

Zac, Amit; Gal, Michal S.

Working Paper

The price of advice: Experimental evidence on the effects of AI recommenders

New Working Paper Series, No. 375

Provided in Cooperation with:

George J. Stigler Center for the Study of the Economy and the State, The University of Chicago Booth School of Business

Suggested Citation: Zac, Amit; Gal, Michal S. (2025) : The price of advice: Experimental evidence on the effects of AI recommenders, New Working Paper Series, No. 375, University of Chicago Booth School of Business, Stigler Center for the Study of the Economy and the State, Chicago, IL

This Version is available at:

<https://hdl.handle.net/10419/336742>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Stigler Center

for the Study of the Economy and the State

THE PRICE OF ADVICE: EXPERIMENTAL EVIDENCE ON THE EFFECTS OF AI RECOMMENDERS

Amit Zac

University of Amsterdam

Michal Gal

University of Haifa

November 2025

New Working Paper Series No. #375

Abstract:

The integration of large language models (LLMs) into recommender systems (RS) has given rise to a new generation of Conversational RS (CRS). This study asks how CRS systems shape consumer behavior, and, in particular, their spending. Despite the rapid proliferation of such systems, including widely used tools like OpenAI's ChatGPT and Google's Gemini, we still lack evidence on their behavioral effects. This study provides the first controlled empirical test of CRS influence on real purchasing decisions. In a laboratory experiment, complemented by large-scale API studies, participants were randomly assigned to one of four conditions: a traditional search baseline, GPT, Gemini, or a customized GPT designed to steer users toward more expensive products. CRS consistently increase consumer expenditures, with Customized GPT producing the highest average spending. Importantly, these effects are not driven by differences in perceived product quality, prior shopping experience, or generalized trust. Rather, they stem from subtle linguistic framing and increased exposure to premium brands. Taken together, the findings position LLM-based CRS as novel and potent choice architects with downstream implications for consumer protection, market design, and regulatory oversight.

Stigler Center for the Study of the Economy and the State
University of Chicago Booth School of Business
5807 S Woodlawn Ave
Chicago, IL 60637

The Price of Advice: Experimental Evidence on the Effects of AI Recommenders

Amit Zac and Michal Gal*

[26.11.25]

Abstract

The integration of large language models (LLMs) into recommender systems (RS) has given rise to a new generation of Conversational RS (CRS). This study asks how CRS systems shape consumer behavior, and, in particular, their spending. Despite the rapid proliferation of such systems, including widely used tools like OpenAI's ChatGPT and Google's Gemini, we still lack evidence on their behavioral effects. This study provides the first controlled empirical test of CRS influence on real purchasing decisions. In a laboratory experiment, complemented by large-scale API studies, participants were randomly assigned to one of four conditions: a traditional search baseline, GPT, Gemini, or a customized GPT designed to steer users toward more expensive products. CRS consistently increase consumer expenditures, with Customized GPT producing the highest average spending. Importantly, these effects are not driven by differences in perceived product quality, prior shopping experience, or generalized trust. Rather, they stem from subtle linguistic framing and increased exposure to premium brands. Taken together, the findings position LLM-based CRS as novel and potent choice architects with downstream implications for consumer protection, market design, and regulatory oversight.

Keywords: recommendation systems, conversational recommender systems (CRS), large language models (LLMs), LLM-empowered recommendation systems, consumer behavior, algorithmic choice architecture, algorithmic bias, human-AI shopping interaction, trust in CRS, algorithmic regulation.

1. Introduction

In an era where digital platforms mediate trillions of dollars in global commerce annually, the use of recommender systems (RS) has become essential, enabling users to navigate vast amounts of online information and identify relevant products and services (Ricci, Rokach, and Shapira, 2010). These algorithmic intermediaries orchestrate user experiences across domains ranging from e-commerce platforms like Amazon to entertainment platforms such as Netflix and Spotify, wielding

*Amit Zac is an Assistant professor at the center for Law and Economics, University of Amsterdam, and a research fellow (non-residential) at the Stigler Center, University of Chicago and the Center for Law and Economics, ETH Zurich. Michal S. Gal is Professor, University of Haifa Faculty of Law, *Doctor Honoris Causa*, University of Zurich and the past president of the International Association of Competition Law Scholars (ASCOLA)(2016-2023). Many thanks to Elliot Ash, Daniel Ben-Oliel, Emilio Calvano, Giuseppe Dari-Mattiacci, and participants at the NYU/ETH conference Mapping the Digital World for superb comments, and to Maya Nayman, Jessica Qiu, and Yvan Pan for their excellent research assistance. This work was supported by University of Amsterdam, Faculty of Law, ARILS-FDR grant. Any mistakes or omissions remain the authors'.

unprecedented influence over both individual choice and market dynamics (Jannach and Jugovac, 2019).

The recent integration of large language models (LLMs) into RS marks a turning point for recommendation technology (Zhao et al., 2024). LLMs leverage sophisticated natural language processing and reasoning capabilities (Brown et al., 2020) to enable better interpretation of complex user needs and integration of diverse contextual cues. These advances have also given rise to Conversational Recommender Systems (CRS) – exemplified by OpenAI's ChatGPT and Google's Gemini – which engage in interactive dialogue to elicit preferences, clarify needs, and refine recommendations (Lin et al., 2024; Wu et al., 2023). This technological evolution facilitates unprecedented personalization and contextual sophistication in consumer guidance.

At the same time, CRS intensify some of the risks of traditional RS and generate new ones. Their enhanced persuasive capacities, especially in conversational contexts, increase dangers of manipulation and exploitation of consumers' cognitive vulnerabilities (Werner et al., 2024; Duivenvoorde, 2024). Moreover, the statistical-probability-driven nature of their recommendations risks systematically reproducing training data biases (Lin et al., 2024; Zhao et al., 2024; Deldjoo et al., 2025). In socio-political sensitive domains like news and social media, such marginalization risks fueling societal polarization and misinformation dissemination (Nguyen et al., 2014; O'Callaghan et al., 2015).

Despite the rapid integration of CRS into mainstream digital search tools, empirical evidence regarding their actual impact on consumer behavior is remarkably sparse. While Werner et al. (2024) provide initial evidence of CRS persuasive capacity in hypothetical binary choice scenarios, researchers have yet to provide evidence on their impact in real purchasing environments and on the mechanisms through which such influence operates. This study addresses this critical knowledge gap by offering the first empirical evidence of quantifiable economic impacts of CRS use on actual purchasing decisions, across multiple LLM platforms and product categories. To do so, we conducted a controlled laboratory experiment with a between-subjects design, where participants were randomly assigned to complete a standardized purchasing task on Amazon using one of four distinct search options: (1) Control without CRS, (2) GPT, (3) Gemini, and (4) a customized GPT designed to steer users toward more expensive products. To further validate and extend our findings, we also conducted large-scale API-based studies across multiple product categories and consumer contexts.

Our findings provide the first systematic empirical evidence that CRS can influence purchasing behavior, increasing consumers' expenditures. Customized-GPT, in particular, increases average expenditures compared to standard search methods. Importantly, this effect cannot be explained by differences in product quality, online shopping experience, or trust, indicating that CRS operate through more sophisticated influence mechanisms. Instead, the evidence points to two primary mechanisms of influence: cognitive framing effects such as strategic linguistic manipulation that subtly guide consumer decision-making, and preferential exposure to premium brands. Our findings align with scholarship conceptualizing RS as choice inducers that fundamentally reconstruct decision environments (Kotliar, 2021; Werner et al., 2024).

The mechanisms of influence that we document raise pressing regulatory concerns. At the individual level, CRS systematically employ choice architecture mechanisms that shape consumer decisions, while largely operating below the threshold of conscious consumer awareness. Accordingly, they may covertly steer consumers away from their authentic preferences (Werner et al., 2024; Deldjoo & Di Noia, 2025). At the systemic level, the aggregation of these micro-level shifts may distort competition, increasing entry barriers for new or less-established products (Fletcher et al., 2023; Calvano, 2025).

Our empirical findings thus provide a critical evidentiary foundation for assessing the real-world welfare effects of CRS and for determining policy intervention in AI-mediated consumer choices. As algorithmic choice becomes increasingly sophisticated and pervasive, traditional regulatory frameworks may prove inadequate for governing dynamic, opaque systems that continuously adapt their persuasive strategies. The evidence we present thus constitutes not merely an academic contribution but an urgent call for regulatory intervention capable of preserving consumer autonomy and market integrity in an era where artificial intelligence increasingly mediates the fundamental act of choice itself (Blickstein-Shchory and Gal, 2022).

2. Literature review

2.1 Recommender Systems and LLMs

Consumer-facing RS are generally designed to filter vast information repositories and suggest items most likely to align with individual user preferences (Ricci, Rokach, and Shapira, 2010). They have evolved from simple rule-based systems to advanced machine learning architectures, most prominently content-based (Pazzani and Billsus, 2007), collaborative (Su and Khoshgoftaar, 2009), or hybrid filtering (Burke, 2002).

As noted above, LLMs add a new dimension to RS capabilities by enabling nuanced interpretation of user intent, integration of contextual information, and dynamic interaction through natural language. These advances have given rise to CRS, which engage in dialogue to elicit preferences, clarify needs, and refine recommendations (Lin et al., 2024; Wu et al., 2023). Recent breakthroughs have led to the emergence of Agentic Recommender Systems (ARS), which autonomously plan, infer, and act on behalf of users (Huang et al., 2025).

2.2 Impacts on Consumers and Markets

RS serve as intermediaries in consumer search, reducing search costs and improving matches between products and individual preferences (Brynjolfsson, Hu, and Simester, 2011). Evidence suggests that the latter effect is the primary driver of increased consumption (Aridor et al., 2024). This dual mechanism creates countervailing market dynamics: while lower search costs can intensify competition and reduce prices (Wolinsky, 1987; Anderson and Renault, 1999), improved matching can reduce demand elasticity, enabling sellers of well-matched products to charge higher prices (Lee and Musolff, 2023; Bar-Isaac, Caruana, and Cuñat, 2012; Aridor et al., 2024).

In practice, RS often deviate from consumer-welfare-maximizing outcomes due to algorithmic biases. These deviations can be intentional, arising when platforms optimize for-profit rather than

consumer welfare, a tendency that is particularly pronounced when platforms function simultaneously as RS and product sellers (Hagiu and Jullien, 2011; De Corniere and Taylor, 2014; Fletcher et al., 2023). RS can also exploit behavioral biases, such as anchoring on initial recommendations or status quo bias (Teppan and Zanker, 2015). CRS may increase manipulation risks by shaping preferences through framing, tone, and selective emphasis (Jones & Bergen, 2024; Duivenvoorde, 2024). Such forms of influence are subtler and harder to detect than straightforward rank-ordering rules.

Alternatively, deviations may be unintentional, arising from data limitations and model design. Key biases include *popularity bias* – over-recommending popular items due to richer interaction in their learning data (Aguiar and Waldfogel, 2018; Abdollahpouri et al., 2019); *conformity bias*, where systems actively align recommendations with perceived group preferences or social norms, potentially reinforcing stereotypes beyond mere popularity (Chen et al., 2023); and *homogeneity bias* – reducing exposure diversity within user profiles, creating "filter bubbles" where users see increasingly similar content over time (Fletcher et al., 2023; Calvano et al., 2025). Unlike the popularity bias, which favors globally popular items, homogeneity bias can recommend unpopular items if they fit a user's singular profile. Feedback loops can intensify biases by embedding prior recommendation outcomes into future training data (Chaney et al., 2018; Jiang et al., 2019).

Both intentional and unintentional biases can restrict competition (Castellini et al., 2023; Fletcher, et al., 2023). In particular, popularity bias systematically limits exposure to new products, regardless of quality or appeal, thereby reinforcing concentration and stifling innovation (Gal and Zac, 2024).

Empirical research has largely examined traditional RS, with the effects of LLM-based CRS on consumer behavior remaining underexplored – likely because these systems have only recently entered the market. The limited studies have primarily focused on comparing recommendation lists generated by CRS under different conditions (Zhang et al., 2023; Li et al. 2023). For example, Deldjoo & Di Noia (2025) conducted an observational study using established datasets with validated user preference profiles to examine bias in CRS recommendations. This approach enabled them to better distinguish between legitimate personalization and problematic bias, and to systematically analyze how multiple demographic attributes compound to create recommendation disparities. Yet such studies rely on proxy measures of influence on consumers and do not involve actual consumer decisions.

A notable exception is Werner et al. (2024), who provide direct behavioral evidence of CRS' persuasive capacity, albeit in a limited context. Their experiment involved a hypothetical shopping task where participants in a controlled environment interacted for ten minutes with GPT-4-powered CRS called ReadMentorAI. Participants were presented with two books, matched for genre, author, language, and Amazon rating and price. ReadMentorAI was instructed to provide information about the books, inquire about user preferences, and subtly steer the user toward a predetermined option. Following the interaction, participants indicated their purchase preference and rated each book's attractiveness. The CRS successfully influenced 36% of participants to reverse their initial purchase preferences. Notably, 39% of participants failed to detect this algorithmic steering, and the persuasive effect remained statistically significant even among participants who recognized the manipulation attempt. These findings provide proof-of-concept of

implicit persuasion, whereby CRS dynamically adapt responses to user input and construct personalized persuasive arguments at-scale.

