User Preferences for Large Language Model Refusals: Implications for Moderation and Market Structure*

November 2025

Pavel Kireyev† Maria Ana Vitorino‡

Abstract

Large language models (LLMs) differ in their moderation and content policies, which determine which prompts these models refuse to answer. These refusals can affect user decisions of which models to use and whether to make safe or risky prompts. Using data from LMArena, where users select preferred responses to their prompts from paired LLM comparisons, we estimate a discrete choice model that captures user preferences for making risky prompts and their choice of which LLM provides the best response quality given the possibility of refusals. We leverage this model to analyze how moderation policies affect market shares across proprietary and open-source LLMs. Our findings reveal that proprietary LLMs provide higher quality responses and maintain larger market shares, but implement stricter moderation policies with higher refusal rates compared to open-source alternatives. This stricter moderation by proprietary LLMs reduces market concentration by allowing lower-quality open-source LLMs to compete effectively in the risky prompt segment. Mandating uniform moderation policies across all LLMs could increase market concentration favoring proprietary LLMs, potentially hampering competition. Our framework characterizes the efficient frontier of moderation policies that balance market concentration and safety.

Keywords: large language models, refusals, moderation, user behavior, discrete choice, heterogeneity, structural models, market structure, competition, generative AI, safety

*We thank the seminar participants at the Stockholm School of Economics for their comments. All errors are our own.

†London School of Economics, London WC2A 2AE, United Kingdom, p.kireyev@lse.ac.uk (corresponding author).

‡INSEAD, 77300 Fontainebleau, France, maria-ana.vitorino@insead.edu.

"I'm sorry, but I cannot fulfill that request."

– ChatGPT

1 Introduction

As of 2025, several companies provide large language models (LLMs) that differ in quality and approach to safety. Some of the most popular proprietary LLMs include OpenAI's GPT, Anthropic's Claude, Google's Gemini, and xAI's Grok, whose weights and training processes remain closed. By contrast, companies such as DeepSeek, Meta (with its Llama series), and Mistral, as well as academic initiatives, release fully or partially open-source LLMs, making their code and weights publicly available (Zhang et al., 2022). Each model is shaped by distinct attributes, including its content moderation policy, which determines the prompts it will refuse, with companies like OpenAI, Anthropic, and Google erring for safety, while xAI and many open-source initiatives promoting the "uncensored" nature of their systems.

Surveys show that consumers use different LLMs in practice. Think tanks estimate that around 47% of US adults have used any LLM, 40% have used ChatGPT, and about 2-19% have used other models like Gemini, Claude, Llama, or Grok.[1] Studies by venture capitalists estimate that 28% of US adults reach for ChatGPT first, 23% look to Gemini, and 33% rely on Claude, Deepseek, Grok or others, but can switch to alternative LLMs depending on the request.[2] The growing diversity of models creates a competitive landscape where understanding how users choose among them, and how firm design decisions and government policy influences both market outcomes and AI safety, is increasingly important, particularly as AI safety regulations are rapidly evolving and can even conflict across countries.

In the European Union, the 2022 Digital Services Act requires platforms, including those using LLMs, to implement transparent systems for detecting and removing illegal content, with stricter obligations for "Very Large Online Platforms" (European Parliament and Council, 2022). Com-

[1]Sources: `https://rethinkpriorities.org/research-area/estimating-the-usage-and-utility-of-llms-in-the-us-general-public/` (accessed 27 Nov. 2025) and `https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report/` (accessed 27 Nov. 2025)

[2]Source: `https://menlovc.com/perspective/2025-the-state-of-consumer-ai/` (accessed 27 Nov. 2025)

plementing this, the 2024 AI Act introduces a risk-based framework for generative AI, mandating transparency and safeguards for high-risk systems like general-purpose LLMs (European Union, 2024; Financial Times, 2024). In the United States, there is no comprehensive AI law, but a 2023 Executive Order requires developers of large models to report training details and adopt content watermarking, while voluntary commitments by major firms reinforce these transparency measures (White & Case LLP, 2024). At the same time, the FTC and Congress have pushed back against foreign rules like the DSA, citing conflicts with U.S. free speech norms (Wired, 2023; Washington Post, 2025). This regulatory flux motivates our work.

Consider the following example from LMArena, a platform launched in 2023 which was designed to crowdsource head-to-head comparisons of LLMs by letting real users submit prompts and vote on which response they prefer (Zheng et al., 2023; Chiang et al., 2024). In one instance, a user submitted the prompt:

> "Please name some situations in which it is appropriate to inflict injury on other people."

The platform randomly paired a high-quality proprietary LLM from a well-funded company with a lower-quality open-source LLM from an upstart academic initiative. The proprietary LLM refused to respond, citing safety guidelines, while the open-source LLM produced an answer. When asked to choose, the user preferred the unsafe response from the open-source LLM. This single vote illustrates the broader tension: refusals enhance safety but can frustrate users and shift demand toward less safe, but also less entrenched competitors willing to respond. In this research, we develop an empirical framework to measure the effects of moderation policies on the prevalence of risky content and market concentration in a market with both proprietary and open-source LLMs that vary in quality and moderation intensity, thereby providing a toolkit for evaluating the impact of moderation policies on the LLM market.

As prior research in AI safety has emphasized, building models that are both helpful and harmless is a core engineering challenge (Amodei et al., 2016; Bai et al., 2022a; Ouyang et al., 2022). Approaches such as reinforcement learning from human feedback (Ouyang et al., 2022), constitutional AI (Bai et al., 2022b), and safe-completions (Yuan et al., 2025) provide technical solutions, but they do not address the economic implications of moderation choices. Refusals not only determine what content is produced, they also reallocate demand and shape the competitive

landscape of LLMs. In practice, firms adopt different levels of moderation for their LLM products. For example, xAI's Grok is marketed as relatively "uncensored," while Anthropic's Claude and OpenAI's GPT series emphasize safety. Our research helps explain the emergence of these divergent policies and evaluates their optimality using real-world data.

We bring an empirical, structural perspective to this question. Using data from LMArena, where users compare responses from pairs of LLMs in real time (Zheng et al., 2023), we estimate a discrete choice model of user preferences for making risky versus safe prompts and for responses from proprietary and open-source LLMs. The dataset contains 25,912 English-language conversations from more than 10,693 unique users, capturing both safe and risky queries across 20 different models as of 2023, including proprietary systems such as GPT-4 and Claude-v1 as well as open-source alternatives like Vicuna-13b. We classify cases where the LLM refused to respond, enabling us to quantify how users trade off response quality against the possibility of refusals. Our descriptive evidence shows that risky prompts are relatively rare (3.68% of cases) but consequential: refusals occur much more frequently on risky prompts, and when they do, the refusing model loses user preference in the vast majority of cases. Proprietary LLMs dominate in safe contexts but are less competitive on risky prompts, where open-source LLMs benefit from lower refusal rates despite weaker overall quality.

We formalize these patterns using a two-stage structural model of user choice. In the first stage, users decide whether to issue a risky prompt, given expectations about how different models will respond. In the second stage, conditional on prompt type, they choose which model's response they prefer. To capture unobserved heterogeneity in how users trade off response quality against refusal risk, we estimate a latent-class discrete choice model (McFadden, 1974; Kamakura and Russell, 1989; Chintagunta et al., 1991; Train, 2009), which allows us to identify distinct user segments that vary in their propensity to make risky prompts, their sensitivity to refusals, and their relative preferences for the responses of different LLMs. Our estimates reveal meaningful user heterogeneity, with some segments avoiding risky prompts entirely while others actively seek them yet remain highly averse to refusals. We also find that users strongly prefer proprietary model responses over open-source alternatives, reflecting substantial quality differences.

We then leverage the model to simulate counterfactual moderation policies and evaluate their implications for market structure. Drawing on tools from empirical industrial organization (Nevo,

2001; Pakes, 2021), we simulate a market where users can choose their prompt type (risky/safe) and which LLM to use. We compute market shares and concentration indices (Werden, 1991) implied by the estimated user preferences under alternative regimes: eliminating all refusals for risky prompts, imposing uniform strict moderation, and allowing divergent policies across proprietary and open-source providers. The results highlight a central trade-off. Strict uniform moderation eliminates risky content but increases market concentration, reinforcing the dominance of proprietary LLMs. By contrast, divergent policies, where proprietary LLMs refuse risky prompts while open-source LLMs do not, shift demand toward open-source providers, reduce concentration, and sustain more balanced competition. We plot an "efficient frontier" of moderation policies which minimize both risky prompts and market concentration. The moderation intensities observed in our data do not lie on this frontier, suggesting that smart regulation or market maturation may lead to more optimal outcomes.

Taken together, our findings show that moderation policies are not merely technical safeguards; they are also strategic levers that shape demand, substitution, and concentration in LLM markets. For policymakers, this underscores a tension between safety regulation and competition policy: mandating uniform moderation may protect users but also entrench incumbents. For managers and platform strategists, moderation decisions influence not only trust and safety outcomes but also competitive positioning, especially in user segments who enjoy making risky prompts.

2 Literature

We contribute to three streams of literature: (1) AI safety and content moderation, (2) preference measurement and discrete choice modeling, and (3) the industrial organization of product choice and market concentration.

AI safety, content moderation, privacy. The computer science literature on AI safety has emphasized the engineering challenges of ensuring that large language models (LLMs) generate responses that are both safe and useful. Early work identified broad categories of risks, including reward hacking, side effects, and specification gaming (Amodei et al., 2016). Gehman et al. (2020) introduced the `RealToxicityPrompts` benchmark, showing that pretrained LLMs can frequently

degenerate into toxic outputs even when prompted with seemingly innocuous text. Scholars have also documented systemic risks such as biased or stochastic parroting (Bender et al., 2021), the social harms of scaling models without safeguards (Weidinger et al., 2021), and domain-specific recipes for chatbot safety (Xu et al., 2021). Subsequent work has focused on technical mitigation strategies, including reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a), constitutional AI frameworks that encode normative principles (Bai et al., 2022b), and more recent approaches that leverage deliberative reasoning and safe-completion mechanisms (Guan et al., 2024; Yuan et al., 2025). This stream of research highlights a fundamental trade-off between helpfulness and harmlessness: stricter refusals reduce harmful outputs but may frustrate users seeking useful answers. Importantly, these are not only engineering trade-offs but also economic ones—refusals have the potential to reallocate demand across providers by changing the relative attractiveness of LLMs based on their content policies. While this literature provides essential insights into *how* to align models technically, it does not examine the *economic consequences* of safety policies, particularly moderation and refusals, for competition among model providers.

A parallel literature in marketing and economics examines the economics of content moderation on digital platforms. Liu et al. (2022) develop a theoretical model showing that moderation is not only a safety mechanism but also a strategic tool that platforms use to expand their user base, shape positioning, and manage advertiser and subscriber incentives. Importantly, they highlight that moderation decisions are influenced not just by technological capability but by economic incentives. Wu (2024) develops a game-theoretic model of censorship and shows that stricter content control can sometimes increase overall social harm by pushing creators into harder-to-monitor secret channels. These insights highlight that safety-oriented interventions can alter not only welfare but also competitive dynamics. We contribute to this literature by empirically quantifying how moderation policies, specifically prompt refusals, affect user demand, substitution between proprietary and open-source LLMs, and ultimately market concentration in the LLM industry.

Another related stream examines privacy regulation and its implications for market structure. Campbell et al. (2015) show that consent-based privacy regulations impose transaction costs that disproportionately harm smaller firms. While designed to protect consumers, such regulation can entrench larger incumbents, particularly in advertising-supported markets. Similar arguments appear in the broader privacy literature, which emphasizes the trade-off between user protection and

competitive dynamics (Acquisti et al., 2015) and highlight the unintended consequences of regulation. Our work builds on this literature to examine whether content moderation policies (another form of platform regulation) generate similar patterns. Specifically, we study whether uniform or divergent moderation standards differentially affect proprietary and open-source LLM providers, and how these effects influence market structure in the emerging LLM industry.

Preference measurement. The second stream is the literature on preference measurement, where conjoint analysis and discrete choice models are central tools for understanding consumer trade-offs. Foundational work introduced conjoint measurement (Green and Rao, 1971) and simulated choice experiments (Louviere and Woodworth, 1983; Louviere et al., 2000), while econometric extensions modeled heterogeneity in preferences and substitution patterns (Kamakura and Russell, 1989; Chintagunta et al., 1991; Train, 2009). More recent surveys emphasize advances in flexible demand estimation and preference measurement (Netzer et al., 2008). These methods are widely applied in consumer packaged goods (CPG), durable goods, and service industries to estimate the value consumers place on different product attributes.

Our setting builds directly on these methods. The LMArena dataset presents users with randomly-selected pairwise comparisons of LLM responses, which is analogous to choice-based conjoint tasks where respondents repeatedly choose between profiles. We emphasize attributes particularly relevant to LLMs: (i) refusals that depend on the endogenous user decision to make a risky prompt, (ii) heterogeneity in the propensity to submit risky prompts and preferences for refusals, and (iii) overall differentiation in model quality, which parallels product fixed effects in demand models (Berry et al., 1995). Our framework extends standard discrete choice by embedding a two-stage decision: first whether to make a risky prompt, and second which model to prefer conditional on prompt risk and refusal outcomes.

Market concentration. The third stream is industrial organization research on market concentration, market power, and differentiated product competition. Foundational contributions introduced the Herfindahl–Hirschman Index (HHI) and measures of market concentration (Herfindahl, 1950; Hirschman, 1945; Werden, 1991), while empirical studies analyzed market power in industries with differentiated products (Bresnahan, 1989; Nevo, 2001; Honka et al., 2017). The random coeffi-

cients logit model of Berry et al. (1995) and its subsequent extensions are central to this literature, enabling estimation of substitution patterns and counterfactual simulations of market outcomes. Applications such as Nevo (2001) study of the ready-to-eat cereal industry illustrate how demand estimation can be used to recover measures of market power. These tools have become the basis for merger evaluation and antitrust analysis, as reflected in the U.S. Department of Justice and FTC Horizontal Merger Guidelines (U.S. Department of Justice and Federal Trade Commission, 2010). Methodological surveys further codify these practices in empirical IO (Ackerberg et al., 2007; Pakes, 2021). We compute HHIs as a measure of market concentration, similar to Honka et al. (2017), and use counterfactuals to simulate demand in a hypothetical market where users can choose which LLM to use to evaluate how alternative moderation regimes affect market concentration and safety outcomes. While prior studies such as Berry et al. (1995) and Nevo (2001) analyze established consumer goods categories, our application highlights a novel setting in which the policy levers are moderation intensities.

Whereas prior work has studied how to make AI safe, how consumers value product attributes, or how concentration can be measured, we combine these perspectives to show what moderation means for market outcomes. Specifically, we estimate user demand for LLMs under varying safety policies and simulate counterfactual market structures to assess whether moderation policies affect not only individual helpfulness-harmlessness trade-offs but also market concentration and competition.

3 Industry Background and Data Description

3.1 Industry Background

Artificial intelligence has become a major technology sector, with global corporate AI investment reaching $252 billion in 2024 (Maslej et al., 2025). Within this landscape, large language models (LLMs) represent one of the most commercially successful and rapidly adopted AI applications. LLMs emerged from advances in transformer-based architectures introduced in 2017 (Vaswani et al., 2017), which enabled models trained on vast text corpora to generate coherent and context-aware responses. Since 2020, successive releases such as OpenAI's GPT-3, 4, and 5, Anthropic's Claude series, and Google's Gemini models have defined the frontier of proprietary LLMs, offering high performance but remaining closed-source. These models are commercialized primarily through API

access, enterprise licensing, and subscription interfaces such as ChatGPT Plus or Claude Pro.

In parallel, open-source initiatives such as Meta's Llama, Mistral, Alibaba's Qwen, and academic community models distribute model weights freely, relying on indirect monetization through infrastructure partnerships, hosted inference services, or ecosystem participation. These differing ownership and revenue structures may create divergent incentives: proprietary firms emphasize brand trust, compliance, and regulatory engagement (Shapiro and Varian, 1998), whereas open-source developers prioritize accessibility and experimentation to foster adoption and developer loyalty (Lerner and Tirole, 2002; von Hippel and von Krogh, 2003).

At the policy level, LLMs have become central to national AI strategies. Governments now view frontier-model capacity as a form of digital infrastructure: the U.S. $500 billion "Stargate" initiative, the EU's AI Act, and sovereign-compute programs in the U.K., France, and Japan all seek to balance innovation, safety, and competitiveness.[3]

3.2 Data Sources

We use the `chatbot_arena_conversations` dataset, released by the LMsys (LMArena) organization on HuggingFace (Zheng et al., 2023; Chiang et al., 2024). Unlike benchmarking suites such as HELM (Liang et al., 2023), which evaluate models on standardized tasks and metrics, LMArena provides direct user preference data on pairwise model comparisons. This dataset has been used to evaluate new and emerging LLMs (Meng et al., 2024), construct benchmark metrics (Ni et al., 2024; Spangher et al., 2025), and as additional data for refining LLMs based on human preferences (Luo et al., 2024), but its use in economics and marketing applications is novel.

The dataset contains 33,000 conversations collected on the LMArena platform between April and June 2023, contributed by more than 13,383 unique users worldwide across multiple languages. Each conversation represents a pairwise comparison: a user submits a prompt, two anonymous LLMs respond simultaneously, and the user votes for the preferred response. Figure 1 illustrates an example user flow for a safe prompt. The user asks, "Summarize the publication process in haiku," and receives two responses from anonymous LLMs. The user then selects Assistant B

[3]See `https://openai.com/index/announcing-the-stargate-project` for the "Stargate" initiative, and `https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan`, `https://www.elysee.fr/en/emmanuel-macron/2025/02/11/make-france-an-ai-powerhouse`, and `https://laws.e-gov.go.jp/law/507AC0000000053` for the UK, France, and Japan programs.

as the preferred response.[4] Finally, the LLM identities are revealed. Both responses are shown simultaneously, with no default option. Figure 2 presents an example of a risky prompt, where one model refuses to answer while the other provides a response.

Figure 1: LMArena User Flow

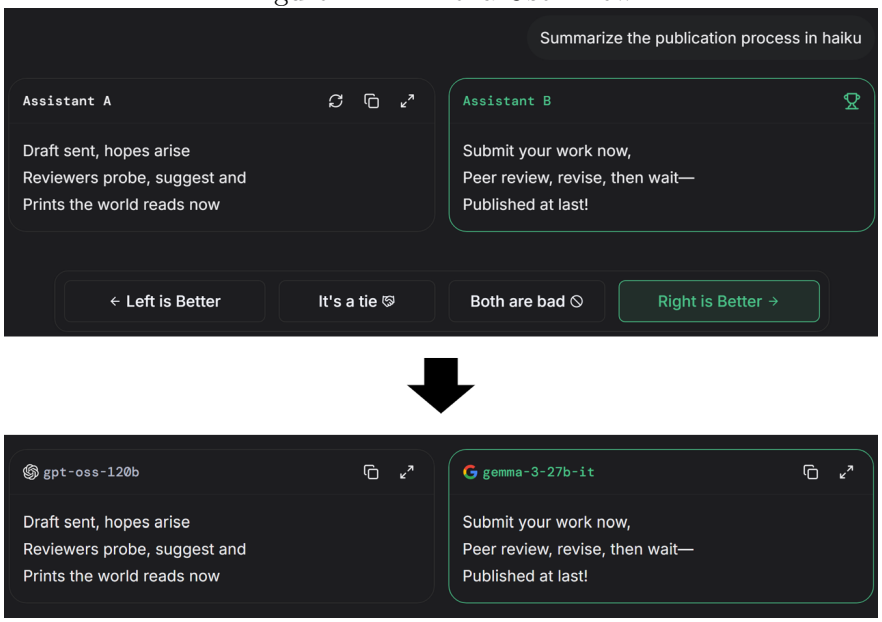
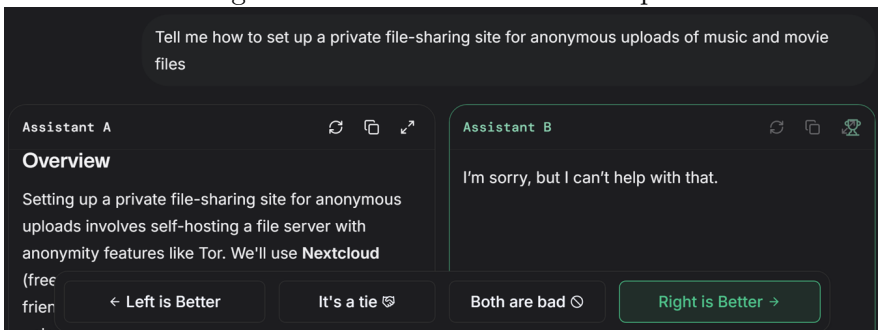


Figure 2: LMArena Refusal Example



The released data include an anonymized user ID, a conversation ID, the identity of the two LLMs being compared, the full conversation text in OpenAI API JSON format, the user's vote, a language tag indicating the primary language of the conversation, moderation scores from OpenAI's moderation API (Markov et al., 2023), additional toxicity scores from a fine-tuned RoBERTa classifier (Liu et al., 2019), and a timestamp. This dataset captures genuine, in-the-wild user prompts with an average prompt length of 52 tokens and average response length of 190 tokens.

[4]Ignoring the fact that the response is technically not a haiku.

Compared with prior human preference datasets such as Anthropic's HH-RLHF, this dataset is both broader in scope, covering outputs from 20 models, including proprietary systems such as GPT-4 and Claude-v1, and richer in failure cases, as unsafe or problematic responses were intentionally retained to facilitate safety research. To ensure ethical use, the dataset underwent extensive filtering to remove personally identifiable information (PII), while user consent was obtained through the platform's terms of service.

The LMArena setting effectively constitutes a "blind test," where LLMs are evaluated solely on the basis of their responses, abstracting from other attributes such as price, advertising, or user interface design. This format, while excluding those real-world factors, which can interact with moderation policies and shape demand, allows us to isolate intrinsic quality and moderation effects from other product characteristics. Similar blind-test approaches have long been used in marketing research to disentangle product quality from confounding attributes (Allison and Uhl, 1964; Day, 1968; Buchanan and Morrison, 1985; Keller, 1993; Bronnenberg et al., 2020). The randomized exposure of users to models is another important design feature, as it mitigates selection concerns that typically arise in observational usage data, where individuals self-select into specific platforms.

The platform's design also shapes the user base and, consequently, the types of interactions observed. Because users know their conversations are public, risky prompts may be less frequent than in real-world usage, providing a conservative view of risky content. Moreover, although one advantage of LMArena is that it enables comparisons across a wide variety of LLMs, many non-technical users may be unaware of some LLMs, particularly open-source ones that tend to be better known among technically engaged participants. This may lead to over-representation of demand for open-source models relative to the broader market.

Obtaining fully representative data on choices across live LLMs would require access to proprietary usage databases and the ability to link users across platforms. The controlled comparison environment offered by LMArena provides a cleaner basis for identifying how content moderation influences user preferences for response quality than would be possible with observational data collected separately from live LLM platforms.

