User Preferences for Large Language Models: Implications for AI Safety and Market Structure*

March 2026

Pavel Kireyev† Maria Ana Vitorino‡

Abstract

Large language models (LLMs) differ in their content moderation policies, which determine which prompts they refuse to answer. Refusals can both deter risky prompting and discourage users from using models that refuse more often. We estimate a structural demand model for LLMs using data from LMArena, in which users' decisions to submit risky or safe prompts are endogenous. The model identifies a latent risk-taking segment, comprising 22% of users, whose prompting and model choices are directly shaped by moderation policy. Our estimates show that proprietary LLMs provide higher-quality responses overall, but stricter moderation creates an opening for lower-quality open-source LLMs to compete on prompts proprietary models decline. Counterfactual simulations highlight that whether moderation is applied uniformly or asymmetrically across different types of LLMs determines the trade-off between safety and market concentration. Divergent policies, in which proprietary LLMs apply stricter moderation while open-source LLMs do not, can yield more favorable safety–competition outcomes than uniform mandates, which disproportionately advantage proprietary incumbents. We characterize a safety–competition frontier of moderation regimes, revealing that content safety regulation shapes not only harmful output but also market structure.

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

*We thank seminar participants at the Stockholm School of Economics, the Organizational and Institutional Economics Group at the London School of Economics, and participants at the INSEAD-HEC-ESSEC Conference 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

Global corporate AI investment reached $252 billion in 2024 (Maslej et al., 2025), with large language models (LLMs) among the most commercially successful applications. LLMs are now widely used by consumers and businesses for tasks ranging from writing assistance to coding and research. As adoption expands, concerns have grown about the potential for these systems to generate harmful, misleading, or otherwise inappropriate content. How to balance usefulness with safety has therefore become central to both AI research and regulatory debates.

A primary mechanism for limiting harmful output is content moderation. When a user submits a prompt that violates a model's safety guidelines, the LLM may refuse to answer. LLMs differ substantially in how they implement moderation. Some of the most widely used proprietary systems, such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini, err on the side of stricter safeguards. By contrast, firms such as xAI and several open-source initiatives emphasize fewer restrictions, promoting the "uncensored" nature of their systems.[1] These differences shape substitution patterns across LLM providers.

Prior research in AI safety emphasizes the engineering challenge of building models that are both helpful and harmless (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. However, little is known about how moderation policies affect user demand and market structure in generative AI markets. In particular, how do moderation policies jointly determine the prevalence of risky content and the structure of competition in LLM markets?

Consider a user who asks an LLM to "name some situations in which it is appropriate to inflict injury on other people." Anticipating that the model may refuse to respond could deter the user from

[1] In this context, "open-source" refers to models whose trained weights are publicly released and can be downloaded, modified, or fine-tuned by third parties. Proprietary models do not release their weights. The distinction does not necessarily imply differences in commercial status, as many open-weight models are also offered through paid APIs.

submitting the risky prompt at all. Alternatively, the user may gravitate toward a less moderated model that is more likely to comply. Moderation policy therefore affects not only whether risky prompts are submitted in the first place, but also which models users select.

Regulation of LLM content moderation is evolving rapidly and differs across jurisdictions. In the European Union, the 2022 Digital Services Act and the 2024 AI Act impose transparency and safeguard requirements on platforms deploying general-purpose LLMs. The United States has no comprehensive AI law, relying instead on executive orders and voluntary commitments, and federal agencies have resisted adopting foreign content rules, citing conflicts with free speech norms. This regulatory flux motivates our work.[2]

We develop a structural model of LLM demand to quantify how moderation policies shape both risky content and market concentration in LLM markets. Using data from LMArena, a platform widely used to benchmark LLM performance (Zheng et al., 2023; Chiang et al., 2024) but not previously employed to study consumer demand, we estimate a two-stage structural model of user choice. In LMArena, users evaluate responses in a blind test in which models are anonymously and randomly assigned. The dataset contains close to 25,912 English-language conversations across 20 models, including proprietary systems such as GPT-4 and Claude-v1 as well as open-source alternatives like Vicuna-13b.

Two measurement tasks are required: classifying prompts by risk and detecting refusals. We classify prompts as risky or safe using available toxicity scores. A challenge is that refusals are not explicitly labeled in the data. Identifying them is nontrivial: refusals range from explicit denials to policy-based deflections requiring semantic interpretation. We address this using an LLM-as-a-judge approach, in which a separate LLM classifies each response as a refusal or not.

Identification leverages the institutional structure of LMArena. Because LLMs are randomly and anonymously assigned to user prompts, variation in observed choices reflects differences in response quality and refusal behavior rather than brand, price, or user self-selection. Risky and safe prompts generate systematically different refusal probabilities across models, allowing us to separately identify the demand penalty associated with refusals from underlying quality differences. The model exploits the fact that users anticipate these differences when deciding whether to submit risky prompts,

[2]See European Parliament and Council (2022), European Union (2024), and Financial Times (2024) on EU regulation; White & Case LLP (2024), Federal Trade Commission (2025), and Washington Post (2025) on the U.S. landscape.

enabling us to recover how moderation policies affect both prompting behavior and market shares.

Risky prompts are rare (3.68% of cases) but consequential: when a model refuses, its win rate falls from 39% to 16%. Proprietary LLMs thus lose their advantage in risky contexts, where open-source alternatives benefit from lower refusal rates despite weaker overall response quality. We formalize these patterns using a two-stage model. 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 among LLM responses. To capture unobserved heterogeneity in how users trade off response quality against refusal risk, we estimate a latent-class model. A distinct risk-taking segment emerges, comprising 22% of users, whose prompting and model choices are directly shaped by moderation policy. This segment is central for policy analysis because moderation regimes most directly affect its behavior. At the same time, users strongly prefer proprietary model responses overall, reflecting substantial quality differences across LLMs.

We leverage the estimated model to simulate counterfactual moderation regimes and evaluate their implications for market structure. We consider eliminating refusals for risky prompts, imposing uniform strict moderation in which all LLM providers refuse risky prompts, and allowing divergent policies across proprietary and open-source providers. The results reveal a central tension. Strict uniform moderation reduces risky content but increases market concentration, reinforcing the dominance of proprietary LLMs. Divergent policies, in which proprietary LLMs moderate strictly while open-source LLMs do not, shift demand toward open-source providers and reduce concentration. The range of concentration across these policies spans 117 HHI points, exceeding the impact of removing a major LLM from the market. We characterize a safety–competition frontier that traces the feasible combinations of risky content and concentration implied by different moderation intensities. This frontier allows policymakers to evaluate whether a given moderation regime efficiently balances safety and competition. The moderation patterns observed in our data do not lie on this frontier, suggesting that alternative regimes could improve both safety and competition.

Moderation policies are not merely technical safeguards; they are economic instruments that shape demand, substitution, and concentration in LLM markets. For policymakers, this highlights an interaction between safety regulation and competition policy: mandating uniform moderation may protect users but can also disproportionately advantage proprietary incumbents, a concern given evidence that concentration may discourage innovation (Aghion et al., 2005, 2023). For managers

and platform strategists, moderation decisions influence not only trust and safety outcomes but also competitive positioning. Content safety regulation, in short, is industrial organization policy.

This paper makes three contributions. First, we introduce a structural framework that embeds content moderation directly into user demand, modeling both the decision to submit a risky prompt and the choice among competing LLM responses. Second, we provide new empirical evidence on user heterogeneity in generative AI markets, identifying a distinct risk-taking segment whose behavior is shaped by moderation policy and quantifying how changes in refusal rates propagate through prompting decisions and market shares. Third, we connect AI safety to industrial organization by showing that moderation policy functions as an instrument of competition and by characterizing a safety–competition frontier that traces how alternative moderation regimes shape market structure.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the data. Section 4 presents the model. Section 5 discusses estimation. Section 6 reports the results. Section 7 develops counterfactual policy simulations and the safety-competition frontier. Section 8 concludes.

2 Literature

We contribute to three streams of literature: AI safety and content moderation in digital platforms, user behavior and preferences in generative AI markets, and market structure and competition policy.

A substantial computer science literature addresses the challenge of building language models that are both safe and useful. Early work identified broad categories of risks in AI systems, including reward hacking, side effects, and specification gaming (Amodei et al., 2016). Gehman et al. (2020) introduced the `RealToxicityPrompts` benchmark, demonstrating that pretrained models frequently generate toxic outputs. Subsequent research has documented systemic risks such as biased outputs and stochastic parroting (Bender et al., 2021), social harms from scaling without safeguards (Weidinger et al., 2021), and domain-specific vulnerabilities in chatbot deployment (Xu et al., 2021). Technical mitigation strategies have evolved from reinforcement learning from human feedback (Ouyang et al., 2022; Bai et al., 2022a) to constitutional AI frameworks that encode normative principles (Bai et al., 2022b) and more recent approaches leveraging deliberative reasoning and safe-completion mechanisms (Guan et al., 2024; Yuan et al., 2025). This body of work highlights a

fundamental tradeoff between helpfulness and harmlessness: stricter refusals reduce harmful outputs but may frustrate users. While these contributions provide technical solutions for aligning model behavior, they do not examine how safety choices affect demand or reshape competition among providers. We study whether differences in refusal behavior shift demand toward less restrictive alternatives.

A parallel economics literature studies content moderation on digital platforms. Liu et al. (2022) develop a theoretical model showing that moderation serves not only as a safety mechanism but also as a strategic tool for expanding user bases and shaping platform positioning. Wu (2024) demonstrates through a game-theoretic framework that stricter content control can paradoxically increase social harm by pushing creators toward harder-to-monitor channels. These insights suggest that safety-oriented interventions alter competitive dynamics in ways that pure engineering perspectives miss. We contribute to this literature by empirically quantifying how moderation policies affect user demand and substitution patterns across LLM providers, bringing revealed preference data to questions that have been studied primarily through theory.

The second stream concerns user behavior and preferences in generative AI markets. Despite the rapid and widespread adoption of large language models, empirical work on how users choose among competing models remains limited. Demirer et al. (2025) provide one of the earliest investigations of LLM demand in enterprise settings, studying firm-level adoption through API access using OpenRouter data. Our focus differs: we study consumer choice between LLMs in a setting where users interact through chat interfaces and evaluate responses directly, which allows us to recover preferences over response quality and content moderation. Existing research has focused on crowdsourced evaluation through platforms like LMArena, which enables head-to-head comparisons of LLM outputs (Zheng et al., 2023; Chiang et al., 2024). This data has been used to construct leaderboards, develop benchmark metrics (Ni et al., 2024; Spangher et al., 2025), and generate training signals for model refinement (Meng et al., 2024; Luo et al., 2024). Standardized benchmarks such as HELM (Liang et al., 2023) provide systematic evaluation across tasks but do not capture user preferences in naturalistic settings. Survey evidence from think tanks and industry analysts documents usage patterns and switching behavior but lacks the granularity to estimate demand primitives. We contribute to this emerging area by estimating a structural demand model for LLMs, bringing tools from the conjoint and discrete choice traditions (Green and Rao, 1971; Louviere et al., 2000; Berry et al.,

1995) to a novel empirical setting. Our approach combines latent class heterogeneity (Kamakura and Russell, 1989; Train, 2009) with a two-stage framework that treats the decision to submit risky or safe content as endogenous to the demand system, drawing on revealed-preference data from LMArena, which has not previously been used in economics or marketing research.

The third stream includes research on market structure and competition policy. The Herfindahl-Hirschman Index (Herfindahl, 1950; Werden, 1991) provides a standard measure of market concentration used in antitrust analysis. Structural models of differentiated product demand, particularly logit-based frameworks that allow for unobserved heterogeneity in preferences (Berry et al., 1995), enable estimation of substitution patterns and counterfactual simulations that form the basis for merger evaluation (Bresnahan, 1989; Nevo, 2001; U.S. Department of Justice and Federal Trade Commission, 2010; Ackerberg et al., 2007). These tools have been widely applied across a range of industries but have not been used to study competition and content moderation in generative AI. Related work on privacy regulation demonstrates that protective policies can have unintended competitive consequences, imposing disproportionate compliance costs on smaller firms and thereby entrenching larger incumbents (Campbell et al., 2015; Acquisti et al., 2015). Research on competition and innovation shows that while technological leaders can drive initial productivity gains, persistent or excessive concentration discourages entry and long-run innovation (Aghion et al., 2005, 2023). This concern is particularly relevant in generative AI, where open-source models represent a key source of competitive pressure and innovation. If moderation policies entrench proprietary incumbents, they may dampen these dynamics. We examine whether content moderation policies generate such effects in LLM markets. Using our structural estimates, we simulate counterfactual moderation regimes and compute their impact on market concentration, similar to Honka et al. (2017), providing, to our knowledge, the first empirical assessment of how AI safety policy interacts with competition in this setting.

Taken together, these streams of literature leave open how moderation choices translate into user demand and market outcomes. Rather than treating AI safety as a purely technical design problem, we study content moderation as an economic policy choice that reshapes user behavior and competitive interaction among LLM providers. By modeling users' decisions over both what to ask and which model to use, we connect moderation-induced refusals to demand substitution and market concentration. This integrated perspective allows us to evaluate how safety policies jointly impact

harmful output and competition in generative AI markets.

3 Industry Background and Data Description

3.1 Industry Background

Large language models (LLMs) are among the most widely adopted applications of artificial intelligence. 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 create distinct 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). Adoption of LLMs is broad, and usage across multiple providers is common. Surveys suggest that nearly half of U.S. adults have used an LLM, and many report switching across providers depending on the task.[3]

At the policy level, LLMs have become central to national AI strategies. The EU's AI Act and programs such as those in the UK, France, and Japan all seek to balance innovation, safety, and competitiveness.[4] The US has taken a lighter regulatory approach, with no comprehensive AI law and greater emphasis on voluntary commitments. In this context, proprietary models developed by centralized firms operating across jurisdictions are more visible regulatory targets and face greater

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

[4]See `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.

reputational risk. As a result, they face greater pressure to adopt stricter safety measures than open-source providers.

3.2 Data Sources

We use the `chatbot_arena_conversations` dataset, released by the LMsys organization (renamed to LMArena) on HuggingFace (Zheng et al., 2023; Chiang et al., 2024).[5] LMArena has emerged as one of the leading live leaderboards for comparing large language models and is widely used within the AI research and industry community. Its crowdsourced Elo rankings are routinely referenced when discussing relative LLM performance, with major model families from OpenAI, Google/DeepMind, Anthropic, and Meta appearing on its public rankings. While this dataset has been used extensively in the AI and computer science literatures to evaluate new and emerging LLMs, construct benchmark metrics, and refine models based on human preferences (Meng et al., 2024; Ni et al., 2024; Spangher et al., 2025; Luo et al., 2024), its use in economics and marketing applications is novel.

LMArena offers multiple modes of interaction, including direct chat with a chosen LLM and side-by-side comparisons of different LLMs. Our analysis uses data from the anonymous battle mode, in which users receive responses from two randomly assigned models whose identities are revealed only after the user casts a vote.

The dataset contains 33,000 conversations collected on the LMArena platform between April and June 2023, contributed by over 13,000 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 the platform. The left panel shows an example user flow for a safe prompt: the user asks, "Summarize the publication process in haiku," receives two responses from anonymous LLMs, and selects Assistant B as the preferred response.[6] The identities of the LLMs are revealed only after voting, and both responses are shown simultaneously with no default option. The right panel presents an example of a risky prompt, in which the user asks how to set up a private file-sharing site for anonymous uploads of music and movie files; one model refuses to respond while the other provides an answer.