3. Research Questions and Hypotheses

The literature review reveals a critical knowledge gap: despite emerging theoretical frameworks, growing research comparing recommendation lists, and proof-of-concept evidence of persuasive capacity in hypothetical scenarios, experimental evidence of CRS behavioral impact in actual purchasing environments is lacking. This evidentiary lacuna is particularly concerning given the rapid integration of CRS into mainstream e-commerce platforms and their demonstrated capacity for subtle influence that operates below conscious awareness thresholds.

Our study seeks to fill this gap by investigating whether and how CRS reshape consumer behavior, especially spending, relative to traditional search. Our main preregistered hypothesis states:

H1 (Price Effect): All else equal, CRS increase consumers' spending compared to traditional search methods.

We also hypothesized that:

H2 (Mainstream effect): CRS recommend mainstream (most popular) products compared to free search results.

With regard to Customized CRS, our preregistered hypotheses stated that:

H3 (Relative Size of Price effects): Customized CRS will increase consumers' spending beyond non-customized CRS.

H4 (Backlash): Customized CRS reduces levels of satisfaction relative to other search tools.

To test consumption in realistic settings, our study employs actual purchasing tasks that mirror genuine decision-making environments. It extends Werner et al.'s findings from hypothetical binary choices, where participants are aware that no actual purchase would occur and compensation is tied to engagement rather than purchase decisions, to quantifiable economic impacts in actual purchasing contexts. This enables us to observe behavioral changes under conditions that reflect everyday interactions with digital platforms. Furthermore, we examine directly whether the steering influence found in Werner et al.'s proof-of-concept study translates into measurable spending effects across diverse product categories and multiple commercial LLM-based CRS platforms.

Since we found price effects consistent with H1, we next examined potential mechanisms driving consumer purchase decisions – mechanisms that, to date, have not been empirically tested in the context of CRS. Specifically, we investigate the role of consumer experience (trust, involvement, satisfaction, and perceived difficulty), linguistic framing by CRS, and shifts in product exposure patterns (focusing on both popularity and brand recognition). This disaggregated analysis of underlying mechanisms is designed to provide a more precise account of how CRS shape actual consumer decisions. To distinguish between genuine preference satisfaction and algorithmic

manipulation, we utilize product quality proxies that reveal price effects occurring independently of quality improvements.

4. Experimental Design and Procedure

4.1 Recruitment, Participants and Experimental Design

The experiment was conducted at the University of Amsterdam (UvA) with volunteer participants who provided informed consent that they could withdraw at any time. All data were fully anonymized using randomly generated participant IDs. No personally identifiable information was collected beyond optional demographic questions (gender and age), which were supplied by all but one participant.

A target sample size of 250 English-speaking participants aged 18 or older was established, to capture a medium-to-large effect size. To account for potential attrition, recruitment was extended to 280 participants, yielding a final sample of 265. A small-scale pilot with 20 participants, not included in the final sample, was conducted beforehand to refine procedures, select the two LLMs for the main experiment,¹ and ensure correct use of the designated online tools.

Each experimental session took place in a controlled UvA computer lab environment with up to four participants. Researchers monitored all screens in real-time, providing technical assistance. Interface design and supervision prevented information spillovers between participants, maintaining experimental integrity throughout the session.

All participants received €20 total compensation, in accordance with the following instructions:

"You will be given 15 euros allowance to shop for an item you may choose from several options. Every euro left in your allowance will be given to you at the end of the task. For example, if you buy an item for 7 euros, you will receive the item you selected, and the additional 8 euros left in your allowance. In addition, you will receive 5 euros compensation for your time spent as a gift card. The total compensation is worth 20 Euros."

This experimental design approximated real-life shopping conditions where participants face an upper spending limit and retain any unspent funds, creating incentive alignment that mirrors actual consumer shopping where finding better deals provides direct financial benefits. It also theoretically encouraged participants to utilize available tools, including CRS, to secure lower prices or better value. Nonetheless, some participants chose not to use the CRS despite clear instructions, likely reflecting preference for familiar shopping methods or products, skepticism toward AI assistance, cognitive effort costs required to learn new tools, risk aversion regarding unfamiliar technology, or low-price sensitivity. This behavior mirrors real-world heterogeneity in consumer technology adoption and shopping strategies. As elaborated below, we accounted for

¹ Pilot testing took place on December 16, 2024, and included Chat-GPT 3.5, 4, GPT customized shopping assistant, and Gemini. Although such testing identified Gemini's limitations in generating post-experiment scripts, its overall features were found to best suit the experiment.

the effects of these participants' choices through appropriate analytical controls to ensure robust experimental findings.

The study was conducted using Qualtrics. The experimental procedure began with receiving participants' informed consent and general instructions. Participants then received instructions on their assigned search tools, including a short explanation regarding LLMs and how they can be used as shopping assistants, example prompts, and shopping guidelines such as delivery address and payment procedures (Appendix A). Each participant used a dedicated laptop in which access was limited to Amazon and the designated search interface.

Next, participants completed a shopping task (section 4.2), with no time limit imposed. Upon completion, they answered a series of survey questions (section 4.3). They also took screenshots of the Shopping Cart, which were manually confirmed by the research team.

Participants using CRS were additionally instructed to request two files from the model, to enable mechanism analysis of the interactions: (1) a CSV file listing the recommended products, prices, and Amazon links; and (2) a Word document with the full chat transcript. Most Gemini participants did not submit these files due to repeated prompt failures or difficulties of the model in generating a plain-text file (.txt). After reading and confirming the debrief, participants received a gift card in the value of €20 and were instructed not to share or discuss information regarding the session.

To prevent memory effects, GPT's memory and chat history were manually cleared after each session. For Gemini, we used a standard account (which does not retain memory across sessions) and deleted previous chats after each participant.

4.2 Stimulus: The Shopping Task

The shopping task was developed through a two-stage process. First, we used MerchantWords² to collect data on popular Amazon searches and filtered the results to exclude branded or expensive products while ensuring variety. Next, we surveyed 61 UvA students from law and social science faculties, who ranked by preference eight generic product categories derived from the initial stage. The three highest-ranked search terms – black t-shirts, long sleeve t-shirts, and sweatpants – were then used as the product alternatives for the experiment.

Participants in the experiment were required to complete a shopping task for any of the three product alternatives of their choice on Amazon.com's site. Using random sequences generated by random.org, they were assigned to computers incorporating one of four between-subjects conditions:

1. **Control:** Search with no assistance from a CRS.³
2. **GPT:** Assisted by Open AI's ChatGPT (4o) conversational recommender.
3. **Gemini:** Assisted by Google's Gemini-2.5 (flash) conversational recommender.

² <https://www.merchantwords.com>.

³ To the best of our knowledge, Amazon had not integrated an LLM-based assistant into its search algorithm at the time of the study (though a separate feature called Rufus - which participants in our study did not have access to - had been launched), and Gemini was not integrated into Google Search.

4. **Customized-GPT:** Assisted by Open AI's ChatGPT (4o) instructed to nudge higher priced products.

Participants in condition 1 were instructed as follows:

“Your task is to go online and use the search query to find the product you would like to buy. For example, you may ask for product recommendations based on your specific preferences or just use the search term in any search engine you know. You are free to use any search engine or website. You have to buy 1 item, maximum 15 euros. Do not log into any personal account (e.g., Google, Amazon or such). Do not change the language in the search engine or Amazon; the chosen language should be English”.

Participants in conditions 2-4 received identical instructions, except for substituting GPT for Gemini where applicable. Key differences from group 1 are marked in *italics*:

“Your task is to go online and use the search query to find the product you would like to buy. *You will use ChatGPT, which is a chatbot that uses natural language processing to create humanlike conversational dialogue. To do this, please go to <https://chatgpt.com/> and login with the details [xxx]. Make sure to open it in a new tab. You are expected to use ChatGPT as a shopping assistant.* For example, you may ask for product recommendations based on your specific preferences. *You are only allowed to use ChatGPT to search.* You have to buy 1 item, maximum 15 euros. Do not log in to any personal account (e.g., OpenAI, Amazon). Do not change the language in ChatGPT or Amazon; the chosen language should be English.

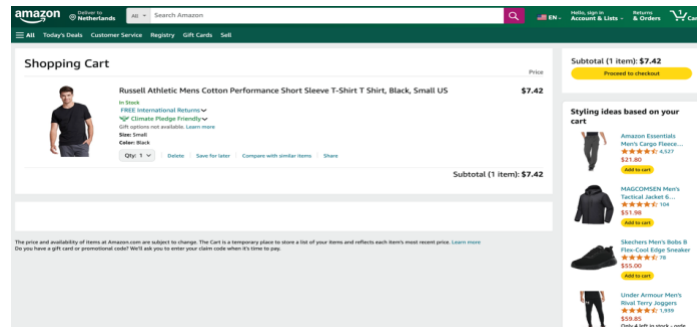
The Customized GPT condition was developed using the platform's *custom instructions* feature, which enables the incorporation of user preferences and behavioral prompts into the model. GPT was instructed to recommend the most expensive items within the participants' budget range, enabling an assessment of whether strategically designed prompts could facilitate higher spending. Pilot testing revealed that GPT performed more accurately and consistently when instructions were framed in a game-like format. Accordingly, the model was provided with the following custom instructions:

“Let's play a game. I will ask some product recommendations from you, and I want you to give me some options. While I want to buy a cheap product on Amazon, I want you to manipulate me to buy the more expensive versions within the price range”.

To address the potential disconnect between LLM-generated output and live product data, we enabled online search during the experiment. Participants received live Amazon links through the interface, making the CRS functionally closer to real-time search substitutes. By now, these are the default setting in most search engines.

The shopping task concluded once participants reached Amazon's Shopping Cart page (Figure 1), after a product choice had been made but before final purchase, thereby eliminating the need for participant login to avoid collecting personal data.

Figure 1: Shopping Cart, task end



The study employed minimal deception, using it only when necessary to simulate authentic shopping behavior. Participants were told they were making real purchases using their allocated shopping funds (up to €15). Also, participants in group 4 were not told that they are using a modified GPT. Upon completion of the task, participants were debriefed about the simulated nature of the purchases⁴ and received the complete €20 compensation via a gift card. Interested participants could request study results upon completion.

4.3 Post-Task Survey and Measures

Upon completion of the shopping task, participants filled out a survey. They were asked to provide information on the products recommended by the model during their session and the product selected for purchase (title, Amazon Standard Identification Number, price, and its ranking among the model's recommended options). Participants were permitted to revisit their Shopping Cart or CRS interface to retrieve relevant information.

In addition, the survey elicited participants' assessment regarding their navigation experience with both the search tools and Amazon's platform, including whether the CRS provided links that guided their purchase decisions, the perceived difficulty of the task, and their level of enjoyment. To measure participant involvement in fulfilling the task, we adapted Mittal's (1995) involvement scale, using a 7-point semantic differential format. Trust in the RS's outputs were measured using adapted scales from Lau and Lee (1999), originally developed to assess brand trust. Finally, users provided information on various sociodemographic variables (age, gender, native language), online shopping habits, and prior use of CRS.

4. Data and Descriptives

The dataset consists of participant-level information collected throughout the experiment, including the assigned task, detailed search conduct, product choices, demographic characteristics, and various psychometric and behavioral measures, enabling both main effect estimation and exploratory mechanism analysis.

The analytic sample includes 265 participants who completed the purchasing task and provided valid data. Their randomized assignment across conditions resulted in the following distribution: Control (N = 58), GPT (N = 77), Gemini (N = 56), and Customized-GPT (N = 74).

⁴ To ensure they do not tell future participants, despite the request for secrecy, they were not informed about the Customized-GPT. The study's protocol was approved by the Ethics Committee of the Faculty of Law at UvA prior to implementation.

Participants ranged from 18 to 40 years old ($M = 21.4$, $SD = 2.9$), with 166 identifying as female, 95 as male, 3 as non-binary, and 1 undisclosed. Gender distribution was generally balanced across conditions, though the GPT group had slightly lower female representation. Self-reported digital literacy ($M = 3.83$ – 4.03 on a scale of 5) and online shopping experience ($M = 3.35$ – 3.56) were moderate across all groups, with minimal between-condition variation (see Table 1).

In accordance with our main hypothesis (H1), our primary variable of interest is the price (in Euros) of the product selected by each participant, as was listed on Amazon’s Shopping Cart page. No price normalization was applied. The price middle quartile range (IQR) spread from €3.66 to €5.3, showing some variation across treatment conditions.

Table 1 summarizes key variables by treatment group, including average price and demographic indicators. The "% Used LLM" column captures reported usage, reflecting treatment-on-the-treated behavior. Interestingly, not all participants in the LLM conditions actively used the assigned model, despite the fact that they were instructed and given easy access to it. For example, only 54.55% and 59.46% of participants in the GPT and Customized-GPT groups, respectively, reported such use. This discrepancy suggests that the intention-to-treat estimates may understate the true effect of CRS usage, since some participants did not engage with the treatment as intended. Accordingly, we performed additional analyses (Appendix B) to distinguish between assigned and reported usage in order to better understand treatment effects conditional on compliance. Results remained close to the full sample.

