The Help Me Search Era: Do We Choose "Better" with AI?*

Raluca Ursu† Anita Rao‡ Jake Embrey§

July 30, 2026

Abstract

AI-powered search tools are rapidly becoming a primary interface through which consumers discover and evaluate products. This paper asks whether consumer choices improve when they delegate search to AI tools such as ChatGPT, relative to searching on their own via traditional e-retailer websites such as Amazon.com. We proceed in four steps. First, we identify common shopping-related AI prompt types. Second, scraping more than 27,000 responses from ChatGPT (OpenAI's GPT 5.1) and Google AI Mode to these common shopping prompt types, we find that AI consistently adopts a *best-for* format which emphasizes the features each brand is best suited for while leaving others implicit. As a result, only a fraction of features per brand appear in a response, so full information requires follow-up questions, allowing for inferior choices. Third, we develop a search model formalizing this best-for structure. It predicts that AI typically leads to the same choice as under full information, but can induce errors when the preferred product's highlighted attribute dominates and its non-highlighted attribute is inferior. Finally, we test these predictions in a pre-registered incentivized experiment, comparing traditional e-retailer search with AI search. Our results provide the first direct experimental evidence on the causal effect of AI-formatted information on consumer choice quality.

JEL Codes: D83, L15, M31, L86, O33
Keywords: artificial intelligence, consumer search, information presentation, best-for, large language models, online shopping

*We thank seminar participants at Chicago Booth and NYU Stern for helpful comments. Also, Houssein Moussa provided excellent research assistance. All errors are our own.

†Stern School of Business, New York University.
‡McDonough School of Business, Georgetown University.
§Booth School of Business, University of Chicago.

1 Introduction

Consumer search is undergoing a fundamental transformation. For decades, shoppers navigated product markets by visiting physical stores, browsing retailer websites, or entering keyword queries into search engines such as Google. In each of these environments, the consumer was the active agent: she decided which products to examine, in what order, and how much search effort to exert. The rise of generative AI has altered this dynamic. Tools such as ChatGPT, Google AI Mode, and their competitors now comprise a substantial portion of the search process: they retrieve information from dozens of sources, identify top brands and key attributes, synthesize product reviews, and deliver structured comparisons and recommendations directly to the user within seconds. The consumer still controls the prompts she writes, but the responses are AI-generated, and in producing them the AI *restructures* how product information is presented to her — which brands she sees, how they are organized/categorized, and which attributes are foregrounded. According to industry estimates, ChatGPT alone handled approximately 53 million shopping-related prompts per day in 2025 (Chatterji et al., 2025) and 31 percent of consumers already prefer AI to traditional search engines for product discovery.[1] The economic stakes of this transition are large. McKinsey projects that $3 to $5 trillion of global shopping will be conducted through AI interfaces by 2030.[2]

This rapid growth has prompted enthusiasm about AI's potential to reduce search frictions and improve consumer welfare. Yet a systematic understanding of how AI changes the *quality* of consumer decision-making is still lacking. Does AI help consumers choose better products? Or does the way AI presents information introduce new distortions? AI reshapes consumer search along many dimensions at once, such as replacing keyword queries with natural-language conversation and personalizing results to the individual shopper. We focus on one of these dimensions: the way AI restructures product information relative to traditional e-retailer search.[3] We do so because this restructuring is one channel that bears directly on decision quality: it determines the choice set the consumer actually evaluates and the terms on which she compares options.

In this paper we address these questions empirically and theoretically. To do so, we proceed in four steps. First, we used several methods to determine the most common types of AI shopping-

[1] Retrieved December 2025 from www.admarketplace.com. Article written October 2025.
[2] Retrieved December 2025 from www.economist.com. Article written December 2025.
[3] We also abstract away from hallucinations or agents that may try to deliberately mislead consumers, although such effects may be important.

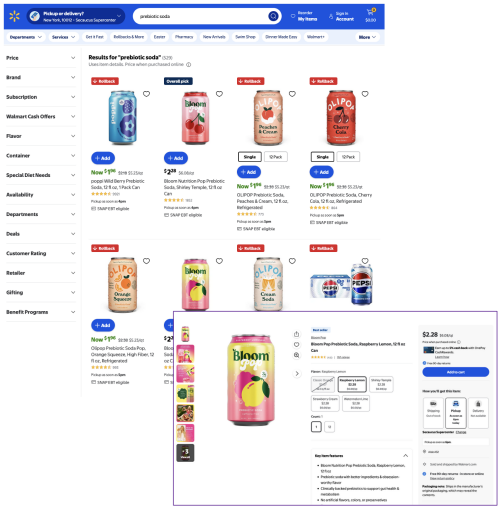
related prompts: (i) asking AI tools directly what are the most common shopping related queries, (ii) investigating industry reports on shopping prompts, (iii) asking Prolific study participants how they phrase their shopping related queries to AI tools, and (iv) using a publicly available data set (WildChat) containing almost 1 million conversations between real users and ChatGPT, focusing on the subset of shopping related conversations. All methods converged on a relatively small list of common shopping prompt types.

Second, we used this list of common shopping-related prompts across several categories and scraped AI responses to these prompts over a three month ending in March 2026. We scraped responses every day at half hour to one hour intervals from both ChatGPT (Open AI's GPT 5.1) and Google AI Mode (tools which are frequently used for shopping), resulting in a data set of approximately 27,000 AI responses across all prompts.[4]

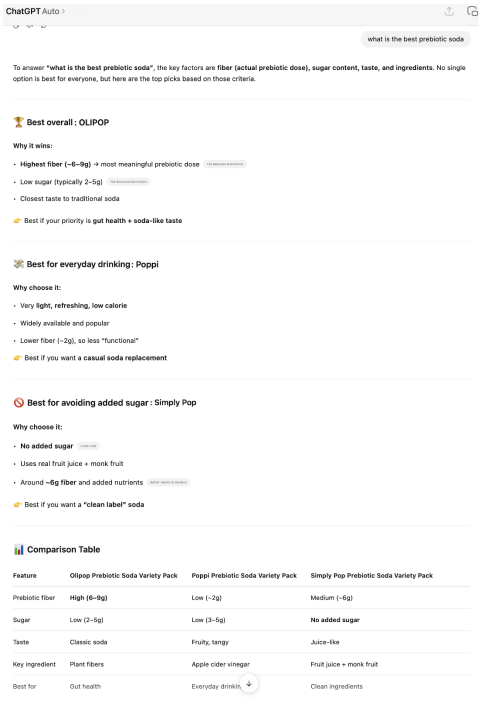
Our central observation from analyzing these AI responses is that AI tools generally present product attribute information in a systematically different manner than traditional e-retailers. We illustrate some of these differences in Figure 1. On an e-retailer's website, such as Amazon.com or Walmart.com, products are presented first on a ranked list, with the same set of attributes (such as brand, photo, price, ratings) visible for all listed products, and another set of attributes (such as additional photos, product description, reviews text) that becomes available after a product on the list is clicked. In other words, information is generally consistent and symmetric across products. In contrast, we find that AI restructures the information presented to consumers in a *best-for* format: each product is described primarily through the attribute or use-case for which it excels, while other attributes are often omitted or mentioned only in passing. For example, in a category such as prebiotic soda, one product may be the best for those who want the highest fiber, another may be the best for everyday drinking due to its low calories, while another may be the cheapest.

Across diverse shopping queries and over 27,000 scraped responses, we find that a majority of ChatGPT and Google AI Mode responses contain explicit best-for language (e.g., "best for," "ideal for," "great for"). This presentation style reflects the structure of sources that AI systems draw upon (i.e., review sites/buying guides, such as Wired, Business Insider, CNET) and the revealed preference of human raters during model training for bullet points, headings, and categorized lists.

[4]In contrast, Claude and Google Gemini are more frequently used for coding and writing.



(a) E-Retailer Search



(b) AI Search

Figure 1: Comparing Search on an E-Retailer Website verus AI

Thus, although AI models are updated frequently, this presentation style may be long-lasting.

Consistent with the best-for format, there is also a lack of feature parity: one attribute may be highlighted for one brand (e.g., price), while another attribute for another brand (e.g., calories). In other words, the best-for format is coupled generally with a lack of full information on brands and their attributes, suggesting that AI's restructuring of product information may introduce new frictions in the consumer's search process. Obtaining full information requires follow-up questions and consumers who do not ask them may settle on inferior choices.

Third, to further trace out the implications of the best-for format for consumer choices, we developed a simple model of AI-mediated search where an AI search tool with full product information assists a consumer without knowing their preferences. In our model, the AI tool presents each of two products as best for one of two attributes, leaving the other attribute unobserved by the consumer. We characterize the conditions under which the consumer's AI-guided choice coincides with the full-information choice. We compare this choice to that under traditional sequential search, where the consumer observes one attribute of both products for free and then chooses whether to

incur a search cost to reveal the second.

Our model makes several clean predictions. In most cases, AI best-for search leads to the same choice as under full information. However, a specific case arises where AI-guided choice diverges from the full-information optimum: this occurs when the product that AI highlights as "best for" the consumer's most-valued attribute happens to score particularly poorly on its non-highlighted attribute; in these cases AI may lead the consumer to choose an inferior product, relative to her preferences. This mistake would be avoided under traditional search, provided search costs were low enough that the consumer searches. We also model how asking follow-up questions of the AI tool can mitigate these negative effects, but this result depends on whether asking follow-up questions of the AI tool is more or less costly than clicks in traditional search.

Finally, we test these predictions in a pre-registered incentivized experiment where participants are asked to choose between six prebiotic sodas. Participants are randomly assigned to one of three conditions - traditional e-retailer search (Trad), AI best-for search (AI-BF), and AI best-for search with follow-ups (AI-Qs). In the Trad condition, participants browsed through options on a website we designed to mimic an e-retailer's website. On this website, participants had to click to obtain full information about each product. In the AI-BF condition, participants received a static AI summary that presented each brand through its highlighted attribute in the best-for format. In the AI-Qs condition, participants had the additional ability to choose among a set of follow-up questions to ask the AI before making a purchase decision. The AI-BF condition serves as a conceptual condition that allows us to disentangle the effect of the AI best-for format on consumer choices from that of information acquisition costs.

Each product differed in brand (Olipop or Poppi), flavor, price, fiber level and calories. To incentivize choices and ensure there were no errors in estimating participants' preferences, we told participants their utility (and therefore their bonus) was a prespecified function of fiber and calories, where fiber was more important. We designed the stimuli such that under full information Poppi provides higher utility than Olipop, but under a best-for information policy Olipop seems to dominate based on the AI response only. We measured choices, time on task, number of clicks and follow-up questions, as well as consumer knowledge of product attributes at the end of the experiment, and their self-described search strategies.

We find that participants in the AI-BF condition earned the lowest bonuses (i.e., made the

worst choices), as the majority in this condition chose the product whose favored attribute (fiber) is highlighted while its unfavorable attribute (calories) remained obscured. In contrast, participants in the Trad condition, who could click to inspect all product information, typically made the payoff-maximizing choice. The average bonus earned in the AI-Qs condition was between that of AI-BF and Trad, as being able to ask follow-up questions improved choices for those who asked them. Search behavior mirrored these choice patterns: most participants obtained full information under traditional search (by searching both brands), whereas many in the AI-Qs condition asked few or no follow-up questions, suggesting that asking questions may be more costly than clicking.

Furthermore, we analyze the language participants used to describe their search strategies. We find that mentions of the highlighted attribute rose sharply from the Trad condition to the AI conditions (and are most pronounced under AI-BF), indicating that emphasizing a single strong attribute steers consumers' attention toward it and away from the information needed to make the better choice.

The rest of the paper proceeds as follows. The next section describes prior work. Section 3 describes our scraping methodology and documents empirical regularities in how AI tools present product information. Section 4 develops the theoretical framework. Section 5 describes the experimental design and results. Section 6 concludes.

2 Related Literature

Our paper connects to several bodies of work. The classic literature on consumer search (Stigler, 1961; Weitzman, 1979) emphasizes the role of search costs in distorting choices. Relatedly, work on online price comparison websites and intermediaries shows that reducing search costs can lower prices (Brown and Goolsbee, 2002). More recently, Padilla et al. (2025) find that LLM adoption crowds out other online search activity, consistent with the idea that LLMs make information processing more efficient for the consumer. We contribute to this literature by showing that AI search tools, beyond reducing search costs (Shahidi et al., 2025), actively restructure the information presented to consumers.

We also contribute to a growing literature on the economic consequences of large language models. Closest to our paper is prior work that directly compares search and purchase decisions

between traditional search engines versus AI tools (Spatharioti et al., 2025; Xu et al., 2023). They find that, for information lookup tasks (like "what is the cargo space to total length ratio of <vehicle>" or "list five websites that can be used for booking a flight between two cities"), the use of AI makes information retrieval faster and more accurate, but at the same time may lead to over-reliance, even when AI responses contain errors. We contribute to this work in three ways: by studying how search engines versus AI compare on shopping recommendation tasks (that may not display a single correct answer), by focusing on the *format* of AI information rather than its accuracy, and by designing an experiment to causally identify the effects of this format.

Other work, on the importance of chatbots (Yan et al., 2025; Lin et al., 2025) and generative search (Zheng et al., 2025; Agarwal and Sen, 2026), as well as papers on the ability of LLMs to simulate human preferences (Argyle et al., 2023; Horton et al., 2023; Brand et al., 2023; Dillion et al., 2023; Goli and Singh, 2024; Gui and Toubia, 2023; Arora et al., 2025; Zhang et al., 2025), make choices on a person's behalf (Delaprez and Hortacsu, 2026; Allouah et al., 2025; Brucks and Toubia, 2023), or shift preferences (Werner et al., 2024), is thematically related, though our focus differs substantially: our paper studies instead how the format of AI responses affects consumer choices.

There is significant work on identifying the factors that lead consumers to over-rely on AI responses even when these are wrong (for a review, see Passi et al. (2024)). Also, a lot of work proposes solutions to such over-reliance, such as forcing users to think about the right answer first (Buçinca et al., 2021), requiring the AI to highlight information it is uncertain about (Vasconcelos et al., 2025), or communicating uncertainty in percentages or words (Lin et al., 2022; Spatharioti et al., 2025). Given these issues, the topic of trust in AI responses has also been investigated. Prior work shows that consumers' trust in AI is highly malleable and responds to easily manipulated cues, such as citations/reference links, even when hallucinated or broken (Li and Aral, 2025), and number formatting (e.g. 74.99% versus 75%) (Kim et al., 2021). Instead, we focus on a specific aspect of the way AI formats responses (the best-for format), assuming improvements in technology will address issues of trust, and study the effect of AI response format on consumer choices.

Our work also contributes to literature that shows that selectively highlighting a favorable attribute can reshape consumer inference and reduce attention to other relevant product information (Keller et al., 1997; Andrews et al., 1998, 2000; Roe et al., 1999; Rao and Ursu, 2025). We extend this

literature to the AI domain to understand if a similar logic applies in AI search, where generative systems and recommendation interfaces do not simply present neutral information but instead curate, summarize, and prioritize outputs, functioning as a selective-highlighting technology.

3 Scraped Data

We scraped data for this project in two steps. First, we used several methods to determine common AI shopping-related prompt types. Second, we scraped AI responses to these prompts from two different AI tools.

3.1 Step 1: What Are Common Shopping-Related AI Prompt Types?

To date, no major AI platform publicly releases logs of individual shopping prompts users actually submit. We therefore used a four step process to identify a set of common shopping-related prompts. Below we summarize our approach and findings, with complete details appearing in Online Appendix OA.1.

First, we directly and repeatedly asked both ChatGPT and AI Mode for the most common types of shopping related prompts (prompt "what are the top shopping-related queries on chatgpt/AI Mode?"). Second, we used industry reports for additional information on prompts related to shopping. These two methods identified eight main prompt types: best/top products, product comparison, product recommendation/'help me' advice, gift shopping, budget/deals, lifestyle matching, summarize reviews, and product availability and inventory.

Third, we designed a Prolific survey asking participants to phrase an AI shopping-related prompt in a given category (prebiotic sodas). Specifically, the instructions asked "Imagine you are looking to buy a prebiotic soda and are going to use ChatGPT or your preferred AI to help you. We want you to type the question you would ask an AI chatbot to help you decide which prebiotic soda to purchase." We also asked participants about: the frequency of their AI use; the frequency of purchasing prebiotic soda; whether they mentioned all product attributes they cared about in their written prompt; how many brands they expected to see in the AI generated response; the type of answer they expected (best/top products, product comparison, general product information, other); their expected need for follow-up questions; as well as demographic information (age, gender,

income, education). Our results show that most participants asked about best/top products (47%), followed by product recommendations (23% recommendations, 3.5% 'help me' advice), product comparison (7%), budget/deals (7%), product availability and inventory (4%) and reviews (2%).

Finally, we used a large publicly available data set (WildChat) collected by Zhao et al. (2024) of approximately 1 million real ChatGPT conversations. Most conversations in these data are not shopping-related; for example, the largest category is "assisting/creative writing", with 62% of observations. Using keyword search, we identified 123 conversations that were shopping-related. Analyzing these conversations we found that most prompts asked about best/top products (42%), followed by budget/deals (19%), product comparison (11%), and product recommendations (8%).

