

Trade-offs in Leveraging External Data Capabilities: Evidence from a Field Experiment in a Search Market

Ananya Sen (CMU)

with Xiaoxia Lei (Shanghai Jiao Tong University) and Yixing Chen (University of Notre Dame)

External Data as a Lever for Product Growth?

- Practitioners highlight the need to use external data:
 - Firms may gain an edge by incorporating external data to build their data ecosystems ([Deloitte Insights 2019](#); [McKinsey 2021](#))
 - Data sharing through large players' application programming interfaces (APIs) is increasingly common ([Fatemi 2019](#)): e.g., Google search API for publishers and developers
- Despite its economic relevance, it is challenging to pin down its causal impact
 - Firms may self-select into API adoption ([Benzell, Hersh, and Van Alstyne 2022](#)).

New Regulations

What the European DSA and DMA proposals mean for online platforms

January 14, 2021 | [Aline Blankertz](#) and [Julian Jaurisch](#)



European Commissioner for a Europe Fit for the Digital Age Margrethe Vestager and European Internal Market Commissioner Thierry Breton attend the presentation of the European Commission's data/digital strategy in Brussels, Belgium February 19, 2020. REUTERS/Yves

The Digital Markets Act imposes obligations on “gatekeepers”:

- Provide to any third-party providers of online search engines with access to ranking, query, click, and view (deidentified) data generated by end users (Article 6.11).

The latest guidelines published by the State Council in China proposes a 20-point agenda around the data economy:

- data sharing to enable growth of small and medium sized companies. Sharing should not compromise personal information or “public interest”.

Recent Court Verdict: US DOJ vs. Google

Google avoids break-up but must share data with rivals

2 September 2025

Share  Save 

Lily Jamali North America Technology Correspondent, San Francisco
and **Rachel Clun** Business reporter, BBC News

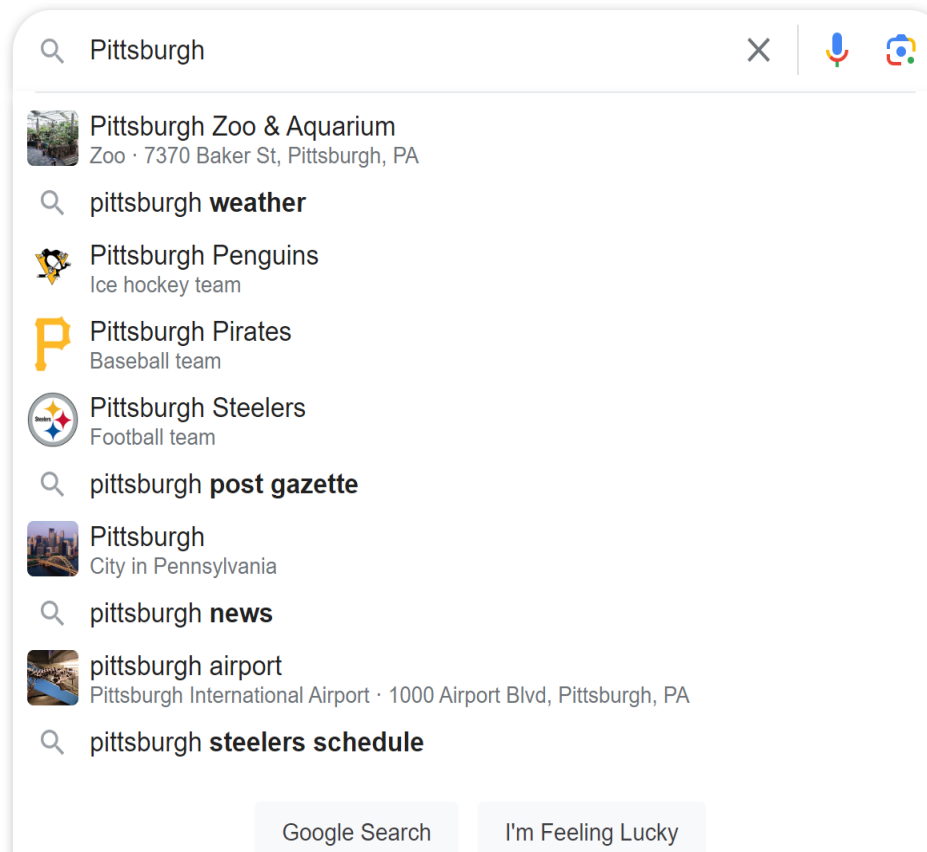
Questions and Overview of Results

1. What is the causal impact of access to the **market leader's data** on the **focal company's** product performance?
 - Removing access to market leader's data leads to a 4.6% decrease in CTR.
 - Downstream Elasticity of 0.12-0.18 (Search Engine Results Page)
2. Does the effect vary across **types of content**?
 - Only popular content affected.
3. Impact in the **short term vs. longer term**?
 - Average effect is much smaller than short-run decline in performance.
 - Using the API can impede improved prediction due to internal data.

Empirical Context

- Partnership with a leading Chinese technology company
 - Millions of monthly active users
- An app with hybrid functions:
 - News feed, streaming, eBooks, search engine, file management
- Our focus: search suggestion, a product developed by the company in 2020.
 - A start-up like team within a larger company.
 - A new product embedded within a “super-app”.

