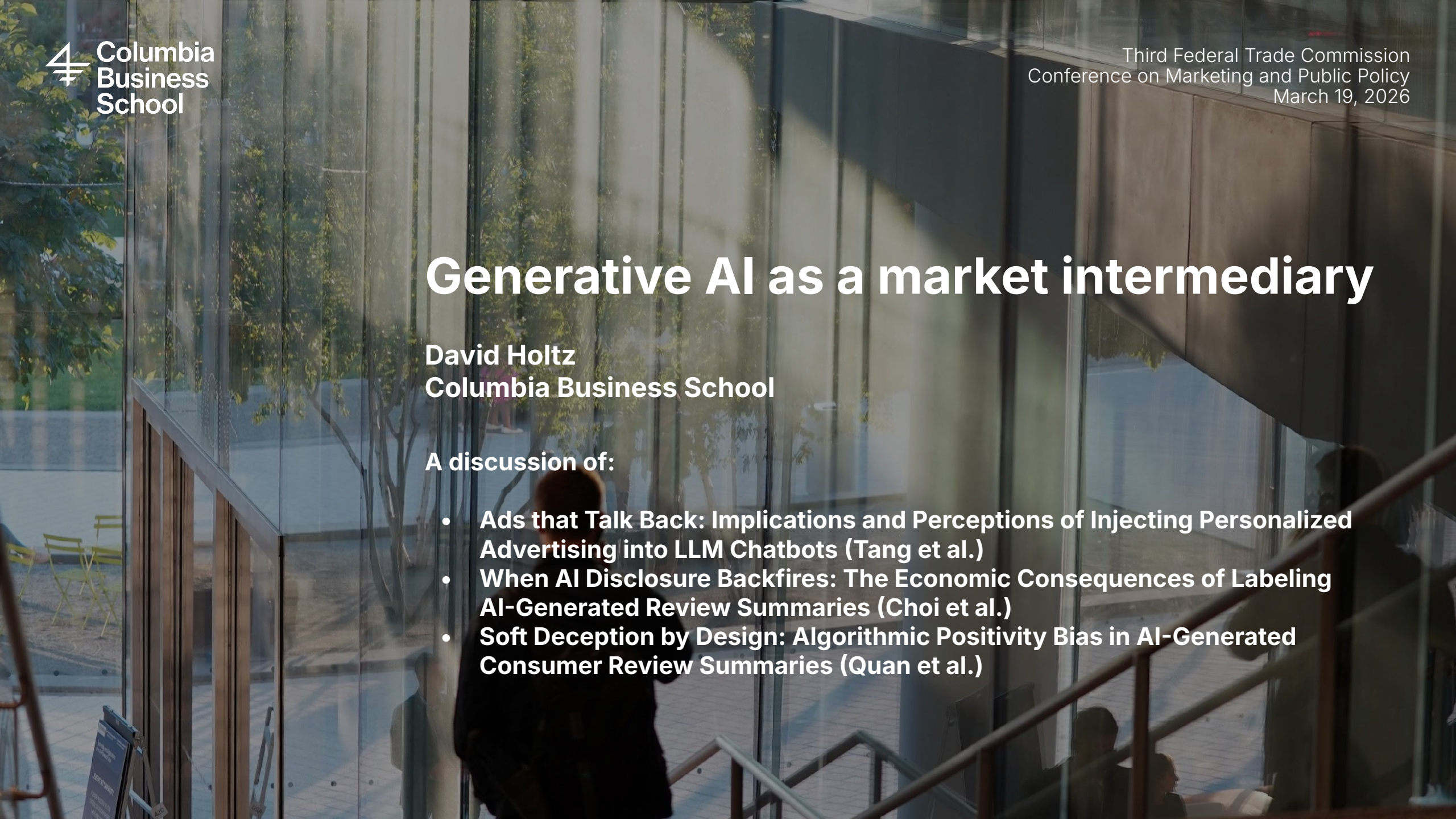


Columbia Business School

Third Federal Trade Commission
Conference on Marketing and Public Policy
March 19, 2026

Generative AI as a market intermediary

David Holtz
Columbia Business School

A discussion of:

