The Value of External Data Capabilities in the Search Market: Evidence from a Field Experiment*

Xiaoxia Lei
Shanghai Jiao Tong University†

Yixing Chen
University of Notre Dame‡

Ananya Sen
Carnegie Mellon University§

June 2024

Abstract

Firms increasingly leverage external entities' data capabilities to unlock improvements in their offerings, but measuring the value of such capabilities is challenging. Collaborating with the search team at a technology company, we analyze a field experiment where we manipulated access to an external, leading search engine's autocomplete API. We measure the causal impact of the removal of access to such an API on the clickthrough rate (CTR) on the focal company's search suggestions and CTR on results on the Search Engine Results Page. We find that, on average, compared to the baseline with access to the external API, API removal leads to a 4.6% decrease in the CTR on search suggestions. Further, exploiting the experimental variation, we use an instrumental variables approach to establish that a 10% increase in CTR on search suggestions leads to a 1.85% increase in CTR on the top-slot search results. These estimates suggest the revenue potential of utilizing external data capabilities for nascent search products. Finally, the negative effect of API removal on search suggestions becomes less negative over time with the effect magnitude in the longer term being half 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 capability limits the improvement of its own autocomplete predictions based on the accumulation of internal data. Apart from its managerial implications, this research sheds light on policies such as the Digital Markets Act that proposes data sharing by large digital platforms.

*We thank Manuela Collis, Hong Deng, Anindya Ghose, John Lalor, Shijie Lu, Alex Moehring, Abhishek Nagaraj, Christian Peukert, Mohammad Rahman, Yoonseock Son, Steve Tadelis, Sonny Tambe, Yun (Alicia) Wang, Joy Wu, Li Zheng, Shuang Zheng and participants at Notre Dame (Marketing), CMU (Tepper), University of Pittsburgh, UC Irvine (Merage), Nanyang Business School, SWUFE, Soochow University, UT Austin (McCombs), Georgia Tech (Scheller), Purdue (Daniels), Cornell (Dyson), UCSD, IU (Kelley), 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, TSE Online Platform Seminar Series, Columbia/Wharton Management, Analytics, and Data Conference, 11th Operations and Supply Chain Workshop, NBER Economics of Artificial Intelligence Conference, INFORMS Annual Meeting, and 34th POMS Conference for helpful suggestions.

†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

With the rise of the internet, search engines have become a valuable vehicle that enables users to navigate the plethora of online information. Given the critical role of search engines, it is vital to have healthy competition in the search market that allows for the best offering for users. Regulators have become increasingly wary of a few dominant players in the market (e.g., Google in the U.S. and Baidu in China). The focal concern is that dominant players possess data capabilities[1] that prevent newer and smaller competitors from gaining a foothold in the search market. Leading search engines utilize data capabilities to streamline the search process by providing more relevant search results at the top of the search engine results page (SERP). Increasingly, a fundamental way in which search engines aim to gain a data-capability-related advantage is query autocomplete, which efficiently generates relevant search suggestions for queries entered in the search bar before users reach the SERP (e.g., Google Autocomplete, Bing Autosuggest).[2] Indeed, providing relevant search suggestions through query autocomplete is fundamental to user experience, so leading search engines devote significant resources to the development of query autocomplete (Cai et al., 2016).

To counter such potential data-capability-related barriers to entry and expansion, new legislation such as the Digital Markets Act that came into effect in November 2022 requires leading search engines to provide smaller players access to privacy-preserving search results to level the playing field.[3] Theoretically, access to external entities' data capabilities is potentially valuable for smaller players in the search market for two reasons. First, queries can be ambiguous, so search engines rely on multiple external sources to generate responses to the query (Kopliku et al., 2014). Second, unlike algorithmic recommendations in e-commerce that rely heavily on personal data (Sun et al., 2023), search engines rely more on the content relevance and popularity (what other users have searched for in aggregate) and only partly on personalization, especially for results at the top of

[1] Both the academic literature (Zeleti and Ojo, 2017) and policy documents (Great Britain. Department for Business and (BIS), 2013; Ubaldi, 2013) have emphasized the concept of data capabilities as part of the open data architecture.
[2] See https://blog.google/products/search/how-google-autocomplete-works-search/ and https://blogs.bing.com/search/March-2013/A-Deeper-Look-at-Autosuggest.
[3] For an overview of the Digital Markets Act, see: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en#documents.

the SERP (Hannak et al., 2013).[4]

Managerially, measuring the value of external data capabilities is essential as practitioners highlight the need to integrate external data capabilities.[5] A prevalent practice is to use Application Programming Interfaces (APIs) to gain access to data capabilities from large players (Benzell et al., 2023; Havakhor et al., 2024; Xue et al., 2019). In the context of search, Google Autocomplete API enables publishers and developers to retrieve search suggestions from Google, and Bing Autosuggest API enables smaller players such as DuckDuckGo to source search suggestions from Bing. Despite their prevalence and potential economic importance, there is limited empirical evidence of the causal impact of external data capabilities on the recipients' product performance.

In this paper, we analyze a large-scale field experiment to measure the impact of the *removal* of access to a leading search engine's autocomplete API (i.e., the market leader's autocomplete API) on the performance of our partner company's search suggestions (i.e., a smaller player; hereafter, "the company"). We further explore which users respond to the removal of API access, as well as how users respond to the experimental variation in the longer term. The primary outcome of interest is clickthrough 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 (Kucharavy et al., 2023; Park and Chiba, 2017; Serban et al., 2016).[6] In line with industry standards, search suggestions are generated by the underlying algorithm that uses multiple sources to predict what users want to click in response to their queries. As an entrant, the company leveraged the market leader's autocomplete API to retrieve candidate suggestions 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 access to the market leader's autocomplete API.

The field experiment lasted 108 days and involved more than 2.3 million users among which each

[4] On average, 11.7% of search results are personalized, and top slots tend to be less personalized than bottom slots. See also: https://www.techpolicy.press/whats-the-difference-between-search-and-recommendation/.

[5] See this report by Deloitte on the opinion of analytics professionals among others as an example https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html.

[6] Indeed, recent applications of generative artificial intelligence (e.g., ChatGPT) are referred to as "autocomplete for everything". For details, see https://www.noahpinion.blog/p/generative-ai-autocomplete-for-everything.

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 'the market leader's data capability'.[7] The queries to the market leader's autocomplete 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 design has three 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 capability. Second, maintaining a consistent treatment for 108 days enables us to estimate longer-term effects of access to the market leader's data capability. Third, this design is closely related to the broader agenda of the Digital Markets Act, which requires privacy-preserving data-sharing agreements touted to have significant benefits for companies in the marketplace.

We report three sets of findings. First, relative to the status quo, API removal leads to a 4.6% decline in average CTR on search suggestions, demonstrating the value of the market leader's data capability. We also document who is more (less) responsive to API removal. API removal has no differential effect across male and female users while users in larger cities are affected more. Additionally, API removal impact is stronger for heavy users relative to light users, suggesting that access to the market leaders' API is more effective in helping user retention by streamlining heavy users' experience than in ameliorating the cold-start problem for light users. Managerially, these heterogeneous treatment effects inform the company's growth and segmentation strategy for a nascent online product based on readily available demographic and usage variables.

Second, leveraging the random assignment as an instrumental variable, we estimate the impact of

[7] We use terms such as external data capabilities, the market leader's data capabilities, or the market leader's candidate items interchangeably throughout the paper.

clicking on search suggestions on clicking on the search result at the top of the SERP. Specifically, we show the elasticity of CTR on the top-slot search result with respect to CTR on search suggestions is 0.185. That is, a 10% increase in CTR on search suggestions leads to a 1.85% increase in CTR on the top-slot search result. This result demonstrates the downstream impact of CTR on search suggestions on user engagement on the SERP.

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 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 capability (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. Furthermore, the treatment effect is less negative for queries that accumulated more searches over time: i.e., (1) those that have been searched by the same user previously and (2) those that have been searched by other users previously. Together, these results suggest that in the absence of the market leader's API, there has been a gradual improvement in the company's search suggestions because of the accumulation of its internal data. This creates a trade-off between the short-run benefits of using the market leader's data capability and longer-term growth through enriching internal data. We further verify that this data pattern is not driven by (1) diminishing returns to external candidate items, (2) user adaptation, (3) differential attrition or dynamic self-selection, or (4) unobserved spillover.

