, so the behavioral effects should be minimal for such users. As a result, examining treatment effects over time among these users should help us detect the presence of gradual improvement over time in the company's algorithmic recommendations based solely on its internal data. Figure 7 shows that the (absolute) magnitude of treatment effects decreases over the course of the experiment, even among new users (red line). Thus, we provide suggestive evidence that estimates are likely to be driven by the incremental improvement in algorithmic recommendations due to continuous incorporation of internal data in

the treatment condition relative to control condition.[23]

Figure 7: Suggestive Evidence of Mechanism: Longer-Term Treatment Effects by User Type



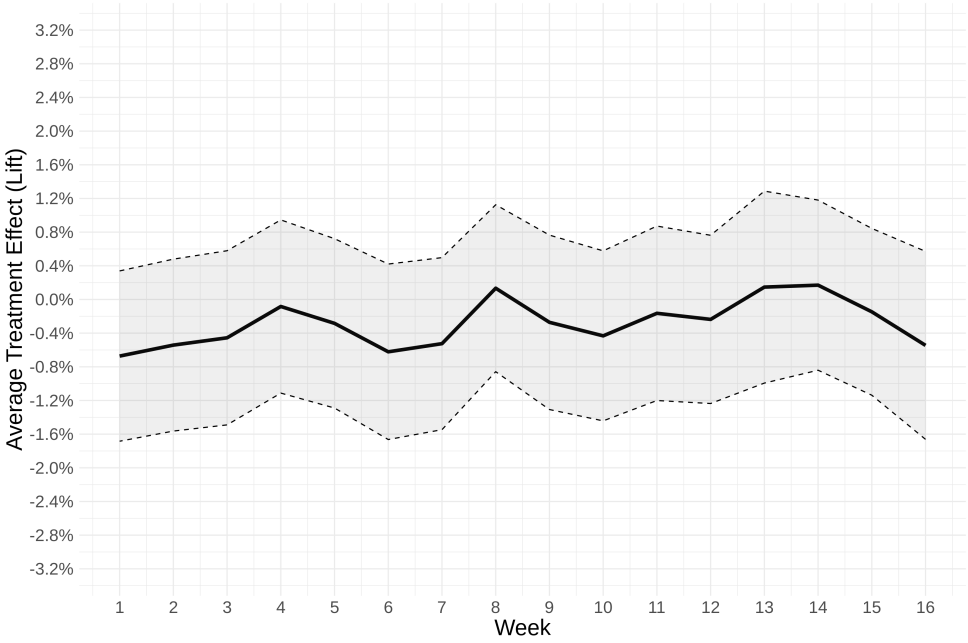
Notes: Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users. The figure plots weekly lift estimates based on linear regressions for sub-samples of new and returning users. Error bands represent 95% confidence intervals.

Furthermore, with such a longer-term analysis, we need to ensure that the estimates are not driven by differential attrition across treatment and control conditions. If users were becoming inactive regarding the search product usage, the estimates would capture the treatment effect due to temporal changes in the sample of users. To alleviate this concern, we examine whether the treatment induces a significant change in search-suggestion usage: i.e., the number of search-suggestion requests made by a user. We plot the weekly estimates of the impact of API removal on search-suggestion usage. Figure 8 shows that API removal does not cause a significant change in search-suggestion usage among treated users relative to control users in any week (confidence inter-

[23]Notably, we find that that the upward trend in Figure 7 has a flatter slope among experienced users (blue line) relative to new users. This suggests that experienced users learned that search suggestions degraded over time and then stopped clicking on them. As a result, such a behavioral mechanism counterbalances the improvement in the company's algorithmic recommendations, making the slope of the upward trend flatter among experienced users. If novelty effect were at play, we should expect a steeper upward trend among experienced users relative to new users because the novelty effect would strengthen the improvement in the company's algorithmic recommendations.

vals consistently contain zero). Together, these results suggest that the decrease in the (absolute) magnitude of lift estimates is unlikely to be driven by differential attrition across two conditions.

Figure 8: Longer-Term Effects of API Removal on Search-suggestion Usage



Notes: Lift refers to the incremental search-suggestion usage among treated users relative to control users as a percentage of search-suggestion usage among control users. The figure plots weekly regression estimates for the full sample. Error bands represent 95% confidence intervals.