Economic Relevance




- An early application of generative AI models. (Serban et al. 2016)
- Bridge the gap between users' intent and content consumption (Agrawal, Gans, and Goldfarb 2018)
- Clicks imply revenue (sponsored words)
- Hence, our outcomes of interest:
 - (a) Click-through rate ($CTR = \text{Clicks} / \text{Exposures}$)
(also probability of click, total number of clicks)
 - (b) Downstream: Top Slot Clicks on Search Results Page (SERP)

Google

coffee

Search

[Gmail](#) [Images](#) [Sign in](#)



coffee

coffee near me

coffee holliston

coffee framingham

coffee shops near me

coffee nearby

coffee table

coffee break

coffee maker

Google Search

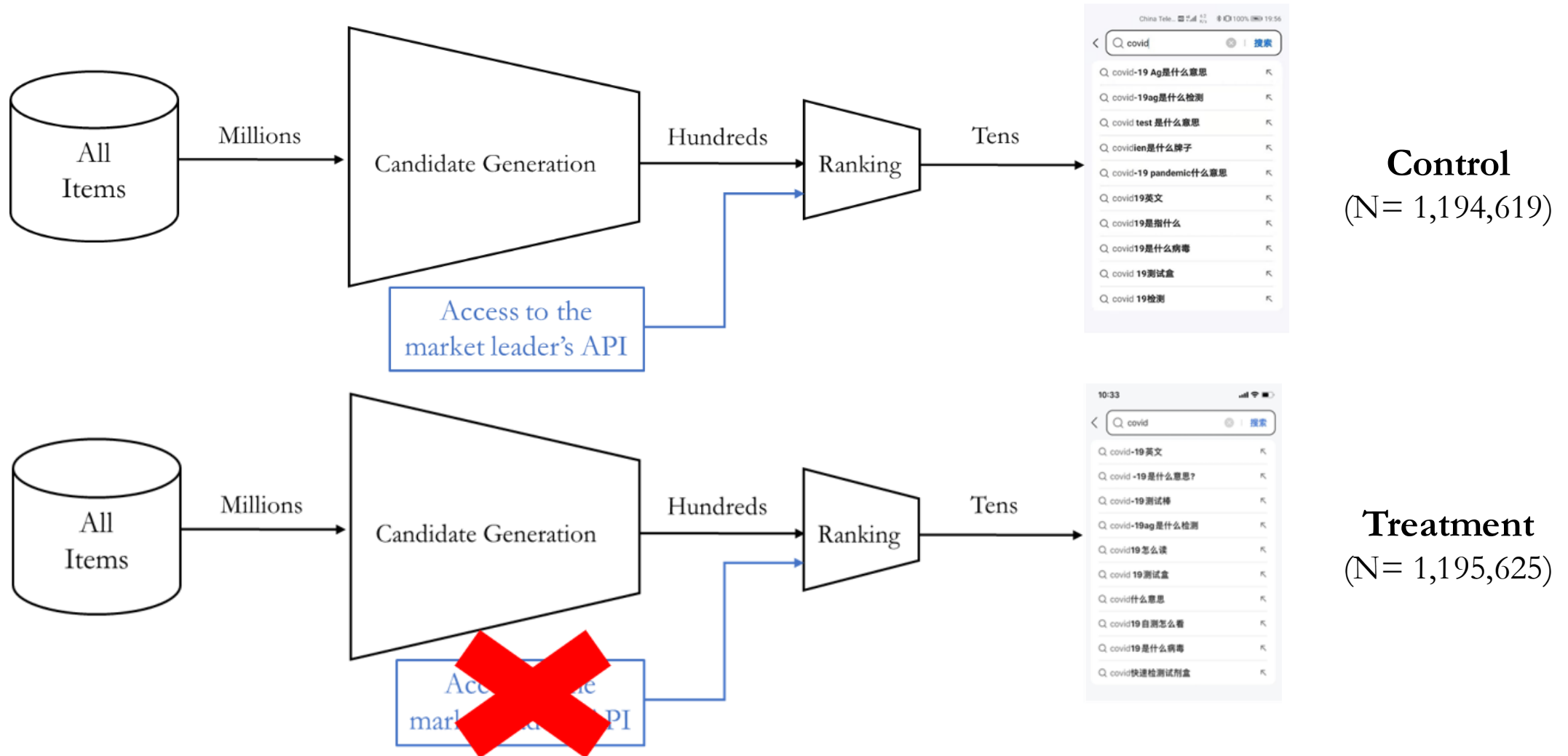
I'm Feeling Lucky

Report inappropriate predictions

Tools

```
    "autocomplete_results_state": "Showing completion results.",
  },
  "suggestions": [
    {
      "value": "coffee near me",
      "relevance": 1250,
      "type": "QUERY",
      "serpapi_link": "https://serpapi.com/search.json?engine=google_autocomplete&q=coffee+near+me"
    },
    {
      "value": "coffee holliston",
      "relevance": 601,
      "type": "QUERY",
      "serpapi_link": "https://serpapi.com/search.json?engine=google_autocomplete&q=coffee+holliston"
    },
    {
      "value": "coffee framingham",
      "relevance": 600,
      "type": "QUERY",
      "serpapi_link": "https://serpapi.com/search.json?engine=google_autocomplete&q=coffee+framingham"
    },
    {
      "value": "coffee shops near me",
      "relevance": 554,
      "type": "QUERY",
      "serpapi_link": "https://serpapi.com/search.json?engine=google_autocomplete&q=coffee+shops+near+me"
    },
    {
      "value": "coffee nearby",
      "relevance": 553,
      "type": "QUERY",
      "serpapi_link": "https://serpapi.com/search.json?engine=google_autocomplete&q=coffee+nearby"
    },
  ],
}
```


Between-Subject (3.5 Month Long) Field Experiment



- Provide to any third-party providers of online search engines with access to **ranking**, query, click, and view data generated by end users (Article 6.11).