Our findings offer clear, practical implications for managers and policymakers. First, we inform managers of the impact of tapping into external data capabilities such as third-party APIs on search product performance. Moreover, we believe our paper is the first to establish the elasticity between CTR on search suggestions and CTR on the search results highlighting how query autocomplete can shape the search process. Our analyses provide 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] More broadly, other platforms such

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

as Reddit and Yahoo Finance have also (sometimes inadvertently) limited access to information through their API, influencing the entities utilizing these APIs.[9] Second, our findings inform policymakers, in particular with regard to the Digital Markets Act. We highlight a critical trade-off where the extensive reliance on the market leader's data capability can limit internal learning, which can impede improvements in autocomplete predictions 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. In this regard, our results show some economically meaningful short-term benefits of privacy-preserving data-sharing for newer players in the search market. More generally, we provide empirical evidence that informs a recent debate on whether big data undermines market competition (Crémer et al., 2019; Krämer et al., 2020; Tucker, 2019). Lastly, understanding the value of privacy-preserving data sharing in search is becoming increasingly important, especially with recent innovations of generative models and user-facing applications in search engines and web browsers (e.g., OpenAI large language model for Bing search engine and Edge browser).

Our paper contributes to four 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 the personalization of search and quantifies significant returns to personal data. Chiou and Tucker (2017) analyze the efficacy of search recommendations when companies such as Yahoo! and Microsoft reduced the amount of individual-level data retention to 90 days and found no change in a user's CTR. Schaefer and Sapi (2023) analyze the role of within and across user learning for algorithms used by search engines to increase engagement. Klein et al. (2022) demonstrate the need for large players to share user-generated data to improve the performance of competing search products. These papers are motivated by privacy regulations and the role of data in conferring market power to search engines. We augment this strand of research by analyzing how access to external data capabilities, a combination of data and inferences based on that data, impacts search product performance using a large-scale, longer-term field experiment. We demonstrate that deper-

[9] See more here: https://www.reddit.com/r/rootgame/comments/14jmfzx/reddit_is_killing_thirdparty_apps_and_itself/ and Havakhor et al. (2024).

sonalized search-suggestion candidates through third-party API access could help search products grow successfully, at least in the short run, which speaks to regulation associated with mandated data sharing across search engines in a privacy-preserving manner.

Second, we contribute to the literature that analyzes different strategies for platform growth. Benzell et al. (2023) use aggregate data on online platforms to demonstrate how a firm can grow by opening itself up to third-party complementors using APIs. Relatedly, Peukert et al. (2023) quantify the value of personal data (or lack thereof) for an algorithm relative to human experts. Sun et al. (2023) simulate a privacy regulation through a field experiment on Alibaba to quantify the value of individual-level data for platform users. They find significant negative effects on engagement and purchase when personal data is not used for product recommendations, especially for niche products. Relative to these papers, and Sun et al. (2023) in particular, we analyze a different yet economically meaningful context of search, which, as summarized in Table A1, is theoretically distinct from e-commerce in terms of their recommender systems and the potential use for external data capabilities. Moreover, whereas Sun et al. (2023) studies the value of personal data in a business-to-consumer (B2C) setting, we study the extent to which access to the market leader's data API influences smaller players in search (i.e., a business-to-business (B2B) data-sharing mechanism). Next, due to our three-month-long experiment relative to a few hours in the case of Sun et al. (2023), we highlight a trade-off where external data capabilities can help product performance initially, but over-reliance in the longer term can inhibit focal prediction development based on internal data. Finally, we take a middle path and analyze a privacy-preserving situation where depersonalized data inputs from external sources could be used for product development.

Third, we contribute to a nascent strand of literature that focuses on the efficacy of text-based search in the context of e-commerce. Zheng et al. (2023) leverage a field experiment on a food delivery platform to understand the impact of 'related product suggestions' in the search bar interface on the amount and diversity of consumption. Fang et al. (2024) analyze the impact of textual refinement and visual cues in the search bar on short- and long-term consumption. Similar to these papers, we analyze the impact of prompts in the search bar in facilitating the search process. Our estimates are in the same ballpark as these two papers. A key differentiator is that we analyze

the role of access to data capability via an external API in improving the search suggestions, which a key managerial and policy question.

Finally, our paper is related to a strand of literature that analyzes the impact of utilizing several forms of third-party data and data-enabled capabilities 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) 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, Chan et al. (2022) show that access to employer-verified employment data via Equifax benefits both auto loan borrowers and lenders. Havakhor et al. (2024) utilize the shutdown of Yahoo Finance API to show that access to high-volume historical and real-time price data cause retail investors to trade excessively with worse outcomes. 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, external source of data capabilities through third-party APIs for online platforms, providing implications for other companies that aim to utilize such a source for their product growth.

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. The app offers its users various in-app products such as news feed, search engine, and video and eBooks streaming. Launched in 2020, the main product is search suggestions, with more than 100 million monthly active users. By design, search suggestions are predictive in nature and function

as query autocomplete, which is similar to Google autocomplete feature offered by Google Chrome. Search suggestions are considered as 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).[10] 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?"). A user can choose to click a matched suggestion, press the search button without adopting search suggestions, delete this query and enter a new one, or quit the session.

Figure 1: An Illustration of Search Suggestions



Search suggestions and related contexts of predictive text completion are important with companies investing significant resources (Agrawal et al., 2018). Search suggestions help streamline the search process, bridging the gap between users' search intent and content consumption. Optimizing search suggestions is important for our partner company because it intends to monetize this process. For example, search suggestions can help users accelerate the process of finding desired search result at the top of the SERP, and they have implications for revenue since clicks on search results at the top of the SERP can lead to substantial revenue for the focal firm (Schaefer and Sapi, 2023; Ursu,

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

2018). Moreover, the search team's medium-term aim is to earn revenue from each click made on search suggestions with links embedded within the suggestions.[11] Indeed, search box optimization is an increasingly popular strategy for companies trying to get additional visibility.[12]

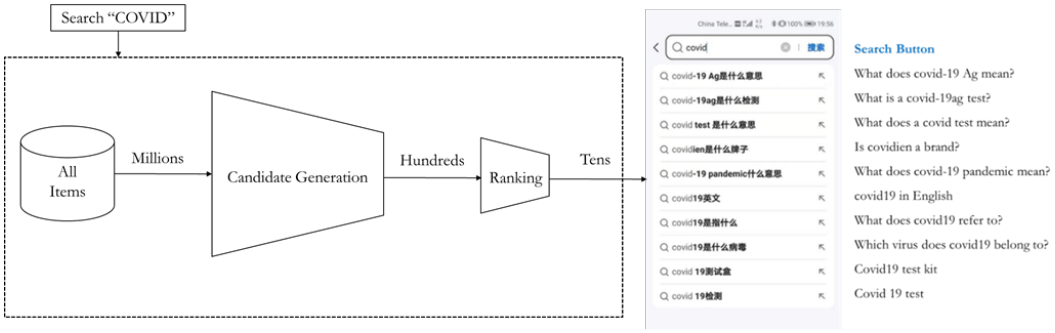
How does the company generate search suggestions in response to user-submitted queries? At a high level, the company uses a proprietary algorithmic funnel, a three-stage architecture similar to a recommender system (Covington et al., 2016): (1) item generation, (2) candidate generation, and (3) ranking. 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, text summarization) based on large language models 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

[11] Indeed, while our experiment was running, the team started deliberating about 'direct reach cards'. 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. In addition, our partner company is not the only one innovating in this search market. New search engines in the United States, e.g. Neeva, highlight how users found these direct reach links in autocomplete suggestions helpful (https://www.theverge.com/23802382/search-engine-google-neeva-android).

[12] https://googlerankingexpert.medium.com/search-box-optimization-ff883b2fb56e

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 potential lack of candidate items with respect to quantity and quality. For example, at the initial development stage, the company's algorithmic funnel may generate a relatively shorter list of search suggestions that could match users' search intent. To address this problem, the company had started to leverage 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 large user base and capability 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 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.[13] Additionally, these candidate items capture both the market leader's data and the inferences based

[13] We believe that the query terms are depersonalized due to privacy concerns. For a similar operation, see https://docs_v3.dataforseo.com/v3/serp-google-autocomplete-overview/

on the market leader's candidate generation algorithm. Since we cannot separate the impact of data from that of the inferences, we refer these candidate items as 'data capabilities', a concept highlighted in the literature and policy discussions as mentioned earlier. 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 an agreement where the company pays service fees to the market leader at an undisclosed rate for each query it makes through the API, and the market leader compensates the company through the volume of search query terms (see Figure A1 in the Appendix).[14] For each query to the API, the focal company pays a fixed amount (typically a fraction of a cent per query, for example, in the Google Autocomplete API)[15]. Hence, for millions and billions of queries, it can become a sizable amount just from one company. If an embedded link is clicked on in the search suggestions, there will be revenue sharing. Note that this revenue sharing was not in effect during our experiment (but was put in place after our experiment).

