

Generative AI as a market intermediary

David Holtz
Columbia Business School

A discussion of:

- **Ads that Talk Back: Implications and Perceptions of Injecting Personalized Advertising into LLM Chatbots (Tang et al.)**
- **When AI Disclosure Backfires: The Economic Consequences of Labeling AI-Generated Review Summaries (Choi et al.)**
- **Soft Deception by Design: Algorithmic Positivity Bias in AI-Generated Consumer Review Summaries (Quan et al.)**

Connecting to the literature on reputation system design

- A large literature focuses on how the design of reputation systems impacts market growth/structure
- This literature has largely focused on the “numerical” aspects of reputation systems
- This is despite a general understanding that text is equally (if not more) important
- AI/LLMs opens up the “design space” – specifically for review text, but also more generally



The screenshot shows the top portion of a journal article page from Marketing Science. The header is dark blue with the journal title 'MARKETING SCIENCE' in white. Below the header is a navigation bar with links for 'JOURNAL HOME', 'ARTICLES IN ADVANCE', 'CURRENT ISSUE', 'ARCHIVES', and 'ABOUT'. There are also 'SUBMIT' and 'SUBSCRIBE' buttons. The article title is 'Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb' by Andrey Fradkin and David Holtz. The page includes a 'View PDF' button, 'Tools' and 'Share' options, and an abstract section. The abstract discusses how online reputation systems often rely on volunteer reviews, which can be self-selected, and how a randomized experiment on Airbnb showed that incentivizing reviews led to more negative feedback without affecting sales or revenue.

MARKETING SCIENCE

JOURNAL HOME ARTICLES IN ADVANCE CURRENT ISSUE ARCHIVES ▾ ABOUT ▾

SUBMIT SUBSCRIBE

Open Access | CC

Home > Marketing Science > Vol. 42, No. 5 >

Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb

Andrey Fradkin , David Holtz 

Published Online: 7 Apr 2023 | <https://doi.org/10.1287/mksc.2023.1439>

Abstract

Many online reputation systems operate by asking volunteers to write reviews for free. As a result, a large share of buyers do not review, and those who do review are self-selected. This can cause the reputation system to miss important information about seller quality. We study the extent to which a platform can improve market outcomes by attempting to increase the amount and quality of information collected by its reputation system. We do so by analyzing a randomized experiment conducted by Airbnb. In the treatment, buyers were offered a coupon to review listings that had no prior reviews. In the control, buyers were not offered any incentive to review. We find that, although the treatment induced additional reviews that were more negative on average, these reviews did not affect the number of nights sold or total revenue. Furthermore, we find that, contrary to the treatment’s intended effect, Airbnb’s incentivized program caused transaction quality for treated sellers to fall. We examine how the quality of the induced reviews, market conditions, and the design of Airbnb’s reputation system can explain our findings.

Go to Section

Abstract

1. Introduction

Connecting to the literature on reputation system design

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT

JOURNAL HOME ARTICLES IN ADVANCE CURRENT ISSUE ARCHIVES ABOUT

SUBMIT SUBSCRIBE Search this Journal

Home > Manufacturing & Service Operations Management > Vol. 23, No. 3 >

Designing Informative Rating Systems: Evidence from an Online Labor Market

Nikhil Garg, Ramesh Johari

Published Online: 16 Dec 2020 | <https://doi.org/10.1287/msom.2020.0921>

View PDF Tools Share

Abstract

Problem definition: Platforms critically rely on rating systems to learn the quality of market participants. In practice, however, ratings are often highly inflated and therefore, not very informative. In this paper, we first investigate whether the platform can obtain less inflated, more informative ratings by altering the *meaning* and *relative importance* of the levels in the rating system. Second, we seek a principled approach for the platform to make these choices in the design of the rating system. **Academic/practical relevance:** Platforms critically rely on rating systems to learn the quality of market participants, and so, ensuring these ratings are informative is of first-order importance. **Methodology:** We analyze the results of a randomized, controlled trial on an online labor market in which an additional question was added to the feedback form. Between treatment conditions, we vary the question phrasing and answer choices; in particular, the treatment conditions include several *positive-skewed verbal rating scales* with descriptive phrases or adjectives providing specific interpretation for each rating level. We then develop a model-based framework to compare and select among rating system designs and apply this framework to the data obtained from the online labor market test. **Results:** Our test reveals that current inflationary norms can be countered by reanchoring the meaning of the levels of the rating system. In particular, positive-skewed verbal rating scales yield substantially deflated rating distributions that are much more informative about seller quality. Further, we demonstrate that our model-based framework for scale design and optimization can identify the most informative rating system and substantially improve the quality of information obtained over baseline designs. **Managerial implications:** Our study illustrates that practical, informative rating systems can be designed and demonstrates how to compare and design them in a principled manner.