Lastly, we look at another target metric, search button usage, to explore whether changes in CTR on search suggestions can be compensated by the utilization of an alternative search function. On average, API removal significantly increases the usage of search button by 5.7% ($\pm$1.1%). Figure A4 in the Appendix shows how the estimated lift in search button usage varies over time, and they suggest that the demand spillover to search button remains positive over the 16-week period.

Collectively, there are several takeaways from this analysis. First, a short-term evaluation would have made us overstate the value of the market leader's data. Second, we provide suggestive evidence that the relative improvement in the company's algorithmic recommendations due to the continuous feedback from internal data in the treatment condition is a plausible mechanism. Third, the longer-term treatment effects are unlikely to be driven by differential attrition across two conditions.

4.5 Robustness Checks

We carry out a variety of checks to test for the robustness of our results (see Table A2 in the Appendix). First, we estimate the treatment effect using only the first-day observation of each user in the experiment (i.e., the first day of the experiment when a user interacts with search suggestions) Second, we estimate a linear regression adjusting for observed user characteristics to potentially improve the precision of the estimate, and check the stability of results among the sample where all user characteristics are observed. Columns 1 and 2 show that the results are qualitatively similar to our baseline estimates with a negative and statistically significant effect. Third, we use alternate operationalizations of the dependent variable, including the click dummy (Column 3) and the logarithm of one plus clicks (Column 4). Our results are robust and qualitatively similar. Lastly, we estimate a model with all moderators in the same linear regression, rather than subsample analysis reported in Figure 5. The results are very consistent (Column 5). In summary, these checks provide an additional degree of confidence in our baseline results.

5 An Extension with Field Experiment 2

In the previous section, we have provided evidence on the impact of the market leader's data on the company's search product performance. In this section, we leverage another field experiment in July 2021 (corresponding to week 10 of the main experiment above) that ran for one day to achieve three objectives: (1) replicate the main results from the first field experiment, (2) validate the manipulation of search suggestions provided by the market leader's API, and (3) assess the relative impact of manipulating the supply of candidates into the ranking algorithm versus manipulating the rank directly in the same experimental setup.[24]

[24] Although the date of this experiment overlaps with the main field experiment, the interference between two experiments is less of a concern because the company conducted the main field experiment in a non-overlapping domain where users do not overlap with the users in this experiment.

5.1 Experimental Design

Similar to field experiment 1, the randomization in this experiment was implemented at the user level (see the randomization checks in the Appendix, Table A3). A total of 250,281 users were randomly assigned to one of three conditions:

Control (N=83,500): This condition is similar to the control condition in field experiment 1. Search suggestions are generated by the proprietary algorithm funnel. At the ranking stage, the ranking algorithm scores candidate items from two data sources: (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. The ranking algorithm scores such candidate items and generates highest-scoring search suggestions in a ranked order.

Rank Adjustment (N=83,517): Identical to the control condition, at the ranking stage, the ranking algorithm scores candidate items from two data sources: (1) those retrieved at the candidate generation stage and (2) those supplied by the market leader's API. In contrast to the control condition, the ranking algorithm lowered (boosted) the rank of the market leader's candidate items (the company's own candidate items) and then generated the final search suggestions. As a result, relative to those in the control condition, users in this condition are less likely to see search suggestions supplied by the market leader in the final list.

Removal of Access to the Market Leader's API (N=83,264): This condition is similar to the treatment condition in field experiment 1. The ranking algorithm (the same as the one used in the control condition) scores only candidate items retrieved from the candidate generation stage and generates highest-scoring search suggestions in a ranked order. As a result, users in this condition never see search suggestions supplied by the market leader.