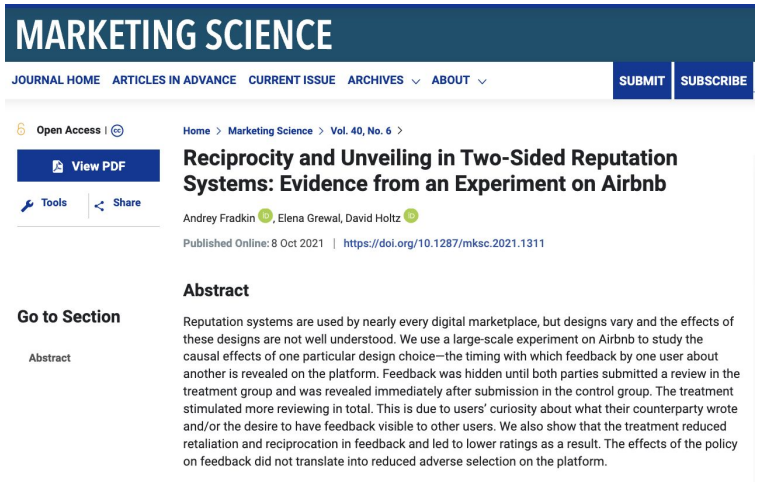
- **Ads that Talk Back: Implications and Perceptions of Injecting Personalized Advertising into LLM Chatbots (Tang et al.)**
- **When AI Disclosure Backfires: The Economic Consequences of Labeling AI-Generated Review Summaries (Choi et al.)**
- **Soft Deception by Design: Algorithmic Positivity Bias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