Article Information

Supplemental Material

Metrics

Downloaded 90 times in the past
Cited 28 times

Information

Received: October 05, 2018
Accepted: June 11, 2020
Published Online: December 16, 2020
Copyright © 2020, INFORMS

Cite as

Nikhil Garg, Ramesh Johari (2020) Informative Rating Systems: Evidence from an Online Labor Market. Manufacturing & Service Operations Management

MARKETING SCIENCE

JOURNAL HOME ARTICLES IN ADVANCE CURRENT ISSUE ARCHIVES ABOUT

SUBMIT SUBSCRIBE

Open Access |

Home > Marketing Science > Vol. 40, No. 6 >

Reciprocity and Unveiling in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb

Andrey Fradkin, Elena Grewal, David Holtz

Published Online: 8 Oct 2021 | <https://doi.org/10.1287/mksc.2021.1311>

View PDF Tools Share

Abstract

Reputation systems are used by nearly every digital marketplace, but designs vary and the effects of these designs are not well understood. We use a large-scale experiment on Airbnb to study the causal effects of one particular design choice—the timing with which feedback by one user about another is revealed on the platform. Feedback was hidden until both parties submitted a review in the treatment group and was revealed immediately after submission in the control group. The treatment stimulated more reviewing in total. This is due to users' curiosity about what their counterparty wrote and/or the desire to have feedback visible to other users. We also show that the treatment reduced retaliation and reciprocation in feedback and led to lower ratings as a result. The effects of the policy on feedback did not translate into reduced adverse selection on the platform.

Go to Section

Abstract

INFORMATION SYSTEMS RESEARCH

JOURNAL HOME ARTICLES IN ADVANCE CURRENT ISSUE ARCHIVES ABOUT

SUBMIT SUBSCRIBE

Home > Information Systems Research > Vol. 16, No. 2 >

Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard

Chrysanthos Dellarocas

Published Online: 1 Jun 2005 | <https://doi.org/10.1287/isre.1050.0054>

View PDF Tools Share

Abstract

This paper offers a systematic exploration of reputation mechanism design in trading environments with opportunistic sellers of commonly known cost and ability parameters, imperfect monitoring of a seller's actions, and two possible seller effort levels, one of which has no value to buyers. The objective of reputation mechanisms in such *pure moral hazard* settings is to induce sellers to exert high effort as often as possible. I study the impact of various mechanism parameters (such as the granularity of solicited feedback, the format of the public reputation profile, the policy regarding missing feedback, and the rules for admitting new sellers) on the resulting market efficiency. I find that maximum efficiency is bounded away from the hypothetical first-best case where sellers can credibly precommit to full cooperation by a factor that is related to the probability that cooperating sellers may receive "unfair" bad ratings. Furthermore, maximum efficiency is independent of the length of past history summarized in a seller's public reputation profile. I apply my framework to a simplified model of eBay's feedback mechanism and conclude that, in pure moral hazard settings, eBay's simple mechanism is capable of inducing the maximum theoretical efficiency independently of the number of recent ratings that are being summarized in a seller's profile. I derive optimal policies for dealing with missing feedback and easy online identity changes. Finally, I show that if the number of buyers is large, the results obtained in the monopoly case are also approximately valid in settings where multiple sellers of different reputations simultaneously offer auctions for identical goods.

