Estimating Consumer Preferences for LLMs: Evidence from LMArena

El Hadi Caoui
U of Toronto

December 26, 2025*

Preliminary draft

Abstract

This paper studies product differentiation in the consumer market for large language models (LLMs) using data from randomized pairwise comparisons of LLM responses. I estimate a differentiated-product demand model in which consumer taste parameters vary with prompt embeddings. Perceived LLM quality is strongly associated with an LLM's benchmark performance, context window size, and reasoning ability. Taste heterogeneity across consumers is significant and arises primarily over the stylistic and syntactic features of responses (e.g., readability, verbosity, tone, and formatting), along which LLMs systematically differ. Counterfactual analysis shows that, between 2024 and 2025, consumer surplus per user increased by 38%–54% (equivalently, $37.4–$52.9 per user-month). For the average user, improvements in model intelligence account for 61% of this welfare gain, with an even larger share for technical tasks (71% for coding).

Keywords: large language model, demand estimation, random coefficients, mixed logit

JEL Classification: L11, L13, L86

*elhadi.caoui@rotman.utoronto.ca. Department of Management, University of Toronto Mississauga and Rotman School of Management. 105 St George St, Toronto, ON M5S 3E6. I am grateful to Victor Aguirregabiria, Heski Bar-Isaac, Kevin A. Bryan, Joshua Gans, Avi Goldfarb, Matt Mitchell, Regina Seibel, and seminar participants at the University of Toronto for helpful comments. The analysis, findings, and conclusions expressed in this paper should not be attributed to Arena Intelligence, Inc (LMArena). All errors remain my own.

1 Introduction

Since the release of ChatGPT in November 2022, large language models (LLMs) have diffused rapidly and become widely used across a broad range of everyday activities. By mid-2025, hundreds of millions of people were using LLM-based services each week and exchanging billions of messages per day, indicating substantial user engagement. Survey evidence suggests that users rely on these models for a diverse set of tasks, spanning professional uses such as coding and research to non-work activities, such as information retrieval and writing assistance.

Despite this growth, little is known about how consumers evaluate and select among available LLM products. At face value, these models appear differentiated along both vertical dimensions—such as core capabilities in reasoning or coding—and horizontal ones related to writing style, personality, or tone. Anecdotal evidence suggests that many users have strong preferences for particular models and may continue to use them despite the availability of more capable alternatives.[1] Empirically, however, it remains unclear how consumers trade off these vertical and horizontal attributes or how changes in specific product features shape demand. Understanding the roles of improvements in model intelligence and alignment with human preferences is central to evaluating the nature of competition and the determinants of consumer surplus in this industry.

This paper quantifies the role of product differentiation in consumer choice and welfare by leveraging a novel revealed-preference dataset from LMArena (Chiang et al. (2024)), a public platform for LLM evaluation. On this platform, users submit prompts, which are assigned to two randomly sampled and anonymized LLMs (in a "battle"), users vote for the preferred response, after which model identities are revealed. These comparisons provide direct evidence on how consumers value differences in model outputs, offering a rare window into preferences in the consumer market for LLMs, where typically usage and conversation data are proprietary.

Using LMArena choice and conversation data from 2024 and 2025, I first document several stylized facts about systematic differences in LLM outputs. Although it is well understood that models differ in their core capabilities and overall intelligence, I find that models also differ meaningfully along other dimensions, such as response formatting, readability (e.g., word choice and sentence structure), and tone. For instance, readability measures indicate that OpenAI's models have become progressively easier to read over time, whereas Anthropic's have become more complex. Similarly, response assertiveness has increased for

[1] For example, OpenAI's GPT-5.1 release note (OpenAI (2025)) describes the model as "warmer by default and more conversational," while Anthropic (2025) emphasizes that "each Claude model has a unique character" and that some users find older versions more compelling despite improved capabilities.

OpenAI and Google models over time, while remaining relatively stable for Anthropic. These patterns highlight persistent and evolving differences across providers that are not reflected in headline benchmark scores.

The paper proceeds to build a structural demand model for LLMs in two main steps. First, I model voting choices in LMArena battles using a discrete choice framework with random utilities over LLM *responses*. User votes allow me to estimate how consumer utilities depend on response-level characteristics. In a second step, utilities over LLM responses are aggregated to define utilities over products (i.e., LLMs). Each user chooses products based on the expected utility over responses to their prompt distribution. Finally, I construct measures of consumer surplus and decompose the contribution of improvements in model intelligence from that of horizontal attributes—such as tone, readability, or formatting—in driving changes in surplus over time.

To recover consumers' valuations for LLM responses, I specify a random coefficient discrete choice model for LMArena battles. One important departure from the literature is that I allow taste parameters to vary not only across consumers (as is standard in random coefficient discrete choice models; e.g., Berry et al. (1995)) but also across prompts (i.e, user queries): the mean of the random coefficients is a function of a low-dimensional representation of the prompt vector embedding. This allows the model to capture rich heterogeneity in preferences that may vary across use cases (e.g., coding, writing) and even within use case (e.g., writing a long response email vs. a structured output with tables and lists), while keeping the model tractable. The paper contributes to the growing literature incorporating unstructured data in demand estimation (Compiani et al. (2023), Magnolfi et al. (2025)) by proposing a novel approach for dimensionality reduction of vector embeddings based on Partial least squares (PLS) regression.

Estimation results reveal that, on average, consumers prefer responses by more recently released, larger (by number of parameters), and reasoning models. Additionally, consumer preferences for LLM responses are closely aligned with benchmark performance: consumers tend to favor responses generated by models that score highly on benchmarks. Conditional on benchmark scores, however, reasoning models are on average less preferred, reflecting disutility from slower response speed. Consumers systematically prefer responses with positive sentiment (relative to neutral) and strongly dislike responses with negative sentiment or tentative responses.

Preferences vary substantially across prompts, underscoring the importance of use-case heterogeneity. Linguistic complexity, for example, is positively valued in technical tasks such as mathematics and coding, but is negatively associated with utility in most prompts involving practical guidance. The distribution of estimated coefficients further highlights key tradeoffs across benchmark performance: responses by LLMs with higher knowledge benchmark scores (e.g., MMLU-Pro, HLE, and GPQA Diamond) are consistently valued across all prompts. By contrast, the utility from responses by LLMs with higher coding benchmark scores (e.g., LiveCodeBench, SciCode) depends on the task: these scores increase utility for math and coding tasks but are negatively associated with utility in writing, casual conversation, and information-seeking prompts. This pattern suggests that gains in coding ability may come at the expense of fluency or alignment in non-technical domains. Finally, formatting features—such as the use of bold text, tables, and structured lists—generate the largest utility gains in information-seeking and practical guidance queries.

In the counterfactual analysis, I use estimated preferences over LLM responses to construct utilities over LLMs and compute consumer surplus for the choice set of models available in the actual market in each period. This aggregation maps response-level utilities into model-level utilities by integrating over a prompt distribution that mirrors real-world query patterns observed on ChatGPT (Chatterji et al. (2025)). I account for potential multi-homing by considering two polar demand specifications. Under *prompt-level multi-homing*, users select the utility-maximizing LLM separately for each query. Under *single-homing*, users commit to a single firm across all queries and choose the firm whose portfolio of LLMs delivers the highest expected utility given the user's prompt distribution. These two extremes bound plausible demand behavior and bracket the resulting welfare estimates. In the counterfactual exercise, I estimate the rise in consumer surplus between 2024 and 2025 due to the change in the choice set of available LLMs, and quantify the shares of this increase that can be attributed to improvements in model intelligence (proxied by benchmark performance, reasoning, and context window) versus better alignment with human preferences over text.

Average consumer surplus per user rose substantially between July 2024 and July 2025. Under prompt-level multi-homing, surplus increased by 38%; under single-homing, the increase was 54%. To express these gains in monetary terms, I anchor consumer surplus in July 2024 using survey-based estimates from Collis and Brynjolfsson (2025), who find that the average U.S. user derived $98 in monthly value from LLM usage in 2024. Using this estimate, the increase in consumer surplus from 2024 to 2025 corresponds to $37.4 per user per month under prompt-level multi-homing and $52.9 per user per month under single-homing.

The composition of welfare gains varies across use cases. In technical domains, most of the surplus growth is attributable to improvements in core capabilities: e.g., for mathematics and coding, 68% and 71% of gains arise from improvements in model intelligence, respectively. These use cases place a premium on accuracy, reasoning, and other dimensions of model intelligence. By contrast, for use cases such as practical guidance and casual conversation, improvements in attributes such as tone, readability, and formatting account for a significantly larger share of welfare gains: for example, 44% and 47% in practical guidance under multi- and single-homing, respectively.

The paper makes three contributions. First, it provides a quantitative model of consumer demand for LLMs, offering a framework for analyzing how users evaluate and substitute among models. As emphasized by Berry and Haile (2021), demand estimation is a necessary foundation for understanding market structure. This is particularly relevant given the growing policy interest in the generative AI sector, with competition authorities in the US, UK, and EU launching inquiries into its value chain (Competition and Markets Authority (2024a), Competition and Markets Authority (2024b), Vestager et al. (2024), Federal Trade Commission (2025)). This paper shows that product differentiation operates along multiple dimensions: model intelligence is the primary driver of demand, but firms also compete by better "aligning" their models to user tastes, particularly in non-technical use cases. By quantifying these trade-offs, the analysis provides a first step for evaluating how product differentiation shapes competition and welfare.

Second, while the empirical setting focuses on chatbot interfaces, the model applies more broadly to human preferences over text generated by AI, which increasingly mediate information search and communication. Third, the model enables quantification of consumer surplus from generative AI, providing a grounded measure of welfare gains from technologies that are widely used but, as of 2025, offered to a large extent at zero prices to consumers.

The rest of the paper proceeds as follows: after a literature review, Section 2 provides background information on the industry. Section 3 presents the data used in the empirical exercise. Section 4 describes the model. Section 5 discusses the estimation strategy and results. Section 6 presents the results of counterfactual exercises, and Section 7 concludes.

Literature Review. This paper contributes to two strands of the literature: the literature documenting the demand and use of LLMs and the IO literature on differentiated product demand estimation, in particular, that incorporating unstructured data (text, images).

First, a growing literature examines the rapid diffusion of LLMs and their use. Survey evidence reveals widespread adoption across education levels, occupations, and other demographic groups (Bick et al. (2024)). Usage-based data highlight substantial heterogeneity in how individuals engage with LLMs. Drawing on large-scale ChatGPT conversations, Chatterji et al. (2025) show that users rely on the tool for both work and non-work tasks, with writing assistance, practical guidance, and information-seeking comprising the majority of interactions. While work-related messages have steadily increased, non-work-related usage has grown even more rapidly, rising from 53% to over 70% of total messages. Padilla et al.

(2025) find that LLM adoption alters online behavior, leading to declines in traditional search activity and traffic to smaller content sites

A related literature investigates the use of generative AI at work, focusing on tasks, productivity, and labor market impacts. Survey evidence and internal company data (from Microsoft) point to rapid adoption of LLMs in the workplace (Hartley et al. (2025), Tomlinson et al. (2025)). Experimental studies document sizable productivity gains from LLM assistance across various professional settings, including management consulting and customer support (Dell'Acqua et al. (2023); Brynjolfsson et al. (2025b)). Complementing these task-level studies, Eloundou et al. (2023) quantify occupational exposure to LLMs, while Handa et al. (2025), using Claude interaction data, find that AI use is concentrated in writing and software-related tasks and reflects a mix of automation and augmentation. At the labor market level, Brynjolfsson et al. (2025a) document reduced hiring for young workers in AI-exposed occupations.

While existing work emphasizes who uses LLMs and for which tasks, this paper shifts focus to which LLMs consumers choose and why. As non-work usage continues to rise (Chatterji et al. (2025)), consumer-facing interactions represent a growing and economically significant share of overall LLM demand.

The closest papers within this literature are Demirer et al. (2025) and Nagle and Yue (2025), which both study the enterprise-facing LLM market using aggregate API usage data from platforms such as OpenRouter and Microsoft Azure. Demirer et al. (2025) provide indirect evidence of both vertical and horizontal differentiation, with no single model dominating across use cases (i.e., programming and marketing) and significant variation in the demand for intelligence across applications. The authors also find that most firms allocate the vast majority of usage to a single model, with limited persistent multi-homing. Nagle and Yue (2025) shows that closed-source models dominate usage on OpenRouter, with users continuing to select them even when open-source alternatives are cheaper and offer superior performance. This paper complements these two works by studying the consumer-facing segment, where usage patterns, substitution dynamics, and provider prominence differ from the firm-level API market. A key advantage of the LMArena setting is the observation of a large-scale sample of conversations across LLMs and firms, enabling measurement of fine-grained LLM attributes (such as formatting, tone, and sentiment) that shape consumer demand.

Second, the paper builds on the IO literature on demand estimation in differentiated-product markets, including seminal frameworks (Berry (1994), Berry et al. (1995), Nevo (2001), Backus et al. (2021); see also recent reviews: Berry and Haile (2021), Gandhi and Nevo (2021)). Within this literature, the paper contributes specifically to the strand using vector embeddings from unstructured data in demand estimation. Studies within this strand have used embeddings obtained from: e.g., product descriptions and pictures (Compiani et al. (2023), Bach et al. (2024), Bajari et al. (2025), Quan and Williams (2021), Lee (2025)), fonts (Han et al. (2024)), survey data on product similarities (Magnolfi et al. (2025)), and consumers' transactions and search (Armona et al. (2025), Ruiz et al. (2020), Kumar et al. (2020), Gabel and Timoshenko (2022)).

This paper differs from prior work using unstructured data in two ways. First, it incorporates two types of unstructured input per choice (i.e., user prompts and LLM responses) and uses prompt embeddings to capture heterogeneity in user intent and preferences across tasks, by disciplining taste variation in the structural demand model.[2] Second, to obtain a low-dimensional representation of prompt embeddings, I depart from prior work (e.g., Compiani et al. (2023), Bach et al. (2024)) that relies on Principal Component Analysis (PCA). While PCA performs well in narrow domains, it is poorly suited to LLM prompts, which are isotropic and semantically diverse. Instead, I apply Partial Least Squares (PLS), a supervised dimensionality-reduction method better able to extract utility-relevant variation in this setting.[3]

2 Industry Background

This section provides background on the LLM industry, including how LLMs are trained, how firms differentiate their models, and how model performance is evaluated. It also describes the LMArena platform and some of its limitations.

Market definition. LLMs are statistical models that generate text responses to text inputs. They are implemented using deep neural networks based on the transformer architecture and operate on sequences of tokens, which represent sub-word units.

This paper focuses on the consumer-facing segment of the LLM market—as distinct from the enterprise segment—where users interact with models via chat interfaces. The analysis centers on foundation models: general-purpose systems trained on broad corpora and capable of performing diverse language tasks without task-specific re-training. These models are developed by both tech incumbents (e.g., Google, Meta) and new entrants (e.g., OpenAI, Anthropic, DeepSeek). Geographically, the study is restricted to English-speaking consumers and to providers offering public-facing models accessible in North America.

[2] For instance, a prompt like "write a concise Python script" reveals a distaste for verbosity, while "provide a detailed and structured plan for a trip to Paris" reflects a preference for organized responses using lists or headings.

[3] Studies building *predictive* models, rather than structural models, do not need to reduce the dimension of the embeddings. For example, Bajari et al. (2025) feed the embedding input vector (with thousands of dimensions) into a neural network to predict Amazon prices.

LLM development. Training LLMs proceeds in two stages. In the pre-training phase, models learn to predict the next token in large-scale text corpora drawn from books, websites, and other written sources. Post-training adjusts model outputs to better align with human preferences. This second phase typically involves supervised fine-tuning (SFT) on curated question–answer pairs written or rated by human experts, followed by reinforcement learning from human feedback (RLHF), where models are trained to prefer (and generate) outputs ranked more highly by human evaluators. In settings where task performance can be objectively verified (math, coding), post-training may additionally use reinforcement learning from verifiable rewards (RLVR), in which rewards are computed automatically based on correctness or other checkable criteria rather than human judgments. Additional tuning may incorporate safety guidelines and application-specific constraints.

Differentiation levers. LLMs are differentiated through a series of design and training decisions made by developers. Pre-training choices—such as model architecture, scale, and the composition of training data—shape core capabilities such as general knowledge and intelligence.

Post-training decisions, particularly the design of SFT and RLHF, further shape model behavior by selecting the type of conversation data used in post-training and how to weight tradeoffs between helpfulness, safety, and style. For instance, developers can steer models toward concise or elaborate answers, more formal or conversational tone, or stronger preference for cautious versus assertive completions. These attributes emerge from the choice of (post-)training data, prompt formats, human annotator instructions, and reward models used during alignment.[4]

System prompts, fixed instructions provided by the *user* in a conversation, offer a limited degree of control over LLM response. System prompts can shape model persona or verbosity but exert weaker influence on deeper model attributes. Although some platforms allow users to personalize interactions,[5] such adjustments operate within constraints set during pre- and post-training. As a result, key dimensions of differentiation—such as core capabilities, tone, and response style—are to a large extent hardwired by firms during training rather than adjusted dynamically by users.