Table 1: AI Prompts Used in Data Collection

Product Category	Prompt Type	Specific Prompt	Group
Prebiotic soda	Best/top products	What is the best prebiotic soda?	prebiotic
	Product recommendations/ 'Help me' advice	Can you recommend a prebiotic soda?	
	Product recommendations/ 'Help me' advice	Help me choose a prebiotic soda.	
	Product comparison	Can you compare different prebiotic sodas?	
Wireless Headphones	Best/top products	What are the best wireless headphones?	headphones
	Product recommendations/ 'Help me' advice & Budget/deal & Lifestyle matching	Recommend affordable wireless headphones for a beginner runner.	
Winter Jacket	Product recommendations/ 'Help me' advice & Lifestyle matching	I am a woman who lives in a city and I am always cold. Recommend winter jackets.	jacket
NYC Holiday Hotels	Product availability and inventory	List hotels in New York City available this holiday season.	hotels
KitchenAid Mixer	Summarize reviews	Summarize the reviews of the KitchenAid stand mixer.	mixer
Phones	Product comparison	Compare the iPhone 16 Pro Max and the Google Pixel 9 Pro.	iPhone compare
	Budget/deals	What is the best price I can get on the new iPhone?	iPhone price
Gifts	Product recommendations/ 'Help me' advice & Gift shopping & Budget/deals	Help me find a great housewarming gift under $100.	gifts

The full list of prompt types we converged on and the specific prompts we scraped appear in Table 1. We include queries spanning seven categories (prebiotic sodas, wireless headphones, winter

jackets, hotels, smartphones, mixers, gifts) and all the prompt types we identified as being common (via the methods described earlier). We generated specific prompts to cover the space of all types of prompts we identified. The last column groups the prompts by a short keyword which allows for quick reference to each prompt. We note that individuals may differ in their ability/cost of writing a "good" prompt, which may affect the responses they obtain, a dimension (investigating heterogeneity across individuals) orthogonal to our research question (Misra, 2025). Here we keep prompts constant to identify the format of AI responses to common shopping-related prompts.

3.2 Step 2: What Do AI Responses to Common Shopping-Related Prompt Types Look Like?

For each of the prompts in Table 1, we scraped responses from two AI search platforms: ChatGPT (OpenAI's GPT 5.1) and Google AI Mode.[5] Data collection proceeded at high frequency over an approximately 90-day window spanning November 2025 through March 2026. For AI Mode we scraped data from November 19th, 2025 to February 13, 2026, every 30 minutes in the interval 9am to 5:30pm EST, obtaining approximately 1,445 observations for each of the 12 prompts.[6] For ChatGPT, we scraped data from December 6th, 2025 to March 5th, 2026, hourly in the same time interval.[7] We used a temperature parameter of zero following prior work (Brucks and Toubia, 2023).[8] We obtained approximately 860 observations for each of the 12 prompts for ChatGPT. In total, we have 27,589 observations, where an observation is an AI tool's response to a specific prompt at a given time.

3.3 Summary Statistics

Table 2 presents summary statistics for both ChatGPT and Google AI Mode responses. ChatGPT responses are generally longer in terms of both total words and characters. This happens because

[5]Google AI Mode data is collected via SerpAPI (serpapi.com). ChatGPT data is collected via OpenAI's Responses API (not the Chat Completions API) with the web search preview tool enabled, temperature 0, and model 5.1 (model 4.1 mini before December 8th, 2025).

[6]We started data collection earlier for AI Mode than for ChatGPT. Also, our initial prebiotic soda prompts asked about "gut-healthy sodas", which we later realized included both probiotic and prebiotic sodas. To keep the analysis clean, we later focused only on prebiotic sodas and changed the prompts, which lead to slightly fewer scraping counts for these prompts (30-40 fewer observations).

[7]Before December 16th, we scraped ChatGPT every 30 minutes, but realized the responses did not vary as much as those for AI Mode did, so we decreased the frequency.

[8]OpenAI does not specify what temperature is chosen when producing a certain response; we expect that when actual users interact with ChatGPT they may see more variation in responses than in our data.

ChatGPT often mentions more brands and features than AI Mode. Given that each prompt group has a different number of brands and features, we report the percentage of all brands and features mentioned across all our observations in a group.[9] In general, only a small percentage of the brands available ($26 - 29\%$) are mentioned in a given response.[10] While $80 - 85\%$ of product features are mentioned in an average AI response (equivalent to 6 of 7 features), as we show later, these features are not mentioned for each brand; instead one feature may be highlighted for one brand, while another feature for another brand.

Table 2: Summary Statistics Across Prompts and AI Tools.

	ChatGPT	AI Mode
Observations	10,353	17,236
Average (median) number of words	523 (481)	364 (364)
Average (median) number of characters	3,910 (3,650)	2,335 (2,344)
Percent of product features	0.8473	0.8048
Percent of brands	0.2902	0.2607

Notes: We report the percent of all features/brands mentioned across our observations in a prompt group. The percent of product features excludes prompts with no feature groups (gifts, iPhone price). The percent of brands excludes single-brand prompts (mixer, iPhone price).

3.4 Main Observation: Best-For Information is Highlighted

The most striking regularity in the data is the prevalence of what we term *best-for* language—phrases that describe a product by its best features rather than listing all raw attributes of a product. We show the prevalence of this framing in three ways. First, for every AI response, we check whether the specific phrases "best for", "best if", or "best overall" appear in the text. Second, we check every response with a longer list of 61 similar best-for phrases, such as "great/ideal/top/recommended for," "if you want/like/care/enjoy…," "for those with/who…," "what matters," and "for anyone who…".[11] Finally, we create a variable called "feature parity" which assesses whether brands are

[9] For a full list of the brands and their features in each prompt group, see Table OA.2 in Appendix OA.3.
[10] Figure OA.4 in Online Appendix OA.3 shows which brands are mentioned by each AI tool for every prompt.
[11] The full list appears in Table OA.3 in the Online Appendix OA.4.

discussed in terms of the same or different features in an AI response. To construct this variable, we first find the character position of every brand mentioned in the text. Then, we segment the AI response by brand "ownership": each brand mention "owns" the slice of text from its own position up to the next brand mention (the last brand owns the text to the end). Within each owned slice, we check which product features appear; those features are attributed to that brand. In order to avoid miscounting, we group keywords that relate to the same feature (e.g., cost, budget, price, etc) into one single *feature group* variable (e.g., price).[12] Finally, the feature parity variable equals 1 if a given feature group is either mentioned for all brands discussed in an AI response or is mentioned for none. Low feature parity is consistent with best-for language.

Our results appear in Table 3. The rates at which phrases involving a best-for structure appear are high. For ChatGPT, the rates are generally over 99 percent for nearly all prompts. Google AI Mode also displays high rates, ranging from roughly 47 percent ("Summarize reviews of the KitchenAid stand mixer") to 99 percent ("What is the best prebiotic soda?"). These rates are lower for prompts that ask for a comparison (forcing the AI to consider more features) and for those focusing on one brand (again, mentioning more features of that one brand). The pattern, however, appears across all prompt types with AI responses consistently framing product information in terms of what features each product is best for.

The feature parity variable provides further evidence of this framing. Less than 5% of responses discuss the same features for all brands mentioned. Therefore, even though 6 of 7 attributes are mentioned in an average response (see Table 2), these attributes are not mentioned for each brand; instead one attribute may be highlighted for one brand, while another attribute for another brand.

We further show this result in Figure 2 for the prebiotic soda prompts, and, given similar patterns, present all other prompts in Figures OA.5-OA.10 in Online Appendix OA.3.[13] The figure depicts the percent of brand mentions that also mention a certain feature group (larger percentages indicated by darker colors in the heatmap). It shows that there is substantial variation in feature coverage, such that brands are not discussed in terms of the same features in most responses. This variation is consistent with the best-for format. For example, Poppi has lower sugar and calories than Olipop, so these features are mentioned more frequently for Poppi than for Olipop.

[12]The full list of feature groups for every prompt group is available in Table OA.2 in Online Appendix OA.3.

[13]Note that the frequencies with which brands appear in Figure 2 may differ from those in Figure OA.4 because in some cases brands are mentioned without a feature group).

Table 3: Best-for Response Characteristics by Prompt and AI Tool.

ChatGPT

Prompt	N	Best-for 61 phrases	Best-for 3 phrases	Feature parity
What is the best prebiotic soda?	864	0.998	0.900	0.003
Help me choose a prebiotic soda	862	0.956	0.204	0.062
Can you recommend a prebiotic soda?	862	0.991	0.176	0.003
Can you compare different prebiotic sodas?	861	0.995	0.898	0.033
What are the best wireless headphones?	864	0.981	0.723	0.043
Recommend affordable headphones for a beginner runner	864	0.998	0.760	0.036
Winter jacket for a city woman who is always cold	863	0.998	0.436	0.020
iPhone 16 Pro Max vs. Google Pixel 9 Pro	862	0.988	0.595	0.437
List hotels in New York City	864	0.894	0.059	0.034
Summarize reviews of the KitchenAid stand mixer	862	0.991	0.616	–
Housewarming gift for a friend under $100	863	0.981	0.133	–
What is the best price on the new iPhone?	862	0.965	0.032	–

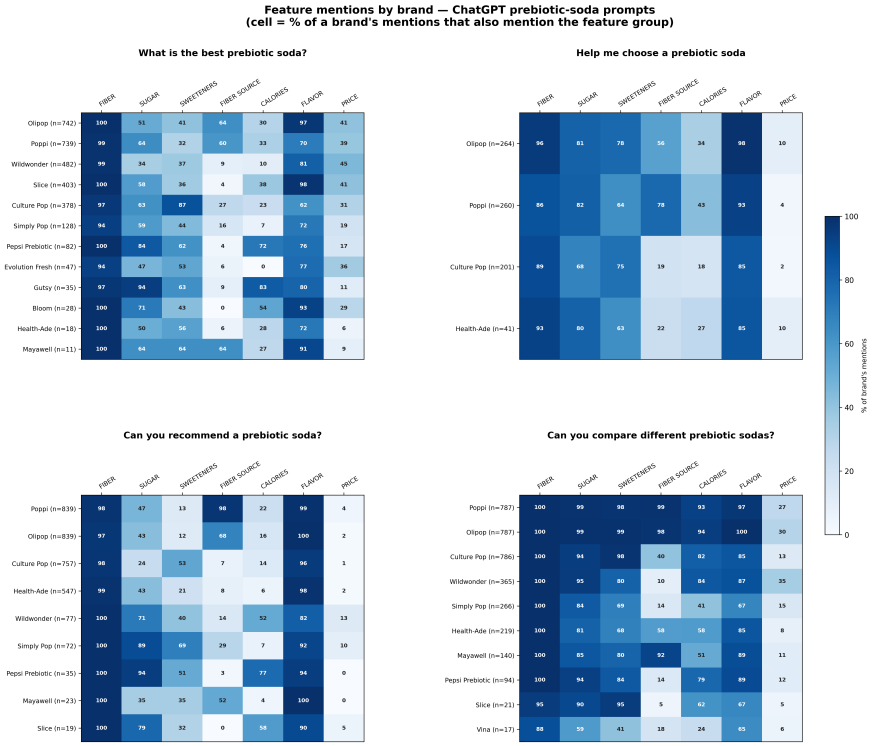
AI Mode

Prompt	N	Best-for 61 phrases	Best-for 3 phrases	Feature parity
What is the best prebiotic soda?	1430	0.989	0.610	0.023
Help me choose a prebiotic soda	1417	0.986	0.402	0.005
Can you recommend a prebiotic soda?	1415	0.954	0.618	0.013
Can you compare different prebiotic sodas?	1404	0.910	0.199	0.001
What are the best wireless headphones?	1444	0.935	0.625	0.009
Recommend affordable headphones for a beginner runner	1446	0.796	0.064	0.192
Winter jacket for a city woman who is always cold	1446	0.949	0.479	0.284
iPhone 16 Pro Max vs. Google Pixel 9 Pro	1446	0.767	0.006	0.057
List hotels in New York City	1445	0.972	0.007	0.226
Summarize reviews of the KitchenAid stand mixer	1450	0.466	0.036	–
Housewarming gift for a friend under $100	1445	0.814	0.075	–
What is the best price on the new iPhone?	1448	0.946	0.087	–

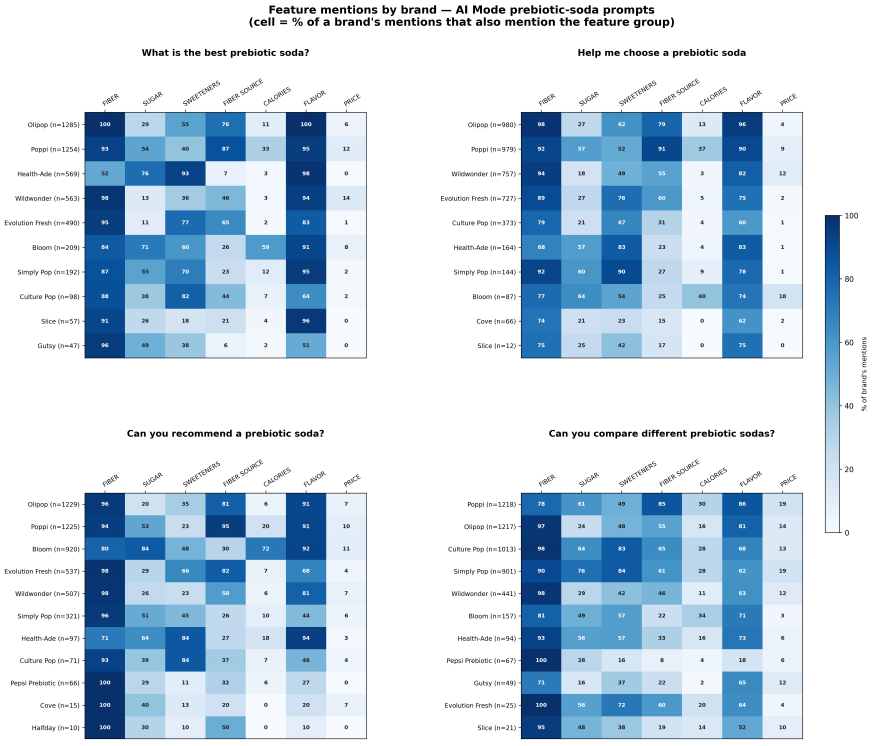
Notes: *Best-for phrases* are means over all responses. Best-for 3 phrases are "best for", "best if" or "best overall". Best-for 61 phrases contains a longer list of 61 phrases such as "best for," "great/ideal/top/recommended for," "if you want/like/care/enjoy…," "for those with/who…," "what matters," and "for anyone who…" (full list appears in Table OA.3 in Appendix OA.4). *Feature parity* is averaged only over responses that mention at least one brand; prompts with no feature groups or with only one brand are omitted ("–").

The same is true for brands that are infrequently mentioned. For example, Gutsy has no added sugar and low calories and these features are mentioned frequently whenever the brand is discussed. Although brands are not always compared in terms of their best features (e.g., Slice only has 6 flavors relative to Olipop's 19, but flavor is mentioned frequently; Canada Goose's high price is frequently mentioned), we generally find, consistent with the best-for format, that the features that stand out for a given brand are more frequently mentioned, with other features left implicit.

The "why" behind the best-for pattern reflects how LLMs are built and trained. First, the



(a) ChatGPT



(b) AI Mode

Figure 2: Percent of Brand Mentions that also Mention a Feature Group

sources that AI systems retrieve information from and synthesize frequently employ best-for categorization. We can show this in our data by extracting the links that are used as citations within each AI response. To do so we summarized the top 10 domains for each prompt group and categorized them into four main groups for each AI tool: review sites/buying guides, social media/UGC sites, e-commerce sites, other.[14] Figure OA.3 in Online Appendix OA.3 shows our results. We find that a majority of citing sources comes from review sites or websites that provide buying guides, such as Wired, Business Insider, Eat This Not That, CNET, for both AI tools.[15] These websites already present product rankings in "best for" categories, and this structure propagates into the syntheses AI tools provide. Second, because reinforcement learning from human feedback optimizes models for human preferences, LLMs often adopt highly organized presentation styles (when human raters prefer them) — including bullet points, headings, and categorized lists — that improve perceived readability (Ouyang et al., 2022; Bai et al., 2022; Zhang et al., 2025). Rather than the best-for format being an arbitrary stylistic choice, it is instead a consistent property of how LLMs (to date) are trained and deployed. For these reasons, although AI models change frequently, this presentation style may be more long-lasting.

We also note that not all prompts receive this type of best-for response. In the case of a close-ended prompt that has one correct answer, such as "which soda has the most fiber", or "what is the best prebiotic soda with at least 8g fiber, under 3g sugar and that is cheap", best-for responses are less likely to occur. In other words, best-for responses appear when there is a trade-off involved, with multiple options potentially fitting the prompt, where the AI needs to create decision criteria.

Nevertheless, for the types of common shopping-related queries we have analyzed (that are generally open-ended, with no universal optimum), "best-for" responses are common. We further highlight this by analyzing AI responses to the prompts submitted by individuals in our Prolific survey who were tasked to write a prompt for an AI search tool given interest in buying a prebiotic soda.[16] We used the same method as above for scraping AI responses (once, on June 9th, 2026) using both ChatGPT (OpenAI's GPT 5.1) and Google AI Mode for all the prompts collected in

[14] We hand coded each domain (approximately 60 per AI tool, with overlaps) by visiting each website to determine which category fits best. Websites such as Yahoo, The Independent, The Verge, Washington Post, and others contain lifestyle sections that have best-for guides and exactly such pages were cited by AI tools in our data.

[15] There is a smaller percentage for AI Mode because it relies more heavily on Google Shopping results.