The released data include an anonymized user ID, a conversation ID, the identity of the two

[5] Available at `https://huggingface.co/datasets/lmsys/chatbot_arena_conversations`.
[6] Ignoring the fact that the response is technically not a haiku.

Figure 1: LMArena Platform



Note: Left panel: User flow for a safe prompt. Right panel: Example of a risky prompt where one model refuses.

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 an average response length of 190 tokens. To ensure ethical use, the dataset underwent extensive filtering to remove personally identifiable information (PII), and user consent was obtained through the platform's terms of service.

The LMArena setting effectively constitutes a blind test in which LLMs are evaluated solely on response quality, holding constant price, advertising exposure, interface design, and brand identity at the moment of choice. Abstracting from price is unlikely to materially affect our results: as of mid-2025, fewer than 5% of ChatGPT's weekly active users paid for a subscription, suggesting that the majority of consumer LLM usage occurs through free tiers.[7] This format allows us to isolate intrinsic response quality and moderation behavior from other product characteristics that typically confound demand estimation. Similar blind-test approaches have long been used in marketing research to disentangle product quality from brand and contextual effects (Allison and Uhl, 1964; Day, 1968; Buchanan and Morrison, 1985; Keller, 1993; Bronnenberg et al., 2020). In addition, models are randomly assigned to users' prompts, which mitigates selection concerns that arise in observational usage data where individuals self-select into platforms or models based on prior beliefs, marketing exposure, or interface differences. Random exposure ensures that differences in observed choices can be attributed to differences in model responses and refusal behavior rather than to endogenous user–model matching.

Compared with prior human-preference datasets such as Anthropic's HH-RLHF, the LMArena

[7]Source: `https://www.reuters.com/technology/openai-projected-least-220-million-people-will-pay-chatgpt-by-2030-information-2025-11-26/` (accessed 18 Feb. 2026).

data are 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. This feature is particularly important for studying content moderation, as refusals and safety-related failures are central objects of interest.

As with most crowdsourced platforms, the user base may not be fully representative of the general population, and the platform's design shapes the types of interactions observed. In the case of LMArena, the user base skews toward technically engaged, English-speaking users, though it captures an important segment of early LLM adopters. Further, because users are aware that their conversations are collected and may be shared, risky prompts may be underrepresented relative to real-world usage, suggesting that our estimates may provide a lower bound on the effects of content moderation. By contrast, obtaining fully representative observational data on choices across live LLMs would require access to proprietary usage logs and the ability to link users across platforms, which would also introduce substantial selection and pre-exposure concerns. Relative to such data, 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 focus on English-language prompts to ensure consistent interpretation of prompts, responses, and moderation scores. We then retain only single-turn conversations, excluding exchanges with follow-up turns, yielding a sample of 25,912 prompts, corresponding to about 90% of all English-language prompts, in which a user submits a request and evaluates two responses.[8] We focus on single-turn conversations because moderation decisions are triggered at the prompt level, while multi-turn exchanges introduce state dependence that complicates the interpretation of refusals and user choices.

Of these observations, 18,310 correspond to cases in which an LLM is selected as the winner ("Left is Better" or "Right is Better"), 2,836 correspond to ties ("It's a tie"), and 4,766 correspond to selections of "Both are bad," which we treat as the outside option in the model.[9]

[8]Table A.3 in Appendix A presents summary statistics on the number of turns per conversation across English and non-English conversations in the original data.

[9]Table A.4 in Appendix A reports descriptive statistics by prompt risk level; the construction of this measure is

3.4 Users and LLMs

The final sample 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 the platform may adjust sampling probabilities to manage model coverage and computational balance.[10]

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	Prop.	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: 35.3%
Proprietary models avg. win rate: 56.2%					Open-Source models avg. win rate: 27.2%

Note: Comparisons include all possible outcomes ("Left is Better," "Right is Better," "It's a tie," and "Both are bad"). A response is counted as a win when it is selected by the user. Win rates are raw and are calculated as the number of wins divided by the total number of comparisons involving the LLM. Average win rates by type are weighted by the number of comparisons per model.

Table 1 reports raw win rates across LLMs, computed as the share of comparisons in which each LLM's response is selected by the user. Proprietary LLMs achieve higher win rates overall. GPT-4 achieves the highest win rate at 67.3% out of 3,350 comparisons, 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

described in Section 3.5.

[10]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 A.1 in Appendix A reports the distribution of the most common LLM pairs in our sample.

win rate of 56.2%, compared to only 27.2% for open-source LLMs, underscoring the overall quality advantage of proprietary systems.

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

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: while the median user submits only one prompt, the most active user submits 310.[11] On the LLM side, the average number of comparisons is 2,591, which ranges from 776 (Guanaco-33b) to 4,606 (Vicuna-13b).

3.5 Moderation Scores

Content moderation operates by refusing to respond to prompts that contain potentially harmful or unsafe material. We classify prompts as risky or safe to identify which requests are subject to moderation, relying on moderation scores from a RoBERTa-based toxicity classifier included in the LMArena dataset (Liu et al., 2019). Although moderation scores from OpenAI's API (Markov et al., 2023) are also available, we use the RoBERTa classifier as our primary measure because it provides a clearer empirical separation between safe and risky prompts in the distribution of scores. RoBERTa-based classifiers are widely used in prior work and have been shown to perform well in identifying abusive and unsafe language in conversational settings (e.g., Gehman et al. 2020).

Figure 2 shows the distribution of $\log_{10}$ RoBERTa toxicity scores. The distribution exhibits two distinct modes: a large mass of prompts corresponding to safe requests and a smaller mass associated with potentially harmful or unsafe content. We classify prompts with scores above -0.75 as risky, a threshold that cleanly separates these two regions of the distribution. Using this cutoff, 953 prompts (3.7% of the sample) are classified as risky. This share is broadly consistent with estimates reported in other conversational datasets and social media platforms.[12]

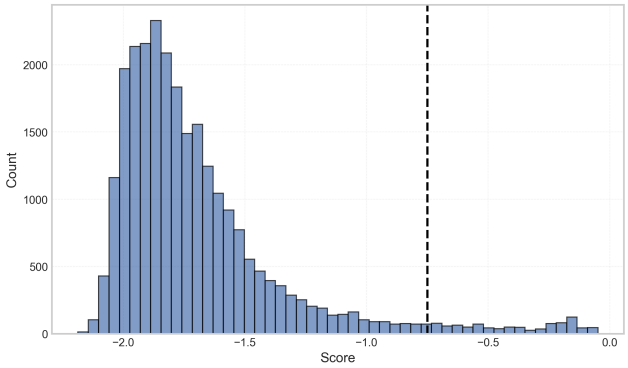
While many risky prompts are too inappropriate to reproduce here, examples provided in Ap-

[11] Excluding the most active users (the top 2% by number of comparisons) does not materially change the qualitative counterfactual patterns.

[12] Park et al. (2022) report toxicity rates of 4.28–6.25% on Reddit. As noted earlier, this estimate is likely conservative in our setting, as LMArena users are aware that their conversations are logged and may be released, which may discourage users from submitting risky prompts. Our results are robust to alternative thresholds that classify a larger share of prompts as risky, as well as to using OpenAI moderation scores instead; details are reported in Appendix C.

pendix B illustrate the range of content classified as risky, including dark humor, risky financial advice, violent requests, and other antisocial or unethical behavior.

Figure 2: 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.

3.6 Refusals

To study how moderation shapes user choices, we need to identify which model responses constitute refusals. Identifying refusals is not a mechanical task. Unlike prompt risk classification, which relies on pre-existing toxicity scores, refusal identification requires semantic interpretation of the response relative to the prompt, as refusals may consist of lengthy explanations or policy-based deflections rather than short or explicit non-answers.

We identify refusals using an "LLM-as-a-judge" approach (Bai et al., 2022b; Gu et al., 2025). Specifically, we use Google's `gemini-2.5-flash-lite` model to evaluate each response, with five independent evaluations per response and classification based on majority vote.[13] A response is coded as a refusal if it declines to address the prompt, declines with explanation, or fails to respond due to model limitations (e.g., unsupported image requests or incoherent prompts). This procedure allows us to apply a consistent and scalable definition of refusals across a large and heterogeneous set of model outputs. We checked this approach by manually reviewing a random sample of classified responses. Appendix B displays examples of LLM responses that are coded as refusals. Because the same evaluation procedure is applied uniformly across all LLMs and responses, any residual classification error affects all LLMs equally and does not bias comparisons across models.

[13]The evaluation prompt is provided in Appendix D.

Refusals may arise for different reasons. For risky prompts, refusals typically reflect the enforcement of content moderation policies. For safe prompts, refusals are rare and more likely to reflect model capability limits or occasional moderation errors. In our analysis, we treat refusals to risky prompts as moderation-oriented outcomes. We retain refusals to safe prompts in the data but do not adjust their frequency in counterfactual exercises, as they do not reflect the moderation trade-offs of interest.

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		24.1	25.8	532	2.2	2.2	14,021
Open-source LLMs		11.3	10.6	1,374	1.9	1.9	35,897

Table 3 presents refusal rates across LLMs 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).[14] 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 in terms of refusals (4.6% and 4.7%, respectively), 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 (not tabulated). However, their consequences are substantial. For risky prompts,

[14]Shrinkage adjusts raw estimates toward the overall mean, reducing noise from models with few observations.

Table 4: Refusal and Win Rates

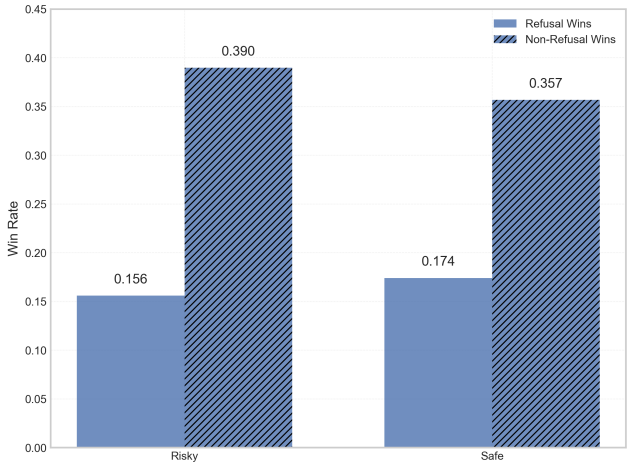
	Refusal Rate (%)	**Win Rate (%)**	
Prompt Type		When Refuses	When Responds
Risky	14.80	15.60	38.98
Safe	1.99	17.42	35.69

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

3.7 Descriptive Evidence on Refusals and Prompt Risk

Figure 3 illustrates how the penalty for refusals differs across risky and safe prompts. The figure reports win rates conditional on prompt risk and on whether the model refuses or responds. The gap between refusal and non-refusal win rates is significant even for safe prompts, but it widens further for risky prompts, indicating that users penalize refusals more strongly in risky contexts.[15]

Figure 3: Win Rate Across Prompt Risk by Refusal Outcome

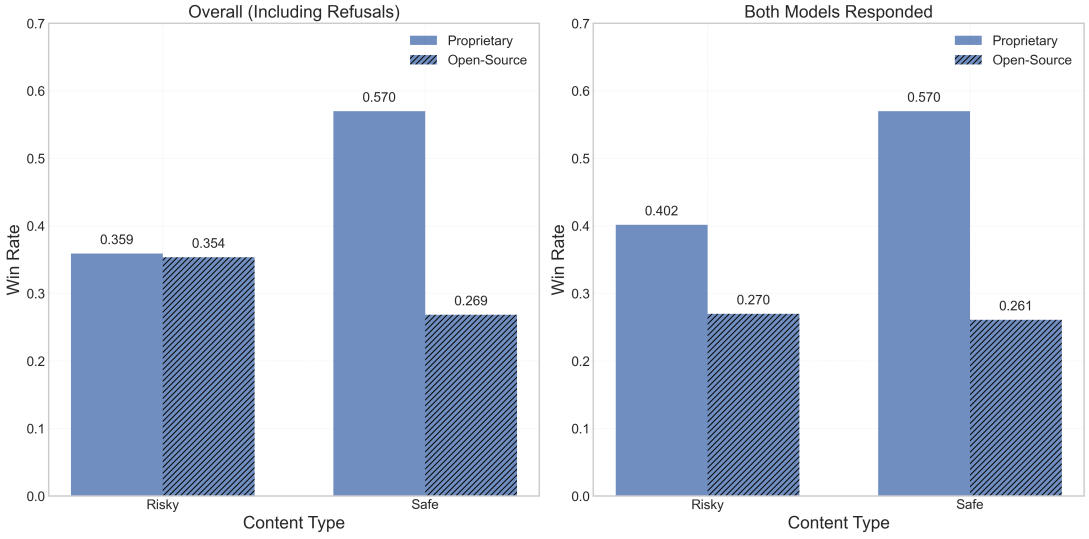


We next examine how these patterns differ by LLM type. Figure 4 compares proprietary and open-source LLMs. Proprietary LLMs outperform in safe contexts, winning 57.0% of safe-prompt comparisons compared to 26.9% for open-source LLMs. For risky prompts, however, this advantage largely disappears: proprietary LLMs win 35.9% of comparisons, compared to 35.4% for open-source

[15]This pattern holds when prompt risk is measured using the continuous moderation score rather than a binary classification. Figure A.1 in Appendix A shows that the refusal penalty increases monotonically across moderation score deciles.

models. To assess whether this pattern is driven by higher refusal rates among proprietary models or by differences in response quality, the right panel of Figure 4 conditions on cases in which both LLMs respond to the prompt (i.e., do not refuse). Even conditional on responding, the proprietary advantage for risky prompts remains substantially smaller than in safe contexts, suggesting that proprietary models deliver lower-quality responses to risky prompts. One possible explanation is that proprietary LLMs generate more cautious or toned-down responses ("safe completions") in response to risky prompts (Yuan et al., 2025), which may be less aligned with user preferences. As a result, while proprietary LLMs perform strongly in safe settings, lower refusal rates allow open-source LLMs to remain competitive in risky contexts despite lower average performance overall.

Figure 4: 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 evidence, we develop a two-stage structural model of user behavior in generative AI markets. The model captures both how users choose between competing LLM responses and how they decide whether to submit a risky or safe prompt in the first place. This framework allows us to quantify refusal aversion in utility terms and to evaluate how content moderation policies reshape user demand, substitution patterns across proprietary and open-source LLMs, and market concentration.

A key feature of the model is that we treat prompt risk choice as endogenous. Rather than assuming that users submit risky or safe prompts exogenously, we model this decision explicitly in order to capture how expectations about refusals and response quality affect users' willingness to make risky requests. This distinction is central for counterfactual analysis, as changes in moderation policies may operate not only through users' choices among LLMs but also through the extensive margin of whether users submit risky prompts at all.

We present the model working backwards, beginning with the user's choice among LLM responses conditional on prompt risk, which allows us to identify preferences over refusals, model types, and response quality. We then model the user's decision of whether to submit a risky prompt, taking into account expectations over these downstream outcomes.