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 their popularity.[16] 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

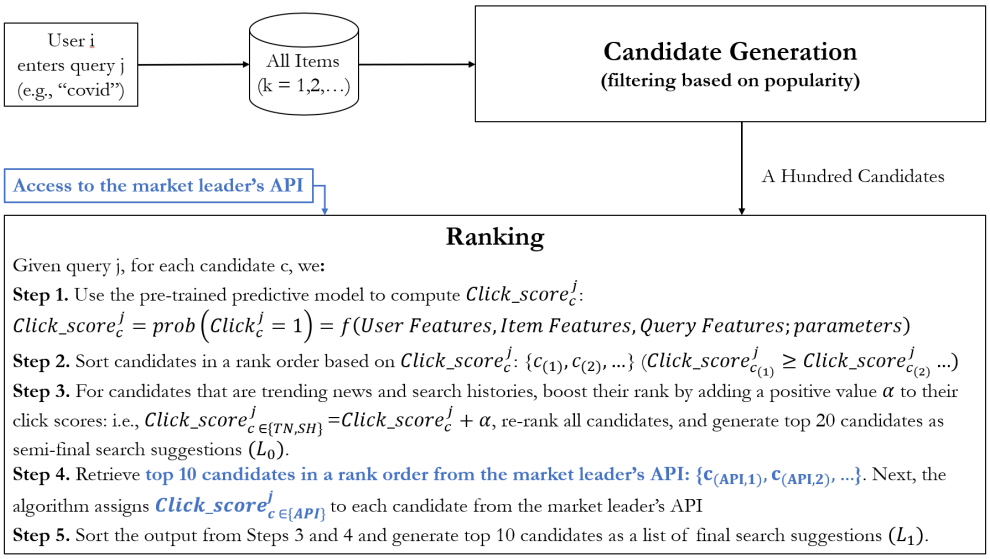
[14] It is pertinent to note that neither we as researchers nor our partner company observe any details of the market leader's database. Hence, it is a 'black box' for us. This elevates the need for our field experiment to understand the economic implications of such data inputs along various dimensions.

[15] https://developers.google.com/maps/documentation/places/web-service/usage-and-billing

[16] 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.

under **Ranking**). Lastly, using standard tools for "crossing streams" in such settings,[17] 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**).[18] 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



Notably, this setup reflects how a specific form 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 (Alrashed et al., 2020; Zaveri et al., 2017). Given its prevalence,

[17] For details about Kafka tools used in such contexts to interleave 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/.

[18] Since the rank was determined based on the click score, a continuous measure based on a variety of input features, it is very unlikely (with probability tending to zero) that there were ties between two search suggestions with respect to their click scores.

external candidate items retrieved from the market leader's API as input into the ranking stage makes it a key data and inference-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. Moreover, since user clicks on the search results on SERP represent a key metric of search quality (Schaefer and Sapi, 2023), we also examine CTR on search results as a downstream outcome of CTR on search suggestions.

However, measuring the value of the market leader's data capability (candidate items retrieved from the market leader's autocomplete API) 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.

3 Field Experiment

3.1 Experimental Design

The randomization was implemented at the user level in real time. That is, as soon as 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 to mitigate imperfect compliance.[19] 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 from May 17, 2021 to September 1, 2021. Figure 4 visualizes our design.

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 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.

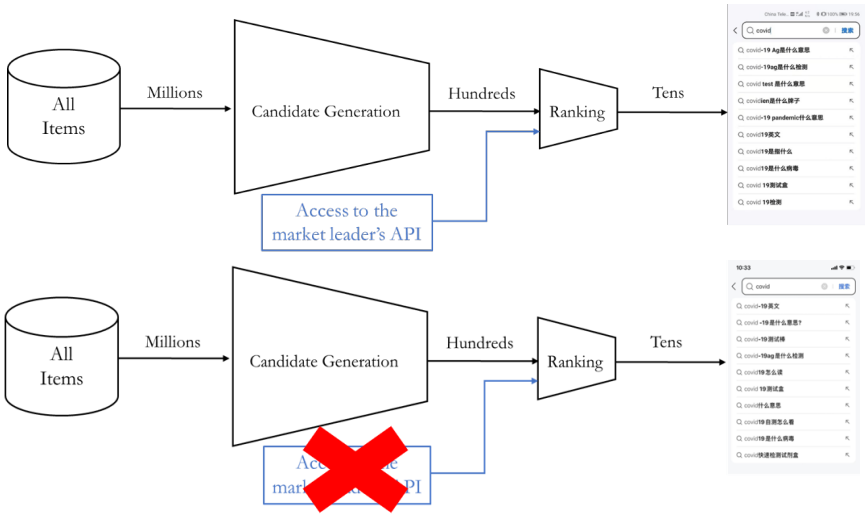
Treatment (N=1,195,625): In response to user-submitted queries, search suggestions are generated by the same proprietary algorithm funnel without access to the market leader's API at the ranking stage. The ranking algorithm (the same as the one used in the control condition) scores only candidate items 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 never see search suggestions supplied by the market leader.[20] Notably, our treatment does not cause any

[19] 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 such as LinkedIn, Google, and Microsoft, the company uses the standard design of overlapping experiments to run experiments simultaneously and efficiently (Figure 2b in Tang et al. (2010)) and hash-based assignment (rather than generating ephemeral random numbers) to ensure no correlation between the assignments with multiple experiments running (Xu et al., 2015).

[20] It is also pertinent to note that any updates to the training data input during the experimental period remains separate for treatment and control conditions. Similar to data-diverted experiments proposed in Holtz et al. (2023), this was explicitly done to prevent Stable Unit Treatment Value Assumption (SUTVA) violations.

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 autocomplete API.

Figure 4: Experimental Design



3.2 Data

The primary data set for analysis is at the user-level (the unit of randomization) 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). About 82% of users have used the search bar prior to the experiment in this data set. The dependent variable, aggregate CTR, is computed for each user as the ratio of the total number of clicks 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 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). Our data suggests that, on average, the number of search suggestions presented was remarkably similar, with 8.80 and 8.68 search suggestions in the control and treatment conditions, respectively.

We further supplement this user-level data set (**Primary Data**) with two additional data sets. First, we collect user-level metrics such as the number of searches conducted on the search engine results page (SERP), the number of clicks on the top-slot search results, and the number of clicks on the other-slot search results (**SERP Data**). Second, we collect detailed log records of a random sample of our experimental subjects (11,630 treated users and 11,913 control users) starting from June 26, 2021 (**Granular Data**).[21] In each log record in **Granular Data**, we observe the query term entered by each user, content category associated with the term, the ordered list of search suggestions provided, and whether and what a user clicked.

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
	(0.155)	(0.154)	
Query views in the past 30 days (search activities)	100	99.383	0.126
	(0.288)	(0.283)	

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.

[21] We were not able to collect log-level data for the entire experiment sample nor data for the random sample prior to June 26, 2021 due to the company's data retention policy.

4 Empirical Analyses

4.1 Empirical Framework

We use a potential outcome framework to specify 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 regression without control variables as the baseline results and use heteroskedasticity-robust standard errors.

Another important consideration is that β itself does not provide a sense of the effect magnitude. Therefore, we report the *lift* estimates 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 Effects

4.2.1 Average Treatment Effect and Treatment Effects by User Characteristics

Our starting point is the estimation of the average treatment effect. As shown in Figure 5, there is a statistically significant and negative treatment effect: on average, API removal decreases the CTR on search suggestions by 4.6% in the treatment condition relative to the control condition (i.e., the average lift is –4.6%). The large sample size gives us a precise estimate with the width of the 95% confidence interval of less than 10% of the absolute value of the point estimate. Notably, the magnitude of the effect is economically significant, especially in the context of search markets.[22] Relative to other academic studies, the magnitude is in line with the 2% effect in Zheng et al. (2023) for the impact of 'related product suggestions' in the search bar interface on sales on an online food delivery platform. Similarly, Fang et al. (2024) find that textual and visual refinement of text search leads to a 3% increase in revenue over 24 weeks on an e-commerce platform.