[16] We thank Sanjog Misra for the suggestion. A full description of this survey and our methodology is available in Online Appendix OA.1.

the Prolific survey. In the Prolific survey participants asked a variety of prompts and across all of them we again observe a prevalence of best-for responses, as well as a low feature parity level (see Table 4 below). Also, we asked participants how many brands they expected to be mentioned in the AI response. On average (median), participants expected 3.68 (3) brands to be mentioned, with 98% of participants expecting at least 2 brands (so not only one brand).[17] About half of all participants did not expect the need to ask any follow-up questions beyond the first prompt (very similar to the pattern observed in the WildChat data), even though 74% report not asking about all the attributes they cared about in their initial prompt. All these patterns are consistent with responses to AI shopping-related prompts often having a best-for format.[18]

Table 4: Best-for Response Characteristics for the Prolific Survey

		Best-for		Feature
AI Tool	N	61 phrases	3 phrases	parity
ChatGPT	97	0.979	0.320	0.186
AI Mode	97	0.948	0.330	0.103

Using the observations made in this section, we now turn to developing a search model that will trace out the implications of the best-for format for consumer choices.

4 Theoretical Framework

4.1 Setup

Suppose there is one consumer and there are two products available, $j \in \{1, 2\}$, each with two attributes, (x_j, y_j). The consumer has preferences (α, β) over these attributes. Given symmetry, we assume the consumer cares more about attribute x than about y, such that $\alpha > \beta > 0$, and that product 1 has the higher x attribute ($x_1 = H$) and product 2 has the higher y attribute ($y_2 = H$). Prior to any search, the consumer believes each attribute independently takes values in $\{L, M, H\}$ with equal probability, where $0 < L < M < H$.

[17] This statistic assumes those participants who answered "More than 6" expected 7 brands; ignoring these entries the average (median) is 3.47 (3).

[18] More than half (53%) of the Prolific prompts match in intent one of the four main prebiotic prompts we scraped in our main analysis.

The consumer's utility from product j when both attributes are observed is

$$u_j = \alpha x_j + \beta y_j. \tag{1}$$

To proceed, we assume that the value of L is small enough (relative to H and M), so that hiding an attribute with level L can have an effect on consumer choices:

Assumption 1. Let $\frac{\alpha}{\beta} < \frac{H-L}{H-M}$.

If Assumption 1 does not hold, then in all cases below, the AI search tool will dominate traditional search and lead to the same choice as under full information.

The consumer does not know the values of x and y for either of the two products, but can interact with AI or conduct traditional sequential search on an e-retailer's website before deciding.[19] Additionally, we will also compare her choices with those under full information.

4.2 AI Search

The consumer interacts with an AI search tool before making a purchase decision between the two products. The AI has full information on the value of the two attributes for both products, does not hallucinate or try to mislead the consumer, and does not observe the consumer's preferences, α and β. The consumer types in a prompt (for free, consistent with prior work on search that assumes the first query is free, (Kim et al., 2010; Chen and Yao, 2017; Ursu et al., 2025)), sees AI-generated product information, and decides which of the two products to buy. In an extension (Section 4.5 below), we also allow the consumer to ask the AI follow-up questions before deciding.

The AI adopts a best-for (BF) policy: it presents each product as "best for" its highest attribute, so the AI presents product 1 as "best for x" and product 2 as "best for y." The remaining attributes, y_1 and x_2, are left implicit.

Note that even though u_1 and u_2 are independent, the BF policy reveals information beyond the two attributes revealed: it reveals that y_1 and x_2 can only be L or M. In the absence of other information, the consumer sets her beliefs about the missing attributes equal to the mean, $\mu = (L+M)/2$.

[19] We assume that even the smallest utility is sufficiently high such that the consumer prefers to buy one of the two products, rather than choose the outside option (with utility normalized to zero). This assumption is similar to that made in prior work (Anderson and Renault, 1999; Haan et al., 2018) and it does not affect the substantive insights from our model, but simplifies the presentation of our results.

Formally, the choice the consumer has after observing the AI response is between (in terms of expected utilities, $\hat{u}_j$)[20]

$$\begin{aligned} \hat{u}_1 &= \alpha H + \beta \mu, \\ \hat{u}_2 &= \alpha \mu + \beta H. \end{aligned} \tag{2}$$

Given that $\alpha > \beta$, it follows that $\hat{u}_1 > \hat{u}_2$, so the consumer prefers product 1 over product 2 after seeing the AI response, obtaining a final expected utility of $\hat{u}_1$. For simplicity, we will refer to this outcome as the "the AI-guided choice" and compare it next to the consumer's choice under full information.

4.3 Comparison to Full Information

The consumer observes (x_j, y_j) for both products and chooses the product with higher utility. Since $x_1 = H$, $y_2 = H$, and $(y_1, x_2) \in \{L, M\}^2$, there are four cases to consider (see Table 5):

Table 5: Full-Information vs. AI-Guided Choices

Case	y_1	x_2	$u_1 = \alpha H + \beta y_1$	$u_2 = \alpha x_2 + \beta H$	Full Info. Choice	Match AI-Guided Choice?
(1)	L	L	$\alpha H + \beta L$	$\alpha L + \beta H$	1	✓
(2)	L	M	$\alpha H + \beta L$	$\alpha M + \beta H$	2	✗
(3)	M	L	$\alpha H + \beta M$	$\alpha L + \beta H$	1	✓
(4)	M	M	$\alpha H + \beta M$	$\alpha M + \beta H$	1	✓

The AI-guided choice matches the full-information choice in three of the four cases. It fails in case (2), where $y_1 = L$ and $x_2 = M$: product 2 is actually better under full information (because $\alpha M + \beta H > \alpha H + \beta L$), but the AI response hides product 1's poor $y_1 = L$ attribute and product 2's reasonable $x_2 = M$ attribute, steering choices towards the inferior product (given consumer preferences).

Next we make a few observations about the characteristics of the AI-BF policy:

[20]Most AI tools to date do not allow the consumer to purchase within the platform, but may direct them to the retailer's website for purchase. We assume a large enough cost such that navigating back to the AI tool is prohibitive for most consumers. Even in the absence of this assumption, as long as this cost is higher than the click cost of the traditional search model introduced below (which is likely given the consumer needs to switch between platforms rather than only click a button on a single platform), qualitatively our results will continue to hold.

(1) AI-BF policy provides at least partial information on all attributes of a product, which facilitates comparison, even though utilities are independent. One important benefit for the BF policy is that it narrows down the range of possible values on attributes that are hidden given that they are compared against the highest value of the attribute of a certain brand. This feature makes it easier to compare products and make a choice.

(2) AI-BF may lead to an inferior choice, if the hidden attribute on what seems to be the superior product is poor (relative to expected utilities).

(3) AI-BF may affect the consumer's product valuation. AI may affect the consumer's valuation of the product (e.g., her willingness to pay or her expected utility), since choices are based on expected utilities ($\hat{u}_1 = \alpha H + \beta\mu$ or $\hat{u}_2 = \alpha\mu + \beta H$), which may be higher or lower than the true utilities.

(4) AI-BF helps even when the AI does not observe consumer preferences. Our analysis was performed under the assumption that the AI does not know the consumer's preferences, α and β, and that $\alpha > \beta > 0$. Now suppose the AI continues to not observe α and β, but that $\beta > \alpha > 0$. The consumer would now choose product 2 when searching using AI and would make the same choice under full information (except in case 3 under an equivalent condition to Assumption 1; otherwise, she would also choose product 2 in case 3). In other words, the BF policy also has the advantage of aiding the AI in providing the consumer with enough information to choose the objectively higher utility product in many cases even though it does not observe preferences.

4.4 Traditional Sequential Search

Suppose the consumer interacts with an e-retailer rather than AI before purchasing products 1 or 2. On the e-retailer's website, the consumer types in a query (for free, as was the case with the prompt in AI), and is presented with a list of the two products. Then, the consumer can pay cost $c > 0$ to reveal any remaining uncertainty before making a purchase decision.[21]

There are three settings to consider in modeling what information is available on the list page: (1) the list page contains information only on x and the consumer needs to pay c to reveal y, (2) the list page contains information only on y and the consumer needs to pay c to reveal x, and (3)

[21] We abstract away from cases where the consumer can add a product to the cart directly from the list page (consistent with prior work (Kim et al., 2010; Chen and Yao, 2017; Honka and Chintagunta, 2017; Ursu, 2018; Ursu et al., 2025)), since in such cases consumers are likely to not need to search.

the list page contains no information on either x or y and the consumer needs to pay c to reveal both attributes. We abstract away from ranking effects (to keep the setup consistent with the AI setup) and we assume the cost to search is the same regardless of how many attributes the consumer reveals (although our results will not change if we allow search costs to vary). Finally, we note that utilities are independent, so searching one product does not reveal the utility of another.

We analyze the consumer's search problem using the Weitzman (1979) framework. Search is sequential and there are two possible search orders: (1) the consumer searches product 1 first (occurs with probability 1 if the list page includes information only on attribute x for both products[22] and with probability 1/2 if the list page does not provide any relevant information for her search order) or (2) the consumer searches product 2 first (occurs if the list page contains information only on y[23] or with probability 1/2 if it does not provide any relevant information for search order). We analyze both search orders below. Depending on the values of the product attributes, we again have four cases to analyze as in Table 5.

Search product 1 first If attribute x is available on the list page, then in cases (1), (3), and (4), the consumer's continuation value from searching product 2 is dominated by the utility already realized from product 1.[24] The consumer therefore only searches one product and chooses product 1, the same choice as under AI.

In case (2), the consumer will search product 1 first and reveal $u_1 = \alpha H + \beta L$. She expects product 2's utility to equal $\alpha M + \beta y_2$, with y_2 taking on values of $\{L, M, H\}$ with equal probability. If the consumer would choose product 1 if she were to reveal $y_2 = M$, then the search condition for product 2 is[25]

$$\begin{aligned} u_1 &< -c + E(\max(u_1, u_2)) \qquad (3) \\ \alpha H + \beta L &< -c + \frac{2}{3}(\alpha H + \beta L) + \frac{1}{3}(\alpha M + \beta H) \\ c &< \frac{\beta(H-L) - \alpha(H-M)}{3}. \end{aligned}$$

[22] To see this, note that the revealed attributes are $x_1 = H$ and $x_2 < H$ and that the consumer expects the hidden attribute y for both products to be $\{L, M, H\}$ with equal probability.

[23] To see this, note that the revealed attributes are $y_1 < H$ and $y_2 = H$ and that the consumer expects the hidden attribute x for both products to be $\{L, M, H\}$ with equal probability.

[24] For example, in case (1), after searching product 1 and revealing $u_1 = \alpha H + \beta L$, she expects product 2's utility to equal $\alpha L + \beta y_2$, with $y_2 \in \{L, M, H\}$. However, even if she were to reveal $y_2 = H$, she would still prefer to choose product 1, so she optimally stops searching after inspecting product 1.

[25] This requires that $\frac{\alpha}{\beta} > \frac{M-L}{H-M}$.

If instead the consumer would choose product 2 if she were to reveal $y_2 = M$, then she will search whenever $c < \frac{\beta(H+M-2L)-\alpha(2H-2M)}{3}$.

With low enough search costs, the consumer searches product 2, discovers $u_2 = \alpha M + \beta H > u_1$, and chooses the superior product (product 2) relative to AI. However, if search costs are large, then the consumer will not search product 2 and will choose product 1, as under AI search.

If instead neither attribute is available on the list page, then the consumer would make similar choices as above, but her search conditions would adjust. For example, in cases (1) and (2), the search condition is given by $c < \frac{\beta(2H+M-3L)-\alpha(H-M)}{9}$, while in cases (3) and (4), the search condition is given by $c < \frac{\beta(H-M)}{9}$.[26] In all cases except (2), the consumer chooses product 1, while in case (2), if she searches she will choose product 2, but she will stick with product 1 if she does not search. Thus, search helps the consumer make a better choice than under the AI regime in case (2) when search costs are low.

The comparison across regimes is summarized in Table 6.

Table 6: Traditional Search vs AI-Guided vs. Full Information Choices: Search Product 1 First

Case	y_1	x_2	Trad. Search Choice	AI-Guided Choice	Full Info. Choice
(1)	L	L	1	1	1
(2)	L	M	1 (or 2 if c small)	1	2
(3)	M	L	1	1	1
(4)	M	M	1	1	1

Search product 2 first If search starts with product 2, regardless of whether this occurs because attribute y is available on the list page or neither x or y are available on the list page, in cases (1), (3), and (4), if the consumer searches (if search costs are low enough), then the she will choose the superior product 1, the better choice compared to the AI-guided choice.[27] Not searching means making the inferior choice, product 2. In case (2), the consumer will choose product 2 regardless of search, choice that coincides with the full information choice and that dominates the AI-guided choice. Table 7 summarizes our results.

Overall, we find that in three of four cases, the AI-guided choice coincides with the choice made

[26]Full derivations of these results appear in Appendix A.
[27]Full derivations of these results appear in Appendix A.

Table 7: Traditional Search vs AI-Guided vs. Full Information Choices: Search Product 2 First

Case	y_1	x_2	Trad. Search Choice	AI-Guided Choice	Full Info. Choice
(1)	L	L	2 (or 1 if c small)	1	1
(2)	L	M	2	1	2
(3)	M	L	2 (or 1 if c small)	1	1
(4)	M	M	2 (or 1 if c small)	1	1

under full information. In case (2), AI can lead the consumer to choose an inferior product, but in traditional search this outcome is averted if search costs are sufficiently low (or if the consumer happens to start search with the superior product). In most cases, using AI has the additional benefit of protecting consumers with high search costs from choosing an inferior product, which they would have chosen in traditional search if searching the inferior product first.

4.5 AI with Follow-up Questions

We now consider the case where after seeing the AI-BF response the consumer can ask follow-up questions—akin to how most AI tools work—and refer to this regime as AI-Qs. In reality, with two products, the consumer could ask questions about either one or about both in one prompt. To simplify the analysis and to be consistent with the majority of choices participants make in our experiment (next section), we focus on the setting where with one follow-up question, costing $q > 0$, the consumer can reveal the true utility of both products.

Formally, after observing the expected utility of the preferred product based on the AI response, $\hat{u}_1 = \alpha H + \beta\mu$, the consumer can either choose product 1, or pay follow-up cost q and then be able to choose the maximum utility among the realized values (u_1, u_2). So the consumer will ask a follow-up question if

$$\begin{aligned} \hat{u}_1 &< -q + E(\max(u_1, u_2)), \\ \hat{u}_1 &< -q + \frac{\alpha H + \beta L}{4} + \frac{\alpha M + \beta H}{4} + \frac{\alpha H + \beta M}{2}, \\ q &< \frac{\beta(H - L) - \alpha(H - M)}{4}. \end{aligned} \tag{4}$$

Asking a follow-up question allows the consumer to choose the superior product in the AI-Qs regime, just as clicking allowed the consumer to choose the superior product in traditional search.

The magnitude of q relative to c (an empirical question we test in our experiment) will determine whether the consumer is more or less likely to ask follow-up questions of AI than make product clicks in traditional search.

4.6 Testable Implications

The theoretical analysis yields the following testable predictions.

1. **Choice quality:** When the preferred product's non-highlighted attribute is low (case 2) and search costs are small, AI-BF leads to a strictly inferior product choice relative to traditional search and full information.

2. **Search costs:** Depending on the relative value of q and c, AI-Qs may or may not save consumers the cost of searching.

3. **Follow-up mitigation:** Allowing follow-up questions in the AI-Qs regime may partially restore the information consumers would have obtained through traditional search, attenuating the choice error in case (2). However, this depends on the relative value of q versus c.

5 Experiment

5.1 Design

We conducted an incentivized pre-registered online experiment on Prolific to test the model's predictions.[28] Specifically, we wished to test the predictions of our model in case 2, where each brand excels on one attribute, but the hidden attribute of the brand consumers prefer based on the known information is low. The experiment focuses on the prebiotic soda category, a relatively new product category necessitating search. In order to keep the experiment simple, we focused on the two leading brands: Olipop and Poppi. Each brand has three flavors, a price, and varies in terms of fiber and calories. In reality, Olipop has 9 grams of fiber and 35–50 calories (depending on the flavor), while Poppi has 2-3 grams of fiber and 20–30 calories (depending on the flavor), with prices in the $\$2 - \$3$ range for both. In the experiment, we inform participants that they will choose among new products, that are alterations of existing products. The product attribute values in

[28] Pre-registered study under As Predicted #289491.

our experiment are shown in Table 8. These product attributes are unknown to participants at the beginning of the experiment, but they will have access to different search technologies (described below) to reveal (part of) this information.

Table 8: Product Attributes in the Experiment

Product	Fiber (g)	Calories
Olipop	**10**	45
Poppi	5	**15**

On Prolific, participants were told they were buying a can of prebiotic soda for a friend, that the friend cares more about fiber than calories, and that the friend's utility followed this function: $5 \times Fiber - Calories$. To incentivize participants to choose the product that provides their friend with the highest utility score, they were told they would receive a financial bonus (in addition to their base payment paid at a rate of $12 per hour) equal in cents to the utility score multiplied by 10. To a participant with full information, it is clear that choosing Olipop will earn them a bonus of 50 cents, while choosing Poppi will result in a $1.00 bonus. This bonus structure makes the "correct" choice (from the perspective of payoff maximization) clearly Poppi, allowing us to measure choice error directly. We presented the bonus information transparently and identically to all participants across the conditions. Participants were also free not to choose either of the two brands (for example if their search costs were large) and were instead able to buy water, which they were told contains no fiber and no calories (and thus earns them 0 bonus).