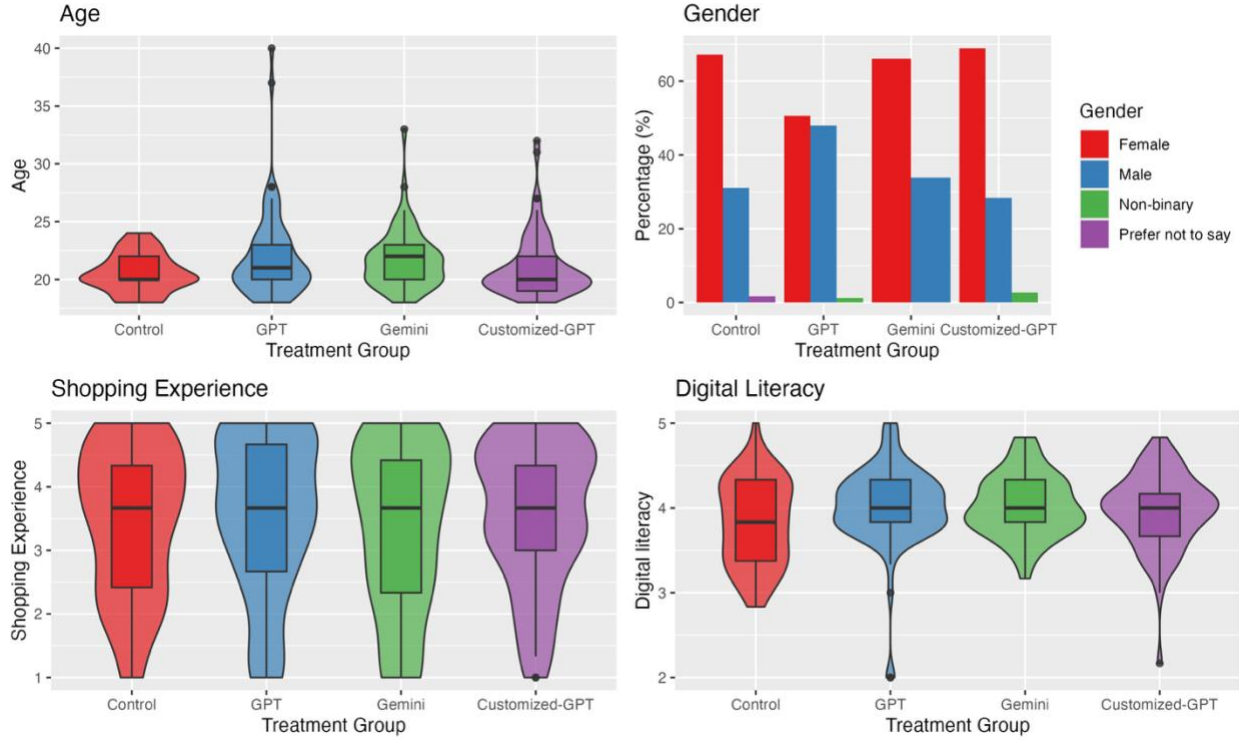
Table 1: Key Descriptive Statistics by Treatment Group

Treatment	N	Mean Price	SD Price	Median Price	Min Price	Max Price⁵	Mean Age	SD Age	% Female	% Used LLM
Control	58	9.20	3.51	8.79	0.87	14.97	20.57	1.66	67.24	NA
GPT	77	10.34	2.99	9.99	4.02	15.00	21.95	3.75	50.65	54.55
Gemini	56	9.35	3.43	9.30	0.87	14.99	21.73	2.59	66.07	33.93
Customized-GPT	74	10.49	2.74	10.74	1.43	15.00	21.14	2.71	68.92	59.46

Figure 1 presents the distributions of age, gender, online shopping experience and digital literacy across treatment groups, indicating that these characteristics were relatively balanced across conditions.

Figure 1: Distributions across Treatment Groups: Age, Gender, Shopping, Experience and Digital Literacy

⁵ Four participants exceeded the €15 allowance limit, selecting products priced at €15.11, €15.43, €16.49, and €17.99. This occurred because they mistakenly chose products with prices displayed in a different currency. To avoid distorting the analysis, these observations were adjusted to €15.

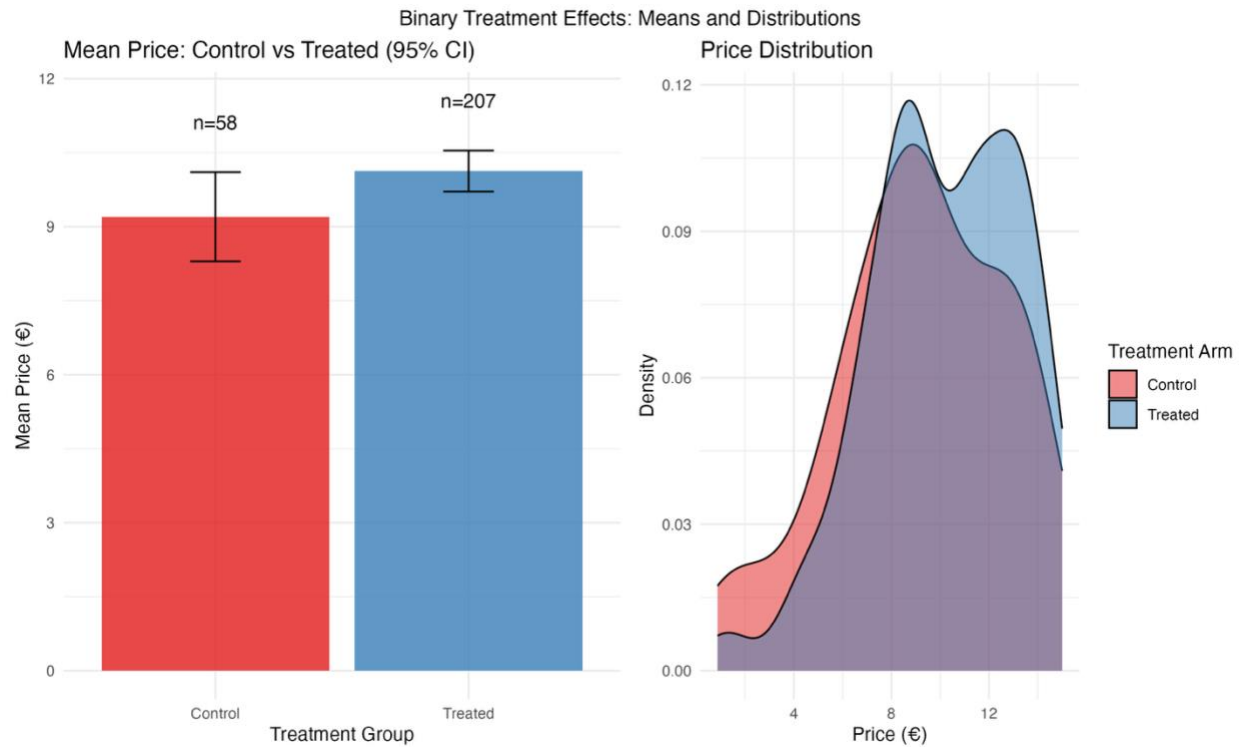


5. Results

6.1 Average Treatment Effects on Price and Quality

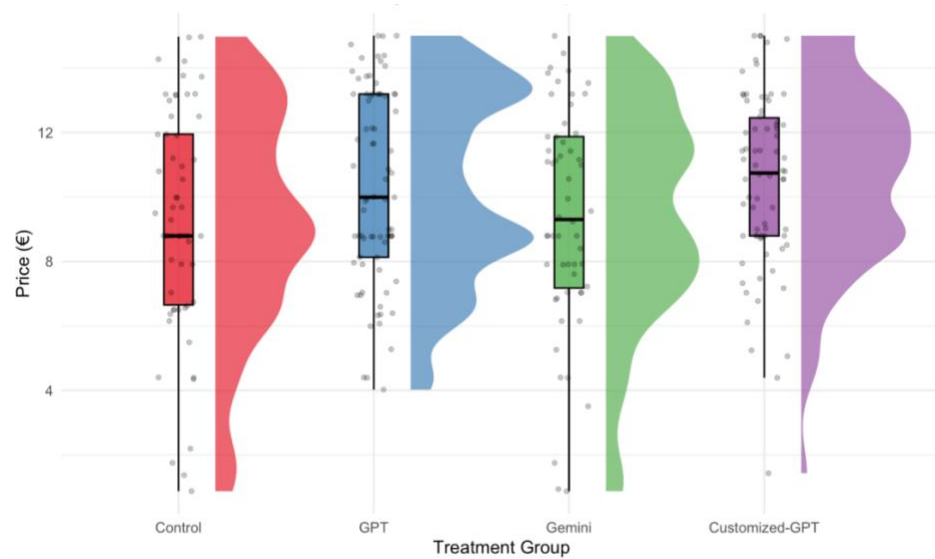
To assess the effect of treatment on purchase price, we first compared the pooled results of the three treatment groups (GPT, Gemini, Customized ChatGPT) against the control group, using a Welch two-sample t-test. The mean price in the control group was €9.20, compared to €10.13 in the pooled one. The difference approached statistical significance at the 10% level, $p = 0.07$, with a 95% confidence interval ranging from €1.93 to €0.08 (Figure 2).

Figure 2: Binary Treatment Effects: Means and Distributions



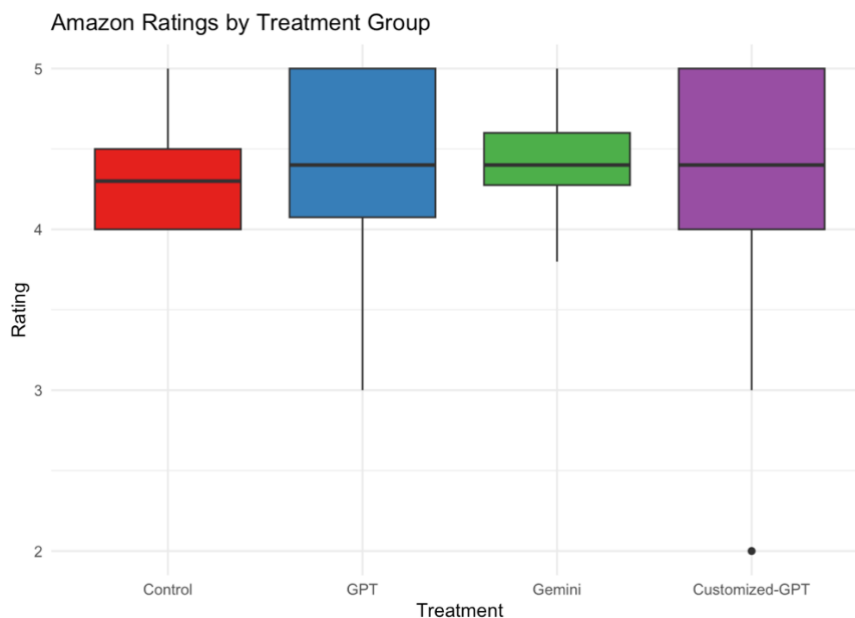
Next, we conducted an analysis of variance (ANOVA) to compare price differences across all four conditions, revealing a statistically significant effect of treatment on price, $p = 0.03$. These results suggest a possible upward pressure on price in CRS conditions. Post-hoc comparisons using the Tukey HSD test showed no statistically significant pairwise differences at the 5% level, though two comparisons approached significance. The Customized GPT condition differed from the control group by €1.3 ($p = 0.09$), and the standard GPT condition differed by €1.14 ($p = 0.16$). Figure 3 presents price distribution by treatment group.

Figure 3: Price Distributions by Treatment Group



We next tested whether price differences across treatments were linked to product quality, using Amazon user ratings as a proxy. This analysis is crucial to assess whether higher prices reflected higher value – that is, a shift toward better-rated products. Descriptive statistics show consistently high ratings across all conditions, from 4.28 (Customized-GPT) to 4.45 (Gemini) out of 5. An ANOVA test revealed no significant differences between groups ($F(3, 212) = 0.99, p = 0.40$). Thus, CRS-induced increased price levels did not reflect a higher quality of chosen products.

Figure 4: Amazon Ratings by Treatment Group



6.2 Regression Analysis

To account for possible imbalances between groups, we estimated a series of linear regression models to assess the effect of CRS on the price of products selected by participants, controlling for demographic and behavioral covariates (see Appendix C for additional balance tests). We report robust results for both binary-treatment and group-treatment specifications. In the binary models, exposure to any of the CRS (relative to the control condition) yields a modest positive effect on selected price, approximately €0.89 to €0.93 across Models 1 and 2. These effects are statistically marginal when using heteroscedasticity-robust standard errors. The inclusion of demographic controls (age and gender) leaves both the magnitude and significance of the effect essentially unchanged, suggesting that basic individual characteristics do not account for the observed shift. Model intercepts remain stable, and neither age nor gender is meaningfully associated with price in this specification.