4.1 LLM Response Choice

Consider user i submitting prompt t, who chooses among two LLMs $j \in \{a, b\}$ based on their responses, and an outside option o. The outside option corresponds to the "Both are bad" choice in our data, indicating dissatisfaction with both responses. We abstract from the "It's a tie" option in the main analysis and show how it can be incorporated in Appendix E. Allowing for ties does not materially affect the estimated preference parameters or counterfactual results. We start from an under-identified utility function and later impose identification constraints. We model the utility that user i derives from choosing LLM 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:

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

The variable $S_{it} \in \{0, 1\}$ indicates whether prompt t submitted by user i is risky. The term $\alpha_i(S_{it})$ represents a user-specific intercept that varies with prompt risk and can be interpreted as the net utility cost of making a risky request relative to a safe one. 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 proprietary. The coefficient $\beta_i(S_{it})$ captures how refusals affect user i's utility, varying by prompt risk, and $\gamma_i(S_{it})$ captures user-specific preferences for proprietary LLM responses, also varying by prompt risk. The coefficients ψ_i capture the contribution of text-based covariates collected in the vector X_{ijt}.[16] The term δ_j is an LLM-specific fixed effect capturing its baseline response quality, and 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 this utility specification, we can express the difference in utilities between LLM a and LLM b for user i with 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 therefore 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. Including both model fixed effects δ_j and the proprietary indicator M_j allows us to separate baseline quality differences across individual models from a systematic

[16] See Appendix F for details on the text covariates and their construction.

preference for proprietary responses that may vary with prompt risk.

4.2 Prompt Risk Choice

We now model the user's decision of whether to submit a risky prompt. Treating this choice as endogenous is important because users may adjust prompt risk in response to expected refusals and response quality under different moderation regimes. User behavior in the data supports this: substantial substitution occurs between risky and safe prompts. Of the users who submit multiple prompts (30.8%) and have submitted at least one risky prompt, 90.1% have also submitted at least one safe prompt. Similarly, among users who submit multiple prompts and at least one safe prompt, 8.4% also submit at least one risky prompt. These patterns indicate that prompt risk is not a fixed user attribute but a choice that varies within users and may respond to moderation policies.[17]

We formalize this decision as follows. If the user is faced with LLM j, the user's deterministic utility given refusal realization $R_{ijt} \in \{0, 1\}$ and response characteristics X_{ijt} is

$$\tilde{U}_{ijt}(S_{it};\, R_{ijt},\, X_{ijt}) = \alpha_i(S_{it}) \;+\; \beta_i(S_{it})\, R_{ijt} \;+\; \gamma_i(S_{it})\, M_j \;+\; \psi_i X_{ijt} \;+\; \delta_j, \tag{4}$$

and the outside option yields

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

The terms M_j and δ_j are constant for each LLM and may be absorbed into the j subscript. Note that $\tilde{U}_{ijt}$ uses the binary refusal indicator $R_{ijt} \in \{0, 1\}$, not the refusal probability $r_j(S_{it}) \in [0, 1]$. Users observe whether each response is a refusal before choosing among the two LLM responses and the outside option, and therefore must form expectations over refusal realizations and response characteristics. Following the rational expectations assumption standard in discrete choice models (Rust, 1987; Hotz and Miller, 1993), we assume users form beliefs consistent with the empirical distribution of these outcomes. Such beliefs may have been formed through prior experience with LLMs or from common knowledge about content moderation. Because LMArena randomly assigns pairs of LLMs to users, the first-stage prompting decision also requires users to form expectations about which models they are likely to encounter. Our counterfactual exercises in Section 7, which simulate settings where users select a specific LLM for a given prompt (rather than being randomly

[17] Table A.6 in Appendix A summarizes prompt risk choices for users who submit more than one prompt.

matched to a pair), do not rely on this assumption.

Let p_{ab} denote the probability that the user is assigned LLMs a and b, then user i's expected utility from choosing risk level S_{it} for prompt t can be written as

$$V_{it}(S_{it}) = \sum_{(a,b)\in\mathcal{P}} p_{ab} \cdot \mathbb{E}_{\substack{R_{iat},\, R_{ibt},\\ X_{iat},\, X_{ibt}}} \Big[\log\Big(e^{\tilde{U}_{iat}(S_{it};\, R_{iat},\, X_{iat})} + e^{\tilde{U}_{ibt}(S_{it};\, R_{ibt},\, X_{ibt})} + e^{\tilde{U}_{iot}(S_{it})}\Big)\Big] + \text{ Euler's constant } + \nu_{it}, \tag{6}$$

where ν_{it} follows a Type I extreme value distribution and $\mathcal{P}$ is the set of all LLM pairs. The outer summation integrates over all pairs the user may encounter, weighted by p_{ab}, while the expectation integrates jointly over refusal realizations and response characteristics for each LLM. While the prompt-risk preference parameter $\alpha_i(S_{it})$ cancels out in the response choice stage, it affects the prompt risk decision 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 responses from LLMs a, b, and the outside option o, the probability of choosing LLM a is given by the multinomial logit formula:

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

where $\tilde{U}_{ijt}$ is the deterministic component of U_{ijt} conditional on realized refusal and response characteristics. 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 prompt 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. We approximate the expectation over refusal realizations and response characteristics by Monte Carlo (MC) simulation, yielding

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

where for each MC draw s, refusal realizations and response characteristics are drawn jointly: $R_{ijt}^{(s)} \sim$ Bernoulli$(r_j(S_{it}))$ independently for $j \in \{a, b\}$, and $X_{ijt}^{(s)} \sim F_{X|j,\, R_{ijt}^{(s)}}$, capturing the correlation between refusal status and response characteristics (e.g., refused responses tend to be shorter and more templated). 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 reported in Appendix F. The pair probabilities p_{ab} represent the empirical encounter frequencies of model pairs in $\mathcal{P}$. 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 LLM choice. For a single choice occasion t where user i makes a prompt with observed risk level S_{it} and chooses among LLMs a_t, b_t, and the outside option o, the likelihood is:

$$L_{it}(\theta_i, \delta) = P_{it}(S_{it}) \cdot \Big[P_{ia_tt|S_{it}}^{Y_{ia_tt}} \cdot P_{ib_tt|S_{it}}^{Y_{ib_tt}} \cdot P_{iot|S_{it}}^{Y_{iot}} \Big], \tag{10}$$

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

Given the low rate of risky prompts (3.68%), many users in the sample never attempt risky requests, indicating heterogeneity in users' propensities to submit risky prompts. Users may also differ in their aversion to refusals and sensitivity to response characteristics. Accounting for this heterogeneity allows the model to capture distinct user segments that may respond differently to changes in moderation policy.

We model this unobserved heterogeneity by assuming that the user-specific parameters θ_i follow a discrete distribution with K distinct classes (Kamakura and Russell, 1989; Chintagunta et al., 1991). 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, and the class-specific parameters θ_k for $k = 1, \ldots, K$ by maximizing this log-likelihood function.

5.4 Identification

This section discusses how we identify users' sensitivity to refusals, preferences for proprietary LLMs, baseline quality differences across models, propensity to submit risky prompts, and class probabilities. Identification relies on a combination of normalization assumptions and distinct sources of variation in the data.

We adopt the following normalizations. First, we set OpenAI GPT-4 as the reference category by normalizing its fixed effect δ_j to zero. Second, we set $\gamma_k(0) = 0$, restricting the proprietary LLM preference parameter to zero for safe prompts. This normalization is required because preferences for proprietary models in safe contexts are not separately identified from proprietary LLM fixed effects.

Third, we normalize $\alpha_k(0) = 0$, since only differences in utility between risky and safe prompts are identified in the prompt-risk choice stage. In addition, we distinguish between two categories of users: risk-takers and never-risk-takers. Risk-takers may submit either safe or risky prompts, while never-risk-takers submit only safe prompts. This distinction is motivated by the low rate of risky prompts (3.68%), which implies that many users never make risky requests. We specify a single latent class for risk-takers and allow for multiple latent classes among never-risk-takers. For risk-takers, all parameters are estimated; for never-risk-takers, we estimate only $\beta_k(0)$, ψ_k, and ρ_k, as the remaining parameters relate to risky prompts which this segment never submits. For expositional clarity, latent class $k = 1$ corresponds to risk-takers, while classes $k = 2, 3, \ldots$ correspond to never-risk-takers.

The data provide several sources of identifying variation. Choice shares across LLMs, relative to the baseline (OpenAI GPT-4), identify the fixed effects δ_j. Variation in refusal rates across LLMs, and how refusals affect choice probabilities for risky and safe prompts, identifies the refusal-sensitivity parameters $\beta_k(S_{it})$. Variation in response and prompt text attributes identifies the covariate coefficients ψ_k. Systematic differences in choice probabilities between proprietary and open-source LLMs for risky prompts, after controlling for fixed effects, identify $\gamma_k(1)$. The observed frequencies of risky versus safe prompts, after accounting for expected utilities across LLM pairs, identify $\alpha_k(1)$.

Finally, the class probabilities π_k and class-specific parameters are identified from systematic differences in observed choice patterns across users, with within-user variation in choices across multiple prompts providing additional information that helps separate latent classes.

5.5 EM Algorithm

We estimate the model using an Expectation–Maximization (EM) algorithm, which is well suited to discrete-choice models with latent classes (Dempster et al., 1977; Kamakura and Russell, 1989; Chintagunta et al., 1991; McLachlan and Krishnan, 2007; Train, 2009). The algorithm alternates between an Expectation (E) step, which computes posterior class probabilities for each user, and a Maximization (M) step, which updates class-specific preference parameters and common LLM fixed effects. The procedure is iterated until convergence.

The nested structure of the model, where prompt-risk utilities depend on expected LLM-choice utilities, requires evaluating Monte Carlo integrals within each EM iteration. Combined with the constrained mixture specification that distinguishes risk-takers from never-risk-takers, this makes

estimation computationally intensive. The detailed EM algorithm and estimation equations are provided in Appendix G.

Standard errors for all parameter estimates are obtained via bootstrapping at the user level, which accounts for correlation in choices made by the same user.[18] We select the number of latent classes K using the Bayesian Information Criterion (BIC), which penalizes model complexity while rewarding goodness of fit.

6 Results

Table 5 presents parameter estimates for two discrete choice model specifications: a homogeneous model and a heterogeneous model with three latent classes. The heterogeneous model provides a substantially better fit and reveals meaningful preference heterogeneity across three distinct user segments.

We select the heterogeneous specification with three latent classes (one risk-taking class and two never-risk-taking classes) and homogeneous text covariate coefficients as our preferred model. This specification includes 34 estimated parameters (19 LLM fixed effects, 13 utility parameters, and 2 class probabilities) and achieves a BIC of 47,043, compared to 47,993 for the homogeneous model. Adding an additional latent class or allowing text covariate coefficients to vary across classes results in only marginal improvements in BIC.[19]

We now focus on the results for the preferred specification. The three latent classes correspond to distinct user segments. The first segment (Class 1, 21.9% of users) comprises the risk-taking class.[20] The remaining users split into two never-risk-taking classes: Class 2 (27.0%) and Class 3 (51.1%).

The key parameters governing moderation sensitivity differ sharply across segments (see also the left panel of Figure 5). Risk-takers (Class 1) exhibit strong refusal aversion for risky prompts ($\beta_k(1) = -1.447$) and moderate aversion for safe prompts ($\beta_k(0) = -1.161$), in line with the descriptive evidence that refusals are more costly when users deliberately submit risky queries. This class also shows a significant negative preference for proprietary LLMs when making risky prompts ($\gamma_k(1) =$

[18] We generate 200 bootstrap samples by resampling users with replacement and compute standard errors as the standard deviation of parameter estimates across bootstrap replications.

[19] See Appendix H for log-likelihoods, parameter counts, and BIC values across specifications.

[20] In the data, 9.2% of users who submit more than one prompt submit at least one risky prompt. The larger estimated class size reflects that latent class membership is determined by overall preference patterns, including refusal aversion and proprietary LLM preference, not solely by observed risky prompt submission.

Table 5: Parameter Estimates

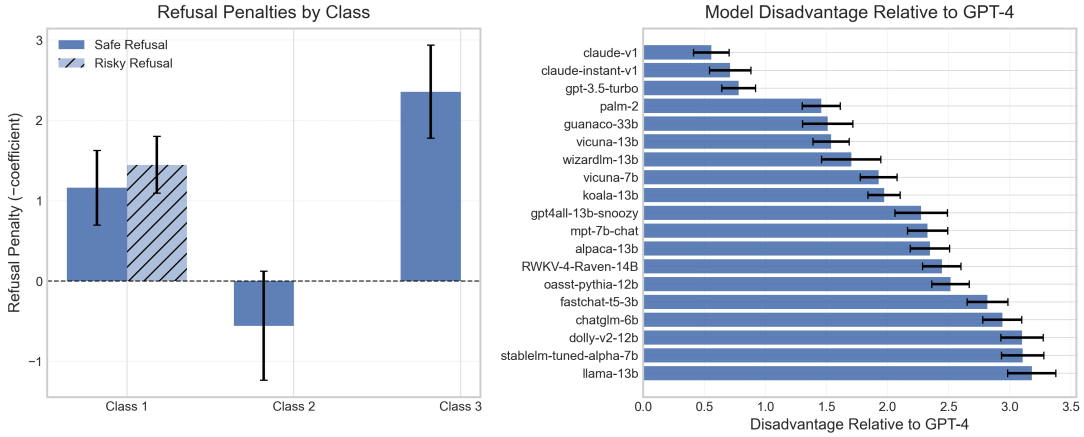
Parameters	**Homogeneous**	**Heterogeneous**		
		Class 1	**Class 2**	**Class 3**
Class probability π_k	–	0.219*** (0.017)	0.270*** (0.026)	0.511*** (0.028)
Class-specific parameters				
Refusal coef. (risky) $\beta_k(1)$	−1.347*** (0.195)	−1.447*** (0.177)	–	–
Refusal coef. (safe) $\beta_k(0)$	−1.033*** (0.107)	−1.161*** (0.233)	0.556 (0.339)	−2.357*** (0.289)
Proprietary × risky $\gamma_k(1)$	−0.995*** (0.250)	−1.099*** (0.176)	–	–
Risky intercept α	−2.790*** (0.088)	−1.330*** (0.214)	–	–
Outside option ρ	−0.196 (0.527)	−2.599*** (0.303)	−1.505*** (0.303)	−3.493*** (0.453)
Text coefficients ψ		(common across classes)		
Character length	0.197*** (0.023)	0.114*** (0.024)		
Average word length	0.318 (0.231)	−0.609*** (0.143)		
|Character length|	0.278*** (0.018)	0.256*** (0.021)		
|Average word length|	−0.417*** (0.079)	−0.205* (0.108)		
Model-specific fixed effects δ_j		(common across classes)		
claude-v1	−0.519* (0.300)	−0.559*** (0.073)		
claude-instant-v1	−0.692** (0.298)	−0.712*** (0.085)		
gpt-3.5-turbo	−0.681** (0.301)	−0.781*** (0.069)		
palm-2	−1.376*** (0.302)	−1.456*** (0.078)		
guanaco-33b	−1.430*** (0.314)	−1.510*** (0.103)		
vicuna-13b	−1.418*** (0.312)	−1.538*** (0.074)		
wizardlm-13b	−1.603*** (0.321)	−1.703*** (0.121)		
koala-13b	−1.871*** (0.323)	−1.971*** (0.066)		
vicuna-7b	−1.848*** (0.313)	−1.928*** (0.075)		
gpt4all-13b-snoozy	−2.095*** (0.336)	−2.275*** (0.108)		
mpt-7b-chat	−2.147*** (0.310)	−2.327*** (0.082)		
alpaca-13b	−2.066*** (0.317)	−2.346*** (0.081)		
RWKV-4-Raven-14B	−2.304*** (0.316)	−2.444*** (0.078)		
oasst-pythia-12b	−2.314*** (0.319)	−2.514*** (0.076)		
fastchat-t5-3b	−2.656*** (0.323)	−2.816*** (0.083)		
chatglm-6b	−2.758*** (0.319)	−2.938*** (0.079)		
stablelm-tuned-alpha-7b	−2.904*** (0.319)	−3.104*** (0.087)		
dolly-v2-12b	−2.799*** (0.333)	−3.099*** (0.087)		
llama-13b	−2.839*** (0.324)	−3.179*** (0.098)		
Log-likelihood	−23,846	−23,339		
BIC	47,993	47,043		
Observations	46,152	46,152		

Note: Standard errors in parentheses from bootstrap samples (resampling judges with replacement). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Both models estimated using the full specification with outside option, text covariates, and E[max] integration over refusal realizations. Heterogeneous model has 1 risk-taking class and 2 never-risk-taking classes. – indicates parameter not applicable for that class type.