Baseline Results

VARIABLES	(1)	(2)
	Lift CTR	Lift CTR
API Removal	-0.0462*** (0.001)	-0.0539*** (0.0013)
Controls	N	Y
Observations	2,390,244	1,932,886

- Back of the envelope (including SERP) suggests this is economically meaningful.

Impact on Types of Content: Popular vs. Niche

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

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

- Popular vs. niche: **due to depersonalization**, the long tail suffers.

Downstream Effects



Google Search

I'm Feeling Lucky

Report inappropriate predictions



Google

Pittsburgh restaurants

×

Pittsburgh Magazine
<https://www.pittsburghmagazine.com> › here-are-the-2...

Here Are The 25 Best Restaurants in Pittsburgh
May 5, 2023 — The List · 40 North at Alphabet City · Alta Via Ristorante · Altius · Apteka · Back to the Foodture · Bar Marco · Butterjoint · Casbah · Chengdu ...
[Back To The Foodture](#) · [Butterjoint](#) · [Dianoia's Italian Eatery](#)

Discover the Burgh
<https://www.discovertheburgh.com> › best-restaurants-i...

A Local's Guide to the Best Restaurants in Pittsburgh
Nov 1, 2023 — Italian · DiAnoia's — Strip District — \$\$ to \$\$\$ · Dish Osteria — South Side — \$\$ to \$\$\$ · La Tavola — Mount Washington — \$\$\$ · Girasole — ...
[Apteka Pittsburgh Review](#) · [Morcilla Pittsburgh Review](#) · [DiAnoia's Review](#)

Tripadvisor
<https://www.tripadvisor.com> › ... › Pittsburgh

THE 10 BEST Restaurants in Pittsburgh ...
Results 1 - 30 of 2220 — **Restaurants in Pittsburgh** ; Coughlin's Law Kitchen and Ale House · 36. American, Bar, Pub ; Pasha Cafe Lounge · 108. Cafe, Mediterranean, ...

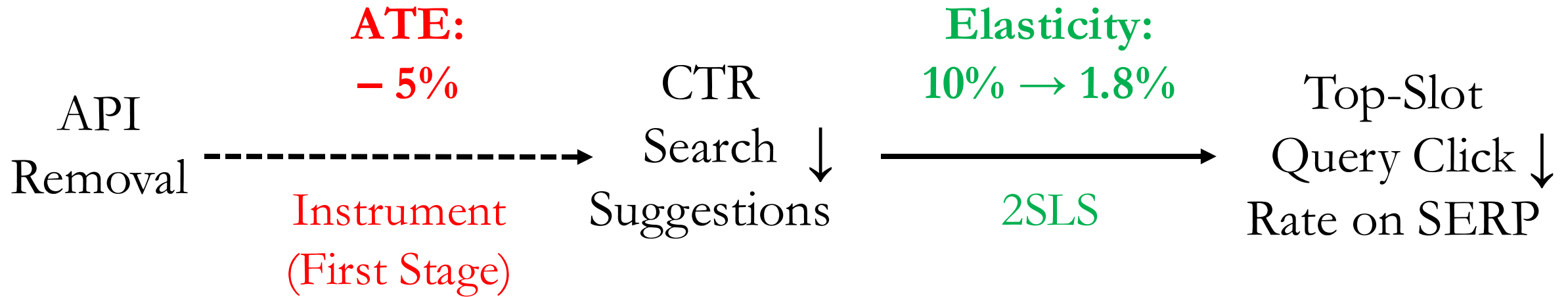
What are the most popular restaurants in Pittsburgh?

What are the best restaurants in Pittsburgh that deliver?

Visit Pittsburgh
<https://www.visitpittsburgh.com> › blog › top-places-to...

Top Places to Eat in Downtown Pittsburgh
Top Places to Eat in Downtown **Pittsburgh** · 1. Con Alma · 2. The Speckled Egg · 3. Bae Bae's Kitchen · 4. Gaucho Parrilla Argentina · 5. Nicky's Thai Kitchen · 6 ...
[3. Bae Bae's Kitchen](#) · [6. Tākō](#) · [16. Alihan's Coffee &...](#)

Implications for Search Engine Results Page (SERP)