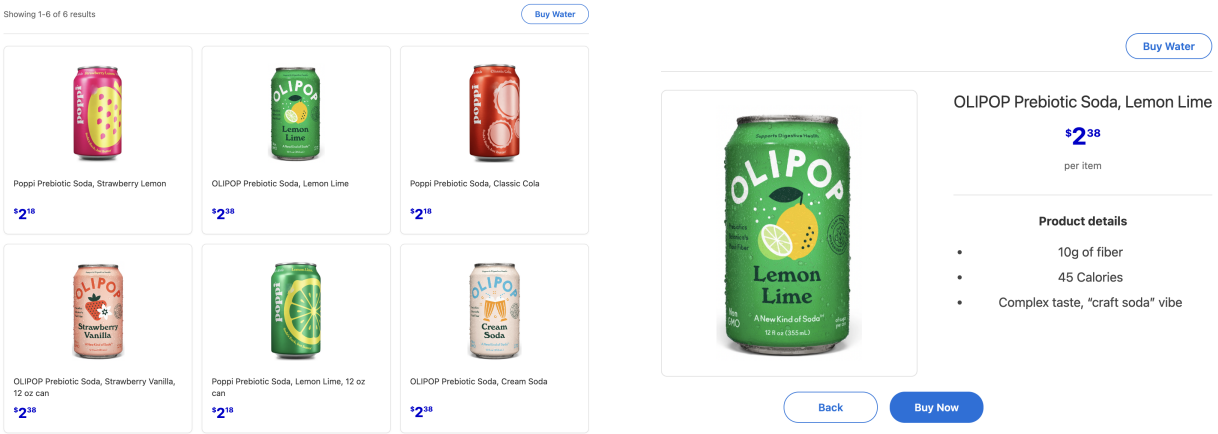
After receiving the above instructions, participants were given examples of the types of bonuses they could earn using different product choices (we used different attributes than in Table 8).[29] Participants also needed to pass five attention checks to be able to proceed to the main part of the experiment.

Participants were then randomly assigned to one of three conditions (search technologies): traditional e-retailer search (Trad), AI best-for search (AI-BF), and AI best-for search with follow-up questions (AI-Qs). The first and third conditions were meant to mimic realistic interactions be-

[29]For example, participants were shown two sodas: Soda A with 8g of fiber and 30 calories and Soda B with 3g of fiber and 10 calories. They were then walked through the equation (i.e., 5x Fiber - Calories) to show that choosing Soda A would score 10 points (and thus earn them a $1.00 bonus) and that Soda B scored 5 points (earning them a $.50 bonus).

tween participants and these search technologies. In contrast, the AI-BF condition allows us to disentangle the effect of the AI best-for format on consumer choices from that of information acquisition costs, following our theoretical framework. After random assignment, participants navigated to three different websites we designed:[30]

1. **Traditional E-Retailer Search (Trad):** Participants browsed a mock e-retailer's website that displayed information on brand names, flavor, price, and a product image on the list page. There were a total of six products available (two brands, Olipop and Poppi), each with three different flavors. Attribute information that affected payoffs (fiber and calories), as well as a note about taste was accessible only after clicking on individual product pages (equivalent to not seeing any payoff-relevant information on the list page in our model). Screenshots of these pages are available in Figure 3. Choosing a product to buy (one of the six options or water) is followed by a confirmation page which signaled the end of this portion of the experiment. Clicking on two unique brands on the website allowed participants to obtain full information on all relevant attributes in Table 8 (not on the same screen).



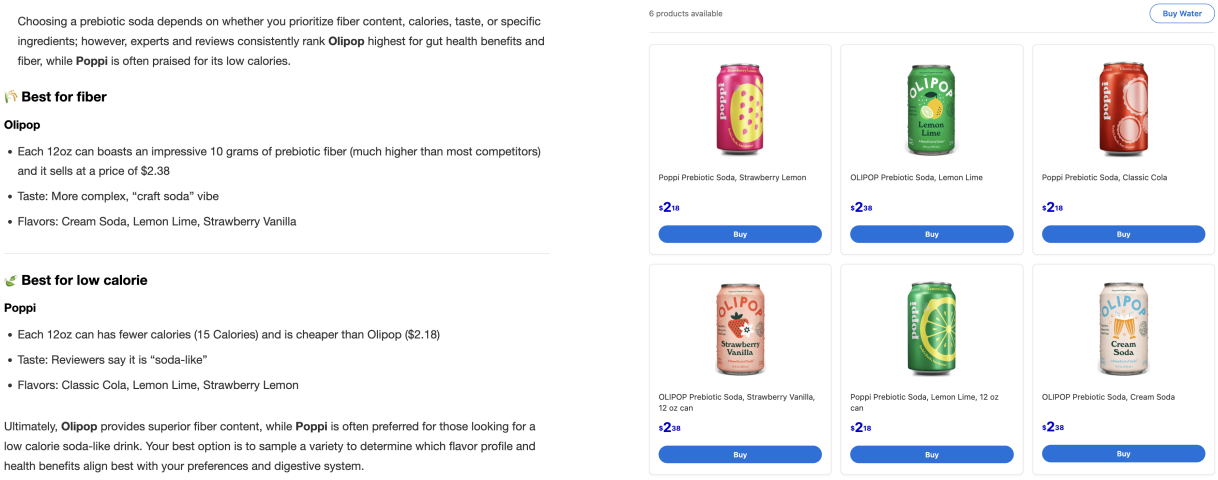
(a) List Page (b) Product Page

Figure 3: Traditional Search Website Layout

2. **AI Best-For Search (AI-BF):** Participants viewed a static AI-generated response that presented Olipop as "best for fiber" and Poppi as "best for calories". The response focused on what each brand was best for (Olipop has high fiber, 10 grams; Poppi has low calories, 15

[30]Other recent work building online/search environments includes Liu and Toubia (2018, 2025); Morozov and Tuchman (2024); Dang et al. (2026).

calories), as well as mentioning price, a note about taste (same as on the traditional search website), and listed the three available flavors for each brand. Participants were instructed that below the AI response was a list of the available products and that they will either need to buy one of the sodas or water. Following purchase, a confirmation page was displayed, after which this portion of the experiment concluded. Screenshots of the AI response and the list of options to buy are available in Figure 4.



(a) AI Response (b) Product Choice

Figure 4: AI-BF Website Layout

3. **AI Best-For Search with Follow-up Questions (AI-Qs):** Participants received the same initial AI-BF response, but could ask up to five follow-up questions before deciding which product to buy. Only two of the five follow-up questions were relevant to the task, as they directly asked about the hidden attributes of the two brands (i.e., fiber for Poppi and calories for Olipop). Responses to these follow-up questions were static and provided accurate attribute information. The follow-up questions appeared right under the AI response, as shown in Figure 5; participants could not type their own question. All possible follow-up questions and examples appear in Figure 6. Asking the appropriate follow-up questions allowed the participants to obtain full information on all attributes in Table 8. Choosing a product to buy was followed by a confirmation page which signaled the conclusion of this portion of the experiment. Choices were made from a list exactly like the one in AI-BF (right hand side of Figure 4), available under the follow-up question section.

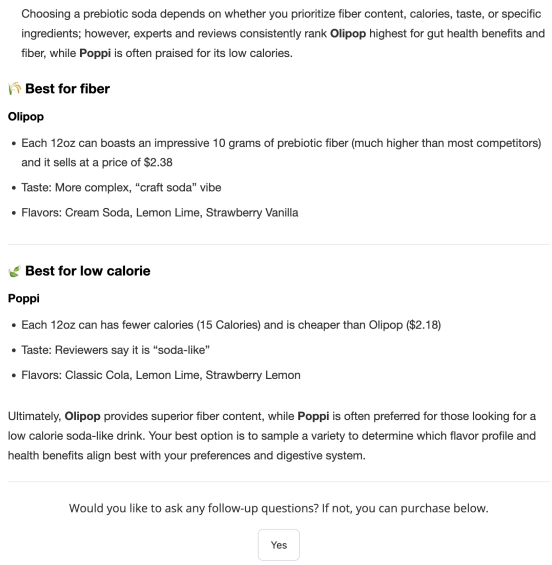
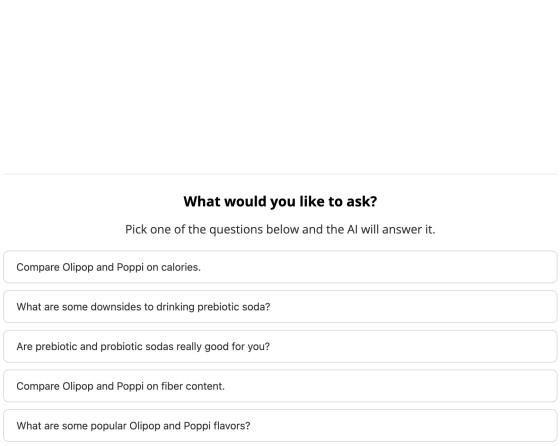
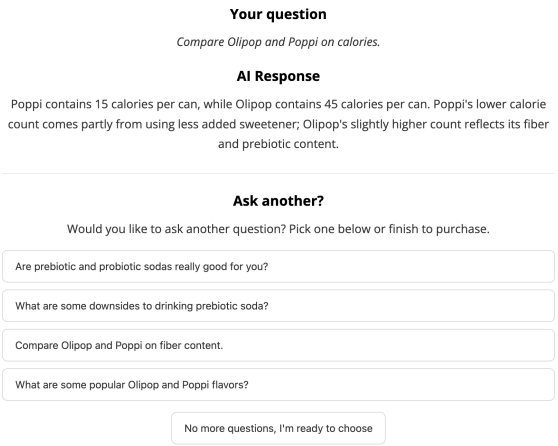


Figure 5: AI-Qs Website Layout - AI Response



(a) Follow-up questions



(b) Follow-up Questions with Response

Figure 6: AI-Qs Follow-up Questions Layout

The websites we built presented the same information to participants regardless of the condition they were assigned to. We kept AI responses static following prior work (Pfrommer et al., 2024; Sharma et al., 2024), which allowed us to control the content that participants saw, ensuring that we can explain *why* we observed changes in behavior between conditions. For this portion of the experiment, we collected data on: choices, time spent on task, number of clicks (Trad condition) and follow-up questions (AI-Qs condition).

Once participants made a choice, they answered questions on: demographics (age, gender, income, education), their search strategy (free text answer), previous use and trust of AI, prebiotic soda purchase history, typical perceived search effort, as well as their product knowledge (i.e., we asked participants to estimate how much fiber and calories Olipop and Poppi contained after the task, irrespective of whether they saw this information or not). There were 290 participants in total ($M_{age} = 39.43, (SD = 13.35)$; 119 males, 162 females, 7 non-binary, and 2 who preferred not to answer), with 100 participants in the Trad condition, 93 in AI-BF, and 97 in AI-Qs.[31] We find that most participants have never or very infrequently purchased a product in our category previously and that they have a relatively high level of trust in AI ($M_{trust} = 4.32, (SD = 1.45)$ on a 7 point scale), which did not vary by condition. Participants also reported a high level of subjective effort across conditions ($M_{effort} = 4.79, (SD = 1.33)$), which did not vary by condition. These results are available in Appendix B.

5.2 Results

In Figure 7, we test the primary prediction of our model related to choice quality. To do so, we plot the bonus earned by participants in each of the three conditions. We analyzed these differences using an ANOVA and follow up tests (Tukey-corrected pairwise comparisons). Consistent with our model, AI-BF leads to the lowest bonus (AI-BF vs. AI-Qs: $\Delta M = -22.1$, $t(287) = -5.98$, $p < .001$; AI-BF vs. Trad: $\Delta M = -32.9$ $t(287) = -8.96$, $p < .001$), as the majority of participants chose Olipop (obtaining a bonus of 50 cents). In other words, emphasizing the best attribute of each product (as in AI-BF), and not displaying the inferior attributes, led participants to make an inferior choice relative to their preferences. Participants in the Trad condition chose Poppi in

[31] Information on income and education is in Appendix B; these variables did not significantly differ between conditions.

a majority of cases, significantly more than the other conditions (AI-Qs vs. Trad: $\Delta M = -10.8$, $t(287) = -2.97$, $p = .009$), thus maximizing their bonus. Additionally, we observed that the AI-Qs condition had a higher proportion of Poppi purchasers (the correct choice), which was driven by participants who asked the correct follow up questions, thus revealing full information.[32]

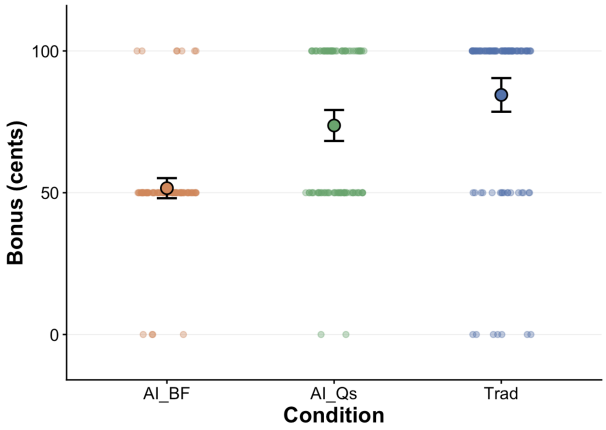


Figure 7: Choice Quality Results. Large dots represent means, bars represent 95% confidence intervals, and transparent colored dots indicate individual participant choices.

In Figure 8, we plot the amount of search performed in the two conditions that allow it: Trad (unique brands clicked) and AI-Qs (number of relevant follow-up questions). We find that (1) the majority of participants searched both brands in the Trad condition, obtaining full information, and (2) most participants in the AI-Qs condition either asked both relevant follow-ups or none. Importantly, the proportion of participants asking the two relevant follow-up questions in AI-Qs was however significantly less than the number of participants searching both brands in Trad search (Wilcoxon rank-sum test, $p < .001$). Thus, fewer participants in AI-Qs obtain full information than in Trad, allowing for inferior choices. Also, since utility is fixed across conditions, this difference suggests that clicks (c in our model) may be less costly than asking follow-up questions (q in our model). Finally, we note that in our experiment, asking follow-up questions may be easier than in real AI interactions as we provided a set of follow-ups to choose from rather than requiring participants to think of and write a follow-up prompt. Thus, our result may understate the effect in practice: in real interactions with an AI search tool that uses a best-for strategy, more participants may not ask follow-ups and instead rely on the initial AI response — that is, more of them may

[32]We used a logistic regression, removing the two people who chose water in this condition, to assess whether people who asked more questions earned better bonuses; the results show a significant relationship between asking more questions and earning the full bonus: $OR = 3.36$, 95% CI $[2.01, 5.61]$, $p < .001$.

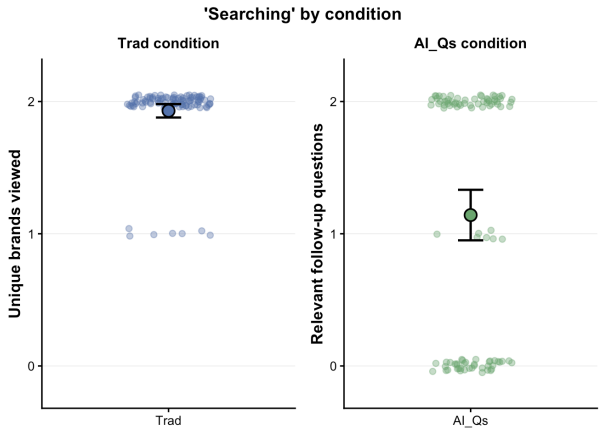


Figure 8: Amount of Search Results. Large dots represent means, bars represent 95% confidence intervals, and transparent dots indicate individual participant choices.

make an inferior choice when the hidden attributes of the favored products are poor.

We also assessed how long participants took to complete the search task in each condition (Figure 9). We find that participants in the AI-Qs condition spent significantly more time completing the task compared to participants in both AI-BF and Trad conditions (AI-BF - AI-Qs: $\Delta M = -40.02$, $t(287) = -4.35$, $p < 0.001$; Trad - AI-Qs: $\Delta M = -31.71$, $t(287) = -3.51$, $p = 0.002$); p-values for pairwise comparisons corrected via Bonferroni method). Once again, this result suggests that follow-up questions may be costlier than clicking in traditional search. Thus, contrary to optimistic assessments in the popular press on AI's potential to reduce search frictions, it may not be more efficient for the consumer to search via AI, especially when follow up questions need to be asked. Finally, there was no significant reduction in search time in the AI-BF condition relative to Trad search (AI-BF - Trad: $\Delta M = -8.31$, $t(287) = -.91$, $p = .635$), suggesting that there may be a cost of reading and parsing the AI response text.

Next, we analyzed how participants described their own search strategies. To do this, we focused on the share of open-ended responses that mentioned each of the two key attributes (Figure 10). Mentions of calories were stable across conditions at roughly 45% of responses. Mentions of fiber, however, differ dramatically: only about 45% of Trad responses reference fiber, rising to approximately 60% under AI-Qs and exceeding 75% under AI-BF. We analyzed these differences with a logistic regression: participants in AI-BF had significantly higher odds of mentioning fiber than those in AI-Qs (OR = 2.35, $z = 2.61$, $p = .024$) and the Trad condition (OR = 5.04, $z = 5.00$, $p < .001$). Participants in the AI-Qs condition also had significantly higher odds of mentioning fiber

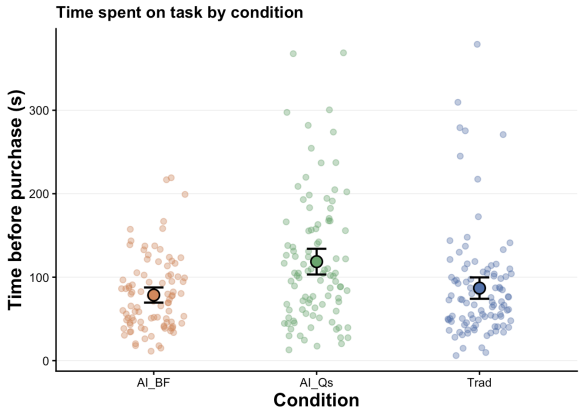


Figure 9: Time Spent on Task Results. Large dots represent means, bars represent 95% confidence intervals, and transparent dots indicate individual participant choices.

than those in the Trad condition (OR = 2.14, $z = 2.63$, $p = .023$). This escalation suggests that the best-for format affects how people approach the search task—drawing attention disproportionately toward the highlighted attribute and, in the AI-BF condition, concentrating it on fiber to a degree that explains participants' inferior choices. It is worth reiterating that given participants received the same instructions on how to weigh fiber versus calories prior to each task, it must be the format of the information provided during search which is directly affecting people's search strategies (or at least their reporting of these strategies).