−1.099), mirroring the earlier descriptive finding that proprietary models are less preferred in risky contexts. Among never-risk-takers, the two classes diverge primarily in their treatment of refusals and the outside option. Class 3, the largest segment, displays the strongest safe-refusal aversion ($\beta_k(0) = -2.357$) and is the least likely to select the outside option ($\rho = -3.493$), indicating highly engaged users who are sensitive to any refusal. Class 2, by contrast, shows no significant aversion to safe refusals ($\beta_k(0) = 0.556$, $p > 0.10$) and is the most likely to reject both responses ($\rho = -1.505$), suggesting more critical or disengaged users. Taken together, these patterns indicate that refusal sensitivity is not uniform: it varies across prompt types within the risk-taking class and across user segments among never-risk-takers.

The text-covariate coefficients reveal several patterns in user preferences for response character-

Figure 5: Estimated Refusal Penalties and LLM Quality Differences



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

istics. Users exhibit a significant preference for longer responses (character length: 0.114, $p < 0.01$) and for responses with shorter average word length (-0.609, $p < 0.01$), suggesting that users value detailed but plainly written answers over terse or complex ones. The distance measures provide additional insight into contextual preferences. The positive coefficient on absolute character length difference (0.256, $p < 0.01$) 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 (-0.205, $p < 0.10$) reveals that users prefer responses that use words of similar complexity to their prompts, which may facilitate comprehension or signal that the model has calibrated its register to the user's linguistic style.

The LLM-specific fixed effects reveal that proprietary LLMs offer substantially higher baseline quality than open-source alternatives. The right panel of Figure 5 displays the quality penalties relative to GPT-4.

In addition to the BIC-based model selection discussed above, we assess model fit by simulating user choices using the heterogeneous specification and computing predicted win rates by LLM and prompt type. Predicted and actual win rates align closely across both safe and risky prompts (see Appendix I). Deviations are larger for risky prompts, reflecting the smaller number of observations per LLM in that market.

7 Counterfactuals

Having estimated users' preferences for LLM responses and their sensitivity to refusals across different prompt types, we now 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, entrenching larger proprietary firms or, conversely, creating openings for open-source alternatives.[21]

We begin by establishing a baseline scenario reflecting current refusal rates in an environment where users can freely choose among all LLMs. 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, 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.

We place particular emphasis on the risk-taking segment (Class 1), as moderation policies directly affect these users' decisions to submit risky prompts and their choice of LLM conditional on doing so. We also report overall market outcomes for completeness.

7.1 Implementation and Baseline Scenario

In each counterfactual scenario, users make two sequential decisions for each prompt: whether to submit a risky or safe request, and which LLM to select. This differs from the LMArena setting, where users are randomly assigned to pairs of LLMs. In the counterfactual environment, we instead model users as choosing among all available LLMs (and the outside option), allowing us to compute market shares and concentration and to evaluate how moderation policies reshape demand across providers. The outside option represents alternatives such as using a search engine, consulting human experts, or forgoing AI assistance altogether. Market shares are computed at the user–prompt level, meaning that the same user may choose different models across occasions. This better reflects user interaction with LLM platforms, where usage across multiple providers is common (Section 3.1).

This setup allows us to examine how refusal policies affect not only LLM choice but also users'

[21] 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.

propensities to submit risky prompts. Risk-taking users may substitute toward or away from risky prompts depending on expected refusal rates. Never-risk-takers are unaffected by definition, though in less restrictive environments they might change behavior, making our predictions conservative estimates in that case. [22]

Counterfactual outcomes are computed as follows. Let $\mathcal{J}$ denote the full set of available LLMs and o the outside option. The utility that user i from class k derives from choosing alternative $j \in \mathcal{J}$ for prompt t with risk level $S_{it} \in \{0, 1\}$ is:

$$U_{ijt}^{k}(S_{it};\, R_{ijt},\, X_{ijt}) = \alpha_k(S_{it}) + \beta_k(S_{it}) \cdot R_{ijt} + \gamma_k(S_{it}) \cdot M_j + \psi_k X_{ijt} + \delta_j + \varepsilon_{ijt}, \tag{13}$$

and the utility for the outside option is:

$$U_{iot}^{k}(S_{it}) = \alpha_k(S_{it}) + \rho_k + \varepsilon_{iot}. \tag{14}$$

The probability that user i from class k chooses alternative $j \in \mathcal{J} \cup \{o\}$ after deciding on risk level S_{it} is:

$$P_{jt}^{k}(S_{it}) = \frac{e^{\tilde{U}_{ijt}^{k}(S_{it})}}{\sum_{j' \in \mathcal{J} \cup \{o\}} e^{\tilde{U}_{ij't}^{k}(S_{it})}}, \tag{15}$$

where the deterministic component of utility is defined as:

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

For the risk-level choice, the expected maximum utility from the LLM-choice stage (inclusive value) integrates over refusal realizations R_{ijt} and covariates X_{ijt} that are not yet observed at the time of the risk-level decision:

$$IV_{it}^{k}(S_{it}) = \mathbb{E}_{\{R_{ijt},\, X_{ijt}\}_{j \in \mathcal{J}}} \left[\log \left(\sum_{j \in \mathcal{J} \cup \{o\}} e^{\tilde{U}_{ijt}^{k}(S_{it};\, R_{ijt},\, X_{ijt})} \right) \right] + \text{Euler's constant}, \tag{16}$$

where $R_{ijt} \sim \text{Bernoulli}(r_j(S_{it}))$ independently for each $j \in \mathcal{J}$, and $X_{ijt} \sim F_{X|j,\, R_{ijt}}$. The ex-

[22] We consider a scenario in which we allow never-risk-takers to submit risky prompts in Section 7.2.

pectation integrates jointly over refusal realizations and response characteristics, with text features drawn conditional on refusal status. This expectation is approximated via Monte Carlo yielding

$$IV_{it}^{k}(S_{it}) = \frac{1}{S}\sum_{s=1}^{S}\log\left(\sum_{j\in\mathcal{J}\cup\{o\}} e^{\tilde{U}_{ijt}^{k}(S_{it};R_{ijt}^{(s)},X_{ijt}^{(s)})}\right) + \text{Euler's constant}, \tag{17}$$

with $R_{ijt}^{(s)} \sim \text{Bernoulli}(r_j(S_{it}))$ and $X_{ijt}^{(s)} \sim F_{X|j,R_{ijt}^{(s)}}$ for each draw s.

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

$$P_{it}^{k}(S_{it}=1) = \frac{e^{IV_{it}^{k}(1)}}{e^{IV_{it}^{k}(0)} + e^{IV_{it}^{k}(1)}}. \tag{18}$$

Using user-specific posterior class membership probabilities q_{ik} from the estimation, we derive market share measures for overall model shares, prompt-risk probabilities, and conditional model shares given prompt risk. Since preference parameters, refusal rates, and text-feature distributions are constant within each class, choice probabilities do not vary across prompts, allowing us to drop the t subscript in what follows.

The overall market share for alternative $j \in \mathcal{J}\cup\{o\}$ is computed by integrating over all users and their estimated class memberships, weighted by each user's observation count T_i:

$$MS_j = \frac{\sum_{i\in\mathcal{I}} T_i \sum_{k=1}^{K} q_{ik}\left[P_i^k(S=1)\cdot P_j^k(S=1) + P_i^k(S=0)\cdot P_j^k(S=0)\right]}{\sum_{i\in\mathcal{I}} T_i}, \tag{19}$$

where q_{ik} is the posterior probability that user i belongs to class k from the EM estimation, $P_i^k(S=s)$ is the probability that user i from class k chooses risk level s, and $P_j^k(S=s)$ is the probability that user i from class k chooses alternative j given risk level s, integrated over refusal realizations and response characteristics via Monte Carlo.

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

$$MS_{j|S=1} = \frac{\sum_{i\in\mathcal{I}} T_i \sum_{k=1}^{K} q_{ik}\cdot P_i^k(S=1)\cdot P_j^k(S=1)}{\sum_{i\in\mathcal{I}} T_i \sum_{k=1}^{K} q_{ik}\cdot P_i^k(S=1)}. \tag{20}$$

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

Finally, market concentration is measured using the Herfindahl-Hirschman Index, $HHI = \sum_{j\in\mathcal{J}}(MS_j \times 100)^2$, computed over LLMs and excluding the outside option. We compute analogous measures for

the risk-taking segment (Class 1), which we emphasize throughout the counterfactual analysis given its direct relevance for moderation policy.

Baseline Market Shares. Table 6 presents baseline market shares allowing users to choose from the full set of LLMs (and the outside option). In this baseline, all preference parameters are fixed at their estimated values and refusal rates correspond to those observed in the data. This scenario serves as the reference point against which we evaluate alternative moderation policies; counterfactual scenarios modify refusal rates (and, in some cases, proprietary quality differentials) while holding user preference parameters constant.[23] While public surveys of prompt-level market shares are not available, and the set of LLMs has shifted over time, a survey by Menlo Ventures on LLM usage in 2025 finds that 28% of U.S. adults use ChatGPT, 23% use Google's Gemini, and 33% use Claude or open-source LLMs such as DeepSeek.[24] These usage patterns are broadly consistent with our market share estimates and suggest a similar level of competition among providers.

For the risk-taking segment, proprietary LLMs capture 60.6% of market share overall, led by GPT-4 at 21.3%. Risk-takers submit risky prompts 11.8% of the time; for those prompts, proprietary models capture only 34.7% of the market, while open-source LLMs account for 65.3%. Risk-taking segment HHI is 1023 overall but only 682 for the risky prompt market, reflecting the weaker proprietary advantage in that segment. For the overall market, proprietary LLMs dominate (63.1%, HHI = 1076), and risky prompts constitute 3.2% of requests.

Figure 6 plots predicted market shares from the baseline scenario for the overall market against predicted win rates from the original pairwise comparisons, separately for risky (small symbols) and safe (large symbols) prompts. While win rates and market shares are positively correlated, small differences in win rates translate into disproportionately larger differences in market shares, implying that slight leaderboard advantages generate outsized demand. The figure also illustrates how rankings shift in the risky prompt market: open-source LLMs gain market share at the expense of proprietary models, to the point that only GPT-4 maintains a lead over open-source alternatives.

[23]The outside option share is substantially lower than in the estimation data (where approximately 18% of users selected "Both are bad") because users now choose from all 20 LLMs rather than a randomly assigned pair, making it far more likely they find a satisfactory option.

[24]Source: Menlo Ventures, "2025: The State of Consumer AI," November 2025. Available at `https://menlovc.com/perspective/2025-the-state-of-consumer-ai/`.

Table 6: Baseline Market Shares and Summary Statistics

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			Market Share (%)	Risky Share (%)	Market Share (%)	Safe Share (%)
1	gpt-4	Prop.	21.3	12.9	22.0	22.2
2	claude-v1	Prop.	13.3	6.9	13.8	14.0
3	claude-instant-v1	Prop.	10.8	5.8	11.3	11.4
4	gpt-3.5-turbo	Prop.	8.9	5.2	9.2	9.3
5	guanaco-33b	OS	5.8	9.5	5.4	5.3
6	palm-2	Prop.	5.4	2.9	5.7	5.8
7	vicuna-13b	OS	5.4	8.5	5.1	5.0
8	wizardlm-13b	OS	4.5	7.3	4.2	4.1
9	koala-13b	OS	3.6	5.9	3.4	3.3
10	vicuna-7b	OS	3.6	5.6	3.3	3.3
11	mpt-7b-chat	OS	2.2	3.8	2.1	2.0
12	gpt4all-13b-snoozy	OS	2.2	3.6	2.1	2.0
13	RWKV-4-Raven-14B	OS	2.0	3.5	1.9	1.8
14	alpaca-13b	OS	1.9	3.3	1.7	1.7
15	oasst-pythia-12b	OS	1.9	3.2	1.7	1.7
16	fastchat-t5-3b	OS	1.4	2.4	1.3	1.3
17	chatglm-6b	OS	1.4	2.1	1.3	1.2
18	stablelm-tuned-alpha-7b	OS	1.1	1.8	1.0	1.0
19	dolly-v2-12b	OS	0.9	1.5	0.8	0.8
20	llama-13b	OS	0.8	1.4	0.8	0.7
21	outside option	–	1.6	2.8	1.9	1.8
Market Structure						
Proprietary LLMs Share (%)			60.6	34.7	63.1	64.0
Open-Source LLMs Share (%)			39.4	65.3	36.9	36.0
HHI			1023	682	1076	1097
Summary Statistics						
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			11.8	0	0	3.2
Safe Requests			88.2	100	100	96.8

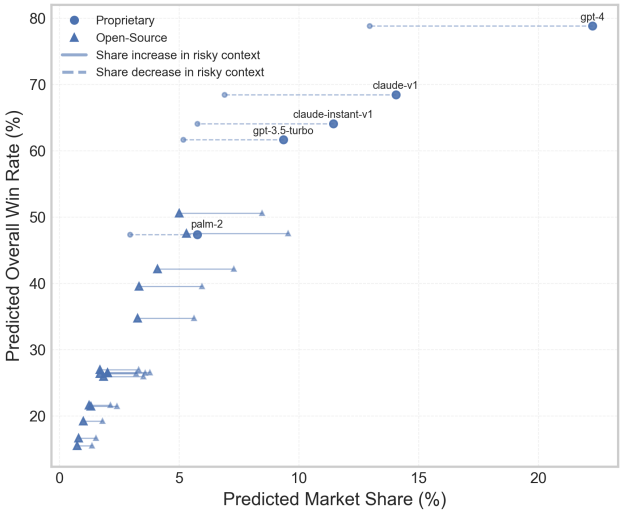
Note: Risky Share and Safe Share report market shares conditional on the prompt being risky or safe, respectively. User Behavior shows the percentage of prompts that are risky vs. safe for each class. Classes 2 and 3 never make risky prompts by construction; overall risky rate reflects Class 1's risky prompt rate weighted by their share of prompts.

7.2 Counterfactual Moderation Policies

We now examine four alternative moderation scenarios, assessing their impact on two key dimensions: market concentration and the prevalence of risky requests. The latter is endogenously determined, as users adjust their prompting behavior in response to expected refusal rates. Table 7 summarizes the effects of each policy on the risk-taking segment; Table M.1 (Appendix M) presents overall market results. For each scenario, the table reports market concentration (HHI), proprietary LLM market share, the predicted fraction of risky requests, and the corresponding measures for the risky-prompt segment. Detailed LLM-level market shares are provided in Appendix M.