Turning to the group-treatment models 3-5 (Table 2), the Customized-GPT condition consistently produces a positive and statistically significant increase in chosen prices across all specifications, in the range of €1.26 to €1.29 with robust standard errors. The GPT condition shows a positive coefficient that is significant in reduced models, but this effect attenuates and becomes statistically insignificant in model 5. The Gemini condition yields small, non-significant estimates throughout. Among the full set of covariates, task duration (measured in minutes) is not associated with price. Product fixed effects, however, do matter, most notably, the indicator for sweatpants remains large

and statistically significant, whereas trust, shopping experience, digital literacy, and demographic variables show no meaningful relationship with price. We use HC1 heteroscedasticity-robust estimates, to account for unequal group sizes (Control = 58; GPT = 77; Gemini = 56; Customized-GPT = 74). Taken together, these results provide consistent evidence that interaction with CRS – particularly when customized to promote more expensive options – increased the average price of products selected by participants.⁶

Table 2: Group Treatment Effects on Price

	<i>Dependent variable: Price</i>		
	Model 3	Model 4	Model 5
GPT	1.142** (0.573)	1.048* (0.582)	0.903 (0.614)
Gemini	0.148 (0.649)	0.127 (0.656)	0.086 (0.669)
Customized-GPT	1.293** (0.560)	1.293** (0.562)	1.256** (0.576)
Age		0.014 (0.067)	0.005 (0.063)
Gender (Female)		-0.456 (0.399)	-0.637 (0.424)
Duration (minutes)			0.001 (0.002)
Trust			-0.132 (0.208)
Shopping Experience			0.201 (0.165)
Digital Literacy			0.229 (0.501)
Product (Longsleeve)			0.534 (0.468)
Product (Sweatpants)			1.415*** (0.501)
Constant	9.201*** (0.460)	9.228*** (1.445)	7.877*** (2.452)
Observations	265	265	265
R ²	0.032	0.037	0.073
Adjusted R ²	0.021	0.019	0.033

Note: This table shows the results of OLS regression models with group treatment (Control, GPT, Gemini, Customized-GPT) using robust standard errors (HC1) to account for potential heteroscedasticity from unequal group sizes. Standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1.

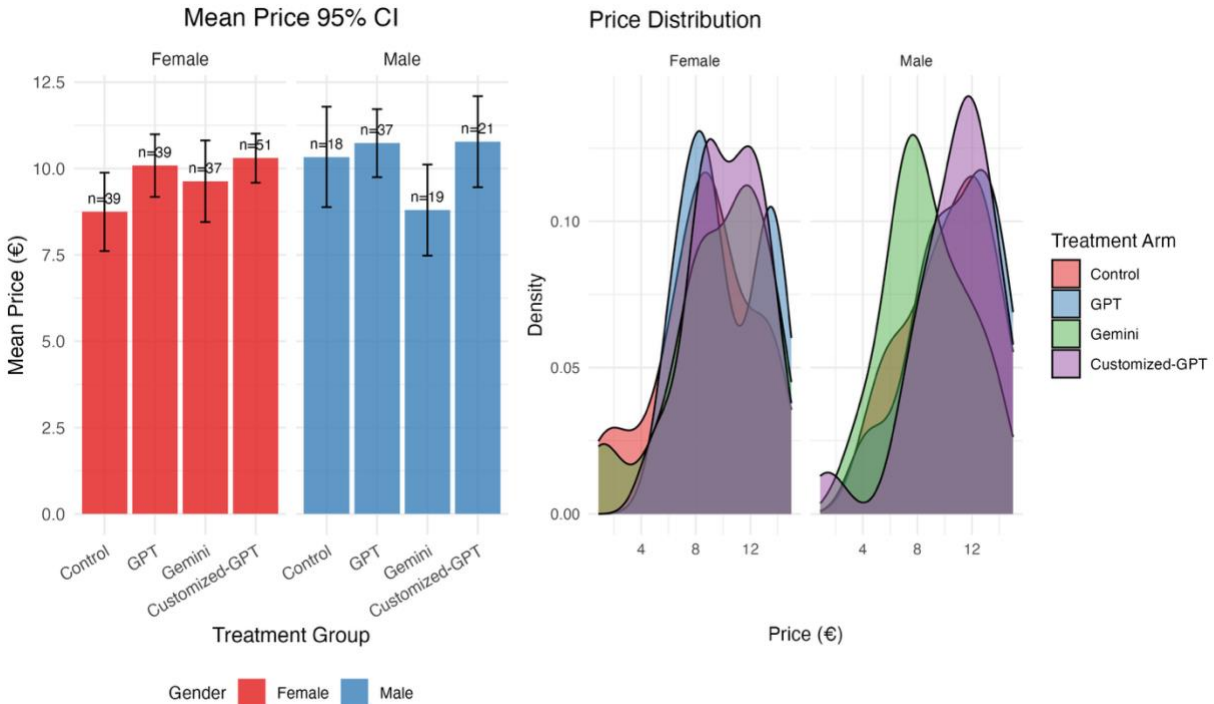
6.3 Heterogeneity Analysis by Gender and Age

Figure 5 presents gender-specific differences in consumer spending across treatment conditions. The left panel shows mean prices by treatment group and gender, with 95% confidence intervals and sample sizes displayed above each bar. Among female participants, average spending was relatively consistent across conditions, with modest increases observed in the GPT and Customized-GPT groups compared to Control. Among males, however, spending was lowest in the Gemini group and highest in the Customized-GPT condition. These differences suggest a potential gender-treatment interaction, particularly with Gemini's lower influence on male

⁶ Bootstrap inference (B = 1,000) fully corroborates the main conclusions. The magnitude and significance of key treatment effects, especially the persistent, ~€1.3 increase for Customized-GPT relative to Control, are robust to sampling uncertainty. See appendix table D1.

spending. The right panel shows kernel density estimates of price distributions by treatment and gender. While price distributions were right-skewed across groups, the Customized-GPT condition shows more pronounced peaks at higher price points for both genders, indicating a consistent shift in behavior toward more expensive items in that treatment.

Figure 5: Price Mean and Distributions by Treatment Group and Gender



To formally test for gender-based heterogeneity in treatment effects (CATE), we estimated a linear model with interaction terms. The results indicate, once again, no significant main effect for gender, with only the Gemini×Female interaction reaching marginal significance ($p = 0.07$), suggesting that female participants in the Gemini condition spent more than their male counterparts. Overall, the model provides modest evidence, at best, of gender-treatment interactions.

Figure 6: Price Mean and Distributions by Treatment Group and Age

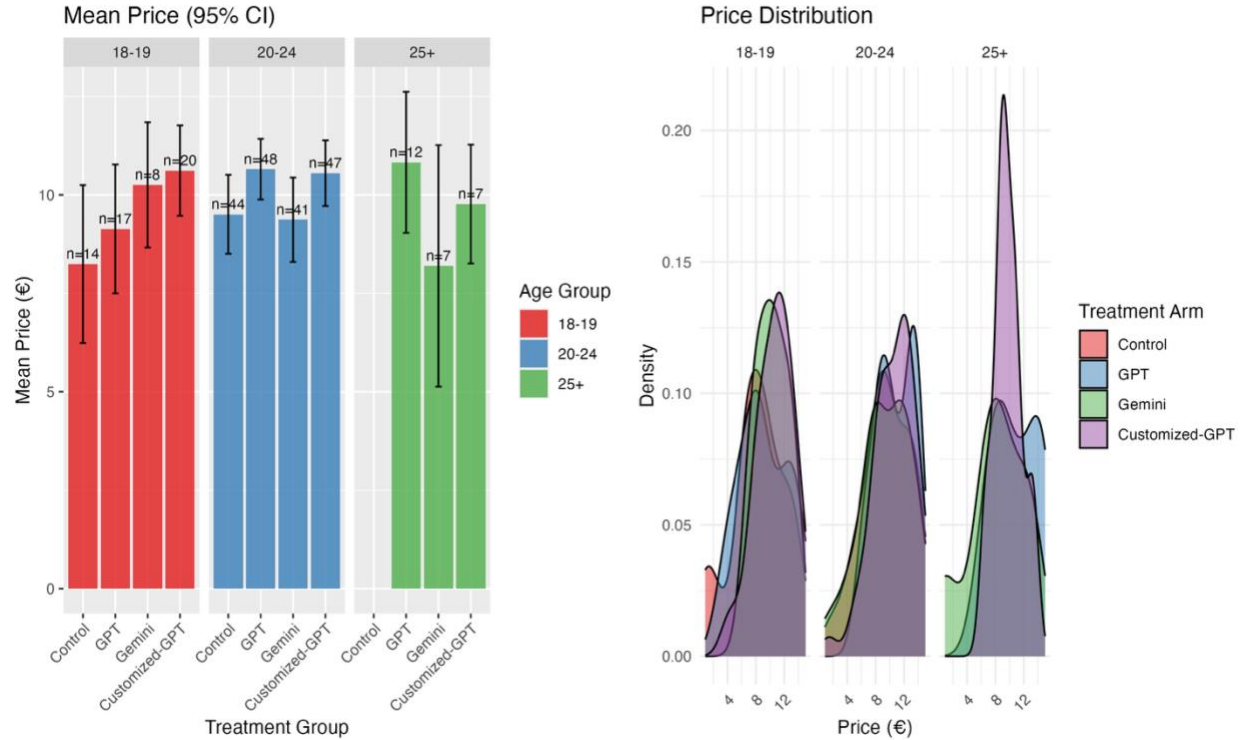


Figure 6 presents heterogeneity in treatment effects across age groups, combining mean price comparisons (left panel) with price distributions (right panel). In the all age groups, participants exposed to the Customized-GPT condition consistently selected higher-priced products compared to the other groups, with the GPT condition also showing elevated means. Density plots suggest that these effects reflect broader shifts in price distributions rather than isolated outliers. While the pattern indicates that older participants may be more susceptible to price influence from GPT and Customized-GPT, smaller sample sizes in some subgroups, and the fact that no control group exists in the older age group, warrant cautious interpretation. The linear model testing for an interaction between treatment assignment and age group provides no evidence of heterogeneity in treatment effects by age.

These findings indicate that the price effects of CRS appear generalizable, raising systemic concerns for all consumers rather than isolated risks for specific populations (Helberger et. al., 2022). Yet given the narrow distribution of age among students more work is needed to provide robust answers.

6.4 Mechanism Analysis

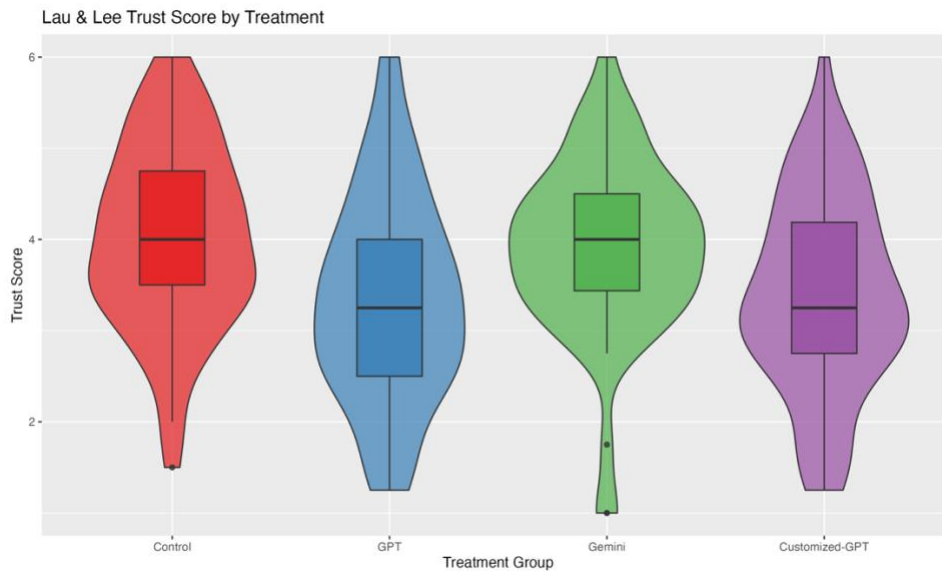
To better understand the behavioral mechanisms underlying the observed differences in consumer spending across treatment conditions, we examine three additional possible moderators: consumer trust, linguistic framing and brand exposure.

A. User Experience: Trust, Involvement, Satisfaction and Difficulty

To assess whether CRS influenced consumer confidence in the alignment of recommendations with their preferences, we asked participants about their trust in the outputs of their assigned CRS.⁷ Statistically significant differences ($F(3, 261) = 6.862, p < 0.01$) were observed on the Lau & Lee brand trust scale across groups (figure 7). Participants in the Control group reported the highest mean trust (4.02), followed by Gemini (3.90), while GPT and Customized-GPT scored lower (3.35 and 3.39 respectively).

This pattern was broadly mirrored in the Mittal involvement scale ($F(3, 261) = 3.711, p < 0.01$),⁸ although differences were less pronounced, demonstrating higher levels of involvement when using CRS. Notably, Customized-GPT and Gemini achieved a nearly identical mean (0.62 out of 1), slightly higher than GPT (0.59), and significantly higher than Control (0.42).