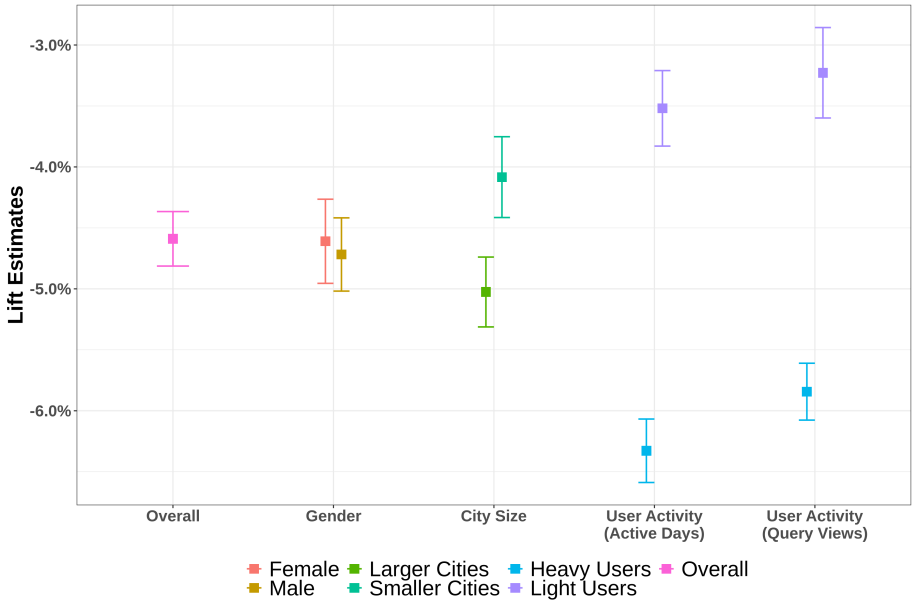
Next, we explore whether treatment effects vary with pre-experimental user characteristics to dig into who is more (less) responsive to the treatment of API removal. The first dimension relates to gender and city of residence as possible sources of heterogeneity. We see that (1) treatment effects are economically and statistically indistinguishable for female and male users, and (2) the treatment effect is significantly larger among users in larger cities than those in small cities. From a managerial perspective, these findings inform the company's growth and segmentation strategy for a nascent online product based on readily available demographic variables.

The second dimension is to examine whether users with different activity levels are more or less responsive to API removal. A priori, it is unclear whether external candidate items will help solve the cold-start problem for light, less active users or provide a more streamlined experience for heavy, more active users, or both. To examine this, we use the number of search-related active days in the past month prior to the experiment to differentiate light 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

[22] 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.

is strictly above the median and zero otherwise. Figure 5 shows that the magnitude of the negative effect of API removal among heavy users is significantly larger than that among light users. This pattern is highly consistent when we use the number of query views in the past month prior to the experiment as an alternate measure. A plausible explanation is that because heavy users are more experienced, 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 capability.[23] In this regard, our results complement Sun et al. (2023) 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.

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. The estimates are based on Equation (1). Overall refers to the estimate of average treatment effect based on the full sample, while other estimates of heterogeneity treatment effects are based on subsamples. Error bars represent 95% confidence intervals.

[23] In Appendix B, we further explore how treatment effects vary by query popularity and find that the negative treatment effect is primarily driven by popular queries. It suggests that due to its data capabilities, the market leader's autocomplete API provides the company with access to useful candidate items with respect to their popularity based on real searches on the market leader's platform.

Next, we look at another target metric, search button usage, to explore whether a decrease in CTR on search suggestions is substituted by the utilization of an alternative search function. In particular, if the search suggestions become less useful for users, they might use the search button as an alternative way to navigate to the SERP. On average, we find that API removal significantly increases the usage of search button by 5.7%. This positive effect on search button usage suggests a positive demand spillover to the search button as a substitute for search suggestions. The magnitude of the change suggests an almost one-to-one substitution between these actions.

Lastly, we carry out a variety of checks to test for the robustness of our baseline estimates (Table A3 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. Third, our results are robust to alternative operationalizations of the dependent variable, including the click dummy (Column (3)) and the logarithm of one plus the number of clicks (Column (4)). Lastly, we estimate a linear regression with moderators rather than subsample analysis reported in Figure 5. The results are consistent (Column (5)). In summary, these checks provide an additional degree of confidence in our baseline results.

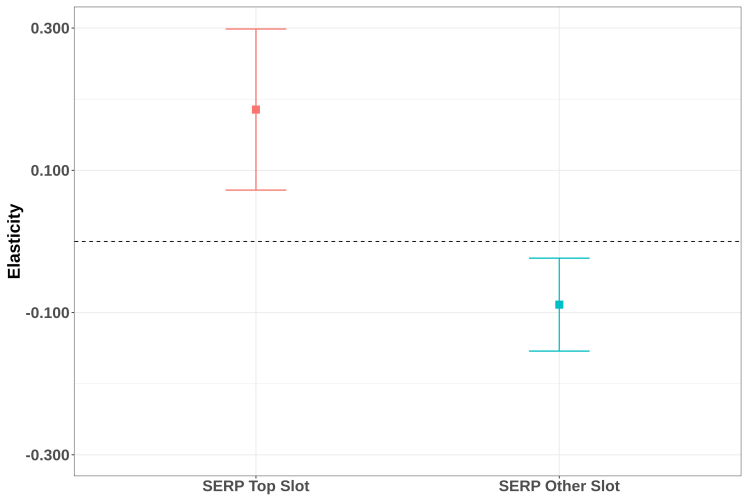
4.2.2 Economic Implications

To gauge the economic implications of the average treatment effect, we establish an empirical link between clicking on search suggestions and clicking on results on the search engine results page (SERP). That is, we translate the decrease in clicking on search suggestions into a decline in clicking on search results. Specifically, we focus on the total number of top-slot search results because the economic significance of top-slot clicks is well established in the literature both in terms of revenue for the company and in streamlining the search process for the users(e.g., Schaefer and Sapi (2023); Ursu (2018); Yoganarasimhan (2020)). Using **SERP Data**, we operationalize the CTR on search

results using (a) click-through rate on the top-slot search result on the SERP and (b) click-through rate on other-slot search results on the SERP.

As shown earlier, API removal exogenously shifts CTR on search suggestions and does not belong directly in the production function of CTR on search results, so it can serve as the instrumental variable (IV) for CTR on search suggestions. Thus, we can estimate the causal relationship between CTR on search suggestions and CTR on the top-slot search result via two-stage least-squares. We log transformed all variables to enable the interpretation of coefficient as elasticity. The first stage F-statistic is 506.85, which is significantly above the standard threshold of 10 (Stock and Yogo, 2002). For the second stage, Figure 6 shows that the elasticity of CTR on the top-slot search result with respect to CTR on search suggestions is 0.185. That is, a 10% increase in CTR on search suggestions leads to a 1.85% increase in the CTR on the top-slot search result. This implies that clicking on search suggestions has a downstream impact on clicking on results at the top of the SERP.

Figure 6: Elasticity of CTR on Search Results with respect to CTR on Search Suggestions



Notes: Error bars represent 95% confidence intervals.

Next, if the CTR on top-slot search results increases, is it a market expansion effect or does it substitute away from other slots? Figure 6 shows that an increase in CTR on search suggestions

decreases CTR on other-slot search results. This implies that search suggestions streamline the search process by facilitating relevant top-slot search results, reducing attention on the lower slots.

Further, we carry out a back-of-the-envelope computation to understand the economic implications of access (or lack thereof) to external data capabilities. The monetary value of the CTR on search suggestions can be separated into two components: (1) query revenue gained from clicking on the search suggestions and (2) revenue from clicking on search results.[24] We will provide a conservative estimate based on internal and industry data. We focus on a monthly user base of 100 million, which approximates the number of users of the search product toward the end of our experiment. This number aligns with global estimates of the small search players such as DuckDuckGo (100 million as of 2023), Bing (500 million as of 2023).[25] We leveraged internal data to estimate the potential revenue loss from clicks on suggestions at the search suggestion stage. Specifically, the unit query revenue is 0.012 Chinese yuan (CNY), which we aggregate to the month level, accounting for 6.45 search-suggestion clicks per month per user. A caveat is that the unit query revenue is on the lower end of related industry estimates exactly because the search product at that stage was a new one with limited monetization. For the 100,000,000 users of the search product, a 4.6% decrease in search-suggestion CTR creates a loss of roughly 356,000 CNY or $50,000 per month (i.e., $6.45 \times 0.012 \times 0.046 \times 100,000,000$). As mentioned earlier, this is the most conservative estimate, with the industry average being at least 10 times higher in terms of the unit query revenue.