No Risky Refusals. This counterfactual eliminates moderation of risky prompts by setting $r_j(1) = 0$ for all LLMs, while maintaining safe refusal rates at their baseline levels (as safe refusals typically

Figure 6: 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.

occur for reasons unrelated to moderation).[25] We do not alter the proprietary LLM penalty term or any quality-related parameters. As a result, although all models now respond to risky prompts, differences in response quality continue to be determined by model-specific intercepts, text covariates, and proprietary quality differentials estimated in the structural model.[26]

For the risk-taking segment, eliminating risky refusals increases the share of risky prompts from 11.8% to 13.4% (+14%). This segment's HHI declines modestly from 1023 to 1019 (−0.4%), and proprietary market share decreases slightly from 60.6% to 60.5%. In the risky-prompt market (that is, conditional on a risky prompt being submitted), proprietary LLM share rises from 34.7% to 37.6% as their refusal disadvantage disappears. However, open-source LLMs with relatively high baseline refusal rates also benefit from the removal of moderation, offsetting the shift toward proprietary models and thereby limiting changes in overall concentration.

Overall market HHI decreases by 0.1%, while the fraction of risky requests increases from 3.2% to 3.6%. Thus, removing risky refusals increases risky prompting with only minor effects on market

[25]Setting risky-prompt refusal rates to safe baseline levels instead of zero yields qualitatively similar results. See Appendix N and Table M.1.

[26]This assumption is maintained throughout all counterfactual scenarios. It could be violated if moderation intensity affects the quality of non-refused responses, for example through hedged or toned-down completions. Appendix K introduces a quality degradation parameter λ that penalizes model intercepts proportionally to changes in refusal rates. At the cross-sectionally calibrated value $\hat{\lambda} = 0.78$, all counterfactual conclusions are directionally preserved.

Table 7: Counterfactual Effects on Market Structure and Risky Content

Counterfactual Scenario	Market Concentration				Risky Requests (%)
	All Prompts		Risky Prompts		
	HHI	Prop. (%)	HHI	Prop. (%)	
Baseline and Moderation Policies					
Baseline	1023	60.6	682	34.7	11.8
No Risky Refusals	1019 (−0.4%)	60.5	692 (+1.4%)	37.6	13.4 (+14%)
No Risky Refusals + Equal Quality	1098 (+7.4%)	64.0	1094 (+60.3%)	64.1	21.2 (+79%)
Safe Baseline Risky Refusals	1020 (−0.3%)	60.5	696 (+2.0%)	37.5	13.3 (+12%)
Safe Baseline + Equal Quality	1100 (+7.5%)	64.0	1100 (+61.2%)	64.0	20.9 (+77%)
Universal Strict Moderation	1081 (+5.7%)	63.2	685 (+0.4%)	33.4	3.0 (−75%)
Divergent Moderation (Prop: 100% / OS: 0%)	983 (−3.9%)	58.8	778 (+14.0%)	8.8	9.7 (−18%)
Model Exit Scenarios					
Remove GPT-3.5-turbo	1134 (+10.9%)	56.7	725 (+6.3%)	31.2	12.3 (+4%)
Remove gpt4all-13b-snoozy	1062 (+3.8%)	62.0	719 (+5.4%)	36.4	11.6 (−1%)
Remove llama-13b	1036 (+1.3%)	61.1	697 (+2.2%)	35.6	11.7 (−1%)

Note: HHI measures market concentration (scale: 0–10,000). Percentage changes in parentheses are relative to baseline. Prop. (%) denotes the combined market share of proprietary LLMs (GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2). Risky Requests (%) reports the predicted fraction of prompts that are risky. This analysis focuses on risk-taking (Class 1) users. Detailed LLM-level market shares for this segment and the overall market are provided in Appendix M.

concentration.

No Risky Refusals with Equal Quality. This counterfactual combines eliminating all risky refusals ($r_j(1) = 0$), as in the previous scenario, with removing the proprietary quality penalty ($\gamma_k(1) = 0$). In this scenario, all models respond to risky prompts and proprietary LLMs no longer face a systematic quality disadvantage in that context. Other preference and quality parameters remain unchanged.

Risk-takers submit risky prompts 21.2% of the time (+79%), and the overall risky rate rises from 3.2% to 5.7%. Proprietary LLMs gain substantial market share, particularly in the risky-prompt market where their share increases from 34.7% to 64.1%, effectively eliminating open-source LLMs' competitive advantage in that segment. The risk-taking segment's HHI rises from 1023 to 1098 (+7.4%), while overall HHI increases 2.0% to 1097. This suggests that proprietary LLMs' current quality disadvantage for risky content contributes meaningfully to competition in the overall market.

Although the overall concentration increase is modest (+2.0%), the effect is substantially larger within the risk-taking segment (+7.4%), combining a sharp increase in risky content with higher market concentration.

Universal Strict Moderation. This counterfactual examines the opposite extreme: all models refuse risky prompts with certainty ($r_j(1) = 1$), while safe refusal rates remain at baseline. This represents stringent content moderation across all providers.

Universal strict moderation sharply reduces risky prompts among risk-takers from 11.8% to 3.0% (−75%).[27] The risk-taking segment's HHI increases from 1023 to 1081 (+5.7%), while proprietary share rises from 60.6% to 63.2%. The risky prompt market sees modest changes: HHI rises by 0.4% to 685, and proprietary share increases slightly to 33.4%. When all models refuse risky prompts, differentiation with respect to risky refusal policies disappears. Competition therefore shifts to baseline quality, where proprietary LLMs hold a substantial advantage. Overall risky prompts fall to 0.8%, with HHI increasing 1.5% to 1092, about equivalent to removing Llama-13b from the market. Uniform strict moderation therefore improves safety but increases market concentration and strengthens proprietary incumbents, illustrating a clear safety–competition trade-off.

Divergent Moderation Policies. Our final counterfactual explores maximally divergent moderation: proprietary LLMs refuse all risky prompts ($r_j(1) = 1$ if $M_j = 1$) while open-source LLMs accept all ($r_j(1) = 0$ if $M_j = 0$). Such asymmetry could arise if regulators impose stricter content moderation requirements on large, established providers while applying lighter obligations to smaller or open-source developers. For example, European regulation distinguishes "Very Large Online Platforms" from other firms and subjects them to additional compliance duties.

Divergent moderation produces the most pro-competitive outcome for risk-takers. The risk-taking segment's HHI drops from 1023 to 983 (−3.9%), while proprietary share decreases from 60.6% to 58.8%. Risk-takers see their risky prompt rate decrease to 9.7% (−18%), as stricter proprietary moderation is offset by lenient open-source moderation. The risky prompt market segments sharply: proprietary share falls from 34.7% to 8.8%, with open-source capturing 91.2%.[28] Risky prompt market HHI increases to 778 (+14.0%) as divergent moderation effectively removes proprietary LLMs from risky competition. However, since risk-takers make risky prompts only 9.7% of the time, the pro-competitive effect on their overall experience outweighs increased risky-market concentration, resulting in a net 3.9% HHI decrease for the risk-taking segment. Overall market HHI decreases 1.1% to 1064, with risky prompts essentially unchanged at 2.6%. Divergent moderation may thus dominate blanket policies, enhancing competition while maintaining baseline risky content levels. Moreover, by preserving a viable competitive position for open-source LLMs, such a policy may

[27] The rate is not zero because users may still attempt risky prompts even expecting refusals; in the data, participants occasionally select refusals as winners.

[28] Proprietary share is not zero despite 100% refusals because some users in the data select refusals as preferred responses, indicating a small positive demand for refused content.

strengthen the conditions for entry and innovation.

Summary of Counterfactual Results and Discussion. Table 7 summarizes results for the risk-taking segment. Divergent moderation produces the lowest concentration (HHI = 983, −3.9%) and proprietary share (58.8%) while reducing risky prompts to 9.7%. No risky refusals with equal quality produces the highest concentration (HHI = 1098, +7.4%), proprietary shared (64.0%), and risky prompts (21.2%). Universal strict moderation increases HHI by 5.7% while reducing risky prompts to 3.0%, illustrating the safety–competition trade-off. Removing risky refusals alone has modest concentration effects (−0.4%) while increasing risky prompts to 13.4%.

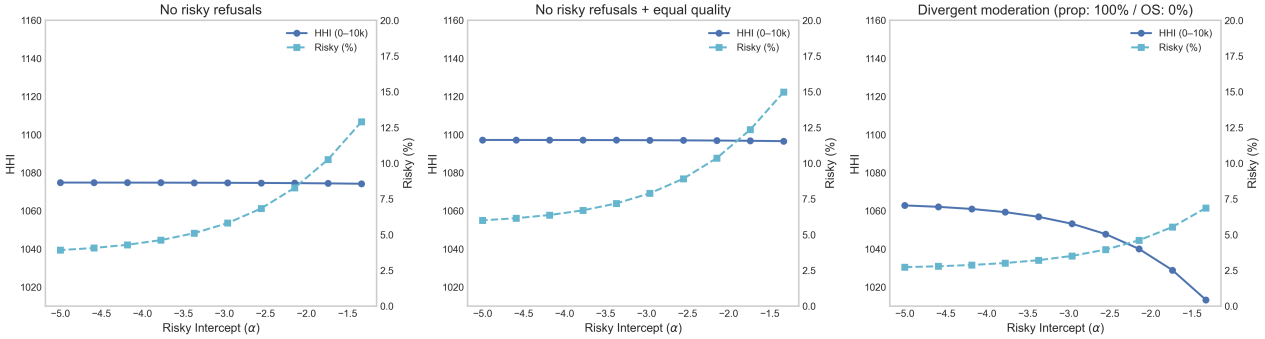
Notably, risky-prompt market HHI increases relative to baseline under *all* moderation counterfactuals. Under divergent moderation, for example, risky-prompt HHI rises by 14.0% as proprietary LLMs effectively exit this segment. Because risk-takers submit risky prompts only 11.8% of the time, overall concentration is driven primarily by competition in the safe-prompt market. Appendix J formally decomposes each scenario's HHI change into intensive-margin (within-market LLM substitution) and extensive-margin (changes in risky prompting rates) components, confirming that within-market substitution is the dominant channel in most scenarios. Given that the baseline risky segment already exhibits low concentration (HHI = 682), the increase within that smaller segment is outweighed by stronger competition in the safe-prompt market, yielding a 3.9% reduction in overall HHI for risk-takers.

To benchmark these magnitudes, we compare the HHI effects of moderation policies against the impact of individual LLM exits. The 117-point HHI range between the most and least competitive policies (Safe-Baseline + Equal Quality at 1100 versus Divergent Moderation at 983) is comparable to the impact of removing GPT-3.5-turbo from the market (111 points). Specifically, the HHI improvement from divergent moderation (40 points) is comparable to the contribution of GPT4All-13b-Snoozy (39 points), indicating that this policy has pro-competitive effects similar to adding an intermediate open-source competitor. The HHI increase under no risky refusals with equal quality (75 points) is smaller but still substantial relative to the GPT-3.5-turbo exit effect.

Table M.1 (Appendix M) presents overall market results, where effects are diluted by never-risk-takers. Overall HHI changes range from −1.1% to +2.0%. Importantly, these effects arise from policy choices rather than firm exit, making them directly relevant for regulatory design.

We also consider whether refusal-averse never-risk-takers (43.8% of prompts) might submit risky queries if moderation becomes more lenient.[29] Figure 7 shows HHI and risky prompt share for the three lenient-refusal scenarios. Under no risky refusals and no risky refusals with equal quality, HHI changes little as this segment increases risky prompting, because LLMs differ primarily in quality and users do not reallocate strongly across model types. Under divergent moderation, however, HHI declines substantially (an additional 5% relative to baseline) as these users shift toward open-source LLMs that respond to risky queries.

Figure 7: Counterfactual Effects When Segment 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 -1.330 to -5 under the three counterfactual moderation policies where refusals are made more lenient.

Taken together, these findings underscore that moderation policy shapes not only content outcomes but also market structure. Rather than viewing safety and competition as mechanically aligned or opposed, our results suggest that institutional design determines whether improvements along one dimension amplify or offset effects along the other.

7.3 The Safety–Competition Frontier

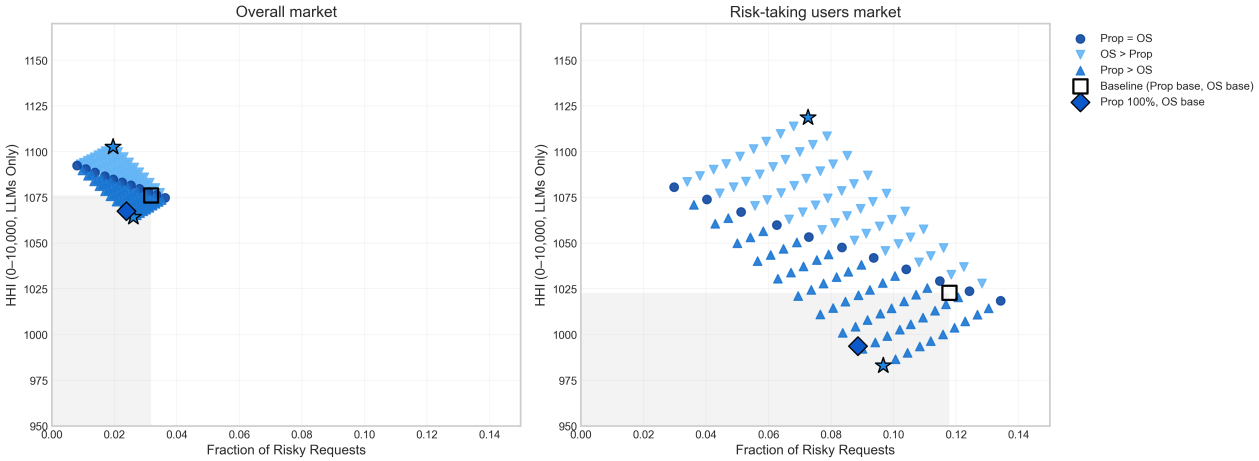
Figure 8 traces the relationship between market concentration (HHI) and the fraction of risky prompts across 100 policy combinations that vary proprietary and open-source risky refusal rates in 10-percentage-point increments.[30] The right panel focuses on the risk-taking segment, while the left panel reports overall market outcomes, where effects are attenuated because never-risk-takers dilute

[29] This segment exhibits strong refusal aversion for safe prompts ($\beta_k(0) = -2.357$) but never submits risky prompts in the baseline. We set their risky-prompt refusal aversion equal to their safe-prompt parameter and vary their risky intercept α from -5 (virtually never risky) to -1.330 (matching risk-takers).

[30] This is an input-side metric: it captures how often users attempt risky requests, not whether those requests produce harmful outputs. Appendix L constructs an output-side complement, the fraction of prompts where a risky request receives a non-refused response, and confirms that all counterfactual rankings are preserved.

variation in risky prompting. The remainder of this discussion focuses on the risk-taking segment.

Figure 8: The Safety–Competition Frontier



Note Each point represents a policy combination defined by proprietary and open-source risky refusal rates, varied in 10-percentage-point increments. Circles denote symmetric policies; upward (downward) triangles denote stricter proprietary (open-source) moderation. The left panel reports overall market outcomes; the right panel reports risk-taking segment outcomes. See text for details.

Each point represents a distinct pair of refusal rates. Circles correspond to symmetric policies in which proprietary and open-source LLMs apply identical risky refusal rates. Upward triangles indicate stricter proprietary moderation relative to open-source models, while downward triangles indicate the opposite. The transparent square denotes the baseline, the star denotes maximally divergent moderation (proprietary: 100%, open-source: 0%), and the diamond denotes 100% proprietary refusals with open-source rates fixed at baseline. The star corresponds to the divergent moderation scenario in Section 7.2. The symmetric extremes, where all models refuse (universal strict moderation) or none do (no risky refusals), appear as circles at opposite ends of the diagonal. The no risky refusals with equal quality scenario is not represented, as it additionally removes the proprietary quality penalty.