Go to Section

Abstract

BOOK CHAPTER

Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system

By Paul Resnick; Richard Zeckhauser

+ Author and Other Information

DOI: [https://doi.org/10.1016/S0278-0984\(02\)11030-3](https://doi.org/10.1016/S0278-0984(02)11030-3)

Published: 2002

Share Tools Cite

© Authors

One of the earliest and best-known Internet reputation systems is run by eBay, which

Buy Print

Core takeaway from each paper

- **Quan and Shen:** Summarizing review text using large language models can be a non-neutral transformation of the underlying raw review data
- **Choi et al.:** The way that AI summaries are displayed and framed can have implications for market outcomes
- **Tang et al.:** As generative models become more capable, we will see more and more search / advertising embedded inside of LLMs.
 - LLMs have the potential to become general purpose online platforms
 - What does this mean for reputation system design / online platform design more generally?

Quan and Shen: *Soft Deception...*

Paper's focus: AI review summaries from a sample of 905 TripAdvisor Hotels

Reviews summary

This summary was created by AI, based on recent reviews.



The Westin Austin Downtown earns praise from many travelers for its convenient, walkable location near Austin's vibrant music scene and attractions. Guests often highlight the hotel's inviting atmosphere, spacious rooms, and comfortable beds.

The rooftop pool and amenities like the fitness center and restaurant receive rave reviews, though some mention crowding and service inconsistencies. While service is frequently described as friendly, issues with housekeeping and unexpected charges lead some to question the value for the price.

[Jump to all reviews ↓](#)

Location Convenient	Atmosphere Inviting
Rooms Spacious	Value Pricey
Cleanliness Mixed	Service Inconsistent
Amenities Well-appointed	

Quan and Shen: *Soft Deception...*

Key (striking!) result: AI summaries have a positive bias relative to the underlying reviews

Measure	M	SD	Min	Max
AI Summary Sentiment (AI_i)	0.94	0.10	-0.57	0.99
Review Sentiment ($Review_i$)	0.71	0.20	-0.30	0.97
Sentiment Gap (Gap_i)	0.23	0.18	-0.83	1.12

Minor question: how robust is this result to different approaches to measuring sentiment?

Quan and Shen: *Soft Deception...*

- **Why this matters:** Summarizing Nosko and Tadelis (2015), a bad reputation system can cause buyers to transact with sellers they believe are good but are not, and those bad experiences then lead buyers to update downward about the marketplace as a whole
- **Nosko and Tadelis' solution:** a new review summary metric, effective percent positive (EPP), that is less susceptible to distortions.
- **Key area for this paper to expand:** Is something similar to “the Nosko and Tadelis approach” possible? In other words, can AI be used to summarize reviews without inducing positivity bias?

Quan and Shen: *Soft Deception...*

- This paper asserts (perhaps implicitly) that AI systems will **always** amplify favorable language
 - This claim seems overly broad based on an empirical analysis of one dataset / one summarization model
- Areas to expand / improve:
 - Broaden analysis to a wider range of datasets beyond TripAdvisor (e.g., Amazon)
 - Take a more “platform / reputation system design” approach
 - How much favorability bias exists when different models are used? When different system prompts are used? etc.

Choi et al.: *When AI Disclosure Backfires...*

Paper's main question: How do the market impacts of AI-generated summaries vary as a function of (a) message framing and (b) source disclosure?

Approach: A long-term RCT with a leading automotive e-commerce platform in Asia

		Source Disclosure		
		Baseline (No Review Summary)	None	Human
Content Framing	Pos + Neg	(3.2.) Review Summary	<i>Human-labeled</i> Review Summary	<i>AI-labeled</i> Review Summary (3.5.)
	Pos	(3.3.) <i>Positive-only</i> Review Summary	<i>Positive-only</i> <i>Human-labeled</i> Review Summary	<i>Positive-only</i> <i>AI-labeled</i> Review Summary

Note: Red dashed arrows in the original image indicate a path from the Baseline cell to (3.2.), then to (3.3.), and finally to (3.4.).