Next, based on our IV estimates, we quantify the revenue loss in downstream SERP clicks due to API removal. Due to unavailability of internal data on this dimension, we leveraged industry-level data to estimate the potential revenue loss from downstream outcomes in search. In China, each search usually generates a revenue of over 0.7 CNY (or 10 cents) for a gatekeeper search engine, while it is 0.35 CNY for a smaller search engine like our partner.[26] Using this estimate and the fact that about 58% of searches end up in clicking on search result, we find that a 4.6% decrease in search-

[24] A key metric for (2) is "unit query revenue", which equals the total search keyword auction revenue divided by the total number of searches. For more details, see https://dataforseo.com/blog/google-autocomplete-api-for-keyword-research-tool and https://www.dragonmetrics.com/guide-to-keyword-research-for-baidu-seo/.

[25] For more details, see https://techreport.com/statistics/duckduckgo-statistics/ and https://backlinko.com/bing-users.

[26] For more details, see the interview https://www.geekpark.net/news/334965 and the industry report from Dongxing Securities https://pdf.dfcfw.com/pdf/H3_AP201509210010852974_1.pdf.

suggestion CTR creates a loss of roughly 1,117,000 CNY or $159,000 per month through SERP results using the elasticity calculated before (i.e., $6.45 \times 0.58 \times 0.35 \times 0.046 \times 100,000,000 \times 0.185$). This implies that the total revenue loss due to API removal is approximately $210,000 per month or about $2.52 million annually. This is an economically significant number for any organization, particularly for a team that launched a new search product. A caveat is that this number is based on the assumption that all revenue is generated from the top slot. While the literature suggests that the top slot gets the maximum clicks and revenue, some of the revenue generated could be from lower slots. In that scenario, we would have to account for the reduction in clicks on other slots on the SERP as a result of the treatment. Accounting for the reduction in CTR of other slots, and making the extreme assumption that all slots on the SERP generate the same revenue, the total revenue would be approximately $1.5 million annually. While this is also an economically significant number, we believe that realistically the estimate would be closer to $2.52 annually.

We do not have proprietary information about the cost per query to the market leader's API, and industry estimates can vary significantly across contexts, scale, and time.[27] Moreover, under the regulations proposed, tapping into external APIs for data-related inputs will not be priced through the market but from an external fund to aid smaller companies.[28] Hence, when such an external fund is set up, access to such an API can be done without cost, leading to the gains quantified above. Based on our revenue estimates of $2.52 million, we can also estimate the cost per query, making it worthwhile for the focal company to query the API each time through a market mechanism. An average user performs 14.5 searches per month and the external API is queried for every search. Based on our lower bound revenue estimates, this implies that for 100 million users, the break-even point is approximately $0.00015. As the search engine matures, as mentioned before, the unit query revenue for search suggestions can grow 10-fold while it can at least double for SERP. This can make the break-even cost be $0.00057. Overall, this exercise allows us to highlight the economic significance of the external data capabilities through the API, which has implications for the focal company, the search engine industry, but also policy makers.

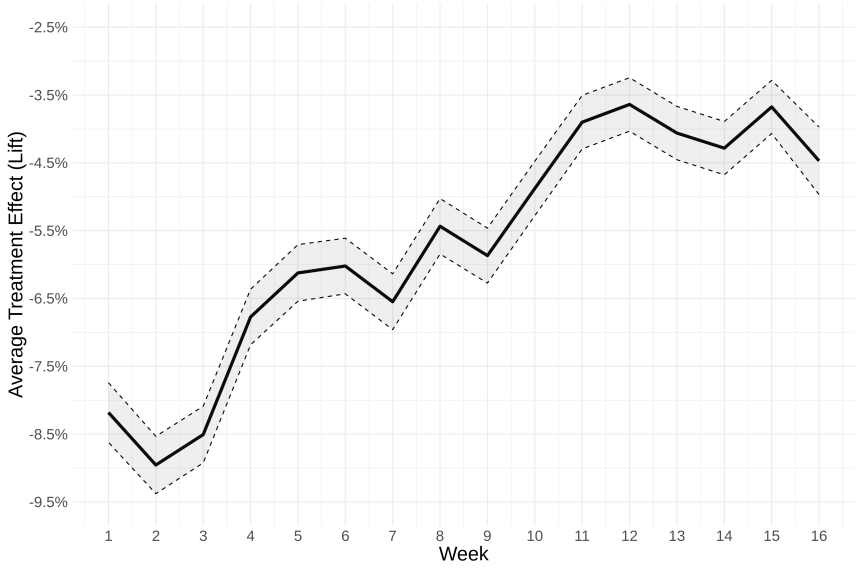
[27] See https://www.geekpark.net/news/334965 for some anecdotes.

[28] See https://en.m.wikipedia.org/wiki/United_States_v._Microsoft_Corp. and https://moglen.law.columbia.edu/publications/lu-18.pdf for some anecdotes.

4.3 Long(er)-Term Effects

A novel aspect of our design is that a consistent 16-week treatment assignment allows us to estimate the 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 7 shows how the lift estimates vary over time, with the solid (dotted) line representing the point estimates (the 95% confidence intervals). The absolute values of the estimates start at about a 8.1%–9%, with those 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 capabilities by a factor of 2.

Figure 7: Longer-Term 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.

4.3.1 Do Our Results Reflect Diminishing Returns to External Candidate Items?

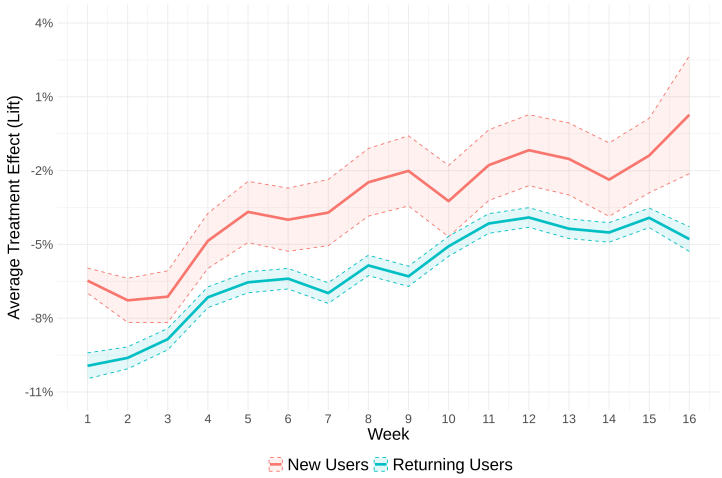
A possible explanation is that the decline in the absolute magnitude over time could reflect diminishing returns to external candidate items (Peukert et al., 2023). That is, the treatment effect

estimates could be less negative over time if the API-generated candidate items become less valuable over time. To examine this possibility, we focus on the trend in CTR among the users in the control condition where the candidate items are supplied by both the focal firm and the market leader. If the observed long-term effect were driven by diminishing returns to external candidate items, the trend in CTR in the control condition would show a concave pattern. In Figure A5, we conduct a calibration analysis to test the curvature of the weekly average CTRs among the users in the control condition (Committee et al., 1994; Johansson, 1979; Little, 1979). This calibration exercise cannot reject the null of a linear trend and hence, these results do not support a concave trend indicating diminishing returns to candidates items from the external API.

4.3.2 Do Our Results Reflect User Adaptation?

Another possible explanation is that users might have a strong initial negative response to the changes in quality of search suggestions due to API removal and then gradually get used to such changes over time—user adaptation. This adaptation might explain the upward trend in the treatment effect estimates. However, this is less plausible based on our institutional setting because our treatment does not involve the disclosure of API items to the user. Moreover, to provide concrete empirical evidence, 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 adaptation should be minimal for new users. Therefore, examining treatment effects over time among new users should help us understand whether our effects are driven by user adaptation over time. Figure 8 shows that the treatment effects become less negative over the course of the experiment, even among new users (red line). Thus, our estimates are unlikely to be driven by user adaptation to API removal.

Figure 8: Longer-Term Treatment Effects on CTR by User Type: New vs. Returning Users



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.

4.3.3 Do Our Results Reflect Dynamic Self-selection?

The third possible explanation is that our results are driven by dynamic self-selection of users. For example, our definition of new users could make the sample of new users became smaller over time, causing the temporal changes in the sample of new users. Similarly, if returning users were becoming inactive regarding their usage, the estimates would capture the treatment effect due to temporal changes in the sample of returning users. First, as shown in Figure A3, our randomization ensures that the proportion of new users assigned to each condition is balanced over time. Second, we examine whether the treatment induces a significant change in query volume. We plot the weekly estimates of the impact of API removal on query volume among new and returning users. Figure 9 shows that API removal does not cause a significant change in query volume among treated users relative to control users in any week for both subgroups (confidence intervals consistently contain zero). The average effects for the whole period are similarly statistically insignificant for both subgroups, with p=0.62 (for returning users) and p=0.36 (for new users).[29] Together, these results

[29] In Figure A4, we verify that API removal does not cause a significant change in the query volume for the entire sample, alleviating the concern of differential attrition (changes in the composition of users).

suggest that the decrease in the magnitude of lift estimates is unlikely to be driven by dynamic self-selection of users.