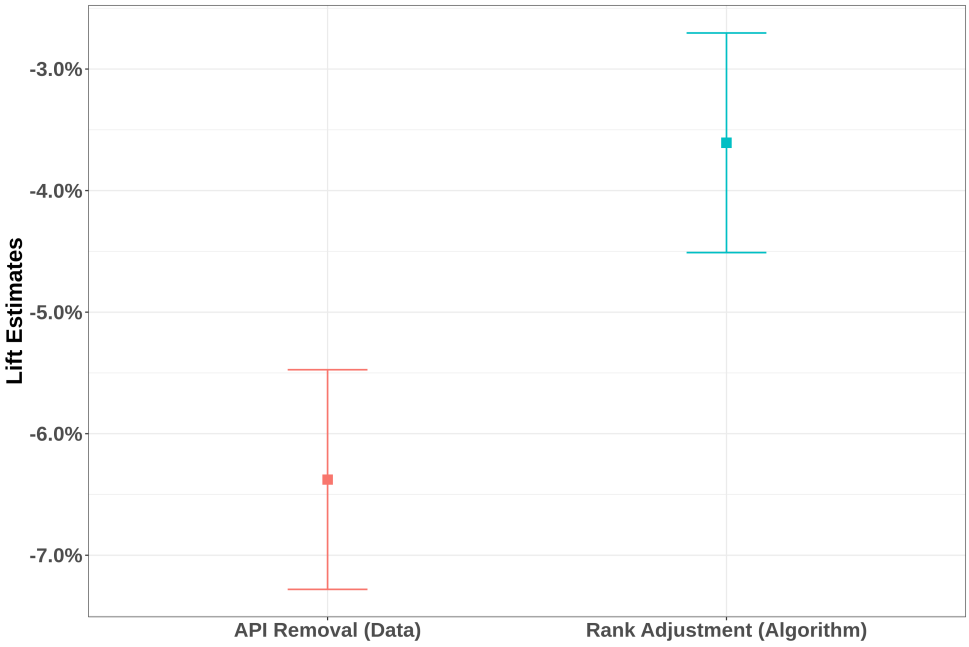
5.2 Results

Figure 9 shows that the removal of the market leader's API leads to a decrease in CTR by 6.4% (±0.9%). This estimate is between the aggregate estimate (i.e., –4.6%) and first-week estimate (i.e., –8.1%) from the main experiment. Additionally, the estimate of the main field experiment for week 10 is about 5%, which is statistically similar to the effect we find in field experiment 2. Hence, we

are able to broadly replicate the main effect reported earlier.

Turning to rank adjustment, we find that pushing down the ranking of search suggestions from the market leader leads to a decrease in CTR by 3.4% (±0.9%). This estimate supports that manipulating the rank directly has a significant impact on product quality. In practice, such an adjustment broadly relates to Google's adjustment in its ranking algorithms. Specifically, Google's white paper notes that "where our algorithms detect that a user's query relates to a "Your Money or Your Life (YMYL)" pages topic, we will give more weight in our ranking systems to factors like our understanding of the authoritativeness, expertise, or trustworthiness of the pages we present in response (Google, 2019)."

Figure 9: The Relative Impact of API Removal and Rank Adjustment



Notes: N= 250,281. Lift refers to the incremental CTR among treated users relative to control users as a percentage of CTR among control users. Error bars represent 95% confidence intervals.

Taken together, these estimates provide suggestive evidence that in our context, manipulating the rank of search suggestions induces a smaller impact on CTR relative to the removal of the market leader's data as candidates for the ranking algorithm. Another takeaway is that these

estimates shed light on the strength of the manipulation across different conditions. Conceptually, the manipulation of pushing down the rank of search suggestions from the market leader should be weaker than completely removing the market leader's candidate items. Therefore, the magnitude of the estimates (6.4% versus 3.4%) increases our confidence in the success of the manipulation.

6 Conclusion

We leverage a large-scale field experiment where we randomize access to the market leader's data as an algorithmic input to generate search suggestions for users. We find that, without access to the market leader's data, users click on search suggestions 4.6% less relative to those users in the condition where the algorithm has access to the external data through the market leader's API. The lack of external data leads to a reduction in the breadth of content consumed as measured by the number of content categories. These content categories are both mainstream and niche in terms of the share of clicks they get from users. The external data streamlines the experience for users that are highly active in using the search suggestions tool. Using the entire length of the large-scale experiment, we document significant dynamics in how users interact with the product when there is no access to external data. In particular, we find that the negative effect of the API removal reduces over time and is likely to be driven by the improvement in the company's algorithmic recommendations due to continual development by incorporating internal data. Finally, using a second (short-term) experiment, we manipulate (i) access to the market leader's API and (ii) the rank of search-suggestion candidates to ensure that it is indeed the market leaders data that leads to the effects we measure.

Our study has clear managerial implications. The results suggest that leveraging external sources of data can provide a significant return for a company, especially in the early stages of new product development. For search products, in particular, we show that access to depersonalized search results can increase engagement. This is important because there are several search products that are launched periodically (e.g., Neeva, Cliqz) that could look to this potential strategy. It could help to provide a good experience by increasing the breadth of both popular and niche content

available to them. We highlight a trade-off between the short-run gains and longer-run product development: i.e., reliance on external data over the longer term can limit organic development of the focal algorithm using only internal data.