Choi et al.: *When AI Disclosure Backfires...*

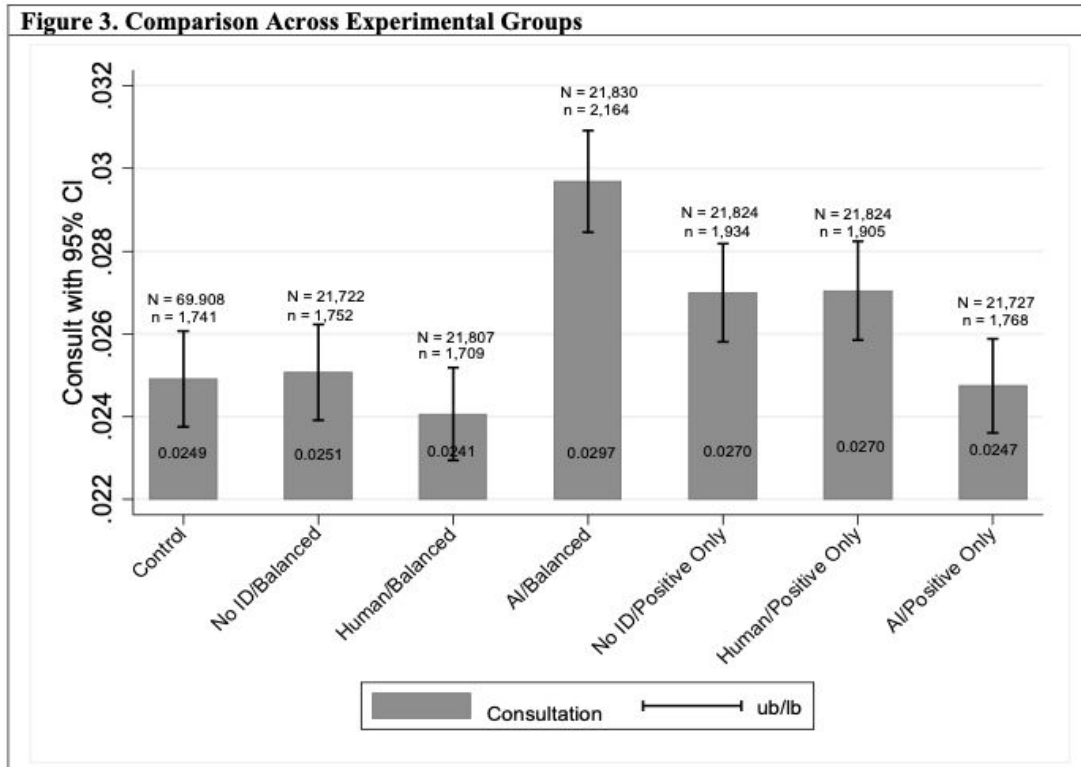
Answer: It's complicated. The authors assert an "AI disclosure paradox:" *disclosing AI authorship drives engagement in early decision stages, but suppressed conversion at later stages (relative to identical balanced summaries that are not labeled)*