Figure 9: Longer-Term Treatment Effects on Query Volume: New vs. Returning Users



Notes: Lift refers to the incremental number of queries entered by treated users relative to control users as a percentage of number of queries entered by control users. Error bands represent 95% confidence intervals.

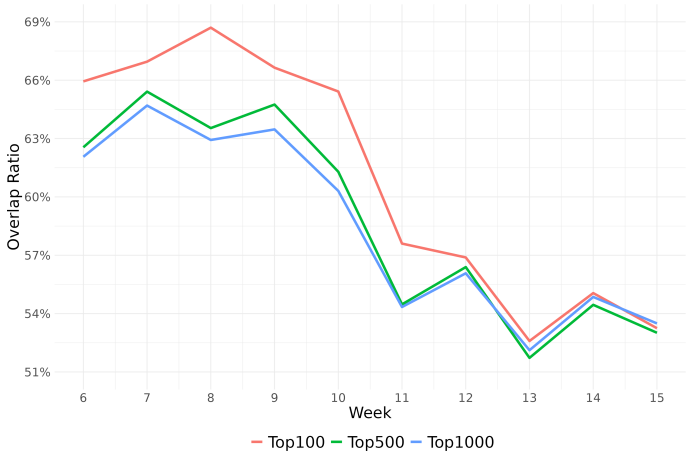
4.3.4 Do Our Results Reflect Violations of SUTVA?

The fourth possible explanation is that the decline in the magnitude of the longer-term effects could be driven by a spillover across the treatment and control conditions. Over time, user-item-algorithm interactions in the treatment (control) condition could affect the algorithm and subsequent search suggestions to the counterparts. As a result, this spillover might make search suggestions more similar across two conditions, gradually reducing the performance gap between the two conditions.

To examine this explanation, we provide both institutional and empirical evidence. First, the algorithm was held constant across two conditions, and the system in each condition was updated based on only data produced by experimental subjects in that condition. Similar to the purpose of "data-diverted experiments" proposed in Holtz et al. (2023), the goal of these system decisions was to prevent SUTVA violations. Second, we utilize the **Granular Data** (detailed log records of 11,630 treated users and 11,913 control users). If there were a spillover, the overlap between search suggestions in the two conditions should potentially increase when users entered the same query

across the two conditions (i.e., common query).

Figure 10: Overlap Ratio of Search Suggestions to Common Queries



In Figure 10, we observe that the overlap ratios between search suggestions in the two conditions decrease for the top 100, top 500, and top 1000 common queries that users enter in both treatment and control conditions. suggesting that the decrease in the magnitude of lift estimates is unlikely to be driven by spillovers between the treatment and control conditions. This exercise also lends credence to the steps taken by our partner team to keep data systems separate for treated and control users. Moreover, this diminishing overlap over time and the treatment effect becoming less negative hints at search suggestions becoming better over time, potentially due to access to more internal data—the mechanism we explore next.

4.3.5 Do Our Results Imply Better Prediction due to Internal Data Accumulation?

So far, we have provided empirical evidence to rule out several alternative explanations of the observed longer-term effects. Our thesis is that the decline in the magnitude over time could be driven by the improvement in the company's search suggestions fueled by the accumulation of internal data in the absence of the market leader's data capability. Specifically, in the absence of the market leader's candidate items, the company's search suggestions could be improved because of additional internal data generated through user activities related to internal candidate items.

Therefore, in the absence of access to the market leader's API, the relative improvement in the company's autocomplete predictions, based on its internal data, may gradually compensate for the negative impact of API removal.

Yet can the company use external candidate items to develop the algorithmic system? In line with industry standards, although the focal company's algorithm has pre trained, offline pipelines to transform internal candidate items into features, it only has access to the output from the market leader's API and thus does not have pre-computing features for API items to be used for training the algorithm in real-time. As a result, this causes an inference and data-sparsity problem for external API-based output (Sarwar, 2001).[30] 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 OpenAI API usage agreement, which "prohibits using output from the API to develop a competing product" and "prohibits reverse engineering the source code, model parameters and algorithm".[31] 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.

Specifically, we examine whether the negative effect of API removal could be weakened if the queries have been searched in the past. The rationale is that if there were indeed improved predictions due to the accumulation of internal data in the treatment condition in the absence of the market leader's API, the negative impact of API removal would be mitigated for queries that accumulated more internal data (e.g., search histories). To dive deeper into this mechanism, we constructed compile the data set at the user-query-day level based on the **Granular Data**. In Column (1) of Table 2, we first replicate the baseline effect in this user-query-day level dataset of the random sample. That is, the API removal leads to a decrease in CTR on search suggestions. In Column (2) of Table 2, we provide evidence to support the hypothesis of improved prediction due to internal data accumulation within the same user. We defined a time-varying indicator variable, **Repeated Queries** (=1 if query j has been searched by user i prior to day t, 0 otherwise). We can see that this interaction term is positive and significant, implying that the negative impact of

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

API removal is smaller for terms queried by the same user before. Furthermore, in Column (3), we analyze whether the negative impact of API removal decreases as the number of times a query term has been input by other users apart from the focal user. We define a time-varying cumulative measure, **Query Histories**: i.e., the logarithm of the frequency of query j that has been searched by other users apart from user i prior to day t. Again, we see that the interaction term is positive and significant, suggesting that the negative impact of API removal is weakened as queries establish their histories via user searches over time. These results, along with Figure 10 where search suggestions across the treatment and control conditions were becoming different over time, suggest the prevalence of within-user and across-user learning based on internal data in the absence of the market leader's API (Hagiu and Wright, 2023; Schaefer and Sapi, 2023).

Table 2: Evidence on the Role of Internal Data

Variables	(1) CTR	(2) CTR	(3) CTR
API Removal	-0.0030***	-0.0171***	-0.0044***
	(0.0005)	(0.0040)	(0.0009)
API Removal × Repeated Query		0.0145***	
		(0.0041)	
API Removal × Query Histories			0.0004**
			(0.0002)
Query fixed effects	✓	✓	✓
Day fixed effects	✓	✓	✓
Observations	1,636,900	1,636,900	1,636,900

* $p<0.1$, ** $p<0.05$, ***$p<0.01$. Intercepts are omitted. Repeated Query is equal to 1 if the query j has been searched by user i prior to day t, 0 otherwise. Query Histories is defined as the logarithm of the frequency of query j that has been searched by other users apart from user i prior to day t. Standard errors reported in parentheses are clustered at the query level.

Collectively, there are several takeaways from this section. First, a short-term evaluation would have made us overstate the value of the market leader's data capability. Second, we validate that a plausible mechanism of the longer-term effects is that the accumulation of internal data over time helps generates better search suggestions. Third, the longer-term effects are unlikely to be driven by (1) diminishing returns to external candidate items, (2) user adaptation, (3) dynamic self-selection, or (4) unobserved spillover.

5 An Extension with Field Experiment 2

In the previous section, we have provided evidence on the impact of the market leader's data capabilities on the company's search product performance. In this section, we leverage another one-day field experiment in July 2021 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.[32]

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 A2). 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 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 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

[32] 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.

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.

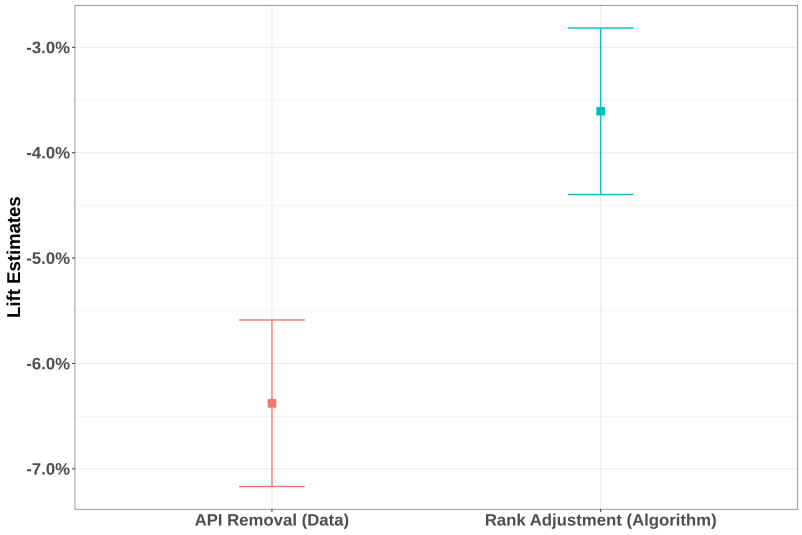
5.2 Results

Figure 11 shows that the removal of the market leader's API leads to a decrease in CTR by 6.4%. This estimate is between the aggregate estimate (i.e., –4.6%) and first-week estimate (i.e., –8.2%) from the main experiment. In addition, the estimate of the main field experiment for week 10 is about 4.9%. 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 3.6%. 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)."