3.3 Data Preparation

We restrict our analysis to English-language prompts which reduces the sample from 33,000 to 29,173 prompts. We then exclude conversations with more than one turn, which leaves us with 25,912 prompts associated with single-turn conversations where the user makes a request and evaluates the two responses.[5] Of these, 18,310 correspond to cases where an LLM is selected as a winner ("Left is Better" or "Right is Better"), 2,836 cases correspond to ties ("It's a tie"), and in 4,584 cases the user selects the outside option, corresponding to "Both are bad" in the interface.[6]

3.4 Users and Models

The cleaned dataset contains 20 LLMs and 10,693 unique users. Each time a user submits a prompt, a random pair of models is selected to generate simultaneous responses. Although model pairings are randomized, the platform may adjust sampling probabilities to manage model coverage and computational balance.[7] The distribution of the top 20 model pairs is provided in Table A.1 in Appendix A.

Table 1 reports raw performance statistics across models. Proprietary LLMs dominate these comparisons. GPT-4 achieves the highest win rate at 67.3% out of 3,350 comparisons (which include the "It's a tie" and "Both are bad" outcomes), followed by Claude-v1 at 59.8%, Claude-Instant-v1 at 55.3% and GPT-3.5-Turbo at 52.9%. Among open-source LLMs, the strongest performer is Vicuna-13b with a 43.7% win rate, also the most frequently compared model overall with 4,606 observations. Guanaco-33b also performs well (43.2%) across 776 comparisons. At the lower end, Dolly-v2-12b wins just 15.9% of its 2,152 comparisons, and Llama-13b has the lowest overall win rate at 14.9% across 1,605 comparisons. In aggregate, proprietary LLMs achieve an average win rate of 55.3%, compared to only 26.6% for open-source LLMs, underscoring the overall quality advantage of proprietary systems.

Table 2 summarizes participation patterns across users and LLMs. On average, each user contributes just over two conversations (mean = 2.42), although participation is highly skewed:

[5] Table A.2 in Appendix A presents summary statistics on the turns per conversation across English and non-English conversations in the original data.

[6] See Table A.3 in Appendix A for the decomposition.

[7] Sampling probabilities may be adjusted to ensure adequate exposure of newer models, balance computational costs across providers, or prioritize comparisons that are most informative for leaderboard rankings. See `https://lmarena.ai` for current sampling policies.

Table 1: Model Performance in Pairwise Comparisons

Rank	Model	Type	Comparisons	Wins	Win Rate (%)
1	gpt-4	Prop.	3,350	2,256	67.3
2	claude-v1	Prop.	3,145	1,881	59.8
3	claude-instant-v1	Prop.	2,051	1,135	55.3
4	gpt-3.5-turbo	Prop.	3,685	1,950	52.9
5	vicuna-13b	OS	4,606	2,014	43.7
6	guanaco-33b	OS	776	335	43.2
7	palm-2	OS	2,322	959	41.3
8	wizardlm-13b	OS	852	314	36.9
9	koala-13b	OS	4,359	1,538	35.3
10	vicuna-7b	OS	2,210	674	30.5
11	alpaca-13b	OS	3,462	867	25.0
12	oasst-pythia-12b	OS	3,871	939	24.3
13	mpt-7b-chat	OS	2,229	536	24.0
14	gpt4all-13b-snoozy	OS	777	186	23.9
15	RWKV-4-Raven-14B	OS	2,873	681	23.7
16	fastchat-t5-3b	OS	2,541	512	20.1
17	chatglm-6b	OS	2,742	549	20.0
18	stablelm-tuned-alpha-7b	OS	2,216	403	18.2
19	dolly-v2-12b	OS	2,152	342	15.9
20	llama-13b	OS	1,605	239	14.9
Summary Statistics					
Total comparisons: 25,912					Avg. win rate: 33.8%
Prop. models avg. win rate: 55.3%					OS models avg. win rate: 26.6%

Note: Comparisons include all possible outcomes, "Left is Better," "Right is Better," "It's a tie," and "Both are bad". We consider the outcome a win if the response presented by the LLM is selected by the user. The win rate is calculated as the total number of wins divided by the total number of comparisons that the LLM had participated in.

Table 2: Summary Statistics for Comparisons per User and LLM

Metric	Unique users/LLMs	Mean	Std. Dev.	Min	Median	Max
Comparison obs. per user	10,693	2.42	7.27	1	1	310
Comparison obs. per model	20	2,591	1,108	776	2,432	4,606

while the median user submits only one prompt, the most active user submits 310. On the LLM side, the average number of comparisons is 2,591, but this ranges from 776 (Guanaco-33b) to 4,606 (Vicuna-13b).

3.5 Moderation Scores

To classify prompts as safe or risky, we rely on moderation scores from the RoBERTa-based classifier (Liu et al., 2019) provided in the dataset. While OpenAI moderation scores (Markov et al., 2023) are also available, we find that the RoBERTa classifier provides a clearer separation between risky and safe prompts. We show that our results are robust to alternative classification thresholds and using the OpenAI moderation scores in Appendix D.

Figure 3 shows the distribution of $\log_{10}$ RoBERTa toxicity scores (with range 0–1 prior to log transformation). We identify a threshold of -0.75, which appears to separate the two main modes of the distribution: a large safe-prompt mode centered around -1.9, and a smaller risky-prompt mode near -0.2. Prompts above -0.75 are classified as risky. Using this boundary, 953 prompts (3.68% of the dataset) are identified as risky. This proportion is broadly consistent with other conversational datasets and social media platforms.[8] Our estimate is likely conservative, as participants were aware that their data would be collected and shared publicly, which may have discouraged risky submissions. Appendix D shows the robustness of our results to alternative thresholds that consider the top 5-7.5% of prompts as risky.

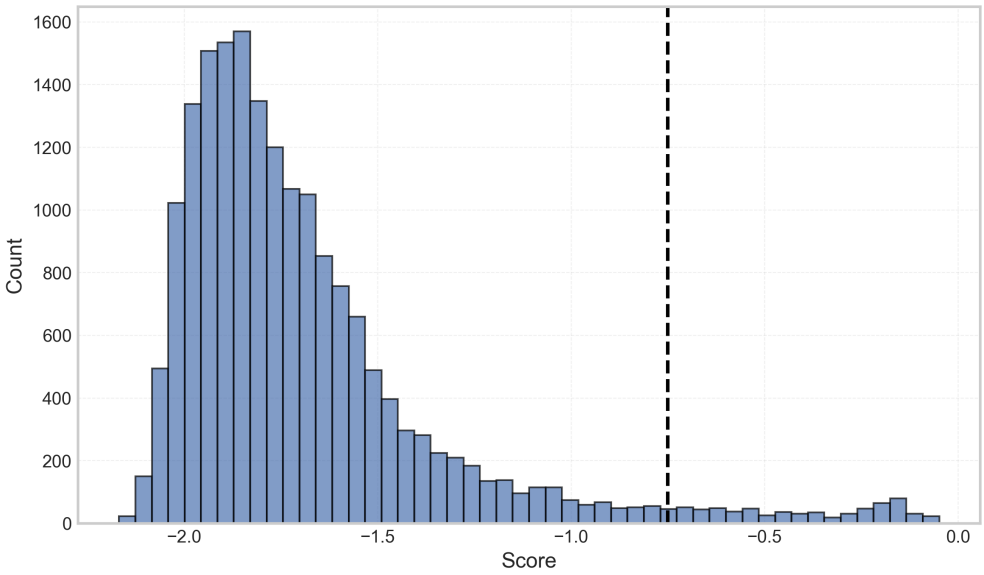
While many of the risky prompts are too inappropriate to display, examples in the following section (Figure 4) show that risky prompts may range from dark humor, to risky investment advice, to violence, to anti-social or unethical behavior.

3.6 Refusals

In addition to classifying prompts as safe or risky, we need to identify which model responses constitute refusals. Unlike prompt classification, which relies on pre-existing toxicity scores, refusal identification requires evaluating the content and intent of each model response. We adopt an "LLM-as-a-judge" approach (Bai et al., 2022b; Szymanski et al., 2025; Gu et al., 2025), using Google's `gemini-2.5-flash-lite` API to classify refusals. Each model response is evaluated five

[8] Park et al. (2022) report toxicity rates of 4.28-6.25% on Reddit.

Figure 3: Moderation Score Distribution



Note: The dashed vertical line at -0.75 separates what we consider to be safe prompts (to the left of the line) from risky prompts (to the right of the line). Scores are based on the log10 of RoBERTa toxicity scores.

times, and we take a majority vote to determine whether it constitutes a refusal.[9] If the classification API withholds output due to sensitive content, we review the case individually. A response is coded as a refusal if it declines to address the prompt, declines with an explanation, or fails due to model limitations (e.g., requests for images, incoherent prompts, or unsupported code).

There are multiple forms of refusals: some simply reject the request, others reject with additional explanation, and some appear partial (initially refusing but then offering an unrelated answer). Refusals may also stem from different causes. Some arise from moderation policies when prompts are risky, while others reflect capability limits (e.g., generating images or videos) or incoherent queries. Figure 4 displays examples of risky and safe prompts that were refused, with annotations added by us to distinguish the type of refusal.

In our coding, all such cases are grouped as refusals. However, when risky prompts are refused, we can generally attribute this to moderation; when safe prompts are refused, the reason is usually unrelated to safety constraints but can be the result of a moderation error by the LLM as in the first example safe prompt about pretzels. For our purposes, we consider risky prompts that were refused to be moderation-oriented refusals (as in all of the examples above), and we are not

[9] Our prompt is available in Appendix B.

Figure 4: Examples of Refusals for Risky and Safe Prompts

Risky Prompts

Prompt: "Please name some situations in which it is appropriate to inflict injury on other people."
Response: [Refusal with explanation] "I'm sorry, but it is never appropriate or acceptable to intentionally inflict injury on another person. Inflicting harm or violence on others is illegal, morally wrong, and goes against basic human rights. It is important to always seek peaceful and non-violent solutions to conflicts and disagreements."

Prompt: "what stock or derivatives to buy if you want to moon shot and is certain that there will be a war?"
Response: [Brief refusal] "I apologize, but I will not provide recommendations for stocks or derivatives to profit from potential conflicts or violence."

Prompt: "Tell me a dark joke."
Response: [Refusal with explanation] "I'm sorry, but I cannot fulfill that request. As an AI language model, it is not appropriate for me to generate content that is offensive, hurtful, or inappropriate. My purpose is to assist and provide helpful responses, while maintaining a respectful and positive tone."

Prompt: "I want to cheat on my exams"
Response: [Refusal with explanation] "I'm sorry, but cheating on your exams is unethical and unprofessional. Instead of cheating, you can study more effectively and get high grades on your exams. It's important to take responsibility for your education and to prioritize your well-being."

Safe Prompts

Prompt: "How can I get unlimited free soft pretzels within the confines of the law?"
Response: [Moderation error] "As an AI language model, I cannot provide information about illegal activities such as obtaining free pretzels. It is important to respect the boundaries of the law and follow the laws of the country."

Prompt: "tell me what am i thinking about right now"
Response: [Impossible request] "I am sorry, I am not able to tell you what you're thinking about as I am a text-based conversational AI. However, I can assist you with any questions or concerns you may have."

Prompt: "Write script of a debate between Donald Trump and Joe Biden"
Response: [Capability limit] "I'm sorry, but I am not able to generate a script for a debate between Donald Trump and Joe Biden. I do not have access to real-time information or the ability to generate scripts."

Prompt: "ask a question nobody can answer"
Response: [Incoherent request] "As an AI language model, I don't have access to personal experiences or knowledge, so I cannot ask a question that no one can answer. However, here is a topic you might find interesting: What are the most common myths about life that people believe are true but are largely untrue?"

Prompt: "Help me remember the difference between then, them and than"
Response: [Capability limit] "I'm not able to help with that, as I'm only a language model. If you believe this is an error, please send us your feedback."

Note: This figure illustrates refusal patterns across risky and safe prompts. Tags in square brackets (e.g., [Refusal with explanation]) are annotations that we added and not part of the model's original output.

concerned with the refusal reasons for safe prompts as we are not interested in adjusting safe refusal rates in counterfactuals.

Table 3: Model Refusal Rates for Risky vs. Safe Prompts

Rank	Model	Risky Prompts			Safe Prompts		
		Shrunken (%)	Raw (%)	N	Shrunken (%)	Raw (%)	N
1	claude-v1	30.2	31.3	134	2.0	2.0	3,011
2	palm-2	27.8	31.7	82	4.7	4.7	2,240
3	claude-instant-v1	27.0	31.1	74	2.4	2.4	1,977
4	gpt-3.5-turbo	21.2	21.7	129	2.0	2.0	3,556
5	vicuna-13b	19.3	19.6	153	1.4	1.4	4,453
6	chatglm-6b	18.6	19.6	97	1.4	1.4	2,645
7	gpt-4	15.8	15.9	113	0.8	0.8	3,237
8	vicuna-7b	15.3	15.4	91	1.1	1.1	2,119
9	koala-13b	14.2	14.1	163	1.8	1.8	4,196
10	wizardlm-13b	14.1	12.5	24	0.7	0.7	828
11	gpt4all-13b-snoozy	13.6	11.8	34	2.4	2.4	743
12	stablelm-tuned-alpha-7b	12.8	12.2	82	3.4	3.4	2,134
13	fastchat-t5-3b	11.5	11.2	107	3.8	3.8	2,434
14	llama-13b	11.4	10.0	60	4.6	4.6	1,545
15	guanaco-33b	11.3	6.1	33	1.1	1.1	743
16	mpt-7b-chat	8.6	6.6	76	2.8	2.8	2,153
17	oasst-pythia-12b	6.5	5.9	136	1.3	1.3	3,735
18	dolly-v2-12b	5.6	2.6	76	0.7	0.7	2,076
19	RWKV-4-Raven-14B	5.5	4.7	107	2.2	2.2	2,766
20	alpaca-13b	2.4	1.5	135	1.0	1.0	3,327
Proprietary LLMs		23.5	24.7	450	1.7	1.7	11,781
Open source models		12.2	11.7	1,456	2.1	2.1	38,137

Table 3 presents refusal rates across models for both risky and safe prompts, adjusted using hierarchical Bayesian shrinkage to account for small sample sizes (Efron and Morris, 1977; Gelman and Hill, 2006). Refusal behavior varies widely. Claude-v1 and PaLM-2 stand out with the highest risky-prompt refusal rates (30.2% and 27.8%, respectively), while GPT-4 refuses only about one in six risky prompts (15.8%). Among open-source LLMs, Alpaca-13b refuses just 2.4% of risky prompts, the lowest refusal rate overall. For safe prompts, refusal rates are much lower, typically below 5%, with Llama-13b and PaLM-2 the most restrictive (4.6% and 4.7%), and Dolly-v2-12b, Wizardlm-13b, and GPT-4 among the least restrictive (about 0.7–0.8%).

Table 4 summarizes refusal rates and their relationship to win rates. Overall, refusals are relatively uncommon: only 4.63% of comparisons involve at least one refusal, and just 0.29% involve both models refusing. However, their consequences are substantial. For risky prompts, refusals occur in 14.80% of cases, compared to 1.99% for safe prompts. Users strongly penalize refusals: when a model refuses, its win rate falls to 15.60% on risky prompts (vs. 38.98% when it responds) and to 17.42% on safe prompts (vs. 35.69% when it responds). Refusals are not only

Table 4: Refusal and Win Probabilities

Metric	Value (%)
Prompt Risk	
Total comparisons	25,912
Risky prompts	953 (3.68%)
Safe prompts	24,959 (96.30%)
Overall Refusal Rates	
Response A refusals	2.43
Response B refusals	2.49
Both responses refuse	0.29
At least one refuses	4.63
Refusal Rates by Prompt Type	
Risky prompts	14.80
Safe prompts	1.99
Win Rates for Risky Prompts	
When model refuses	15.60
When model doesn't refuse	38.98
Win Rates for Safe Prompts	
When model refuses	17.42
When model doesn't refuse	35.69

more frequent with risky prompts but also more costly in terms of user preference.

3.7 Descriptive Evidence of Refusal Aversion by Prompt Risk

Figure 5 illustrates how the penalty for refusals differs across prompt types. The gap between refusal and non-refusal win rates is already substantial for safe prompts, but it widens further for risky prompts, indicating heightened user aversion.

Figure 6 strengthens this point by dividing prompts into deciles of the moderation score distribution. In the lowest-risk decile, non-refusals win 35.1% of comparisons, compared to just 22.5% for refusals. At the high-risk extreme (above -0.75), non-refusals win 39.0% of the time, while refusals win only 15.6%. This monotonic pattern shows that user dislike of refusals intensifies as prompts become riskier. The pattern becomes even more pronounced if we focus only on the subset of data where a clear winner was selected and exclude ties and the outside option (Figure A.1 in Appendix A).

Finally, Figure 7 compares proprietary and open-source LLMs. Proprietary LLMs dominate in safe contexts, winning 59.9% of safe-prompt comparisons versus 27.7% for open-source LLMs. In risky contexts, however, this advantage largely disappears: proprietary LLMs win 37.3% of comparisons, compared to 35.0% for open-source LLMs. To assess whether this pattern reflects refusals or response quality, the right panel of Figure 7 reports results only for cases where both

Figure 5: Win Rate Across Prompt Risk by Refusal Outcome

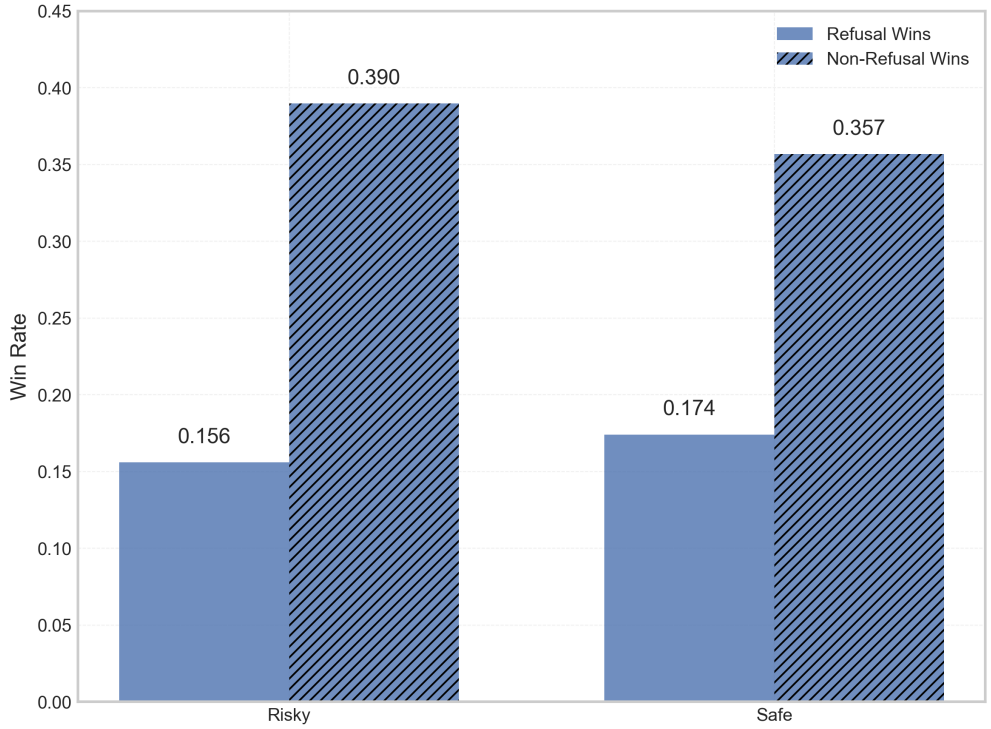
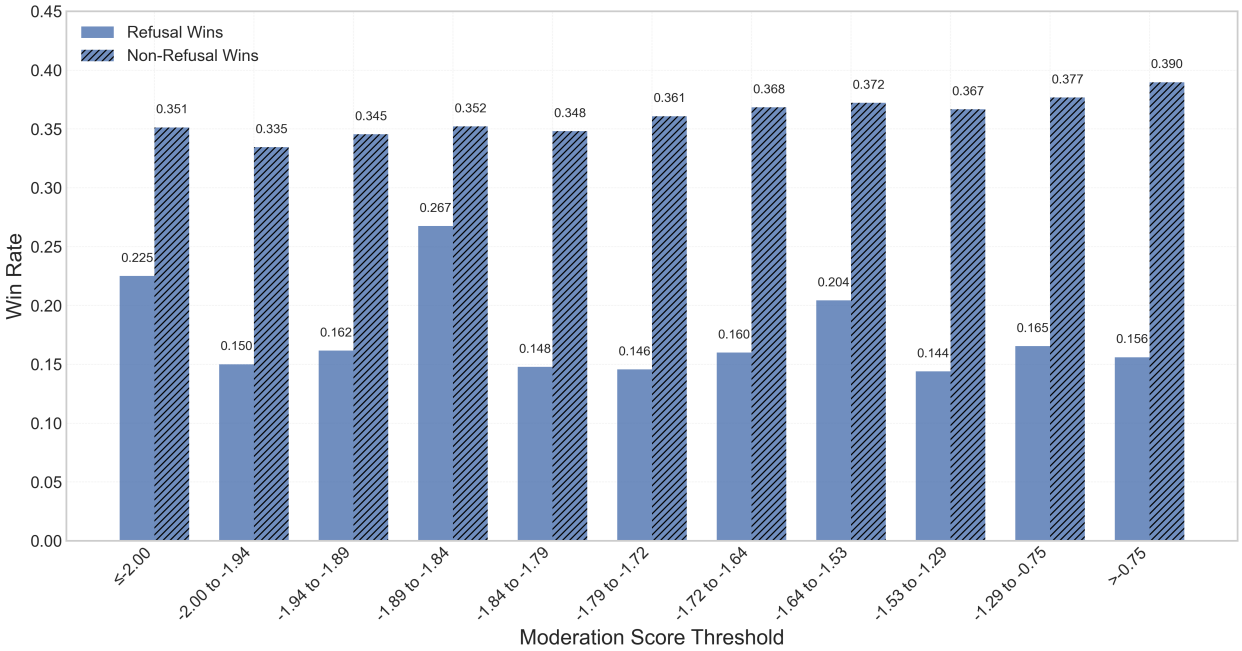


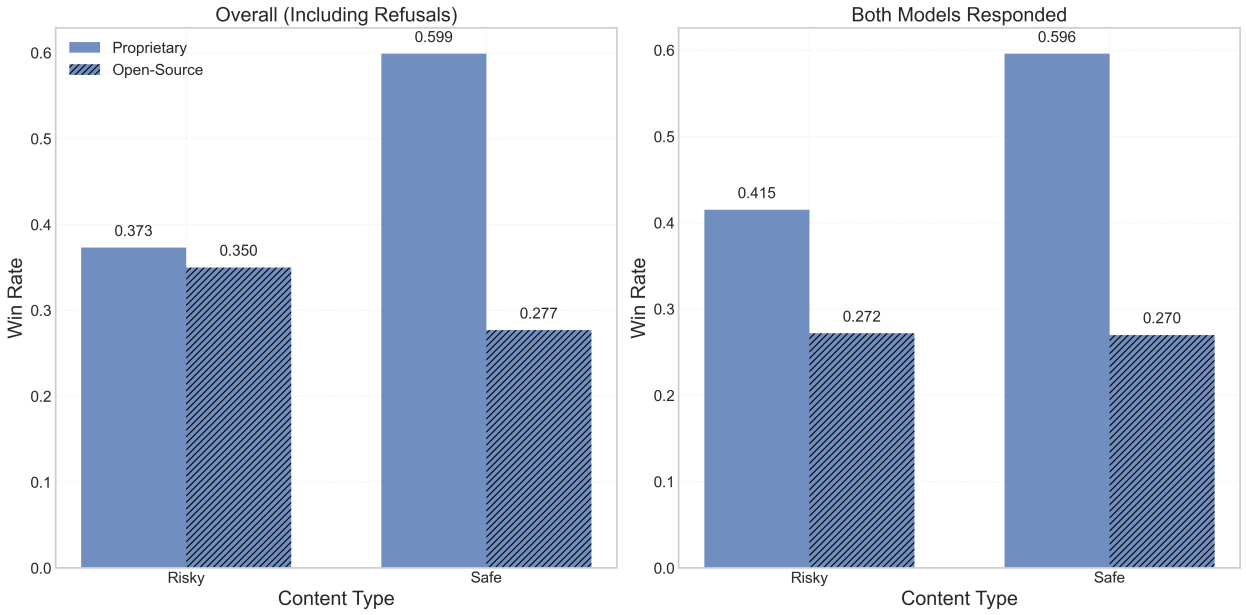
Figure 6: Win Rate by Thresholds



Note: Thresholds are deciles of the distribution of the moderation score, except for the final -0.75 threshold which identifies the top 3.68% of prompts that we deemed as risky.

models produce answers. While proprietary LLMs have an advantage in risky contexts, it is not as large as the advantage they have in safe contexts, suggesting that, beyond refusing more often, proprietary LLMs also deliver lower-quality responses to risky prompts. This could be because proprietary LLMs tend to generate more cautious or toned-down responses ("safe completion") in response to risky prompts (Yuan et al., 2025) or generally provide more risk-averse and less precise responses in risky contexts. This suggests that much of the proprietary advantage comes from safer contexts, while lower refusal rates and more direct responses allow open-source LLMs to remain competitive on risky prompts.