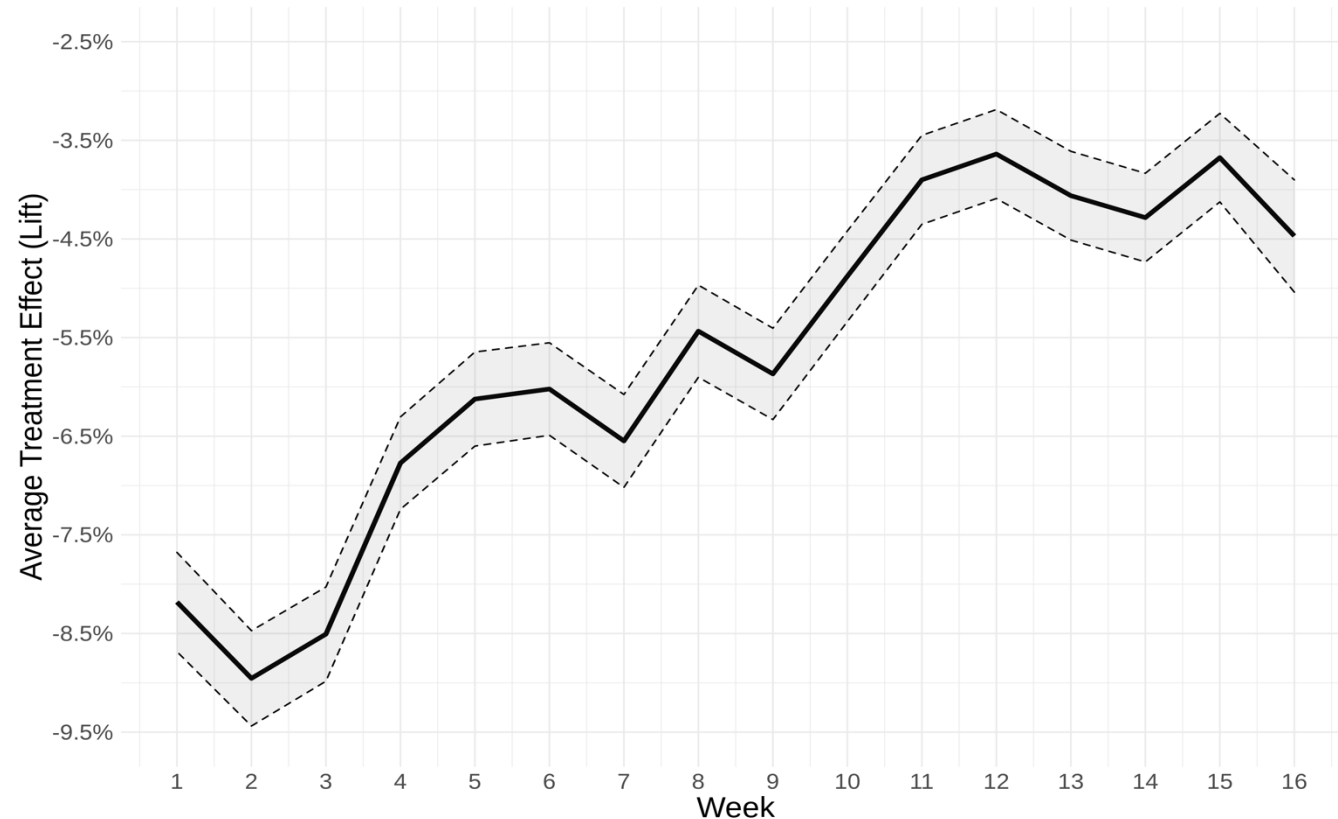


$$\log(SERP_i) = \mu_1 + \mu_2 \log(\widehat{SUGG}_i) + \vartheta_i$$

$$\log(SUGG_i) = \gamma_0 + \beta \times APIRemoval_i + \epsilon_i$$

- Fang, Chen, Farronato, Yuan (2023): 3.2% in orders due to text-based search aid.
- Burtch, Kwon, and Tong (2023): 1.2% increase in sales due to keyword recommender system.

Longer-Term Effects: CTR



- Magnitude of the effect is half as large in longer run relative to first few weeks.
- We rule out differential attrition, diminishing returns, other behavioral adjustments.
- We posit the role of improved prediction due to internal data.