We believe our results also shed light on policy issues being debated currently. The benefit derived by our partner company demonstrates that depersonalized search results being mandated by regulations, such as the Digital Markets Act, might help provide a leg up for nascent search products. The reduction in mainstream and niche content consumption demonstrates the power of access to a great breadth of data despite the absence of the use of personal data. These suggest that external data input at the ranking stage, as is prevalent in the search market, could help shore up engagement for startup search products at least in the short run.

Our study is not without limitations. Like other studies with field experiments, we are only able to look at one setting. It would be prudent to try and replicate the general tenor of our findings in other contexts. This would, of course, depend on whether the opportunity exists to leverage such data sharing agreements. Our study is also a partial equilibrium analysis and further research could take into account more general equilibrium dimensions. For example, there could be strategic responses from other (competing) platforms in the face of such data partnerships. Analyzing how the platform ecosystem evolves with such data partnerships or through regulations would be a fruitful next step.

References

Agrawal, A., J. Gans, and A. Goldfarb (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press. 1, 2

Alrashed, T., J. Almahmoud, A. X. Zhang, and D. R. Karger (2020). Scrapir: making web data apis accessible to end users. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–12. 2

Bar-Gill, S., E. Brynjolfsson, and N. Hak (2021). When small businesses become data-driven: A field experiment. *Working Paper*. 4.2

Benzell, S., J. Hersh, and M. Van Alstyne (2022). How apis create growth by inverting the firm. *Working Paper*. 1

Beraja, M., D. Y. Yang, and N. Yuchtman (2022). Data-intensive innovation and the state: Evidence from ai firms in china. *The Review of Economic Studies*. 1

Berman, R. and A. Israeli (2022). The value of descriptive analytics: Evidence from online retailers. *Marketing Science 41*(6), 1074–1096. 4.2

Brynjolfsson, E. and K. McElheran (2016). Data in action: data-driven decision making in us manufacturing. *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-06, Rotman School of Management Working Paper* (2722502). 4.2

Casella, G. and R. L. Berger (2002). *Statistical inference* (Second ed.). Duxbury Press: Pacific Grove, CA. 4.1

Chan, T., N. Hamdi, X. Hui, and Z. Jiang (2022). The value of verified employment data for consumer lending: Evidence from equifax. *Marketing Science 41*(4), 795–814. 1

Chiou, L. and C. Tucker (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research. 1

Claussen, J., C. Peukert, and A. Sen (2019). The editor vs. the algorithm: Targeting, data and externalities in online news. *Data and Externalities in Online News (June 5, 2019)*. 1

Covington, P., J. Adams, and E. Sargin (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198. 2

Crémer, J., Y.-A. de Montjoye, and H. Schweitzer (2019). Competition policy for the digital era. *Report for the European Commission*. 1

Deng, A., U. Knoblich, and J. Lu (2018). Applying the delta method in metric analytics: A practical guide with novel ideas. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 233–242. 4.1

Economist, T. (2017). The world's most valuable resource is no longer oil, but data. 1

Google (2019). How google fights disinformation. *White Paper*. 5.2

Gordon, B. R., R. Moakler, and F. Zettelmeyer (2022). Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*. 4.1

Huang, J., D. H. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio. *IO: Empirical Studies of Firms & Markets eJournal*. 4.4

Kim, H. (2019). The value of competitor information: Evidence from a field experiment. Technical report, Working Paper Harvard Business School. 4.2

Klein, T. J., M. Kurmangaliyeva, J. Prüfer, P. Prüfer, and N. N. Park (2022). How important are user-generated data for search result quality? experimental evidence. Technical report, TILEC Discussion Paper No. 1

Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1168–1176. 18

Krämer, J., D. Schnurr, and S. B. Micova (2020). *The role of data for digital markets contestability: case studies and data access remedies*. Centre on Regulation in Europe asbl (CERRE). 1

Kucharavy, A., Z. Schillaci, L. Maréchal, M. Würsch, L. Dolamic, R. Sabonnadiere, D. P. David, A. Mermoud, and V. Lenders (2023). Fundamentals of generative large language models and perspectives in cyber-defense. *arXiv preprint arXiv:2303.12132*. 1