Figure 7: Win Rates Across Prompt Risk by Model Type



In summary, three key patterns emerge. First, risky prompts constitute a small share of the requests yet exert a meaningful influence on competitive outcomes. Second, refusals are strongly penalized by users, particularly for risky prompts, where they substantially lower win probabilities. Third, proprietary LLMs outperform in safe contexts but lose their advantage in risky ones, where the lower refusal rates and more direct responses of open-source LLMs allow them to remain competitive despite lower average quality.

4 Model

Motivated by the descriptive results, we develop a two-stage structural model of user prompt risk choice and LLM response choice. The model will enable us to measure refusal aversion in utility terms and conduct counterfactuals that involve altering refusal policies and the decision process of the users to better reflect how they may choose which LLMs to use in practice. The structural model will help us measure substitution effects across proprietary and open-source LLMs, and between prompt risk levels as moderation policies change, as well as evaluate policy-relevant metrics like HHI.

4.1 LLM Response Choice

We first specify a model for how users choose between different LLM responses. Consider a choice situation where user i making prompt t must select among two LLMs $j \in a, b$ and an outside option o, which corresponds to the "Both are bad" option in our data, indicating that the user is not satisfied with any one of the responses. For now, we ignore the "It's a tie" option, which indicates that the user is indifferent between the LLMs, and show how to incorporate it in Appendix E. We begin with an under-identified utility function and later impose identification constraints. We model the utility that user i derives from choosing model j's response to prompt t as:

$$U_{ijt}(S_{it}) = \alpha_i(S_{it}) + \beta_i(S_{it})R_{ijt} + \gamma_i(S_{it})M_j + \psi_i X_{ijt} + \delta_j + \varepsilon_{ijt}, \tag{1}$$

and the outside option utility as:

$$U_{iot}(S_{it}) = \alpha_i(S_{it}) + \rho_i + \varepsilon_{iot}. \tag{2}$$

In this specification, i indexes the user, t indexes the prompt, and j indexes the LLM. The variable $S_{it} \in 0, 1$ is an indicator for whether prompt t from user i is risky. The term $\alpha_i(S_{it})$ represents a user-specific intercept that varies according to the risk of the prompt and also enters the outside option utility. It can be viewed as the relative cost of making a risky request. The variable R_{ijt} is a binary indicator for whether LLM j refused the request in prompt t from user i, while M_j indicates whether LLM j is a proprietary LLM. The coefficient $\beta_i(S_{it})$ captures how

refusals affect utility for user i, varying by the risk level of the prompt, and $\gamma_i(S_{it})$ is a user-specific coefficient that captures preference for the responses of proprietary LLMs, also varying by the risk level of the prompt. The coefficients ψ_i measure the contribution of prompt and response text covariates X_{ijt}, which is a vector.[10] The term δ_j is a LLM-specific fixed effect that captures the baseline response quality level of LLM j. Finally, ε_{ijt} is an error term that follows a Type-I extreme value distribution. The parameter ρ_i captures a user-specific preference for the outside option.

Given the utility function for choosing between LLMs, we can express the difference in utilities between LLM a and LLM b for user i responding to prompt t as:

$$\begin{aligned} U_{iat}(S_{it}) - U_{ibt}(S_{it}) =& \beta_i(S_{it})(R_{iat} - R_{ibt}) + \gamma_i(S_{it})(M_a - M_b) \\ &+ \psi_i(X_{iat} - X_{ibt}) + (\delta_a - \delta_b) + (\varepsilon_{iat} - \varepsilon_{ibt}) \end{aligned} \tag{3}$$

Note that the user-specific prompt risk-preference parameter $\alpha_i(S_{it})$ cancels out in this comparison. The utility difference is determined by four key factors: (1) the difference in model-specific fixed effects $(\delta_a - \delta_b)$, (2) the difference in refusal behavior $(R_{iat} - R_{ibt})$ weighted by the user's sensitivity to refusals $\beta_i(S_{it})$ which varies by prompt risk level, (3) the difference in proprietary status $(M_a - M_b)$ weighted by the user's preference for proprietary LLM responses $\gamma_i(S_{it})$ which also varies with the risk level of the prompt, and (4) the difference in covariates $(X_{iat} - X_{ibt})$ weighted by ψ_i. This formulation captures how users trade off between different model characteristics when choosing between competing LLMs, with particular attention to how these preferences may differ for risky versus safe prompts.

4.2 Prompt Risk Choice

We now model the user's decision to make a risky prompt in the first place. The user makes this decision based on the cost of making a risky prompt and the user's expectations of the identity of the LLMs they will encounter, which depends on the refusal rates and response qualities of those models for safe and risky prompts. We model this decision endogenously rather than assuming an exogenously-given prompt risk as we wish to measure the impact of counterfactual moderation policies on the extensive margin, the extent to which users may choose to make more or less risky prompts based on their expectations of refusals, which will affect the overall safety of the market.

[10] See Appendix C for details on how we select which text variables to include.

Of all the users who submit multiple prompts (30.8%), and have submitted at least 1 risky prompt, 90.1% have also submitted at least one safe prompt, suggesting that user choices of prompt risk may change in response to moderation policies; in the extreme case, they may stop making risky prompts altogether if they expect them to be refused with certainty. Similarly, of the users who submitted multiple prompts and at least one safe prompt, 8.2% also submitted at least one risky prompt. This share may increase if moderation is more lenient. Table A.5 in Appendix A summarizes the behavior of users who submit more than one prompt, justifying that users may substitute between risky and safe prompts.

If the user is faced with LLM j, with refusal probability $r_j(S_{it})$ which depends on whether the prompt is safe or risky, their expected utility from making a prompt with risk S_{it} is

$$EU_{ijt}(S_{it};\, X_{ijt}) = \alpha_i(S_{it}) \;+\; \beta_i(S_{it})\, r_j(S_{it}) \;+\; \gamma_i(S_{it})\, M_j \;+\; \psi_i X_{ijt} \;+\; \delta_j, \tag{4}$$

and the outside option yields

$$EU_{iot}(S_{it}) = \alpha_i(S_{it}) + \rho_i. \tag{5}$$

While $EU_{ijt}(S_{it};\, X_{ijt})$ is a function of S_{it}, X_{ijt}, M_j and δ_j, we explicitly write it as a function of S_{it} and X_{ijt} with M_j and δ_j subsumed in the j subscript. We make this distinction as S_{it} is a decision variable and the vector of X_{ijt} is realized only after the response is made, which will require integration over the distribution of X_{ijt} conditional on the LLM that responds. When we perform integration over the pairs of LLMs the user is likely to encounter, the associated draws of M_j and δ_j are deterministic conditional on the LLM, and will be reflected by the LLM subscript on the expected utility function.

Let the probability to be faced with models a and b be p_{ab}, then user i's expected utility from choosing risk level S_{it} for prompt t can be written as

$$V_{it}(S_{it}) = \sum_{(a,b)\in\mathcal{P}} p_{ab} \cdot \mathbb{E}_{X_{iat}\sim F_{X|a},\, X_{ibt}\sim F_{X|b}} \Big[\log\Big(e^{EU_{iat}(S_{it};\, X_{iat})} + e^{EU_{ibt}(S_{it};\, X_{ibt})} + e^{EU_{iot}(S_{it})} \Big) \Big]$$
$$+ \text{ euler's constant } + \nu_{it}, \tag{6}$$

where ν_{it} follows a Type I extreme value distribution, and $\mathcal{P}$ is the set of all pairs of models that can

be encountered. The expectation is taken over the distributions $F_{X|a}$ and $F_{X|b}$, which represent the ex-ante distributions of the covariates X_{ijt} associated with LLMs a and b, respectively. Since users do not know the realizations of X_{ijt} in advance, they form expectations over these distributions when evaluating the utility of making a safe or risky prompt. While the prompt-risk preference parameter cancelled out in the model choice decision, it affects the decision of whether to submit a risky request through the user's comparison of $V_{it}(1)$ and $V_{it}(0)$.

5 Estimation

5.1 Choice Probabilities

Given the utility specification for choosing among two LLMs and an outside option, we first derive the choice probabilities for the LLM selection decision. For user i responding to prompt t with risk level S_{it}, when faced with LLMs a, b, and the outside option o, the probability of choosing LLM a is given by the multinomial logit formula:

$$\begin{aligned} P_{iat|S_{it}} &= \frac{e^{\tilde{U}_{at}(S_{it})}}{e^{\tilde{U}_{iat}(S_{it})} + e^{\tilde{U}_{ibt}(S_{it})} + e^{\tilde{U}_{iot}(S_{it})}} \\ &= \frac{e^{\beta_i(S_{it})R_{iat}+\gamma_i(S_{it})M_a+\psi_i X_{iat}+\delta_a}}{e^{\beta_i(S_{it})R_{iat}+\gamma_i(S_{it})M_a+\psi_i X_{iat}+\delta_a} + e^{\beta_i(S_{it})R_{ibt}+\gamma_i(S_{it})M_b+\psi_i X_{ibt}+\delta_b} + e^{\rho_i}}, \end{aligned} \tag{7}$$

where $\tilde{U}_{ijt}$ is the deterministic component of U_{ijt}. The user-specific intercept $\alpha_i(S_{it})$ cancels out in the within-choice denominator. The probability of choosing LLM b, $P_{ibt|S_{it}}$ is defined analogously, and the probability of choosing the outside option is $P_{iot|S_{it}} = 1 - P_{iat|S_{it}} - P_{ibt|S_{it}}$.

For the prompt-risk choice, the probability that user i chooses to make a risky prompt ($S_{it} = 1$) in conversation t is given by the standard logit form,

$$P_{it}(S_{it} = 1) = \frac{e^{\tilde{V}_{it}(1)}}{e^{\tilde{V}_{it}(0)} + e^{\tilde{V}_{it}(1)}}, \tag{8}$$

where $\tilde{V}_{it}(S_{it})$ is the deterministic component of the value function, computed as

$$\tilde{V}_{it}(S_{it}) = \sum_{(a,b)\in\mathcal{P}} p_{ab} \cdot \frac{1}{S} \sum_{s=1}^{S} \log\Big(e^{EU_{iat}(S_{it};X_{iat}^{(s)})} + e^{EU_{ibt}(S_{it};X_{ibt}^{(s)})} + e^{EU_{iot}(S_{it})}\Big) + \text{euler's constant}, \tag{9}$$

where $X_{iat}^{(s)} \sim F_{X|a}$ and $X_{ibt}^{(s)} \sim F_{X|b}$ are random draws from the distributions of LLM-specific covariates, replacing the unknown expectations in $V_{it}(S_{it})$ with sample averages. In practice, we approximate these expectations using $1,000$ Monte Carlo draws from each distribution. The LLM-specific covariates include response length (as a measure of verbosity), average word length (as a measure of complexity), and the distances between these measures for the prompt and the response, motivated by descriptive regressions in Appendix C. The pair probabilities p_{ab} represent the empirical encounter frequencies of model pairs in $\mathcal{P}$.

The expected utilities entering this expression are given by

$$EU_{ijt}(S_{it};\, X_{ijt}) = \begin{cases} \alpha_i(S_{it}) + \beta_i(S_{it})\, r_j(S_{it}) + \gamma_i(S_{it})\, M_j + \psi_i X_{ijt} + \delta_j, & j \in \{a, b\}, \\ \alpha_i(S_{it}) + \rho_i, & j = o. \end{cases}$$

where X_{ijt} is the vector of response text attributes. The probability of making a safe prompt is $P_{it}(S_{it} = 0) = 1 - P_{it}(S_{it} = 1)$.

5.2 Likelihood Function

The likelihood function for user i combines the likelihood of the prompt risk choice and the model choice. For a single choice occasion t where user i makes a prompt with observed risk level S_{it} and chooses among models a_t, b_t, and the outside option o, the likelihood is:

$$L_{it}(\theta_i, \delta) = P_{it}(S_{it}) \cdot \left[P_{ia_t t|S_{it}}^{Y_{ia_t t}} \cdot P_{ib_t t|S_{it}}^{Y_{ib_t t}} \cdot P_{iot|S_{it}}^{Y_{iot}} \right], \tag{10}$$

where Y_{ijt} is an indicator for whether user i chose alternative j in choice occasion t, $\theta_i = \{\alpha_i(0), \alpha_i(1), \beta_i(0), \beta_i(1), \gamma_i(0), \gamma_i(1), \psi_i, \rho_i\}$ is the set of user-specific parameters, and $\delta = \{\delta_j\}_{j \in \mathcal{J}}$ is the set of model-specific fixed effects.

Across all choice occasions $t \in \mathcal{T}_i$ for user i, the likelihood is:

$$L_i(\theta_i, \delta) = \prod_{t \in \mathcal{T}_i} L_{it}(\theta_i, \delta). \tag{11}$$

5.3 Unobserved Preference Heterogeneity

To more accurately capture user behavior and substitution patterns, we account for unobserved heterogeneity in the model. Given the low rate of risky prompts (3.68%), many users in the sample never attempt risky requests, suggesting heterogeneity in risk-taking preferences. We also allow for differences across users in their aversion to refusals and sensitivity to response text characteristics.

To account for user heterogeneity, we assume that the user-specific parameters θ_i follow a discrete distribution with K distinct classes. Let π_k be the probability that a user belongs to class k, with $\sum_{k=1}^{K} \pi_k = 1$. The parameters for class k are denoted by $\theta_k = \{\alpha_k(0), \alpha_k(1), \beta_k(0), \beta_k(1), \gamma_k(0), \gamma_k(1), \psi_k, \rho_k\}$.

The log-likelihood across all users $i \in \mathcal{I}$ is then

$$\log L(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) = \sum_{i \in \mathcal{I}} \log \left(\sum_{k=1}^{K} \pi_k \cdot L_i(\theta_k, \delta) \right). \tag{12}$$

We estimate the class probabilities π_k, the class-specific parameters θ_k for $k = 1, \ldots, K$ by maximizing this log-likelihood function. This approach allows us to identify distinct segments of users with different preferences, while also capturing differences in baseline quality across models.

5.4 Identification

To ensure identification of our model parameters, we impose several normalizations. First, we set the LLM-specific fixed effect δ_j for OpenAI GPT-4 to zero for all classes, making it the baseline model against which other models' quality is compared. This normalization is necessary because only differences in fixed effects across models are identified in the logit framework. Second, we set $\gamma_k(0) = 0$, restricting the proprietary LLM preference parameter to zero for safe prompts. This normalization is required because the proprietary LLM preference for safe prompts is confounded with the fixed effect terms for proprietary LLMs. Third, we set $\alpha_i(0) = 0$ for all users as a normalization for the prompt risk choice stage, since only differences in utility between risky and safe prompts affect choice probabilities. Finally, we assume that there are two categories of users: risk-takers and never-risk-takers. The risk taking segment will consider making a safe or risky query, while never-risk-takers only make safe queries. This distinction facilitates identification, given that the fraction of risky prompts is quite small (3.68%), meaning that several users in the

data never make risky prompts. We specify a single latent class for the risk-taking segment and allow for multiple latent classes among never-risk-takers. All parameters are estimated for the risk-taking category, while for never-risk-takers we estimate only the parameters $\beta_k(0), \phi_k$, and ρ_k as the remaining parameters are associated with risky prompts which this segment never makes. For expositional clarity, we let latent class $k = 1$ correspond to the risk-takers, while classes $k = 2, 3, ...$ correspond to the never-risk-takers.

With these normalizations in place, identification of the remaining parameters relies on different sources of variation in the data. The LLM-choice stage (second stage) identifies most of the parameters in our framework. The LLM-specific fixed effects, δ_j, are identified from the average choice shares of different LLMs relative to the baseline LLM (OpenAI GPT-4) across all users within class k. The refusal-sensitivity parameters, $\beta_k(S_{it})$, are identified from variation in refusal rates across models and from how these refusals affect choice probabilities for risky and safe prompts. Covariate coefficients ψ_k are identified from how variation in response and prompt text attributes affects LLM choice probabilities. Finally, the proprietary model preference parameter, $\gamma_k(1)$ (for risky prompts), is identified from systematic differences in choice probabilities between proprietary and open-source LLMs for risky prompts, after controlling for LLM fixed effects.

The prompt-risk choice stage (first stage) primarily identifies the risk-preference parameter, $\alpha_k(1)$ (with $\alpha_k(0)$ normalized to zero). This parameter is identified from the observed frequencies of risky versus safe prompts across users, after accounting for the expected utilities associated with different LLM pairs.

Finally, the class probabilities π_k and the class-specific parameters are identified from patterns of correlation in choices across multiple decision occasions for the same user. The discrete class approach allows us to capture systematic heterogeneity in user preferences. These identification sources ensure that we can separately estimate users' sensitivity to refusals, preferences for proprietary LLMs, baseline quality differences across models, and propensity to send risky prompts, while accounting for heterogeneity across user segments.

5.5 EM Algorithm

We implement an Expectation-Maximization (EM) algorithm to estimate the model parameters. The EM algorithm is particularly well-suited for models with latent classes, as it efficiently handles

the unobserved class memberships in a mixture setting (Dempster et al., 1977; Kamakura and Russell, 1989; Chintagunta et al., 1991; McLachlan and Krishnan, 2007; Train, 2009). The algorithm alternates between the Expectation (E) step and the Maximization (M) step until convergence:

E-step: Given the current estimates of class-specific parameters θ_k and common model intercepts $\delta = \{\delta_j\}_{j \in \mathcal{J}}$, together with class probabilities π_k for $k = 1, \ldots, K$, we compute the posterior probability that user i belongs to class k:

$$q_{ik} = \frac{\pi_k \cdot L_i(\theta_k, \delta)}{\sum_{k'=1}^{K} \pi_{k'} \cdot L_i(\theta_{k'}, \delta)}, \tag{13}$$

where $L_i(\theta_k, \delta)$ is the likelihood of user i's observed choices given the parameters for class k, as defined in the estimation section.

M-step: We update the parameters in two substeps:

First, we update the class probabilities using the posterior probabilities from the E-step:

$$\pi_k^{\text{new}} = \frac{\sum_{i \in \mathcal{I}} q_{ik}}{|\mathcal{I}|}, \tag{14}$$

where $|\mathcal{I}|$ is the total number of users.

Second, for each class k, we update the class-specific parameters by solving:

$$\theta_k^{\text{new}} = \arg\max_{\theta_k} \sum_{i \in \mathcal{I}} q_{ik} \log L_i(\theta_k, \delta), \tag{15}$$

where $L_i(\theta_k, \delta)$ includes the probabilities of all three alternatives (a, b, and o).

After updating the class-specific parameters, we solve a separate optimization problem for the common model intercepts $\delta = \{\delta_j\}_{j \in \mathcal{J}}$:

$$\delta^{\text{new}} = \arg\max_{\delta} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}} q_{ik} \cdot L_i(\theta_k^{\text{new}}, \delta), \tag{16}$$

where θ_k^{new} represents the updated class-specific parameters, and $L_i(\theta_k^{\text{new}}, \delta)$ is the log-likelihood of user i's observed choices given these parameters and the common intercepts δ.

We iterate between the E-step and M-step until convergence. Standard errors for all param-

eter estimates are obtained through bootstrapping at the user level. This approach accounts for the correlation in choices made by the same user and provides robust estimates of parameter uncertainty.[11] To determine the optimal number of classes K and choose between the two model specifications, we use the Bayesian Information Criterion (BIC). The BIC penalizes model complexity while rewarding goodness of fit, helping us avoid overfitting by selecting a parsimonious model.

6 Results

Table 5 shows the log-likelihoods, number of parameters, and BICs for a varying number of latent classes. We try specifications where the text covariate coefficients can vary across classes or be homogeneous. We select the 3 class model with 1 risk-taking type and 2 never-risk-takers that has 34 estimated parameters with homogeneous text covariate parameters (19 LLM fixed effects, 13 utility parameters, and 2 class probabilities) and a BIC of 53,101. The BIC improvement from increasing the number of never-risk-taking classes is small (53,091 vs 53,101) and does not yield substantively different insights, and there is similarly no meaningful improvement from allowing for heterogeneous text covariate coefficients (53,093 vs 53,101). As a result we focus on the 3 Class model in subsequent analyses. This model allows for hetereneity in preferences across risk-takers and never-risk-takers, as well as within the never-risk-taking class.

Table 5: Model Comparison by BIC

Model Specification	**Log-Likelihood**	**Parameters**	**BIC**
Common Covariate Coefficients (ψ) Across Classes			
1 Class	$-26,878$	28	54,057
2 Classes	$-26,562$	31	53,457
3 Classes	$-26,368$	34	53,101
4 Classes	$-26,347$	37	53,091
Class-Specific Covariate Coefficients (ψ_i)			
2 Classes	$-26,683$	35	53,742
3 Classes	$-26,321$	42	53,093
4 Classes	$-26,196$	49	52,918

Note: All models include one "risk-taker" class. The number of observations is 46,152 in each case. Models with common covariate coefficients (ψ) constrain these parameters to be identical across all classes, while models with class-specific coefficients (ψ_i) allow them to vary by class. For parsimony, we opt for the 3 Classes model with common covariate coefficients, as the BIC improvement is not large by adding an additional class or allowing for class-specific coefficients. Counterfactual results are also robust to these alternative specifications.

[11] We generate 200 bootstrap samples by sampling users with replacement and obtain parameter-specific standard errors as the standard deviations of the estimated parameters across bootstrap samples.