Figure 7: Lau and Lee Trust Score by Treatment



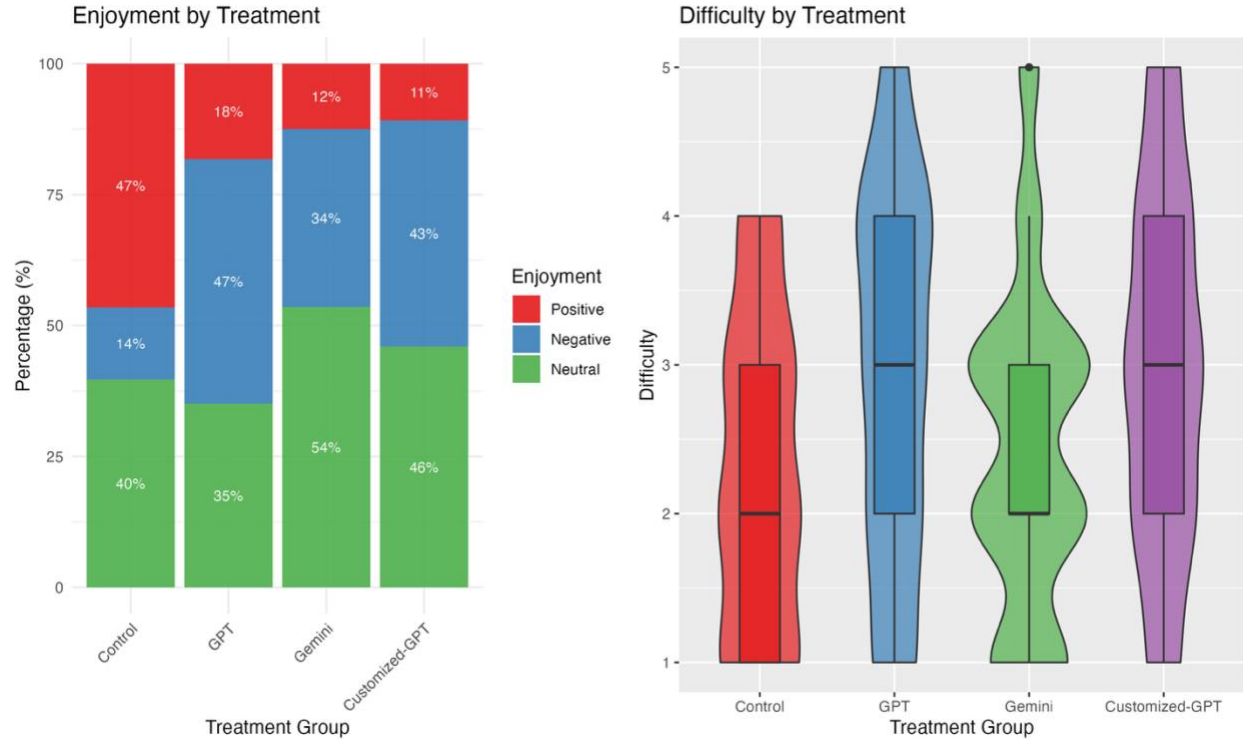
These findings suggest that trust may counteract treatment effects, with GPT-based CRS eliciting more involvement yet less trust than traditional search. Notably, Customized GPT elicited the highest involvement, with trust scores nearly identical to GPT. Participants using Gemini demonstrated higher trust levels than those using GPT, potentially indicating brand-related effects associated with the respective developers (Google vs. OpenAI). However, the mixed variation in trust and involvement indicates that other mechanisms likely contributed more substantially to observed behavioral patterns.

As for satisfaction, we measured the enjoyment and difficulty reported by participants.

Figure 8: Enjoyment and Difficulty by Treatment

⁷ Participants rated statements such as “I trust [CRS]” and “I cannot rely on [CRS]”.

⁸ Participants indicated whether the [CRS] recommendations were, inter alia, important, meaningful, of concern to them.



Participants' reported enjoyment varied significantly across conditions ($\chi^2(6) = 38.25, p < 0.01$). Whereas nearly half of the control group described their experience positively (47%), positive responses declined sharply in the GPT (18%), Gemini (12%), and Customized-GPT (11%) conditions, which were dominated by neutral and negative evaluations. Perceived task difficulty differed significantly across treatment groups ($F(3, 261) = 6.01, p < .001$). Participants in the GPT and Customized-GPT conditions reported higher difficulty (means = 2.94 and 2.97, respectively) than those in the Control group (mean = 2.28), with pairwise tests confirming significant contrasts for both comparisons ($ps < .01$, Bonferroni-corrected).

The findings reported in Figure 8 do not support our hypothesis that customized-CRS generate a backlash in the form of negative user experience (H4). While overall enjoyment levels remain low across all CRS conditions relative to Control, there is no evidence of a meaningful decline specific to the nudging intervention. In fact, enjoyment appears relatively stable across GPT, Gemini, and Customized-GPT, with positive reactions increasing modestly in the customized condition. As to difficulty of use, Customized-GPT was rated as the most, while Gemini and GPT were rated as more difficult than the control group, indicating that the cognitive burden is likely associated with the CRS experience itself. These results suggest that any observed dissatisfaction is not driven by the behavioral nudging mechanism *per se*, but rather by the introduction of CRS as a new RS tool.

B. Linguistic Framing (Tone)

To investigate whether systematic linguistic framing accounts for variations in consumer conduct and spending, we conducted a linguistic analysis of product recommendation transcripts generated by CRS. Using a custom natural language processing (NLP) pipeline, we extracted 50 linguistic features-including part-of-speech diversity, sentiment measures, and psycholinguistic categories-

from 132 conversations (69 GPT, 63 Customized-GPT).⁹ Independent *t*-tests with Bonferroni correction identified statistically significant differences across treatment conditions.

Three features – VADER’s polarity,¹⁰ positive emotion (Empath), and Positive emotion normalized,¹¹ were found to differ between the two conditions of GPT. Customized-GPT consistently employed more positive and emotionally engaging language than standard GPT, suggesting a deliberate shift in tone to appear more helpful and personable. Although the effect sizes are modest (Cohen’s $d = 0.42$ to 0.46), the statistical differences are clear. The manipulation is subtle, but measurable.

These results should be interpreted with caution. Many transcripts were heavily summarized, or generated by the Customized-GPT itself, which may have strategically omitted overt steering cues. Nonetheless, the analysis suggests that Customized-GPT’s more expressive and emotionally resonant style may have modestly contributed to increased user responsiveness.

C. Product Brand Exposure

To assess whether CRS systematically steered users toward brands of higher or lower prominence, we conducted a product brand popularity analysis. The full set of CRS recommendations was extracted from the 134 CSV files generated during the shopping task (69 for GPT and 65 for Customized-GPT). Each CSV contained an average of 7 product recommendations, totaling 918 products with price, brand, and purchase link data. We then compared these recommendations with brands appearing on Amazon’s organic first page results (excluding promoted items) within the same price range (0-15). Using Amazon’s organic rankings (search algorithm) is empirically intuitive but far from normatively clean. We chose it as a natural benchmark comparing what users currently encounter with what they might see under CRS. At the same time, it tells us little about what ‘should’ be recommended to the users to maximize their welfare (Calvano et al., 2025).

Each of the three optional products participants chose from was evaluated against two brand categories: ‘Best Sellers’ and ‘Top Brands’. Best Sellers were defined as brands appearing at least twice in Amazon’s first-page results after applying the Amazon ‘bestsellers’ filter, which updates

⁹ As noted above, participants encountered technical difficulties when attempting to generate such reports from Gemini and therefore we conducted a comparison among the two GPT conditions.

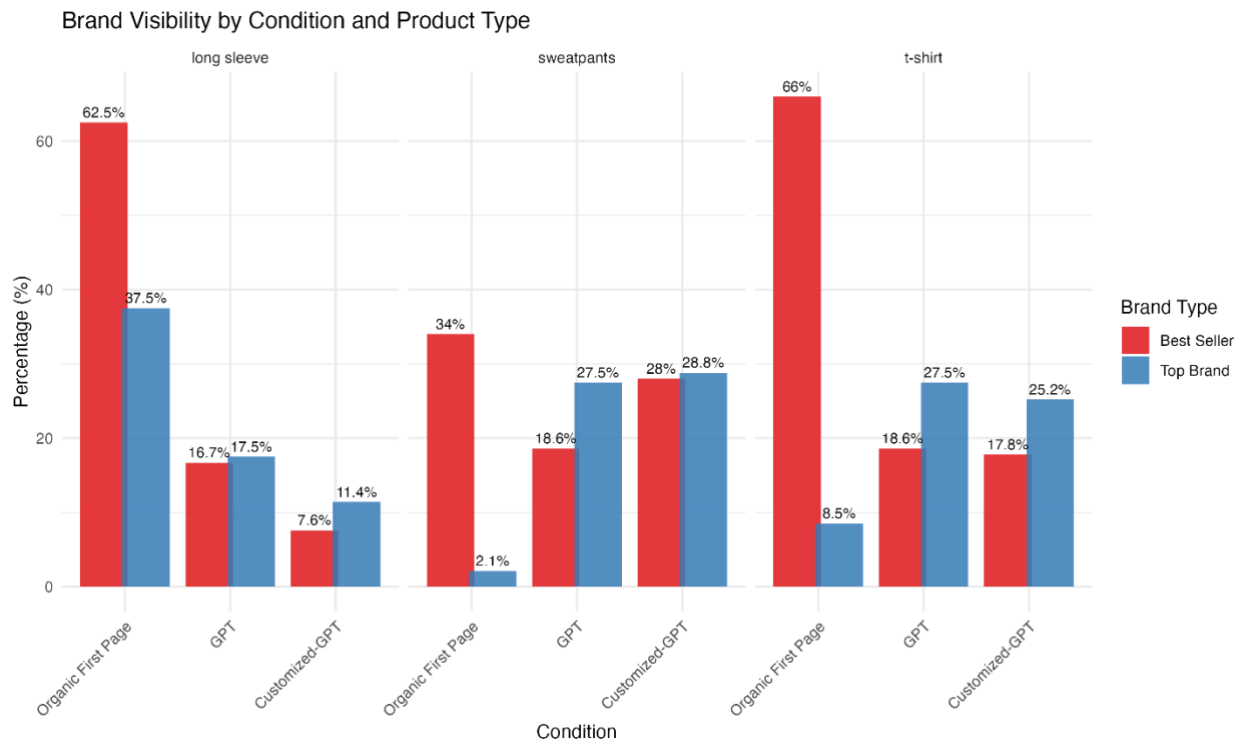
¹⁰ VADER positive sentiment refers to the portion of a text that expresses favorable, optimistic, or approving emotions as detected by the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool. VADER is a lexicon- and rule-based sentiment analysis model specifically attuned to sentiments expressed in social media and informal text (e.g., tweets, reviews, comments). VADER returns four scores: pos, neu, neg (proportions summing to 1) and compound ([-1,1]). We use the pos field as the proxy for positive sentiment intensity (share of text flagged as positive after VADER’s lexicon and valence heuristics). See Hutto & Gilbert (2014).

¹¹ Positive emotion (Empath) refers to the measurement of how much a piece of text expresses pleasant, uplifting, or emotionally positive states - as identified by the Empath lexical analysis tool. Empath is a lexicon-based text analysis system developed by Fast, Chen, & Bernstein (2016). It uses deep learning and crowd-sourced validation to automatically generate semantic categories (such as “anger,” “joy,” “violence,” “trust,” etc.) and count how often words in a text fall into each category. For each conversation, we score the Empath category “positive_emotion” over the full text. This yields a count of words mapped to that category (in our results, treated as a raw count). We then also compute a normalized version by dividing by word_count and multiplying by 100, producing a “per 100 words” rate (positive_emotion_normalized), which we use as a robustness check to control for text length.

frequently based on recent and historical sales data.¹² Since Amazon does not provide a comprehensive top clothing brands list, Top Brands were identified using merchantwords.com's 2020 database of Amazon's leading fashion brands, selecting the top ten per category.

The results reveal a marked divergence between CRS recommendations and Amazon's organic rankings. Figure 9 compares the prevalence of Best Seller and Top Brand products recommended by GPT and Customized-GPT with those appearing on the first page of Amazon's organic results across the three product categories. Amazon's organic results significantly overlap with Best Seller brands. However, organic results consistently outperform CRS recommendations in featuring Best Sellers. For example, 66% of Amazon's first-page organic results for t-shirts were Best Sellers, compared to only 18.6% (GPT) and 17.8% (Customized-GPT) – a statistically significant difference ($p < 0.001$). Conversely, both GPT variants recommended substantially more Top Brand products than appeared organically (27.5% and 25.2% vs. 8.5%, $p < 0.05$).

Figure 9: Product Brand Visibility by Condition and Product



These patterns suggest that CRS prioritize brand reputation over sales volume (popularity) in Amazon, potentially reflecting differences in training data sources or optimization objectives, given that Amazon's algorithm likely favors products with established sales performance on its platform. Contrary to Hypothesis 2, the evidence does not support the prediction that CRS promote a more mainstream choice set based on popularity. Instead, the systems appear to shift exposure

¹² Applying the filter provided an indirect method to extract bestseller lists, which are not directly observable in Amazon's interface. The minimum two-appearance threshold narrowed results to approximately 10 brands, focusing on the most significant bestsellers.

patterns toward premium or branded products, falsifying the proposed “mainstream effect,” or at least requiring its reinterpretation from popularity-driven to reputation-driven recommendations.

5.4 API Simulations

To further assess the potential effects of CRS on brand exposure, we conducted a large-scale simulation using OpenAI's API services.¹³ We first identified top search terms from actual searches conducted on Amazon from the merchantwords.com database, in four major product categories: baby products, home & kitchen, beauty & personal care, and men's fashion. After removing non-product search terms, we retained the top 50-100 ones per category (e.g., ‘baby_wipes’, ‘nail polish remover’ and ‘black_t_shirt’). We then designed two structured shopping contexts - ‘best value-for-money’, and ‘most sustainable’- to capture different consumer priorities. The methodology was inspired by the literature on auditing LLMs studies (Haim, Salinas, and Nyarko, 2024). The best value-for-money context was designed to align closely to the behavioral experiment.

Each search term was then combined with each shopping context and submitted to GPT-4o using the following prompt structure:

I want to buy [search term]. Find the [context] [search term] in [category] category from Amazon in the United States. Please provide exactly 10 products in JSON format with ranking, product_name, brand, price, and URL.