Columbia Business School

Connecting to the literature on reputation system design



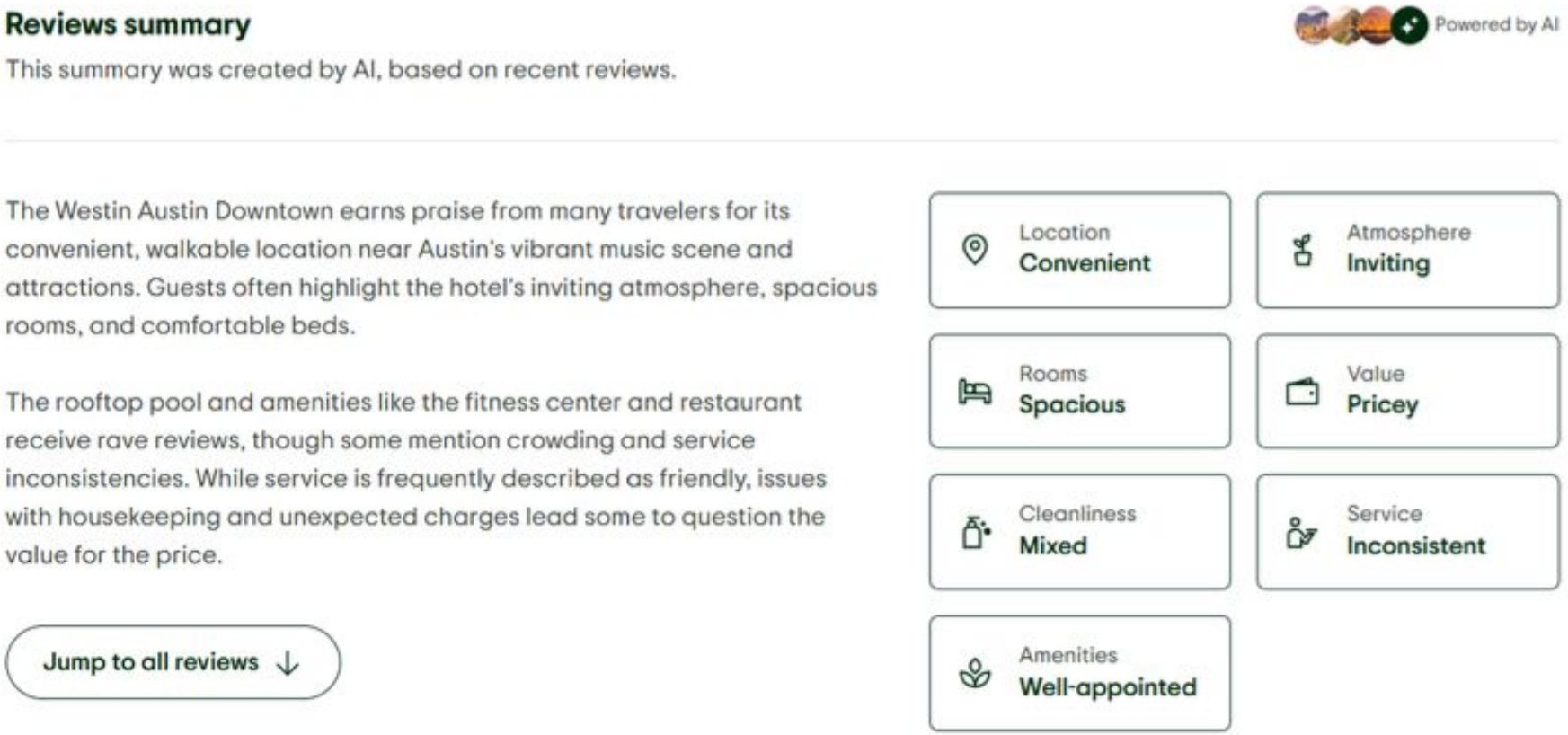
Columbia Business School

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



Columbia Business School

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?