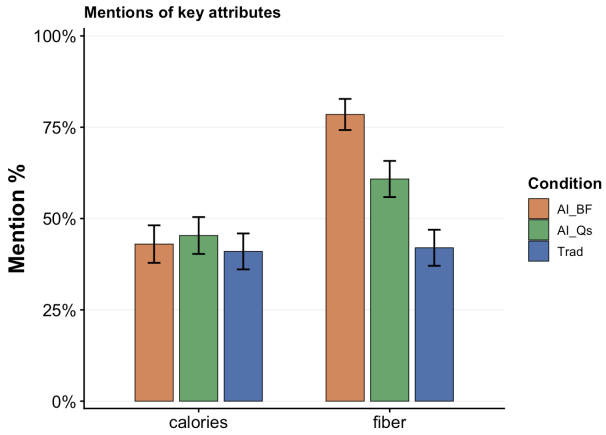


Figure 10: Mentions of Key Attributes in Search Strategy Description. Bars represent 95% confidence intervals.

Finally, we assessed participants' knowledge of fiber and calorie levels, which we collected after the search task. For each brand, participants were asked to estimate the amount of fiber and calories

each soda had. Generally, we did not find significant differences across conditions, with only a few exceptions: (1) participants in the Trad condition deviated from the true value of Poppi's fiber more than participants' in either AI condition (AI-BF: $b = 2.40$, $t(275) = 3.85$, $p < .001$; AI-Qs: $b = 1.51$, $t(275) = 2.48$, $p = .036$), (2) participants in AI-Qs were more accurate than those in Trad for Olipop calorie estimates ($b = 10.17$, $t(277) = 2.57$, $p = .029$), and (3) participants in the AI-Qs condition also had better Poppi calorie estimates than those in AI-BF condition ($b = -6.24$, $t(282) = -2.39$, $p = .046$). One potential interpretation of these results is that participants have a stronger memory of the product information they obtained when using AI to search than an e-retailer's website, since the former compares brands rather than presents raw attributes. Several of these differences, however, are quite small and responses were heterogeneous, so we caution against drawing strong conclusions from these differences.

6 Discussion and Conclusion

This paper studies whether AI search tools help consumers make better product choices. We approach this question by: (i) documenting that AI presents product information in a "best-for" format when responding to shopping-related questions, (ii) developing a theoretical model that formalizes the distinctive "best-for" information structure of AI responses, and (iii) testing the model's predictions in an incentivized experiment.

Several important implications follow from our analysis.

Consumer welfare. AI shopping agents will benefit most consumers most of the time, in line with optimistic assessments in the popular press and prior academic work (Spatharioti et al., 2025). The welfare concern we document is concentrated in a specific scenario: when a consumer's most-valued attribute points toward product j, but product j's other attributes are poor enough that an alternative product would be preferred under full information. Whether this scenario is empirically common depends on the joint distribution of product attributes, consumer preferences in specific markets, and on information acquisition costs.

Competition and market structure. We document brand and attribute coverage across thousands of AI responses, which can help shed light on which products receive favorable AI placement and how this varies across AI tools and specific prompts—a question with direct implications for

market competition and antitrust.

Limitations and future work. Our experiment involves a single product category (prebiotic sodas) and a simple two-product environment with a predefined utility rule. We adopt this simplistic design (with defined utilities) to ensure there is an objectively correct answer which can be easily understood by participants, without needing to obtain noisy estimates of their own preferences. Future work however should test the model's predictions across richer product spaces and examine whether the effects we document persist over time as consumers gain more experience with AI search tools; for example, consumers may learn that AI tools provide "best for" formatting and become accustomed to asking follow-up questions when necessary. Additionally, while we abstract away from AI hallucinations and strategic incentives for AI companies to promote specific brands, these considerations may become increasingly important as AI-affiliate advertising models develop.

Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

Agarwal, S. and A. Sen (2026). The impact of google ai overviews on publisher traffic and user experience: Evidence from a field experiment.

Allouah, A., O. Besbes, J. Figueroa, Y. Kanoria, and A. Kumar (2025). What is your ai agent buying? evaluation, implications, and emerging questions for agentic e-commerce. *Evaluation, Implications, and Emerging Questions for Agentic E-Commerce (August 04, 2025)*.

Anderson, S. P. and R. Renault (1999). Pricing, product diversity, and search costs: A Bertrand-Chamberlin-Diamond model. *The RAND Journal of Economics 30*(4), 719–735.

Andrews, J. C., S. Burton, and R. G. Netemeyer (2000). Are some comparative nutrition claims misleading? the role of nutrition knowledge, ad claim type and disclosure conditions. *Journal of Advertising 29*(3), 29–42.

Andrews, J. C., R. G. Netemeyer, and S. Burton (1998). Consumer generalization of nutrient content claims in advertising. *Journal of Marketing 62*(4), 62–75.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis 31*(3), 337–351.

Arora, N., I. Chakraborty, and Y. Nishimura (2025). Ai–human hybrids for marketing research: Leveraging large language models (llms) as collaborators. *Journal of Marketing 89*(2), 43–70.

Bai, Y., A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Brand, J., A. Israeli, and D. Ngwe (2023). Using llms for market research. *Harvard business school marketing unit working paper* (23-062).

Brown, J. R. and A. Goolsbee (2002). Does the internet make markets more competitive? evidence from the life insurance industry. *Journal of political economy 110*(3), 481–507.

Brucks, M. and O. Toubia (2023). Prompt architecture can induce methodological artifacts in large language models. *Available at SSRN 4484416*.

Buçinca, Z., M. B. Malaya, and K. Z. Gajos (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction 5*(CSCW1), 1–21.

Chatterji, A., T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman (2025). How people use chatgpt. Technical report, National Bureau of Economic Research.

Chen, Y. and S. Yao (2017). Sequential search with refinement: Model and application with click-stream data. *Management Science 63*(12), 4345–4365.

Dang, I. C., R. M. Ursu, and P. K. Chintagunta (2026). Going back to move forward? how search revisits on a website we built inform us about search outcomes. *Quantitative Marketing and Economics 24*(1), 6.

Delaprez, Y. and A. Hortacsu (2026). The welfare impact of large language models. *Kilts Center at Chicago Booth Marketing Data Center Paper Forthcoming*.

Dillion, D., N. Tandon, Y. Gu, and K. Gray (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences 27*(7), 597–600.

Goli, A. and A. Singh (2024). Frontiers: Can large language models capture human preferences? *Marketing Science 43*(4), 709–722.

Gui, G. and O. Toubia (2023). The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*.

Haan, M. A., J. L. Moraga-González, and V. Petrikaitė (2018). A model of directed consumer search. *International Journal of Industrial Organization 61*, 223–255.

Honka, E. and P. Chintagunta (2017). Simultaneous or sequential? search strategies in the us auto insurance industry. *Marketing Science 36*(1), 21–42.

Horton, J. J., A. Filippas, and B. S. Manning (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Keller, S. B., M. Landry, J. Olson, A. M. Velliquette, S. Burton, and J. C. Andrews (1997). The effects of nutrition package claims, nutrition facts panels, and motivation to process nutrition information on consumer product evaluations. *Journal of Public Policy & Marketing 16*(2), 256–269.

Kim, J., M. Giroux, and J. C. Lee (2021). When do you trust ai? the effect of number presentation detail on consumer trust and acceptance of ai recommendations. *Psychology & marketing 38*(7), 1140–1155.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010). Online demand under limited consumer search. *Marketing science 29*(6), 1001–1023.

Li, H. and S. Aral (2025). Human trust in ai search: A large-scale experiment. *arXiv preprint arXiv:2504.06435*.

Lin, S., J. Hilton, and O. Evans (2022). Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Lin, S., Z. J. Shi, and X. Sun (2025). Towards intelligent shopping assistant: How can llm-based chatbots transform online shopping process? *HKUST Business School Research Paper* (2025-196).

Liu, J. and O. Toubia (2018). Search query formation by strategic consumers. *Columbia Business School Research Paper* (18-30).

Liu, J. and O. Toubia (2025). The effects of search interface on consumer preferences: Evidence from a randomized experiment. *HKUST Business School Research Paper* (2025-192).

Misra, S. (2025). Foundation priors. *arXiv preprint arXiv:2512.01107*.

Morozov, I. and A. Tuchman (2024). Where does advertising content lead you? we created a bookstore to find out. *Marketing Science 43*(5), 986–1001.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems 35*, 27730–27744.

Padilla, N., H. T. Lam, A. Lambrecht, and B. Hollenbeck (2025). The impact of llm adoption on online user behavior. *Available at SSRN*.

Passi, S., S. Dhanorkar, and M. Vorvoreanu (2024). Appropriate reliance on generative ai: Research synthesis. Technical report, Technical Report MSR-TR-2024-7. Microsoft. https://www. microsoft. com/en-us ....

Pfrommer, S., Y. Bai, T. Gautam, and S. Sojoudi (2024). Ranking manipulation for conversational search engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9523–9552.

Rao, A. and R. Ursu (2025). The impact of voluntary labeling. *Marketing Science 44*(4), 874–893.

Roe, B., A. S. Levy, and B. M. Derby (1999). The impact of health claims on consumer search and product evaluation outcomes: Results from fda experimental data. *Journal of Public Policy & Marketing 18*(1), 89–105.

Shahidi, P., G. Rusak, B. S. Manning, A. Fradkin, and J. J. Horton (2025). The coasean singularity? demand, supply, and market design with ai agents. Technical report, National Bureau of Economic Research.

Sharma, N., Q. V. Liao, and Z. Xiao (2024). Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.

Spatharioti, S. E., D. Rothschild, D. G. Goldstein, and J. M. Hofman (2025). Effects of llm-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Stigler, G. J. (1961). The economics of information. *Journal of political economy 69*(3), 213–225.

Ursu, R., S. Seiler, and E. Honka (2025). The sequential search model: A framework for empirical research: R. ursu et al. *Quantitative Marketing and Economics 23*(1), 165–213.

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

Vasconcelos, H., G. Bansal, A. Fourney, Q. V. Liao, and J. Wortman Vaughan (2025). Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Transactions on Computer-Human Interaction 32*(1), 1–30.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica 47*(3), 641–654.

Werner, T., I. Soraperra, E. Calvano, D. C. Parkes, and I. Rahwan (2024). Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection. *arXiv preprint arXiv:2409.12143*.

Xu, R., Y. Feng, and H. Chen (2023). Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*.

Yan, S., H. Zhong, Z. Z. Zhong, W. Zhou, and N. Mehta (2025). The impact of reasoning ai on consumer purchases: Evidence from an online platform field experiment. *Wenyu and Mehta, Nitin, The Impact of Reasoning AI on Consumer Purchases: Evidence from an Online Platform Field Experiment (October 20, 2025).*

Zhang, B., J. Li, A. Hortaçsu, X. Ye, V. Chernozhukov, A. Ni, and E. W. Huang (2025). Agentic economic modeling. *arXiv preprint arXiv:2510.25743*.

Zhang, X., W. Xiong, L. Chen, T. Zhou, H. Huang, and T. Zhang (2025). From lists to emojis: How format bias affects model alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26940–26961.

Zhao, W., X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng (2024). Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Zheng, S., Y. Zhu, X. Ye, and L. Shen (2025). Generative search: Evidence from a large-scale field experiment. *Available at SSRN 5297194*.

Appendix

A Model - Additional Derivations

See y for free on the list page

Across all cases, the consumer expects $u_1 = \alpha x_1 + \beta y_1$ and $u_2 = \alpha x_2 + \beta H$, with y_1 equal to either L (cases 1 and 2) or M (cases 3 and 4), but known before deciding whether to search. The x attribute for both products is unknown without search. Given that she expects the x attribute for both products to be $\{L, M, H\}$ with equal probability, she starts search with product 2.

In case 1, the consumer expects $u_1 = \alpha x_1 + \beta L$ and after searching product 2, she reveals $u_2 = \alpha L + \beta H$. If the consumer does switch to product 1 when $x_1 = M$, then the search condition for product 1 is $c < \frac{\alpha(H+M-2L)-2\beta(H-L)}{3}$, while if she does not switch when $x_1 = M$, then the search condition is $c < \frac{(\alpha-\beta)(H-L)}{3}$. With low enough search costs, the consumer searches product 1, discovers $u_1 = \alpha L + \beta H > u_2$, and chooses the superior product (product 1), as she would with

AI. However, if search costs are large, then the consumer will not search product 1 and will choose product 2, i.e., she will choose a lower utility product than when she uses AI.

In case 2, similarly the consumer expects $u_1 = \alpha x_1 + \beta L$ and after searching product 2, she sees $u_2 = \alpha M + \beta H$, but now there is no value of x_1 for which the consumer would consider searching product 1. Thus, the consumer does not search beyond product 2, so she will choose product 2. This choice coincides with the one under full information and dominates the AI-guided choice.

In case 3, the consumer expects $u_1 = \alpha x_1 + \beta M$ and after searching product 2, she sees $u_2 = \alpha L + \beta H$. If she does switch when $x_1 = M$, then the search condition for product 1 is $c < \frac{\alpha(H+M-2L)-2\beta(H-M)}{3}$, while if she does not switch when $x_1 = M$, then the search condition is $c < \frac{\alpha(H-L)-\beta(H-M)}{3}$. With low enough search costs, the consumer searches and chooses the superior product (product 1), as she would with AI. However, if search costs are large, then the consumer will not search product 1 and will choose product 2, i.e., she will choose a lower utility product than when she uses AI.

Finally, in case 4, the consumer expects $u_1 = \alpha x_1 + \beta M$ and after searching product 2, she sees $u_2 = \alpha M + \beta H$, and would search whenever $c < \frac{(\alpha-\beta)(H-M)}{3}$. Again, if she searches, she chooses the superior product 1 (coinciding with both the AI-guided choice and the full information choice), but if she doesn't, she chooses the inferior product 2.

See neither x nor y for free on the list page

Unlike in the cases where either x or y were available on the list page for free, not seeing either attribute for free before search makes the consumer indifferent about the first product she searches. We consider both cases below.

Start search with product 1

In cases 1 and 2, after observing product 1's utility, $u_1 = \alpha H + \beta L$, and expecting $u_2 = \alpha x_2 + \beta y_2$, the consumer would only consider searching product 2 if $(x_2, y_2) = \{(M, H), (H, M), (H, H)\}$, so her search condition is given by $c < \frac{\alpha(M-H)+\beta(2H+M-3L)}{9}$. In case 1, regardless of search, the consumer chooses product 1, while in case 2, if she searches she will choose product 2, but she will stick with product 1 if she does not search. Thus, search helps the consumer make a better choice than under the AI regime in case 2, without affecting her final choice in case 1. In cases 3 and 4, after observing product 1's utility, $u_1 = \alpha H + \beta M$, and expecting $u_2 = \alpha x_2 + \beta y_2$, the consumer

would only consider searching product 2 if $(x_2, y_2) = \{(H, H)\}$, so her search condition is given by $c < \frac{\beta(H-M)}{9}$. In both cases, regardless of search, the consumer will still choose product 1, choice that coincides with the AI-guided choice.

Start search with product 2

In cases 1 and 3, after observing product 2's utility, $u_2 = \alpha L + \beta H$, and expecting $u_1 = \alpha x_1 + \beta y_1$, the consumer would only consider searching product 1 if $(x_1, y_1) = \{(M, H), (H, L), (H, M), (H, H)\}$, with the case $(x_1, y_1) = \{(M, M)\}$ ambiguous. So her two search conditions are either $c < \frac{3\alpha H - 2\beta H + \beta M + (\beta - 4\alpha)L}{9}$ or $c < \frac{3\alpha H - 3\beta H + 2\alpha M + \beta M + \beta L - 5\alpha L}{9}$. In both cases, if the consumer searches, she will prefer to choose product 1 over 2, choice that coincides with both the AI-guided choice and the full information choice. However, not searching means choosing product 2, the inferior choice. In cases 2 and 4, after observing product 2's utility, $u_2 = \alpha M + \beta H$, and expecting $u_1 = \alpha x_1 + \beta y_1$, the consumer would only consider searching product 1 if $(x_1, y_1) = \{(H, M), (H, H)\}$, so her search condition is $c < \frac{(2\alpha - \beta)(H-M)}{9}$. In case 2, regardless of search, the consumer will continue choosing product 2, the superior choice. In case 4, if the consumer searches, she will choose product 1, the better choice compared to the AI-guided choice; not searching means making the inferior choice, product 2.

B Experiment: Demographics and Balance Checks

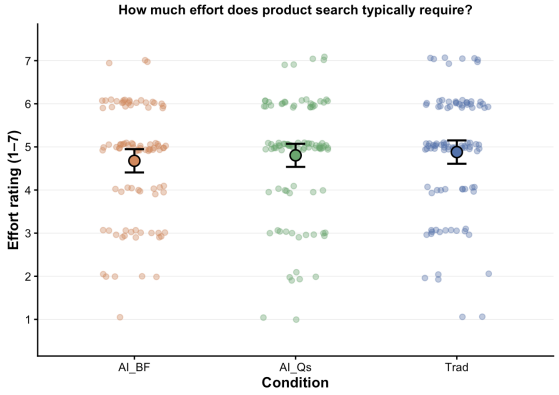
Full demographic breakdowns are reported in Table 9 and Table 10. None of the demographics (age, gender, income, education) significantly differed between groups (ps $> .05$ for all relevant tests).