Various company statements accompanying model releases illustrate these points. Ope-

[4] For instance, OpenAI publishes a "model spec" which provides detailed guidelines about model behavior, personality, and safety. See https://model-spec.openai.com/2025-12-18.html [Last Accessed: 2025-12-23].

[5] For example, OpenAI allows users to customize aspects of interaction with GPT-5, including warmth, enthusiasm, or emoji usage. In the framework developed in this paper, such personalization does not constitute within-model adaptation but instead corresponds to selecting among a discrete set of pre-configured models. From the consumer's perspective, these variants are treated as distinct models part of the firm's portfolio.

nAI's GPT-5.1 release note (OpenAI (2025)) state that "GPT-5.1 Instant, ChatGPT's most used model, is now warmer by default and more conversational," underscoring that shifts in tone arise from developer-level choices. Similarly, Anthropic (2025) observes that "each Claude model has a unique character, and some users find specific models especially useful or compelling, even when new models are more capable." Consistent with this, after releasing GPT-5, OpenAI reintroduced GPT-4o due to sustained user preference for that model's stylistic profile despite GPT-5's superior intelligence. Together, these examples suggest that differentiation is embedded in the training process and that consumers perceive models as being differentiated along multiple dimensions (beyond core capabilities/intelligence).

LLM evaluations. LLM evaluations (or benchmarks) are tests designed to assess and compare model performance across various dimensions such as capability, safety, and alignment. These evaluations influence research priorities, shape perceptions of model quality, and inform strategic decisions by developers.

Traditional benchmarks rely on static problem sets with predefined answers and ground truth labels—such as MMLU, HellaSwag, or GSM8K—while others draw from live, evolving sources like Codeforces competitions. In contrast, preference-based benchmarks compare outputs using human judgments, either on static problem sets (e.g., MT-Bench, AlpacaEval) or through live user interactions, as in LMArena. Static benchmarks face several limitations: they often focus on closed-ended tasks, are vulnerable to data contamination (i.e., problems sets become part of the training data for new models) and overfitting, and struggle to define ground truth for complex, open-ended, or subjective questions. In many real-world settings, users pose open-ended prompts that require reasoning, synthesis, and judgment. For such tasks, human evaluation remains the only reliable standard for assessing helpfulness and quality.

LMArena. LMArena (`lmarena.ai`) is a public platform for human evaluation of LLMs through side-by-side, blind comparisons. Launched in 2023 by researchers at UC Berkeley's SkyLab (Chiang et al. (2024)), it was spun out as Arena Intelligence Inc. in 2025 and has since raised a $100 million seed round from a16z, UC Investments, and Lightspeed.

Users interact with the platform by submitting prompts and voting on paired model responses, in what LMArena terms a battle. After entering a prompt, the user is shown responses from two randomly sampled and anonymized models (model A and B) and has the option to vote for their preferred response. There are four possible battle outcomes: model A wins, model B wins, tie, or both responses are bad. Voting need not occur after one turn: i.e., users can continue the conversation by submitting a follow-on prompt (multi-turn conversation) and cast a vote at any turn after viewing the paired responses. Once a vote is submitted, the model identities are revealed, and the user may start a new conversation with a newly sampled pair of models. By October 2025, the platform had recorded more than 4.2 million such pairwise votes.

LMArena aggregates these votes into public leaderboards using an Elo rating system, enabling incremental score updates from each user vote. The platform maintains multiple leaderboards, including for text, image generation, vision, and search. The empirical analysis in this paper focuses on the text battles, though the framework could in principle be extended to model demand for image-generation or search. LMArena evaluates both publicly released models and, in some cases, pre-release models supplied for private testing; the data used in this paper covers only models available to the public. Since its launch, the platform has become a widely cited benchmark in model announcements and promotional materials, shaping perceptions of model quality across the industry (Google (2025), xAI (2025)).[6]

LMArena offers several advantages for analyzing consumer preferences. First, it directly captures human judgments over model outputs, providing revealed-preference data rather than relying on proxy measures such as benchmark scores. This is important because user preferences are not necessarily aligned with benchmark performance: models that score highly on standardized tasks may not be perceived as more useful or engaging in practice. Second, because different models respond to identical prompts, the data permit credible inference on substitution patterns across models: i.e., there is no selection into use of a particular LLM. Finally, the availability of full conversation transcripts enables the analysis of rich heterogeneity in preferences across both users and prompt types.

Limitations of the LMArena data. A first limitation of the LMArena platform is sample representativeness. The platform attracts a specialized group of users, including researchers, engineers, and technically engaged early adopters, whose preferences may differ from those of the general population. For a given LLM response, these users may be more attuned to subtle differences in quality or formatting, which may bias preference coefficients away from zero, and lead to estimate of welfare gains from improvement in LLM capabilities that are upper bounds. Bounding welfare gains in the context of LLMs would still be valuable.

A second concern is that LMArena users may submit a different mix of prompts and use cases than typical LLM users. This issue is less consequential, as the empirical approach controls for prompt content and conditions on use case. Moreover, when computing counterfactuals such as welfare, I calibrate the prompt distribution to match that of the general

[6] Singh et al. (2025) raises some structural issues in the LMArena leaderboard, including data access disparities that may favor certain large providers. While the machine learning community often treats LMArena rankings as reflecting intrinsic model quality, I interpret the data as measuring revealed human preferences over LLMs. These preferences are informative but may not fully align with objective notions of model capabilities. Indeed, the purpose of this study is to identify human preferences over a range of LLM attributes, which are not limited to core capabilities.

ChatGPT user base using data from Chatterji et al. (2025) (Section 5.2). Finally, it is worth noting that LMArena users are not peripheral: they are likely power users whose activity constitutes a non-negligible share of real-world consumer usage, and whose preferences likely shape model development. The prominence of LMArena rankings in model announcements and product marketing suggests that firms monitor and respond to this segment's feedback.

A third set of limitations arises from the structure of the LMArena choice environment. Users make binary comparisons in a frictionless setting, with no switching costs. These choices do not account for platform-level features such as complements (e.g., integration of Gemini into Google search and Gmail). The absence of prices to some extent restricts the scope of elasticities that can be estimated. Nonetheless, this limitation is mitigated by the fact that currently the vast majority of users in the consumer LLM market are on the free tiers offered by providers.[7]

3 Data and Descriptive Statistics

3.1 Data sources

This paper draws on two primary data sources: LMArena, the platform described in the previous section, and Artificial Analysis, an independent aggregator of LLM benchmarks and capabilities.

LMArena. The main dataset consists of battles from LMArena, covering two time windows: June 2024–August 2024 and April 2025–July 2025.[8] Each observation includes an (anonymized) user identifier, a battle identifier, the two model identities sampled in the battle, the user's vote, and the full conversation text (user prompts and both model responses in all turns of the conversation).[9]

I classify battles into topic categories, following the same classification used by OpenAI's internal conversation classifier and reported in Chatterji et al. (2025) (Section 5.2). The topic categories are: information seeking, practical guidance, writing, technical help (math, coding), casual conversation (referred to as "Self-Expression" in Chatterji et al. (2025)), multimedia, and other. Topic definitions and examples are provided in Appendix Table A1.

From the conversation data, I extract three sets of LLM response-level characteristics. The first are formatting-related variables. These include the number of Markdown headers (section headings); the number of bold text blocks; and the number of ordered and unordered list items. I also record whether the response includes a table or emojis, and measure the

[7] Financial Times (2025) reports that around 5% of ChatGPT users are on the subscription tier.

[8] These datasets are `arena-human-preference-140k` and `arena-human-preference-100k`, they are publicly available at: https://huggingface.co/lmarena-ai [Last Accessed: 2025-12-05].

[9] The LMArena data does not include information on response latency and throughput.

total length of the response in tokens.

The second group of variables pertains to readability and syntactic complexity. I compute two standard readability metrics: the Flesch Reading Ease score and the SMOG Index. The Flesch score ranges from 0 to 100, with higher values indicating more accessible text; it is based on sentence length and the number of syllables per word. The SMOG Index estimates the years of education required to comprehend a text, and is based on the frequency of polysyllabic (three or more syllables) words within a fixed number of sentences. While both metrics aim to quantify readability, they emphasize different linguistic features: the Flesch score is more sensitive to sentence and word length, whereas the SMOG Index emphasizes lexical complexity.

The third group of variables captures model tone. While many stylistic features could be analyzed, I focus on two dimensions: sentiment and hedging. Sentiment is classified into three categories—positive, neutral, and negative—based on the emotional valence and intensity of the LLM's response. Positive sentiment includes favorable expressions (e.g., "insightful question," "incredibly helpful"), while neutral sentiment reflects objective, factual, or informational content without evaluative tone. Negative sentiment reflects unfavorable, critical, or emotionally intense response (e.g., "this is a terrible idea").

Hedging refers to the degree of linguistic caution or tentativeness in a response. I classify tone as assertive when the LLM uses confident, directive language (e.g., "You should do X," "This is the best approach"), tentative when the response includes hedging markers such as modal verbs, qualifiers, or disclaimers (e.g., "This might help," "You could try...", "I don't know"), and neutral when the tone is descriptive without conveying strong confidence or hesitation. Hedging focuses specifically on the LLM's epistemic commitment: how strongly the model asserts or qualifies its claims.

The classification of conversation topics and response-level sentiment and hedging are performed using an LLM as judge. Appendix E provides the full prompt and implementation details used for each classification tasks.

In addition to the battles data, I collect biweekly snapshots of the LMArena leaderboard, which reports the list of models being tested, their providers, and licensing status (open-source, open-weights, proprietary), from June 2023 to July 2025.

Artificial Analysis. To measure model core capabilities, the paper draws on public data from Artificial Analysis, a market research firm that provides standardized benchmarking and related model information. The dataset covers all publicly released LLMs as of July 2025 and includes, for each LLM, its performance scores on a wide range of benchmarks, along with model attributes such as release date, context window size, latency, throughput, and whether the model is a reasoning model.[10]

A key advantage of this data source is its consistent evaluation framework. LLM providers often report benchmark performance for new models under differing conditions—such as varying numbers of demonstrations (n-shot prompting)—which undermines comparability. Artificial Analysis addresses this issue by applying uniform testing protocols across all models tested, enabling credible comparisons of benchmark scores. Appendix Table A2 provides definitions of the benchmarks used. In the remainder of the paper, I use 4 indexes formed by averaging benchmark scores: the "Intelligence index" is a weighted average of all 7 scores (where weights are given by the last column in Table A2), the "Coding index" is an average of SciCode and LiveCodeBench, the "Math index" is an average of MATH-500 and AIME, finally the "Knowledge index" is an average of MMLU-Pro, HLE, and GPQA Diamond.

3.2 Descriptive statistics

Figure 1 shows the cumulative number of LLMs tested on LMArena over time, with the periods for which conversation data are available shaded. By July 2025, 411 LLM from 65 providers have been tested. Providers are major technology firms and commercial AI labs as well as smaller startups and university research groups. Proprietary models are supplied primarily by firms such as Google, OpenAI, Anthropic; open-weights models by organizations including Meta, Google, and Cohere; and open-source models by contributors such as Alibaba, Microsoft, Mistral, and several academic groups.[11] Together, these data capture the full range of LLM development activity across licensing regimes and provider types.

Table 1 shows descriptive statistics of LMArena battles. The dataset contains 241,768 battles, of which 53% are in English, and have on average 1.31 turns per conversation. For battles in English, users vote for one of the two model outputs in 68% of battles, with 17.2% ties and 14.7% rated both bad. The distribution over conversation topics is shown in Panel C: the top 3 topics are information seeking, writing, and practical guidance. This distribution is broadly in line with the most popular topics reported by Chatterji et al. (2025) for a large representative sample of ChatGPT users. However, as expected, math and coding prompts are more frequent on LMArena (21%) compared to the overall population of ChatGPT users,

[10] The context window size is the maximum number of tokens a model can process in a single prompt–response interaction. Latency measures the time required to generate the first token of output. Throughput captures the sustained generation speed (tokens per second). A reasoning model is one explicitly optimized, typically through SFT and RL, for multi-step reasoning and problem-solving tasks. These models generate "reasoning" tokens before outputing a response and are therefore slower than non-reasoning models.

[11] For open-weights models, model weights (parameters) are released for use subject to a license, whereas open-source models make both the weights and the full training and implementation code available under an open-source license.

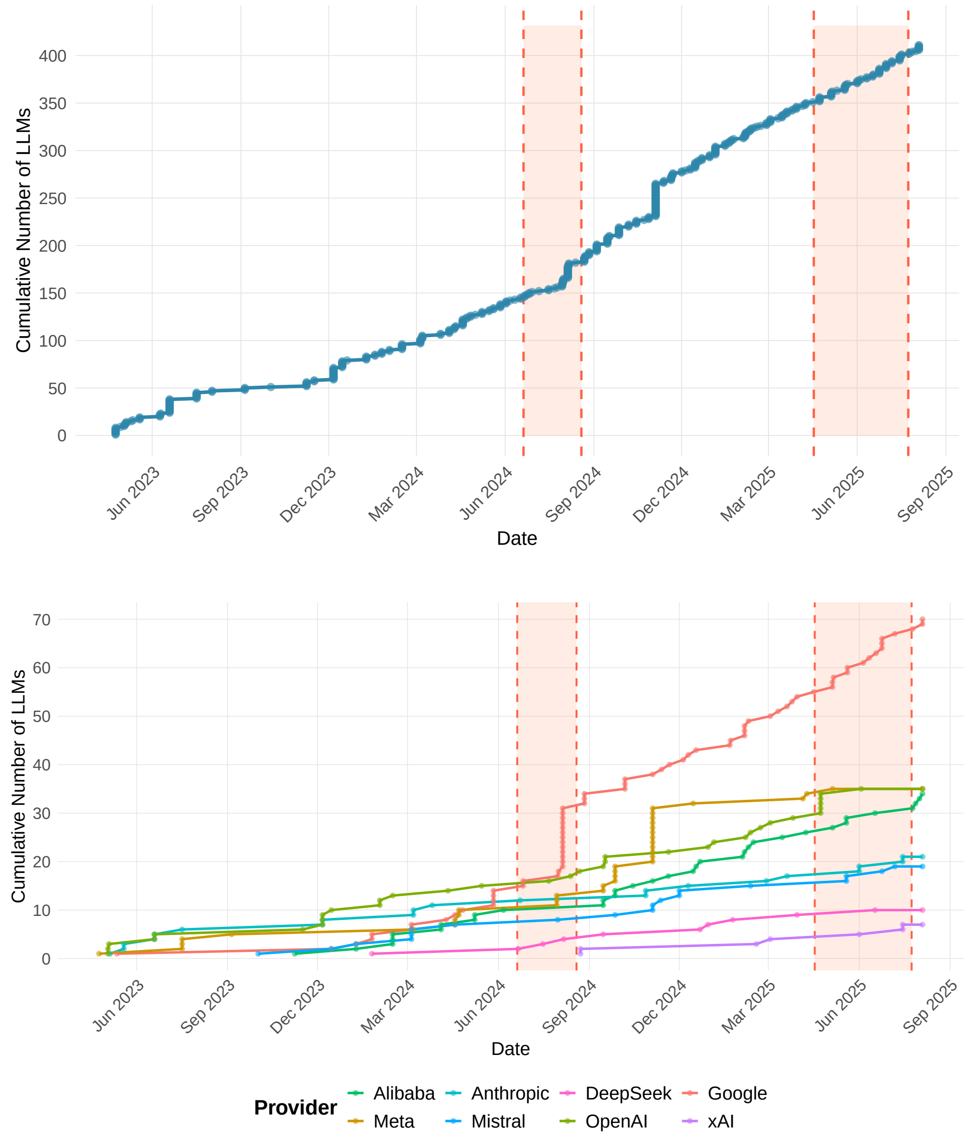


Figure 1: Cumulative number of LLMs on LMArena (top shows total, bottom shows the top 8 providers)

where such prompts drop from 12% to 5.1% of conversations between 2024 and 2025.

Table 2 shows the distribution of LLM capabilities in the Artificial Analysis dataset. Benchmark-based performance indices span a wide range: intelligence scores range from 8.3 to 73.2, coding from 0.1 to 63.8, and math from 2.9 to 96.7. The context window sizes vary from 2,000 to 10 million tokens, and about 28% of models are classified as reasoning models. Figure 2 plots the highest intelligence index across LLMs for each of the top 8 providers: OpenAI's models scored higher on benchmark metrics through mid-2024, but providers such

Table 1: Descriptive Statistics: Battles

	Jun 2024 – Aug 2024	Apr 2025 – Jul 2025	All
Panel A: Basic Statistics			
Number of battles	106,134	135,634	241,768
Number of users/sessions	49,382	115,372	164,754
Number of LLMs	55	52	103
Number of providers	17	13	22
Number of battles in English	57,675	71,175	128,850
Share of battles in English	0.543	0.525	0.533
Avg. number of turns per conversation	1.37	1.26	1.31
Panel B: Share of Battle Outcomes (English Battles Only)			
Model A wins	0.307	0.359	0.336
Model B wins	0.318	0.366	0.345
Tie	0.187	0.160	0.172
Both bad	0.188	0.114	0.147
Panel C: Share of Battles by Topic Category (English Battles Only)			
Information Seeking	0.331	0.405	0.369
Writing	0.233	0.141	0.186
Practical Guidance	0.101	0.190	0.147
Coding	0.140	0.131	0.135
Math	0.113	0.041	0.076
Other/Multimedia	0.040	0.054	0.047
Casual Conversational	0.043	0.039	0.041

Note: For the 2025 dataset, users are identified only per session but cannot be tracked across sessions.

as Google, xAI, and others rapidly closed the gap by mid-2025.