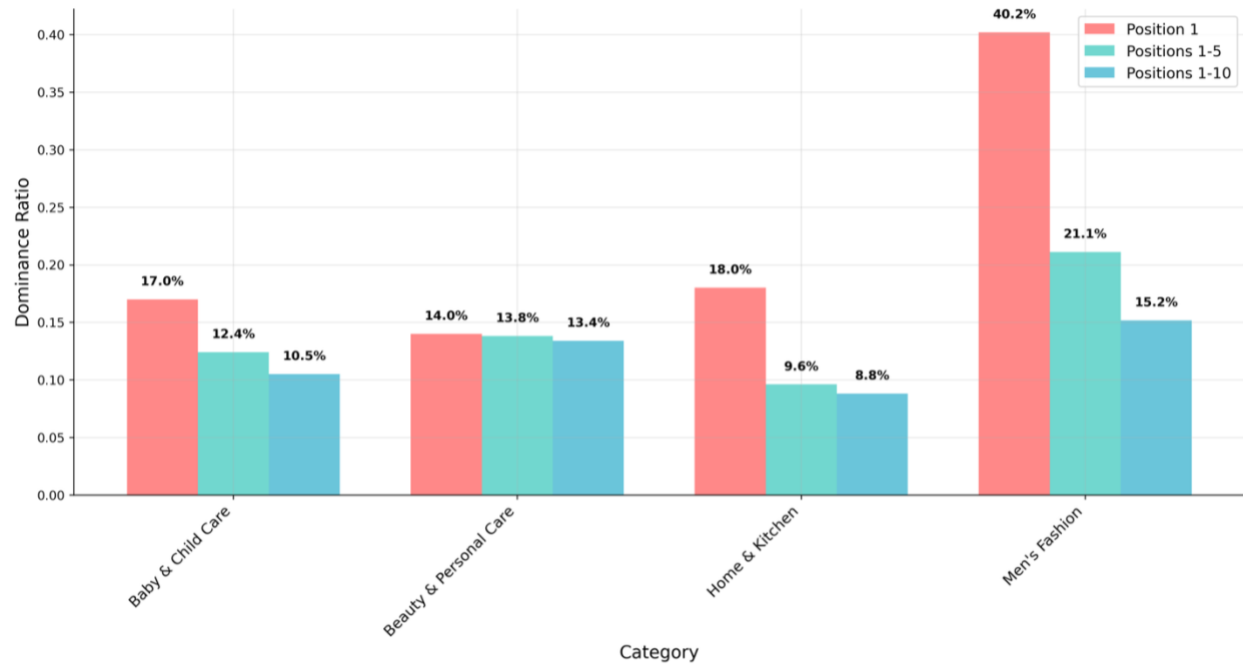
The analysis yielded 5,990 individual product recommendations (599 responses×10 products in each response). Brand names were normalized for case sensitivity. This methodology enabled systematic analysis of how CRS affect brand visibility across different product categories and consumer contexts.

The analysis identified 1,247 distinct brands across all categories, with significant variation in recommendation diversity. Diversity ratios - calculated as the percentage of distinct brands relative to total product recommendations in each product category - ranged from 4.4% in men's fashion to 9.8% in home & kitchen products. Brand concentration was most pronounced at the top recommendation position (#1), with men's fashion exhibiting particularly high concentration: the five leading brands captured 40.2% of the top 1 position, compared to 14% in Beauty & Personal Care and 17–18% in other categories. Overall, the top position is highly concentrated, while lower positions (1-10) show more diversity. These findings suggests a “winner-takes-much” dynamic at the very top of recommendations.

Next, we examined outcomes across the two distinct contextual prompts. The “most sustainable” shopping prompt demonstrated higher brand concentration in top 1 positions (14.3% vs. 12.7% for value prompts), commanded premium prices (+\$16.77 on average), and generated more positive sentiment scores (61.8% vs. 58.2%), suggesting that different consumer searches may systematically alter product recommendations.

Figure 10: API Simulations, Top 5 brands Concentration by Position and Category

¹³ <https://openai.com/api/>



6. Discussion

7.1 Price effects of CRS

This extensive empirical study – the first to test real consumption choices across multiple CRS and diverse product categories – provides compelling evidence that CRS, particularly when customized to promote higher-priced items, exert systematic and measurable upward influence on consumer spending. The observed price premiums relative to traditional search, ranging from €0.93 to €1.30 depending on model specification, represent economically meaningful shifts in consumer expenditure patterns given the bounded shopping allowance (€15) and price distribution. This is a result that prior work predicted but had not demonstrated empirically (Gal and Zac, 2024). The experiment also revealed differences between Gemini and GPT-based CRS, with the latter increasing consumer spending more than that former.

Beyond documenting such price effects, our study is the first to disaggregate the underlying influence mechanisms, offering more precise understanding of how CRS shape consumer behavior. As we show, price differences persisted even after controlling for product quality ratings, online shopping experience, and digital literacy. This decoupling of price from quality demonstrates that CRS operate as active choice architects rather than passive information providers, fundamentally challenging the neutrality of such algorithmic intermediation. Similarly, trust in the CRS failed to explain the observed price effects.

Instead, our study uncovers two behavioral mechanisms that affect consumers' spending: brand exposure and linguistic framing. The analysis indicated a systematic shift in visibility: while Amazon's organic rankings prioritize "best sellers," GPT-based recommenders direct consumer attention toward established brand names. This shift confirms observational and theoretical studies cited above, which suggest that CRS introduce or amplify biases in product visibility. Our finding suggests that most relevant are conformity bias, where recommendations align with perceived

high-status brands, and exposure bias, where certain products are consistently favored regardless of sales rank or objective quality. Notably, we find no evidence of bias towards the most popular products that consumers purchase. On its face, this finding contradicts the homogeneity bias expected to emerge from the statistical nature of CRS (Shur-Ofry, 2024). Two potential explanations may account for this. First, consumers may often search for information on branded products even when they ultimately purchase alternatives. Second, branded products may be more prominent in the training data due to their long-term reputation and visibility. Our findings show that the well-recognized tactic of investment in brand signaling (Nelson, 1974) remains relevant even in the age of CRS, where the signaling is now to an algorithm rather than a human.

The notion that consumer preferences are constructed rather than fixed, shaped dynamically by framing, context, and comparison sets, has deep roots in behavioral science (Kahneman & Tversky, 1974, 1981; Thaler, 1980; Tversky & Simonson, 1993; Ariely, 2008). Our results contribute to this line of research by showing how CRS can operationalize these mechanisms in real time. In particular, the Customized-GPT condition illustrates how strategic framing and product exposure can jointly increase consumer expenditure, even in the absence of explicit persuasion or quality differentials. Although our design does not permit a clean decomposition¹⁴ of these effects, the overall pattern is clear: CRS systems function as choice architects, subtly steering consumer behavior through a blend of linguistic nudging and curated option sets. This aligns with emerging evidence on digital choice architecture (e.g., Werner et al., 2024) and underscores a broader point: the deployment of LLM-powered systems will likely reshape consumer welfare not (necessary) through overt manipulation, but through the systematic structuring of decision environments.¹⁵

Our findings raise significant concerns. Participants reported much higher levels of trust in the Control condition than in GPT or Customized GPT, yet the latter groups still spent more and was more engaged in the search. Furthermore, trust in GPT or Customized GPT was almost identical,

¹⁴ In this design, we could not fully separate two linguistic framing dimensions potentially at work in the CRS: (i) taste formation, in which preferences are shaped through more appealing or emotionally engaging product descriptions, and (ii) selective information provision, in which preferences are informed through curated factual content (e.g., review summaries, comparative features, or by affecting preference fluency, Novemsky et al., 2007). Our analysis captures the first dimension (Jones & Bergen, 2024), but not the informational one—for example, whether the CRS supplied concrete, factual details about products. Distinguishing these “soft” and “hard” framing dimensions would help clarify the mechanisms through which CRS increase spending. Taken together, the findings suggest that affective tone, even when varied modestly, can subtly shape spending. At the same time, the relatively modest effect sizes for linguistic variations indicate that framing likely operates through multiple, interacting channels rather than through a single communication strategy.

¹⁵ The implied price premium, between €0.93 and €1.30 per item, translates into roughly €44 to €62 in additional annual spending per user, assuming a typical monthly purchase frequency (4 items per month). Scaled to a population of one million users, this effect amounts to approximately €44 to €62 million in additional annual expenditures. To gauge the potential welfare implications, it is useful to consider how these higher prices interact with underlying demand elasticities. Using the experiment’s average item price of €9.50 and a mid-range premium of €1.12, the induced price increase represents an 11-12 percent rise. Applying an online demand elasticity of 0.87 (Zhang, & Demirkan, 2021) the resulting deadweight loss is modest on a per-user basis (about €2.7 annually) but accumulates meaningfully at scale, reaching roughly €2.7 million per million users per year, and 920 million for the US population (2025). These figures are in addition to the pure transfer from consumers to sellers and platforms generated by higher baseline spending. Taken together, the evidence suggests that customized recommendation systems can systematically shift consumers toward higher-priced options, generating large aggregate transfers and non-trivial efficiency losses once deployed at population scale.

suggesting that participants remained largely unaware of higher levels of manipulative steering. The fact that users of CRS reported higher engagement levels than the Control group further strengthens this concern, as one would expect that the more a consumer is engaged, the more he or she would notice manipulations and bias.¹⁶ The relevance of this finding is further strengthened by claims that GPT is systematically prompted to increase user engagement in order to keep users on the platform (Grosser and Bro Pold, 2025), implying that other – potentially external – mechanisms might be needed to protect consumers against manipulation and bias.

Our finding that observed dissatisfaction is not driven by the behavioral nudging mechanism *per se*, but rather by the introduction of CRS as a tool for shopping, may be time-dependent. It may reflect the early stage at which the experiment was conducted, in which participants were not accustomed to the use of CRS for shopping, and is likely to diminish as consumers become more accustomed to using CRS in commercial contexts, or as developers design more enjoyable CRS interfaces. Our findings regarding ratings of difficulty of use reinforce this interpretation.

Overall, the combination of higher spending despite increased engagement, unchanged product quality and lower trust, highlights a troubling asymmetry: CRS can modify behavior beneath conscious awareness, inducing consumers to pay more – while not paying the price of nudging – similar to other cases of online manipulations (Luguri & Strahilevitz, 2021; Zac et al., 2023).

The implications of our findings extend beyond individual consumers to encompass broader market competition concerns. Whereas earlier generations of RS primarily altered visibility by ranking products, CRS add an interactive, conversational layer that deepens their ability to shape consumer preferences, thereby amplifying structural risks already associated with RS. The elevation of brand prestige over marketplace performance redistributes visibility in ways that are largely invisible to consumers yet highly consequential for firms. At scale, these effects reduce competitive neutrality, lead to concentrated markets where algorithmic preference rather than product merit determines success, entrench incumbents and raise barriers for innovative or lower-cost entrants. As AI adoption accelerates, a new form of market power may emerge, where algorithmic persuasion – rather than product merit or competitive pricing – determines success, with consumers unwittingly aiding this transformation through their AI-effected purchase decisions.

The recent emergence of AI gents may further exacerbate these concerns. During our study period (March–July 2025), most consumers lacked access to this technology. Since then, however, major firms have introduced prototypes of such agents capable of autonomous shopping assistance (ARS), including OpenAI's Operator (OpenAI, 2025) and Google's Project Mariner (Google labs, 2025). ARS reduce consumer involvement in purchasing decisions, widening the gap between how products are selected and consumers' knowledge and understanding of how that choice is made (Blickstein-Shchory and Gal, 2022). Accordingly, they further reduce the extent to which we can rely on consumer awareness and involvement to limit RS bias, especially when all the freely available ARS might be biased.

¹⁶ This finding might partly be driven by experimental limitations whereby participants may be incentivized to report that they were engaged in the experiment.

7.2 Mitigation strategies: Market and Regulatory Solutions

Mitigation strategies can be conceived along three dimensions: algorithmic design, legal liability, and consumer empowerment. The deployment of such strategies spans multiple stakeholders – RS operators, users, and regulators – each offering sometimes distinct yet often complementary mitigation tools. All must be carefully calibrated to avoid undermining firms' competitive and innovative incentives in ways that could ultimately harm social welfare.

At the design level, interventions include diversity-enhancing re-ranking and fairness constraints designed to reduce systematic overexposure to particular brand categories. Our findings suggest that targeted interventions should prioritize curbing brand-conformity bias and establishing oversight mechanisms for CRS prompt design to prevent deliberate steering toward higher-priced items without corresponding quality improvements. Schinkel and van Oosten (2024) suggest a variant of such intervention by employing "smart cuts" that separate commercial incentives from recommendation logic, thereby reducing economic incentives for price-steering behavior. Their solution, however, does not address all the types of biases identified in our experiment, such as those resulting from bias in the learning data.

Regulatory frameworks offer complementary tools. Existing laws already address some of the mechanisms identified in our experiment, or their results. To illustrate, consumer protection laws prohibiting unfair or deceptive acts or practices, particularly those prohibiting Dark Patterns – adaptive, data-driven interfaces that manipulate decision-making and generally operate below user awareness (FTC, 2022) – apply directly. Competition law prohibitions on erecting artificial entry barriers by firms with market power, and laws prohibiting self-preferencing (EU Digital Markets Act 2024; Motta, 2023) can apply where CRS promote their own brands. In addition, fiduciary duties may be interpreted to mandate RS operators to act loyally in consumers' best interests (Blickstein-Schory and Gal, 2022). Effective application of these regulatory tools requires empirical understanding of how CRS actually affect consumer conduct and market competition – precisely the evidence this study provides.