Choi et al.: *When AI Disclosure Backfires...*

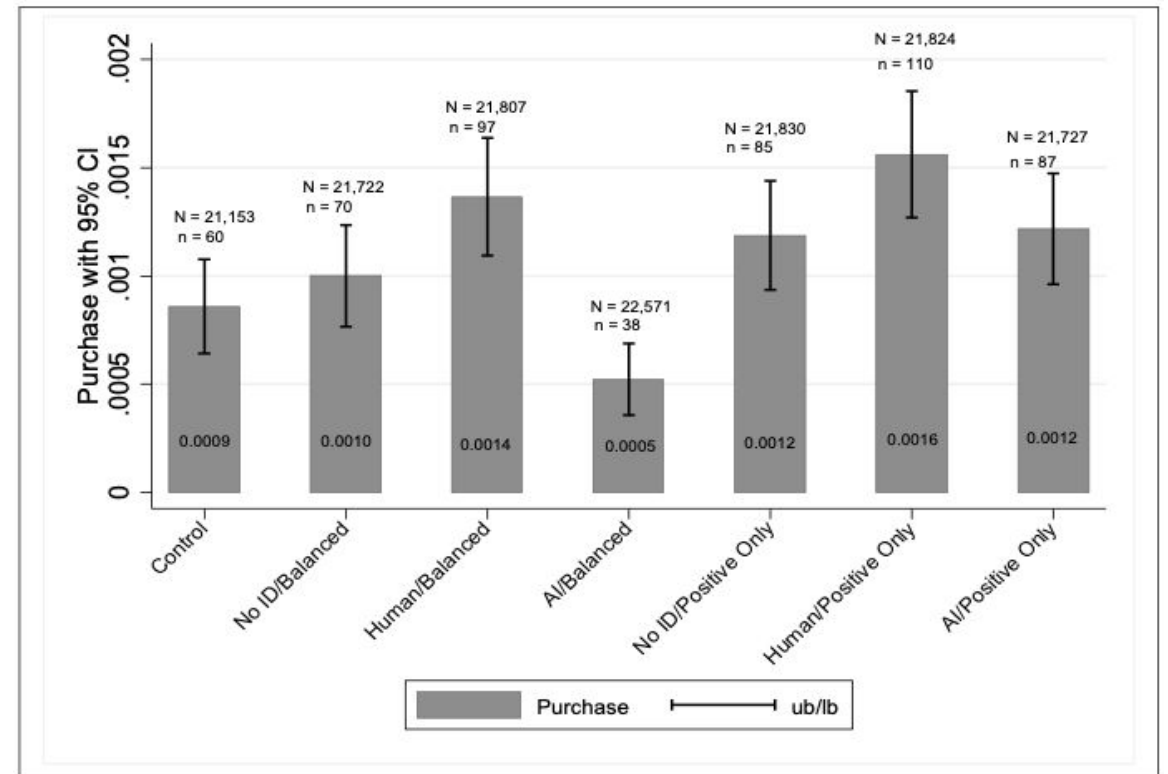
Answer: It's complicated. The authors assert an "AI disclosure paradox:" *disclosing AI authorship drives engagement in early decision stages, but suppressed conversion at later stages (relative to identical balanced summaries that are not labeled)*

Choi et al.: *When AI Disclosure Backfires...*

Model free evidence (consultations)



Model free evidence (purchases)



Choi et al.: *When AI Disclosure Backfires...*

Things to like about this paper

- The paper engages with these “reputation system design” aspects of LLM summaries of text:
 - How is the text transformed? How is that transformed text framed?
- The paper links these changes to system design back to *market outcomes*
 - This is, in my view, a weakness of many papers in the reputation systems literature
- Conducting a long-running RCT with a large online platform
 - This is super hard!

Choi et al.: *When AI Disclosure Backfires...*

Main area for improvement: statistical inference and framing

- The paper seems to struggle to synthesize results across so many treatment arms
 - “AI disclosure paradox” feels like it only speaks to a couple of treatment arms; might not be a general effect
- With so many treatment arms, multiple comparisons is a concern
 - Not clear how randomization inference solves this, Appendix E not included
- Model-free results don’t take into account correlation structure in the data
- Should model specifications also cluster standard errors at the automobile level?

Suggestion: Take advantage of the factorial design, and estimate a model based on variable levels, rather than treatment arm indicators