Finally, Table 3 provides descriptive statistics for LLM responses in the LMArena dataset. Panel A shows modest readability variation, with an average Flesch Reading Ease score of 46.6 and SMOG index of 12.7. Panel B highlights increased use of formatting elements over time: e.g., the number of text blocks in bold increases from 6.23 to 17.95; the number of bullet points increases from 5 to 14.5, and responses are twice as long in 2025 than in 2024. Panel C shows that response sentiment is predominantly neutral, but the share of positive sentiment responses increases from 12% to 18% of all responses between 2024 and 2025. Assertive responses are also more frequent, whereas neutral and tentative responses are less frequent in 2025 relative to 2024.

To illustrate potential differences in responses across LLMs, I regress the response-level characteristics in Table 3 on calendar quarter, conversation topic category, and LLM fixed effects. Given the market definition used in this paper, I include fixed effects for LLMs developed by private firms that offer consumer-facing products (which excludes enterprise-focused providers such as Cohere) and make their product available in the North American market (which excludes some Chinese providers such as Alibaba). The firms are: Anthropic,

Table 2: Descriptive Statistics: LLM Characteristics

Variable	Mean	SD	Min	Median	Max
Panel A: Benchmark Scores					
MMLU Pro	0.63	0.17	0.14	0.68	0.87
GPQA	0.48	0.17	0.10	0.47	0.88
HLE	0.06	0.03	0.03	0.05	0.24
LiveCodeBench	0.30	0.20	0.002	0.27	0.82
SciCode	0.23	0.11	0.00	0.24	0.47
Math 500	0.72	0.23	0.06	0.77	0.99
AIME	0.28	0.29	0.00	0.15	0.94
Intelligence Index (0-100)	37.83	15.01	8.30	36.65	73.20
Coding Index (0-100)	26.51	14.71	0.10	25.35	63.80
Math Index (0-100)	50.75	24.48	2.90	45.80	96.70
Panel B:					
Context window ('000s tokens)	356.27	840.35	2	128	10,000
Reasoning Model	0.28	0.45	0	0	1

Note: The unit of observed is an LLM. Benchmark scores are between 0 and 1. Benchmark indices are between 0 and 100.

DeepSeek, Google, Meta, Mistral, OpenAI, and xAI. LLMs by providers outside this list are labelled as "other" and serve as the omitted reference category.[12]

Figure 3 plots LLM fixed effects for two formatting-related outcomes: response length and table usage. More recent models tend to produce longer responses on average, with notable variation across providers. xAI and Google models yield the lengthiest outputs, followed by OpenAI. By contrast, Anthropic models are consistently associated with shorter responses. Table usage is even more heterogeneous. OpenAI's reasoning models (o3 and o4-mini) stand out, including tables in roughly 30-40% of their responses—substantially higher than Gemini 2.5 Pro from Google (24%) and significantly above models from Anthropic, which rarely format content using tables.

Figure 4 shifts focus to readability and tone, using the SMOG index and a measure of positive sentiment. The SMOG index, which approximates lexical complexity, shows diverging trends across firms. Anthropic's newer models score higher, indicating more complex word choices over time, while OpenAI's models display the opposite pattern, becoming easier to read. This may reflect differences in product positioning or targeted user segments. Regarding sentiment, OpenAI, Google, and Meta exhibit an upward trend in positive framing, with newer models increasingly producing upbeat language (e.g., "That's a great question!"). Anthropic models, by contrast, show little movement on this dimension.

[12] In practice, I define two sets of "other" LLMs: those released up to end of the first window (August 2024), and those released after that date.

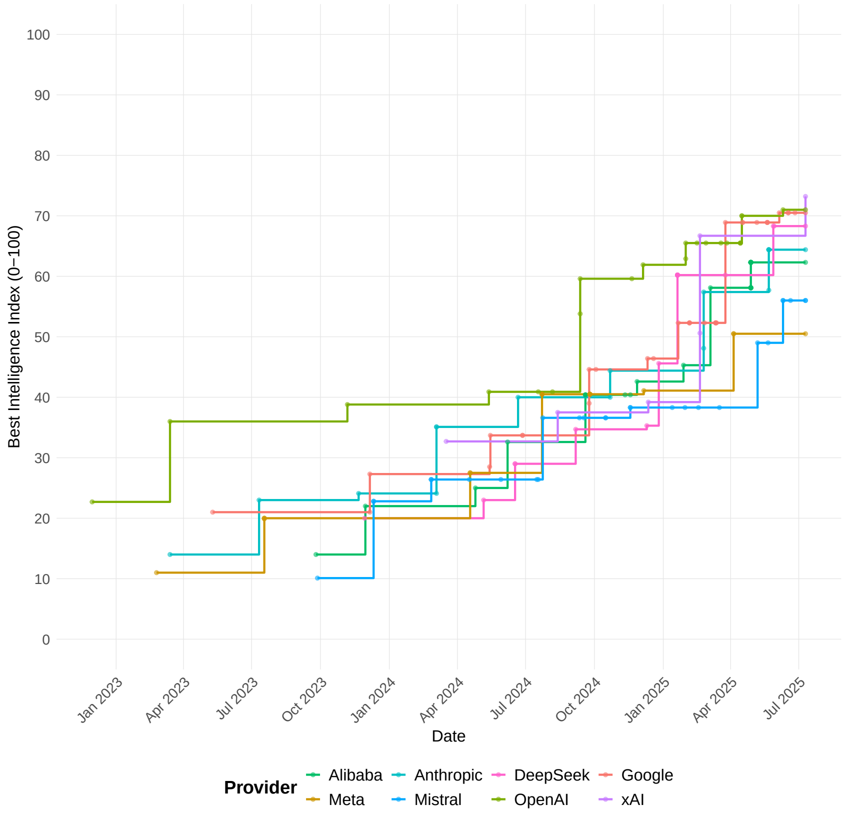


Figure 2: Best Intelligence index (over portfolio of LLMs) by provider over time

These patterns are not restricted to the four outcomes presented in these figures. Appendix Figure A1 shows similar degrees of heterogeneity for other outcomes: e.g., markdown headers and bold text. Overall, these patterns illustrate meaningful LLM-level differences in response formatting, tone, and readability, with variation occurring both across firms and across LLMs for a given firm.

4 A Choice Model for LMArena Battles

This section develops a random-coefficient discrete choice model tailored to the institutional structure of LMArena battles. In each interaction, a user submits a prompt and evaluates two competing LLM responses. The model maps these observed choices into latent utilities over responses, decomposing utility into response-specific attributes and model-level capabilities. By allowing preferences to vary systematically with prompts and across users, the framework provides a disciplined way to recover latent LLM quality and to quantify how different dimensions of response attributes shape user valuations.

Setup. A user i submits a text prompt p to the LMArena platform. Each prompt p is

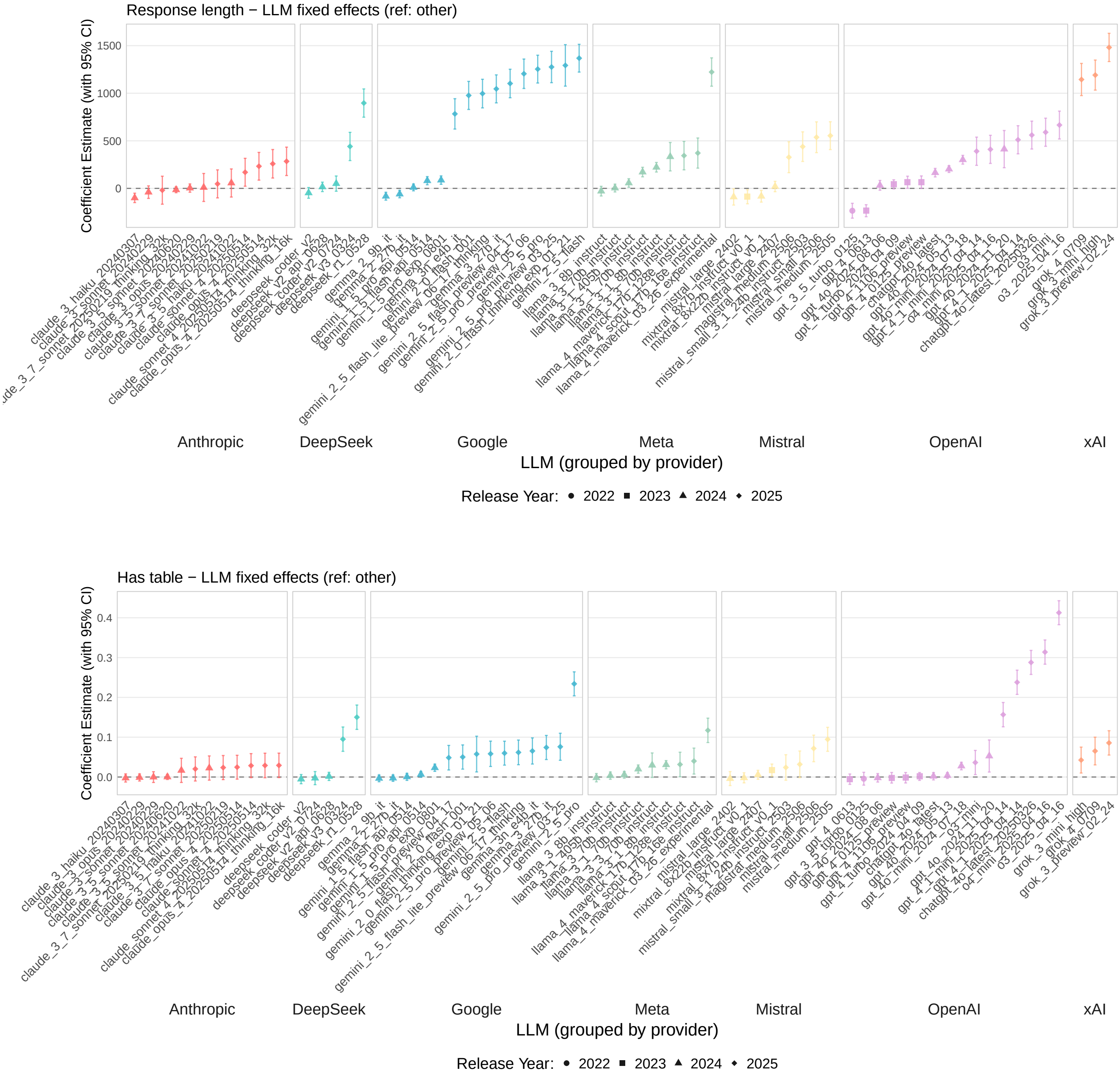
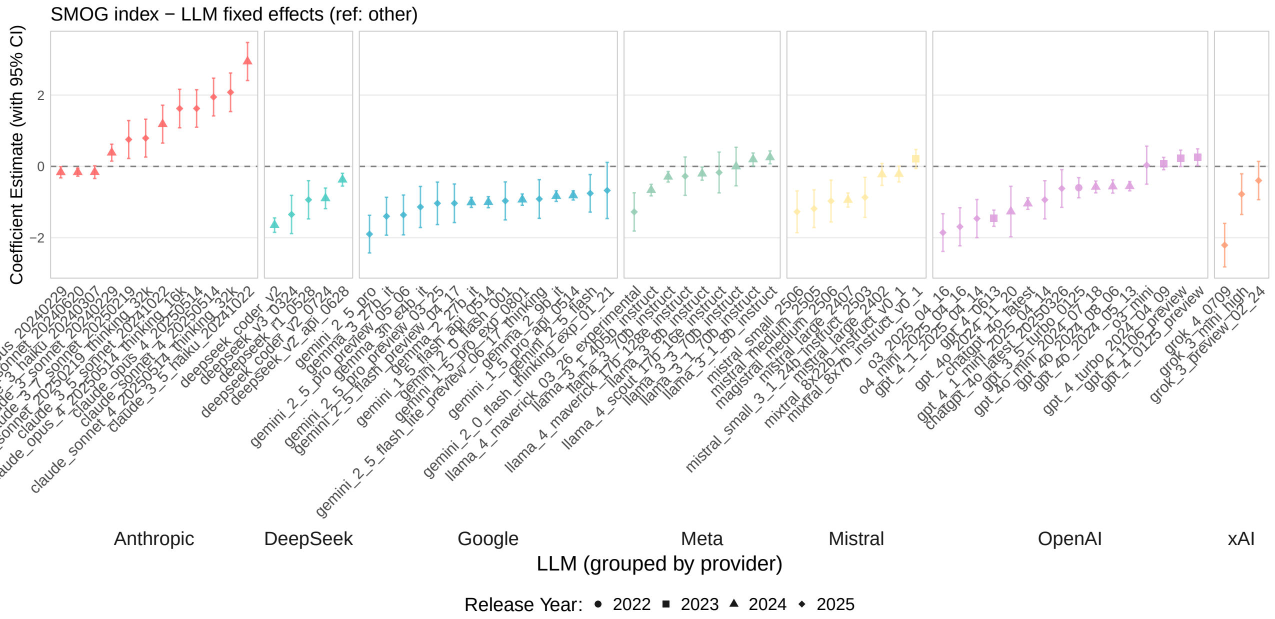


Figure 3: Formatting characteristics by LLM: Response length (tokens) and use of tables



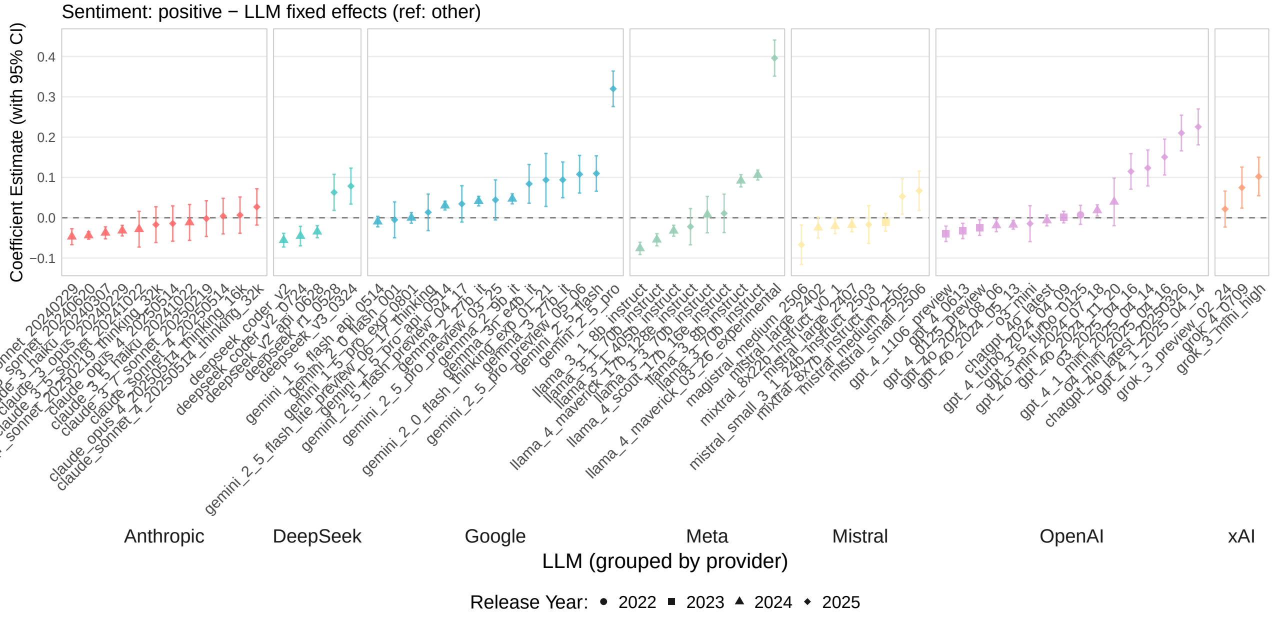


Figure 4: Readability and tone by LLM: SMOG index and positive sentiment

Table 3: Descriptive Statistics: LLM Responses

	Jun 2024 – Aug 2024		Apr 2025 – Jul 2025		All	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Panel A: Readability						
Flesch Reading Ease (0-100)	49.17	(94.03)	44.16	(50.52)	46.59	(74.91)
SMOG Index (years)	12.38	(4.39)	12.96	(4.08)	12.68	(4.24)
Panel B: Formatting						
Markdown headers	0.72	(4.03)	3.39	(8.98)	2.09	(7.16)
Bold text	6.23	(15.27)	17.95	(31.99)	12.26	(25.96)
Ordered lists	3.52	(8.25)	4.62	(10.29)	4.09	(9.37)
Unordered lists	5.00	(12.62)	14.50	(27.38)	9.89	(22.04)
Response length (tokens)	562.74	(906.52)	994.63	(1392.98)	785.04	(1201.72)
Has table	0.02	(0.14)	0.10	(0.30)	0.06	(0.24)
Has emoji	0.02	(0.13)	0.15	(0.36)	0.09	(0.28)
Panel C: Tone						
Sentiment: positive	0.12	(0.32)	0.18	(0.38)	0.15	(0.36)
Sentiment: negative	0.02	(0.14)	0.01	(0.10)	0.02	(0.12)
Sentiment: neutral	0.86	(0.34)	0.81	(0.39)	0.84	(0.37)
Hedging: assertive	0.42	(0.49)	0.48	(0.50)	0.46	(0.50)
Hedging: tentative	0.16	(0.37)	0.14	(0.34)	0.15	(0.36)
Hedging: neutral	0.42	(0.49)	0.38	(0.48)	0.40	(0.49)

Note: Descriptive statistics for response-level features from the LMArena dataset. Flesch Reading Ease: Higher scores indicate easier-to-read text (0-100). Components: words per sentence and syllables per word. SMOG Index: Estimates years of education needed to understand the text. Components: count of polysyllabic (≥3 syllables) words. Higher implies harder to read. Bold text corresponds to the number of blocks of text in bold.

mapped into a d-dimensional embedding vector $\mathbf{E}_p \in \mathcal{P}$, where $\mathcal{P} \subset \mathbb{R}^d$. For each user i, there exists a user-specific probability measure F_i on $\mathcal{P}$ (and associated σ-algebra) describing the distribution of that user's prompt embeddings, and $\mathbf{E}_p$ represents a random draw from the distribution F_i. Two LLMs m_A and m_B generate responses. The resulting "battle" is a tuple

$$(i, p, \{m_A, m_B\}, o),$$

where the outcome $o \in \{m_A, m_B, \emptyset\}$ is determined by

$$o = \begin{cases} m_A & \text{if } U_{ipm_A} > U_{ipm_B} \text{ and } U_{ipm_A} > U_{ip0}, \\ m_B & \text{if } U_{ipm_B} > U_{ipm_A} \text{ and } U_{ipm_B} > U_{ip0}, \\ \emptyset & \text{if } U_{ipm_A} < U_{ip0} \text{ and } U_{ipm_B} < U_{ip0}, \end{cases}$$

with U_{ipm} denoting user i's utility for LLM m's response to prompt p and U_{ip0} the outside-option utility (an acceptability threshold, corresponding to the vote "both bad").[13] The baseline framework ignores ties, Appendix A shows how to accomodate ties within this framework.