Mitra, B. and N. Craswell (2015). Query auto-completion for rare prefixes. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1755–1758. 2

Nagaraj, A. (2022). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science 68*(1), 564–582. 1

Nandy, P., D. Venugopalan, C. Lo, and S. Chatterjee (2021). A/b testing for recommender systems in a two-sided marketplace. *Advances in Neural Information Processing Systems 34*, 6466–6477. 2

Neumann, N., C. E. Tucker, and T. Whitfield (2019). Frontiers: How effective is third-party consumer profiling? evidence from field studies. *Marketing Science 38*(6), 918–926. 1

Rubin, E. and A. Rubin (2021). On the economic effects of the text completion interface: empirical analysis of financial markets. *Electronic Markets 31*, 717–735. 10

Sarwar, B. M. (2001). *Sparsity, scalability, and distribution in recommender systems*. University of Minnesota. 4.4

Serban, I., A. Sordoni, Y. Bengio, A. Courville, and J. Pineau (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 30. 1, 2

Stoica, I., D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. E. Gonzalez, et al. (2017). A berkeley view of systems challenges for ai. *arXiv preprint arXiv:1712.05855*. 21

Sun, T., Z. Yuan, C. Li, K. Zhang, and J. Xu (2021). The value of personal data in internet commerce: A high-stake field experiment on data regulation policy. *Available at SSRN 3962157*. 1, 4.2

Tang, D., A. Agarwal, D. O'Brien, and M. Meyer (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 17–26. 16

Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization 54*(4), 683–694. 1

Valavi, E., J. Hestness, N. Ardalani, and M. Iansiti (2020). Time and the value of data. *Working Paper*. 1

Wernerfelt, N., A. Tuchman, B. Shapiro, and R. Moakler (2022). Estimating the value of offsite data to advertisers on meta. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (114). 1

Xue, L., P. Song, A. Rai, C. G. Zhang, and X. Zhao (2019). Implications of application programming interfaces for third-party new app development and copycatting. *Production and Operations Management*, 1887–1902. 1

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science 66*(3), 1045–1070. 1, 2

Zaveri, A., S. Dastgheib, C. Wu, T. Whetzel, R. Verborgh, P. Avillach, G. Korodi, R. Terryn, K. Jagodnik, P. Assis, et al. (2017). smartapi: towards a more intelligent network of web apis. In *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part II 14*, pp. 154–169. Springer. 2

Appendix A

Figure A1: How Does the Market Leader's API Work?