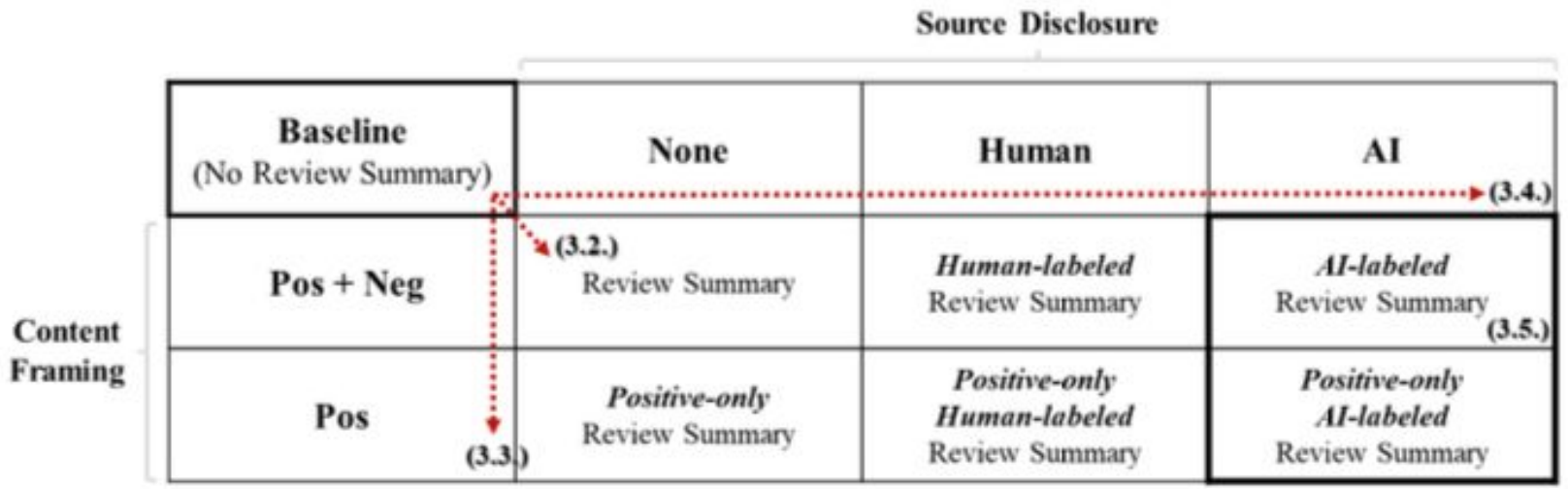
Quan and Shen: *Soft Deception…*

- 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



Columbia Business School

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

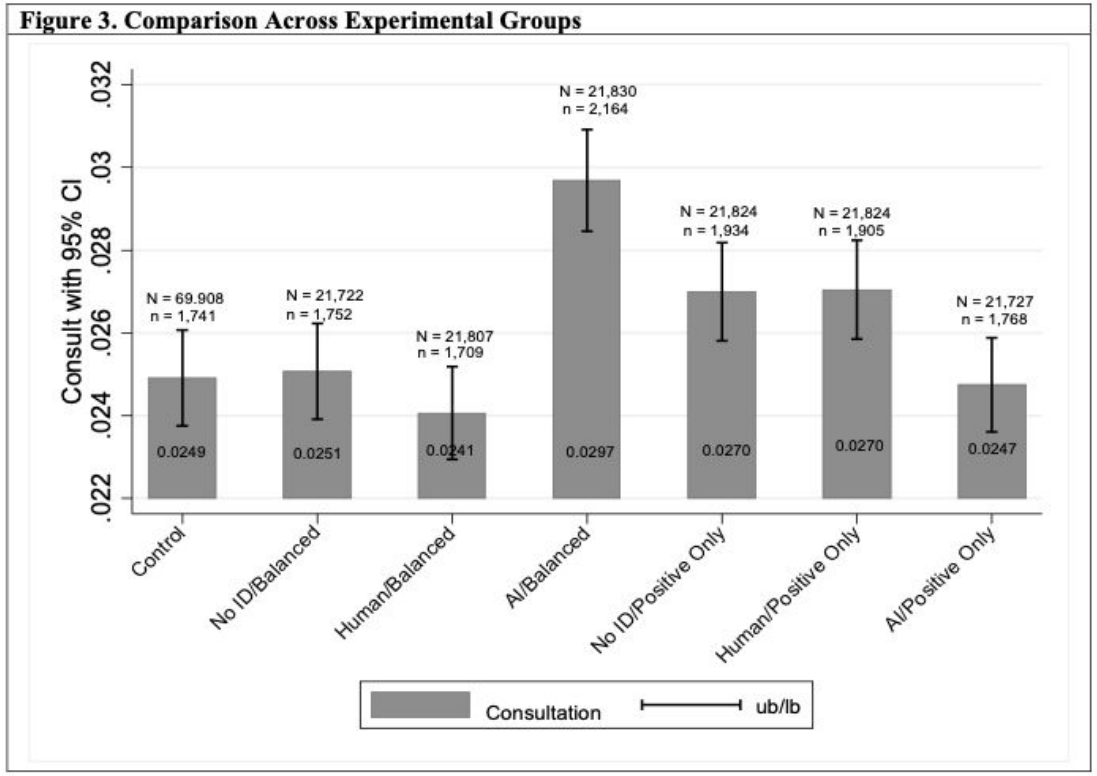
Answer: It's complicated. The authors assert an "AI disclosure paradox:" *disclosing AI authorship drives engagement in early decision stages, but suppressed conversion at later stages (relative to identical balanced summaries that are not labeled)*