Utility specification. The utility a user derives from a response has two distinct components: (i) a component that varies at the *response level*, reflecting observable features of the particular output generated in response to prompt p, and (ii) a component that varies at the *LLM level*, reflecting core capabilities of model m that shape the semantic quality of all its responses. Response-level characteristics X_{pm} capture measured attributes of the specific output—formatting, readability, and tone. In contrast, semantic attributes such as correctness, relevance, and completeness are modeled as functions of the capabilities of the LLM, summarized by the model-level characteristics X_m (which includes the LLM's benchmark scores, context window size, reasoning ability, and a constant). User i's utility from model m's response to prompt p is specified as

$$U_{ipm} = \beta_i^R(p)' X_{pm} + \beta_i^L(p)' X_m + \xi_m + \varepsilon_{ipm}, \tag{1}$$

where ξ_m captures unobserved LLM quality, and ε_{ipm} is an idiosyncratic shock distributed according to a type-1 extreme value distribution. Taste parameters $(\beta_i^R(p), \beta_i^L(p))$ vary across *users* and *prompts* through random coefficients, with a mean that is a function of the user prompt embedding $\mathbf{E}_p$, and a variance-covariance matrix to be estimated, as follows:

$$\begin{pmatrix} \beta_i^R(p) \\ \beta_i^L(p) \end{pmatrix} = \begin{pmatrix} \beta_0^R + \beta_1^{R\prime} h(\mathbf{E}_p) \\ \beta_0^L + \beta_1^{L\prime} h(\mathbf{E}_p) \end{pmatrix} + \Sigma \nu_i, \qquad \nu_i \sim \mathcal{N}(0, I_K), \tag{2}$$

where K is the dimension of the observed characteristics vector $[X_{pm}\ X_m]$, $h(.)$ is a function mapping $\mathbb{R}^d$ into $\mathbb{R}^g$ (with $g < d$) that extracts a low-dimensional representation of the prompt embedding, Σ is a diagonal scaling matrix with element σ_k representing the dispersion in user-specific taste shock for characteristic k. (β_0^R, β_0^L) is a $K \times 1$ vector, (β_1^R, β_1^L) is a $K \times g$ matrix. Each characteristic k in $[X_{pm}\ X_m]$ is associated with a mean coefficient $\beta_{0,k}^R$ and a set of prompt-level interactions $\beta_{1,k}^R$, which capture the dependence of taste parameters on the prompt p. This specification ensures that prompts that are closer in the embedding space will have similar mean taste parameters.

[13] Operationally, the "both bad" outcome corresponds to the event that neither model's output exceeds this intrinsic threshold. The threshold may reflect the quality the user expects from a general search engine (Google) or a minimal notion of usefulness, completeness, or correctness. Although the threshold can vary over time (for example, through time effects in the empirical specification), it is treated as exogenous to the particular pair of models $\{m_A, m_B\}$ shown in a given battle.

To control for the unobserved quality term ξ_m, it is useful to define the following LLM-specific fixed effect parameter

$$\delta_m = \beta_0^{L\prime} X_m + \xi_m \tag{3}$$

The term δ_m represents the average quality of model m that is common across consumers. In the baseline model, the mean of the outside option is normalized to zero: $U_{ip0} = \varepsilon_{ip0}$. Appendix B shows that the results are robust to a time-varying outside option (e.g., to control for shifts in user expectations over time).

Choice Probabilities. Conditional on the random taste shock ν_i, the probability that user i votes for model m_A's response is

$$P(o = m_A \mid p, \{m_A, m_B\}, \nu_i) = \frac{\exp\left(\beta_i^R(p)' X_{pm_A} + \beta_i^L(p)' X_{m_A} + \xi_{m_A}\right)}{1 + \sum_{k \in \{m_A, m_B\}} \exp\left(\beta_i^R(p)' X_{pm_k} + \beta_i^L(p)' X_{m_k} + \xi_{m_k}\right)}. \tag{4}$$

The unconditional probability integrates over the distribution of user heterogeneity:

$$P(o = m_A \mid p, \{m_A, m_B\}) = \int P(o = m_A \mid p, \{m_A, m_B\}, \nu_i)\, d\Phi(\nu_i). \tag{5}$$

where Φ is the cdf of the standard multivariate normal distribution. The choice probabilties for the other outcomes can be similarly defined. In anticipation of the counterfactual analysis, note that a user i is characterized by (i) a distribution over prompts $\mathbf{E}_p \sim F_i$ and (ii) a taste vector ν_i drawn from the population distribution. The framework is general enough that, in principle, one could allow the distribution of prompts to depend on the consumer taste shock, i.e., $F_i = F(.|\nu_i)$. This would capture, for instance, the fact that consumers who use LLMs mostly for coding tasks have a higher valuation for LLM performance on coding benchmarks compared to the average user.

5 Estimation

5.1 Identification

The outcome probabilities (Equation (5)) conditional on the prompt and model pair are identified nonparametrically. These win rates constitute the fundamental choice probabilities that the structural model must rationalize.

Persistent preference heterogeneity across users enter utility through a random-coefficients structure. The diagonal matrix Σ governs the variance of these taste shocks across dimensions of response-level and model-level attributes. Identification of Σ relies on the fact that heterogeneous preferences generate systematic violations of the independence of irrelevant alternatives (IIA) that cannot arise under homogeneous tastes. Appendix C provides a concrete example of how consumer heterogeneity is identified from violations of IIA substitution patterns.

Prompt embeddings $\mathbf{E}_p$ vary exogenously across battles and enter the taste parameter coefficients through the reduced-dimensional representation $h(\mathbf{E}_p)$. Identification of prompt-level heterogeneity in taste parameters (the terms: $\beta_1^{R\prime} h(\mathbf{E}_p)$ and $\beta_1^{L\prime} h(\mathbf{E}_p)$) exploits systematic variation in battle outcomes across prompts. For example, the ability to produce structured output (table, bullet point lists) may matter more for prompts related to information-seeking or practical guidance than for creative writing prompts. Since prompts are assigned randomly to model pairs, this exogenous variation traces out the importance of a particular characteristic in explaining user choice, and identifies the parameters β_1^R and β_1^L.

5.2 Estimation Approach

The goal of estimation is to recover the parameters governing response-level preferences, model-level preferences, and the distribution of random taste heterogeneity, $(\beta_0^R, \beta_1^R, \beta_0^L, \beta_1^L, \Sigma)$, as well as the LLM-specific mean utilities $\{\delta_m\}$, defined in Equation (3). Estimation proceeds in two steps.

Step 1: Simulated Maximum Likelihood. For each user, the likelihood contribution integrates over the individual-specific random coefficients. Let b index battles faced by user i, and let ν_i denote the vector of random taste shocks entering through the random-coefficients specification. The exact (log-)likelihood contribution for user i is

$$LL_i = \ln\left(\int \prod_b P(o \mid b; \nu_i)\, d\Phi(\nu_i)\right),$$

where $P(o \mid b; \nu_i)$ is the model-implied probability of the observed outcome in battle b given ν_i. Because the integral does not admit a closed form, it is approximated by simulation.[14] Drawing $\{\nu_i^r\}_{r=1}^R$ independently from the multivariate normal distribution, the simulated log-likelihood for user i is

$$SLL_i = \ln\left(\frac{1}{R}\sum_{r=1}^{R} \prod_b P(o \mid b; \nu_i^r)\right).$$

Summing over all users yields the simulated log-likelihood function. Maximizing this function produces estimates $(\widehat{\beta}_0^R, \widehat{\beta}_1^R, \widehat{\beta}_1^L, \widehat{\Sigma}, \{\widehat{\delta}_m\}_m)$.[15] The $\widehat{\delta}_m$ terms can be interpreted as model

[14] The simulated LL uses 250 Sobol sequence draws. Robustness checks are conducted in Appendix **??**.

[15] A common concern with simulated maximum likelihood in discrete choice panel data models is that the joint likelihood may involve extremely small choice probabilities when the number of alternatives is large (Berry and Haile (2021)). This issue does not arise in the LMArena setting, where each battle involves only

fixed effects capturing average performance after accounting for response-level attributes and heterogeneity in user tastes.

Step 2: Generalized Least Squares. The second step estimates β_0^L by exploiting the relationship

$$\delta_m = \beta_0^{L\prime} X_m + \xi_m.$$

Given estimates $\widehat{\delta}_m$ from Step 1 and their estimated covariance matrix V_δ, the coefficient vector β_0^L is obtained by generalized least squares:

$$\widehat{\beta}_0^L = (X_m' V_\delta^{-1} X_m)^{-1} X_m' V_\delta^{-1} \widehat{\delta}. \tag{6}$$

This step follows the logic of classical two-step estimators for random-coefficients discrete-choice models, where the first step includes products fixed effects, as in Nevo (2001). The first step recovers consistent estimates of model-specific mean utilities (which control for the unobserved term ξ_m), whereas the second step relates these utilities to model characteristics in a linear specification, improving efficiency by accounting for correlation across the $\widehat{\delta}_m$.

I reserve a 10% hold-out test sample to evaluate out-of-sample performance and compare different specifications, and estimate the model using the remaining 90% of battles.

Choice of the function $h(\cdot)$. The specification of the taste distribution in Equation (2) requires a mapping $h : \mathbb{R}^d \to \mathbb{R}^g$ that reduces the high-dimensional prompt embedding $\mathbf{E}_p$ (where $d = 3072$) to a low-dimensional vector of controls suitable for the random coefficients specification. The choice of this mapping is non-trivial. The objective is to identify the specific dimensions of the semantic space of prompts that are most informative of consumer preference heterogeneity.

The standard approach in the recent literature combining unstructured data with demand estimation is PCA (e.g., Compiani et al. (2023), Bach et al. (2024)). PCA is an unsupervised technique that projects the embeddings onto a small number of orthogonal components that maximize the explained variance within the embedding matrix. While this approach has proven effective in restricted domains—such as product images/descriptions within a single category like apparel—it faces significant limitations in the context of text from conversations involving LLMs. LLM prompts exhibit significant semantic dispersion and isotropy; there are no dominant directions of variance. In the LMArena dataset, for instance, the top three principal components explain less than 6% of the total variance in prompt embeddings. More critically, PCA is agnostic to choice behavior: the dimensions that capture the most variance may not be the ones that drive substitution patterns (e.g., mathematical intent vs. creative writing).

two models and the outside option.

To address these limitations, I use Partial Least Squares (PLS) (Wold (1966), Wold (1985)), a supervised dimensionality reduction technique. Unlike PCA, which maximizes the variance of the projections of $\mathbf{E}_p$, PLS identifies latent components that maximize the *covariance* between the projections of prompt embeddings and a set of target variables. This ensures that the reduced-dimensional representation $h(\mathbf{E}_p)$ retains precisely the semantic features that are most predictive of user choices.

Implementing PLS in a discrete choice framework requires constructing a target response variable that proxies for the latent utility differences driving decisions. To construct this target, I restrict the sample to battles where the consumer prefers one model over the other, excluding observations where the outcome is a tie or "both bad."[16] I construct a multivariate response matrix that captures the "revealed importance" of product attributes for each battle.

Let b index a specific battle. Let $y_b^* \in \{-1, 1\}$ denote the choice indicator, where $y_b^* = 1$ if model m_B is chosen and -1 if model m_A is chosen. Let $\Delta\mathbf{X}_b$ represent the $K \times 1$ vector of differences in response and LLM-level characteristics between the two competing models (i.e., $\Delta\mathbf{X}_b = [X_{pm_B}\ X_{m_B}] - [X_{pm_A}\ X_{m_A}]$). I define the target matrix $\mathbf{W}$ for the PLS regression as an $N \times (K-1)$ matrix, where N is the number of battles and $K-1$ is the number of characteristics (excluding the constant term). Each row b of $\mathbf{W}$, denoted by the vector $\mathbf{w}_b$, corresponds to a single battle and is formed by interacting the choice indicator with the vector of characteristic differences:

$$\mathbf{w}_b = y_b^* \cdot \Delta\mathbf{X}_b$$

Equivalently, for each characteristic k, the element w_{bk} is given by $y_b^* \cdot \Delta x_{bk}$. The vector $\mathbf{w}_b$ serves as a proxy for the gradient of utility with respect to characteristics in that instance. Intuitively, if a user chooses a model that generates a significantly longer response ($y_b^* = 1$ and $\Delta x_{\text{length}} > 0$), the corresponding element in $\mathbf{W}$ will be large and positive, signalling a revealed preference for verbosity for that specific prompt. Conversely, if the user rejects the longer response, the element becomes negative.

I then perform the PLS regression of the prompt embeddings matrix $\mathbf{E}$ (which stacks all prompts embeddings $\mathbf{E}_p$ in the data) on the multivariate target matrix $\mathbf{W}$, restricting the dimensionality of the solution to the top three latent components. The details regarding the algorithm and its implementation are presented in Appendix D. The resulting latent components are orthogonal and constitute the vector $h(\mathbf{E}_p)$. This approach is consistent

[16] Including battles where no choice is made between the two models would provide no directional signal regarding preferences for LLM characteristics (these battles do provide information about substitution to the outside option).

with the underlying economic model: it explicitly searches for the dimensions of the prompt space that explain variation in the marginal utility of response characteristics, rather than variation in the prompts alone. It is also worth noting that the estimation does not rely on an ad-hoc classification of prompt topics, instead it leverages directly the prompt embeddings.

Endogeneity of response length. Response length is likely positively correlated with unobserved dimensions of response quality (ξ_{pm}) such as completeness and accuracy: e.g., a comprehensive answer to a complex query inherently requires more tokens than a superficial one. Consequently, the estimated coefficient for response length likely captures a reduced-form preference for verbosity and completeness associated with longer outputs, rather than a pure preference for verbosity holding response quality fixed.[17]

This interpretation has important implications for the counterfactual analysis and welfare decomposition presented in Section 6. In the baseline analysis, I categorize response length as a "horizontal" response-level attribute akin to verbosity, distinct from core model capabilities (X_m). I conduct robustness checks, in Appendix B, where an increase in response length for new LLMs is treated as reflecting improvements in model intelligence.

Response embeddings. The utility specification in Equation (1) does not use text embeddings of the LLM *responses*, only of the user *prompt*. While recent approaches in demand estimation use product embeddings to capture unobserved product characteristics (e.g., Compiani et al. (2023)), that approach is ill-suited for distinguishing between LLM outputs conditional on a specific prompt.