(Some) Suggestive Evidence for Learning Effects

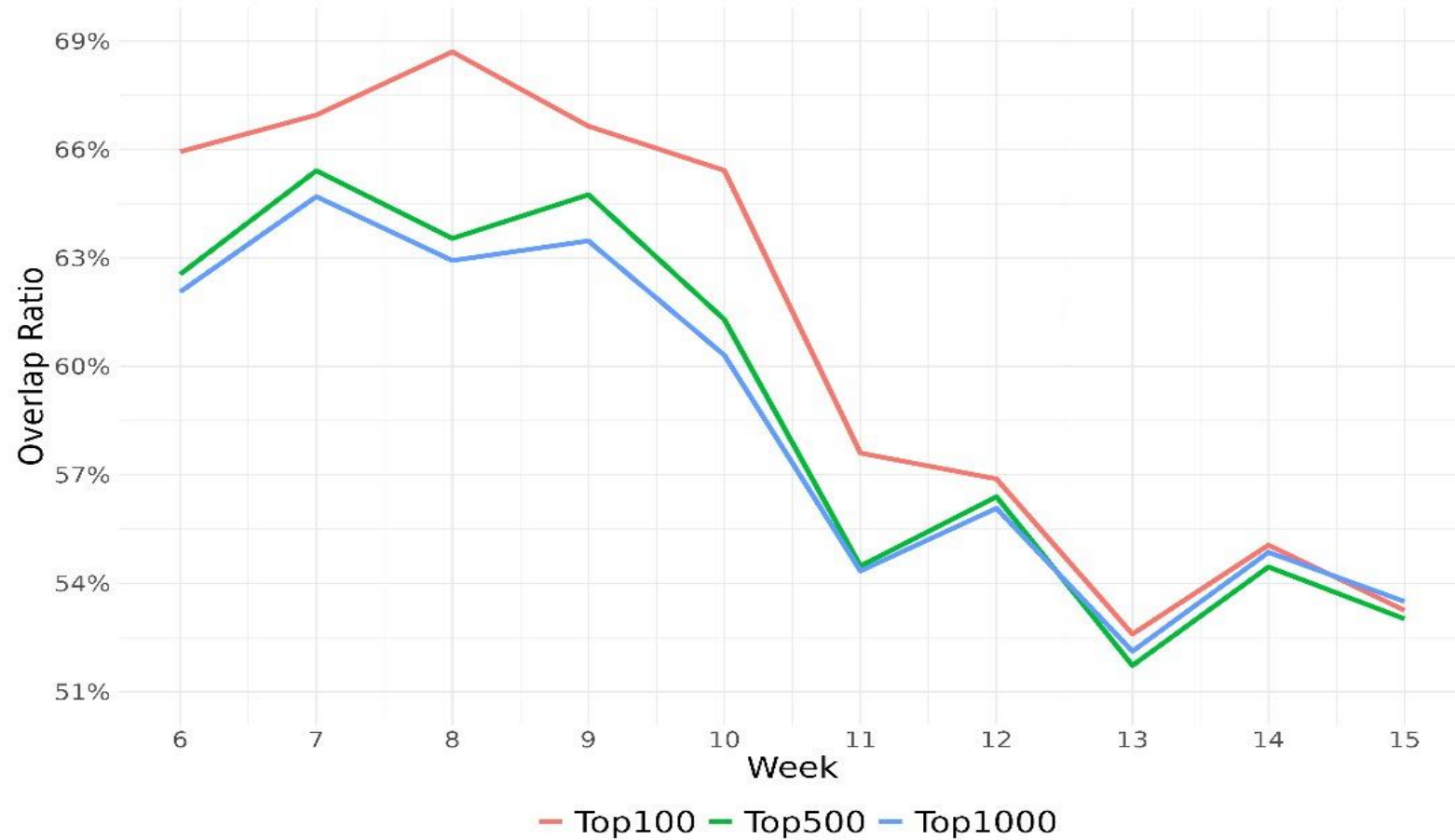
Contextual Information: Learning from API Candidates?

- All else equal, if the company's algorithmic system is learning from API candidates over time, the longer-term effect should **become more negative**.
- Lack of features of API candidates makes model training challenging (Duan and Lalor, 2023).
- Further, even if training is feasible, it can be legally prohibited. e.g., OpenAI API.



(c) Restrictions. You may not (i) use the Services in a way that infringes, misappropriates or violates any person's rights; (ii) reverse assemble, reverse compile, decompile, translate or otherwise attempt to discover the source code or underlying components of models, algorithms, and systems of the Services (except to the extent such restrictions are contrary to applicable law); (iii) use output from the Services to develop models that compete with OpenAI; (iv) except as permitted through the API, use any automated or programmatic method to extract data or output from the Services, including scraping, web harvesting, or web data extraction; (v) represent that output from the Services was human-generated when it is not or otherwise violate our Usage Policies; (vii) buy, sell, or transfer API

Improved Prediction based on Internal Data?



- Suggestions in T become more different relative to C over time.
- Combined with increased CTR might suggest improved prediction based on internal data