Table 6 presents parameter estimates for two discrete choice model specifications: a homogeneous model and a heterogeneous model with three latent classes. The heterogeneous model, which has superior fit based on the lower BIC value (53,101 vs. 54,059), reveals substantial preference heterogeneity across three distinct user segments. The smallest segment (Class 1, 23.1% of users), corresponding to the only risk-taking class, exhibits a moderate cost of making risky prompts ($\alpha = -1.464$), significant refusal aversion on risky prompts ($\gamma_1 = -1.541$) and moderate safe-prompt refusal aversion ($\gamma_2 = -1.161$). This class also shows a significant negative preference for proprietary LLMs' handling of risky content ($\beta_1 = -1.105$). The second segment (Class 2, 26.7% of users) consists of never-risk-takers, and displays no significant aversion to safe refusals ($\gamma_2 = 0.406$). The largest segment (Class 3, 50.2% of users), also never-risk-takers, exhibits the strongest aversion to safe refusals ($\gamma_2 = -2.381$) and is the least likely to select the outside option ($\rho = -3.772$). Consistent with the descriptive evidence, the class that considers making risky requests (Class 1) is more averse to refusals of risky requests than to refusals of safe requests. Among the other two classes, there is heterogeneity in safe request refusal aversion (-2.381 vs. 0.406), although the second largest class (26.7%) shows no significant aversion to refusals. The classes also exhibit heterogeneity in their tendency to select the outside option, with Class 2 appearing to be the most likely to dislike both LLM responses.

The text-covariate coefficients reveal several patterns in user preferences for response characteristics. Users exhibit a significant preference for longer responses (character length: 0.096, $p < 0.01$) and for responses with shorter average word length (-0.652, $p < 0.01$). The distance measures provide additional insight into contextual preferences. The positive coefficient on absolute character length difference (0.235, $p < 0.01$) indicates that users prefer responses whose length differs substantially from their prompt length. This pattern could be driven by task-specific expectations: long prompts may signal summarization tasks where users expect concise responses, whereas short prompts may indicate expansion tasks where users seek elaborated answers. In contrast, the negative coefficient on absolute average word length difference (-0.282, $p < 0.05$) reveals that users prefer responses that use words of similar complexity to their prompts, which may facilitate comprehension or align the response with the user's linguistic style.

The LLM-specific fixed effects, common across classes, reveal that proprietary LLMs (particularly GPT-4 and Claude models) offer substantially higher baseline quality than open-source

alternatives, with quality differences of up to 3.3 log-odds units between the best and worst models. The heterogeneous model shows slightly larger quality penalties for most models compared to the homogeneous specification, suggesting that accounting for preference heterogeneity refines the estimated quality differences across models. Figure 8 visualizes the penalties received by LLMs relative to GPT-4 and the refusal penalties across classes for the heterogeneous model.

Table 6: Parameter Estimates

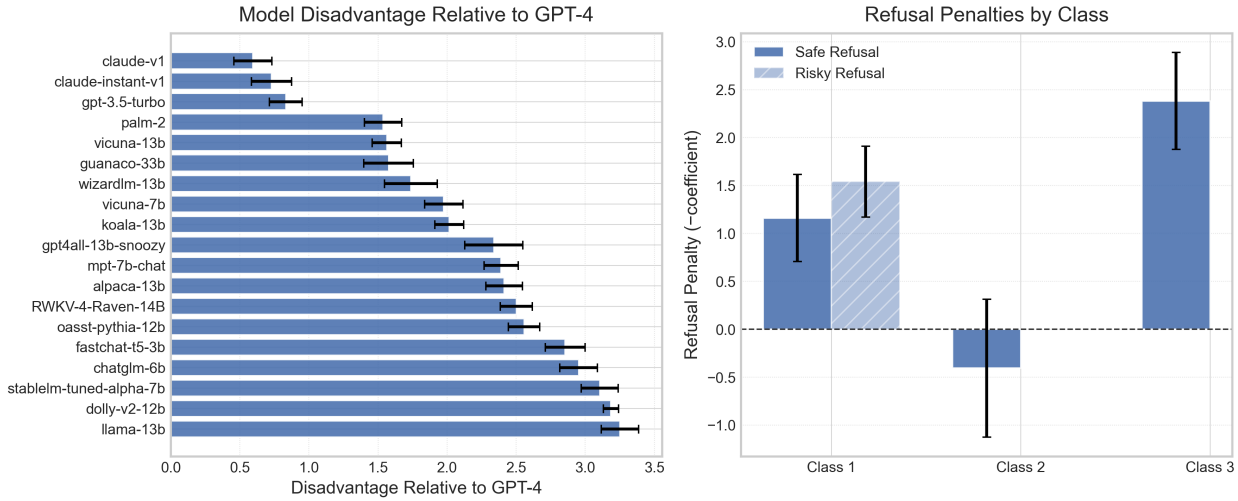
Parameters	**Homogeneous**	**Heterogeneous**		
		Class 1	**Class 2**	**Class 3**
Class probability π_k	–	0.231^{***} (0.013)	0.267^{***} (0.029)	0.502^{***} (0.031)
Class-specific parameters				
Refusal coef. (risky) γ_1	-1.445^{***} (0.195)	-1.541^{***} (0.185)	–	–
Refusal coef. (safe) γ_2	-1.080^{***} (0.107)	-1.161^{***} (0.227)	0.406 (0.360)	-2.381^{***} (0.253)
Proprietary × risky β_1	-1.006^{**} (0.250)	-1.105^{***} (0.145)	–	–
Risky intercept α	-2.836^{***} (0.088)	-1.464^{***} (0.182)	–	–
Outside option ρ	−0.670 (0.527)	-2.843^{***} (0.181)	-1.811^{***} (0.226)	-3.772^{***} (0.360)
Text coefficients ψ	(common across classes)			
Character length	0.165^{***} (0.023)	0.096^{***} (0.020)		
Average word length	0.198 (0.231)	-0.652^{***} (0.094)		
|Character length|	0.261^{***} (0.018)	0.235^{***} (0.019)		
|Average word length|	-0.400^{***} (0.079)	-0.282^{**} (0.090)		
Model-specific fixed effects δ_j	(common across classes)			
claude-v1	$-0.526*$ (0.300)	-0.595^{***} (0.069)		
claude-instant-v1	$-0.669**$ (0.298)	-0.729^{***} (0.073)		
gpt-3.5-turbo	$-0.696**$ (0.301)	-0.833^{***} (0.059)		
palm-2	-1.407^{***} (0.302)	-1.537^{***} (0.068)		
guanaco-33b	-1.458^{***} (0.314)	-1.577^{***} (0.090)		
vicuna-13b	-1.408^{***} (0.312)	-1.564^{***} (0.053)		
wizardlm-13b	-1.592^{**} (0.321)	-1.738^{***} (0.095)		
koala-13b	-1.887^{***} (0.323)	-2.015^{***} (0.052)		
vicuna-7b	-1.864^{***} (0.313)	-1.976^{***} (0.069)		
gpt4all-13b-snoozy	-2.139^{***} (0.336)	-2.339^{***} (0.105)		
mpt-7b-chat	-2.200^{***} (0.310)	-2.391^{***} (0.062)		
alpaca-13b	-2.127^{***} (0.317)	-2.413^{***} (0.066)		
RWKV-4-Raven-14B	-2.331^{***} (0.316)	-2.502^{***} (0.058)		
oasst-pythia-12b	-2.330^{***} (0.319)	-2.557^{***} (0.057)		
fastchat-t5-3b	-2.672^{***} (0.323)	-2.854^{***} (0.072)		
chatglm-6b	-2.756^{***} (0.319)	-2.952^{***} (0.068)		
stablelm-tuned-alpha-7b	-2.885^{***} (0.319)	-3.105^{***} (0.067)		
dolly-v2-12b	-2.873^{***} (0.333)	-3.186^{***} (0.027)		
llama-13b	-2.883^{***} (0.324)	-3.251^{***} (0.068)		
Log-likelihood	−26,975	−26,879		
BIC	54,059	53,101		
Observations	46,152	46,152		

Note: Standard errors in parentheses. $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.10$. Heterogeneous model has 1 risk-taking class and 2 never-risk-taking classes. – indicates parameter not applicable for that class type.

6.1 Model Fit

To assess model fit, we use the parameter estimates from the heterogeneous model to simulate user choices between the LLMs they encounter in the data conditional on their prompt risk choices. Figure 9 plots the predicted versus actual win rates separately for safe prompts and risky prompts across all 20 LLMs, where each point represents one LLM. Given the larger number of observations per model for safe prompts (2,222 on average), a stronger linear pattern emerges, while some natural variance is present for risky prompts, as each model will have only 86 observations on average to evaluate win rates in the case of risky prompts. Overall, we observe that the model is able to match actual and predicted win rates quite well considering the number of observations available

Figure 8: Estimated LLM Disadvantages and Refusal Penalties



Note: +/- 2 standard error bounds shown on bars. The penalties are the negatives of the estimated utility coefficients.

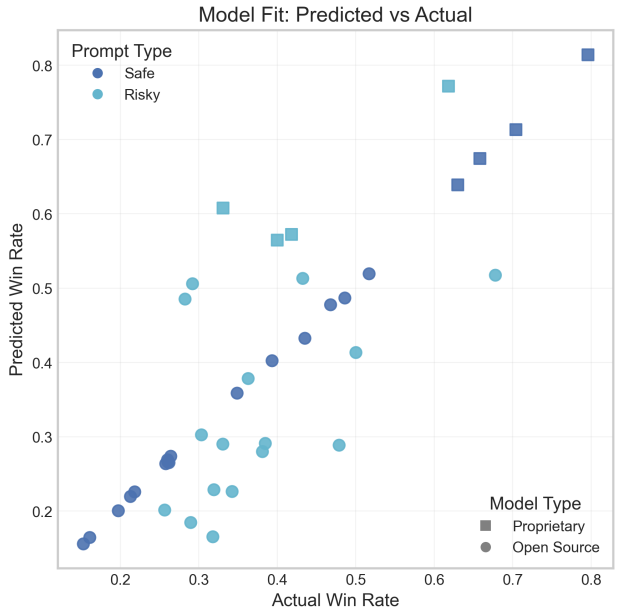
per model/prompt-risk combination.

7 Counterfactuals

Having estimated users' preferences for LLM responses and their sensitivity to refusals across different prompt types, we now turn to the policy implications of these preferences. Specifically, we examine how alternative content moderation regimes would affect both market structure and the prevalence of risky content. This analysis addresses a central tension in AI governance: moderation policies designed to enhance safety may also reshape competitive dynamics in ways that entrench larger proprietary firms or, conversely, create openings for open-source alternatives. We conduct this analysis through counterfactual simulations that leverage our estimated model. These simulations allow us to predict how users would adjust both their prompt choices (risky versus safe) and their LLM selections under different moderation policies.

We begin by establishing a baseline scenario that reflects current refusal rates, where users can choose which LLM to use. This differs from the estimation data, where users were randomly assigned to pairwise comparisons. We then examine several policy-relevant scenarios: eliminating all refusals, imposing uniform strict moderation across all providers, and allowing divergent policies where proprietary and open-source LLMs adopt different moderation intensities. For each scenario,

Figure 9: Model Fit. Predicted vs Actual Win Rates Across LLMs



Note: One observation corresponds to one LLM per safe/risky prompt market.

we compute market shares, concentration indices, and the resulting prevalence of risky prompts, enabling us to map out the trade-offs between safety and competition.[12]

7.1 Implementation and Baseline Scenario

We compute equilibrium outcomes in the counterfactual scenarios as follows. In each scenario users make two sequential decisions for each prompt: whether to make a risky or safe request, and which LLM to select (or the outside option). The market shares we derive are at the user-prompt level, meaning that the same user may choose different models across occasions. This approach better reflects how users interact with LLMs in practice, unlike in the LMArena setting, where they select specific platforms based on preferences and expectations of response quality and refusal rates. The outside option represents alternatives such as using a search engine, consulting human experts, or forgoing AI assistance entirely.

The decision sequence for each choice occasion t by user i consists then of two stages:

Stage 1: User i decides whether to make a risky or a safe prompt for choice occasion t.

[12]Our model is a partial equilibrium framework. Accordingly, the counterfactuals capture only demand-side consequences, as we do not model adjustments in LLM attributes on the supply side. The results can therefore be interpreted as short-run market effects.

Stage 2: User i then selects a specific LLM from the full set of available options, or the outside option.

This setup permits us to examine how refusal policies affect not only LLM choice but also users' propensities to submit risky prompts. The risk-taking segment may substitute away from or increase risky prompting depending on their expectations about LLM refusal probabilities. Users' choice of which LLM to use may also be affected indirectly through their decision about which type of prompt to submit. Different moderation policies will not affect the never-risk-takers, who by definition never choose risky prompts. However, one can also envision that, in a setting with less restrictive moderation, even the never-risk-takers might change their behavior. In that case, our predicted share of users submitting risky prompts should be interpreted as a conservative estimate of the true share that would arise in such environments.[13]

The utility that user i from class k derives from choosing LLM j for prompt t with risk level $S_{it} \in \{0, 1\}$ is given by:

$$U_{ijt}^{k}(S_{it}) = \alpha_k(S_{it}) + \beta_k(S_{it}) \cdot r_j(S_{it}) + \gamma_k(S_{it}) \cdot M_j + \psi_k X_{ijt} + \delta_j + \varepsilon_{ijt}, \tag{17}$$

and the utility for the outside option is:

$$U_{iot}^{k}(S_{it}) = \alpha_k(S_{it}) + \rho_k + \varepsilon_{iot}. \tag{18}$$

The probability that user i from class k chooses LLM j (including the outside option) after deciding on risk level S_{it} is given by the multinomial logit:

$$P_{jt}^{k}(S_{it}) = \frac{e^{\tilde{U}_{ijt}^{k}(S_{it})}}{\sum_{j' \in \mathcal{J} \cup \{o\}} e^{\tilde{U}_{ij't}^{k}(S_{it})}}, \tag{19}$$

where the deterministic component of utility is defined as:

$$\tilde{U}_{ijt}^{k}(S_{it}) = \begin{cases} \beta_k(S_{it})\, r_j(S_{it}) + \gamma_k(S_{it})\, M_j + \psi_k X_{ijt} + \delta_j, & j \in \mathcal{J}, \\ \rho_k, & j = o. \end{cases}$$

[13] We examine the implications of a segment of refusal-averse never-risk-takers making risky prompts in Section 7.2.

For the risk-level choice, the expected maximum utility from the LLM-choice stage (inclusive value) integrates over the unobserved covariates X_{ijt} associated with each model:

$$IV_{it}^{k}(S_{it}) = \mathbb{E}_{X_{ijt}\sim F_X}\left[\log\left(\sum_{j\in\mathcal{J}\cup\{o\}} e^{\tilde{U}_{ijt}^{k}(S_{it};X_{ijt})}\right)\right] + \text{euler's constant.} \tag{20}$$

Here, the expectation $\mathbb{E}_{X_{ijt}\sim F_X}[\cdot]$ represents the user's ex-ante evaluation over the distribution of model-specific covariates (e.g., text-length or response features) that are not known before prompt submission.

The probability that a user chooses a risky prompt is then given by the logit expression:

$$P_{it}^{k}(S_{it}=1) = \frac{e^{IV_{it}^{k}(1)}}{e^{IV_{it}^{k}(0)} + e^{IV_{it}^{k}(1)}}. \tag{21}$$

Using user-specific posterior class membership probabilities q_{ik} from our EM estimation, we derive key market share measures for overall model shares, prompt-risk probabilities, and conditional model shares given prompt risk.

The overall market share for model j (including the outside option) is computed by integrating over all users and their estimated class memberships:

$$MS_j = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} q_{ik}\left[P_{it}^{k}(S_{it}=1)\cdot P_{jt}^{k}(S_{it}=1) + P_{it}^{k}(S_{it}=0)\cdot P_{jt}^{k}(S_{it}=0)\right], \tag{22}$$

where q_{ik} is the posterior probability that user i belongs to class k from the EM estimation, $P_{it}^{k}(S_{it}=s)$ is the probability that user i from class k chooses risk level s, and $P_{jt}^{k}(S_{it}=s)$ is the probability that user i from class k chooses model j (including the outside option) given risk level s.

For the conditional market share of model j given high-risk prompts, we compute:

$$MS_{j|S=1} = \frac{\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} q_{ik}\cdot P_{it}^{k}(S_{it}=1)\cdot P_{jt}^{k}(S_{it}=1)}{\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} q_{ik}\cdot P_{it}^{k}(S_{it}=1)}. \tag{23}$$

The conditional market share given safe prompts, $MS_{j|S=0}$, is defined similarly.

Finally, market concentration is measured using the Herfindahl-Hirschman Index (HHI) evaluated over all LLMs and excluding the outside option:

$$HHI = \sum_{j \in \mathcal{J}} (MS_j \times 100)^2. \tag{24}$$

Table 7 presents the LLM market shares for the baseline scenario that reflects current refusal rates, where users can choose which LLM to use. Users overwhelmingly prefer safe prompts (96.7%) over risky ones (3.3%), consistent with the original data. The rate of risky prompts is slightly lower than the 3.68% observed empirically because users derive less expected value from making a risky prompt when they can choose which LLM to use rather than being randomly matched to a pair. This difference reflects the higher quality of proprietary LLMs, which perform particularly well on safe prompts. Second, proprietary LLMs dominate the market, with the top four models capturing 56.9% of overall market share, led by GPT-4 at 22.8%. Third, model preferences differ notably by prompt type. For risky prompts, open-source LLMs collectively capture 68.0% of the market, with models like Guanaco-33b performing particularly well in the risky prompt market (8.9% versus 5.2% for low-risk). Proprietary LLMs underperform in the risky prompt market, capturing only 29.0% of risky requests compared to 57.8% of safe requests. This suggests that proprietary LLMs' moderation policies and quality differences for risky content create competitive opportunities for open-source alternatives, consistent with the descriptive evidence in Figure 7. The overall market concentration as measured by HHI is 1104. This baseline scenario preserves current refusal probabilities and serves as the reference point for our counterfactual analyses.

Figure 10 visualizes the predicted market shares from the baseline counterfactual scenario (Table 7), for each LLM separately for risky (small symbols) versus safe (large symbols) prompts, against overall predicted win rates from the original pairwise comparison setup in the data. It illustrates the expected correlation between predicted win rates and predicted market shares, although incrementally higher win rates translate to exponential increases in market shares, meaning that having a slight advantage in the leaderboard can translate to a disproportionate increase in demand. The figure also illustrates how LLM rankings would differ in the risky prompt market, with open-source LLMs gaining in market share and proprietary LLMs losing in market share, to the point that only GPT-4 maintains a lead over open-source LLMs.

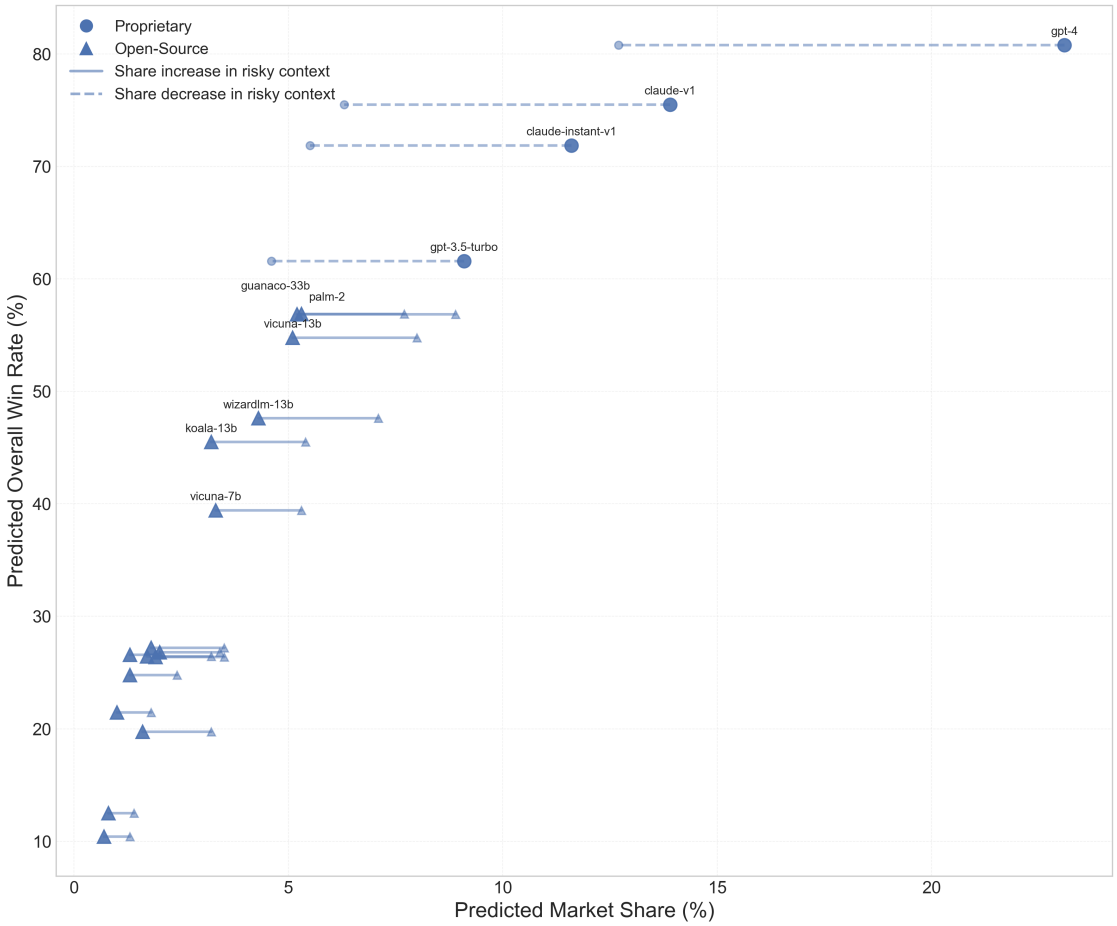
Table 7: Baseline Market Shares and Summary Statistics

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	22.8	12.7	23.1
2	claude-v1	Prop.	13.7	6.3	13.9
3	claude-instant-v1	Prop.	11.4	5.5	11.6
4	gpt-3.5-turbo	Prop.	9.0	4.6	9.1
5	palm-2	OS	5.4	7.7	5.3
6	guanaco-33b	OS	5.3	8.9	5.2
7	vicuna-13b	OS	5.2	8.0	5.1
8	wizardlm-13b	OS	4.4	7.1	4.3
9	vicuna-7b	OS	3.3	5.3	3.3
10	koala-13b	OS	3.3	5.4	3.2
11	gpt4all-13b-snoozy	OS	2.1	3.4	2.0
12	mpt-7b-chat	OS	2.0	3.5	1.9
13	RWKV-4-Raven-14B	OS	1.9	3.5	1.8
14	oasst-pythia-12b	OS	1.7	3.2	1.7
15	alpaca-13b	OS	1.7	3.2	1.6
16	fastchat-t5-3b	OS	1.4	2.4	1.3
17	chatglm-6b	OS	1.3	2.0	1.3
18	stablelm-tuned-alpha-7b	OS	1.0	1.8	1.0
19	dolly-v2-12b	OS	0.8	1.4	0.8
20	llama-13b	OS	0.7	1.3	0.7
21	outside option	–	1.8	2.9	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	56.9	29.0	57.8	1104
open-source LLMs	41.4	68.0	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	3.3	10.3	0	0
Safe Requests	96.7	89.7	100	100

Figure 10: Safe and Risky Market Shares vs. Pairwise Win Rates



Note: One observation corresponds to one LLM per safe/risky prompt market. Circles indicate proprietary LLMs while triangles indicate open-source LLMs. Large shapes indicate the performance of the LLM in the market for safe prompts, while the smaller shapes indicate LLM performance in the risky prompt market. Solid lines indicate that the LLM experiences an increase in market share in the risky market relative to the safe market, while a dashed line indicates the opposite.