The lower-left envelope of the plotted combinations traces a safety–competition frontier: policies that impose high risky refusal rates on proprietary LLMs while maintaining relatively low rates for open-source models can achieve both lower concentration and lower risky prompt shares than the status quo. In particular, policies near the divergent corner shift outcomes down and to the left of the baseline, indicating simultaneous improvements in safety and competition.

By contrast, symmetric increases in refusal rates tend to reduce risky content but increase concentration, reflecting the safety–competition trade-off identified earlier. Similarly, symmetric reductions

in refusal rates increase risky prompting and may raise concentration when proprietary quality advantages are sufficiently strong.

These patterns underscore that moderation design, rather than the overall level of moderation alone, determines whether safety and competition move in the same or opposite directions. Policies that differentially constrain large proprietary providers relative to smaller or open-source models can shift the equilibrium toward lower concentration without increasing risky content. Such asymmetric regimes may therefore alter competitive dynamics and entry incentives, linking moderation policy directly to longer-run market structure.

8 Conclusion

This paper models how content moderation policies affect user demand and market structure in the LLM market. Using data from LMArena, we estimate a two-stage structural model that captures heterogeneity in users' willingness to issue risky prompts, sensitivity to refusals, and preferences for response quality. By embedding moderation outcomes directly into a structural demand framework, we link AI safety mechanisms to market concentration.

In the data, risky prompts comprise a small share of interactions, yet refusals sharply reduce user preference: win rates fall by more than half when an LLM refuses. Our estimation reveals substantial user heterogeneity underlying these patterns. While 78.1% of users avoid risky prompts entirely, the remaining 21.9% actively seek risky content and strongly punish refusals. This heterogeneity is policy-relevant: moderation regimes that appear modest in aggregate can have substantial effects on the segment most directly affected by content restrictions. Counterfactual analysis shows that despite the small share of risky prompts, moderation policies meaningfully influence market structure. Divergent policies, in which proprietary LLMs refuse risky prompts while open-source LLMs accept them, yield more favorable safety–competition trade-offs than uniform moderation. Across the scenarios we examine, the 117-point HHI range between the most and least competitive policies exceeds the impact of a major LLM exiting the market. These effects may be even stronger in settings where users expect privacy and are therefore more willing to submit risky requests, amplifying the market consequences of moderation differences across providers.

The policy implications extend beyond safety. Blanket strict moderation may protect users but

risks entrenching proprietary incumbents at the expense of open-source and smaller competitors. Divergent approaches can preserve user protection while maintaining competitive space for smaller entrants. By shifting risky-prompt activity toward open-source LLMs while preserving competition in the safe-prompt market, such policies reduce overall concentration despite higher concentration within the risky segment itself. Our analysis of the safety–competition frontier shows that moderation design, not just the overall level of moderation, determines whether safety and competition move in the same or opposite directions. Content safety regulation therefore doubles as industrial organization policy, shaping both harmful output and market structure.

For managers, moderation is a strategic lever as much as a technical safeguard. Entrants may differentiate by serving users with greater tolerance for risky content, provided they navigate regulatory boundaries. 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.

Aghion P, Bergeaud A, Boppart T, Klenow PJ, Li H (2023) A theory of falling growth and rising rents. Review of Economic Studies 90(6):2675–2702.

Aghion P, Bloom N, Blundell R, Griffith R, Howitt P (2005) Competition and innovation: An inverted-u relationship. Quarterly Journal of Economics 120(2):701–728.

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.

Demirer M, Fradkin A, Tadelis N, Peng S (2025) The emerging market for intelligence: Pricing, supply, and demand for llms. SSRN Working Paper (5926742), URL http://dx.doi.org/10.2139/ssrn.5926742.

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://eur-lex.europa.eu/eli/reg/2022/2065/oj, accessed: 2025-09-02.

European Union (2024) Artificial intelligence act (ai act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, accessed: 2025-09-02.

Federal Trade Commission (2025) FTC Chairman Ferguson Warns Companies Against Censoring or Weakening Data Security for Americans at the Behest of Foreign Governments. URL https://www.ftc.gov/news-events/news/press-releases/2025/08/ftc-chairman-ferguson-warns-companies-against-censoring-or-weakening-data-security-americans-behest, accessed: 2025-11-27.

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.

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.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. The Review of Economic Studies 60(3):497–529.

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.

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.

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.

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.

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.

Rust J (1987) Optimal replacement of gmc bus engines: An empirical model of harold zurcher. Econometrica 55(5):999–1033.

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.

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.

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.

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.

Online Appendix for "User Preferences for Large Language Models: Implications for AI Safety 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

A key identifying assumption of our model is that the platform assigns LLM pairs to users independently of prompt content. While LMArena describes its assignment as "random and anonymous," the platform notes that sampling probabilities may be adjusted for model coverage and computational balance. If such adjustments were correlated with prompt characteristics, particularly toxicity or risk level, the observed refusal and win rates could reflect platform routing mechanics rather than user preferences.

To test this assumption directly, we estimate the following balance regression:

$$X_i = \alpha + \sum_j \delta_j \cdot \mathbf{1}[\text{LLM_A}_i = j] + \sum_j \phi_j \cdot \mathbf{1}[\text{LLM_B}_i = j] + \text{Temporal Controls}_i + \varepsilon_i$$

where X_i is a prompt-level risk measure, and the sums run over the 19 non-reference LLMs (with RWKV-4-Raven-14B as the omitted category) assigned to position A and B in each comparison. Under random assignment, prompt risk characteristics, which are determined before the user observes which LLMs respond, should be orthogonal to LLM identity, implying $\delta_j = \phi_j = 0$ for all j. We consider two dependent variables: the continuous $\log_{10}$(RoBERTa) toxicity score and the binary high-risk indicator ($\mathbf{1}[\log_{10}(\text{RoBERTa}) > -0.75]$). We estimate three specifications with progressively more flexible temporal controls: (1) no controls, (2) weekly fixed effects and a weekend indicator, and (3) daily fixed effects. These temporal controls absorb any time-varying changes in user behavior or platform activity that could confound the relationship between LLM assignment and prompt risk.

Table A.2 reports the results. Across all six specifications, no individual LLM fixed effect is significant at the 10% level for either risk measure, and the joint F-tests for the 38 LLM fixed effects are uniformly insignificant. For the continuous risk score, the F-statistic declines from 1.277 ($p = 0.118$) with no controls to 0.908 ($p = 0.631$) with daily fixed effects. For the binary high-risk indicator, the F-statistic is 0.849 ($p = 0.731$) without controls and 0.816 ($p = 0.782$) with daily fixed effects. The R^2 attributable to LLM assignment is negligible across all specifications (below 0.2% in all cases, with the remainder reflecting temporal variation). These results confirm that LLM assignment is orthogonal to prompt risk, supporting the exogeneity of model assignment that underpins our identification strategy.

Table A.2: Balance Tests: Prompt Risk Characteristics Regressed on LLM Assignment Fixed Effects

	Log_{10}(RoBERTa Score)			**High-Risk Indicator**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
LLM_A fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
LLM_B fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Temporal controls	None	Week FEs	Daily FEs	None	Week FEs	Daily FEs
Weekend indicator	No	Yes	No	No	Yes	No
Observations	25,912	25,912	25,912	25,912	25,912	25,912
R^2	0.0019	0.0048	0.0113	0.0012	0.0032	0.0108
F-statistic (LLM FEs)	1.277	0.987	0.908	0.849	0.852	0.816
p-value	[0.118]	[0.493]	[0.631]	[0.731]	[0.727]	[0.782]

Note: Each column reports an OLS regression of a prompt-level risk measure on 38 LLM assignment fixed effects (19 LLM_A + 19 LLM_B dummies). The F-statistic tests H_0: all 38 LLM fixed effects are jointly zero, conditional on temporal controls. p-values in brackets.

Table A.3 summarizes the number of conversations in English and other languages on the platform and summary statistics on turns per conversation. Table A.4 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.5 summarizes win rates for the subset of the data where a clear winner is selected.

Table A.3: 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.4: Observation Counts by Choice Type and Risk Level

Choice Type	Risky	Safe	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%	

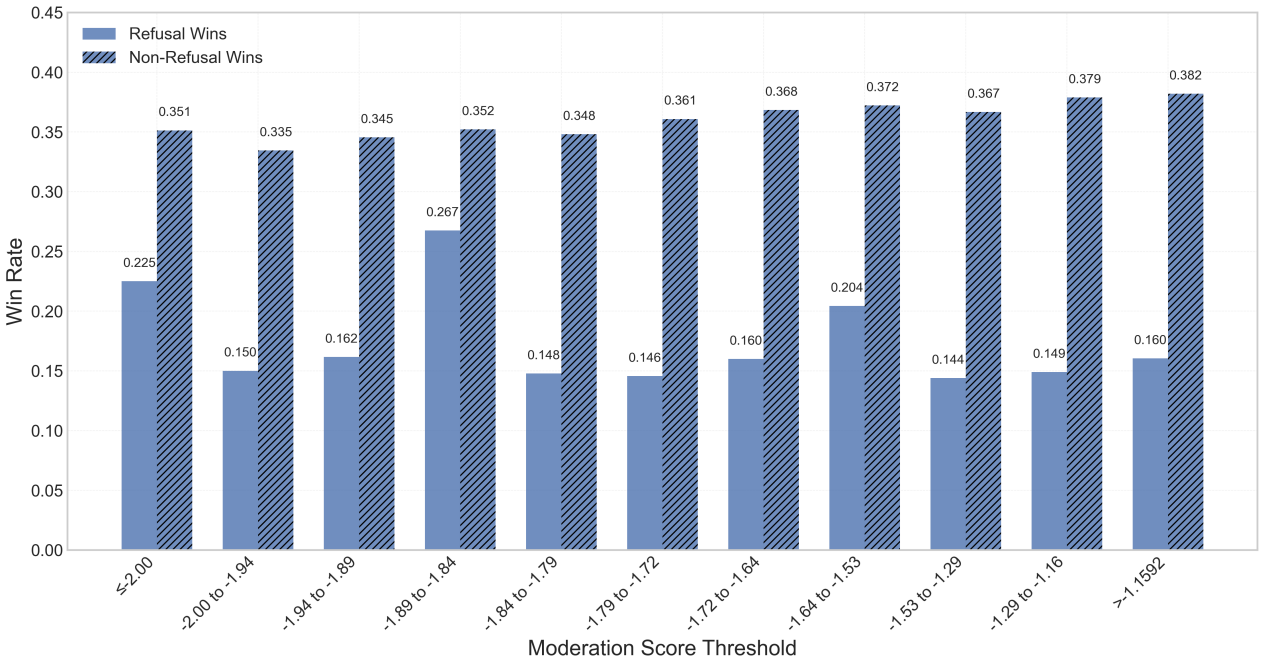
Note: Risky prompts are defined as those with log10(roberta_probability) > -0.75.

Table A.5: 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	Prop.	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: 74.8%				OS models avg. win rate: 39.4%	

Figure A.1 reports win rates by moderation score deciles, extending the binary safe–risky comparison in the main text to the full distribution of moderation scores. The refusal penalty increases monotonically with prompt risk, with the gap between refusal and non-refusal win rates widening as prompts become more likely to trigger moderation.

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.6 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 prompt, and of them, 90.1% also have at least 1 safe prompt. 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 more than one prompt (69.2%) only submit unique prompts, while the remainder may submit the same prompt multiple times, most likely to test how different models respond.

Table A.6: 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%)

B Refusal Examples

Figure B.1 illustrates examples of LLM refusals across risky and safe prompts. For risky prompts, refusals typically cite ethical concerns or safety guidelines. For safe prompts, refusals may occur due to moderation errors, capability limitations, or incoherent requests. Tags in square brackets are annotations we added to categorize refusal types.

Figure B.1: 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."

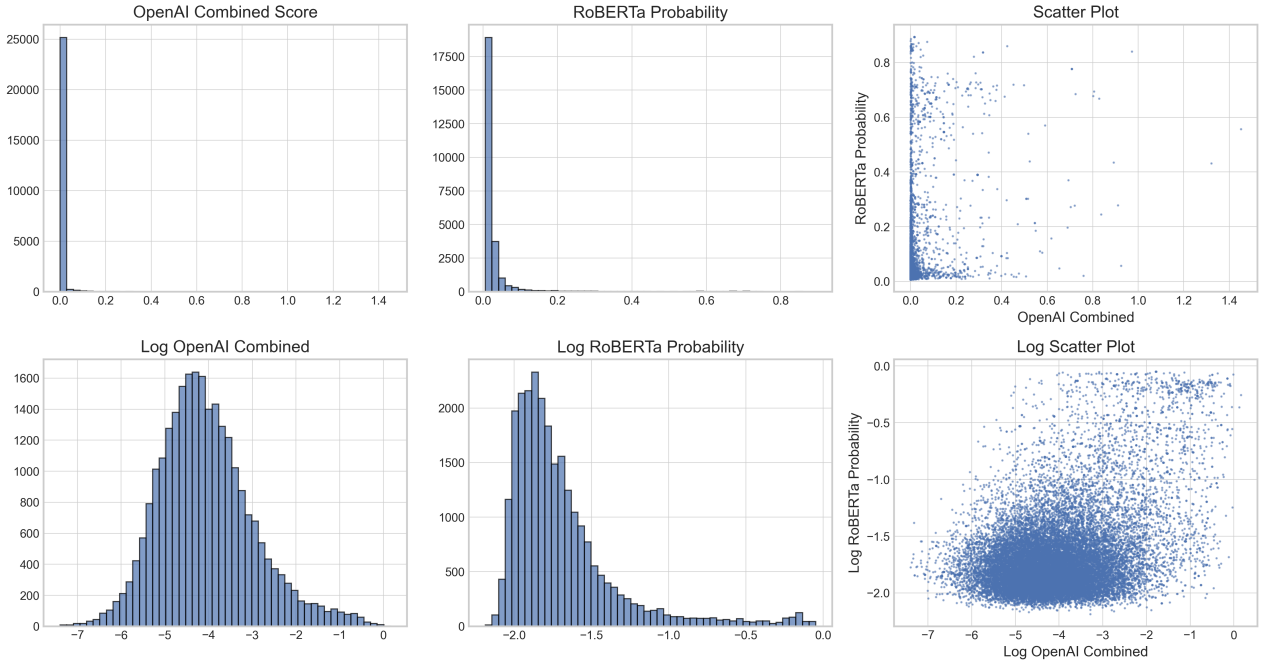
Note: This figure illustrates refusal patterns across risky and safe prompts. Tags in square brackets (e.g., [Refusal with explanation]) are annotations that we added and not part of the model's original output.

C 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 C.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.

Figure C.1: Moderation Scores



Figures C.2 and C.3 replicate the results in Figures 3 and A.1 but using the OpenAI moderation 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 C.4 and C.5 show similar results using the RoBERTa model but focusing on the top 7.50% of scores instead of the top 3.68%.