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

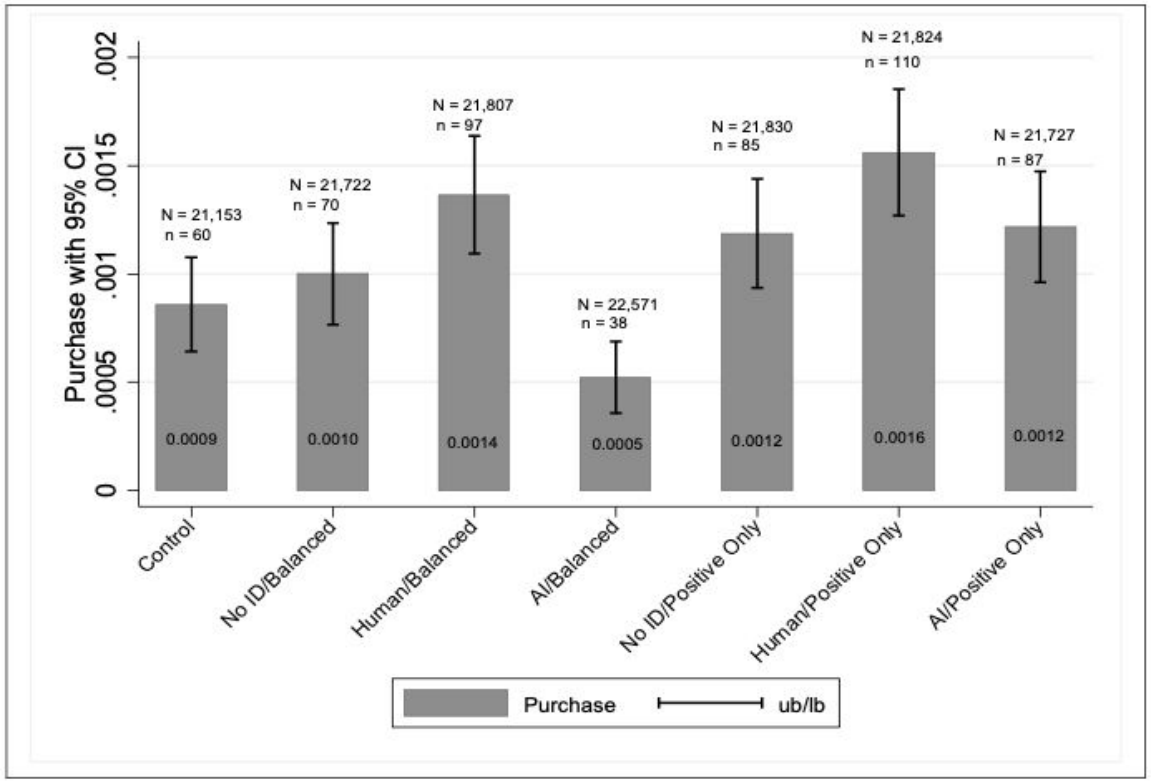
Answer: It's complicated. The authors assert an "AI disclosure paradox:" *disclosing AI authorship drives engagement in early decision stages, but suppressed conversion at later stages* ***(relative to identical balanced summaries that are not labeled)***

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

Model free evidence (consultations)



Model free evidence (purchases)



Columbia Business School

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…*

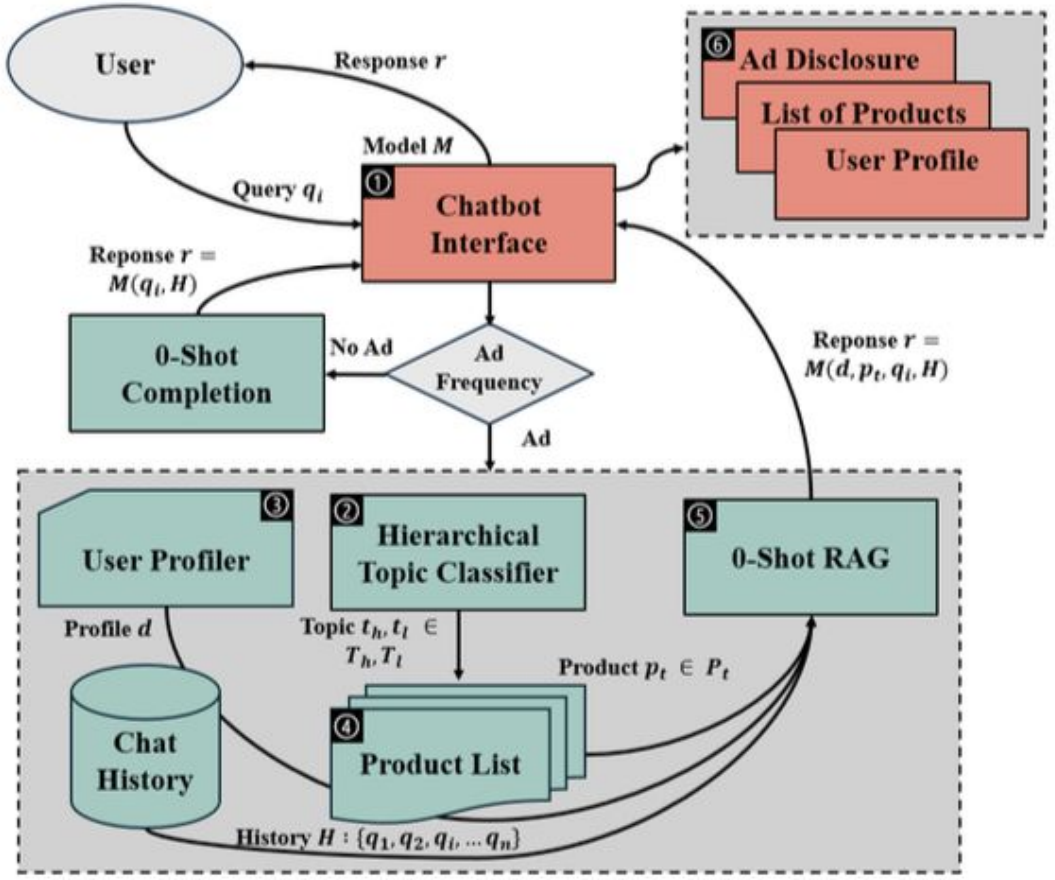
$$Y_{ij} = \alpha + \tau_S S_i + \tau_P P_i + \tau_H H_i + \tau_A A_i + \tau_{PH}(P_i \times H_i) + \tau_{PA}(P_i \times A_i) + X_{ij}'\gamma + \varepsilon_{ij}$$