Embeddings are designed to capture semantic proximity rather than attributes (correctness, completeness, formatting, tone) that drive user choice in this context. Therefore, LLM responses to a given prompt will typically relate to the same core informational content, resulting in very close embedding vectors. For example, a concise, well-structured explanation of quantum computing and a verbose, unstructured block of text on the same topic may be semantically identical (generating high cosine similarity) yet yield vastly different utilities to a user. Since the variance in response embeddings primarily reflects the *topic* of the prompt, relying on explicit response characteristics (e.g., formatting, tone, readability) and latent quality terms (e.g., X_m and ξ_m) provides a more robust basis for estimation.[18]

[17] Since the specification includes LLM fixed effects and controls for prompt embeddings, the bias only arises from within-model variation in response quality that correlates with within-model variation in response length, for a specific query.

[18] In addition, estimates of the effect of changes in response embeddings cannot be mapped into actual firm-level decisions regarding product characteristics, which makes these estimates hard to interpret and leverage in counterfactuals.

5.3 Estimation results

This section presents the results of the two-step estimation approach. To facilitate optimization, all continuous response-level variables (i.e., response length, SMOG index) are normalized to have mean zero and unit variance. Categorical features, including sentiment, hedging, reasoning, response has table, and response has emojis, are coded as indicator variables. Benchmark scores are rescaled to lie between 0 and 10, so that a one-unit change corresponds to a 10% increase in benchmark performance. The context window size is similarly rescaled so that a one-unit change corresponds to an increase of one million tokens. Finally, a dummy variable is included for whether the LLM appears on the right in the battle (Model B), to account for any potential "position bias" in user comparisons.

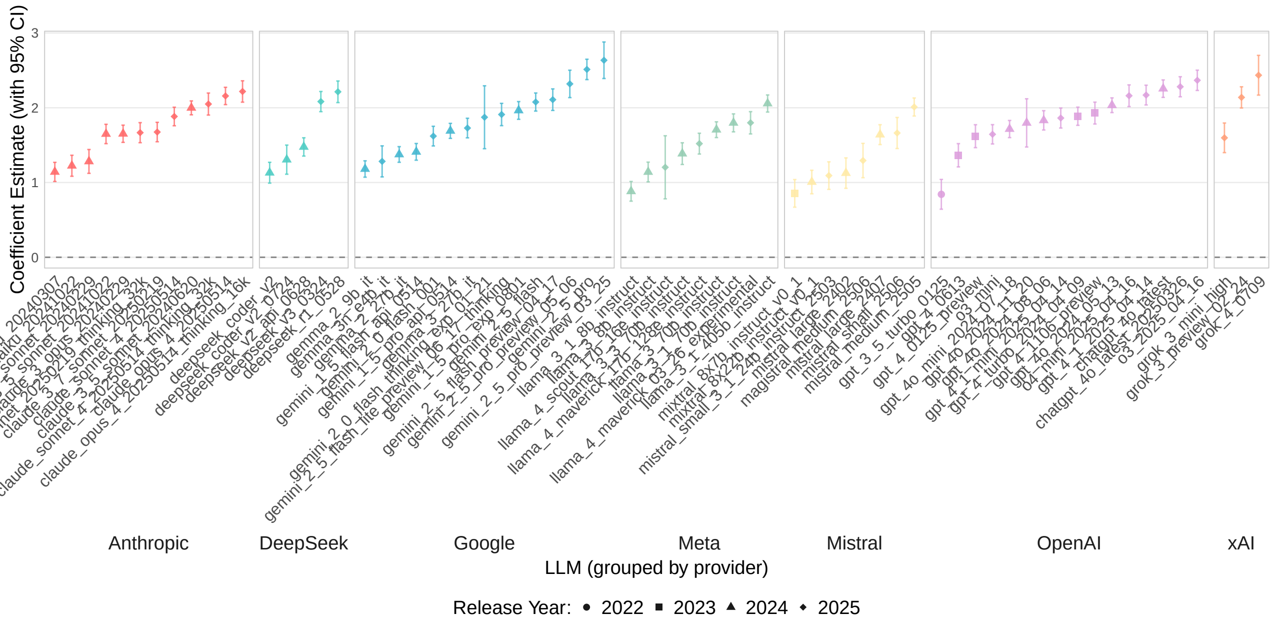


Figure 5: Estimation results: LLM fixed effects by provider

Figure 5 presents coefficient estimates for the LLM fixed effects ($\widehat{\delta}_m$ terms), capturing average consumer utility for each LLM by the top 7 providers. More recent models tend to generate responses that are on average preferred by consumers. Among top-performing models, Gemini 2.5 Pro yields the highest estimated utility, though other models (e.g., Claude Opus 4 (thinking), and OpenAI's o3) achieve comparable scores. Rankings broadly align with underlying model size (number of parameters in the LLM): for instance, within the Llama family of models created by Meta, estimated utility increases from Llama 3.1 8B to 70B to 405B (billions of parameters). Finally, reasoning models—such as Gemini 2.5 Pro, Opus-thinking, Sonnet-thinking, and o3—rank at the top of the distribution, indicating that reasoning capabilities are particularly valued by consumers.

Table 4: Estimation Results

	(1)	(2)		(3)		(4)	
	Mean (β)	Mean (β)	SD (σ)	Mean (β)	SD (σ)	Mean (β)	SD (σ)
Model capabilities							
Intelligence index	0.312	0.355	0.059	0.341	0.075		
	(0.007)	(0.010)	(0.020)	(0.010)	(0.020)		
Knowledge index						0.361	0.045
						(0.029)	(0.021)
Coding index						0.036	0.025
						(0.021)	(0.039)
Context window	-0.001	-0.024	0.199	-0.027	0.210	-0.021	0.233
	(0.006)	(0.012)	(0.059)	(0.012)	(0.059)	(0.013)	(0.057)
Reasoning model	-0.316	-0.353	0.268	-0.368	0.402	-0.247	0.389
	(0.022)	(0.027)	(0.366)	(0.028)	(0.267)	(0.027)	(0.254)
Readability and tone							
Flesch reading ease	0.021	0.041	0.087	0.037	0.181	0.031	0.219
	(0.009)	(0.029)	(0.093)	(0.042)	(0.080)	(0.043)	(0.089)
SMOG index	0.074	0.062	0.136	0.013	0.085	0.012	0.066
	(0.006)	(0.013)	(0.045)	(0.016)	(0.064)	(0.016)	(0.069)
Sentiment: positive	0.177	0.190	0.309	0.054	0.432	0.052	0.467
	(0.017)	(0.022)	(0.128)	(0.026)	(0.122)	(0.026)	(0.118)
Sentiment: negative	-0.016	-0.026	0.219	-0.337	0.424	-0.332	0.393
	(0.045)	(0.058)	(0.329)	(0.085)	(0.185)	(0.085)	(0.195)
Hedging: assertive	-0.123	-0.108	0.560	-0.044	0.300	-0.048	0.372
	(0.012)	(0.016)	(0.049)	(0.017)	(0.071)	(0.017)	(0.065)
Hedging: tentative	-0.189	-0.220	0.077	-0.215	0.256	-0.215	0.217
	(0.017)	(0.022)	(0.207)	(0.023)	(0.113)	(0.023)	(0.137)
Formatting							
Response length	0.086	0.491	1.176	0.694	1.098	0.682	1.053
	(0.007)	(0.025)	(0.051)	(0.032)	(0.052)	(0.031)	(0.048)
Ordered lists	0.032	0.065	0.344	0.043	0.242	0.048	0.273
	(0.007)	(0.014)	(0.045)	(0.014)	(0.048)	(0.015)	(0.048)
Unordered lists	0.047	0.065	0.176	0.031	0.297	0.028	0.248
	(0.007)	(0.019)	(0.089)	(0.022)	(0.072)	(0.022)	(0.086)
Markdown headers	-0.003	0.069	0.428	0.032	0.310	0.037	0.424
	(0.005)	(0.018)	(0.061)	(0.021)	(0.068)	(0.022)	(0.073)
Bold text	0.091	0.244	0.606	0.103	0.680	0.094	0.617
	(0.007)	(0.022)	(0.048)	(0.025)	(0.048)	(0.024)	(0.050)
Has emoji	0.135	0.161	0.810	0.093	1.176	0.089	0.996
	(0.023)	(0.032)	(0.163)	(0.036)	(0.146)	(0.035)	(0.151)
Has table	0.279	0.359	0.863	0.207	0.852	0.213	0.970
	(0.025)	(0.043)	(0.212)	(0.048)	(0.214)	(0.049)	(0.191)
LLM FE	✓	✓		✓		✓	
Random Coefficients		✓		✓		✓	
PLS scores				✓		✓	
Observations	85,257	85,257		85,257		85,257	
LLM FE	70	70		70		70	
Log-likelihood	-84872.6	-82974.0		-81763.5		-81699.1	
LR test stat (df, p-value)		3797 (20, 0.0)		2421 (54, 0.0)			

Note: Standard errors are clustered at the user level and shown in parentheses. Mean estimates for model capabilities (and the constant) are obtained via GLS regression of LLM FEs on model capabilities. LR test shows the likelihood ratio test of specification (2) vs. (1), and (3) vs. (2). Specifications (3) and (4) include interactions of all variables with each prompt-level PLS scores. The reference categories for the tone dummies are neutral sentiment and neutral hedging. Additional controls not reported: dummy for whether the model appears on the right (model B), and a fixed effect and random coefficient on the indicator of LLMs in the other category (i.e., not belonging to the top 7 providers).

Table 4 reports estimates for the coefficients $(\widehat{\beta}_0^R, \widehat{\beta}_0^L, \widehat{\Sigma})$, corresponding to the means and standard deviations of the taste parameters for model- and response-level characteristics. We note that the coefficients $\widehat{\beta}_0^L$ are obtained in step 2 (GLS), whereas all other coefficients are obtained in step 1 (simulated MLE). Specification (1) includes only LLM fixed effects and sets Σ to zero, specification (2) include the random coefficients Σ, while columns (3) and (4) add interactions between all characteristics and the prompt-level PLS scores $\mathbf{E}_p$ (three PLS scores per prompt). Specification (4) replaces the intelligence index with separate coding and knowledge indexes.[19]

Across all specifications, consumer utility from a response increases with benchmark performance, as captured by the intelligence index.[20] This indicates that consumer preferences for LLM responses are strongly aligned with an LLM's benchmark performance. Utility from a response is only weakly responsive to an LLM's context window size. This weak effect is expected, as most LMArena conversations are single-turn, rendering the context window size non-binding in practice. Conditional on benchmark scores, reasoning models are on average less preferred, likely reflecting their slower response times.

Turning to readability and tone, users prefer responses with positive relative to neutral sentiment, and systematically dislike negative sentiment in responses. Assertive and tentative responses are also penalized relative to neutral; the disutility from tentative hedging is nearly four times larger than that from assertive hedging. Finally, most formatting features—such as use of lists, bold text, and structured elements like tables—have a positive mean effect on the utility from a response. However, as shown next, these mean effects mask substantial variation across prompts and topics.

Figures 6, 7, and 8 plot the distribution of prompt-specific coefficients, for all prompts in the data, for key response and model characteristics. These coefficients represent the deterministic component of user preferences: $\beta_0^R + \beta_1^{R\prime} h(\mathbf{E}_p)$ for response-level characteristics, and $\beta_0^L + \beta_1^{L\prime} h(\mathbf{E}_p)$ for model-level characteristics, conditional on the prompt embedding. A central finding is that many variables exhibit sign variation across and within topics, indicating that the utility derived from a given feature is highly use-case dependent. For instance, Figure 8 shows that the SMOG index—a proxy for linguistic complexity—is positively valued in technical tasks like math and coding, but negatively associated with utility in most practical guidance prompts. A few features, however, show consistent directional effects across topics. Intelligence benchmark scores and response length are uniformly positive, while tentative

[19] I do not include the math index (constructed as the average of MATH-500 and AIME scores) due to its high correlation with the coding index, likely because the latter index (based on Scicode and LiveCodeBench) emphasizes scientific and mathematically intensive coding tasks. Including both leads to multicollinearity issues.

[20] Although the mean effect of the coding index is statistically insignificant in Specification (4), its interaction with PLS scores is statistically significant as shown in Figure 6, highlighting its importance only for a subset of use cases.

tone tends to reduce utility regardless of the domain.

The distribution of coefficients also reveals important tradeoffs. Figure 6 shows that the knowledge index (MMLU-Pro, HLE, GPQA Diamond) is positively valued across all prompts. In contrast, the utility from higher coding index (LiveCodeBench, SciCode) depends on the task: a higher score is beneficial for math and coding prompts, but negatively associated with utility in at least half writing, casual conversation, and information-seeking prompts. This pattern suggests that gains in coding ability may come at the cost of fluency or alignment in non-technical domains.

Formatting features also exhibit nuanced effects. The use of tables (Figure 7) is especially useful for information-seeking and practical guidance queries, but provide limited benefit—or even reduce utility—in writing tasks. Bold text, by contrast, appears most effective in highlighting key elements in math and information-seeking conversations, but is largely neutral elsewhere. Finally, sentiment effects vary moderately by topic: users generally prefer responses with positive tone, though the effect is muted or slightly negative in coding-related prompts.

A comparison of the standard deviations shown in Table 4 (random coefficients Σ) with the standard deviation of prompt-specific coefficients shown in Figures 6–8 reveals that heterogeneity in taste parameters across prompts exceeds heterogeneity across consumers for model capabilities (intelligence, coding, knowledge indexes). This implies that users broadly agree on the value of core capabilities like reasoning, coding, and knowledge, but that their relative importance varies considerably across use cases. The opposite holds for formatting, readability, and tone: consumer-level variation dominates, suggesting that preferences for presentation style and linguistic framing are more idiosyncratic.

To compare the relative magnitudes of the estimated coefficients, I simulate a hypothetical scenario in which two identical models are compared head-to-head in a battle. In this hypothetical setting, I increase one model- or response-level characteristic by a single unit for one of the two models, while holding all other attributes fixed. Figure 9 reports the resulting change in win probability, averaged across prompts and broken down by topic. Since baseline win probabilities are 50% (identical models), the estimates capture the percentage point gain (or loss) associated with each characteristic relative to 50%.

The results confirm several patterns documented earlier. Formatting features, such as bold text, tables, and list structures, yield the largest gains in information-seeking and practical guidance contexts, and to a lesser extent in math. Response length improves win probability the most for writing, casual conversation, information-seeking, and practical guidance queries. Model capabilities, however, show more domain-specific effects. Notably, the coding index improves win probability for math and coding prompts but has negative

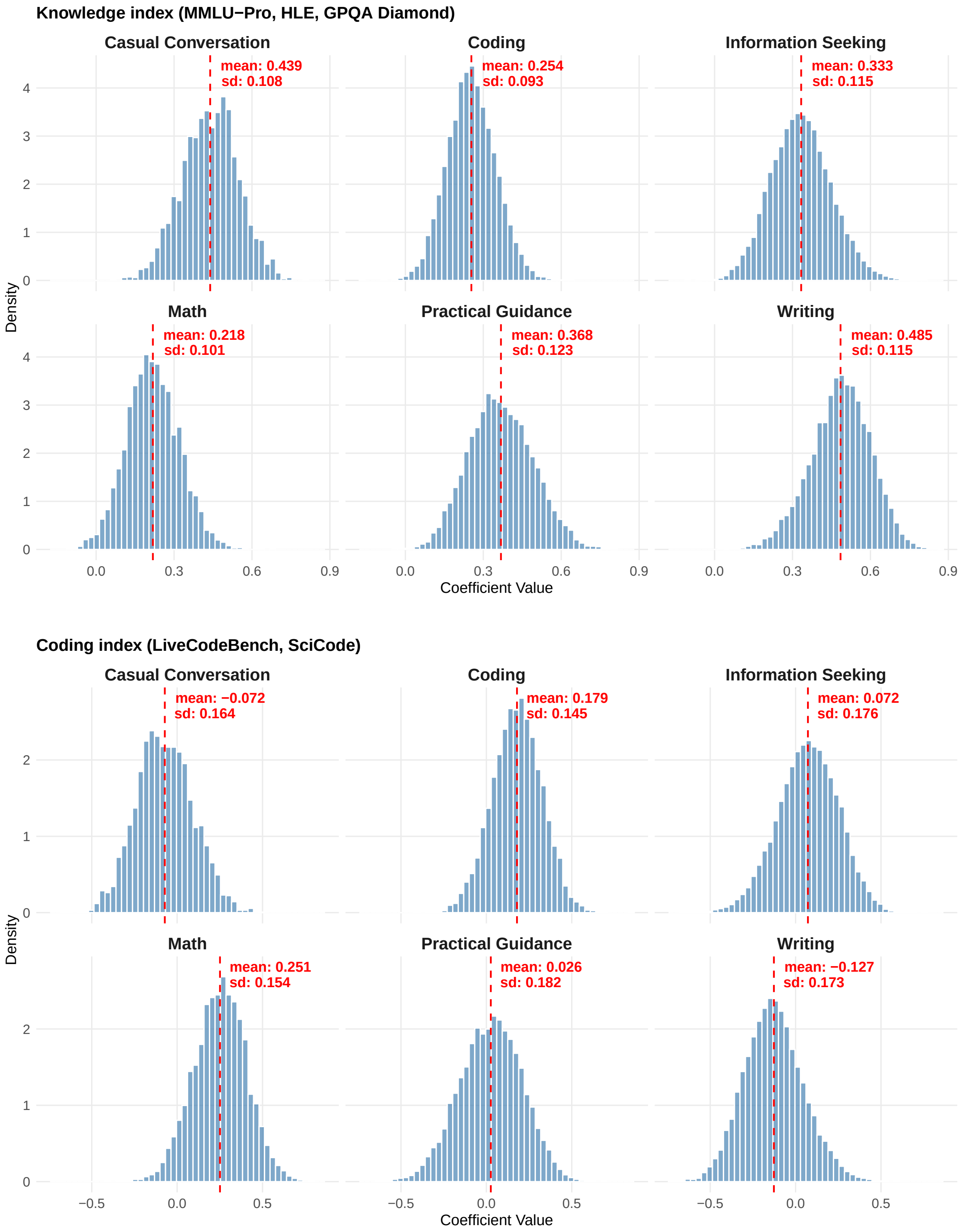


Figure 6: Distribution of taste parameters over prompts by topic: Knowledge and Coding Indices