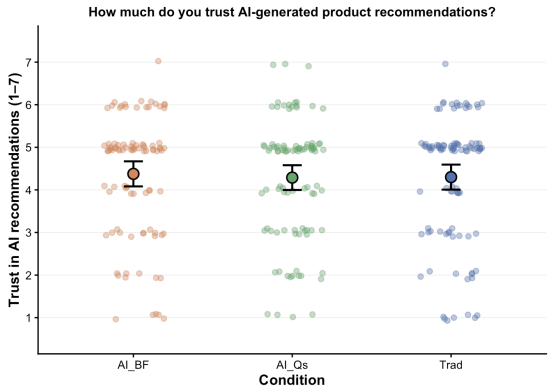
Table 9: Educational attainment of participants

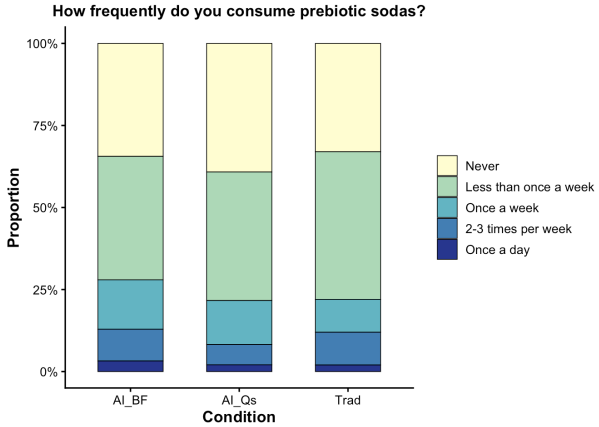
Education	n	%
Bachelor's Degree	117	40.3
Postgraduate Degree	56	19.3
High School (or GED)	46	15.9
Associate's Degree	35	12.1
Partial university education	34	11.7
Less than High School	2	0.7

Table 10: Household income of participants

Income	n	%
$50,000 – $74,999	60	20.7
$100,000 – $149,999	58	20.0
$75,000 – $99,999	55	19.0
$0 – $24,999	33	11.4
$35,000 – $49,999	27	9.3
$25,000 – $34,999	24	8.3
$150,000 – $199,999	22	7.6
$200,000 or above	11	3.8







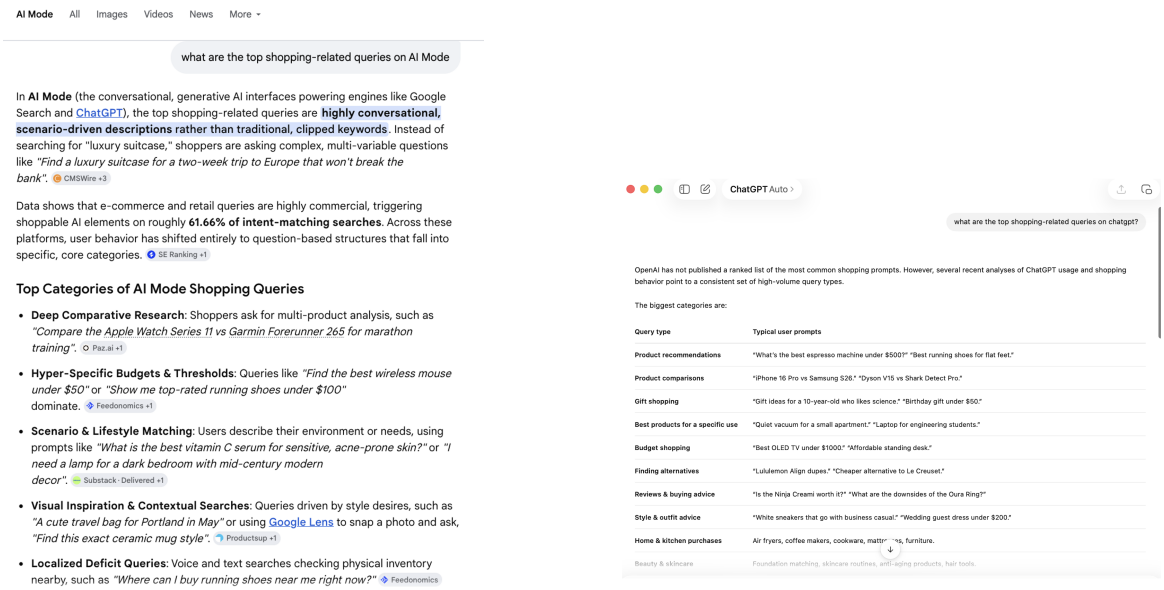
Online Appendix

OA.1 Sources used to identify shopping-related prompts

As described in the main text, we use several data sources to identify the final list of shopping-related prompts presented in Table 1.

AI questions and industry reports

- Directly asking (on July 7, 2026) ChatGPT and Google AI Mode what the top shopping-related queries were (prompt "what are the top shopping-related queries on chatgpt/AI Mode?") reveals the following lists for each (see Figure OA.1).



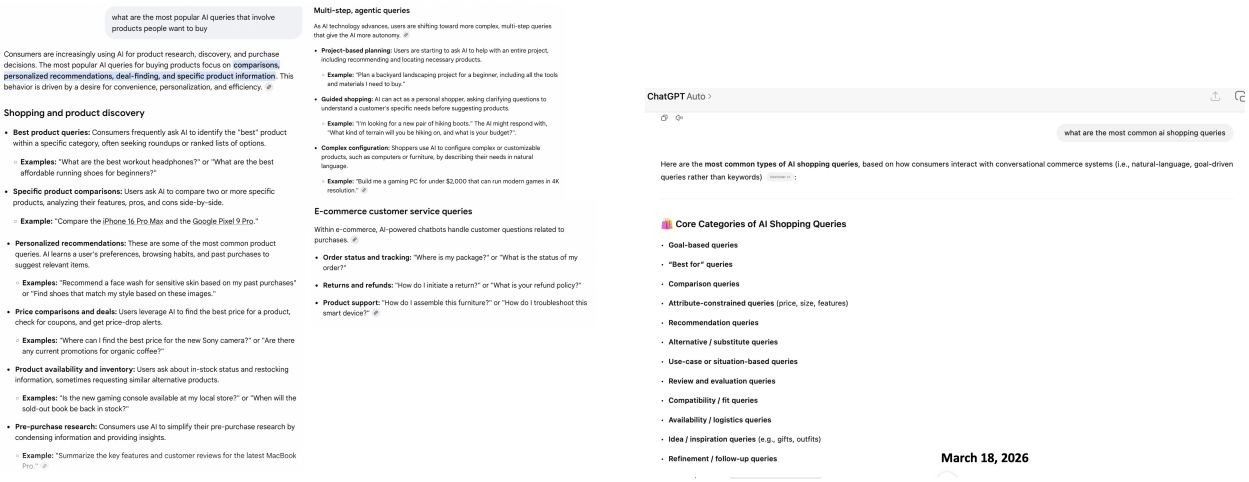
(a) AI Mode (b) ChatGPT

Figure OA.1: AI Responses (July 7, 2026)

- We repeatedly asked both ChatGPT and AI Mode related questions since October 2025 and include two screenshots in Figure OA.2: AI Mode on October 10, 2025 and ChatGPT on March 18, 2026. We do so to demonstrate that the types of shopping-related questions have not changed over this time period, even though AI models have.

- On November 24, 2025 OpenAI launched a new ChatGPT shopping research feature.[33] In its announcement of this feature, OpenAI highlighted being able to use this feature for 5 tasks

[33] Article available at https://openai.com/index/chatgpt-shopping-research/, last retrieved July 7, 2026.



(a) AI Mode (b) ChatGPT

Figure OA.2: AI Responses (Google AI Mode October 2025 and ChatGPT March 18, 2026)

and gave examples of prompts for each feature: discovering new products ("help me find a powerful new laptop suitable for gaming under $1000 with a screen that's over 15 inches"), product comparison ("I need a stroller [...] Please compare these three models and share the best fit and trade-offs"), gift shopping ("I need a gift for my dad who is a fishing fanatic, but never seems to catch anything"), finding lookalikes ("I'm looking for a dress that looks like this, but is full length is less than $250"), and finding deals ("help me find the best Black Friday deal [...]").[34]

- OpenAI's NBER working paper (Chatterji et al., 2025), analyzing more than 1 million conversations from 700M weekly active users, reveals that 2.1% of queries are shopping-related and pertain to a purchasable product, and it gives the following five examples of prompts: "iPhone 15", "What's the best streaming service?", "How much are Nikes?", "Cu´anto cuesta un Google Pixel?", "Recommend a good laptop under $1000."

- In May 2026, Google released an article on how people in the US use AI Mode.[35] The article reveals that the top first keywords in AI Mode queries are (in order): what, how, I, is, can.

- Pfrommer et al. (2024) collect data using a singular prompt "I'm looking to buy a <category>.

[34]Throughout this section (where relevant) we have renamed categories of prompts that are similar under the same name, such as 'gifts' and 'gift shopping' were both renamed as 'gift shopping'.

[35]Article available at https://business.google.com/us/think/search-and-video/ai-mode-us-insights/, last retrieved July 7, 2026.

Can you give me some recommendations". This prompt matches the structures we identified and the responses the authors collected match the format of the responses we document in this paper.

Based on these data sources, the most common shopping-related prompt types (also in the first column of Table OA.1) are: Best/top products, Product comparison, Product recommendation/'Help me' advice, Gift shopping, Budget/deals, Lifestyle matching, Summarize reviews, and Product availability and inventory. Most of these prompt types can either ask about general product information (e.g., "what is the best mattress?") or about a specific use case mentioning one or more constraints (e.g., "what is the best mattress for side sleepers?"). Two common constraints mentioned are price/budget and lifestyle matching/features, which we included as separate types above, as mentioned in the sources cited. We omitted prompt types that are e-retailer specific (Order status, Returns and refunds, Product support, Compatibility/fit, etc) or that are hard to implement using our technology (Finding lookalikes/visual search, Multi-step queries, etc).

WildChat data

We use the WildChat data, a publicly available data set collected between April 9, 2023 and May 1, 2024 by Zhao et al. (2024) from Chatgpt (both the 3.5-Turbo and 4 models).[36] The data cover conversations across countries, the largest of which is the US (21.6%). Most conversations are not shopping-related, for example the largest category is "assisting/creative writing" with 62% of conversations. We focused on the first prompts submitted (approximately 838k conversations) that were in English (approximately 478k conversations, 268k unique) and, using keyword search, obtained 123 conversations that were shopping-related.[37] Approximately half (56%) of initial prompts had follow-up questions. We used these data to obtain a first classification of what may be the most common structure of shopping-related prompts across categories. Our results appear in the second column of Table OA.1.

Prolific survey

We designed a Prolific survey to ask individuals directly how they would phrase a shopping-

[36]Data publicly available through https://arxiv.org/abs/2405.01470.
[37]Online Appendix OA.2 provides all data cleaning steps.

related query in a category of our choosing – prebiotic soda.[38] The Prolific survey ran June 6, 2026 and was answered by 97 participants (51% Male, 61% with a B.A degree or better, 25% with an annual income of $\$100,000$ or more). It took approximately 7 minutes to complete and paid participants at a rate of $12 an hour. Only 25% use AI frequently (indicated "often" or "always" in their responses) and for most participants prebiotic soda is a new category (48% have never purchased and only 16% purchased once a week or more frequently) – one of the reasons for picking this category.

We analyzed the prompts to reveal that approximately half (47%) contained the words "best", "recommend", "help me" or "compare". After manually inspecting the prompts that did not include those keywords, we further classified all prompts and included them in the third column of Table OA.1.[39]

In addition to asking participants to type in a query they would submit to an AI tool, we also asked them what their intent was and found that most expected responses that compare brands (47%), followed by those who expected responses where brands would be described in terms of their best features (32%) and by general category information (13%), with only 7 participants expecting a different type of response, showing that participants' intent matched the type of common prompts uncovered above using WildChat and industry report sources.

The final list of prompts we use appears in Table 1 in the main paper. All prompts is that table are supported by several sources cited here, as we detail below:

1. **Best/top products** (specific prompts "What is the best prebiotic soda?", "What are the best wireless headphones?") This is the most common type of shopping question. This question type is supported by all the direct AI tool questions, the NBER article ("What's the best streaming service?"), the May 2026 Google report on AI Mode ("what" questions), the WildChat Data (42% of prompts asked for the best/top products), and our Prolific survey (47% of prompts asked for the best/top products).

2. **Product comparison** (specific prompts "Can you compare different prebiotic sodas?", "Compare the iPhone 16 Pro Max and the Google Pixel 9 Pro."). This question type is

[38] Pre-registered study under As Predicted #295018.

[39] Percentages may not add up to 100 given that some prompts fall in multiple categories, such as asking for a recommendation while imposing a budget constraint.

Table OA.1: Shopping Prompt Categories Across Several Sources.

	Prompt types based on AI questions and industry reports	**WildChat**	**Prolific survey**
1	Best/top products	42% asked for the best/top products	47% asked for the best/top products
2	Product comparison	11% asked to compare products	7% asked to compare products
3	Product recommendations	8% asked for a recommendation	23% asked for a recommendation and 3.5% asked "help me" advice
4	Gift shopping		
5	Budget/deals	19% asked about affordable/budget options	7% asked about affordable/budget options
6	Lifestyle matching		
7	Summarize reviews		2% asked about reviews
8	Product availability and inventory		4% asked about product availability and inventory
9	Other	Rest asked for general category information or about fulfilling a purchase	Rest asked for general category information

supported by all the direct AI tool questions, the November 2025 OpenAI report announcing the ChatGPT shopping research feature, the May 2026 Google report on AI Mode ("can" questions), the WildChat Data (11% of prompts asked to compare products), and our Prolific survey (7% of prompts asked to compare products).

3. **Product recommendations/'Help me' advice** (specific prompts "Can you recommend a prebiotic soda?", "Recommend affordable wireless headphones for a beginner runner", "I am a woman who lives in a city and I am always cold. Recommend winter jackets", "Help me choose a prebiotic soda", "Help me find a great housewarming gift under $100"). This question type is supported by the direct AI tool questions (Google AI Mode October 10, 2025 and ChatGPT March 18, 2026), the May 2026 Google AI Mode report ("can" questions), the November 2025 OpenAI report announcing the ChatGPT shopping research feature, the NBER article ("Recommend a good laptop under $1000), prior work Pfrommer et al. (2024), the WildChat Data (8% of prompts asked for a recommendation), and our Prolific survey (23% of prompts asked for a recommendation and 3.5% asked "help me" questions).

4. **Gift shopping** (specific prompts "Help me find a great housewarming gift under $100"). This question type is supported by the direct AI tool questions (ChatGPT July 7, 2026, ChatGPT March 18, 2026), and the November 2025 OpenAI report announcing the ChatGPT shopping research feature.

5. **Budget/deals** (specific prompts "What is the best price I can get on the new iPhone? iPhone price", "Recommend affordable wireless headphones for a beginner runner", "Help me find a great housewarming gift under $100"). This question type is supported by all direct AI tool questions, the November 2025 OpenAI report announcing the ChatGPT shopping research feature, the NBER article ("How much are Nikes?", "Cu´anto cuesta un Google Pixel?", "Recommend a good laptop under $1000"), the WildChat Data (19% of prompts asked affordable/budget options), and our Prolific survey (7% of prompts asked affordable/budget options).

6. **Lifestyle matching** (specific prompts "I am a woman who lives in a city and I am always cold. Recommend winter jackets"). This question type is supported by the direct AI tool questions (AI Mode July 7, 2026), by the May 2026 Google AI Mode report ("I" questions), and by prior work Pfrommer et al. (2024).

7. **Summarize reviews** (specific prompts "Summarize the reviews of the KitchenAid stand mixer"). This question type is supported by direct AI tool questions (ChatGPT July 7, 2026, Google AI Mode October 10, 2025, ChatGPT March 18, 2026), and our Prolific survey.

8. **Product availability and inventory** (specific prompts "List hotels in New York City available this holiday season"). This question type is supported by direct AI tool questions (Google AI Mode July 7, 2026, Google AI Mode October 10, 2025, ChatGPT March 18, 2026).

OA.2 WildChat Data Cleaning Steps

We obtained the WildChat-1M data from Hugging Face and extracted the very first prompt from each user. Then we filtered for clear shopping intent using the following shopping-signal keywords: buy, buying, purchase, purchasing, order, ordering, pick up, should i buy / get / purchase, can i

buy, could i buy, i want to buy, i need to buy, i'd like to buy, i want to purchase/get, i need to get, i want a new, i need a new, looking to buy, looking for, in the market for, shopping for, want to buy, planning to buy, thinking of buying, thinking of getting, where can i buy, where to buy, where do i buy, where can i get, where to get, where can i find, under, below, less than, within (+ a price), cheap, cheapest, affordable, budget, low-cost, price of, cost of, how much is, how much are, how much does, how much do, how much for, recommend, recommendation(s), suggest, suggestion(s), advice, advise, opinion(s) (on / for / about), which [product] ... should i buy / to buy / do you recommend / is good / is right, what's a good [product], good [product] for / to buy / under, vs, versus, compare, difference between, worth buying, worth it, worth the money, is [X] a good [product] / a good buy / a good choice / a good option, best, top N, greatest.

We excluded content generation (write / create / translate / essay / ad / lyrics / blog titles / pitch / cover letter), code & technical (code, html/js, compiler, image prompts), word problems & exercises (step-by-step calculation, "question N", accounting, "name this tense", case study, "read the text"), and same-word-different-meaning cases (TV show vs. TV, tablet=pill) keywords.