Improved Prediction based on Internal Data? Within and Across-User Learning

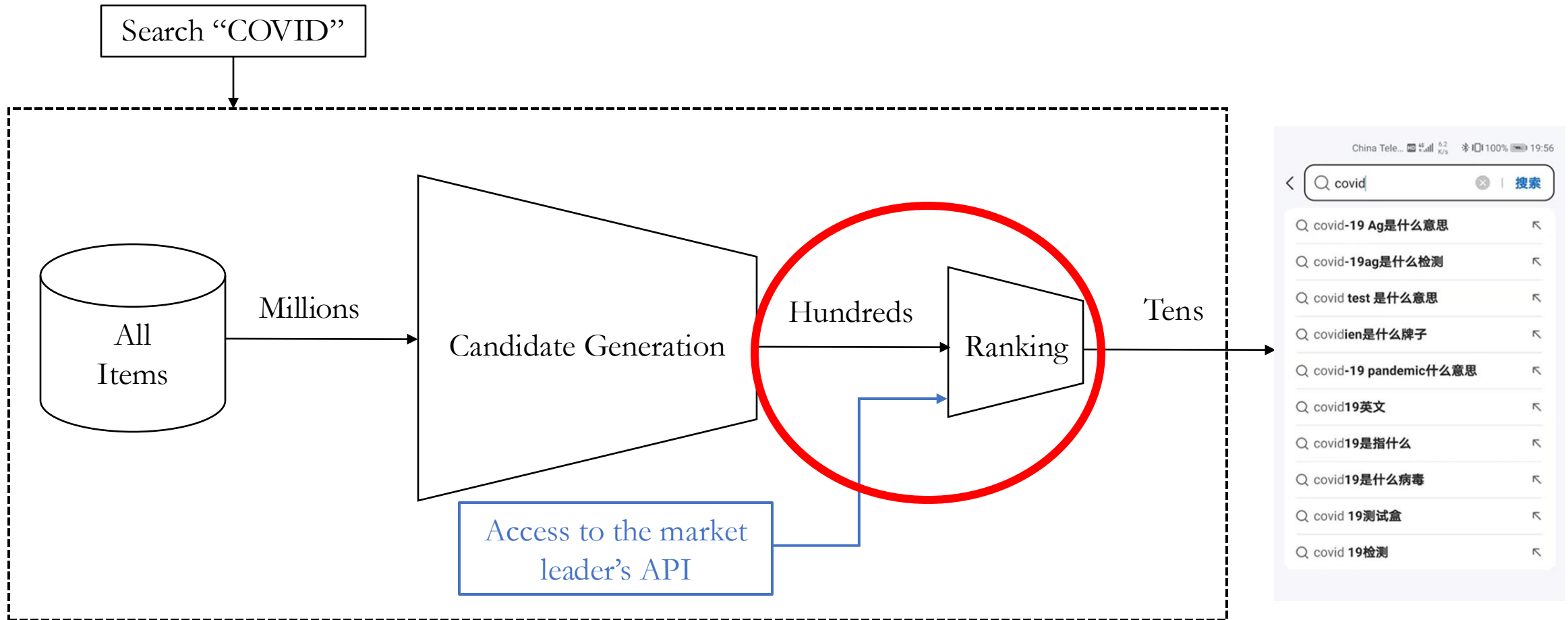
User-query-day data

Variables	(1) CTR	(2) CTR	(3) CTR
API Removal	-0.0030*** (0.0005)	-0.0171*** (0.0040)	-0.0044*** (0.0009)
API Removal \times Repeated Query		0.0145*** (0.0041)	
API Removal \times Query Histories			0.0004** (0.0002)
Unit of analysis	User-Query-Day	User-Query-Day	User-Query-Day
Query fixed effects	✓	✓	✓
Day fixed effects	✓	✓	✓
R^2	0.2289	0.2289	0.2289
Observations	1,636,900	1,636,900	1,636,900

- Treatment effect is significantly smaller for repeat queries within and across individuals.
- Suggests learning in the treated group.

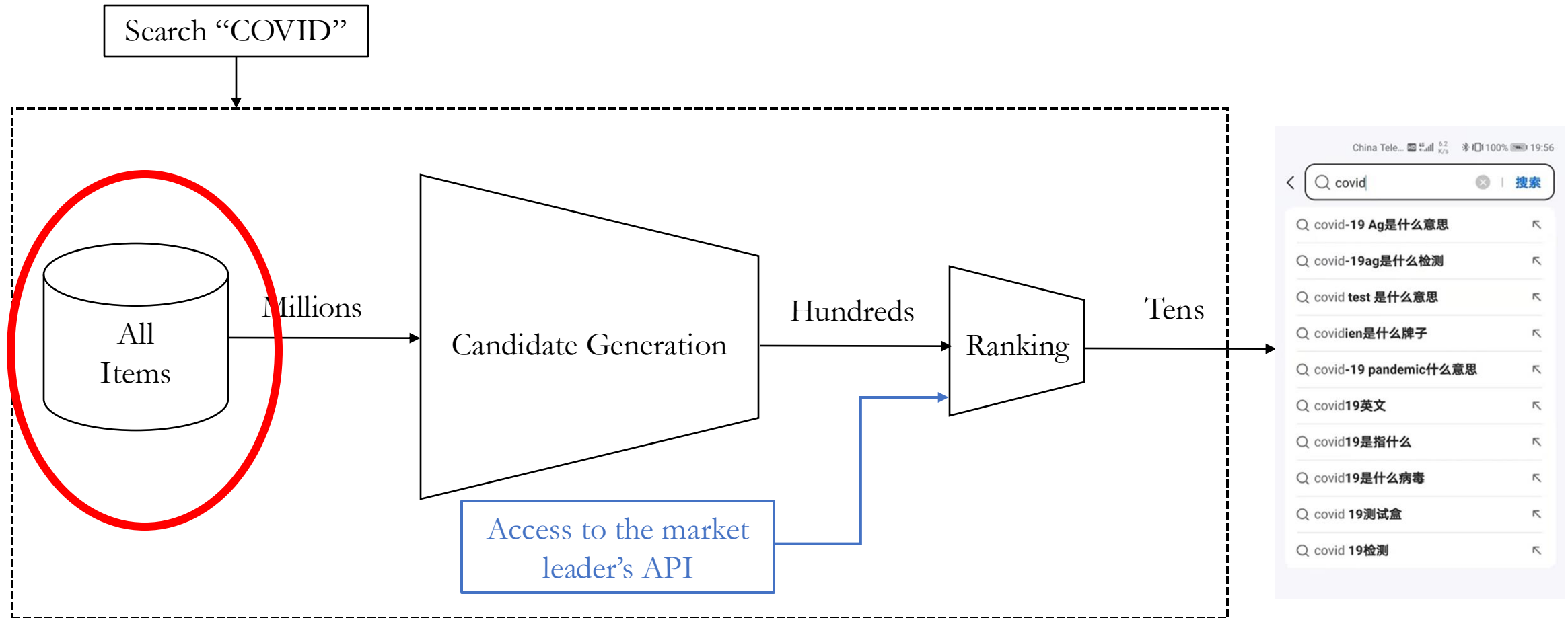
Some Additional Field Experiments: Training Data

Till Now



1. Candidate generation uses simpler algorithms to evaluate a larger set of items, while ranking uses more sophisticated and computationally intensive algorithms (Nandy et al. 2021).
2. A sizable proportion of final search suggestions come from the market leader's API.