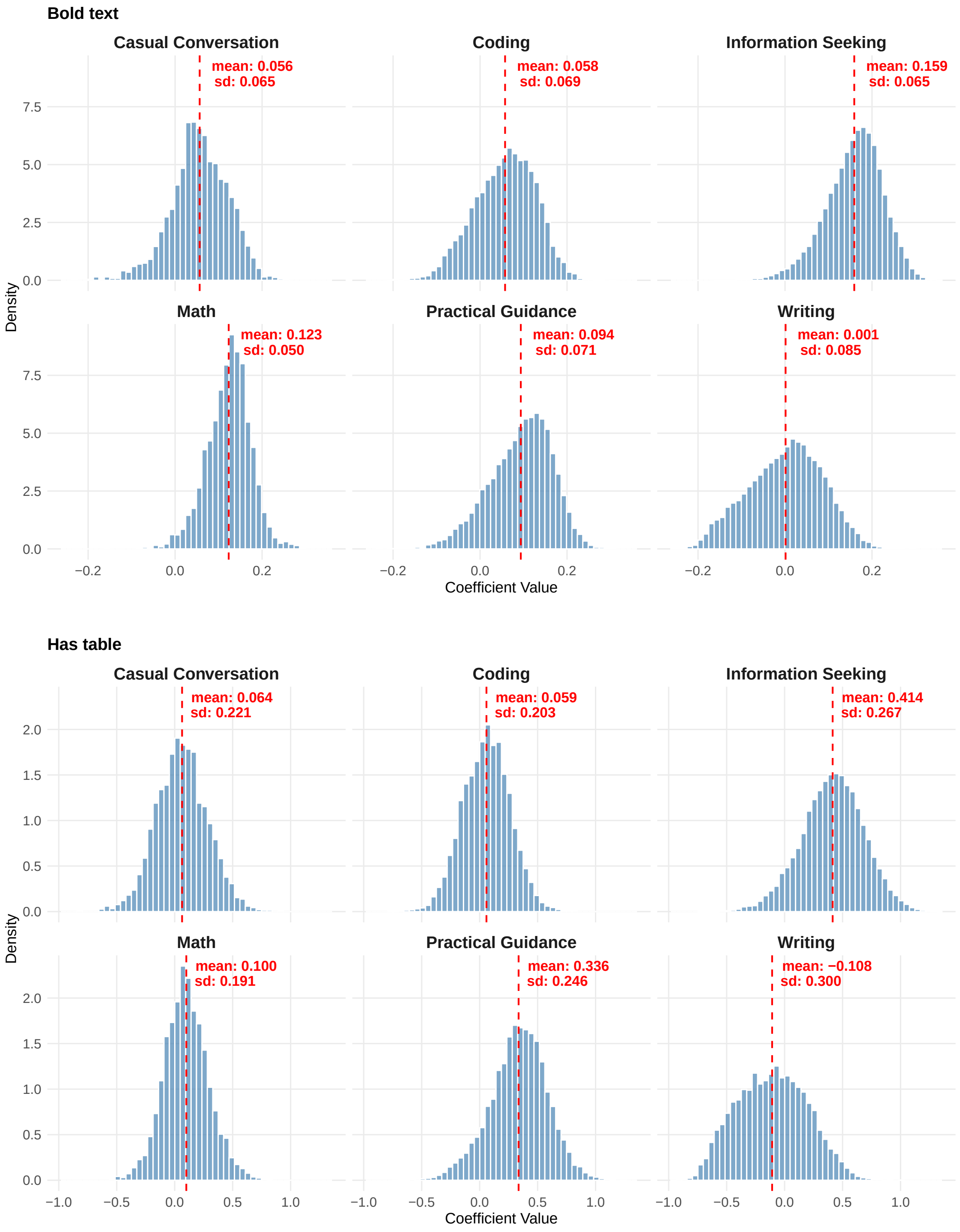


Figure 7: Distribution of taste parameters over prompts by topic: Table use and bold text

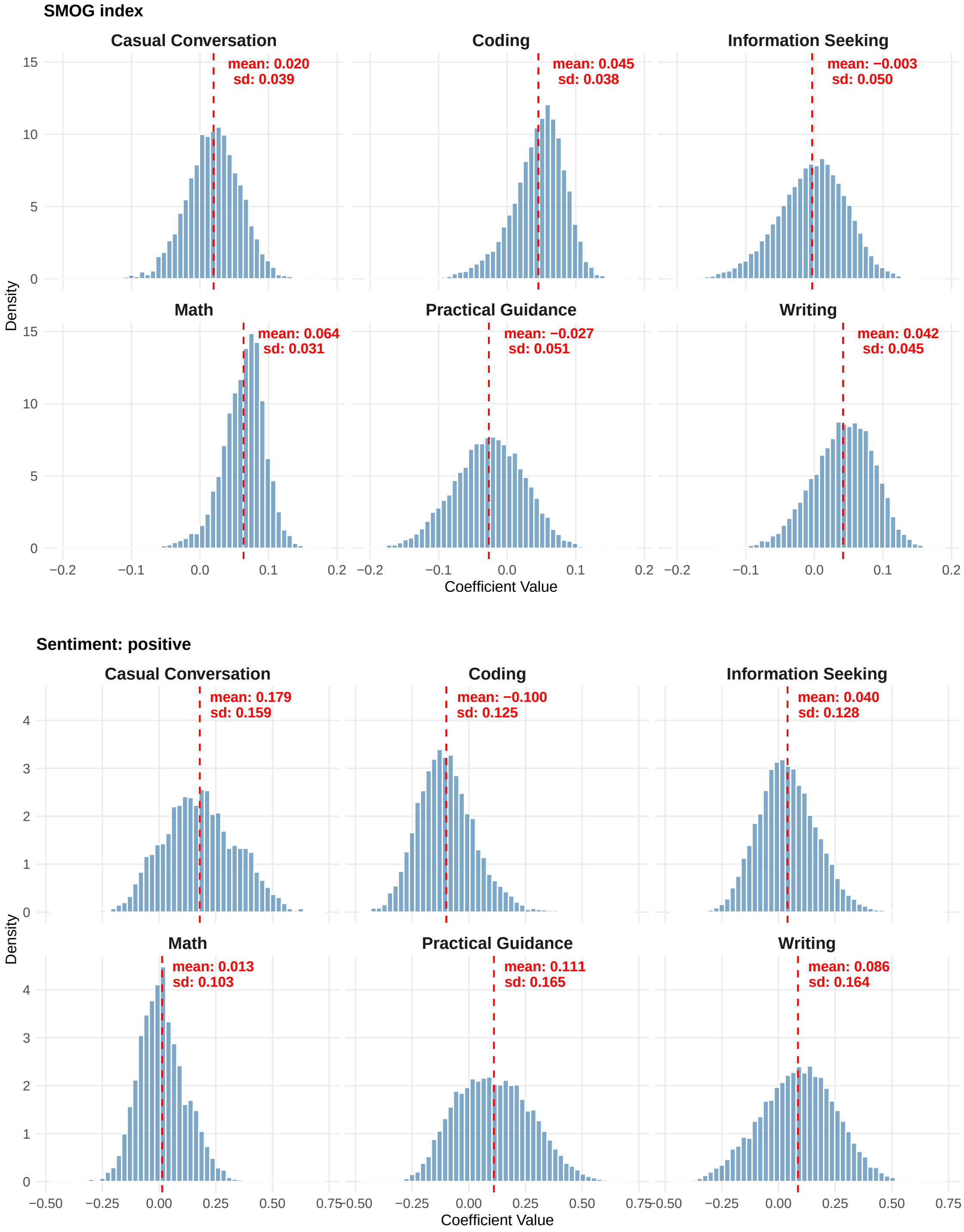
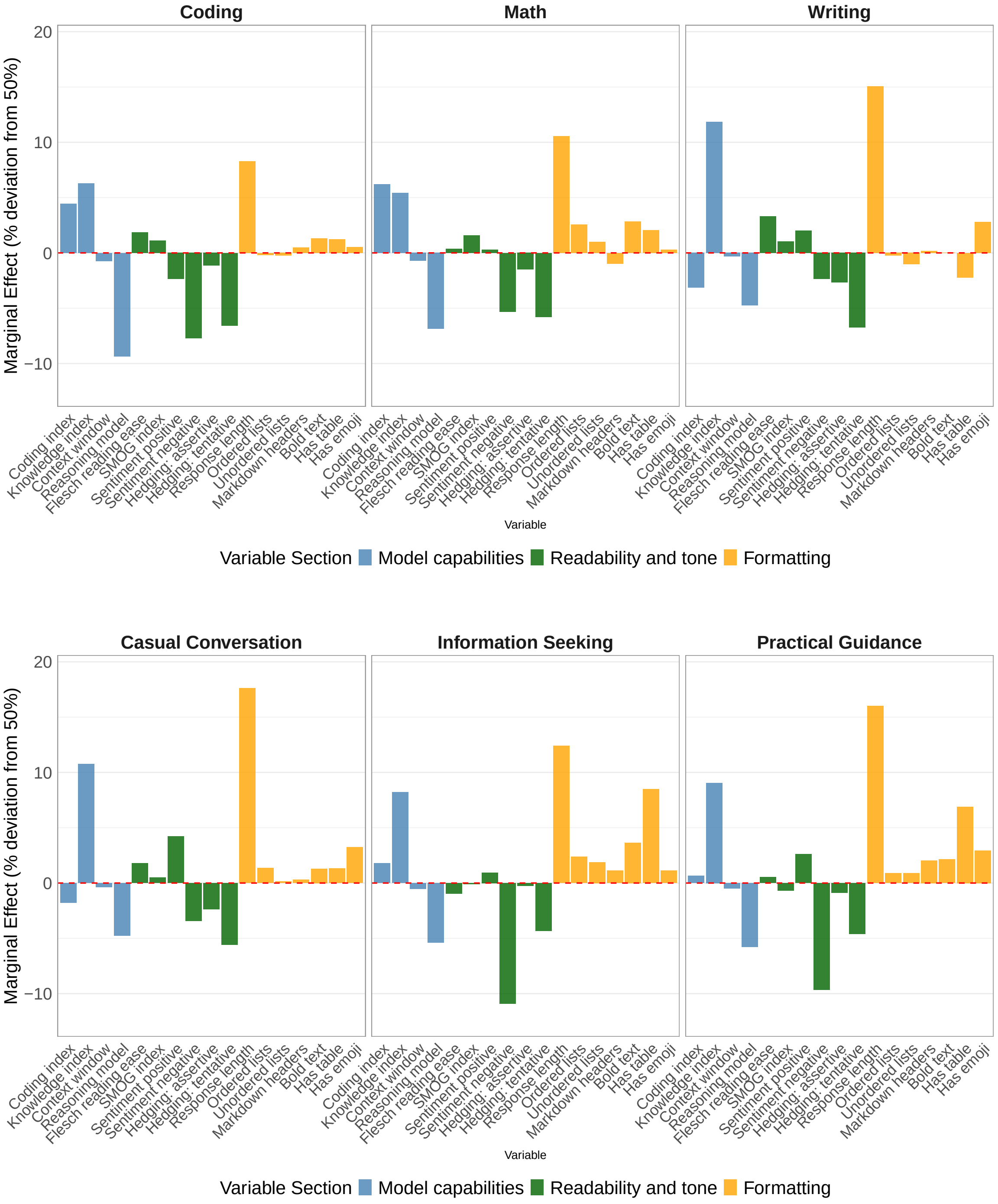


Figure 8: Distribution of taste parameters over prompts by topic: SMOG index and positive sentiment



Note: Marginal effects on win probability from a unit change in each regressor in a symmetric battle. For readability, tone and formatting, a unit change corresponds to a one standard deviation increase, except for the dummy variables (sentiment, hedging, has table, has emoji) which are not rescaled. For model capabilities, the effects are calculated based on a 10% increase in the knowledge or coding index and a 1 million token increase in the context window size. Reasoning is a dummy variable.

Figure 9: Marginal effects on win probability in symmetric battle, by topic

marginal effects in other domains such as writing and casual conversation. This divergence indicates that, whereas model improvements in benchmarks may enhance performance for specific tasks, they can reduce performance elsewhere.

6 Counterfactual Analysis

This section develops a structural model of LLM demand to quantify changes in consumer surplus over time. I embed response-level utility estimates from LMArena into a discrete choice demand model over LLMs in the actual consumer market. Under two polar assumptions about user behavior (prompt-level multi-homing and single-homing), I compute the gains in consumer surplus over time and decompose them into components driven by improvements in core model intelligence and changes in response-level characteristics.

6.1 A Demand Model for LLMs

The central premise of the demand model is that consumers select among LLMs based on the expected utility they derive from model responses to their prompts. In period t, let $\mathcal{J}_t$ denote the set of LLMs available in the market, where each LLM is associated with a provider (e.g., Google, OpenAI) and characterized by observable attributes X_m. For example, in July 2025, $\mathcal{J}_t$ includes models such as Gemini 2.5 Flash and Pro (Google) and Claude Sonnet 4 and Opus 4 (Anthropic).

An important modeling choice concerns the extent to which consumers multi-home—i.e., use multiple LLMs and providers across queries. In practice, some users may switch models on a prompt-by-prompt basis, while others may remain committed to a single provider. To capture this heterogeneity, I consider two polar cases for aggregating response-level utilities. In the first case, *prompt-level multi-homing*, users select the utility-maximizing LLM for each individual query. In the second case, *single-homing*, users commit to a single provider across all queries and select the provider whose portfolio of LLMs delivers the highest expected utility over the user's prompt distribution. These two extremes bound the range of plausible demand responses and bracket the resulting welfare estimates.

In the remainder of this section, I rewrite the utility function (Equation (1)) as:

$$U_{ipm} = \delta_{ipm} + \epsilon_{ipm}. \tag{7}$$

Prompt-level multi-homing. For each user i and prompt p (with corresponding prompt embedding $\mathbf{E}_p$), the expected utility conditional on choice set $\mathcal{J}_t$ is given by the standard logit inclusive value:

$$\textit{Expected utility for prompt } p\text{:} \quad U_{ip} = \mathbb{E}\left[\max_{m \in \mathcal{J}_t} \{\delta_{ipm} + \epsilon_{ipm}\}\right] = \ln\left[\sum_{m \in \mathcal{J}_t} \exp(\delta_{ipm})\right] + \gamma$$

User i's utility over prompts is obtained by integrating over her prompt distribution $F_i(\mathbf{E}_p)$:

$$\textit{Aggregate utility for user } i\text{:} \quad U_i = \int U_{ip} \, dF_i(\mathbf{E}_p)$$

Finally, to compute the aggregate utility across users, I integrate over the population distribution $H(i)$, which governs both heterogeneity in user preferences ν_i and prompt distributions F_i:

$$\textit{Aggregate utility across users:} \quad CS(\mathcal{J}_t) = \int U_i \, dH(i)$$

Single-homing. In single-homing specification, each user selects a single provider to handle all prompts. Let $\mathcal{J}_t = \{\mathcal{J}_{ft}\}_{f \in \mathcal{F}}$ denote the partition of available LLMs in period t across firms $f \in \mathcal{F}$, where $\mathcal{J}_{ft} \subseteq \mathcal{J}_t$ is the portfolio offered by firm f. For each user i and prompt p, the expected utility from selecting firm f is given by the logit inclusive value over LLMs in $\mathcal{J}_{ft}$:

$$U_{ipf} = \mathbb{E}\left[\max_{m \in \mathcal{J}_{ft}} \{\delta_{ipm} + \epsilon_{ipm}\}\right] = \ln\left[\sum_{m \in \mathcal{J}_{ft}} \exp(\delta_{ipm})\right] + \gamma$$

Aggregating over prompts using user i's prompt distribution $F_i(\mathbf{E}_p)$ gives the expected utility of choosing firm f:

$$\textit{Expected utility of user i if choosing firm } f\text{:} \quad U_{if} = \int U_{ipf} \, dF_i(\mathbf{E}_p)$$

Each user selects the provider that delivers the highest expected utility across their prompt distribution:

$$\textit{Firm-level choice for user } i\text{:} \quad U_i = \max_{f \in \mathcal{F}} U_{if}$$

Finally, the average utility per user is obtained by integrating over the population:

$$\textit{Aggregate utility across users:} \quad CS(\mathcal{J}_t) = \int U_i \, dH(i).$$

Beliefs about response-level characteristics. In the LMArena platform, users evaluate responses after observing them; however, in real-world applications, users must select an LLM before observing its output. This distinction raises the question of how to evaluate utility terms that depend on response-level characteristics X_{pm}, which are not known at the time of choice.[21]

To address the unobservability of X_{pm} at choice time, I assume users have rational expectations and evaluate LLMs based on the expected value of X_{pm} conditional on the LLM m and the topic category c of the prompt. That is, when constructing the δ_{ipm} terms entering the inclusive values, I replace X_{pm} with $\mathbb{E}[X_{pm}|m, c_p]$, where c_p is the topic category of prompt p.[22] This captures the idea that users expect, for example, that Claude models tend to use more complex syntax on writing prompts, or that Google and OpenAI models produce warmer or more enthusiastic responses on practical guidance queries. Because prompts are randomly assigned to models in LMArena, I can obtain unbiased estimates of these conditional expectations by regressing X_{pm} on LLM fixed effects and topic category indicators.