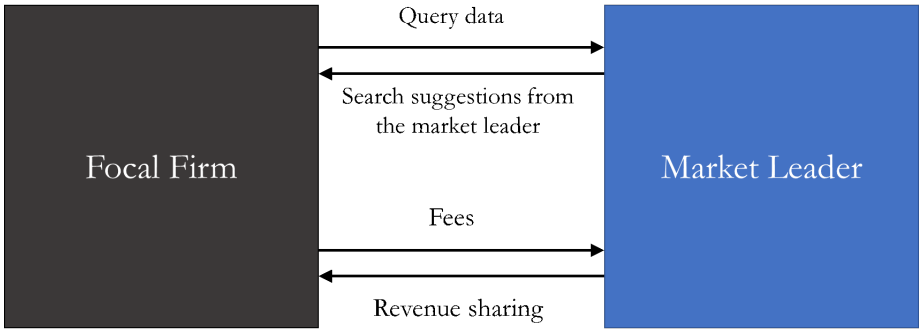


Figure A2: Proportion of Users Assigned to the Treatment Condition Over Time

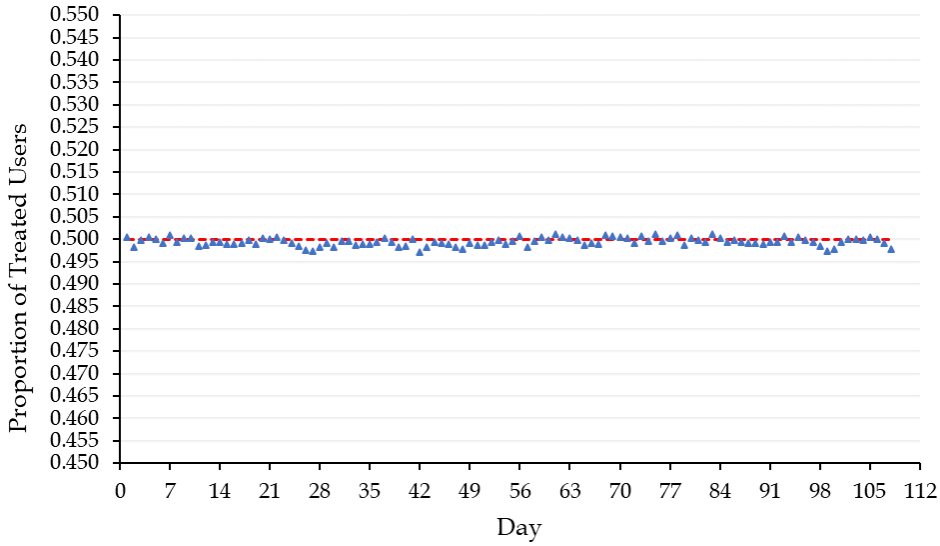


Figure A3: Proportion of New Users Assigned to the Treatment Condition Over Time

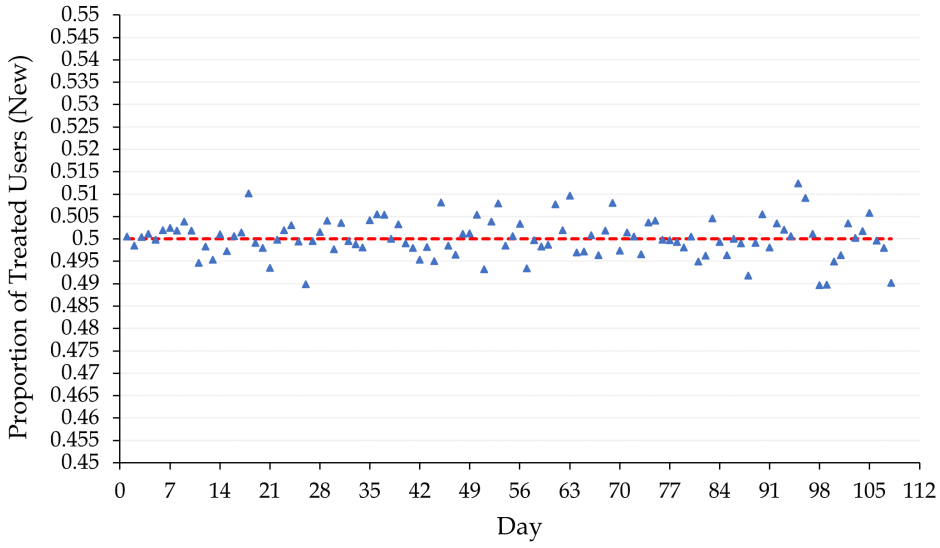
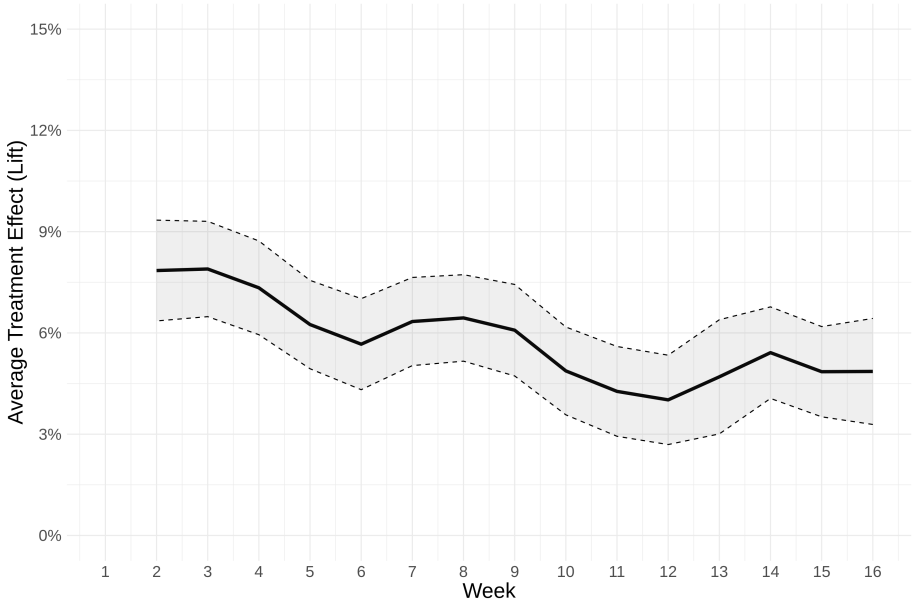


Figure A4: Search Button Usage



Notes: Lift refers to the incremental search button usage among treated users relative to control users as a percentage of search button usage among control users. Error bars represent 95% confidence intervals. Due to a technical glitch, we do not have accurate log records to construct the search button usage in the first week.