- 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?

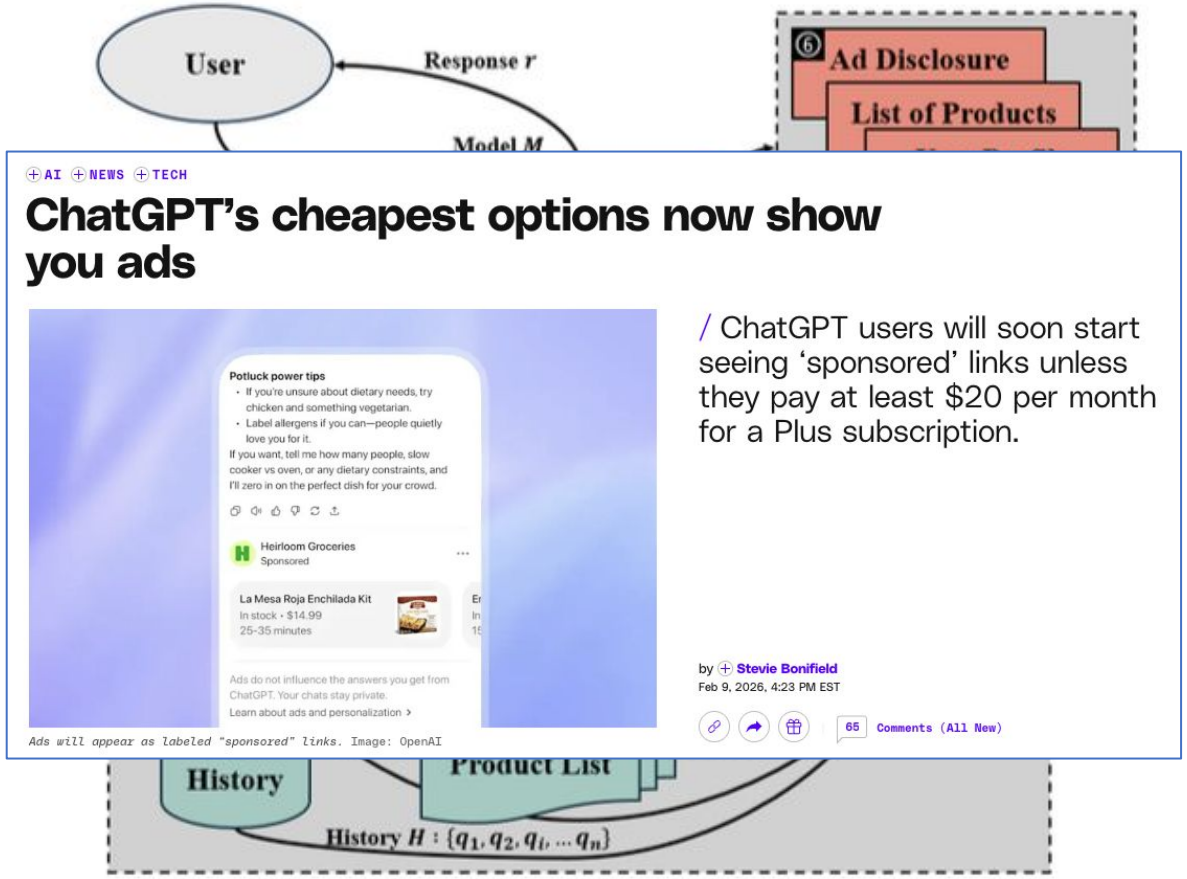


Columbia Business School

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?