Training Data Experiments



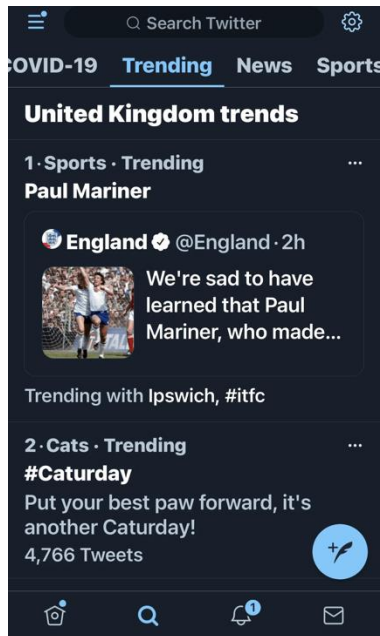
1. Candidate generation uses simpler algorithms to evaluate a larger set of items, while ranking uses more sophisticated and computationally intensive algorithms (Nandy et al. 2021).
2. A sizable proportion of final search suggestions come from the market leader's API.

Training Data : Public Data versus Internal Data

Public Items

Focal Firm's
Internal Items

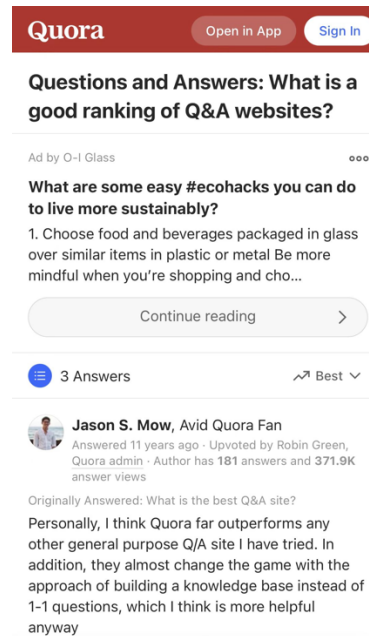
Trending News



UGC



Q&A

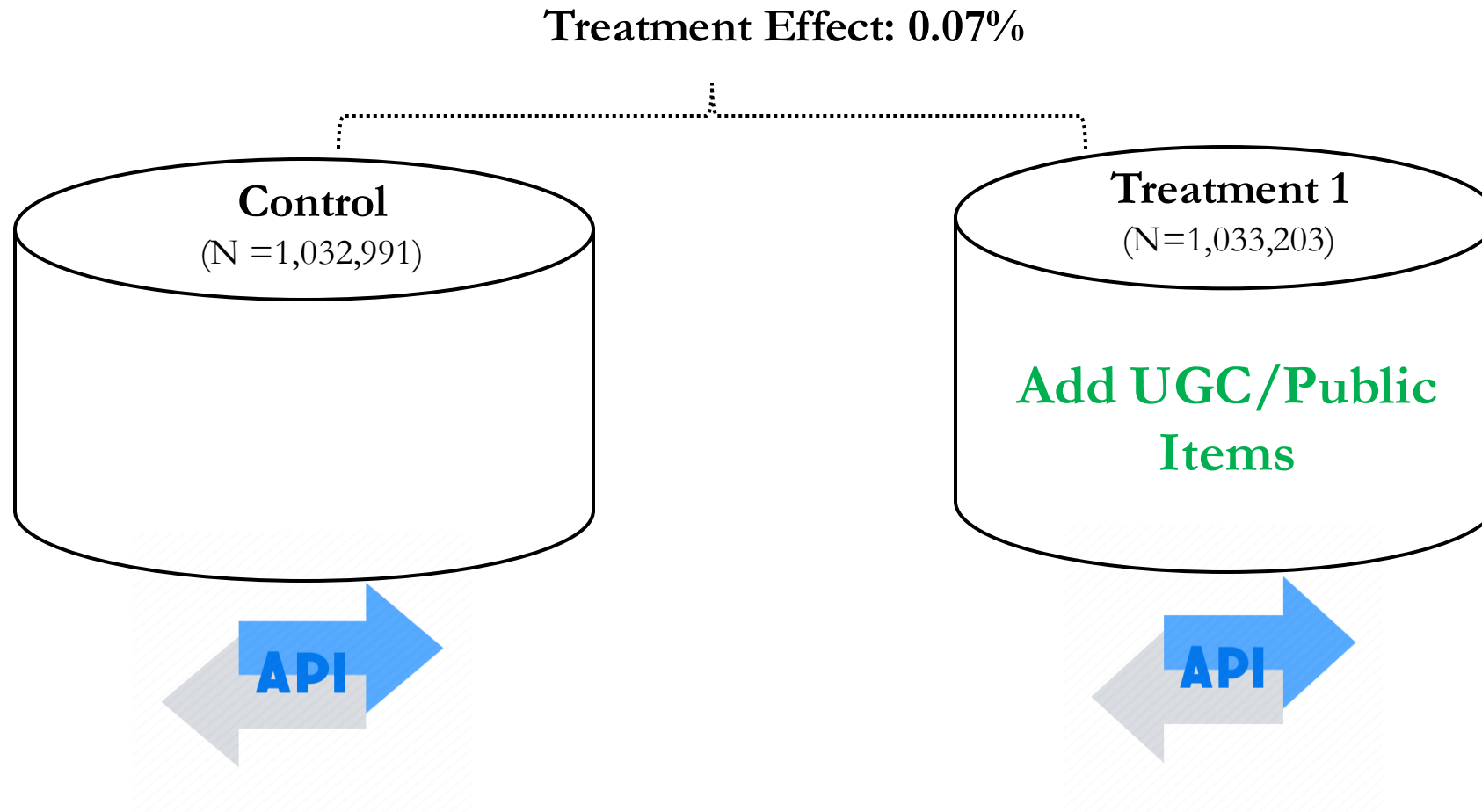


Content resources by the focal firm
([update real-time](#)): e.g.,

- Articles and video clips by content creators
- Videos (e.g., TV shows)
- Real-time new resources from other focal firm's products (e.g., news feed, eBooks streaming)
- Users' search histories