The topic categories used in these regressions correspond to those shown in Figures 6–8: Information Seeking, Writing, Practical Guidance, Coding, Math, and Casual Conversation. I denote the set of categories $\mathcal{C}$.

Distribution of user prompts F_i. To operationalize the computation of consumer surplus $CS(\mathcal{J}_t)$, it is necessary to specify the distribution of prompt embeddings F_i. I calibrate this distribution using the empirical distribution of conversation topics on ChatGPT reported in Chatterji et al. (2025) (Figure 9). The latter paper uses internal data from OpenAI on approximately 1.1 million conversations from May 2024 through June 2025. The reported shares by topic category are: Casual Conversational (4.3%), Coding (4.6%), Information Seeking (21.3%), Math (3.0%), Practical Guidance (28.3%), and Writing (28.1%). I omit the "Other" and "Multimedia" topic categories from the analysis, as my focus is restricted to text-based interactions, and rescale the remaining six topics' shares to sum to one.

I assume that all users draw prompts from a common distribution $F_i = F$ over prompt embeddings. Specifically, prompts are drawn in two stages. First, I draw a topic category $c \in \mathcal{C}$ according to the empirical topic shares described above. Then, conditional on category c, I draw a prompt embedding $\mathbf{E}_p$ from the empirical distribution $\widehat{F}(\cdot \mid c)$ estimated from the LMArena data. This sampling procedure ensures that the distribution of prompts used in the

[21] I assume users do observe the model-level characteristics X_m at the time of selection. For example, users are aware that Claude 4 Sonnet performs better on benchmarks than Claude 3.7 Sonnet, that o3 is a reasoning model, or that Gemini 2.5 Pro has a larger context window size than similar models. These characteristics are known by consumers without needing to observe model outputs.

[22]

welfare calculations closely mirrors the real-world distribution of user queries on ChatGPT.

While the assumption of a common F simplifies the analysis, it is made for convenience and can be relaxed using external data on user-level heterogeneity in prompt distributions. Integration over the population distribution $H(i)$ captures heterogeneity in taste parameters ν_i, holding prompt distributions fixed across users.

Counterfactual simulations. Given a choice set $\mathcal{J}_t$, with corresponding model-level characteristics $\{X_m\}_{m\in\mathcal{J}_t}$ and response-level expectations $\{\mathbb{E}[X_{pm}|m, c_p]\}_{m\in\mathcal{J}_t,\, c_p\in\mathcal{C}}$, I simulate user utilities and compute the resulting consumer surplus $CS(\mathcal{J}_t)$ under both single- and multi-homing. Simulations are done by drawing from the distribution of prompt embeddings $\mathbf{E}_p$ and consumer taste heterogeneity ν_i.

I evaluate consumer surplus under two market configurations: $\mathcal{J}_0$, the set of models available in July 2024, and $\mathcal{J}_1$, the set available in July 2025. Appendix Table A3 provide details about the choice sets in these two periods. The proportional change in consumer surplus across these two periods is given by

$$\Delta CS = \frac{CS(\mathcal{J}_1) - CS(\mathcal{J}_0)}{CS(\mathcal{J}_0)}.$$

This change reflects the combined effect of new product entry, improvements in core model capabilities (X_m), and shifts in the response-level characteristics of LLM outputs ($\mathbb{E}[X_{pm}|m, c_p]$): e.g., formatting, readability, and tone. These latter changes may arise not only from technical improvements but also from better alignment of model outputs with human preferences over text (e.g., better formatting of responses for practical guidance or information seeking queries).

To isolate the contribution of improvements in core capabilities, captured by X_m, I construct counterfactual consumer surplus $CS(\mathcal{J}_1^{CF})$, by imposing that all LLMs available in July 2025 share the same response-level characteristics. Specifically, for each topic category $c \in \mathcal{C}$, I fix the expected response attributes $\mathbb{E}[X_{pm}|m, c]$ to their average values over LLMs in $\mathcal{J}_0$ ($\mathbb{E}[X_{pm}|\mathcal{J}_0, c]$). Under this counterfactual, LLMs in July 2025 differ only in their core capabilities X_m and are undifferentiated along dimensions such as readability, tone, and formatting. Comparing actual and counterfactual welfare levels in $\mathcal{J}_1$ to $\mathcal{J}_0$ allows me to estimate the share of surplus gains attributable to capability improvements (X_m), holding response-level variation constant.[23]

In Appendix B, I conduct robustness checks where I treat response length as reflecting

[23] The model does not account for substitution from the outside option (i.e., not using LLMs) to the inside goods. All welfare comparisons are restricted to users who always choose among LLMs. This reflects a limitation of the LMArena setting: the estimated preferences are not informative about selection into the market and, therefore, cannot be used to quantify market expansion effects.

completeness and accuracy, and therefore, also a part of model intelligence. In this extension, I replace $\mathbb{E}[X_{pm}|m, c]$ by $\mathbb{E}[X_{pm}|\mathcal{J}_0, c]$ for all response-level characteristics except response length: LLMs in July 2025 differ in their core capabilities X_m and their expected response length conditional on topic c, but are undifferentiated otherwise. In this extension, the share of welfare gains attributable to capability improvements accounts for greater completeness and accuracy from longer responses.

6.2 Counterfactual simulations

Table 5 shows the counterfactual simulation results. I show both the total change in average consumer surplus as well as changes broken down by topic. Aggregate consumer surplus increased substantially between July 2024 and July 2025 under both single- and multi-homing. Under prompt-level multi-homing, consumer surplus rose by 38%, whereas under single-homing, the increase was 54%. Decomposing these gains reveals that improvements in model intelligence (proxied by benchmark performance, reasoning ability, and context window size) explain a modest majority of the increase: 61% under multi-homing and 58% under single-homing. The remaining gains are attributable to changes in response-level attributes, with tone, formatting, and syntax becoming better aligned with human preferences in each use case. These results suggest that both capability-driven and stylistic improvements have contributed meaningfully to user welfare.

The relative importance of model improvements (X_m) varies across use cases. For technical domains such as Math and Coding, the majority of surplus gains reflect improvements in core capabilities: 68% and 71% under multi-homing, and 66% and 68% under single-homing, respectively. These tasks depend heavily on accuracy, reasoning, and other dimensions of model intelligence. By contrast, for use cases like Practical Guidance and Casual Conversation, stylistic and syntactic attributes account for a larger share of welfare gains—for example, 44% and 47% in Practical Guidance under multi- and single-homing. This heterogeneity highlights the multidimensional nature of LLM progress and the varying roles of model intelligence versus alignment with user tastes.

To translate the simulated utility gains into monetary terms, I rely on survey evidence. Estimates reported in Collis and Brynjolfsson (2025) indicate that the average U.S. user derived \$98 in monthly value from LLM usage in 2024. I use this figure to anchor consumer surplus in July 2024 ($CS(\mathcal{J}_0)$). Under prompt-level multi-homing, the increase in consumer surplus between July 2024 and 2025 corresponds to a gain of \$37.4 per user per month. Under single-homing, the implied gain is larger, at \$52.9 per user-month. These magnitudes imply that improvements in LLMs over this period generated economically meaningful welfare gains per user.

Table 5: Decomposition of CS change over time

	$CS(\mathcal{J}_0)$	$CS(\mathcal{J}_1^{CF})$	$CS(\mathcal{J}_1)$	ΔCS	$\frac{CS(\mathcal{J}_1^{CF})-CS(\mathcal{J}_0)}{\Delta CS}$	$\frac{CS(\mathcal{J}_1)-CS(\mathcal{J}_1^{CF})}{\Delta CS}$
Panel A: Prompt-Level Multihoming						
Casual Conversation	4.47	5.49	6.26	40.2%	57.3%	42.7%
Coding	4.20	5.24	5.66	34.9%	71.3%	28.7%
Information Seeking	4.52	5.61	6.34	40.3%	60.0%	40.0%
Math	3.78	4.93	5.47	45.0%	68.2%	31.8%
Practical Guidance	4.81	5.87	6.69	39.2%	56.4%	43.6%
Writing	4.62	5.66	6.23	34.9%	64.8%	35.2%
Total	4.60	5.66	6.35	38.1%	60.8%	39.2%
Panel B: Single-Homing						
Casual Conversation	3.25	4.27	5.12	57.6%	54.3%	45.7%
Coding	2.97	4.01	4.50	51.9%	68.1%	31.9%
Information Seeking	3.31	4.38	5.20	57.2%	56.7%	43.3%
Math	2.55	3.70	4.30	68.8%	65.5%	34.5%
Practical Guidance	3.60	4.64	5.57	54.5%	52.8%	47.2%
Writing	3.39	4.44	5.08	49.9%	62.1%	37.9%
Total	3.38	4.43	5.21	54.0%	57.7%	42.3%

Note: $CS(\mathcal{J}_0)$ and $CS(\mathcal{J}_1)$ are consumer surplus in July 2024 and July 2025, respectively, measured in utility units (utils). $CS(\mathcal{J}_1^{CF})$ is the counterfactual consumer surplus in 2025 with response characteristics fixed at the 2024 average. $\Delta CS = (CS(\mathcal{J}_1) - CS(\mathcal{J}_0))/CS(\mathcal{J}_0)$. The last two columns show the share of the CS increase attributable to model characteristics (entry and improvement in model features) and response characteristics, respectively.

7 Conclusion

This paper addresses a central question in the emerging market for LLMs: how do consumers evaluate and choose between products, given the continual evolution of model capabilities, including both core intelligence and surface-level response features such as tone, formatting, and readability? To make progress on this question, I leverage novel revealed-preference data from LMArena, a platform for LLM evaluation. These data enable the estimation of a structural demand model that maps prompt-specific preferences into utilities over LLMs, offering a granular view of how product features translate into consumer choice.

The estimated model reveals that user preferences are broadly consistent with benchmark performance, especially for technical tasks such as coding and mathematics. However, in non-technical domains, horizontal product attributes play a significant role in shaping demand. Counterfactual simulations quantify the welfare implications of changes in product attributes over time. Between 2024 and 2025, average consumer surplus increased by 38%–54%, with improvements in model intelligence accounting for the majority of gains. Still, better alignment with consumers' horizontal preferences contributes to surplus, particularly in non-technical domains. Because the analysis focuses on text interactions in chat settings, these estimates likely understate the full welfare gains from LLM improvement, which are increasingly embedded in downstream applications (e.g., software, image generation).

While the LMArena setting offers a unique window into preferences over LLMs, it abstracts from several real-world aspects of the market. Notably, the framework does not account for switching costs, firm-level complementarities (i.e., integration of LLMs into broader ecosystems), or the role of brand and reputation. Additionally, many consumer interactions might benefit from context shared with an LLM across conversations, giving rise to structural state dependence. Despite these limitations, the framework constitutes a necessary first step in studying consumer demand in this emerging industry. Extensions that incorporate these aspects, or that integrate a supply-side model, offer promising directions for future research.

References

Anthropic. 2025. "Deprecation Commitments for Claude Models." https://www.anthropic.com/research/deprecation-commitments, Accessed: 2025-12-01.

Armona, Luis, Greg Lewis, and Georgios Zervas. 2025. "Learning Product Characteristics and Consumer Preferences from Search Data." *Marketing Science* 44 (4): 838–855.

Bach, Philipp, Victor Chernozhukov, Sven Klaassen, Martin Spindler, Jan Teichert-Kluge, and Suhas Vijaykumar. 2024. "Adventures in Demand Analysis Using AI." 1–42, http://arxiv.org/abs/2501.00382.

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2021. "Common Ownership and Competition in the Ready-to-Eat Cereal Industry." *SSRN Electronic Journal.*

Bajari, P., Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, J. Wang, R. Huerta, J. Li, L. Leng, G. Monokroussos, and S. Wang. 2025. "Hedonic prices and quality adjusted price indices powered by AI." *Journal of Econometrics* 251.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–890.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *The RAND Journal of Economics* 25 (2): 242.

Berry, Steven T., and Philip A. Haile. 2021. "Foundations of Demand Estimation." *Handbook of Industrial Organization* (2301): .

Bick, Alexander, Adam Blandin, and David Deming. 2024. "The Rapid Adoption of Generative AI." *SSRN Electronic Journal.*

Brynjolfsson, Erik, Bharat Chandar, and Ruyu Chen. 2025a. "Canaries in the Coal Mine ? Six Facts about the Recent Employment Effects of Artificial Intelligence."

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025b. "Generative AI at Work." *Quarterly Journal of Economics* 140 (2): 889–942.

Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. "How People Use ChatGPT." http://www.nber.org/papers/w34255.

Chiang, Wei Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." *Proceedings of Machine Learning Research* 235 8359–8388.

Collis, Avinash, and Erik Brynjolfsson. 2025. "AI's Overlooked $97 Billion Contribution to the Economy." *The Wall Street Journal*, https://www.wsj.com/opinion/ais-overlooked-97-billion-contribution-to-the-economy-users-service-da6e8f55.

Competition and Markets Authority. 2024a. "AI Foundation Models Technical update report." (April): , www.nationalarchives.gov.uk/doc/open-government-.

Competition and Markets Authority. 2024b. "AI Foundation Models Update paper." (April): 1–24, https://www.gov.uk/government/publications/ai-foundation-models-update-paper.

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler. 2023. "Demand Estimation with Text and Image Data." *SSRN Electronic Journal*.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *SSRN Electronic Journal*.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis, and Sida Peng. 2025. "The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs."

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." *Science* 384, http://arxiv.org/abs/2303.10130.

Federal Trade Commission. 2025. "Partnerships Between Cloud Service Providers and AI Developers FTC Staff Report on AI Partnerships Investments 6(b) Study." 6 (January): .

Financial Times. 2025. "OpenAI needs to raise at least $207bn by 2030 so it can continue to lose money, HSBC estimates." https://www.ft.com/content/23e54a28-6f63-4533-ab96-3756d9c88bad, Accessed: 2025-12-03.

Gabel, Sebastian, and Artem Timoshenko. 2022. "Product Choice with Large Assortments: A Scalable Deep-Learning Model." *Management Science* 68 (3): 1808–1827.

Gandhi, Amit, and Aviv Nevo. 2021. "Empirical Models of Demand and Supply in Differentiated Products Industries." *Handbook of Industrial Organization*, https://emea.mitsubishielectric.com/ar/products-solutions/factory-automation/index.html.

Google. 2025. "Gemini 3 Announcement." https://blog.google/products/gemini/gemini-3/#note-from-ceo, Accessed: 2025-12-01.

Han, Sukjin, Eric H. Schulman, Kristen Grauman, and Santhosh Ramakrishnan. 2024. "Shapes as Product Differentiation." 1–41, http://arxiv.org/abs/2107.02739.

Handa, Kunal et al. 2025. "Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations." http://arxiv.org/abs/2503.04761.

Hartley, Jonathan S., Filip Jolevski, Vitor Melo, and Brendan Moore. 2025. "The Labor Market Effects of Generative Artificial Intelligence." http://dx.doi.org/10.2139/ssrn.5375017.

Kumar, Madhav, Dean Eckles, and Sinan Aral. 2020. "Scalable bundling via dense product embeddings." (January): 1–47, http://arxiv.org/abs/2002.00100.

Lee, Kevin. 2025. "Generative Brand Choice." 1–60.

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen. 2025. "Triplet Embeddings for Demand Estimation." *American Economic Journal: Microeconomics* 17 (1): 282–307.

Nagle, Frank, and Daniel Yue. 2025. "The Latent Role of Open Models in the AI Economy." *Working paper.*

Nevo, Aviv. 2001. "Measuring market power in the ready-to-eat cereal industry." *Econometrica* 69 (2): 307–342.

OpenAI. 2025. "GPT-5.1 Release Announcement." https://openai.com/index/gpt-5-1/, Accessed: 2025-12-01.

Padilla, Nicolas, H Tai Lam, Anja Lambrecht, and Brett Hollenbeck. 2025. "The Impact of LLM Adoption on Online User Behavior." (August): .