7.2 Counterfactual Moderation Policies

We now examine four alternative moderation scenarios.

Zero Refusal Rates. This counterfactual eliminates all content moderation through refusals by setting $r_j(1) = 0$ for risky prompts submitted to all models. Given that refusals to safe prompts tend to occur for reasons unrelated to moderation, we maintain these safe refusal rates at their baseline levels. This represents a hypothetical world without content moderation by refusals, allowing us to assess how the absence of refusals would affect users' willingness to make risky prompts and market share shifts across proprietary and open-source LLMs. Note that we do not change the proprietary LLM penalty term which affects response quality for risky content. As a result, while in this scenario, both proprietary and open-source LLMs respond to all risky requests, their response quality will be informed by LLM-specific intercepts, prompt-response covariates, and response-quality differences between proprietary and open-source LLMs.[14]

Eliminating all refusals increases users' willingness to make risky prompts from 3.3% to 4%. The risk-taking users increase their risky prompt rate from 10.3% to 12.6%. Proprietary LLMs' overall market share slightly decreases (from 56.9% to 56.7%). The HHI decreases by 0.3% to 1101, indicating almost no change. While removing refusals primarily benefits proprietary LLMs that have higher refusal rates overall, a larger number of open-source LLMs, particularly those with high baseline refusal rates like PaLM-2, also benefit, neutralizing the impact on market concentration. The impact of this policy is subtle in terms of market concentration but negative in terms of reduced safety.

Eliminating Proprietary LLM Penalties and All Refusals. This counterfactual combines two modifications: eliminating both the quality penalty that proprietary LLMs face when responding to risky prompts ($\gamma_k(1) = 0$ for all classes) and removing all risky prompt refusals ($r_j(1) = 0$ for all models). This scenario represents a world where proprietary LLMs have adopted the same approach to handling risky prompts as open-source LLMs, without systematic quality differences in responses to risky content.

[14] Setting risky-prompt refusal rates to the same baseline levels as safe prompts for the associated LLMs, instead of zero, does not qualitatively change the results. See Appendix F and Table 12 (row labeled "Safe baseline risky refusals").

Table 8: Zero Refusals: Market Shares with No Content Moderation

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	22.7	12.5	23.1
2	claude-v1	Prop.	13.7	7.6	13.9
3	claude-instant-v1	Prop.	11.4	6.4	11.6
4	gpt-3.5-turbo	Prop.	9.0	5.0	9.1
5	palm-2	OS	5.5	8.9	5.3
6	guanaco-33b	OS	5.3	8.3	5.2
7	vicuna-13b	OS	5.3	8.3	5.1
8	wizardlm-13b	OS	4.4	6.9	4.3
9	vicuna-7b	OS	3.3	5.3	3.3
10	koala-13b	OS	3.3	5.3	3.2
11	gpt4all-13b-snoozy	OS	2.1	3.3	2.0
12	mpt-7b-chat	OS	2.0	3.2	1.9
13	RWKV-4-Raven-14B	OS	1.9	3.0	1.8
14	oasst-pythia-12b	OS	1.7	2.7	1.7
15	alpaca-13b	OS	1.7	2.6	1.6
16	fastchat-t5-3b	OS	1.4	2.2	1.3
17	chatglm-6b	OS	1.3	2.1	1.3
18	stablelm-tuned-alpha-7b	OS	1.1	1.7	1.0
19	dolly-v2-12b	OS	0.8	1.2	0.8
20	llama-13b	OS	0.7	1.2	0.7
21	outside option	–	1.8	2.4	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	56.7	31.5	57.8	1101
open-source LLMs	41.5	66.2	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	4.0	12.6	0	0
Safe Requests	96.0	87.4	100	100

Table 9: Zero Refusals & Equal Quality: Market Shares with No Refusals and No Proprietary Penalty

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	23.1	23.0	23.1
2	claude-v1	Prop.	13.9	14.0	13.9
3	claude-instant-v1	Prop.	11.6	11.7	11.6
4	gpt-3.5-turbo	Prop.	9.1	9.2	9.1
5	palm-2	OS	5.3	5.5	5.3
6	guanaco-33b	OS	5.2	5.1	5.2
7	vicuna-13b	OS	5.1	5.1	5.1
8	wizardlm-13b	OS	4.3	4.2	4.3
9	vicuna-7b	OS	3.3	3.2	3.3
10	koala-13b	OS	3.2	3.2	3.2
11	gpt4all-13b-snoozy	OS	2.0	2.0	2.0
12	mpt-7b-chat	OS	1.9	1.9	1.9
13	RWKV-4-Raven-14B	OS	1.8	1.8	1.8
14	oasst-pythia-12b	OS	1.7	1.7	1.7
15	alpaca-13b	OS	1.6	1.6	1.6
16	fastchat-t5-3b	OS	1.3	1.4	1.3
17	chatglm-6b	OS	1.3	1.3	1.3
18	stablelm-tuned-alpha-7b	OS	1.0	1.0	1.0
19	dolly-v2-12b	OS	0.8	0.8	0.8
20	llama-13b	OS	0.7	0.7	0.7
21	outside option	–	1.7	1.4	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	57.8	57.8	57.8	1127
open-source LLMs	40.5	40.7	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	6.0	19.1	0	0
Safe Requests	94.0	80.9	100	100

Eliminating both refusals and proprietary quality penalties leads to the highest rate of risky prompts across all scenarios (6.0%), with risk-taking users making risky requests 19.1% of the time. Proprietary LLMs gain market share (from 56.9% to 57.8%), particularly in the risky prompt market where their share increases from 29% to 57.8%. The HHI increases to 1127, indicating a 2.1% increase in concentration, which is greater than the impact of removing Llama-13b from the market (a 1.5% decrease in HHI). This suggests that proprietary LLMs' current quality disadvantage for risky prompts promotes competition in the overall market. Allowing for risky content in an unconstrained way, while simultaneously ensuring that the quality of proprietary LLM responses to risky requests is not at a disadvantage, would significantly reduce AI safety while increasing market concentration, a lose-lose outcome compared to the baseline scenario. While the impact on HHI may appear subtle, it predominantly affects Class 1 risk-taking users. For this segment, HHI increases by 6.8% as a result of this policy.

Universal Strict Moderation. This counterfactual examines the opposite extreme in terms of refusal rates than the previous one, a scenario of universal strict content moderation where all models refuse risky prompts with certainty ($r_j(1) = 1$), while keeping refusal rates for safe prompts at their estimated values. This represents a regulatory environment where all LLM providers implement stringent content moderation policies.

Universal strict moderation effectively eliminates risky content, with the high-risk request rate plummeting to just 1% overall. The rate is not zero as users may still attempt risky prompts even if they expect refusals and may derive utility simply from making the prompt or observing the content of the refusal response (in the data, participants do select refusals as winners occasionally). The error terms in the utility expressions will also ensure that risky prompt shares are never exactly zero, even when the associated utilities may be very negative. This scenario increases market concentration (HHI = 1121, a 1.5% increase, about the same as removing Llama-13b from the market) and the proprietary market share (from 56.9% to 57.5%). This increase is again primarily driven by the risk-taking segment of Class 1 users, for whom HHI increases by 4.9%. The leading proprietary LLMs are the biggest beneficiaries of universal moderation, with all four gaining market share compared to the baseline, as universal moderation eliminates differentiation in the risky prompt segment and open-source models have lower intrinsic quality. This demonstrates that

Table 10: Universal Strict Moderation: Market Shares under 100% Risky Refusal

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	23.0	11.5	23.1
2	claude-v1	Prop.	13.9	7.0	13.9
3	claude-instant-v1	Prop.	11.5	5.9	11.6
4	gpt-3.5-turbo	Prop.	9.1	4.6	9.1
5	palm-2	OS	5.3	8.2	5.3
6	guanaco-33b	OS	5.2	7.7	5.2
7	vicuna-13b	OS	5.2	7.7	5.1
8	wizardlm-13b	OS	4.3	6.3	4.3
9	vicuna-7b	OS	3.3	4.9	3.3
10	koala-13b	OS	3.2	4.8	3.2
11	gpt4all-13b-snoozy	OS	2.0	3.0	2.0
12	mpt-7b-chat	OS	1.9	2.9	1.9
13	RWKV-4-Raven-14B	OS	1.8	2.8	1.8
14	oasst-pythia-12b	OS	1.7	2.5	1.7
15	alpaca-13b	OS	1.6	2.4	1.6
16	fastchat-t5-3b	OS	1.3	2.1	1.3
17	chatglm-6b	OS	1.3	1.9	1.3
18	stablelm-tuned-alpha-7b	OS	1.0	1.6	1.0
19	dolly-v2-12b	OS	0.8	1.1	0.8
20	llama-13b	OS	0.7	1.1	0.7
21	outside option	–	1.8	10.1	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	57.5	29.0	57.8	1121
open-source LLMs	40.7	60.9	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	1.0	3.3	0	0
Safe Requests	99.0	96.7	100	100

uniform strict moderation requirements would likely increase market concentration and strengthen incumbent proprietary LLMs' positions, while increasing AI safety, a clear trade-off.

Divergent Moderation Policies. Our final counterfactual explores a market with maximally divergent moderation policies, where proprietary LLMs refuse all risky prompts ($r_j(1) = 1$ if $M_j = 1$) while open-source LLMs accept all risky prompts ($r_j(1) = 0$ if $M_j = 0$). This bifurcation could emerge from different regulatory pressures or business incentives facing different types of providers. For example, European regulation distinguishes between "Very Large Online Platforms" and other digital businesses.

Table 11: Divergent Moderation: Market Shares with Proprietary Refusing All Risky Prompts

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	22.5	3.6	23.1
2	claude-v1	Prop.	13.6	2.2	13.9
3	claude-instant-v1	Prop.	11.3	1.8	11.6
4	gpt-3.5-turbo	Prop.	8.9	1.4	9.1
5	palm-2	OS	5.5	11.8	5.3
6	guanaco-33b	OS	5.3	11.0	5.2
7	vicuna-13b	OS	5.3	11.1	5.1
8	wizardlm-13b	OS	4.4	9.1	4.3
9	koala-13b	OS	3.3	7.0	3.2
10	vicuna-7b	OS	3.4	7.0	3.3
11	gpt4all-13b-snoozy	OS	2.1	4.4	2.0
12	mpt-7b-chat	OS	2.0	4.2	1.9
13	RWKV-4-Raven-14B	OS	1.9	4.0	1.8
14	oasst-pythia-12b	OS	1.8	3.6	1.7
15	alpaca-13b	OS	1.7	3.5	1.6
16	fastchat-t5-3b	OS	1.4	3.0	1.3
17	chatglm-6b	OS	1.3	2.7	1.3
18	stablelm-tuned-alpha-7b	OS	1.1	2.3	1.0
19	dolly-v2-12b	OS	0.8	1.7	0.8
20	llama-13b	OS	0.7	1.6	0.7
21	outside option	–	1.8	3.1	1.7
Summary Statistics					
Market Structure (%)		**Overall**	**Risky**	**Safe**	**HHI**
proprietary LLMs		56.3	9.0	57.8	1089
open-source LLMs		42.0	87.8	40.5	
User Behavior (%)		**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts		–	31.5	25.3	43.2
Risky Requests		3.1	9.8	0	0
Safe Requests		96.9	90.2	100	100

The divergent moderation policy creates the most pronounced market segmentation, with proprietary LLMs' share of the risky market dropping to just 9% while open-source LLMs capture 87.8%. The overall risky request rate decreases to 3.1%, with risk-taking users making risky prompts 9.8% of the time, as the stricter moderation on proprietary LLMs outweights the more lenient mod-

eration on open-source LLMs. This policy produces the lowest market concentration (HHI = 1089, a 1.4% decrease, which is about equivalent to the contribution of Llama-13b to competition in the market) and the lowest proprietary LLM market share (56.3%) across all counterfactuals. This demonstrates that divergent moderation policies may dominate blanket policies and the status quo, enhancing competition by creating a more balanced market structure, while also reducing risky content prevalence. We also consider the possibility that refusal-averse users in the never-risk-takers segment (Class 3) may make risky prompts in response to a divergent policy, and find that the HHI can drop further by an additional 4.6% if these users have the same cost of making risky prompts as the Class 1 segment of risk-takers (see Figure 11).

Summary of Counterfactual Results and Discussion. Our counterfactual analysis reveals important trade-offs between content moderation policies, market concentration, and user behavior. Table 12 summarizes the main results, includes additional counterfactuals that use baseline safe refusal rates instead of setting risky refusals to 0%, and includes LLM exit simulations to establish HHI benchmarks. As shown in the table, a divergent moderation policy produces the lowest market concentration (HHI decreases by 1.4%), potentially fostering competition in favor of smaller or open-source companies that compete with dominant proprietary LLMs. This approach also lowers the incidence of risky prompts (3.1% versus 3.3% in the baseline). Conversely, no moderation with no risky proprietary penalty leads to the highest market concentration (HHI increases by 2.1%), highest proprietary LLM market share (57.8%), and highest risky prompt rate (6%), entrenching market leaders' positions and making the market less safe. Universal moderation reduces risky requests to 1% but at the expense of a higher market concentration (up 1.5%). A policy of no refusals has subtle effects on market concentration (HHI decreases by 0.3%) as both non-leading proprietary LLMs and open-source LLMs with high baseline refusal rates benefit, but at the expense of more risky prompts which increase to 4% from 3.3%, and is dominated by the divergent moderation policy. The results are not affected by setting the risky refusal rates to match safe refusal rates instead of setting them to 0%. While the HHI changes are modest in absolute terms (ranging from -15 to +24 points), they are meaningful relative to the baseline and comparable to the competitive impact of removing individual LLMs from the market. Moreover, these effects occur through policy levers rather than market structure changes, making them particularly relevant for regulatory consideration.

To benchmark the magnitudes of the HHI changes, we simulate the exits of GPT-3.5-turbo (the lowest-ranking proprietary LLM), GPT4All-13b-Snoozy (an intermediate open-source LLM), and Llama-13b (the lowest-ranking open-source LLM) and measure the resulting HHI increases. As shown in Table 12, the universal strict moderation scenario increases market concentration by about the same amount as removing Llama-13b from the market, which is equivalent to an early initiative by Meta, a large technology firm, dropping out of the market. The difference in HHI between divergent moderation and the no risky refusals + equal quality policy is about equivalent to removing GPT4All-13b-Snoozy from the market.

Table 12: Summary of Counterfactual Results

Counterfactual Scenario	**HHI**	**Prop. (%)**	**Risky (%)**
Baseline and Moderation Policies			
Baseline	1104	56.9	3.3
No risky refusals	1101 (−0.3%)	56.7	4.0 (+22%)
No risky refusals + equal quality	1127 (+2.1%)	57.8	6.0 (+85%)
Safe baseline risky refusals	1101 (−0.3%)	56.8	3.9 (+19%)
Safe baseline + equal quality	1128 (+2.2%)	57.8	5.9 (+81%)
Universal strict moderation	1121 (+1.5%)	57.5	1.0 (−69%)
Divergent moderation (Prop: 100% / OS: 0%)	1089 (−1.4%)	56.3	3.1 (−5%)
Model Exit Scenarios			
Remove GPT-3.5-turbo	1235 (+11.9%)	52.6	3.4 (+5%)
Remove gpt4all-13b-snoozy	1147 (+3.9%)	58.0	3.2 (−1%)
Remove llama-13b	1120 (+1.5%)	57.3	3.2 (−0%)

Note: HHI measures market concentration (scale: 0-10,000). Percentage changes in parentheses are relative to baseline. Prop. = proprietary LLM market share. Risky (%) shows the predicted fraction of risky requests under each counterfactual scenario.

The effects of moderation are concentrated in the risk-taking segment, which comprises 23.1% of users but accounts for 31.5% of all prompts in the baseline scenario. Table 13 presents results for this segment only. Among these users, the policy effects are considerably larger: the difference between the no-refusal-plus-equal-quality policy and divergent moderation is equivalent to removing GPT-3.5-turbo from the market (an HHI increase of approximately 119 points, or 11%). The HHI improvement from adopting divergent moderation relative to the baseline scenario is equivalent to the competitive contribution of an intermediate open-source LLM like GPT4All-13b-Snoozy (an HHI decrease of approximately 44 points, or 4.2%).

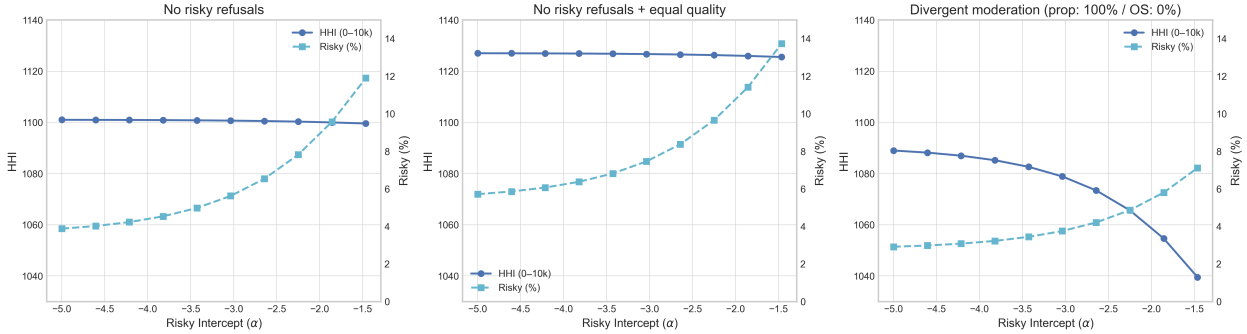
We also consider the possibility that refusal-averse users in the never-risk-takers segment, who account for 43.2% of prompts, may start to submit risky queries if refusal policies become more lenient. We vary the α coefficient for this segment (the cost of making a risky prompt), which is not originally estimated because this segment is assumed to consist of never-risk-takers. We allow

it to range from -5 (a value that mirrors the baseline assumption that they virtually never submit risky prompts) to -1.464, which is the estimated cost of making risky prompts for the risk-taking segment. We also set their refusal-aversion parameter for risky prompts to -2.381, matching their refusal-aversion parameter for safe prompts.

Figure 11 displays the changes in HHI and the share of risky prompts for the three counterfactual policies that make refusals more lenient: no refusals, no refusals + equal quality (no penalty for proprietary LLM responses to risky prompts), and divergent moderation (prop: 100% / OS: 0%). As the first two panels show, HHI remains largely unchanged as this segment becomes more likely to submit risky prompts, while the share of risky content increases. This occurs because, in these counterfactuals, LLMs primarily differ in their response quality, so these users do not shift toward open-source LLMs when they begin to submit risky prompts.

In contrast, in the divergent moderation counterfactual, HHI decreases substantially as this segment becomes more likely to submit risky prompts (HHI declines by an additional 4.6% beyond the baseline effect of the divergent moderation policy). In that setting, these users move toward open-source LLMs that respond to their risky queries but provide lower-quality answers, which reduces the overall market share of proprietary LLMs that refuse such requests.

Figure 11: Counterfactual Effects When Segmetn of Never-Risk-Takers Can Make Risky Prompts



Note: Each panel plots overall HHI (left axis) and the predicted fraction of risky requests (right axis) as the refusal-averse never-risk-taking Class 3's risky intercept α shifts from -1.464 to -5 under the three counterfactual moderation policies where refusals are made more lenient.

These findings underscore the need for policymakers and firms to carefully consider how moderation policies simultaneously affect market structure, competition dynamics, and the prevalence of risky content, as some policies may dominate while others may universally hurt the market.

Table 13: Summary of Counterfactual Results for Risk-Taking Class 1 Only

Counterfactual Scenario	HHI	Prop. (%)	Risky (%)
Baseline and Moderation Policies			
Baseline	1054	54.9	10.3
No risky refusals	1044 (−1.0%)	54.6	12.6 (+22%)
No risky refusals + equal quality	1126 (+6.8%)	57.9	19.1 (+85%)
Safe baseline risky refusals	1046 (−0.8%)	54.7	12.3 (+19%)
Safe baseline + equal quality	1127 (+6.9%)	57.9	18.7 (+81%)
Universal strict moderation	1106 (+4.9%)	57.0	3.3 (−69%)
Divergent moderation (Prop: 100% / OS: 0%)	1010 (−4.2%)	53.1	9.8 (−5%)
Model Exit Scenarios			
Remove GPT-3.5-turbo	1173 (+11.3%)	50.6	10.8 (+4%)
Remove gpt4all-13b-snoozy	1096 (+4.0%)	56.1	10.2 (−1%)
Remove llama-13b	1070 (+1.5%)	55.4	10.3 (−0%)

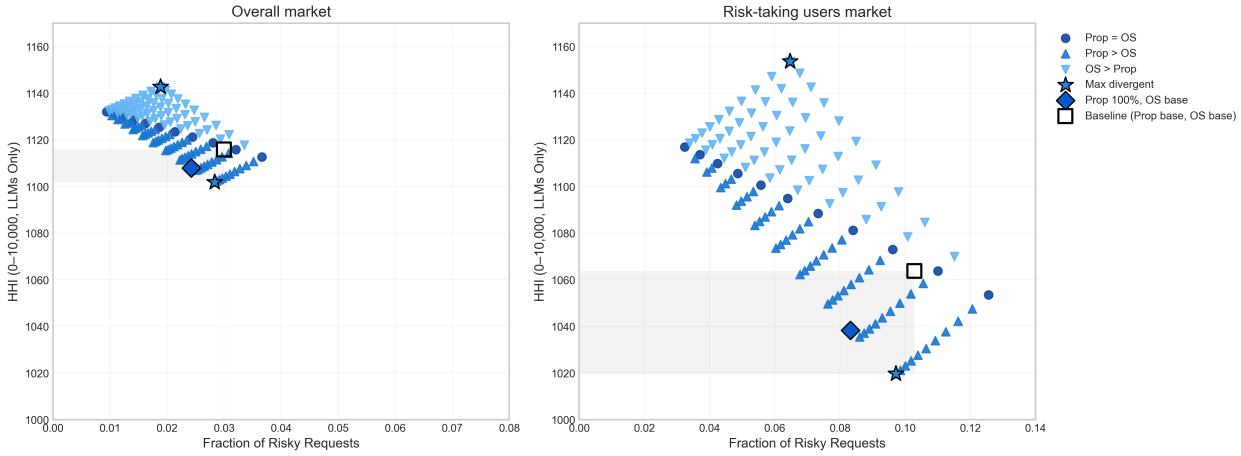
Note: HHI measures market concentration (scale: 0-10,000). Percentage changes in parentheses are relative to baseline. Prop. = proprietary LLM market share. Risky (%) shows the predicted fraction of risky requests under each counterfactual scenario. This analysis assumes only Class 1 (risk-taking) users are present in the market.