Table A1: Randomization Checks (Subsample of Popular and Niche Content)

User Characteristics	Control	Treatment	p. value
Male	0.530	0.530	0.300
	(0.001)	(0.001)	
Larger Cities	0.553	0.552	0.209
	(0.001)	(0.001)	
Smaller Cities	0.438	0.439	0.156
	(0.001)	(0.001)	
Mobile Operating System: Apple iOS	0.125	0.125	0.503
	(0.000)	(0.000)	
Mobile Operating System: Android	0.869	0.869	0.336
	(0.000)	(0.000)	
Active days in the past 30 days (search activities)	100	99.905	0.688
	(0.008)	(0.008)	
Query views in the past 30 days (search activities)	100	99.525	0.263
	(0.069)	(0.066)	

Notes. This table shows the balance between users in the treated relative to control groups along several observable dimensions for those who have clicked on a search suggestion atleast once during the sample period. Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, values reported in the last two rows were normalized so that the variable means in the control condition are 100.

Table A2: Robustness Checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	First Day	Controls	LPM			2nd Level	Niche (10%)	Popular (90%)
Variables	CTR	CTR	Click Dummy	Log(1+Clicks)	CTR	Unique Cat. CTR	CTR	CTR
API Removal	-0.0459***	-0.0539***	-0.0103***	-0.0329***	-0.0124***	-0.0037***	-0.0205***	-0.0134***
	(0.0014)	(0.0013)	(0.0009)	(0.002)	(0.001)	(0.0006)	(0.0053)	(0.0015)
API Removal×Female					-0.0016			
					(0.001)			
API Removal×Smaller Cities					0.0013			
					(0.001)			
API Removal×New User					0.0068**			
					(0.002)			
API Removal×Active Days					-0.0085***			
					(0.001)			
API Removal×Query Views					-0.0104***			
					(0.001)			
Observations	2,388,377	1,932,886	2,390,244	2,390,244	1,932,886	1,415,958	1,415,958	1,415,958

* $p<0.05$, ** $p<0.01$, *** $p<0.001$. Robust standard errors in parentheses. Column (1) uses only first user-day in the sample, column (2) controls for observable characteristics such as gender, city size, user activity in terms of query views and active days. Column (3) uses a linear probability model while column (4) uses the logarithm of (1+clicks) as the dependent variable. Column (5) includes all interactions with treatment status while column (6) has the CTR for second level categories. Columns (7) and (8) have the CTR for Niche and Popular categories at 10% and 90% thresholds as the dependent variable. The estimates are based on OLS regressions as in equation (1).

Table A3: Randomization Checks: Field Experiment 2

User Characteristics	Control (C)	Algorithm (T1)	Data (T2)	*p.* value Diff (T1,C)	*p.* value Diff (T2,C)
Male	0.571	0.567	0.567	0.073	0.076
	(0.002)	(0.002)	(0.002)		
Larger Cities	0.633	0.631	0.633	0.523	0.771
	(0.002)	(0.002)	(0.002)		
Smaller Cities	0.329	0.328	0.327	0.615	0.298
	(0.002)	(0.002)	(0.002)		
Active days in the past 30 days (search activities)	100	100.053	99.866	0.861	0.657
	(0.033)	(0.033)	(0.033)		
Query views in the past 30 days (search activities)	100	100.436	100.399	0.547	0.577
	(0.560)	(0.573)	(0.560)		

Notes: This table shows the balance between users in the treated relative to control groups along several observable dimensions. Following the hierarchical classification of Chinese cities, larger cities include tier 1 to 4 cities (e.g., tier 1: largest cities such as Beijing), whereas smaller cities refer to tier 5 cities and below. p-value is obtained based on a two-sided t-test on the equality of means with unequal variances. For confidentiality purposes, numbers in the last two rows were normalized so that the variable means in the control condition are 100.