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 candidate items 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 3.6%) increases our confidence in the success of the manipulation.

Figure 11: 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.

6 Discussion

We leverage a large-scale field experiment where we randomize access to the market leader's data capabilities as an algorithmic input to generate search suggestions for users. We find that, without access to the market leader's data capabilities, users click on search suggestions 4.6% less relative to those users in the condition where the algorithm has access to the external data capabilities through the market leader's API. Moreover, we find that a 10% increase in CTR on search suggestions leads to a 1.8% increase in CTR on the top slot link on the SERP. 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 capabilities. 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 improved predictions by the incorporation of 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 capabilities that lead to the effects we measure.

Our study has clear managerial implications. The results suggest that leveraging external data capabilities 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 not only on the focal product but also downstream on the SERP. This is important because there are several search products that are launched periodically (e.g., Neeva, Cliqz) that could look to this potential strategy. We highlight a trade-off between the short-run gains and longer-run product development: i.e., reliance on external data capabilities 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 that are currently being debated. 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 increase in top slot clicks on the SERP downstream, as a result, 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 capabilities-based 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 has limitations. Like other studies with field experiments, we can only look at one setting. Hence, our estimates require assumptions to extrapolate to other contexts. Given that our field experiment varies external candidate items at only one stage of the algorithmic funner, it would be prudent to replicate the general tenor of our findings in other contexts. This would, of course, depend on whether the opportunity exists to leverage such data and data-based inference-sharing agreements. Our study is also a partial equilibrium analysis, and further research could consider 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 focused 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. 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

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. B

Benzell, S. G., J. Hersh, and M. Van Alstyne (2023). How apis create growth by inverting the firm. *Management Science*. 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

Cai, F., M. De Rijke, et al. (2016). A survey of query auto completion in information retrieval. *Foundations and Trends® in Information Retrieval 10*(4), 273–363. 1

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

Committee, A. M. et al. (1994). Is my calibration linear? *Analyst 119*(11), 2363–2366. 4.3.1

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

Fang, L., Y. Chen, F. Chiara, Z. Yuan, and Y. Wang (2024). Platform information provision and consumer search: A field experiment. Technical report, National Bureau of Economic Research. 1, 4.2.1

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

Great Britain. Department for Business, I. and S. (BIS) (2013). Seizing the data opportunity: a strategy for uk data capability. 1

Hagiu, A. and J. Wright (2023). Data-enabled learning, network effects, and competitive advantage. *The RAND Journal of Economics 54*(4), 638–667. 4.3.5

Hannak, A., P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson (2013). Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 527–538. 1

Havakhor, T., M. S. Rahman, T. Zhang, and C. Zhu (2024). Tech-enabled financial data access, retail investors, and gambling-like behavior in the stock market. *Management Science 0*(0), null. 1, 9

Holtz, D., J. Brennan, and J. Pouget-Abadie (2023). A study of" symbiosis bias" in a/b tests of recommendation algorithms. *arXiv preprint arXiv:2309.07107*. 20, 4.3.4

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*. 4.3

Johansson, J. K. (1979). Advertising and the s-curve: A new approach. *Journal of Marketing Research 16*(3), 346–354. 4.3.1

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. 22

Kopliku, A., K. Pinel-Sauvagnat, and M. Boughanem (2014). Aggregated search: A new information retrieval paradigm. *ACM Computing Surveys (CSUR) 46*(3), 1–31. 1

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

Little, J. D. (1979). Aggregate advertising models: The state of the art. *Operations research 27*(4), 629–667. 4.3.1

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

Park, D. H. and R. Chiba (2017). A neural language model for query auto-completion. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1189–1192. 1

Peukert, C., A. Sen, and J. Claussen (2023). The editor and the algorithm: Recommendation technology in online news. *Management Science 0*(0), null. 1, 4.3.1

Sarwar, B. M. (2001). *Sparsity, scalability, and distribution in recommender systems*. University of Minnesota. 4.3.5

Schaefer, M. and G. Sapi (2023). Complementarities in learning from data: Insights from general search. *Information Economics and Policy 65*, 101063. 1, 2, 2, 4.2.2, 4.3.5

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

Stock, J. H. and M. Yogo (2002). Testing for Weak Instruments in Linear IV Regression. Technical report, National Bureau of Economic Research, Inc. 4.2.2

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*. 30

Sun, T., Z. Yuan, C. Li, K. Zhang, and J. Xu (2023). The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy. *Management Science*. 1, 4.2.1

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. 19

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

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. 1

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science 37*(4), 530–552. 2, 4.2.2

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

Xu, Y., N. Chen, A. Fernandez, O. Sinno, and A. Bhasin (2015). From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2227–2236. 19

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, 4.2.2

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

Zeleti, F. A. and A. Ojo (2017). Open data value capability architecture. *Information Systems Frontiers 19*(2), 337–360. 1

Zheng, S., S. Tong, H. E. Kwon, G. Burtch, and X. Li (2023). Recommending what to search: Sales volume and consumption diversity effects of a query recommender system. *Available at SSRN 4667778*. 1, 4.2.1

Appendix A Additional Tables and Figures

A.1 Tables

A.1.1 Comparison between Search and Recommendation

Table A1: Search Engine vs. Homepage Recommendations

	Search Engine	Homepage Recommendations
Basis	Rank items based on degree of match with explicit search query	Rank items based on implicitly collected engagement data
Explicitness of ranking signals	High	Low
Use of personalization	Low/Moderate	High

A.1.2 Randomization Checks: Field Experiment 2

Table A2: 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
	(0.214)	(0.216)	(0.214)		
Query views in the past 30 days (search activities)	100	100.436	100.399	0.547	0.577
	(0.506)	(0.518)	(0.506)		

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.

A.1.3 Robustness Checks

Table A3: Robustness Checks

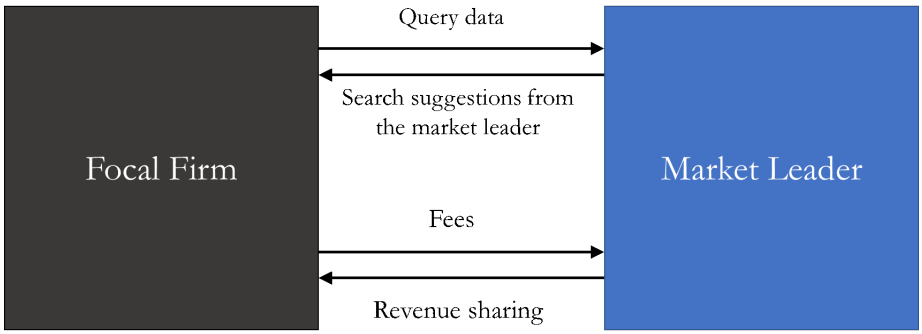
Variables	(1) First Day CTR	(2) Controls CTR	(3) LPM Click Dummy	(4) Log(1+Clicks)	(5) CTR
API Removal	-0.0459***	-0.0539***	-0.0103***	-0.0329***	-0.0124***
	(0.0014)	(0.0013)	(0.0009)	(0.002)	(0.001)
API Removal×Female					-0.0016
					(0.001)
API Removal×Smaller Cities					0.0013
					(0.001)
API 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

* $p<0.1$, ** $p<0.05$, *** $p<0.01$. 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.