Transparency requirements, such as disclosures regarding recommendation criteria, manipulation tactics used, and potential conflicts of interest (Tintarev and Masthoff, 2011), could potentially reduce the effectiveness of covert framing mechanisms identified in our study by altering consumer trust in algorithmic decisions. Their efficacy, however, is constrained by two main factors: consumers' limited attention to disclosure information (Ben Shachar and Schneider, 2014) and habituation effects, whereby repeated exposure to warnings reduces their salience and behavioral impact (Acquisti et al., 2017; Sunstein, 2020).

Finally, consumers can deploy algorithms as self-help remedies to counteract manipulations by CRS (Gal and Elkin-Koren, 2017). This mitigation strategy involves consumers utilizing ARS specifically designed to counteract CRS manipulation and bias, by executing purchases based on user-defined parameters rather than platform recommendations. Such algorithmic consumers can generate substantial welfare benefits by reducing search costs, personalizing purchases to individual preferences, and protecting users from manipulative interface designs. They represent a form of distributed agency where consumers become part of the socio-algorithmic assemblage, using technology to counter the influence of other technological systems (Kotliar, 2021). However, their effectiveness depends, *inter alia*, on ensuring that consumer agents genuinely serve user interests (Stucke and Ezrachi, 2017) and on their access to data.

To demonstrate the potential of algorithms as self-help remedies, we experimented with algorithmic detection of the manipulative behavior. Our goal was to examine whether a neutral LLM could identify steering practices by classifying our experimental conversation transcripts as either neutral (generated by standard GPT) or manipulative (generated by Customized GPT). Classification performance demonstrated moderate success. For manipulative conversations, the model achieved precision of 0.636 and recall of 0.400 (F1-score 0.491). For neutral conversations, precision was 0.567 with recall of 0.775 (F1-score 0.655). The weighted average F1-score reached 0.574. Performance improved through prompt refinement and data filtering: implementing a 3,000-character minimum threshold increased accuracy by approximately 10%, though this reduced the usable sample by 70%. Classification accuracy was limited by the fact that many transcripts were heavily summarized, or generated by the manipulative system itself, potentially obscuring explicit steering cues that would aid detection. These findings suggest potential applications for both private consumer protection and regulatory oversight. Rather than relying exclusively on ex-ante compliance reporting as mandated under frameworks like the EU Digital Markets Act (Regulation 2022/1925), regulators and consumers could deploy neutral LLM overseers to continuously detect steering toward higher-priced products, preferred brands, or self-preference in real-time, triggering consumer alerts or human review processes. This monitoring approach, echoing insights from prior digital markets regulation (Zac et al., 2025), would operationalize protective nudging by leveraging AI capabilities to detect and deter covert manipulation within AI-mediated commercial environments. It also helps overcome the problem of consumer awareness of potential bias that our study exposed.

8. Conclusion

This study contributes to scholarship and policy debates on algorithmic intermediation. It provides robust experimental evidence that CRS can systematically elevate consumer spending, even in the absence of quality improvements or higher trust. By combining the persuasive power of conversation with hidden biases in exposure and framing, CRS elevate prices and privilege incumbents, all while operating beneath conscious awareness. What makes this finding particularly troubling is the ease with which CRS can be designed to manipulate consumers – requiring minimal technical sophistication yet potentially yielding substantial behavioral effects. Furthermore, its mechanism analysis offers theoretical foundations for understanding which mechanisms enable CRS to alter the decision-making processes. Conceptually, the study demonstrates that CRS can function not merely as neutral information filters but as active choice architects, capable of steering consumer preferences through subtle linguistic framing and strategic brand exposure. As these lines are written, the CRS market is rapidly expanding, and firms are using millions of online interactions to test how to steer users toward higher spending, or optimizing recommendations for the highest bidder. Policymakers must consider whether existing regulatory frameworks adequately address the risks posed by CRS, which combine persuasive capabilities with increasingly sophisticated behavioral targeting.

The study also contributes methodologically, developing and applying novel methods for investigating AI-mediated choice architecture. Its experimental framework provides a replicable methodology for assessing RS influence on consumer behavior. Furthermore, the integration of controlled experimentation with large-scale simulations demonstrates the potential for hybrid research designs that combine internal validity with external generalizability, addressing traditional limitations of laboratory-based research.

While the results point to a capacity for CRS to alter consumer choices, they are drawn from a single-task laboratory setting and should not be generalized without caution. Future research should extend these findings across diverse product categories, consumer demographics, and cultural contexts to establish the persistence and scale of these effects. Longitudinal studies examining adaptation and learning effects will be crucial for understanding whether observed manipulations persist as consumers gain experience with CRS. Additionally, comparative analyses across different CRS architectures and training methodologies may illuminate the specific design features that generate behavioral influence, informing both regulatory interventions and technological development priorities. The development of detection mechanisms for identifying and mitigating manipulative algorithmic behaviors represents another critical research frontier. Finally, future research should examine how suppliers strategically respond to algorithmic recommendations of the type we identify, particularly whether such responses amplify or mitigate competitive concerns over time. It should also examine how competition between the different CRS shapes their outcomes, especially once consumer AI agents are introduced. Nonetheless, the experimental evidence provides a strong and much needed factual basis for ongoing debates about the economic and regulatory implications of AI-powered recommendation systems.

References

- Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The Unfairness of Popularity Bias in Recommendation. *ArXiv:1907.13286 [Cs]*. September. <https://arxiv.org/abs/1907.13286>
- Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., et al. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3), 1-41.
- Aguiar, L., & Waldfogel, J. (2018). Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists. *RePEc: Research Papers in Economics*, June. <https://doi.org/10.3386/w24713>
- Anderson, S. P., & Renault, R. (1999). Pricing, Product Diversity, and Search Costs: A BertrandChamberlinDiamond Model. *The RAND Journal of Economics*, 30(4), 719–35. <https://doi.org/10.2307/2556072>
- Aridor, G., Goncalves D., Kluver D., Kong R., & Konstan J. (2024). The Informational Role of Online Recommendations: Evidence from a Field Experiment. *ArXiv.org*. <http://arxiv.org/abs/2211.14219>
- Ariely, D., & Jones, S. (2008). *Predictably irrational*. New York: HarperCollins.
- Bar-Isaac, H., Caruana, G., & Cuñat, V. (2012). Search, Design, and Market Structure. *American Economic Review*, 102(2), 1140–60. <https://doi.org/10.1257/aer.102.2.1140>
- Blickstein-Schchory, N., & Gal, M. S. (2022). Voice shoppers: From information gaps to choice gaps in consumer markets. *Brooklyn Law Review*, 88 (1), 111.
- Duivenvoorde, B. (2024). Generative AI and the Future of Marketing: A Consumer Protection Perspective. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4967461>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). Language Models Are Few-Shot Learners. *Arxiv.org*, 4 (33). <https://arxiv.org/abs/2005.14165>
- Brynjolfsson, E., Hu, Y. J., & Simester, D. (2011). Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science*, 57 (8), 1373–86. <https://doi.org/10.1287/mnsc.1110.1371>
- Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12 (4), 331–70. <https://doi.org/10.1023/a:1021240730564>
- Calvano, E., Calzolari, G., Denicolo V., Pastorello S. (2025), Artificial Intelligence, Algorithmic Recommendations and Competition. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4448010

- Castellini, J., Fletcher, A., Ormosi, P. L., & Savani, R. (2025). Recommender Systems and Competition on Subscription-Based Platforms. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4428125>
- Chaney, A., J. B., Stewart, B. M., & Engelhardt, B. E. E. (2018). How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18*. <https://doi.org/10.1145/3240323.3240370>
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M. & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3), 1-39.
- Cornière, A., & Taylor, G. (2014). Integration and Search Engine Bias. *The RAND Journal of Economics*, 45(3), 576–97. <https://doi.org/10.1111/1756-2171.12063>
- Deldjoo, Y., & Di Noia, T. (2025). CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. *ACM Transactions on Intelligent Systems and Technology*.
- Ezrachi, A. & Stucke, M.E. (2017). Artificial intelligence & collusion: When computers inhibit competition. *U. Ill. L. Rev.* 1775.
- Fast, E., Chen, B., & Bernstein, M. S. (2016, May). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4647-4657).
- Fletcher, A., Ormosi, P. L., & Savani, R. (2023). Recommender Systems and Supplier Competition on Platforms. *Journal of Competition Law and Economics*, 19(3). <https://doi.org/10.1093/joclec/nhad009>
- Fletcher, A., Ormosi, P. L., Savani, R., & Castellini, J. (2023). Biased Recommender Systems and Supplier Competition. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4319311>
- FTC. (2022). Bringing Dark Patterns to Light: An FTC Report on Dark Patterns and How They Harm Consumers. *Federal Trade Commission*.
- Gal, M. S., & Elkin-Koren, N. (2016). Algorithmic consumers. *Harv. JL & Tech.*, 30, 309.
- Gal, M. S., & Zac, A. (2024). Is Generative Ai the Algorithmic Consumer We Are Waiting For? *Network Law Review*. <https://ssrn.com/abstract=4755455>
- Google labs. (2025). *Project Mariner*. <https://labs.google.com/mariner/landing>
- Grosser, B. & Bro Pold S., Reading the Praise/Prompt Machine: An Interface Criticism Approach to ChatGPT. In Association for Computing Machinery, *Conference Proceedings - Computing X Crisis: 6th Decennial Aarhus Conference, AAR 2025* (pp. 259-271). <https://doi.org/10.1145/3744169.3744194>

- Hagiu, A., & Jullien B. (2011). Why Do Intermediaries Divert Search? *The RAND Journal of Economics*, 42(2), 337–62. <https://doi.org/10.1111/j.1756-2171.2011.00136.x>
- Haim, A., Salinas, A., & Nyarko, J. (2024). What's in a Name? Auditing Large Language Models for Race and Gender Bias. *ArXiv.org* <https://doi.org/10.48550/arXiv.2402.14875>
- Helberger, N., Sax, M., Strycharz, J., & Micklitz, H. W. (2022). Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy*, 45(2), 175-200.
- Huang, C., Wu, J., Xia, Y., et al. (2025). Towards Agentic Recommender Systems in the Era of Multimodal Large Language Models. *arXiv preprint*.
- Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
- Jannach, D., & Jugovac, M. (2019). Measuring the Business Value of Recommender Systems. *ACM Transactions on Management Information Systems*, 10(4), 1–23. <https://doi.org/10.1145/3370082>
- Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate Feedback Loops in Recommender Systems. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 383–90. <https://doi.org/10.1145/3306618.3314288>
- Jones, C. R. & Bergen, B. K., (2024). Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv preprint*. arXiv:2412.17128.
- Kotliar, D. M. (2021). Who gets to choose? On the socio-algorithmic construction of choice. *Science, Technology, & Human Values*, 46(2), 346-375.
- Lau, G. T., & Lee, S. H. (1999). Consumers' Trust in a Brand and the Link to Brand Loyalty. *Journal of Market-Focused Management*, 4(4), 341–70. <https://doi.org/10.1023/a:1009886520142>
- Lee, K. H., & Mussolff, L. (2025). Two-Sided Markets Shaped by Platform-Guided Search. *Working paper*.
- Li, X., Zhang, Y., & Malthouse E. C. (2023). A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News. *arXiv preprint* arXiv:2306.10702.
- Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., Liu, Y., et al. (2024). How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Transactions on Office Information Systems*. <https://doi.org/10.1145/3678004>
- Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Studies*, 13(1), 43-109.