Columbia Business School

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

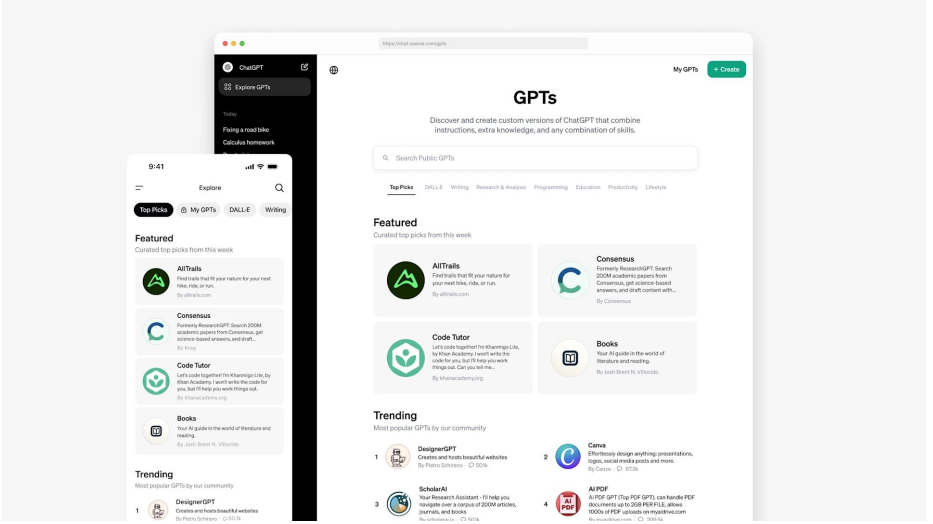
Tang et al.: *Ads that Talk Back…*

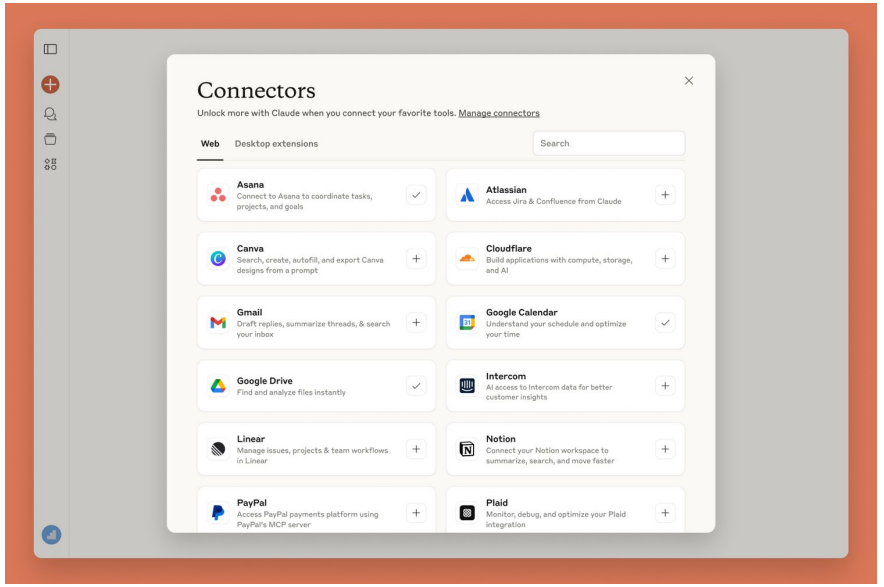
Areas to strengthen the paper:

- 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**





Columbia Business School

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, Delty, Approval AI, Tandem, Rewind AI (now Limitless), and Prezent. 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.