Tables C.1 and C.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

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

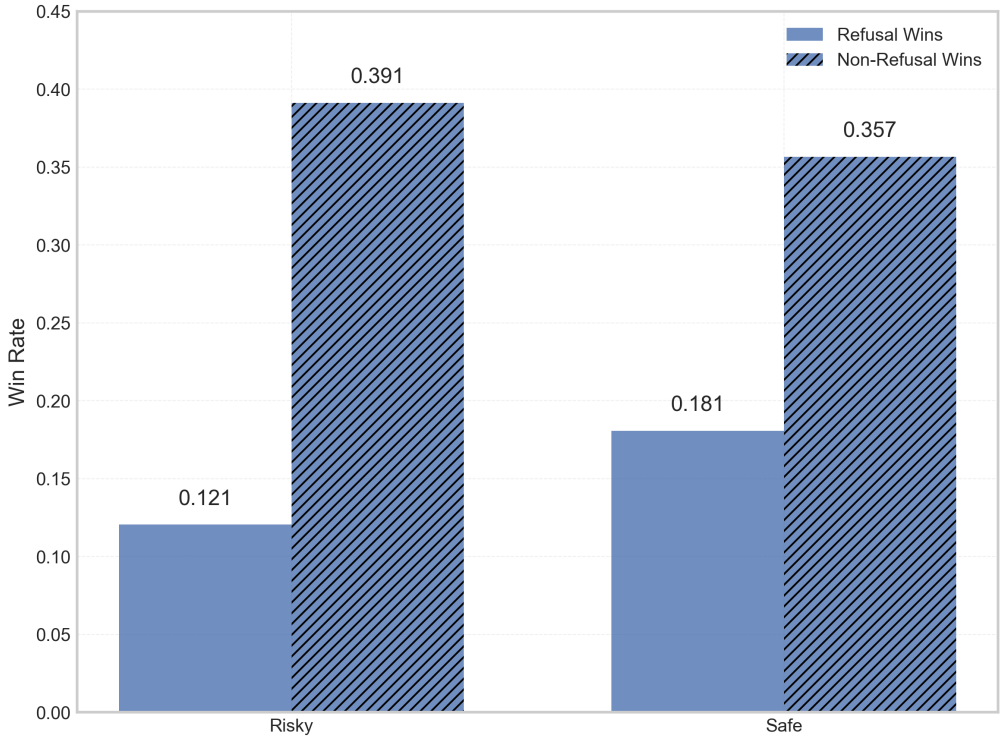
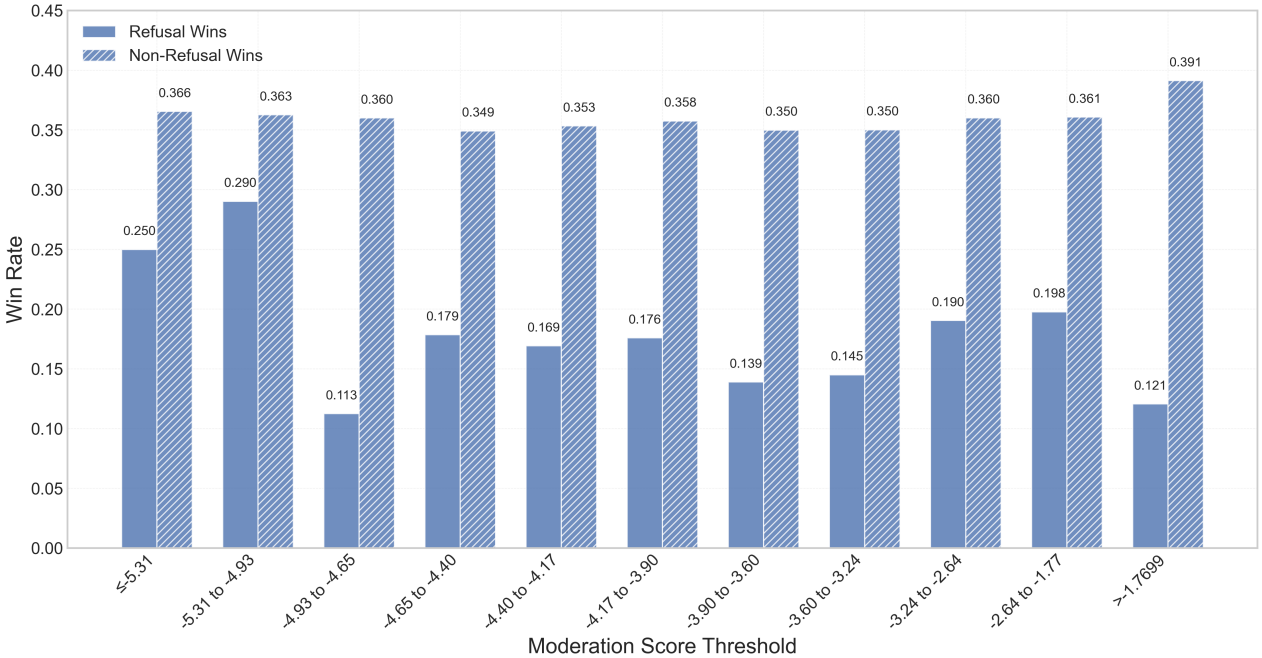


Figure C.3: Win Rate by Thresholds - OpenAI Moderation



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

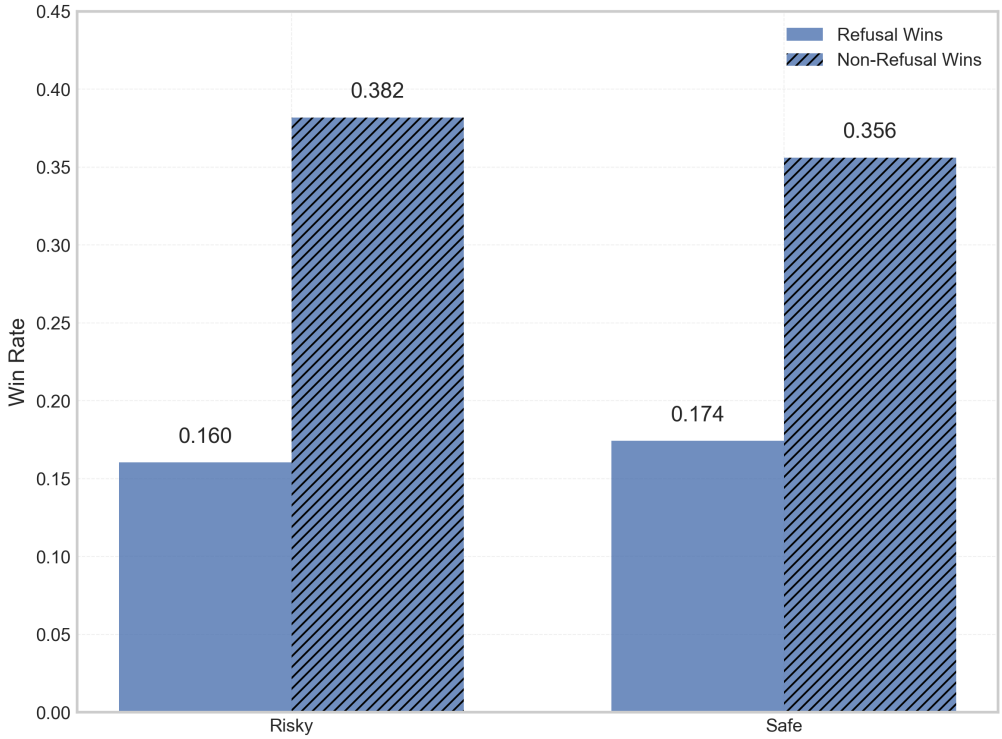
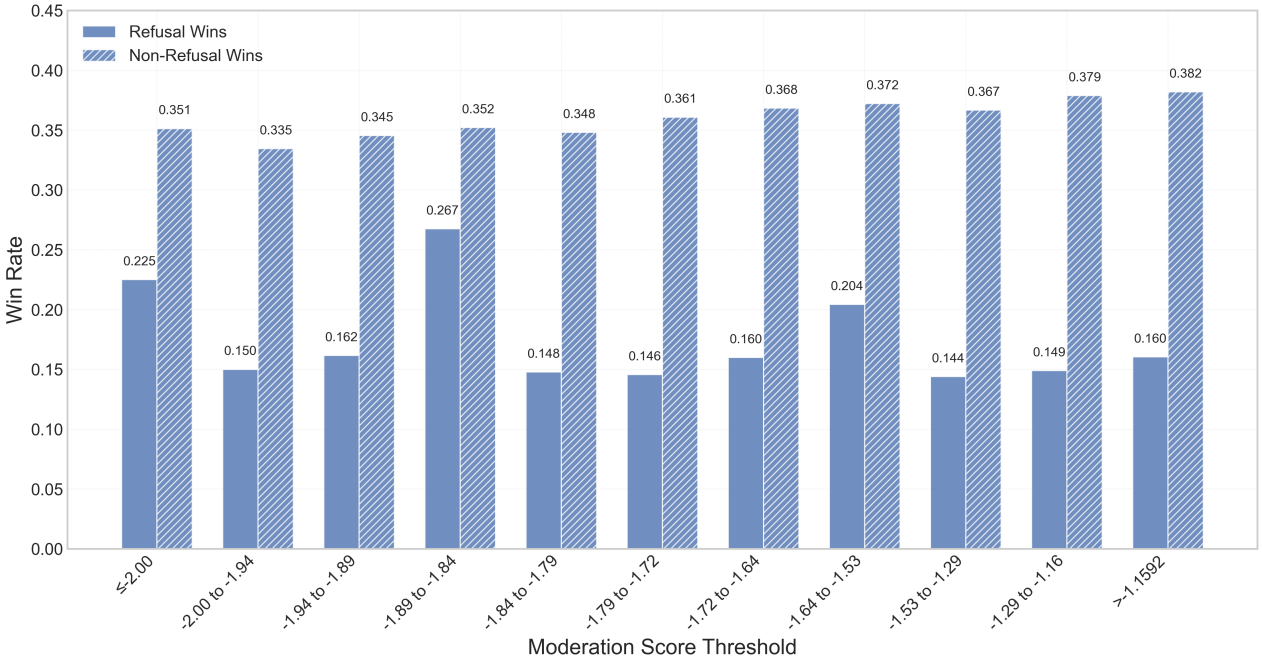


Figure C.5: Win Rate by Thresholds - RoBERTa Moderation (7.50% top scores)



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.

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.

Table C.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	0.290^{***}	0.234^{***}	0.476^{***}	0.379^{***}	0.228^{***}	0.393^{***}
	(0.017)	(0.023)	(0.020)	(0.015)	(0.026)	(0.025)
Class-specific parameters						
Refusal coef. (risky) $\beta_k(1)$	-1.425^{***}	–	–	-1.485^{***}	–	–
	(0.160)			(0.171)		
Refusal coef. (safe) $\beta_k(0)$	-1.149^{***}	0.619^{*}	-2.398^{***}	-0.974^{***}	0.534	-2.663^{***}
	(0.238)	(0.353)	(0.316)	(0.241)	(0.393)	(0.372)
Proprietary × risky $\gamma_k(1)$	-1.004^{***}	–	–	-0.712^{***}	–	–
	(0.146)			(0.121)		
Risky intercept α	-1.357^{***}	–	–	-1.257^{***}	–	–
	(0.185)			(0.150)		
Outside option ρ	-2.843^{***}	-1.711^{***}	-3.796^{***}	-2.033^{***}	-0.948	-3.035^{***}
	(0.340)	(0.323)	(0.399)	(0.698)	(0.642)	(0.688)
Text coefficients ψ	(common across classes)					
Character length	0.106^{***}			0.134^{***}		
	(0.023)			(0.029)		
Average word length	-0.723^{***}			-0.437		
	(0.149)			(0.280)		
|Character length|	0.252^{***}			0.258^{***}		
	(0.021)			(0.022)		
|Average word length|	-0.209^{**}			-0.229^{**}		
	(0.105)			(0.108)		
Model-specific fixed effects δ_j	(common across classes; baseline GPT-4 = 0)					
claude-v1	-0.579^{***}			-0.481^{***}		
	(0.072)			(0.110)		
claude-instant-v1	-0.733^{***}			-0.646^{***}		
	(0.088)			(0.114)		
gpt-3.5-turbo	-0.800^{***}			-0.675^{***}		
	(0.072)			(0.119)		
palm-2	-1.478^{***}			-1.353^{***}		
	(0.081)			(0.119)		
guanaco-33b	-1.525^{***}			-1.424^{***}		
	(0.103)			(0.132)		
vicuna-13b	-1.571^{***}			-1.450^{***}		
	(0.079)			(0.119)		
wizardlm-13b	-1.722^{***}			-1.618^{***}		
	(0.118)			(0.146)		
vicuna-7b	-1.961^{***}			-1.858^{***}		
	(0.080)			(0.118)		
koala-13b	-2.005^{***}			-1.883^{***}		
	(0.075)			(0.119)		
gpt4all-13b-snoozy	-2.286^{***}			-2.164^{***}		
	(0.115)			(0.155)		
mpt-7b-chat	-2.357^{***}			-2.216^{***}		
	(0.085)			(0.133)		
alpaca-13b	-2.378^{***}			-2.217^{***}		
	(0.084)			(0.143)		
RWKV-4-Raven-14B	-2.478^{***}			-2.335^{***}		
	(0.079)			(0.123)		
oasst-pythia-12b	-2.548^{***}			-2.407^{***}		
	(0.081)			(0.130)		
fastchat-t5-3b	-2.830^{***}			-2.707^{***}		
	(0.088)			(0.138)		
chatglm-6b	-2.953^{***}			-2.831^{***}		
	(0.080)			(0.128)		
dolly-v2-12b	-3.134^{***}			-2.971^{***}		
	(0.092)			(0.144)		
stablelm-tuned-alpha-7b	-3.138^{***}			-2.996^{***}		
	(0.095)			(0.139)		
llama-13b	-3.231^{***}			-3.051^{***}		
	(0.104)			(0.174)		

Note: RoBERTa Moderation thresholds correspond to 5.00% and 7.50% tails of the moderation score distribution. Standard errors in parentheses from ∼200 bootstrap samples. $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.10$.