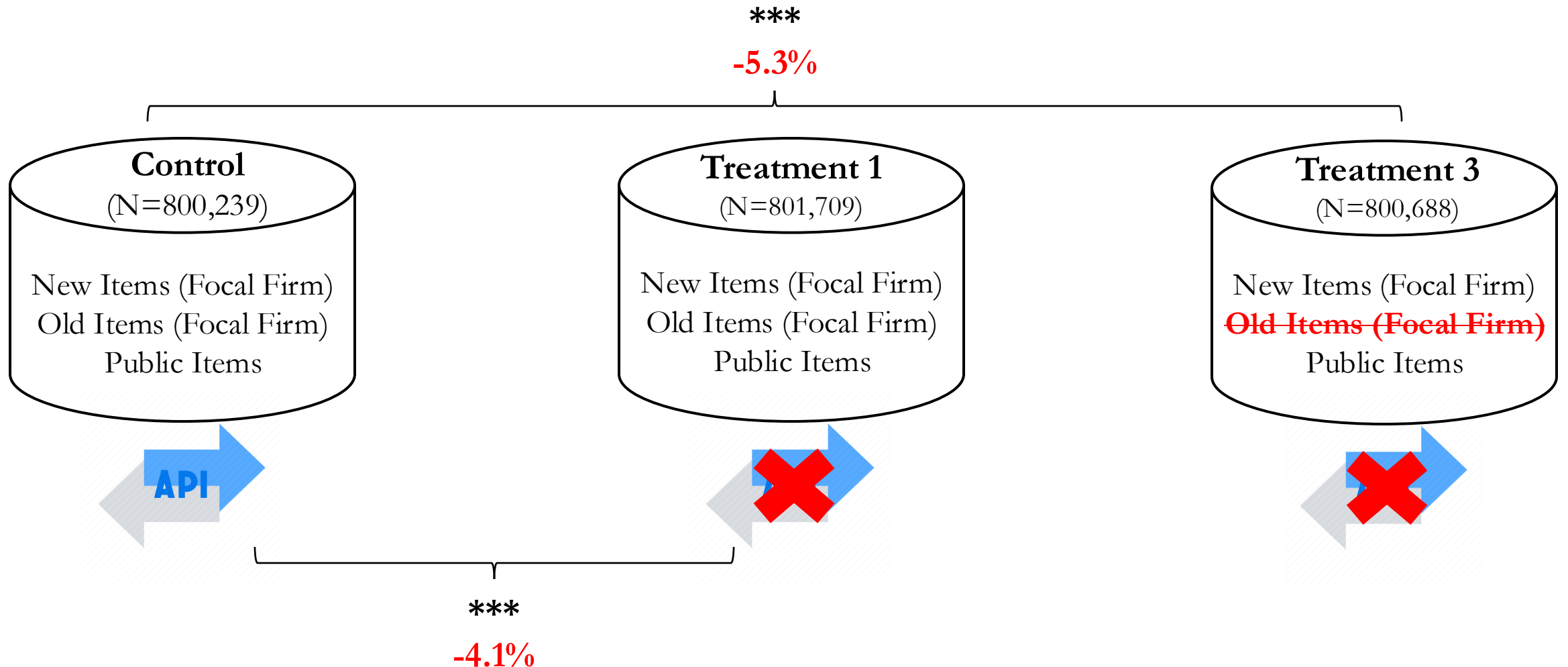
Experiment 2: Add Public Data & Remove Competitor Data

- Addition of User Generated Content doesn't seem to impact performance.



Experiment 3: Remove Competitor Data & Own Data

- Removal of older proprietary data had modest impact.




EVIDENCE-DRIVEN POLICY FRAMEWORKS TO UNLOCK THE POWER OF DATA

BLOG

EVIDENCE-DRIVEN POLICY FRAMEWORKS TO UNLOCK THE POWER OF DATA

September 15, 2025

 [Ananya Sen](#), [Michael D. Smith](#) and [Rahul Telang](#)

ANTITRUST AND COMPETITION

During the past two decades, recommender systems and targeted advertising have used personal data collected at scale in their quest to transform how we consume news, shop, and interact online. The more relevant and high-quality the data, the better the results—or at least that’s been the guiding assumption. That principle now appears to hold true for the new wave of GenAI products as well. But if data matters, then so do the questions of when,

RECOMMENDED READS



NOVEMBER 25, 2024

**TIMELINE OF MAJOR ANTITRUST
ACTIONS, LAWS, AND REPORTS
IN THE US AND EU THROUGH
2024**

COMPILATIONS

MAY 5, 2010

**GOOGLE'S ACQUISITION OF
ADMOB**