Choi et al.: *When AI Disclosure Backfires...*

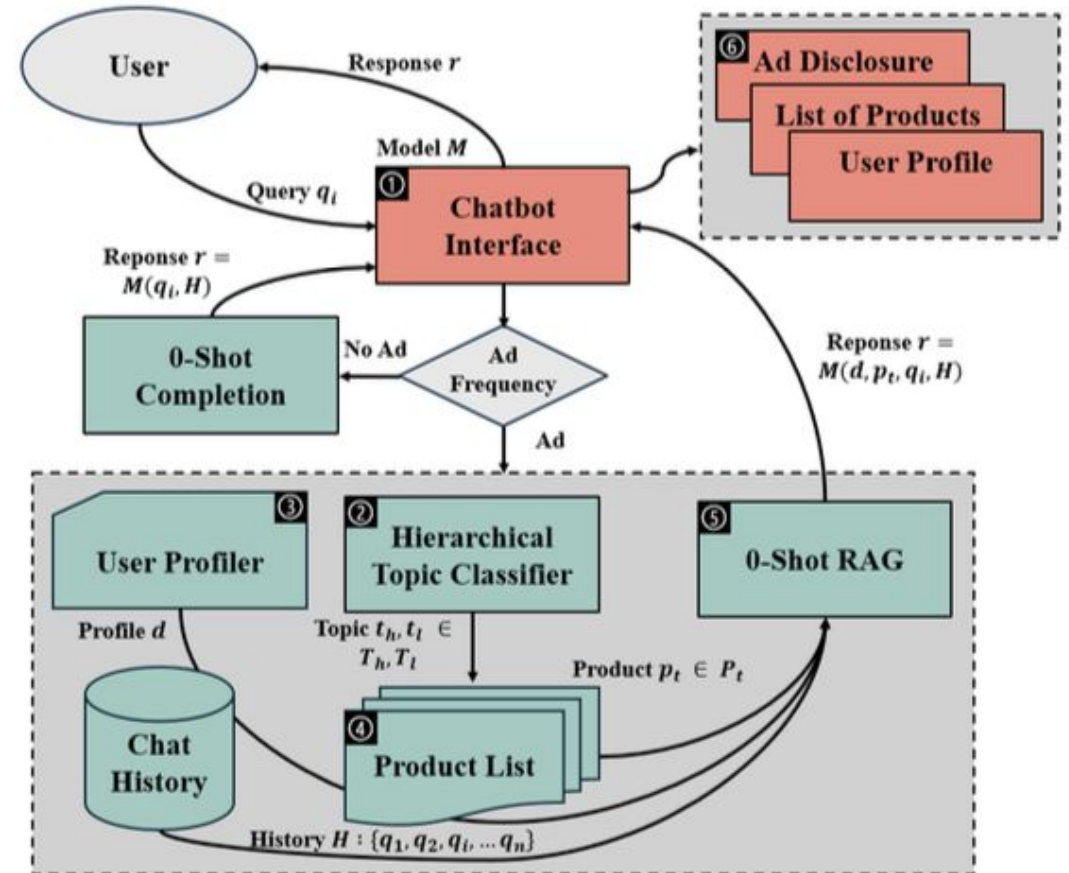
$$Y_{ij} = \alpha + \tau_S S_i + \tau_P P_i + \tau_H H_i + \tau_A A_i + \tau_{PH} (P_i \times H_i) + \tau_{PA} (P_i \times A_i) + X_{ij}' \gamma + \varepsilon_{ij}$$

- $S_i = 1$ if the user sees **any** summary, and 0 for the no-summary control.
- $P_i = 1$ if the summary is **positive-only**, and 0 if it is **balanced**.
- $H_i = 1$ if the source label is **human**.
- $A_i = 1$ if the source label is **AI**.
- Control group: $A_i = H_i = P_i = S_i = 0$.

Tang et al.: *Ads that Talk Back...*

What this paper aims to achieve (a lot of things!):