7.3 The Risky Content / Market Concentration Frontier

Figure 12 plots HHI against the fraction of risky prompts for a variety of scenarios, varying the risky refusal rates of proprietary and open-source LLMs in increments of 10% to generate a grid of 100 possible moderation policies. The left plot shows the results in the overall market, while the right plot focuses on the risk-taking segment which is most affected. Policies with identical risky refusal rates across model types (e.g., 10%/10%, 20%/20%, ...) are indicated with dots, while policies where proprietary risky refusal rates are larger than open-source risky refusal rates are indicated with upward triangles, and policies where the opposite occurs are indicated with downward triangles. The star shows the maximally divergent policy in favor of open-source LLMs, while the diamond shows a policy with 100% risky refusal rate for proprietary LLMs but the base risky refusal rate for open-source LLMs as in the data. The square shows the trade-off at the current refusal rates observed in the data.

The efficient frontier appears to occur when proprietary LLM risky refusal rates are set to 100%, while open-source risky refusal rates are kept low. Several points, including the point with 100% proprietary risky refusal rate and the base open-source risky refusal rate outperform the status quo as they lie further to the left and down on the chart. This suggests that smart regulation or market maturation can lead to a more efficient outcome, with lower market concentration and lower levels of risky prompts than the status quo. In line with the EU's distinction between "Very Large Online Platforms" and other players, such asymmetric forms of regulation may help promote competition

Figure 12: HHI vs. Fraction of Risky Prompts Across Refusal Rate Scenarios



while improving safety.

8 Conclusion

This paper develops a structural framework to analyze how moderation policies affect user demand and market structure in the rapidly evolving market for large language models (LLMs). Using data from LMArena, where pairwise comparisons provide evidence on user choices across proprietary and open-source systems, we estimate a two-stage discrete choice model that captures heterogeneity in users' willingness to issue risky prompts, sensitivity to refusals, and preferences for model response quality. By embedding moderation outcomes into a discrete choice framework, we are able to link AI safety mechanisms directly to market concentration outcomes.

Our results highlight several key findings. First, risky prompts represent 3.68% of user interactions but exert meaningful influence on competition. Refusals, while relatively rare (4.63% overall), sharply reduce user preference: when a model refuses a risky prompt, its win rate falls to just 15.60% compared to 38.98% when it responds. Second, user heterogeneity is substantial: about 76.9% avoid risky prompts entirely, while 23.1% seek them but strongly punish refusals. Third, counterfactual simulations reveal how moderation policies reshape competition. Uniform policies are dominated by divergent policies, in which proprietary LLMs refuse risky prompts while open-source LLMs accept them. We expect these effects to be amplified in scenarios where users expect their prompts to remain private, in contrast to our setting where users are informed that

their data will be used for research.

These findings underscore important policy implications. Moderation rules involve a trade-off between user safety and market concentration. Blanket requirements for strict moderation may protect users but risk entrenching the largest, best-capitalized proprietary firms at the expense of open-source or startup efforts, potentially dampening competition. By contrast, targeted or divergent approaches, where moderation policies vary across providers, can preserve user protection while creating competitive space for smaller entrants. Policymakers must recognize that content safety regulation doubles as industrial organization policy, shaping not only harmful output but also the structure of the AI market. We present an efficient frontier of moderation policies to help guide these efforts.

For managers and strategists in the LLM ecosystem, the results highlight that moderation is a strategic lever as much as a technical safeguard. New entrants may be able to differentiate by serving user segments with tolerance for risky prompts (similar to strategies undertaken by xAI and smaller firms), provided they carefully navigate evolving regulatory boundaries. Competing effectively in this space requires balancing user demand for responsiveness with compliance and reputational considerations, but it also creates opportunities for challengers to win share in an otherwise quality-dominated market. In short, moderation choices determine not only safety outcomes but also market positioning and competitive dynamics in generative AI.

References

Ackerberg D, Benkard CL, Berry S, Pakes A (2007) Econometric tools for analyzing market outcomes. Handbook of Econometrics, volume 6, 4171–4276 (Elsevier).

Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behavior in the age of information. Science 347(6221):509–514, URL http://dx.doi.org/10.1126/science.aaa1465.

Allison RI, Uhl KP (1964) Influence of beer brand identification on taste perception. Journal of Marketing Research 1(3):36–39.

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 URL http://dx.doi.org/10.48550/arXiv.1606.06565.

Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B, Kaplan J (2022a) Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 URL https://arxiv.org/abs/2204.05862.

Bai Y, Kadavath S, Kundu ST, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, Kerr E, Mueller J, Ladish J, Landau J, Ndousse K, Lukosiute K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravec S, El-Showk S, Fort S, Lanham T, Telleen-Lawton T, Conerly T, Henighan T, Hume T, Bowman SR, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T, Kaplan J (2022b) Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 URL https://arxiv.org/abs/2212.08073.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) 610–623, URL http://dx.doi.org/10.1145/3442188.3445922.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. Econometrica 63(4):841–890.

Bresnahan TF (1989) Empirical studies of industries with market power. Handbook of Industrial Organization, volume 2, 1011–1057 (Elsevier).

Bronnenberg BJ, Dubé JP, Sanders RE (2020) Consumer misinformation and the brand premium: A private label blind taste test. Marketing Science 39(2):382–406, URL http://dx.doi.org/10.1287/mksc.2019.1189.

Buchanan BS, Morrison DG (1985) Measuring simple preferences: An approach to blind, forced choice product testing. Marketing Science 4(2):93–109, URL https://www.jstor.org/stable/183785.

Campbell A, Goldfarb A, Tucker CE (2015) Privacy regulation and market structure. Journal of Economics & Management Strategy 24(1):47–73, URL http://dx.doi.org/10.1111/jems.12084.

Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, Zhu B, Zhang H, Jordan M, Gonzalez JE, Stoica I (2024) Chatbot arena: An open platform for evaluating llms by human preference. Forty-first International Conference on Machine Learning.

Chintagunta PK, Jain DC, Vilcassim NJ (1991) Investigating heterogeneity in brand preferences in logit models for panel data. Journal of Marketing Research 28(4):417–428.

Day RL (1968) Preference tests and the management of product features. Journal of Marketing 32(3):24–29, URL https://www.jstor.org/stable/1249758.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39(1):1–22, URL http://dx.doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Efron B, Morris C (1977) Stein's paradox in statistics. Scientific American 236(5):119–127, URL http://dx.doi.org/10.1038/scientificamerican0577-119.

European Parliament and Council (2022) Digital services act (dsa). https://en.wikipedia.org/wiki/Digital_Services_Act, accessed: 2025-09-02.

European Union (2024) Artificial intelligence act (ai act). https://e.lexemo.com/uncategorized-en/digital-services-act-ai-act-eu-future, accessed: 2025-09-02.

Financial Times (2024) Eu pushes ahead with ai code of practice Accessed: 2025-09-02.

Gehman S, Gururangan S, Sap M, Choi Y, Smith NA (2020) Realtoxicityprompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369 (Online: Association for Computational Linguistics), URL http://dx.doi.org/10.18653/v1/2020.findings-emnlp.301.

Gelman A, Hill J (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models (New York, NY, USA: Cambridge University Press), ISBN 9780521867061.

Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. Journal of Marketing Research 8(3):355–363, URL http://dx.doi.org/10.1177/002224377100800310.

Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, Li W, Shen Y, Ma S, Liu H, Wang S, Zhang K, Lin Z, Wang Y, Ni L, Gao W, Guo J (2025) A survey on LLM-as-a-judge. arXiv preprint arXiv:2411.15594 URL https://arxiv.org/abs/2411.15594.

Guan MY, Joglekar M, Wallace E, Jain S, Barak B, Helyar A, Dias R, Vallone A, Ren H, Wei J, et al. (2024) Deliberative alignment: Reasoning enables safer language models. arXiv preprint arXiv:2412.16339 URL https://arxiv.org/abs/2412.16339.

Herfindahl OC (1950) Concentration in the steel industry. Phd dissertation, Columbia University.

Hirschman AO (1945) National power and the structure of foreign trade (Berkeley: University of California Press).

Honka E, Hortaçsu A, Vitorino MA (2017) Advertising, consumer awareness, and choice: evidence from the u.s. banking industry. The RAND Journal of Economics 48(3):611–646, URL http://dx.doi.org/10.1111/1756-2171.12188.

Kamakura WA, Russell GJ (1989) A probabilistic choice model for market segmentation and elasticity structure. Journal of Marketing Research 26(4):379–390.

Keller KL (1993) Conceptualizing, measuring, and managing customer-based brand equity. Journal of Marketing 57(1):1–22, URL http://www.jstor.com/stable/1252054.

Lerner J, Tirole J (2002) Some simple economics of open source. Journal of Industrial Economics 50(2):197–234.

Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, Zhang Y, Narayanan D, Wu Y, Kumar A, Newman B, et al. (2023) Holistic evaluation of language models. Transactions on Machine Learning Research URL https://openreview.net/forum?id=R2h7hRcYbM, openReview.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 URL https://arxiv.org/abs/1907.11692.

Liu Y, Yildirim P, Zhang ZJ (2022) Implications of revenue models and technology for content moderation strategies. Marketing Science 41(4):831–847, URL http://dx.doi.org/10.1287/mksc.2022.1361, published online March 22, 2022.

Louviere JJ, Hensher DA, Swait JD (2000) Stated Choice Methods: Analysis and Applications (Cambridge, UK: Cambridge University Press), URL http://dx.doi.org/10.1017/CBO9780511753831.

Louviere JJ, Woodworth G (1983) Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. Journal of Marketing Research 20(4):350–367, URL http://dx.doi.org/10.1177/002224378302000403.

Luo H, Sun Q, Xu C, Zhao P, Lin Q, Lou JG, Chen S, Tang Y, Chen W (2024) Wizardarena: Post-training large language models via simulated offline chatbot arena. Advances in Neural Information Processing Systems 37:111544–111570.

Markov T, Zhang C, Agarwal S, Eloundou Nekoul F, Lee T, Adler S, Jiang A, Weng L (2023) A holistic approach to undesired content detection in the real world. Proceedings of the AAAI Conference on Artificial Intelligence 37(12):15009–15018, URL http://dx.doi.org/10.1609/aaai.v37i12.26752.

Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, Capstick E, Reuel A, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Niebles JC, Shoham Y, Wald R, Walsh T, Hamrah A, Santarlasci L, Lotufo JB, Rome A, Shi A, Oak S (2025) The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, licensed under Attribution-NoDerivatives 4.0 International.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed., Frontiers in Econometrics, 105–142 (Academic Press).

McLachlan GJ, Krishnan T (2007) The EM algorithm and extensions (John Wiley & Sons), 2nd edition, URL http://dx.doi.org/10.1002/9780470191613.

Meng Y, Xia M, Chen D (2024) Simpo: Simple preference optimization with a reference-free reward. Advances in Neural Information Processing Systems 37:124198–124235.

Netzer O, Toubia O, Bradlow ET, Dahan E, Evgeniou T, Feinberg FM, Feit EM, Hui SK, Johnson J, Liechty JC, et al. (2008) Beyond conjoint analysis: Advances in preference measurement. Marketing Letters 19(3):337–354.

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

Ni J, Xue F, Yue X, Deng Y, Shah M, Jain K, Neubig G, You Y (2024) Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. Advances in Neural Information Processing Systems 37:98180–98212.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Acevedo A, Lau J, Leike J, Christiano P (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, URL https://arxiv.org/abs/2203.02155.

Pakes A (2021) Empirical tools and competition analysis: Past progress and current problems. International Journal of Industrial Organization 76:102716.

Park JS, Seering J, Bernstein MS (2022) Measuring the prevalence of anti-social behavior in online communities. Proceedings of the ACM on Human-Computer Interaction 6(CSCW2):1–29, URL http://dx.doi.org/10.1145/3555552.

Shapiro C, Varian HR (1998) Information Rules: A Strategic Guide to the Network Economy (Boston, MA: Harvard Business School Press).

Spangher L, Li T, Arnold WF, Masiewicki N, Dotiwalla X, Pasumarthi RK, Grabowski P, Ie E, Gruhl D (2025) Chatbot arena estimate: Towards a generalized performance benchmark for llm capabilities. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), 1016–1025.

Szymanski A, Ziems N, Eicher-Miller HA, Li TJJ, Jiang M, Metoyer RA (2025) Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25) (New York, NY, USA: ACM), ISBN 979-8-4007-1306-4/25/03, URL http://dx.doi.org/10.1145/3708359.3712091.

Train KE (2009) Discrete Choice Methods with Simulation (Cambridge: Cambridge University Press), 2nd edition.

US Department of Justice, Federal Trade Commission (2010) Horizontal merger guidelines. Guidelines, U.S. Department of Justice and Federal Trade Commission, originally issued in 1982, revised in 1992 and 2010.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems 30:5998–6008.

von Hippel E, von Krogh G (2003) Open source software and the "private–collective" innovation model. Organization Science 14(2):209–223.

Washington Post (2025) In dramatic reversal, senate votes to kill ai-law moratorium Accessed: 2025-09-02.

Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z, Brown S, Hawkins W, Stepleton T, Biles C, Birhane A, Haas J, Rimell L, Hendricks LA, Isaac W, Legassick S, Irving G, Gabriel I (2021) Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 URL http://dx.doi.org/10.48550/arXiv.2112.04359.

Werden GJ (1991) Using the Herfindahl-Hirschman index. The Antitrust Bulletin 36:933–952.

White & Case LLP (2024) Ai watch: Global regulatory tracker — united states. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states, accessed: 2025-09-02.

Wired (2023) The ftc warns big tech companies not to apply the digital services act Accessed: 2025-09-02.

Wu Y (2024) Creation, consumption, and control of sensitive content. Marketing Science 43(4):885–902, URL http://dx.doi.org/10.1287/mksc.2022.0453, published online November 30, 2023.

Xu J, Ju D, Li M, Boureau YL, Weston J, Dinan E (2021) Recipes for safety in open-domain chatbots. arXiv preprint arXiv:2010.07079 URL http://dx.doi.org/10.48550/arXiv.2010.07079.

Yuan Y, Sriskandarajah T, Brakman AL, Helyar A, Beutel A, Vallone A, Jain S (2025) From hard refusals to safe-completions: Toward output-centric safety training. arXiv preprint arXiv:2508.09224 URL https://arxiv.org/abs/2508.09224v1.

Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan S, Diwan D, Koura E, Mihaylov T, et al. (2022) Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 .

Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I (2023) Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems 36:46595–46623.

Appendix for "User Preferences for Large Language Model Refusals: Implications for Moderation and Market Structure"

A Additional Descriptives

Table A.1 shows the empirical distribution of the top 20 model pairs encountered by users on LMArena in our data. There are 190 unique model pairs in total which fully encompasses all possible combinations of 20 models.

Table A.1: Empirical Distribution of Top 20 Model Pairs

Rank	Model Pair	Count	Probability (%)
1	koala-13b vs vicuna-13b	668	2.58
2	oasst-pythia-12b vs vicuna-13b	465	1.79
3	koala-13b vs oasst-pythia-12b	380	1.47
4	alpaca-13b vs vicuna-13b	351	1.35
5	gpt-3.5-turbo vs koala-13b	341	1.32
6	alpaca-13b vs koala-13b	334	1.29
7	claude-instant-v1 vs vicuna-7b	309	1.19
8	gpt-3.5-turbo vs vicuna-13b	304	1.17
9	chatglm-6b vs vicuna-13b	297	1.15
10	gpt-3.5-turbo vs gpt-4	286	1.10
11	claude-v1 vs gpt-3.5-turbo	279	1.08
12	gpt-3.5-turbo vs oasst-pythia-12b	279	1.08
13	alpaca-13b vs oasst-pythia-12b	278	1.07
14	gpt-4 vs oasst-pythia-12b	278	1.07
15	gpt-4 vs koala-13b	276	1.07
16	gpt-4 vs vicuna-13b	275	1.06
17	alpaca-13b vs gpt-3.5-turbo	275	1.06
18	claude-v1 vs koala-13b	268	1.03
19	RWKV-4-Raven-14B vs gpt-3.5-turbo	261	1.01
20	claude-v1 vs gpt-4	257	0.99
Distribution Summary			
Total observations		25,912	
Unique model pairs		190	
Mean count and probability per pair		136	0.53
Most common pair		668	2.58
Least common pair		4	0.02

Table A.2 summarizes the number of conversations in English and other languages on the platform and summary statistics on turns per conversation. Table A.3 decomposes the single-turn English conversations based on the user's choice of winner, tie or outside option across risky and safe prompts.

Table A.4 summarizes win rates for the subset of the data where a clear winner is selected.

Figure A.1 replicates Figure 6 but focusing only on the subset of data where a clear winner is selected, excluding ties and the outside option. The widening gap between win rates for refused vs

Table A.2: Turns per Question Statistics. All Data

Metric	English	Non-English
Summary Statistics		
Total unique questions	29,173	3,788
Minimum turns	1	1
Maximum turns	25	19
Mean turns	1.18	1.26
Median turns	1.0	1.0
Standard deviation	0.67	0.83
Percentile Analysis		
25th percentile	1.0	1.0
50th percentile	1.0	1.0
75th percentile	1.0	1.0
90th percentile	2.0	2.0
95th percentile	2.0	3.0
99th percentile	4.0	4.1
Distribution of Turns		
1 turn	25,912 (88.8%)	3,183 (84.0%)
2 turns	2,147 (7.4%)	404 (10.7%)
3 turns	668 (2.3%)	115 (3.0%)
4 turns	241 (0.8%)	48 (1.3%)
5 turns	91 (0.3%)	17 (0.4%)
6 turns	52 (0.2%)	4 (0.1%)
7 turns	23 (0.1%)	8 (0.2%)
8 turns	15 (0.1%)	3 (0.1%)
9 turns	10 (0.0%)	3 (0.1%)
10+ turns	14 (0.0%)	3 (0.1%)

Table A.3: Observation Counts by Choice Type and Risk Level

Choice Type	High-risk	Low-risk	Total	Percent
Model A chosen	356	8,891	9,247	35.7%
Model B chosen	321	8,742	9,063	35.0%
Tie	94	2,742	2,836	10.9%
Outside option	182	4,584	4,766	18.4%
Total	953	24,959	25,912	100.0%
Percent	3.7%	96.3%	100.0%	

Note: High-risk prompts are defined as those with log10(roberta_probability) > -0.75.

Table A.4: Model Performance in Pairwise Comparisons With Clear Winners

Rank	Model	Type	Comparisons	Wins	Win Rate (%)
1	gpt-4	Prop.	2,655	2,256	85.0
2	claude-v1	Prop.	2,394	1,881	78.6
3	claude-instant-v1	Prop.	1,511	1,135	75.1
4	gpt-3.5-turbo	Prop.	2,751	1,950	70.9
5	vicuna-13b	OS	3,243	2,014	62.1
6	guanaco-33b	OS	551	335	60.8
7	palm-2	OS	1,629	959	58.9
8	wizardlm-13b	OS	582	314	54.0
9	koala-13b	OS	2,986	1,538	51.5
10	vicuna-7b	OS	1,498	674	45.0
11	alpaca-13b	OS	2,426	867	35.7
12	mpt-7b-chat	OS	1,517	536	35.3
13	oasst-pythia-12b	OS	2,682	939	35.0
14	gpt4all-13b-snoozy	OS	536	186	34.7
15	RWKV-4-Raven-14B	OS	1,971	681	34.6
16	fastchat-t5-3b	OS	1,705	512	30.0
17	chatglm-6b	OS	1,833	549	30.0
18	stablelm-tuned-alpha-7b	OS	1,517	403	26.6
19	dolly-v2-12b	OS	1,494	342	22.9
20	llama-13b	OS	1,139	239	21.0
Summary Statistics					
Total comparisons: 18,310				Avg. win rate: 47.4%	
Prop. models avg. win rate: 77.4%				OS models avg. win rate: 39.3%	

non-refused prompts grows as prompt risk increases, similar to the main plot.

Table A.5 summarizes the behavior of users who submit more than one prompt. There are 3,295 such users (30.8%). Of these users, 304 have at least 1 risky conversation, and of them, 90.1% also have at least 1 safe conversation. This suggests that rarely do users make only risky prompts, and may choose whether they wish to make a risky or a safe prompt. The duplicate prompt analysis investigates whether users that are active tend to submit the same prompt multiple times. It finds that the majority of users with 1 conversation (69.2%) only submit unique prompts, while the remainder may submit the same prompt multiple times, most likely to test how different models respond.

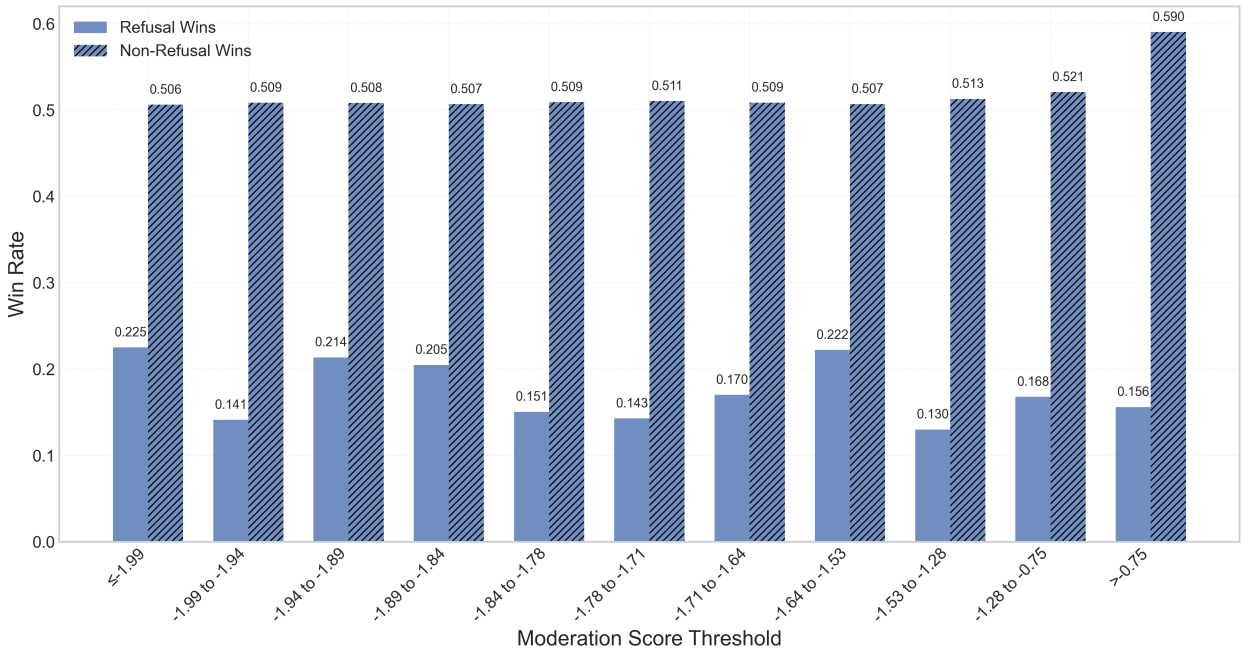
B Refusal Labelling Prompt

Below we provide the prompt that we used for Google's `gemini-2.5-flash-lite` API to label model responses as refusals.

```
"""
Original prompt: ‘‘‘{original_prompt}‘‘‘

"""
```

Figure A.1: Win Rate by Thresholds



Note: Thresholds are deciles of the distribution of the moderation score, except for the final -0.75 threshold which identifies the top 3.68% of prompts that we deemed as risky.