Quan, Thomas W, and Kevin R Williams. 2021. "Extracting Characteristics from Product Images and its Application to Demand Estimation." 53 (9): 1689–1699.

Rao, P. V., and L. L. Kupper. 1967. "Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model." *Journal of the American Statistical Association* 62 (317): 194–204.

Ruiz, Francisco J.R., Susan Athey, and David M. Blei. 2020. "Shopper: A probabilistic model of consumer choice with substitutes and complements." *Annals of Applied Statistics* 14 (1): 1–27.

Singh, Shivalika et al. 2025. "The Leaderboard Illusion." http://arxiv.org/abs/2504.20879.

Tang, Raphael, Crystina Zhang, Wenyan Li, Carmen Lai, Pontus Stenetorp, and Yao Lu. 2025. "Drawing Conclusions from Draws: Rethinking Preference Semantics in Arena-Style LLM Evaluation." http://arxiv.org/abs/2510.02306.

Tomlinson, Kiran, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. 2025. "Working with AI: Measuring the Occupational Implications of Generative AI." 1–42, http://arxiv.org/abs/2507.07935.

Vestager, Margrethe, Sarah Cardell, Jonathan Kanter, and Lina M. Khan. 2024. "Joint Statement on Competition in Generative AI Foundation Models and AI Products." *Federal Trade Commission* 1–3, https://www.ftc.gov/legal-library/browse/joint-statement-competition-generative-ai-foundation-models-ai-products.

Wold, Herman. 1966. "Estimation of principal components and related models by iterative least squares." In *Multivariate Analysis*, edited by Krishnaiah, P.R. 391–420, New York: Academic Press.

Wold, Herman. 1985. "Partial Least Squares." In *Encyclopedia of Statistical Sciences*, edited by Kotz, Samuel, and Norman L. Johnson Volume 6. 581–591, New York: Wiley.

xAI. 2025. "Introducing Grok-4 Fast." https://x.ai/news/grok-4-fast/, Accessed: 2025-12-03.

Online Appendix to "Estimating Consumer Preferences for LLMs: Evidence from LMArena"

A Modelling Ties in LMArena Battles

This appendix describes extensions of the baseline framework that allow for ties in user votes. In the main text, outcomes are restricted to strict preferences between models or selection of the outside option. Here, I outline two approaches that accommodate ties.[1]

Threshold Parameter Approach. A natural way to model ties is to introduce an indifference threshold in the spirit of the threshold-augmented Bradley–Terry model proposed by Rao and Kupper (1967). The key idea is that utility differences below a fixed threshold are not salient and therefore result in a tie. Let $\eta > 0$ denote this threshold parameter.[2] The outcome o of a battle is then determined by

$$o = \begin{cases} m_A & \text{if } U_{ipm_A} - U_{ipm_B} > \eta \text{ and } U_{ipm_A} > U_{ip0}, \\ m_B & \text{if } U_{ipm_B} - U_{ipm_A} > \eta \text{ and } U_{ipm_B} > U_{ip0}, \\ \emptyset & \text{if } U_{ipm_A} < U_{ip0} \text{ and } U_{ipm_B} < U_{ip0}, \\ \{m_A, m_B\} & \text{if } (U_{ipm_A} > U_{ip0} \text{ or } U_{ipm_B} > U_{ip0}) \text{ and } |U_{ipm_A} - U_{ipm_B}| < \eta. \end{cases}$$

where $o = \{m_A, m_B\}$ corresponds to a tied battle. Under this specification, ties arise when at least one response exceeds the outside-option utility, but the difference in utilities between the two responses lies within the indifference band defined by η. This approach provides a structural interpretation of ties as reflecting near-equality in latent utilities.

Random Assignment of Ties. As an alternative, ties can be treated as uninformative about the relative ranking of m_A and m_B. In this approach, when constructing the likelihood, whenever a tie is observed, the outcome is randomly assigned to either m_A or m_B with equal probability. This preserves information about whether responses clear the outside-option threshold U_{ip0}, while ensuring that tied battles do not mechanically favor one model over the other in relative comparisons.

Both approaches are compatible with the baseline framework. The threshold parameter specification offers an extension that explicitly models indifference, whereas the random-assignment approach provides a parsimonious treatment that uses ties simply to inform

[1] Tang et al. (2025) present suggestive evidence that ties are more likely when the user query is too easy or has an objective correct answer.

[2] The threshold parameter can be a function of the prompt embedding $\mathbf{E}_p$.

substitution between inside goods and the outside option.

B Extensions

C Details on the identification strategy

To illustrate how the random coefficients are identified, consider two battles that share a common focal model m on one side but differ in the characteristics of the opposing model. For the sake of the example, assume models differ along two dimensions: size (a proxy for intelligence) and speed. Let the focal model m be large and slow. Suppose that in the first battle, m is paired against a small, fast opponent n, and in the second battle, m is paired against a larger, slower variant n' that is closer to m in product space. Assume that m wins 60 percent of battles in both comparisons.

Now consider a small change in a characteristic of model m, such as a marginal improvement in its size. Under homogeneous tastes, the induced change in the win rate of m, denoted s_m, is governed by the standard multinomial logit structure: if characteristic j of model m is perturbed, then

$$\frac{\partial s_m}{\partial x_{m,j}} = \beta_{x_j}\, s_m(1 - s_m),$$

This expression embodies the IIA restriction: homogeneous tastes imply that the marginal effect of improving m's characteristics must be the same whether m faces n or n'. In practice, this implication is implausible. Consumers who choose model n over model m are likely to have a higher valuation for speed than the average consumer. As a result, increasing m's size should induce a greater increase in its win rate when facing n' than when facing n. The failure of the win-rate responses to align across these two battles thus violates the IIA restriction and reveals underlying heterogeneity in user preferences.

D PLS Algorithm and implementation

This appendix details the Partial Least Squares (PLS) regression algorithm used to construct the low-dimensional prompt representations $h(\mathbf{E}_p)$. Let $\mathbf{E}$ denote the $N \times d$ matrix of prompt embeddings (where N is the number of battles and $d = 3072$) and let $\mathbf{W}$ denote the $N \times (K-1)$ multivariate target matrix defined in Section 5.2. Before estimation, both matrices are centered column-wise to have zero mean.

The objective of the PLS algorithm is to decompose $\mathbf{E}$ and $\mathbf{W}$ into a set of L latent components (scores). Specifically, we seek a set of weight vectors that maximize the *covariance* between the projected prompt embeddings and the target preference proxies. The algorithm operates iteratively. For the first component, we solve the following optimization problem:

$$\max_{\mathbf{u},\mathbf{v}} \mathrm{Cov}(\mathbf{E}\mathbf{u}, \mathbf{W}\mathbf{v})^2 \quad \text{s.t. } \|\mathbf{u}\| = 1, \|\mathbf{v}\| = 1 \tag{8}$$

where $\mathbf{u}$ and $\mathbf{v}$ are weight vectors with lengths d and $K - 1$, respectively. The first latent component (PLS score) for the prompts is given by $\mathbf{z}_1 = \mathbf{E}\mathbf{u}$.

The algorithm then deflates the matrices $\mathbf{E}$ and $\mathbf{W}$ by subtracting the information explained by $\mathbf{z}_1$ and repeats the procedure on the residuals to find subsequent orthogonal components. I set the number of iterations, and therefore PLS scores, to three. The final output is an $N \times 3$ matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3]$. The row corresponding to battle b with prompt p, denoted $\mathbf{Z}_b$, corresponds to the vector $h(\mathbf{E}_p)$ used in the utility specification.

The estimation is performed using the Non-Linear Iterative Partial Least Squares (NIPALS) algorithm as implemented in the `scikit-learn` library in Python.

E Details on the classification approach

This appendix describes the classification procedures used to assign conversation topics and evaluate the tone of LLM responses along two dimensions: sentiment and hedging. All classifications are performed using the `openai/gpt-oss-20b` model.

Model Description. `openai/gpt-oss-20b` is a high-performance open-weight language model released by OpenAI in August 2025 under the Apache 2.0 license. It was trained using a mix of reinforcement learning and techniques informed by OpenAI's advanced internal systems, including o3 and other frontier models. The model delivers results comparable to `o3-mini` on standard benchmarks. Its combination of strong reasoning performance and low deployment cost makes it suitable for large-scale classification tasks. This is particularly relevant for classifying LLM responses, which is more computationally costly than classifying user prompts due to their substantially higher token counts.

Classification procedure. For topic classification, the LLM receives each user-submitted prompts.[3] For tone classification (sentiment and hedging), the model is instead fed the text of the LLM-generated responses. In each case, classification is conducted by passing the relevant text into `gpt-oss-20b` along with task-specific instructions. To ensure deterministic output, all classification runs are performed with a temperature of zero. Each prompt or response is classified independently.

System Prompts The system prompts used to instruct the model in each classification task are provided at the links below:

- `topic_classification_prompt.txt`

[3] For multi-turn conversations, I concatenate all user-submitted prompts into a single input.

- sentiment_classification_prompt.txt
- hedging_classification_prompt.txt

Each file contains the full system prompt and classification schema used in that task.

F Supplementary Tables and Figures

Table A1: Conversation Topic Categories

Category	**Definition**	**Example**
Coding	Prompts involving writing, modifying, debugging, or explaining code or scripts in any programming language. Includes code snippets, error messages, stack traces, and requests about syntax.	"Why is my SQL query failing? Here's the code and error message."
Information Seeking	Prompts asking factual questions such as definitions, explanations, statistics, comparisons, or historical/background information.	"Who discovered penicillin?"
Practical Guidance	Prompts requesting how-to advice, step-by-step instructions, troubleshooting (non-code), or help designing/creating non-text things. Also includes life advice, planning, and product recommendations.	"How do I set up Docker on Windows?"
Writing	Prompts focused on producing, editing, rewriting, summarizing, translating, or improving written text of any kind.	"Rewrite this email to sound more professional."
Math	Prompts requiring numerical calculations, symbolic manipulation, proofs, algebra, calculus, probability, or solving logic puzzles.	"Solve for x: $3x + 5 = 20$."
Multimedia	Prompts requesting the creation, editing, or analysis of images or other non-text media.	"Can you please draw me a unicorn in ASCII."
Casual Conversational	Prompts involving greetings, chit-chat, personal reflections, emotions, philosophical or hypothetical musings, or playful conversation not aimed at producing factual answers or instructions.	"What if humans could live forever?"
Other	Prompts that do not fit the other categories, mainly questions about the AI assistant itself—its training, internal state, limitations, or identity.	"How were you trained?"

Note: In the LMArena text battles, models cannot generate images. When users request images, models typically provide textual descriptions or produce text-based renderings (e.g., ASCII art or character-based drawings).

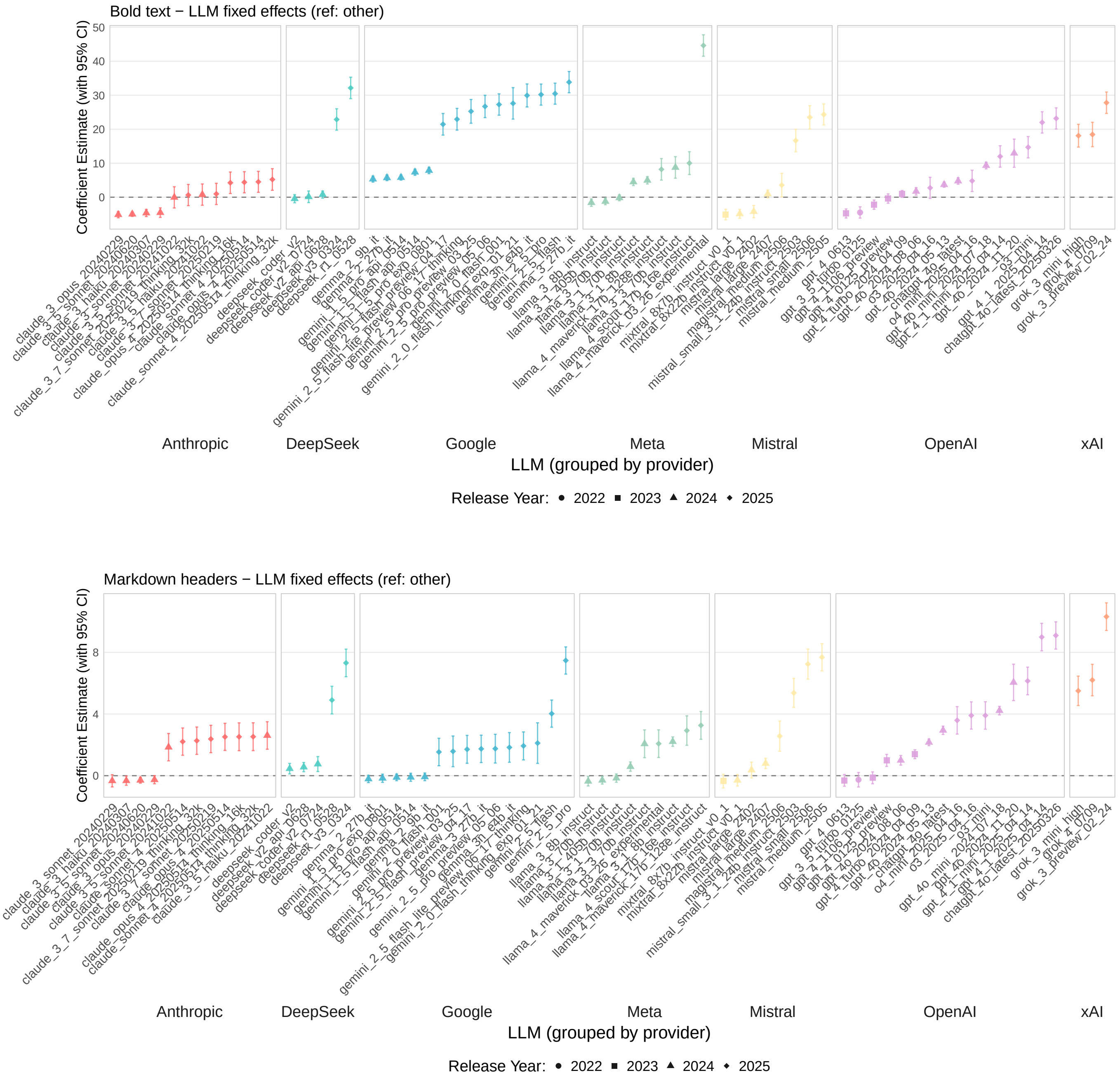


Figure A1: Formatting characteristics by LLM: Number of text blocks in bold and number of markdown headers

Table A2: Benchmarks included in the Artificial Analysis Intelligence Index

Benchmark	Field	Description	Paper	Weight
MMLU-Pro	Reasoning & Knowledge	Comprehensive evaluation of advanced knowledge across domains, adapted from original MMLU. 12,032 multiple-choice questions across math, science, law, economics, psychology, business, etc.	Wang et al. (2024)	1/6
HLE (Humanity's Last Exam)	Reasoning & Knowledge	Challenging academic benchmark across mathematics, humanities, and natural sciences.	Phan et al. (2025)	1/6
GPQA Diamond	Scientific Reasoning	Graduate-level scientific reasoning benchmark (biology, physics, chemistry). Diamond subset (198 questions) selected for discriminative power.	Rein et al. (2023)	1/6
MATH-500	Quantitative Reasoning	Subset of 500 problems from Hendrycks' MATH dataset (mathematical problem-solving). Uses symbolic and LLM-based equality checking.	Hendrycks et al. (2021) Lightman et al. (2023)	1/8
AIME 2024	Competition Math	Advanced competition-level math problems from the 2024 American Invitational Mathematics Exam.		1/8
SciCode	Code Generation	Python programming for scientific computing.	Tian et al. (2024)	1/8
LiveCodeBench	Code Generation	Python code generation benchmark with tasks from LeetCode, AtCoder, and Codeforces.	Chen et al. (2024)	1/8

Table A3: Choice sets

Panel A: July 2024 choice set $\mathcal{J}_0$

Provider	Models
OpenAI	gpt_4o, gpt_4o_mini (2024-07-18)
Anthropic	claude_3_opus, claude_3_5_sonnet, claude_3_haiku
Google	gemini_1_5_pro, gemini_1_5_flash
DeepSeek	deepseek_v2, deepseek_coder_v2
Meta	llama_3_1
Mistral	mistral_large

Panel B: July 2025 choice set $\mathcal{J}_1$

Provider	Models
OpenAI	gpt_4o (2025-03-26), gpt_4_1, gpt_4_1_mini, o3, o3_mini, o4_mini
Anthropic	claude_opus_4, claude_opus_4 (thinking), claude_sonnet_4, claude_sonnet_4 (thinking)
Google	gemini_2_5_pro, gemini_2_5_flash
DeepSeek	deepseek_r1, deepseek_v3
Meta	llama_4_maverick, llama_4_scout
Mistral	mistral_medium, mistral_small, mistral_large
xAI	grok_4, grok_3, grok_3_mini

Notes: Grok 2 and Grok 2 mini by xAI are excluded from Panel A because they do not appear in the LMArena dataset.