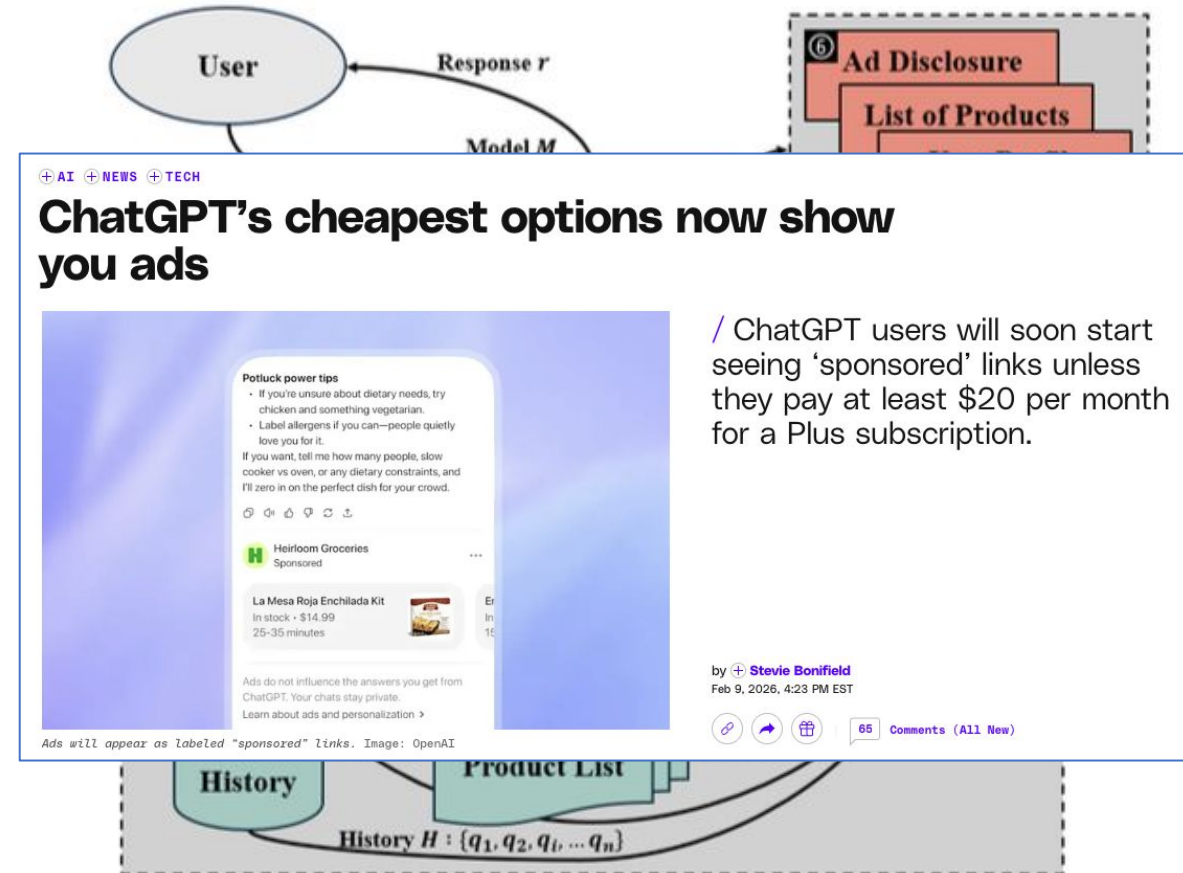
- How might we develop an LLM chatbot with embedded ads?
- How well can such an LLM profile and target users?
- Does embedding ads degrade on performance pre-existing benchmarks?
- How does the presence of ads affect user experience?



Tang et al.: *Ads that Talk Back...*

What this paper aims to achieve (a lot of things!):

- How might we develop an LLM chatbot with embedded ads?
- How well can such an LLM profile and target users?
- Does embedding ads degrade on performance pre-existing benchmarks?
- How does the presence of ads affect user experience?



Tang et al.: *Ads that Talk Back...*

What I liked about this paper:

- Moves beyond thinking about LLMs as a *component/feature* of platforms and asks, “what if LLMs are the platform?”
 - An interesting possibility to consider alongside with other visions of agentic commerce (e.g., Shahidi et al. 2025)
- Description of how the chatbot ad engine is designed lays bare all of the (sometimes subjective) choices that go into platform design
 - These choices have implications for external validity of the study’s findings
 - But open up a lot of interesting future “platform design” style research questions