A.2 Figures

A.2.1 Mechanics of the Market Leader's API

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



A.2.2 Randomization Checks

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

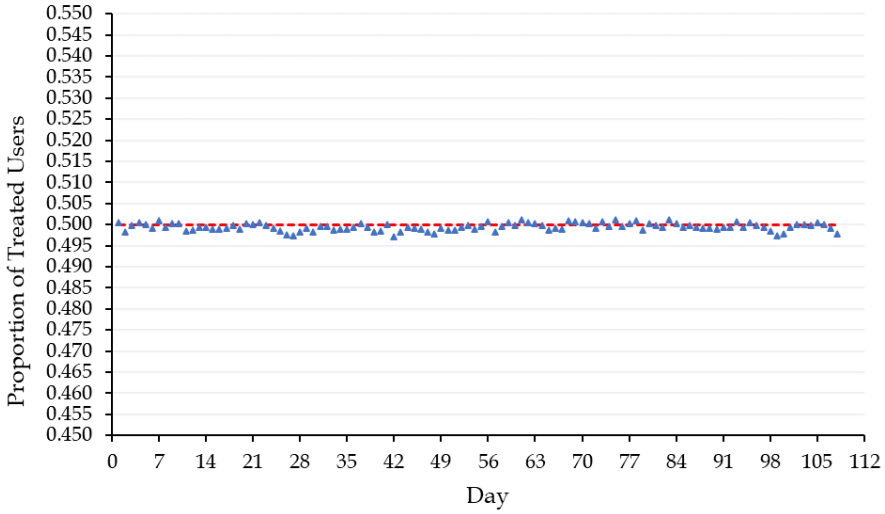
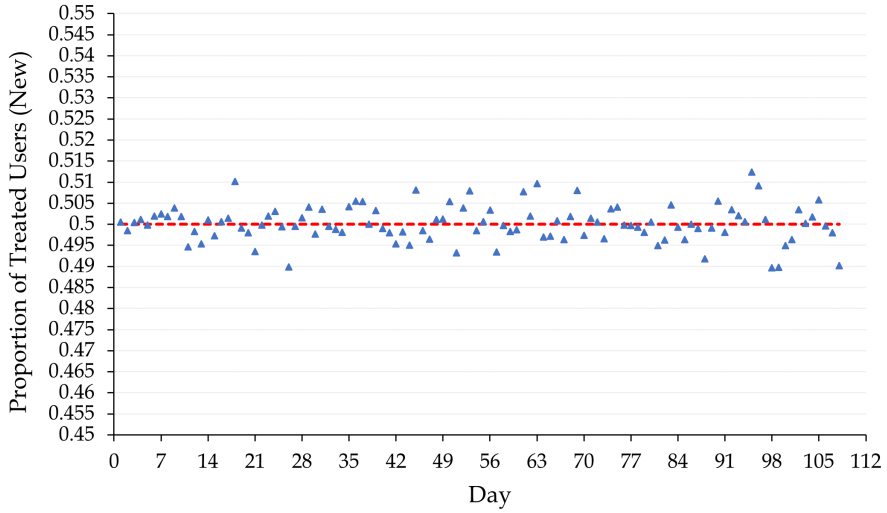
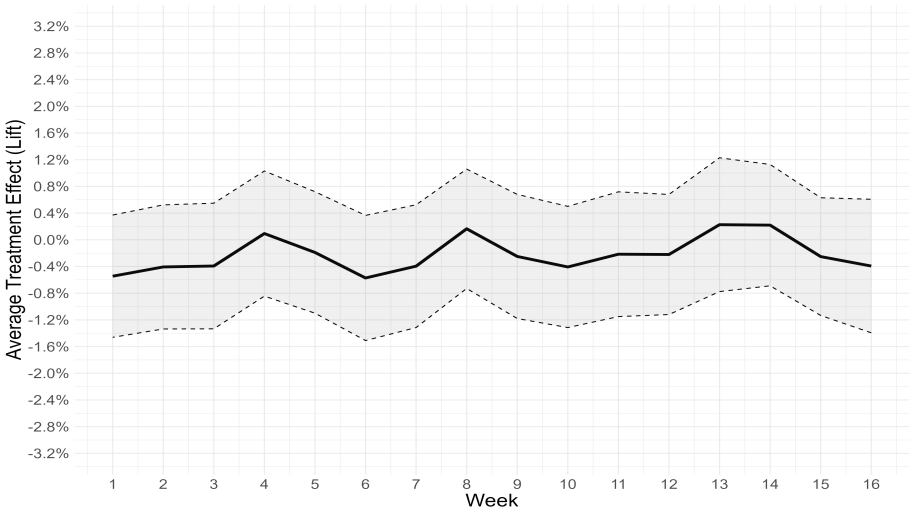


Figure A3: Proportion of New Users Assigned to the Treatment Condition Over Time



A.2.3 Ruling out Attrition

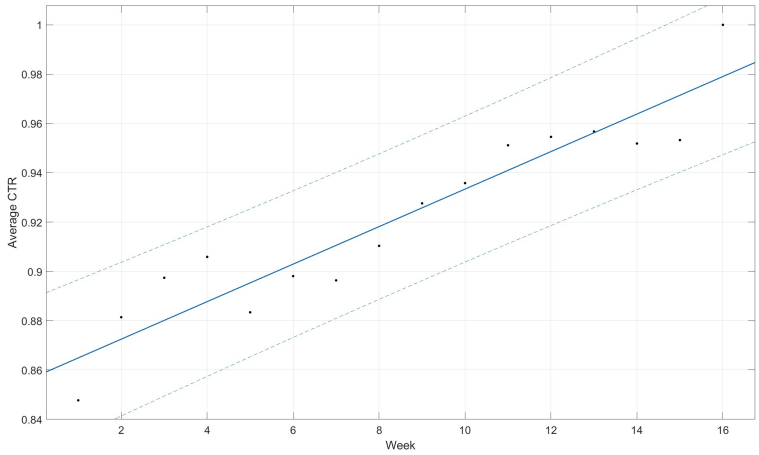
Figure A4: Longer-Term Effects of API Removal on Query Volume



Notes: Error bands represent 95% confidence intervals.

A.2.4 Temporal Variation in CTR among Users in the Control Condition

Figure A5: Temporal Variation in CTR among Users in the Control Condition



Notes: For confidentiality purposes, we added an undisclosed constant to the weekly average clickthrough rates (CTRs). Error bands represent 95% confidence intervals.

Appendix B Treatment Effects by Query Type

In this section, we explore how treatment effects vary by query popularity. When users enter a query term, the company's natural language processing tool can classify the content category of each term into 31 first-level categories (e.g., Health) and 167 second-level categories (e.g., Health-Disease). Conceptually, we can define the query popularity based on the number of searches for a specific content category (e.g., query in a popular or niche content category).

However, we did not observe the information about query type before the experiment (unlike demographics, active days), so we cannot use pre-experimental query popularity as a potential source of heterogeneity. Using the post-treatment query popularity might cause a bad control problem (Angrist and Pischke, 2009). Therefore, we constructed the popular-niche threshold and split the sample based on clicks from users in the control condition on the first day of the experiment. Further, we examine only the first search occasion of users to restrict our analysis to their first exposure to the treatment. Our rationale is that, like any random sample of the user base of the focal search product, users in the control condition represent the status quo. Therefore, under the assumption of no spillovers across the two conditions, the variation in query popularity in the control condition shall mirror the population of interest and not be affected by the treatment. Further, focusing on the first day of the experiment alleviates the concern of violating the assumption of no spillovers across the two conditions.

Therefore, we use the search records of users that clicked at least once and define the content categories that generate 75% of the clicks in the control condition in the first day of the experiment 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). We verified that the user demographics and usage activities are balanced among users in the treatment and control conditions in this subsample (see Table B1). Columns (1) and (2) suggest that the negative treatment effect is driven by popular queries (see Table B2). A plausible explanation is that the candidate items from the market leader's API are generated based on real searches by all users on the market leader's platform but are not personalized. In contrast, clicking niche content is not affected by API removal because recommending niche

content relies more on personal data. This demonstrates that external candidate items informed by the market leader's real searches may provide better suggestions for popular queries that requires less personalization.

Table B1: Randomization Checks

User Characteristics	Control	Treatment	p. value
Male	0.515	0.513	0.152
	(0.001)	(0.001)	
Larger Cities	0.539	0.538	0.679
	(0.001)	(0.001)	
Smaller Cities	0.456	0.456	0.787
	(0.001)	(0.001)	
Mobile Operating System: Apple iOS	0.116	0.116	0.541
	(0.001)	(0.001)	
Mobile Operating System: Android	0.883	0.884	0.539
	(0.001)	(0.001)	
Active days in the past 30 days (search activities)	100	99.579	0.212
	(0.239)	(0.238)	
Query views in the past 30 days (search activities)	100	99.877	0.841
	(0.430)	(0.435)	

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 at least once in the first day for each user. 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 B2: Treatment Effects by Query Type

Variables	(1) Lift in CTR (Popular)	(2) Lift in CTR (Niche)
API Removal	-0.0134***	0.0021
	(0.0017)	(0.0050)
Observations	764,119	764,119

* $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.