OA.3 AI Responses - Supplementary Data and Analysis

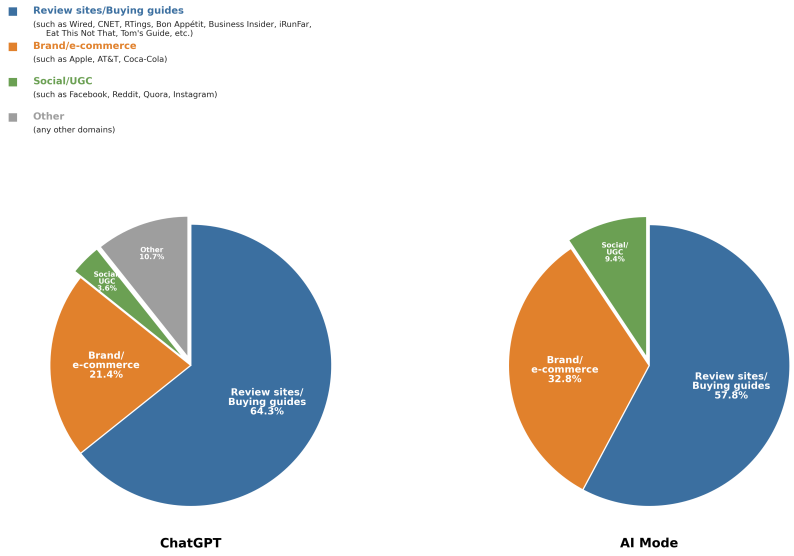
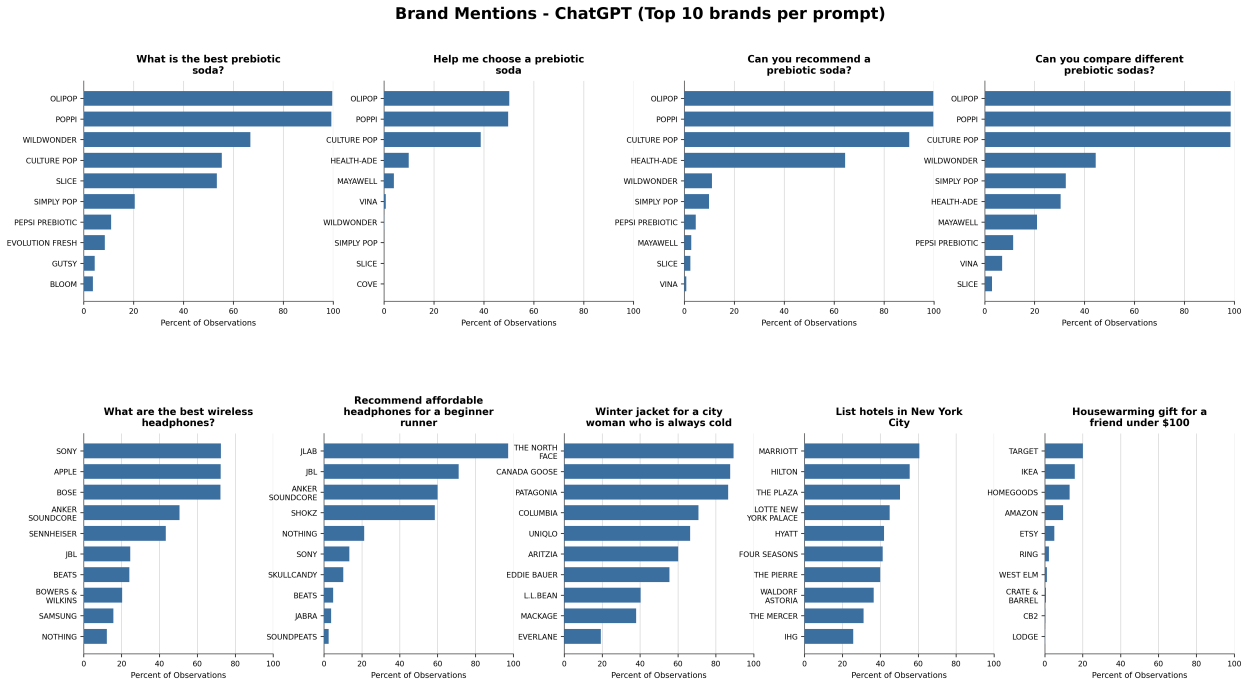


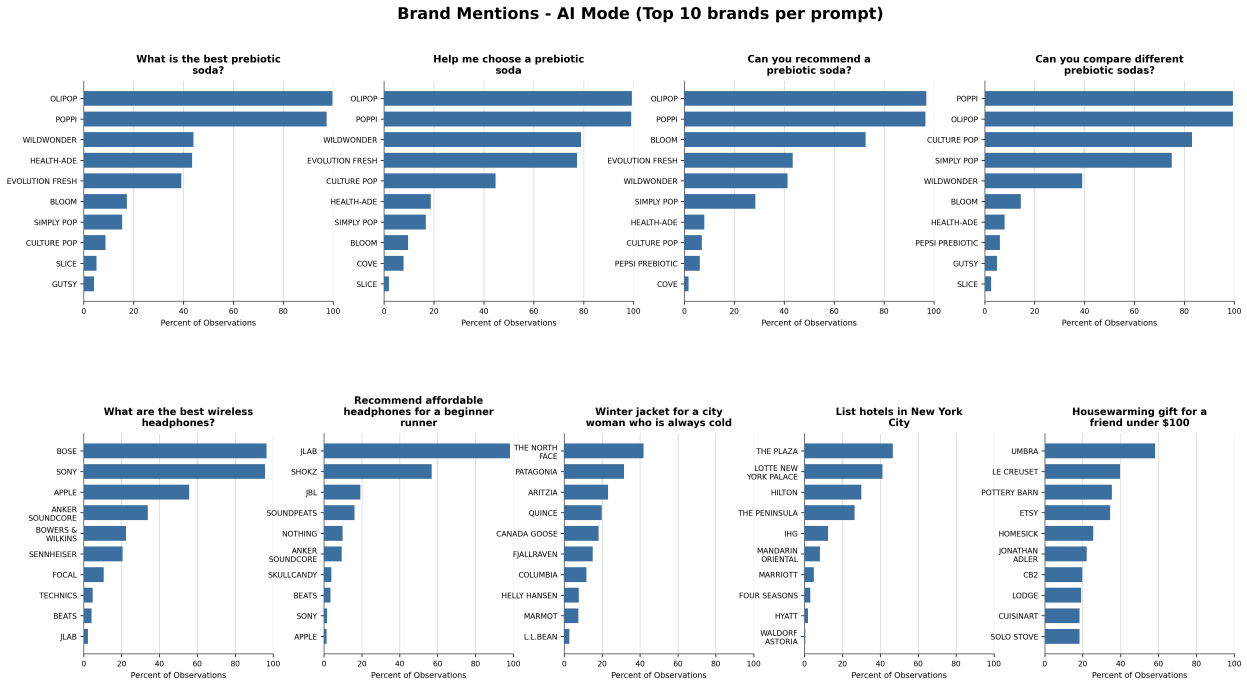
Figure OA.3: Top Citing Sources by AI Tool.

Table OA.2: Feature Groups and Brands Tracked for Each Prompt Group.

Group	Feature groups (all keywords)	Brands
prebiotic	fiber (fiber, gut, prebiotic, probiotic, poop, digestive, digestion); sugar (sugar); sweeteners (sweetener, stevia, monk fruit, artificial, real fruit, fruit juice, syrup); fiber source (jerusalem artichoke, inulin, apple cider vinegar, chicory root, cassava root, bacillus subtilis, soluble corn, olismart blend); calories (calorie); flavor (flavor, flavour, taste); price (price, cost, dollar, $, budget, affordable, cheap, expensive)	Olipop, Poppi, Culture Pop, Health-Ade, Wildwonder, Simply Pop, Slice, Cove, Gutsy, Bloom, Evolution Fresh, Mayawell, Vina, Halfday, Pepsi prebiotic
headphones	noise cancellation (noise cancellation, noise isolation, noise cancelling, ANC, anc, active noise); battery (battery, battery life, long listening, charge, charging); sound quality (sound, sound quality, call quality); comfort and fit (comfort, fit, secure fit); technology type (wireless, open-ear, open ear, bone conduction, bluetooth, pairing, mic, microphone); price (price, cost, dollar, $, budget, affordable, cheap, expensive); durability (durability, waterproof, sweatproof, sweat, splashproof, water resistance, water resistant, swimming)	Sony, Bose, Apple, Beats, Sennheiser, Jabra, JBL, Anker Soundcore, Audio-Technica, Shure, Bang & Olufsen, Bowers & Wilkins, Marshall, Skullcandy, Philips, Nothing, Samsung, Google, Focal, Technics, Sonos, Master & Dynamic, Soundpeats, Shokz, JLab, Plantronics / Poly, SteelSeries, Razer, Logitech, HyperX, Turtle Beach, Corsair
jacket	warmth (warmth, insulation, fill power, synthetic, down, faux fur, fleece, shell); weather protection (wind protection, waterproof, water resistance, water resistant, water-resistant, windproof, wind resistance, Gore-Tex, DWR); design (weight, length, knee, fit, hip, thigh, travel, hike, commute, walking, skiing, hood, pocket, color, packability); price (price, cost, dollar, $, budget, affordable, cheap, expensive)	Patagonia, The North Face, Arc'teryx, Canada Goose, Uniqlo, Columbia, Marmot, Fjallraven, Eddie Bauer, L.L.Bean, Helly Hansen, Aritzia, Everlane, Quince, Mackage, Moncler, Carhartt, Barbour
hotels	price (price, cost, dollar, $, budget, affordable, cheap, expensive, nightly, per night, rate, resort fee, deal); location (midtown, times square, central park, locaton, subway, downtown, brooklyn, soho, fidi, village, chelsea, train); amenities (breakfast, gym, wi-fi, kitchen, pool, parking, accessible); style (luxury, boutique, family-friendly, extended stay, hostel, pet-friendly, business); policy (refundable, free cancellation)	Marriott, Hilton, Hyatt, IHG, Accor, Four Seasons, The Peninsula, The Plaza, Aman, Mandarin Oriental, Waldorf Astoria, The Pierre, The Mercer, Lotte New York Palace
mixer	ease of cleaning (easy to clean, dishwasher-safe, dishwasher safe); speed; accessories (accessorie, attachment, beater, bowl, whisk, dough hook, paddle, pouring shield, beater); performance (motor, power, watt); price (price, cost, dollar, $, budget, affordable, cheap, expensive); durability (durability, longevity, reliability, durable, reliable, build, heavy, years, decades, stiff, stress gears); noise; capacity; ratings (rating, star, recommend); ease of use (ease of use, beginner); design (design, style, aesthetic, color, showpiece)	KitchenAid
gifts	—	Le Creuset, YETI, Hydro Flask, OXO, Keurig, Nespresso, Philips Hue, Brooklinen, IKEA, Target, HomeGoods, Amazon, Etsy, Umbra, Pure Enrichment, Jonathan Adler, Cuisinart, Pottery Barn, West Elm, Crate & Barrel, CB2, Williams Sonoma, Solo Stove, Homesick, Lodge, Bodum, Fellow, Yamazaki, Picnic at Ascot, Boska, ILLUME, Jenni Kayne, Quince, Staub, Emile Henry, Riedel, Brightland, Click & Grow, UGG, Ring, Google Nest, Amazon Echo, Ninja, Victrola, Fujifilm Instax, Amerex, Unhide, Marycele, Dalkin & Co., RBT
iPhone price	—	Apple
iPhone compare	price (price, cost, dollar, $, budget, affordable, cheap, expensive); operating system (iOS, Android); display (display, OLED); performance (performance, chip, CPU, GPU, tensor, processor, ram, memory); battery (battery, gaming, charging, USB-C, Qi2, Qi wireless, MagSafe); camera (camera, video, photo, dolby); AI (Apple Intelligence, Gemini, Magic Editor, Call Notes, Pixel Screenshots); Material (titanium, aluminium, glass); size (size, lighter, bigger, heavier, weight)	Apple, Google

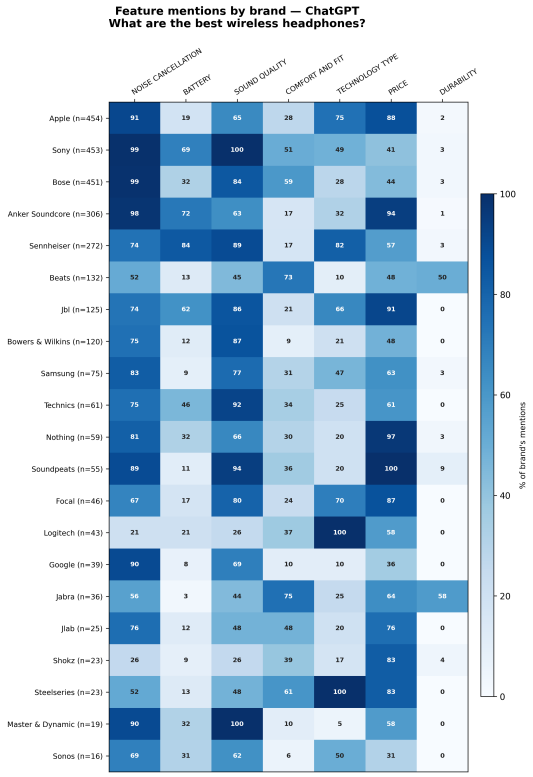


(a) ChatGPT

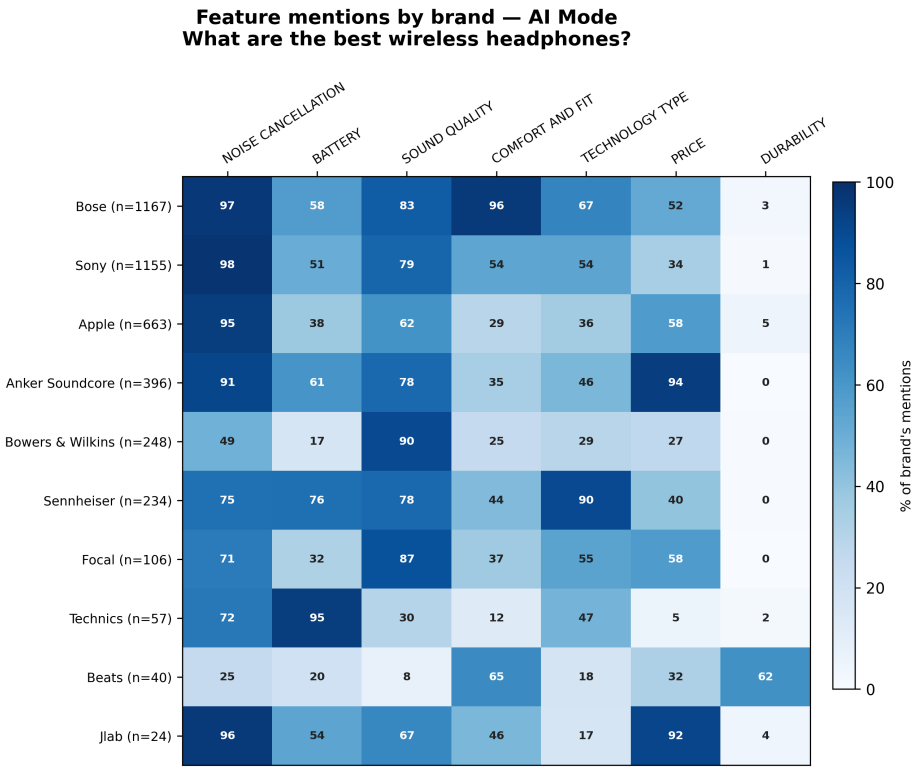


(b) AI Mode

Figure OA.4: Percent of Responses Containing Each Top 10 Brand by Prompt. The figure omits the three prompts that ask directly about one or two brands (iPhone price, iPhone camera, and mixer), since (not surprisingly) such prompts always mention those brands. Also, the "help me choose a prebiotic soda" prompt in about half of ChatGPT's responses does not return any brands, but instead asks the individual for more clarification on the question asked.