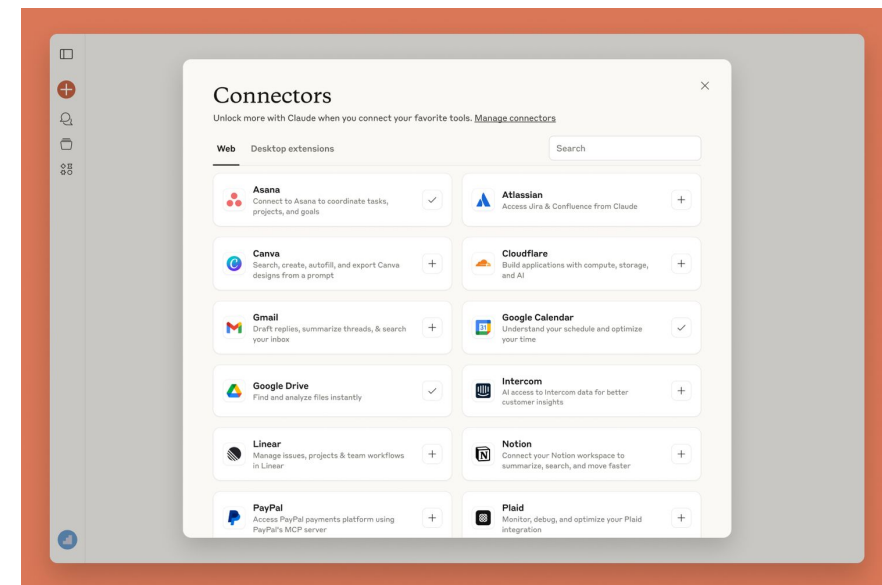
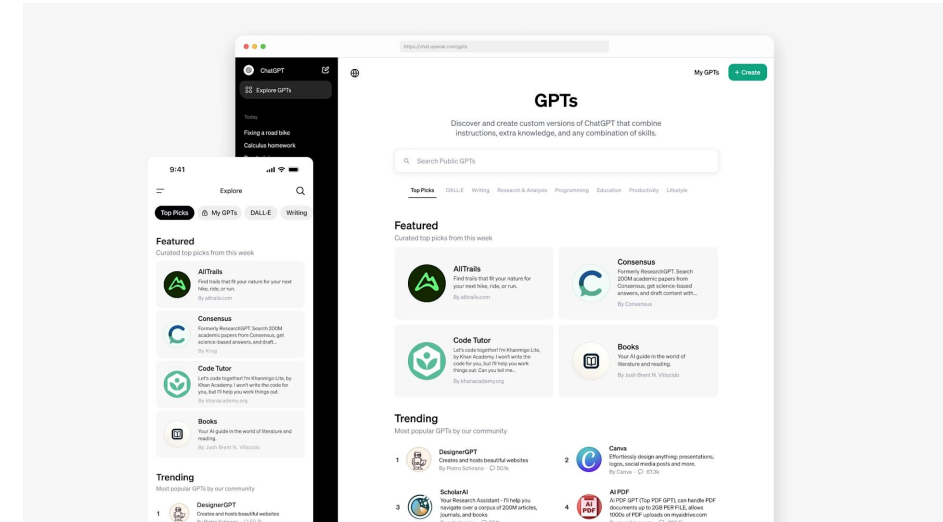
Tang et al.: *Ads that Talk Back...*

Areas to strengthen the paper:

- Experiment sample is relatively small (although not relative to many studies in HCI and related fields)
 - Makes some of the null findings less compelling – is this due to lack of statistical power?
- Similar to Choi et al., may be possible to get more statistical power and easier theory building with model specifications that account for factorial design
- To what extent are these results specific to LLM chatbots? If we ran a similar experiment around sponsored search, would we expect to see different results?

Wrapping up

- Collectively, these papers speak to an interesting direction for the literatures on reputation system design and platform design – how do LLMs and LLM-based products transform and control the diffusion of information?
- **Key question: How should AI intermediaries be designed when they summarize, rank, disclose, and persuade on users' behalf?**
 - **Important insight:** We should study generative AI less as a monolithic technology, and more as a **designed market institution**
- Quan and Shen and Choi et al. speak to AI/LLM as platform component, but Tang et al. points to an interesting question that is perhaps more important topic moving forward: **LLMs as online marketplaces/platforms**



Thank you!

Email: david.holtz@columbia.edu

Disclosures: D.H. is currently a paid part-time visiting researcher at OpenAI, and has made angel investments in several AI-related startups: Trapeze Health, Delyt, Approval AI, Tandem, Rewind AI (now Limitless), and Present. Some of D.H.'s research is or has been funded by Microsoft, and D.H. has an ongoing unpaid research project using data provided by Optimizely. D.H. has been an organizer of the Conference on Digital Experimentation (CODE@MIT) since 2022, which has been sponsored by Meta, Amazon, Netflix, Booking.com, Eppo, DataDog, and Statsig.