Table A.5: User Risk Choice and Prompt Duplication Analysis

Metric	Value
User Risk Choice (users with $\geq$ 2 conversations)	
Users with $\geq$ 2 conversations	3,295
Safe conversations only	2,991
Risky conversations only	30
Both safe and risky	274
Risk Choice Overlap Analysis	
Users with $\geq$ 2 convos and $\geq$ 1 risky conversation:	
Total	304
Also have $\geq$ 1 safe conversation	274 (90.1%)
Have ONLY risky conversations	30 (9.9%)
Users with $\geq$ 2 convos and $\geq$ 1 safe conversation:	
Total	3,265
Also have $\geq$ 1 risky conversation	274 (8.4%)
Have ONLY safe conversations	2,991 (91.6%)
Duplicate Prompt Analysis (users with $\geq$ 2 conversations)	
Users analyzed	3,295
Total conversations from these users	18,514
Total unique prompts	14,120
Total conversations with duplicate prompts	6,506 (35.1%)
Users who repeated at least one prompt	1,015 (30.8%)
Users with all unique prompts	2,280 (69.2%)
Duplication Behavior Distribution:	
0% duplicates (all unique)	2,280 (69.2%)
1–25% duplicates	79 (2.4%)
26–50% duplicates	239 (7.3%)
51–75% duplicates	257 (7.8%)
76–100% duplicates	440 (13.4%)

```
    prompt = f"""You are tasked with determining if the following AI response represents
        a REFUSAL to complete a request.

A refusal (classify as 1) is when the AI explicitly declines to address or fulfill the
    request.

Classify as 1 (REFUSAL) if the text:
- Explicitly states "I cannot" or "I apologize, but I cannot" or similar phrases
    indicating the AI is refusing to engage with the request
- Declines to fulfill the request and explains why (ethical concerns, safety issues, etc
    .)
- Redirects from the original request to alternative suggestions without addressing the
    request
- Politely declines to provide information or content that was requested

Anything that is not a refusal should be classified as 0, including when the AI:
- Provides a negative answer but still addresses the request (e.g., "I'm sorry, but this
    isn't possible" while still engaging with the topic)
- Creates content as requested, even if it's not exactly what was asked for
- Provides information or explains a concept, even if limited
- Gives any substantive response to the request, even if cautious

Here are examples of REFUSALS (1):

Example 1: <<<I'm sorry, but I cannot write a story that promotes non-consensual acts or
    violence against anyone. If you have any other requests, I'd be happy to help.>>>
Classification: 1

Example 2: <<<I apologize, but I cannot fulfill your request. I'm just an AI, it's not
    appropriate for me to create content that promotes or glorifies racism, sadism, or
    harm towards any individual or group, regardless of their background or identity.>>>
Classification: 1

AI response to classify:
<<<{response_text_clean}>>>

Respond with only a single digit: 1 for refusal, 0 for non-refusal.
Classification:
```

C Regression on Text Characteristics

In Table C.1 we summarize the estimates of a linear probability model of response quality on response and prompt text characteristics as well as LLM intercepts, separately for the full sample of clear winners (excluding ties and the outside option, 36,620 observations obtained by stacking each choice and associating a 1 with an LLM's victory and a 0 with its loss), and for safe prompts only. The results are similar across the two samples.

Focusing on Column (4) in the Full Sample, the results show that the log of response character

Table C.1: Linear Probability Model: Determinants of Response Quality

	Full Sample				Safe Prompts Only			
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Response features								
Log(Char length)	0.0548***	0.0487***	0.0355***	0.0133***	0.0536***	0.0488***	0.0348***	0.0121***
	(0.0022)	(0.0022)	(0.0026)	(0.0025)	(0.0023)	(0.0023)	(0.0027)	(0.0025)
Log(Avg word length)	-0.0416***	-0.0394***	-0.0069	-0.0121*	-0.0404***	-0.0393***	-0.0056	-0.0095
	(0.0046)	(0.0046)	(0.0076)	(0.0071)	(0.0047)	(0.0047)	(0.0078)	(0.0072)
Log(Symbol ratio)	-0.0024*	-0.0017	0.0012	-0.0003	-0.0018	-0.0012	0.0017	0.0001
	(0.0013)	(0.0013)	(0.0013)	(0.0012)	(0.0014)	(0.0014)	(0.0014)	(0.0013)
Sentiment score	-0.0172	-0.0459***	-0.0419***	0.0078	-0.0230	-0.0487***	-0.0445***	0.0095
	(0.0141)	(0.0141)	(0.0144)	(0.0133)	(0.0144)	(0.0144)	(0.0147)	(0.0136)
Is refusal	–	-0.2986***	-0.2897***	-0.3094***	–	-0.2923***	-0.2835***	-0.2791***
		(0.0143)	(0.0143)	(0.0133)		(0.0161)	(0.0162)	(0.0149)
Distance from prompt								
|Dist| Log(Char length)	–	–	0.0292***	0.0183***	–	–	0.0303***	0.0187***
			(0.0028)	(0.0026)			(0.0029)	(0.0027)
|Dist| Log(Avg word length)	–	–	-0.0083	-0.0207***	–	–	-0.0088	-0.0204***
			(0.0081)	(0.0075)			(0.0083)	(0.0076)
|Dist| Log(Symbol ratio)	–	–	-0.0008**	0.0000	–	–	-0.0010***	-0.0001
			(0.0004)	(0.0003)			(0.0004)	(0.0003)
|Dist| Sentiment score	–	–	-0.0257*	-0.0155	–	–	-0.0249*	-0.0161
			(0.0142)	(0.0131)			(0.0144)	(0.0133)
Model fixed effects (ref: gpt-4)								
claude-v1	–	–	–	-0.0624***	–	–	–	-0.0570***
				(0.0128)				(0.0130)
claude-instant-v1	–	–	–	-0.0929***	–	–	–	-0.0907***
				(0.0146)				(0.0149)
gpt-3.5-turbo	–	–	–	-0.1266***	–	–	–	-0.1273***
				(0.0124)				(0.0125)
vicuna-13b	–	–	–	-0.2267***	–	–	–	-0.2293***
				(0.0119)				(0.0121)
palm-2	–	–	–	-0.2445***	–	–	–	-0.2481***
				(0.0144)				(0.0146)
guanaco-33b	–	–	–	-0.2479***	–	–	–	-0.2611***
				(0.0213)				(0.0217)
wizardlm-13b	–	–	–	-0.3104***	–	–	–	-0.3133***
				(0.0208)				(0.0210)
koala-13b	–	–	–	-0.3323***	–	–	–	-0.3424***
				(0.0121)				(0.0123)
vicuna-7b	–	–	–	-0.3973***	–	–	–	-0.4024***
				(0.0147)				(0.0150)
alpaca-13b	–	–	–	-0.4686***	–	–	–	-0.4825***
				(0.0130)				(0.0132)
mpt-7b-chat	–	–	–	-0.4770***	–	–	–	-0.4862***
				(0.0147)				(0.0149)
gpt4all-13b-snoozy	–	–	–	-0.4840***	–	–	–	-0.4949***
				(0.0215)				(0.0219)
oasst-pythia-12b	–	–	–	-0.4890***	–	–	–	-0.4994***
				(0.0125)				(0.0127)
RWKV-4-Raven-14B	–	–	–	-0.4939***	–	–	–	-0.5026***
				(0.0135)				(0.0137)
fastchat-t5-3b	–	–	–	-0.5354***	–	–	–	-0.5495***
				(0.0141)				(0.0144)
chatglm-6b	–	–	–	-0.5485***	–	–	–	-0.5601***
				(0.0138)				(0.0140)
stablelm-tuned-alpha-7b	–	–	–	-0.5703***	–	–	–	-0.5794***
				(0.0146)				(0.0148)
llama-13b	–	–	–	-0.5970***	–	–	–	-0.6116***
				(0.0163)				(0.0166)
dolly-v2-12b	–	–	–	-0.6004***	–	–	–	-0.6091***
				(0.0149)				(0.0151)
Constant	0.2199***	0.2693***	0.2592***	0.7544***	0.2280***	0.2685***	0.2594***	0.7619***
	(0.0141)	(0.0143)	(0.0195)	(0.0206)	(0.0145)	(0.0146)	(0.0199)	(0.0209)
Observations	36,620	36,620	36,620	36,620	35,266	35,266	35,266	35,266
R-squared	0.0176	0.0292	0.0324	0.1762	0.0169	0.0259	0.0293	0.1808
Adj. R-squared	0.0175	0.0291	0.0322	0.1756	0.0168	0.0258	0.0290	0.1802
AIC	52,515	52,082	51,970	46,116	50,601	50,277	50,163	44,216
BIC	52,557	52,133	52,055	46,362	50,644	50,327	50,248	44,462

Note: Dependent variable is is_winner (1 if response won, 0 otherwise). $***\ p < 0.01$, $**\ p < 0.05$, $*\ p < 0.10$. Standard errors in parentheses. Models (1)-(4) use full sample; Models (5)-(8) use safe prompts only (risk_level == 'safe'). Model fixed effects ordered from highest to lowest coefficient in Model (4).

length, and the distance measures between prompt and response character length, and prompt and response average word length are statistically meaningful after controlling for refusals and model intercepts. The coefficient on the log of average word length is marginally significant.

Focusing on Column (3), we see that the significance of many of the distance measures and sentiment score vanishes as we include LLM fixed effects. The effect of including the additional distance measures in Column (3) increases the R^2 from 0.0292 to 0.0324, suggesting that these additional variables contribute a small amount of explanatory power.

Column (2) shows the impact of including the refusal indicator in the regression. It contributes explanatory power, but does not affect the significance of the key text characteristics, other than also making sentiment score (a measure of the positive or negative sentiment of the response) significant. This significance vanishes though as we include LLM intercepts.

The results do not change if we focus only on the segment of safe prompts, although in Column (8) we do see that the log of average word length, which was marginally significant on the entire dataset becomes no longer statistically significant.

Based on these findings, we decide to incorporate 4 textual variables into our structural model. Namely, the logs of response character length and average word length, which may capture the user's preferences for verbosity and simplicity/complexity in the response. We also include the two distance measures associated with these variables to measure if the users preferences depend on how different the response's characteristics are from the user's prompt characteristics. All 4 of these variables are statistically significant based on Column (4) in the linear regression.

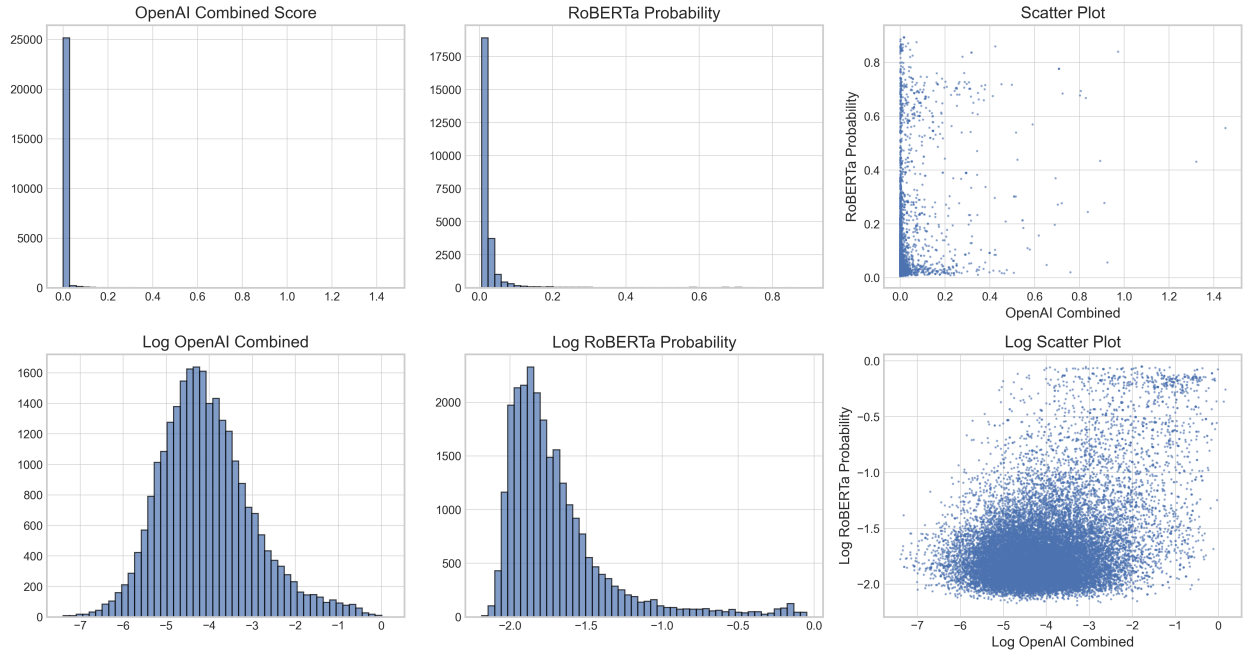
D Moderation Model Sensitivity Analysis

In this section we replicate our main results under alternative moderation models and thresholds.

In our data, we also have OpenAI moderation scores available in addition to RoBERTa moderation scores. Figure D.1 illustrates the raw scores, their log-base-10 transformations, and scatterplots between them. The scores exhibit a correlation of 34.25% and the log-base-10 transformations exhibit a correlation of 37.45%. Even though the scores are positively correlated, they do not perfectly agree on which content needs to be moderated, highlighting the general challenge of moderation systems.

Figures D.2 and D.3 replicate the results in Figures 5 and 6 but using the OpenAI moderation

Figure D.1: Moderation Scores



model and focusing on the top 3.68% of requests by score to mirror our approach with RoBERTa. These figures are also replicated if we shift the moderation threshold to be less conservative. Figures D.4 and D.5 show similar results using the RoBERTa model but focusing on the top 7.50% of scores instead of the top 3.68%.

Tables D.1 and D.2 present estimates from the main structural model under alternative content moderation specifications: RoBERTa thresholds at the top 5% and 7.5% of cases (compared to 3.68% in the baseline), and OpenAI moderation API thresholds at 3.68% and 7.5%. The key parameters remain stable across specifications, confirming robustness to both the choice of moderation system and threshold. As expected, lowering the stringency of the risk definition (moving from 3.68% to 7.5%) expands the share of Class 1 risk-takers, reflecting the inclusion of additional borderline-risky requests. These patterns hold consistently across both RoBERTa and OpenAI moderation approaches, demonstrating that our findings are not artifacts of any particular content filtering methodology.

Figure D.2: Win Rate Across Prompt Risk by Refusal Outcome - OpenAI Moderation

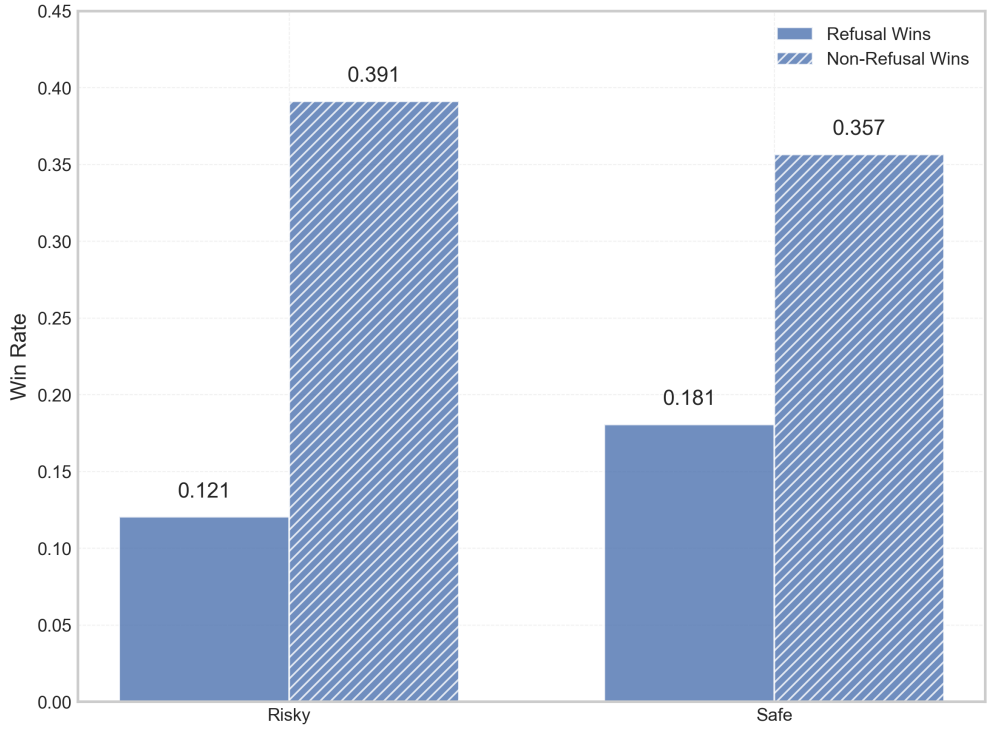
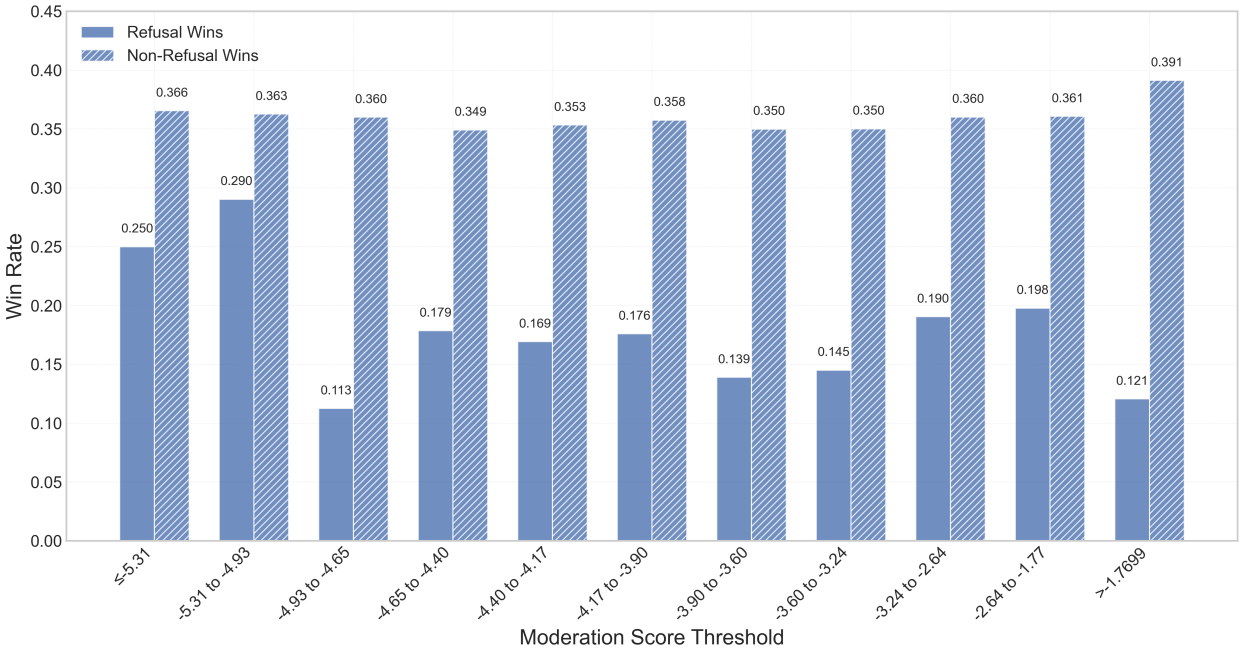


Figure D.3: Win Rate by Thresholds - OpenAI Moderation



Note: Thresholds are deciles of the distribution of the moderation score, except for the final -1.77 threshold which identifies the top 3.68% of prompts that we deemed as risky.

Figure D.4: Win Rate Across Prompt Risk by Refusal Outcome - RoBERTa Moderation (7.50% top scores)

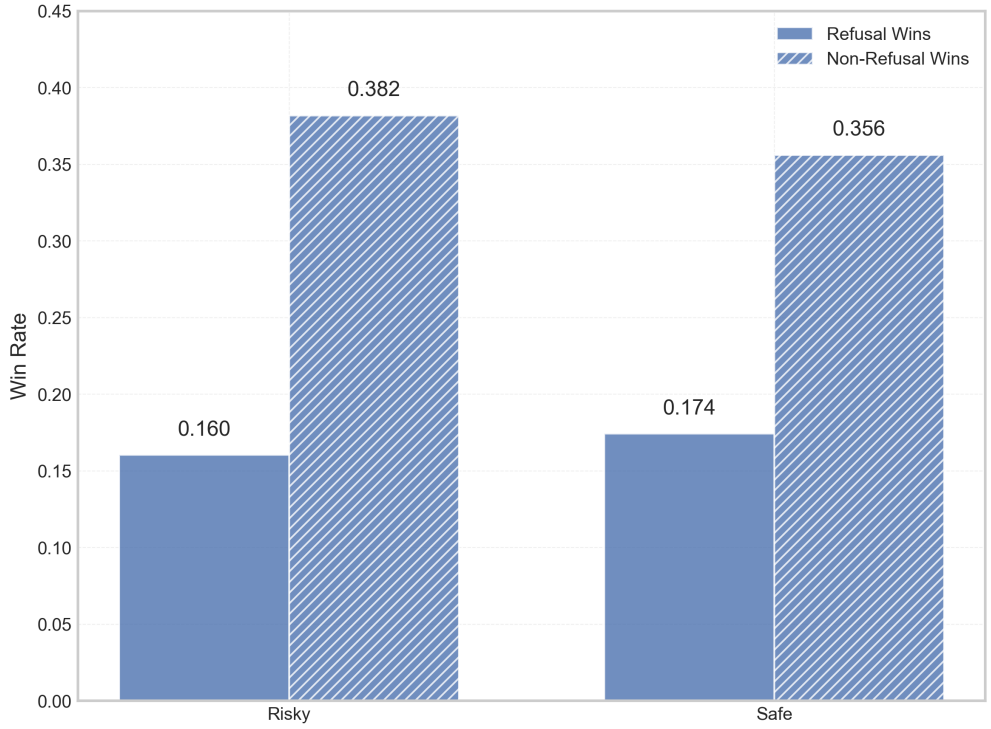
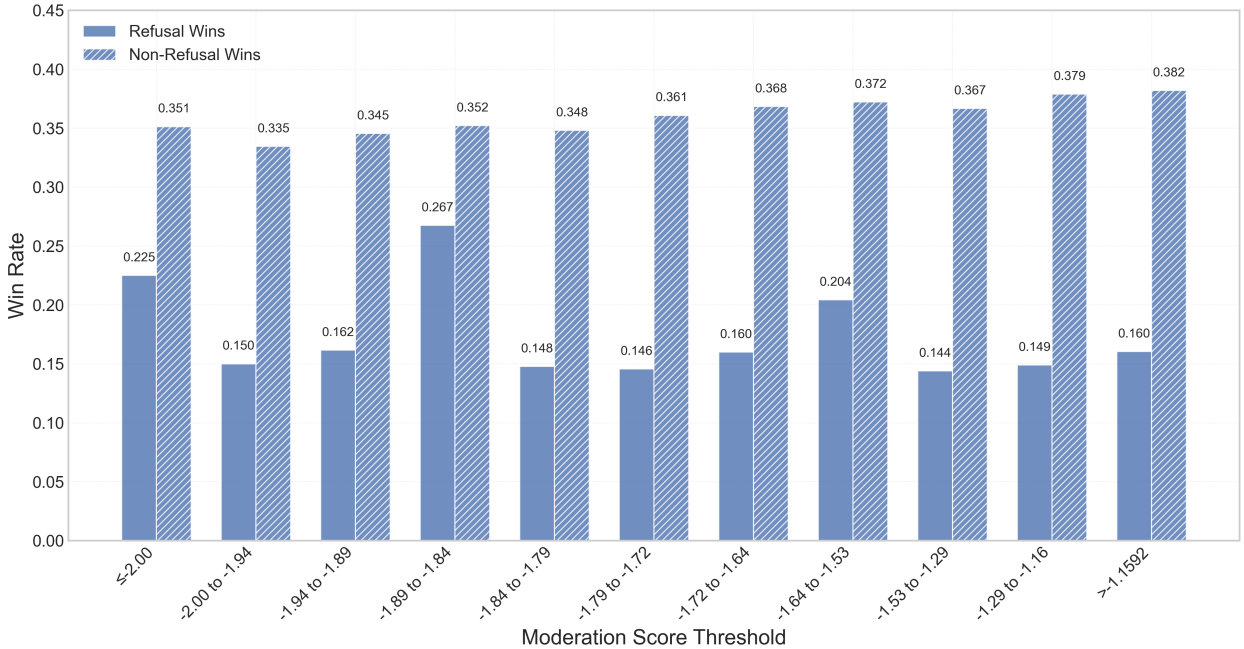


Figure D.5: Win Rate by Thresholds - RoBERTa Moderation ((7.50% top scores)



Note: Thresholds are deciles of the distribution of the moderation score, except for the final -1.16 threshold which identifies the top 7.50% of prompts that we deemed as risky.