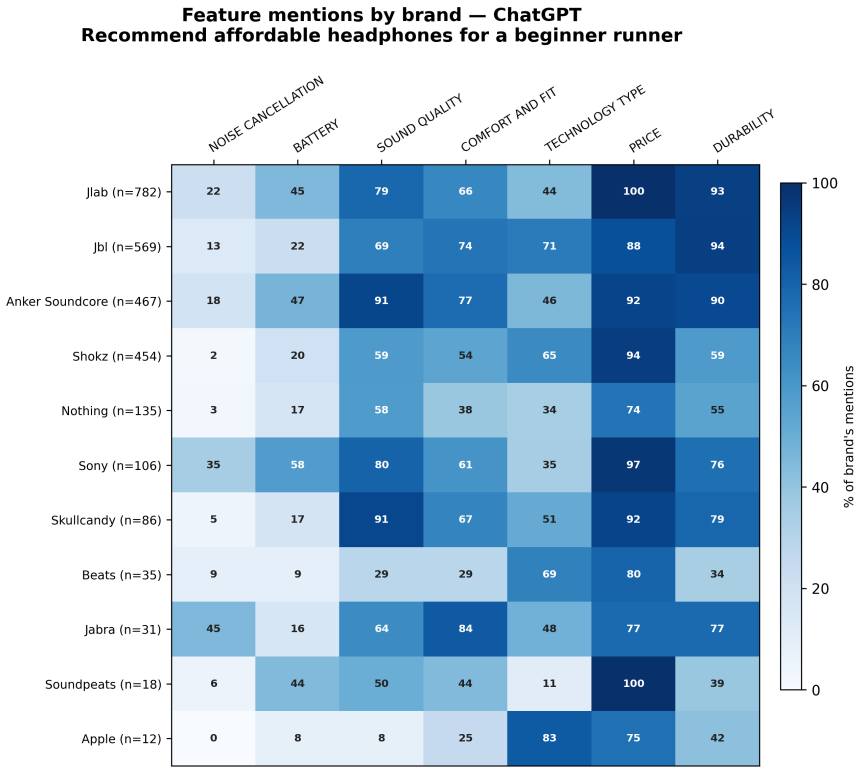


(a) ChatGPT

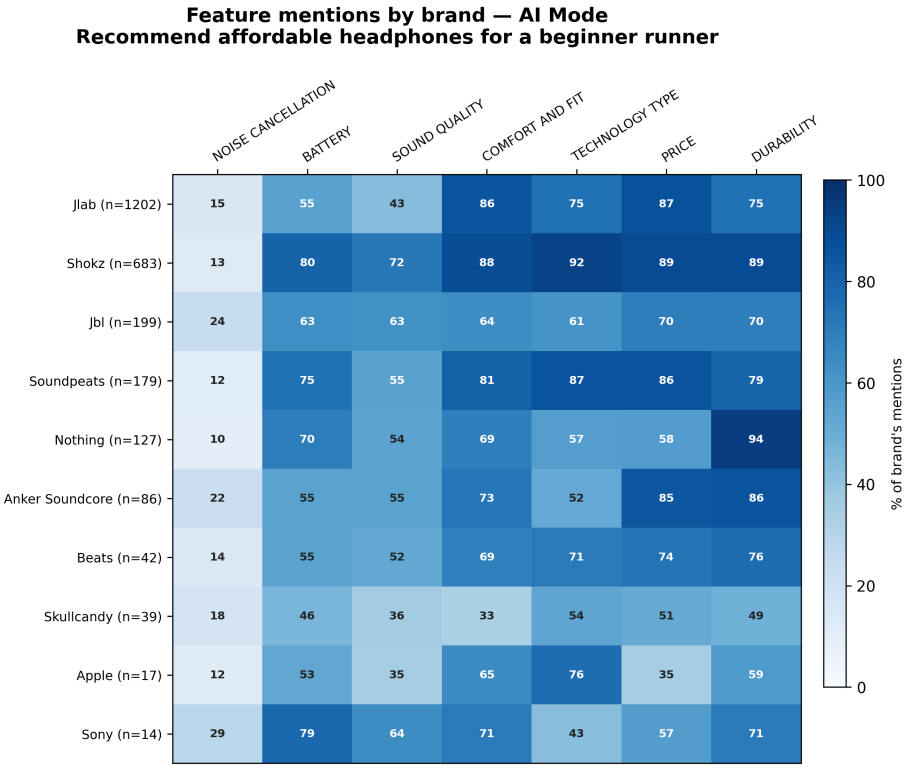


(b) AI Mode

Figure OA.5: Percent of Responses Containing Each Top-10 Brand by Prompt- Best Headphones.

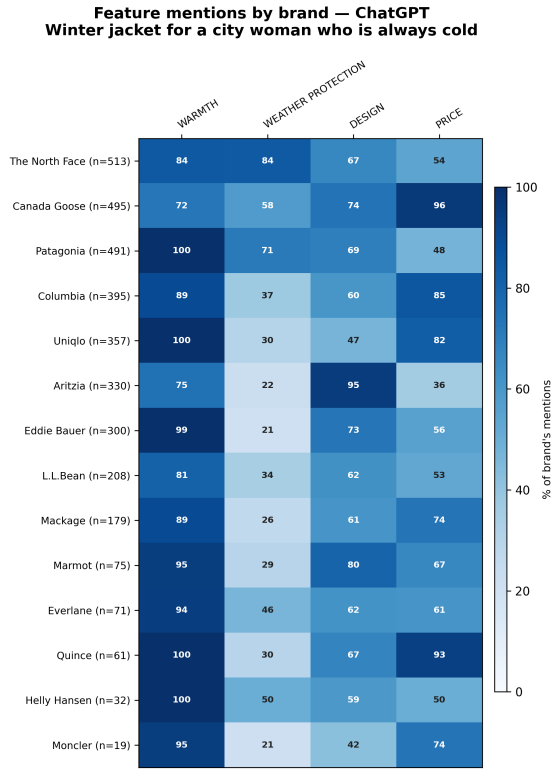


(a) ChatGPT

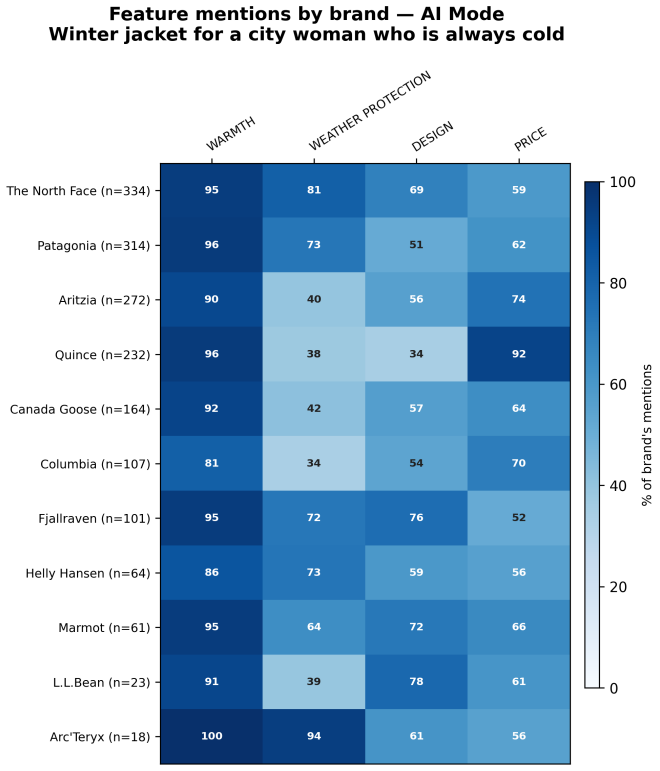


(b) AI Mode

Figure OA.6: Percent of Responses Containing Each Top-10 Brand by Prompt- Runner Headphones.

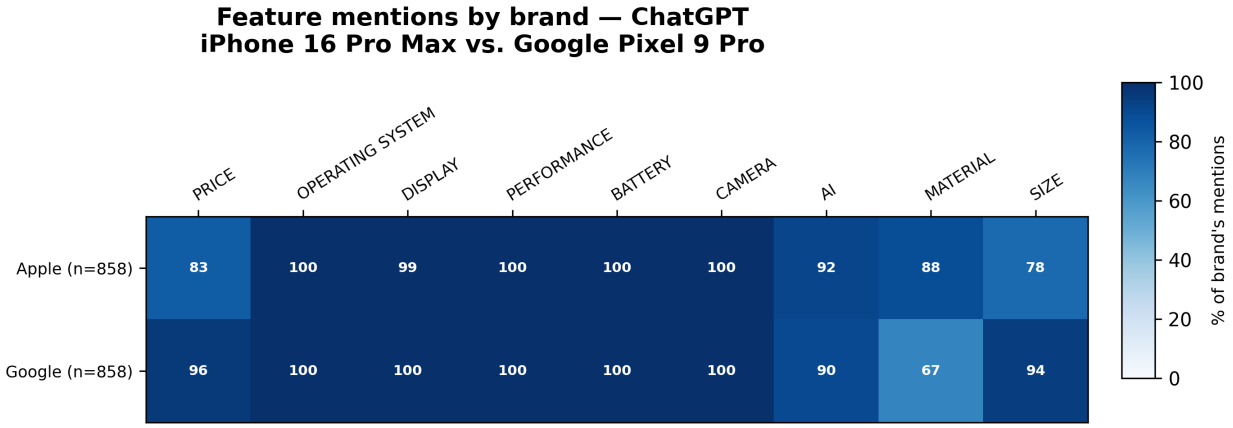


(a) ChatGPT

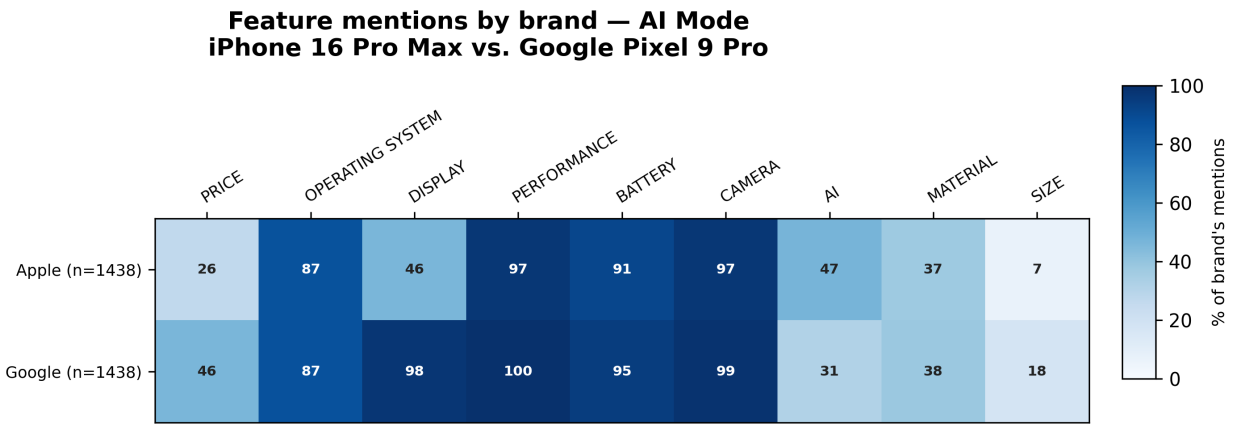


(b) AI Mode

Figure OA.7: Percent of Responses Containing Each Top-10 Brand by Prompt- Winter Jacket.

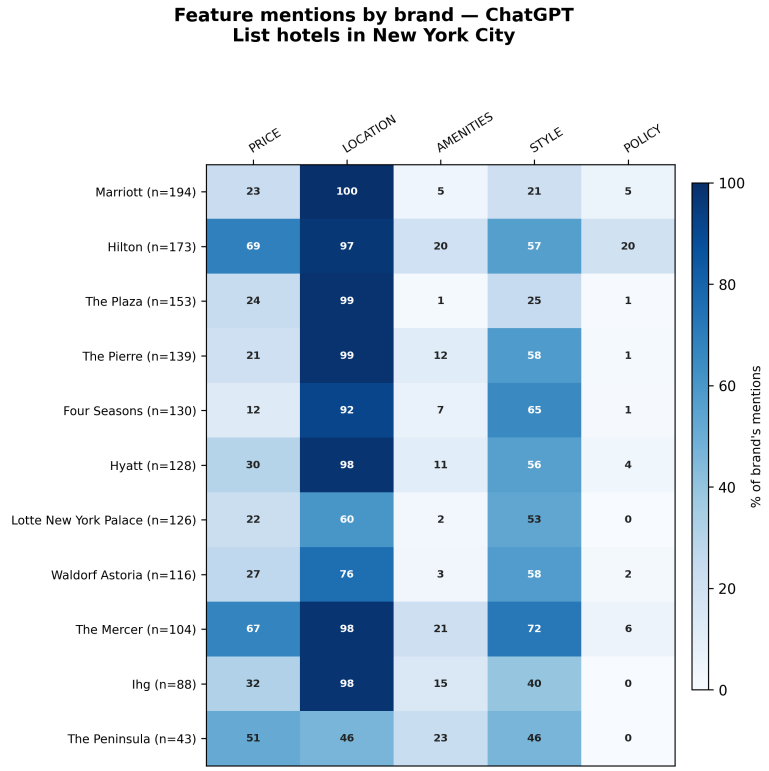


(a) ChatGPT

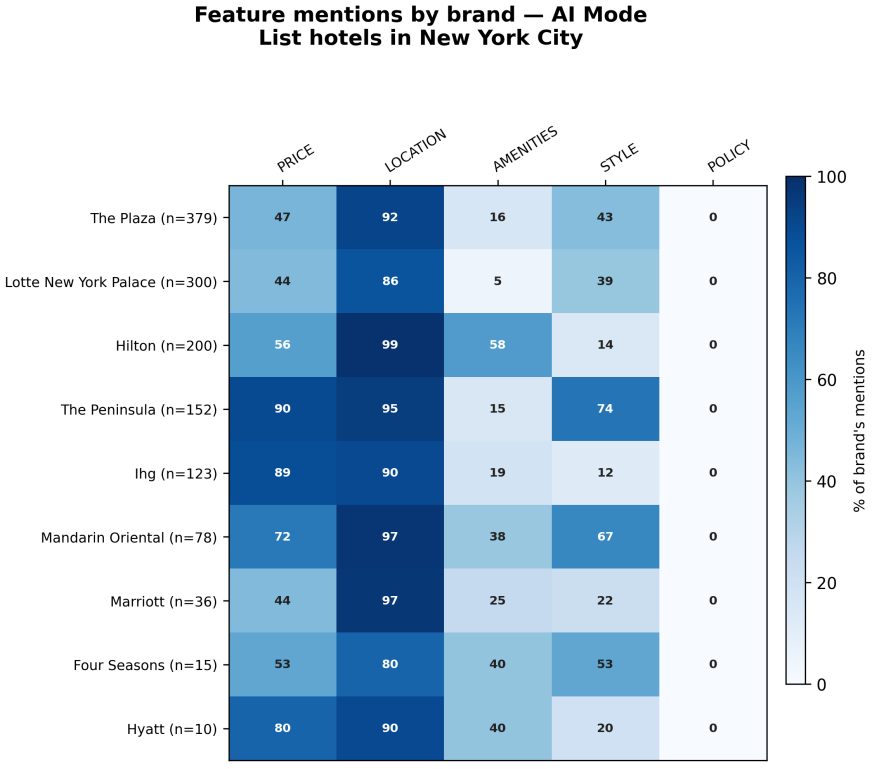


(b) AI Mode

Figure OA.8: Percent of Responses Containing Each Top-10 Brand by Prompt- iPhone Compare.

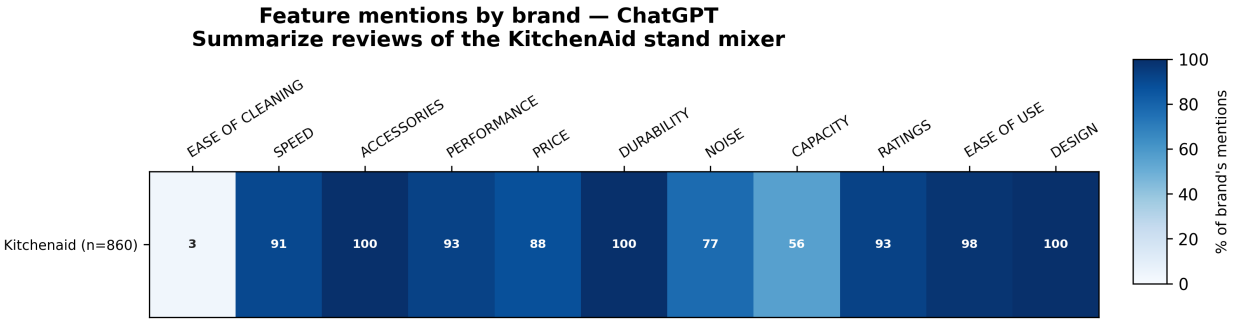


(a) ChatGPT

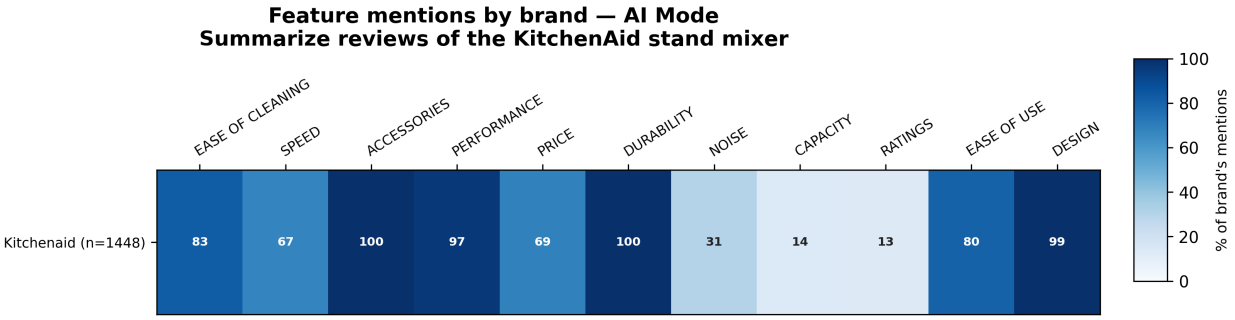


(b) AI Mode

Figure OA.9: Percent of Responses Containing Each Top-10 Brand by Prompt- Hotels.



(a) ChatGPT



(b) AI Mode

Figure OA.10: Percent of Responses Containing Each Top-10 Brand by Prompt- Mixer.

OA.4 Best-for Phrases

Table OA.3: Full list of 61 best-for phrases. Matching is case-insensitive and applied on word boundaries; parentheses and | denote regular-expression alternatives.

Phrase	Phrase
if you are someone who	fit you
if you're someone who	fits you
if you are	fit what you
if you	"best"
if you're	whether you
if you like	whether you're
if you love	When choosing
if you enjoy	highlights
if you want	key benefits
if you need	best for
if you prefer	ideal for
if you often	perfect for
if you plan	great for
if you value	top for
for those	recommended for
for those who	for fans? of
for anyone who	to suit (your|various|different)
for people who	based on your (preference|interest|taste|genre|needs?|budget)
ideal if	tailored (to|for) (your|you)
perfect if	catering to
great if	no matter (your|what your)
recommended if	whatever your
best if	for every (taste|reader|traveler|viewer|genre|style)
you care	something for (every|everyone|all)
depends on	is often praised
to you	is praised
matter	praised for
matters	is known for
for you	known for
are you looking for	good option if
tell me	