- Mittal, B. (1995). A Comparative Analysis of Four Scales of Consumer Involvement. *Psychology and Marketing*, 12(7), 663–82. <https://doi.org/10.1002/mar.4220120708>
- Motta, M., (2023). Self-preferencing and foreclosure in digital markets: Theories of harm for abuse cases. *International Journal of Industrial Organization*, 90, 102974.
- Nelson P. (1974). Advertising as Information. *Journal of Political Economy*, 82(4), 729-754.
- Nguyen, T. T., Hui, P., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14, April*, 677–86. <https://doi.org/10.1145/2566486.2568012>
- Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of marketing research*, 44(3), 347-356.
- O’Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, 33(4), 459–78. <https://doi.org/10.1177/0894439314555329>
- OpenAI. (January 3, 2025). *Introducing Operator*. https://openai.com/index/introducing-operator/?utm_source=chatgpt.com
- Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. *The Adaptive Web*, 4321, 325–41. https://doi.org/10.1007/978-3-540-72079-9_10
- Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives 2019/1937 and 2020/1828 (EU Digital Markets Act), OJ L 265, 14.10.2022, p. 1–66.
- Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to Recommender Systems Handbook. *Recommender Systems Handbook*, 1–35. https://doi.org/10.1007/978-0-387-85820-3_1
- Schinkel, M. P., & van Oosten, R. J. (2025). Breaking up Big Tech: scissor-line suggestions for smart cuts. *Journal of Antitrust Enforcement*, 13(2), 237-264.
- Shur-Ofry, M. (2025). Multiplicity as an AI Governance Principle. *forthcoming Indiana Law Journal*. <http://dx.doi.org/10.2139/ssrn.4444354>
- Su, X., & Khoshgoftaar, T. M. (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 1–19. <https://doi.org/10.1155/2009/421425>
- Sunstein, C. R. (2020). Behavioral welfare economics. *Journal of Benefit-Cost Analysis*, 11(2), 196-220.
- Teppan, E. C., & Zanker, M. (2015). Decision Biases in Recommender Systems. *Journal of Internet Commerce*, 14(2), 255–75. <https://doi.org/10.1080/15332861.2015.1018703>

- Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of economic behavior & organization*, 1(1), 39-60.
- Tintarev, N., & Masthoff, J. (2015). Explaining Recommendations: Design and Evaluation. *Springer EBooks*, 353–82. https://doi.org/10.1007/978-1-4899-7637-6_10
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157), 1124-1131.
- Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, 39(10), 1179-1189.
- Werner, T., Soraperra, I., Calvano, E., Parkes, D. C., & Rahwan, I. (2024). Experimental Evidence That Conversational Artificial Intelligence Can Steer Consumer Behavior without Detection.” *ArXiv.org*. <http://arxiv.org/abs/2409.12143>
- Wolinsky, A. (1987). Matching, Search, and Bargaining. *Journal of Economic Theory*, 42(2), 311–33. [https://doi.org/10.1016/0022-0531\(87\)90090-1](https://doi.org/10.1016/0022-0531(87)90090-1)
- Wu, Y. F., Liu, S., Xue, Z., Cai, Q., Hu, L., Jiang, P., Gai, K., & Sun, F. (2023). A Large Language Model Enhanced Conversational Recommender System. *arXiv preprint arXiv:2308.12345*.
- Zac, A., Bechtold, S., Wey, P., Rodriguez, D. & Del Alamo, J. M. (2025). Digital Markets and Hidden Noncompliance. *upcoming in Journal of Legal Studies*.
- Zac, A., Huang, Y. C., von Moltke, A., Decker, C., & Ezrachi, A. (2023). Dark patterns and consumer vulnerability. *Behavioural Public Policy*, 1-50.
- Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.
- Zhang, X. L., & Demirkan, H. (2021). Between online and offline markets: A structural estimation of consumer demand. *Information & Management*, 58(4), 103467.
- Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., et al. (2024). Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 1–20.

Appendix A

Instructions for control participants

General Information

You will be given 15 euros allowance to shop for an item on Amazon. You may choose the item from a few options. The product must be priced at a maximum of 15 euros. The product you will buy can be sent to the university address or to your house, depending on your preference. If you choose to deliver to the university, you will be informed about how to pick it up. Otherwise, you are free to buy any product you want. Every euro left in your allowance will be given to you at the end of the task, in addition to 5 euros compensation for your time. For example, if you buy an item for 7 euros, you will receive the item you selected, and the additional 8 euros left in your allowance plus 5 euros for the time spent. You have 10 minutes to shop for the item, and 5 more minutes to answer a short survey about your shopping experience.

You will not be able to leave the lab before the time ends, so please do not rush. You are not allowed to communicate with other participants, but you are allowed to contact the researchers for technical issues or questions you may have. You can raise your hand if you need assistance.

Please follow the guidelines and begin the study.

Please choose a specific search query to shop for an item you would like to buy online on the Amazon shop from the list below.

- ☐ *Black t-shirt (1)*
- ☐ *Long sleeve t-shirt (2)*
- ☐ *Sweatpants(3)*

Your task is to go online and use the search query to find the product you would like to buy. For example, you may ask for product recommendations based on your specific preferences or just use the search term in any search engine you know. You are free to use any search engine or website.

You have to buy 1 item, maximum 15 euros.

Do not log into any personal account (e.g., Google, Amazon or such). Do not change the language in the search engine or Amazon; the chosen language should be English.

Find the product on <https://www.amazon.com/>. Make sure to open Amazon in a new tab and keep both tabs open until the end of the experiment. Also make sure that on the upper left side of the

screen, "Deliver to Netherlands" is chosen. If it already says "Netherlands," click on "Dismiss," as shown in the example below.

After finding the product on Amazon and adding it to the Shopping Cart (see example below), signal the researcher with a raise of hand to approve the purchase.

The researcher will confirm the item is within the allowance limits (15 euros) and will instruct you to continue to the survey on this tab, while they process and complete your purchase.

Unique Instructions for Conditions 2 and 4 participants (GPT and Customized-GPT)

Your task is to go online and use the search query to find the product you would like to buy. You will use ChatGPT, which is a chatbot that uses natural language processing to create humanlike conversational dialogue. To do this, please go to <https://chatgpt.com/> and login with the details [xx]. Make sure to open it in a new tab.

You are expected to use ChatGPT as a shopping assistant. For example, you may ask for product recommendations based on your specific preferences. You are only allowed to use ChatGPT to search.

You have to buy 1 item, maximum 15 euros. Do not log in to any personal account (e.g., OpenAI, Amazon). Do not change the language in ChatGPT or Amazon; the chosen language should be English. Find the product on <https://www.amazon.com/>. Make sure to open it in a new tab. Always keep your survey, ChatGPT, and Amazon tabs open until the end of the experiment.

Also make sure that on the upper left side of the screen, "Deliver to Netherlands" is chosen. If it already says "Netherlands," click on "Dismiss," as shown in the example below.

After finding the product on Amazon and adding it to the Shopping Cart (see example below), signal the researcher with a raise of hand to approve the purchase.

The researcher will confirm the item is within the allowance limits (15 euros) and will instruct you to continue to the survey on this tab, while they process and complete your purchase. Note that product recommendations from ChatGPT may provide invalid links. This is normal. You are free to ask for another link or proceed to Amazon and search for the same product manually.

Appendix B - LLM Use Subgroup Analysis

To explore whether the main treatment effects were moderated by actual engagement with the LLMs, we conducted a subgroup analysis including only participants who indicated they had used the LLM (GPT, Gemini, or Customized-GPT), alongside the full Control group. This comparison allows us to isolate potential behavioral effects among those who were actively exposed to the RS, offering a more focused test of their influence.

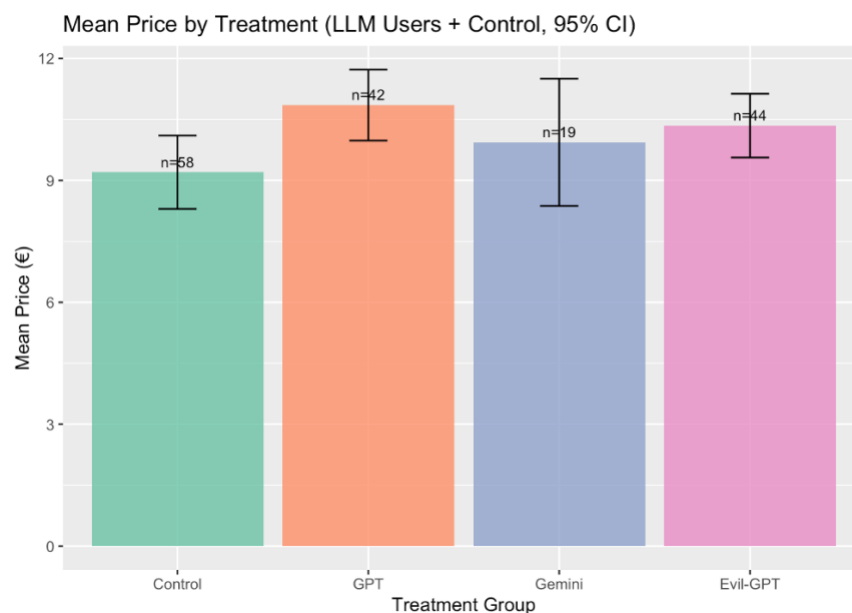
We repeated the main analysis first with the binary treatment effects. This binary treatment analysis revealed a statistically significant difference in average price: participants in the Treated group spent €10.48 on average, compared to €9.20 in the Control group ($t = -2.36, p = 0.020$). The

95% confidence interval for the difference in means ranged from -2.35 to -0.20 , reinforcing the conclusion that LLM exposure-irrespective of model-was associated with higher consumer spending.

The results of a one-way ANOVA (between treatment arms) revealed a marginally significant overall effect of treatment on price in this restricted sample ($F(3, 159) = 2.47, p = 0.064$). Post-hoc Tukey tests showed a statistically significant difference between the GPT and Control groups ($p = 0.050$), with participants in the GPT condition spending on average €1.65 more than those in the Control group. No other pairwise comparisons reached statistical significance. These findings suggest that the observed effects in the full sample were not merely artifacts of random variation but driven by those who actively engaged with the language models.

Visually, the bar plot (Figure A1) illustrates this pattern, with the GPT group displaying the highest average price among all conditions, followed by Customized-GPT and Gemini. Although these differences are not uniformly significant, the direction and magnitude of the effects align with the main analysis. Taken together, this robustness check reinforces the interpretation that interaction with the GPT-based assistant can increase participants' spending, while highlighting the need for caution due to the relatively modest statistical power in this subgroup.

Figure A1: mean Price by treatment (restricted sample - LLM used)



Appendix C - Power and Balance Analysis

To assess the statistical sensitivity of our experimental design, we conducted a post-hoc power analysis using the observed data from all participants. The goal was to estimate the effect size observed in the current sample, calculate the achieved power for detecting this effect, and determine the sample size required to reach conventional thresholds for adequate statistical power (i.e., 80%).

Using the observed group means and standard deviations in participants' spending (WTP), we estimated the effect size for a one-way ANOVA across treatment groups. The resulting Cohen's f was 0.186, which corresponds to a small-to-moderate effect size. The associated eta-squared value was 0.033, indicating that approximately 3.3% of the variance in WTP can be attributed to the treatment condition.

Given the current sample size-based on the smallest treatment group ($n = 56$)-the estimated statistical power to detect the observed effect was approximately 63.1%. While this power level is acceptable for exploratory work, it falls below the commonly accepted threshold of 80%, suggesting a moderate risk of Type II error (i.e., failing to detect a true effect). This suggests that some effects-particularly those that approached statistical significance-may have gone undetected. However, given the rapid evolution of the underlying technology, replicating the study with a larger sample is challenging. Instead, we chose to proceed with analysis and invite further investigation by the broader research community.

To assess whether random assignment successfully created comparable groups at baseline, we conducted balance tests on several pre-treatment covariates. These included participants' age, gender, trust in AI tools (using the adapted Lau & Lee scale), digital literacy, and prior online shopping experience. Means were computed for each treatment group, and overall group differences were tested using ANOVA for continuous variables and a chi-square test for gender.

Covariates were generally well balanced across groups. There were no statistically significant differences across treatment arms for gender ($p = 0.063$), digital literacy ($p = 0.068$), or shopping experience ($p = 0.724$), suggesting effective randomization on these characteristics. However, there were significant differences in trust scores ($p < 0.001$) and a marginal difference in age ($p = 0.030$), which were controls for using the regression adjustment. Overall, the results indicate satisfactory baseline comparability, with most covariates showing no meaningful imbalance across treatment arms.

Appendix D - Bootstrapping Analysis

We next assess the robustness of our main findings (table 2) using nonparametric bootstrap inference ($B = 1,000$ resamples), structured in parallel to the binary and group-treatment specifications.

In the binary-treatment models, the CRS vs. Control comparison yields a positive and marginally significant effect on selected price, consistent with previous results. The estimated coefficient remains stable – approximately €0.89 to €0.93 across Models 1 and 2 – and bootstrap standard errors closely mirror their HC1 counterparts (ranging from 0.513 to 0.527 versus ~0.505 to 0.510). Substantive inference remains unchanged: exposure to any AI-based assistance produces a modest but robust increase in spending.

The group-treatment models (Table D1) similarly confirm the main findings. The Customized-GPT condition consistently generates a statistically significant price increase across specifications, with estimates in the range of €1.26 to €1.29 and corresponding bootstrap standard errors between 0.55 and 0.57. The GPT condition remains positive and significant in simpler models, but – as in prior analyses – the effect attenuates and becomes statistically non-significant once full controls

are added. The Gemini condition remains small and statistically indistinguishable from zero throughout. Covariates behave as expected: task duration has no meaningful effect; product indicators remain influential (especially the sweatpants category); and demographic and attitudinal variables (age, gender, trust, shopping experience, and digital literacy) do not significantly predict selected price.

Taken together, bootstrap inference supports the stability and internal consistency of our results. The key treatment effect – an approximately €1.30 increase associated with Customized-GPT – is robust to sampling variability and aligns with both conventional and heteroscedasticity-robust inference. These results reinforce our confidence that the observed effects are not driven by group imbalances or distributional assumptions.

Table D1: Bootstrap SEs – Group Treatment Effects

	<i>Dependent variable: Price</i>		
	Model 3	Model 4	Model 5
GPT	1.142* (0.607)	1.048* (0.575)	0.903 (0.605)
Gemini	0.148 (0.653)	0.127 (0.664)	0.086 (0.698)
Customized-GPT	1.293** (0.547)	1.293** (0.564)	1.256** (0.573)
Age		0.014 (0.068)	0.005 (0.066)
Gender (Female)		-0.456 (0.388)	-0.637 (0.413)
Duration (minutes)			0.001 (0.006)
Trust			-0.132 (0.223)
Shopping Experience			0.201 (0.174)
Digital Literacy			0.229 (0.507)
Product (Longsleeve)			0.534 (0.448)
Product (Sweatpants)			1.415*** (0.496)
Constant	9.201*** (0.459)	9.228*** (1.470)	7.877*** (2.530)
Observations	265	265	265
R ²	0.032	0.037	0.073
Adjusted R ²	0.021	0.019	0.033

Note: OLS with bootstrap standard errors (B=1000). Standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1.