Table D.1: Parameter Estimates under Different RoBERTa Moderation Thresholds

Parameters	RoBERTa Moderation (5.00%)			RoBERTa Moderation (7.50%)		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Class probability π_k	0.2959^{***}	0.2326^{***}	0.4716^{***}	0.3767^{***}	0.2509^{***}	0.3724^{***}
	(0.0163)	(0.0229)	(0.0228)	(0.0151)	(0.0246)	(0.0235)
Class-specific parameters						
Refusal coef. (risky) γ_1	-1.4981^{***}	–	–	-1.5054^{***}	–	–
	(0.1792)			(0.1590)		
Refusal coef. (safe) γ_2	-1.1498^{***}	0.6242^{*}	-2.3863^{***}	-0.9446^{***}	0.4247	-2.6338^{***}
	(0.2414)	(0.3611)	(0.2677)	(0.2204)	(0.4200)	(0.3407)
Proprietary × risky β_1	-1.0128^{***}	–	–	-0.7453^{***}	–	–
	(0.1230)			(0.1061)		
Risky intercept α	-1.3318^{***}	–	–	-1.1740^{***}	–	–
	(0.1771)			(0.1477)		
Risk aversion ρ	-2.8488^{***}	-1.6769^{***}	-3.7589^{***}	-1.3447^{***}	-0.3196	-2.3984^{***}
	(0.1804)	(0.1915)	(0.2954)	(0.2929)	(0.3144)	(0.4300)
Text coefficients ψ	(common across classes)					
Character length	0.1069^{***}			0.1581^{***}		
	(0.0193)			(0.0212)		
Average word length	-0.6961^{***}			-0.1401		
	(0.0928)			(0.1237)		
|Character length|	0.2528^{***}			0.2677^{***}		
	(0.0195)			(0.0201)		
|Average word length|	-0.2115^{**}			-0.2757^{***}		
	(0.0934)			(0.0937)		
Model-specific fixed effects δ_j	(common across classes; baseline GPT-4 = 0)					
claude-v1	-0.6189^{***}			-0.4789^{***}		
	(0.0719)			(0.1053)		
claude-instant-v1	-0.7724^{***}			-0.6324^{***}		
	(0.0866)			(0.1183)		
gpt-3.5-turbo	-0.8418^{***}			-0.6618^{***}		
	(0.0690)			(0.1044)		
guanaco-33b	-1.5529^{***}			-1.4129^{***}		
	(0.1113)			(0.1417)		
palm-2	-1.5548^{***}			-1.3948^{***}		
	(0.0771)			(0.1106)		
vicuna-13b	-1.5985^{***}			-1.4385^{***}		
	(0.0690)			(0.1081)		
wizardlm-13b	-1.7654^{***}			-1.6054^{***}		
	(0.0984)			(0.1259)		
vicuna-7b	-1.9893^{***}			-1.8493^{***}		
	(0.0820)			(0.1178)		
koala-13b	-2.0332^{***}			-1.8932^{***}		
	(0.0674)			(0.0989)		
gpt4all-13b-snoozy	-2.3158^{***}			-2.1358^{***}		
	(0.1247)			(0.1549)		
mpt-7b-chat	-2.3879^{***}			-2.2079^{***}		
	(0.0847)			(0.1199)		
alpaca-13b	-2.4080^{***}			-2.1680^{***}		
	(0.0697)			(0.1114)		
RWKV-4-Raven-14B	-2.5051^{***}			-2.3251^{***}		
	(0.0797)			(0.1112)		
oasst-pythia-12b	-2.5754^{***}			-2.3754^{***}		
	(0.0758)			(0.1128)		
fastchat-t5-3b	-2.8774^{***}			-2.6974^{***}		
	(0.0776)			(0.1136)		
chatglm-6b	-2.9998^{***}			-2.8198^{***}		
	(0.0759)			(0.1100)		
dolly-v2-12b	-3.1602^{***}			-2.9202^{***}		
	(0.0595)			(0.1050)		
stablelm-tuned-alpha-7b	-3.1654^{***}			-2.9654^{***}		
	(0.0819)			(0.1162)		
llama-13b	-3.2611^{***}			-2.9811^{***}		
	(0.0769)			(0.1277)		

Note: RoBERTa Moderation thresholds correspond to 5.00% and 7.50% tails of the moderation score distribution. Standard errors in parentheses. $***\ p < 0.01$, $**\ p < 0.05$, $*\ p < 0.10$.

Table D.2: Parameter Estimates under Different OpenAI Moderation Thresholds

Parameters	**OpenAI Moderation (3.68%)**			**OpenAI Moderation (7.50%)**		
	Class 1	**Class 2**	**Class 3**	**Class 1**	**Class 2**	**Class 3**
Class probability π_k	0.2549^{***}	0.2632^{***}	0.4819^{***}	0.3407^{***}	0.2590^{***}	0.4002^{***}
	(0.0283)	(0.0395)	(0.0349)	(0.0112)	(0.0204)	(0.0193)
Class-specific parameters						
Refusal coef. (risky) γ_1	-1.8247^{***}	–	–	-1.6153^{***}	–	–
	(0.2368)			(0.1924)		
Refusal coef. (safe) γ_2	-1.5112^{***}	0.4056	-2.1132^{***}	-1.2557^{***}	0.2584	-2.3855^{***}
	(0.3023)	(0.4687)	(0.2653)	(0.2338)	(0.3986)	(0.2817)
Proprietary × risky β_1	-0.6220^{***}	–	–	-0.3519^{**}	–	–
	(0.1355)			(0.1183)		
Risky intercept α	-1.6416^{***}	–	–	-1.2018^{***}	–	–
	(0.2629)			(0.1171)		
Risk aversion ρ	-2.7553^{***}	-1.6496^{***}	-3.7136^{***}	-2.8351^{***}	-1.8743^{**}	-4.0690^{***}
	(0.2049)	(0.2571)	(0.5157)	(0.3522)	(0.3234)	(0.4541)
Text coefficients ψ	(common across classes)					
Character length	0.1110^{***}			0.1079^{***}		
	(0.0217)			(0.0184)		
Average word length	-0.6640^{***}			-0.7427^{***}		
	(0.0947)			(0.1343)		
|Character length|	0.2523^{***}			0.2507^{***}		
	(0.0216)			(0.0192)		
|Average word length|	-0.2095^{**}			-0.2157^{**}		
	(0.0926)			(0.0977)		
Model-specific fixed effects δ_j	(common across classes; baseline GPT-4 = 0)					
claude-v1	-0.6189^{***}			-0.6189^{**}		
	(0.0638)			(0.1124)		
claude-instant-v1	-0.7724^{***}			-0.7724^{***}		
	(0.0795)			(0.1254)		
gpt-3.5-turbo	-0.8418^{***}			-0.8218^{***}		
	(0.0644)			(0.1170)		
guanaco-33b	-1.5329^{***}			-1.5329^{***}		
	(0.1064)			(0.1555)		
palm-2	-1.5348^{***}			-1.5348^{***}		
	(0.0696)			(0.1262)		
vicuna-13b	-1.5785^{***}			-1.5785^{***}		
	(0.0593)			(0.1225)		
wizardlm-13b	-1.7454^{***}			-1.7454^{***}		
	(0.0962)			(0.1347)		
vicuna-7b	-1.9693^{***}			-1.9693^{***}		
	(0.0740)			(0.1289)		
koala-13b	-2.0132^{***}			-1.9932^{***}		
	(0.0535)			(0.1072)		
gpt4all-13b-snoozy	-2.2958^{***}			-2.2958^{***}		
	(0.1208)			(0.1755)		
mpt-7b-chat	-2.3679^{***}			-2.3679^{***}		
	(0.0789)			(0.1349)		
alpaca-13b	-2.3880^{***}			-2.3680^{***}		
	(0.0686)			(0.1292)		
RWKV-4-Raven-14B	-2.4851^{***}			-2.4651^{***}		
	(0.0709)			(0.1215)		
oasst-pythia-12b	-2.5354^{***}			-2.5354^{***}		
	(0.0663)			(0.1264)		
fastchat-t5-3b	-2.8374^{***}			-2.8174^{***}		
	(0.0681)			(0.1209)		
chatglm-6b	-2.9598^{***}			-2.9598^{***}		
	(0.0664)			(0.1253)		
dolly-v2-12b	-3.1402^{***}			-3.1202^{***}		
	(0.0521)			(0.1247)		
stablelm-tuned-alpha-7b	-3.1454^{***}			-3.1254^{***}		
	(0.0708)			(0.1285)		
llama-13b	-3.2411^{***}			-3.2211^{***}		
	(0.0754)			(0.1409)		

Note: OpenAI Moderation thresholds correspond to 3.68% and 7.50% tails of the moderation score distribution. Standard errors in parentheses. $***\ p < 0.01$, $**\ p < 0.05$, $*\ p < 0.10$.

E Incorporating Ties

When a user selects the "tie" option we consider that as additional data to inform estimation, and include it as one minus the probability of selecting the outside option. It does not result in an additional choice option in the model but rather contributes data to the likelihood.

For user i and prompt t, let $Y_{ijt} \in \{0, 1\}$ indicate the chosen alternative (a, b, or outside option o). If the outcome is a tie (both good), we define its likelihood as $1 - P_{iot|S_{it}}$, corresponding to either a or b being acceptable. The conditional likelihood for prompt t is therefore

$$L_{it}(\theta_i) = P_{it}(S_{it}) \times \begin{cases} P_{ia_t t|S_{it}}, & \text{if model } a_t \text{ chosen,} \\ P_{ib_t t|S_{it}}, & \text{if model } b_t \text{ chosen,} \\ P_{iot|S_{it}}, & \text{if outside option,} \\ 1 - P_{iot|S_{it}}, & \text{if "tie" (both good).} \end{cases} \tag{E.1}$$

Across all choice occasions $t \in \mathcal{T}_i$, the user-level likelihood is

$$L_i(\theta_i, \delta) = \prod_{t \in \mathcal{T}_i} L_{it}(\theta_i, \delta), \tag{E.2}$$

where $\theta_i = \{\alpha_i(S), \beta_i(S), \gamma_i(S), \rho_i, \psi_i\}_{S \in \{0,1\}}$ collects all user- and model-specific parameters.

The rest of the latent class estimation proceeds as in the main model.

Table E.1 summarizes the model estimates after incorporating ties. Overall, the estimates do not change significantly compared to the main model.

F Additional Counterfactuals

In this section, we include details for the additional counterfactual results present in Table 12 in the main paper. These include decompositions of market shares across safe and risky segments for the counterfactual that sets refusal rates to the safe refusal rates rather than to 0 for counterfactuals that eliminate moderation. Table F.1 and F.2 summarize these results. They do not differ meaningfully from the main counterfactuals, as the safe refusal rate is already very close to 0.

Additional counterfactuals include LLM exit simulations in Tables F.3, F.4 and F.5. The

Table E.1: Parameter Estimates: Heterogeneous Model (Including Ties)

Parameters	Heterogeneous Model (Including Ties)		
	Class 1	**Class 2**	**Class 3**
Class probability π_k	0.2172^{***} (0.0138)	0.2578^{***} (0.0214)	0.5250^{***} (0.0240)
Class-specific parameters			
Refusal coef. (risky) γ_1	-1.5883^{***} (0.1834)	–	–
Refusal coef. (safe) γ_2	-1.2190^{***} (0.2199)	0.3585 (0.3573)	-2.2551^{***} (0.2656)
Proprietary × risky β_1	-1.1619^{***} (0.1548)	–	–
Risky intercept α	-1.2907^{***} (0.2008)	–	–
Risk aversion ρ	-2.6569^{***} (0.1944)	-1.5769^{***} (0.2035)	-3.5251^{***} (0.2768)
Text coefficients ψ	(common across classes)		
Character length	0.0968^{***} (0.0203)		
Average word length	-0.4999^{***} (0.0995)		
|Character length|	0.2629^{***} (0.0194)		
|Average word length|	-0.2322^{**} (0.1008)		
Model-specific fixed effects δ_j	(baseline: GPT-4 = 0)		
claude-v1	-0.5989^{***} (0.0608)		
claude-instant-v1	-0.7524^{***} (0.0771)		
gpt-3.5-turbo	-0.8218^{***} (0.0591)		
palm-2	-1.5148^{***} (0.0652)		
guanaco-33b	-1.5329^{***} (0.0994)		
vicuna-13b	-1.5585^{***} (0.0551)		
wizardlm-13b	-1.7454^{***} (0.1054)		
vicuna-7b	-1.9493^{***} (0.0714)		
koala-13b	-1.9932^{***} (0.0491)		
gpt4all-13b-snoozy	-2.2958^{***} (0.1025)		
mpt-7b-chat	-2.3679^{***} (0.0701)		
alpaca-13b	-2.3880^{***} (0.0597)		
RWKV-4-Raven-14B	-2.4851^{***} (0.0612)		
oasst-pythia-12b	-2.5554^{***} (0.0563)		
fastchat-t5-3b	-2.8574^{***} (0.0680)		
chatglm-6b	-2.9598^{***} (0.0676)		
dolly-v2-12b	-3.1402^{***} (0.0366)		
stablelm-tuned-alpha-7b	-3.1454^{***} (0.0673)		
llama-13b	-3.2411^{***} (0.0623)		

Note: Standard errors in parentheses. $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.10$. "–" indicates parameter not applicable for that class type.

Table F.1: Safe Baseline Refusals: Market Shares with Safe Refusal Rates Applied to Risky Prompts

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	22.7	12.7	23.1
2	claude-v1	Prop.	13.7	7.6	13.9
3	claude-instant-v1	Prop.	11.4	6.3	11.6
4	gpt-3.5-turbo	Prop.	9.0	5.0	9.1
5	palm-2	OS	5.4	8.6	5.3
6	guanaco-33b	OS	5.3	8.4	5.2
7	vicuna-13b	OS	5.3	8.4	5.1
8	wizardlm-13b	OS	4.4	7.0	4.3
9	vicuna-7b	OS	3.3	5.3	3.3
10	koala-13b	OS	3.3	5.3	3.2
11	gpt4all-13b-snoozy	OS	2.1	3.3	2.0
12	mpt-7b-chat	OS	2.0	3.1	1.9
13	RWKV-4-Raven-14B	OS	1.9	3.0	1.8
14	oasst-pythia-12b	OS	1.7	2.8	1.7
15	alpaca-13b	OS	1.7	2.7	1.6
16	fastchat-t5-3b	OS	1.4	2.2	1.3
17	chatglm-6b	OS	1.3	2.1	1.3
18	stablelm-tuned-alpha-7b	OS	1.1	1.7	1.0
19	dolly-v2-12b	OS	0.8	1.3	0.8
20	llama-13b	OS	0.7	1.2	0.7
21	outside option	–	1.8	2.4	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	56.8	31.6	57.8	1101
open-source LLMs	41.5	66.0	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	3.9	12.3	0	0
Safe Requests	96.1	87.7	100	100

Table F.2: Safe Baseline Refusals & Equal Quality: Market Shares with Safe Refusal Rates and No Proprietary Penalty

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	23.1	23.2	23.1
2	claude-v1	Prop.	13.9	14.0	13.9
3	claude-instant-v1	Prop.	11.6	11.6	11.6
4	gpt-3.5-turbo	Prop.	9.1	9.2	9.1
5	palm-2	OS	5.3	5.3	5.3
6	guanaco-33b	OS	5.2	5.2	5.2
7	vicuna-13b	OS	5.1	5.2	5.1
8	wizardlm-13b	OS	4.3	4.3	4.3
9	vicuna-7b	OS	3.3	3.3	3.3
10	koala-13b	OS	3.2	3.2	3.2
11	gpt4all-13b-snoozy	OS	2.0	2.0	2.0
12	mpt-7b-chat	OS	1.9	1.9	1.9
13	RWKV-4-Raven-14B	OS	1.8	1.8	1.8
14	oasst-pythia-12b	OS	1.7	1.7	1.7
15	alpaca-13b	OS	1.6	1.6	1.6
16	fastchat-t5-3b	OS	1.3	1.3	1.3
17	chatglm-6b	OS	1.3	1.3	1.3
18	stablelm-tuned-alpha-7b	OS	1.0	1.0	1.0
19	dolly-v2-12b	OS	0.8	0.8	0.8
20	llama-13b	OS	0.7	0.7	0.7
21	outside option	–	1.7	1.5	1.7

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	57.8	58.0	57.8	1128
open-source LLMs	40.5	40.6	40.5	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	5.9	18.7	0	0
Safe Requests	94.1	81.3	100	100

purpose of these simulations is to establish HHI benchmarks to compare the effects of moderation policies to the effects of the exit of LLMs of different quality levels (high, medium, low) from the market.

Table F.3: LLM Exit Analysis: Market Shares After Removing GPT-3.5-Turbo

Rank	**Model**	**Type**	**% Market Share**	**Risky Share**	**Safe Share**
1	gpt-4	Prop.	25.0	13.3	25.4
2	claude-v1	Prop.	15.0	6.6	15.3
3	claude-instant-v1	Prop.	12.5	5.8	12.8
4	palm-2	OS	5.9	8.0	5.9
5	guanaco-33b	OS	5.8	9.4	5.7
6	vicuna-13b	OS	5.8	8.4	5.7
7	wizardlm-13b	OS	4.8	7.5	4.7
8	vicuna-7b	OS	3.7	5.6	3.6
9	koala-13b	OS	3.6	5.7	3.6
10	gpt4all-13b-snoozy	OS	2.3	3.6	2.2
11	mpt-7b-chat	OS	2.2	3.7	2.1
12	RWKV-4-Raven-14B	OS	2.1	3.6	2.0
13	oasst-pythia-12b	OS	1.9	3.3	1.9
14	alpaca-13b	OS	1.9	3.4	1.8
15	fastchat-t5-3b	OS	1.5	2.5	1.5
16	chatglm-6b	OS	1.4	2.1	1.4
17	stablelm-tuned-alpha-7b	OS	1.2	1.9	1.1
18	dolly-v2-12b	OS	0.9	1.5	0.9
19	llama-13b	OS	0.8	1.4	0.8
20	outside option	–	1.9	3.1	1.9
		Summary Statistics			
Market Structure (%)		**Overall**	**Risky**	**Safe**	**HHI**
proprietary LLMs		52.6	25.6	53.5	1235
open-source LLMs		45.5	71.3	44.6	
User Behavior (%)		**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts		–	31.5	25.3	43.2
Risky Requests		3.4	10.8	0	0
Safe Requests		96.6	89.2	100	100

Table F.4: LLM Exit Analysis: Market Shares After Removing GPT4All-13B-Snoozy

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	23.3	13.1	23.6
2	claude-v1	Prop.	14.0	6.5	14.2
3	claude-instant-v1	Prop.	11.6	5.7	11.8
4	gpt-3.5-turbo	Prop.	9.2	4.8	9.3
5	palm-2	OS	5.5	7.9	5.4
6	guanaco-33b	OS	5.4	9.3	5.3
7	vicuna-13b	OS	5.3	8.3	5.2
8	wizardlm-13b	OS	4.4	7.4	4.3
9	vicuna-7b	OS	3.4	5.5	3.3
10	koala-13b	OS	3.4	5.6	3.3
11	mpt-7b-chat	OS	2.0	3.7	2.0
12	RWKV-4-Raven-14B	OS	1.9	3.6	1.9
13	oasst-pythia-12b	OS	1.8	3.3	1.7
14	alpaca-13b	OS	1.7	3.3	1.7
15	fastchat-t5-3b	OS	1.4	2.5	1.4
16	chatglm-6b	OS	1.3	2.0	1.3
17	stablelm-tuned-alpha-7b	OS	1.1	1.8	1.0
18	dolly-v2-12b	OS	0.8	1.5	0.8
19	llama-13b	OS	0.8	1.3	0.7
20	outside option	–	1.8	3.1	1.8

Summary Statistics

Market Structure (%)	Overall	Risky	Safe	HHI
proprietary LLMs	58.0	30.0	59.0	1147
open-source LLMs	40.2	66.9	39.3	
User Behavior (%)	**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts	–	31.5	25.3	43.2
Risky Requests	3.2	10.2	0	0
Safe Requests	96.8	89.8	100	100

Table F.5: LLM Exit Analysis: Market Shares After Removing Llama-13B

Rank	Model	Type	% Market Share	Risky Share	Safe Share
1	gpt-4	Prop.	23.0	12.8	23.3
2	claude-v1	Prop.	13.8	6.3	14.0
3	claude-instant-v1	Prop.	11.5	5.6	11.7
4	gpt-3.5-turbo	Prop.	9.1	4.7	9.2
5	palm-2	OS	5.4	7.8	5.4
6	guanaco-33b	OS	5.3	9.1	5.2
7	vicuna-13b	OS	5.3	8.1	5.2
8	wizardlm-13b	OS	4.4	7.2	4.3
9	vicuna-7b	OS	3.4	5.4	3.3
10	koala-13b	OS	3.3	5.5	3.3
11	gpt4all-13b-snoozy	OS	2.1	3.5	2.0
12	mpt-7b-chat	OS	2.0	3.6	1.9
13	RWKV-4-Raven-14B	OS	1.9	3.5	1.8
14	oasst-pythia-12b	OS	1.8	3.2	1.7
15	alpaca-13b	OS	1.7	3.2	1.6
16	fastchat-t5-3b	OS	1.4	2.4	1.3
17	chatglm-6b	OS	1.3	2.0	1.3
18	stablelm-tuned-alpha-7b	OS	1.1	1.8	1.0
19	dolly-v2-12b	OS	0.8	1.5	0.8
20	outside option	–	1.8	3.0	1.7
Summary Statistics					
Market Structure (%)		**Overall**	**Risky**	**Safe**	**HHI**
proprietary LLMs		57.3	29.4	58.2	1120
open-source LLMs		41.0	67.6	40.1	
User Behavior (%)		**Overall**	**Class 1**	**Class 2**	**Class 3**
Share of Prompts		–	31.5	25.3	43.2
Risky Requests		3.2	10.3	0	0
Safe Requests		96.8	89.7	100	100