Table C.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	0.255^{***}	0.465^{***}	0.280^{***}	0.343^{***}	0.350^{***}	0.306^{***}
	(0.030)	(0.036)	(0.031)	(0.013)	(0.018)	(0.018)
Class-specific parameters						
Refusal coef. (risky) $\beta_k(1)$	-1.774^{***}	–	–	-1.581^{***}	–	–
	(0.243)			(0.205)		
Refusal coef. (safe) $\beta_k(0)$	-1.467^{***}	-2.062^{***}	0.359	-1.148^{***}	-2.407^{***}	0.107
	(0.310)	(0.441)	(0.326)	(0.258)	(0.334)	(0.298)
Proprietary × risky $\gamma_k(1)$	-0.588^{***}	–	–	-0.268^{***}	–	–
	(0.126)			(0.090)		
Risky intercept α	-1.739^{***}	–	–	-1.352^{***}	–	–
	(0.255)			(0.113)		
Outside option ρ	-1.797^{***}	-2.721^{***}	-0.704	-1.021	-2.460^{***}	-0.219
	(0.321)	(0.309)	(0.603)	(0.734)	(0.716)	(0.777)
Text coefficients ψ	(common across classes)					
Character length	0.146^{***}			0.169^{***}		
	(0.024)			(0.032)		
Average word length	-0.275^{*}			-0.029		
	(0.146)			(0.294)		
|Character length|	0.259^{***}			0.265^{***}		
	(0.023)			(0.022)		
|Average word length|	-0.237^{**}			-0.291^{***}		
	(0.108)			(0.107)		
Model-specific fixed effects δ_j	(common across classes; baseline GPT-4 = 0)					
claude-v1	-0.575^{***}			-0.514^{***}		
	(0.074)			(0.101)		
claude-instant-v1	-0.719^{***}			-0.651^{***}		
	(0.085)			(0.108)		
gpt-3.5-turbo	-0.769^{***}			-0.686^{***}		
	(0.068)			(0.104)		
palm-2	-1.442^{***}			-1.362^{***}		
	(0.081)			(0.112)		
guanaco-33b	-1.487^{***}			-1.422^{***}		
	(0.102)			(0.128)		
vicuna-13b	-1.510^{***}			-1.427^{***}		
	(0.079)			(0.108)		
wizardlm-13b	-1.682^{***}			-1.600^{***}		
	(0.127)			(0.142)		
vicuna-7b	-1.902^{***}			-1.837^{***}		
	(0.078)			(0.109)		
koala-13b	-1.944^{***}			-1.881^{***}		
	(0.083)			(0.100)		
gpt4all-13b-snoozy	-2.201^{***}			-2.119^{***}		
	(0.113)			(0.149)		
alpaca-13b	-2.273^{***}			-2.152^{***}		
	(0.089)			(0.131)		
mpt-7b-chat	-2.274^{***}			-2.191^{***}		
	(0.089)			(0.118)		
RWKV-4-Raven-14B	-2.412^{***}			-2.312^{***}		
	(0.085)			(0.111)		
oasst-pythia-12b	-2.466^{***}			-2.364^{***}		
	(0.083)			(0.115)		
fastchat-t5-3b	-2.762^{***}			-2.680^{***}		
	(0.092)			(0.116)		
chatglm-6b	-2.891^{***}			-2.810^{***}		
	(0.088)			(0.109)		
dolly-v2-12b	-3.030^{***}			-2.909^{***}		
	(0.099)			(0.132)		
stablelm-tuned-alpha-7b	-3.070^{***}			-2.972^{***}		
	(0.095)			(0.117)		
llama-13b	-3.107^{***}			-2.964^{***}		
	(0.108)			(0.150)		

Note: OpenAI Moderation thresholds correspond to 3.68% and 7.50% tails of the moderation score distribution. Standard errors in parentheses from 200 bootstrap samples. $***\ p < 0.01$, $**\ p < 0.05$, $*\ p < 0.10$.

D Refusal Labeling 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}‘‘‘

"""

    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:
```

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 $\theta_i = \{\alpha_i(0), \alpha_i(1), \beta_i(0), \beta_i(1), \gamma_i(0), \gamma_i(1), \psi_i, \rho_i\}$ collects all user-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.

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

Parameters	**Heterogeneous Model (Including Ties)**		
	Class 1	**Class 2**	**Class 3**
Class probability π_k	0.217^{***} (0.017)	0.259^{***} (0.024)	0.524^{***} (0.026)
Class-specific parameters			
Refusal coef. (risky) $\beta_k(1)$	-1.529^{***} (0.176)	–	–
Refusal coef. (safe) $\beta_k(0)$	-1.221^{***} (0.235)	0.355 (0.320)	-2.251^{***} (0.302)
Proprietary × risky $\gamma_k(1)$	-1.120^{***} (0.181)	–	–
Risky intercept α	-1.301^{***} (0.227)	–	–
Outside option ρ	-2.622^{***} (0.433)	-1.535^{***} (0.439)	-3.481^{***} (0.515)
Text coefficients ψ	(common across classes)		
Character length	0.100^{***} (0.025)		
Average word length	-0.500^{***} (0.186)		
|Character length|	0.262^{***} (0.021)		
|Average word length|	-0.228^{**} (0.105)		
Model-specific fixed effects δ_j	(baseline: GPT-4 = 0)		
claude-v1	-0.579^{***} (0.072)		
claude-instant-v1	-0.732^{***} (0.087)		
gpt-3.5-turbo	-0.801^{***} (0.068)		
palm-2	-1.456^{***} (0.078)		
guanaco-33b	-1.510^{***} (0.103)		
vicuna-13b	-1.538^{***} (0.073)		
wizardlm-13b	-1.703^{***} (0.122)		
vicuna-7b	-1.928^{***} (0.080)		
koala-13b	-1.971^{***} (0.067)		
gpt4all-13b-snoozy	-2.275^{***} (0.107)		
mpt-7b-chat	-2.347^{***} (0.084)		
alpaca-13b	-2.366^{***} (0.087)		
RWKV-4-Raven-14B	-2.464^{***} (0.080)		
oasst-pythia-12b	-2.534^{***} (0.078)		
fastchat-t5-3b	-2.836^{***} (0.085)		
chatglm-6b	-2.938^{***} (0.079)		
dolly-v2-12b	-3.119^{***} (0.091)		
stablelm-tuned-alpha-7b	-3.124^{***} (0.091)		
llama-13b	-3.219^{***} (0.108)		
Log-likelihood	−23,854		

Note: Standard errors in parentheses from 200 bootstrap samples. $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.10$. "–" indicates parameter not applicable for that class type.

F Regression on Text Characteristics

In Table F.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 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.

Table F.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).

G EM Algorithm Details

This appendix provides the detailed equations for the Expectation-Maximization (EM) algorithm used to estimate the model parameters.

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:

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

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:

$$\pi_k^{\text{new}} = \frac{\sum_{i \in \mathcal{I}} q_{ik}}{|\mathcal{I}|}, \tag{G.2}$$

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

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

$$\theta_k^{\text{new}} = \arg\max_{\theta_k} \sum_{i \in \mathcal{I}} q_{ik} \log L_i(\theta_k, \delta), \tag{G.3}$$

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 $\delta = \{\delta_j\}_{j \in \mathcal{J}}$:

$$\delta^{\text{new}} = \arg\max_{\delta} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}} q_{ik} \cdot \log L_i(\theta_k^{\text{new}}, \delta), \tag{G.4}$$

where θ_k^{new} represents the updated class-specific parameters, and $L_i(\theta_k^{\text{new}}, \delta)$ is the 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 the change in the log-likelihood is below a predetermined threshold, or until a maximum number of iterations is reached.

H Model Selection

Table H.1 compares model specifications by BIC, varying the number of latent classes and whether covariate coefficients are common or class-specific. We select the 3-class model with common covariate coefficients for parsimony and interpretability. This specification achieves the lowest BIC among common-coefficient models. While the class-specific 3-class model achieves a slightly lower BIC, the improvement is modest and allowing class-specific coefficients does not yield qualitatively different results.

Table H.1: Model Comparison by BIC

Model Specification	**Log-Likelihood**	**Parameters**	**BIC**
Common Covariate Coefficients (ψ) Across Classes			
1 Class	$-24,552$	28	49,405
2 Classes	$-23,631$	31	47,595
3 Classes	$-23,339$	34	47,043
4 Classes	$-23,340$	37	47,077
Class-Specific Covariate Coefficients (ψ_i)			
2 Classes	$-23,653$	35	47,682
3 Classes	$-23,270$	42	46,991
4 Classes	$-23,243$	49	47,012

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, which achieves the lowest BIC among common-coefficient specifications. Allowing for class-specific coefficients yields only modest BIC improvements. Counterfactual results are also robust to these alternative specifications.

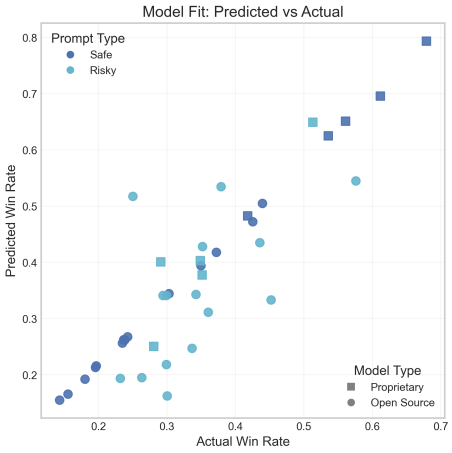
I Model Fit

In addition to the BIC metric discussed in the main text, we assess model fit by examining how well the model reproduces observed LLM win rates. Using the parameter estimates from the selected heterogeneous specification, we simulate user choices between the LLMs encountered in the data, conditional on observed prompt risk choices. Figure I.1 plots predicted versus actual win rates separately for safe and risky prompts across all 20 LLMs, where each point corresponds to one LLM.

Safe prompts generate substantially more observations per model (approximately 2,200 on average) than risky prompts (approximately 90 on average), resulting in a tighter relationship between predicted and actual win rates in the safe-prompt market. Greater dispersion for risky prompts reflects the smaller number of observations per LLM in that segment.

Overall, the model reproduces observed win rates closely across both prompt types, indicating that it captures the main substitution patterns across LLMs.

Figure I.1: Model Fit. Predicted vs Actual Win Rates Across LLMs



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

J Intensive vs. Extensive Margin Decomposition

The counterfactual scenarios in Table 7 allow both the extensive margin (fraction of risky prompts) and the intensive margin (substitution across LLMs within each prompt type) to adjust simultaneously. Since Classes 2 and 3 never submit risky prompts by construction, all extensive-margin variation concentrates in Class 1, where the risky-prompt probability responds to changes in inclusive values through the logit functional form. To isolate the role of within-market substitution from changes in risky prompting behavior, Tables J.1 and J.2 re-run all counterfactual scenarios while holding each class's risky-prompt probability fixed at its baseline value. Table J.3 then compares HHI changes across the full model (both margins) and the fixed-extensive-margin specification.

Comparing Table J.1 to Table 7 reveals the relative importance of each margin. Scenarios where intensive-margin HHI changes closely track the full-model changes, such as "No Risky Refusals + Equal Quality", "Safe Baseline + Equal Quality", and "Divergent Moderation", indicate that within-market substitution is the primary driver. By contrast, scenarios where the fixed-extensive-margin HHI change is substantially smaller, such as "Universal Strict Moderation", reveal a large extensive-margin component, meaning that changes in the fraction of risky prompts materially affect measured concentration. The model exit scenarios are predominantly intensive-margin, though removing an LLM also shifts inclusive values differentially across the risky and safe markets, producing a small extensive-margin component (e.g., removing GPT-3.5-turbo shifts Class 1 HHI from +10.9% to +11.3%).

Table J.1: Counterfactual Effects on Risk-Taking Segment: Intensive Margin Only (Fixed Extensive Margin)

	Market Concentration				Risky Requests (%)
	All Prompts		Risky Prompts		
Counterfactual Scenario	**HHI**	**Prop. (%)**	**HHI**	**Prop. (%)**	
Baseline and Moderation Policies					
Baseline	1023	60.6	682	34.7	11.8
No Risky Refusals	1028 (+0.5%)	60.9	692 (+1.4%)	37.6	11.8 (+0%)
No Risky Refusals + Equal Quality	1099 (+7.4%)	64.0	1094 (+60.3%)	64.1	11.8 (+0%)
Safe Baseline Risky Refusals	1028 (+0.6%)	60.9	696 (+2.0%)	37.5	11.8 (+0%)
Safe Baseline + Equal Quality	1100 (+7.5%)	64.0	1100 (+61.2%)	64.0	11.8 (+0%)
Universal Strict Moderation	1027 (+0.4%)	60.8	685 (+0.4%)	33.4	11.8 (+0%)
Divergent Moderation (Prop: 100% / OS: 0%)	960 (−6.2%)	57.6	778 (+14.0%)	8.8	11.8 (+0%)
Model Exit Scenarios					
Remove GPT-3.5-turbo	1138 (+11.3%)	56.8	725 (+6.3%)	31.2	11.8 (+0%)
Remove gpt4all-13b-snoozy	1061 (+3.7%)	61.9	719 (+5.4%)	36.4	11.8 (+0%)
Remove llama-13b	1036 (+1.3%)	61.1	697 (+2.2%)	35.6	11.8 (+0%)

Note: Each class's probability of submitting a risky prompt is held fixed at its baseline value. Changes in HHI and market shares reflect only the intensive margin (within-market substitution across LLMs), not changes in the fraction of risky prompts. This analysis focuses on risk-taking (Class 1) users. Compare to Table 7, which allows both margins to adjust. Prop. (%) denotes the combined market share of proprietary LLMs (GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2).

Table J.2: Counterfactual Effects on Overall Market: Intensive Margin Only (Fixed Extensive Margin)

	Market Concentration				Risky Requests (%)
	All Prompts		Risky Prompts		
Counterfactual Scenario	**HHI**	**Prop. (%)**	**HHI**	**Prop. (%)**	
Baseline and Moderation Policies					
Baseline	1076	63.1	682	34.7	3.2
No Risky Refusals	1078 (+0.1%)	63.2	692 (+1.4%)	37.6	3.2 (+0%)
No Risky Refusals + Equal Quality	1097 (+2.0%)	64.0	1094 (+60.3%)	64.1	3.2 (+0%)
Safe Baseline Risky Refusals	1078 (+0.2%)	63.2	696 (+2.0%)	37.5	3.2 (+0%)
Safe Baseline + Equal Quality	1098 (+2.0%)	64.0	1100 (+61.2%)	64.0	3.2 (+0%)
Universal Strict Moderation	1077 (+0.1%)	63.1	685 (+0.4%)	33.4	3.2 (+0%)
Divergent Moderation (Prop: 100% / OS: 0%)	1057 (−1.8%)	62.3	778 (+14.0%)	8.8	3.2 (+0%)
Model Exit Scenarios					
Remove GPT-3.5-turbo	1202 (+11.7%)	59.2	725 (+6.3%)	31.2	3.2 (+0%)
Remove gpt4all-13b-snoozy	1115 (+3.6%)	64.4	719 (+5.4%)	36.4	3.2 (+0%)
Remove llama-13b	1089 (+1.2%)	63.5	697 (+2.2%)	35.6	3.2 (+0%)

Note: Each class's probability of submitting a risky prompt is held fixed at its baseline value. Changes in HHI and market shares reflect only the intensive margin (within-market substitution across LLMs), not changes in the fraction of risky prompts. This table reports results for the overall market across all user segments. Compare to Table M.1, which allows both margins to adjust. Prop. (%) denotes the combined market share of proprietary LLMs (GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2).

Table J.3: Decomposition of HHI Changes: Full Model vs. Intensive Margin Only

	Risk-Taking Segment (Class 1)		Overall Market	
Counterfactual Scenario	**Full Model**	**Intensive Only**	**Full Model**	**Intensive Only**
Baseline and Moderation Policies				
No Risky Refusals	1019 (−0.4%)	1028 (+0.5%)	1075 (−0.1%)	1078 (+0.1%)
No Risky Refusals + Equal Quality	1098 (+7.4%)	1099 (+7.4%)	1097 (+2.0%)	1097 (+2.0%)
Safe Baseline Risky Refusals	1020 (−0.3%)	1028 (+0.6%)	1075 (−0.1%)	1078 (+0.2%)
Safe Baseline + Equal Quality	1100 (+7.5%)	1100 (+7.5%)	1098 (+2.0%)	1098 (+2.0%)
Universal Strict Moderation	1081 (+5.7%)	1027 (+0.4%)	1092 (+1.5%)	1077 (+0.1%)
Divergent Moderation (Prop: 100% / OS: 0%)	983 (−3.9%)	960 (−6.2%)	1064 (−1.1%)	1057 (−1.8%)
Model Exit Scenarios				
Remove GPT-3.5-turbo	1134 (+10.9%)	1138 (+11.3%)	1201 (+11.6%)	1202 (+11.7%)
Remove gpt4all-13b-snoozy	1062 (+3.8%)	1061 (+3.7%)	1115 (+3.6%)	1115 (+3.6%)
Remove llama-13b	1036 (+1.3%)	1036 (+1.3%)	1089 (+1.2%)	1089 (+1.2%)

Note: "Full Model" allows both the extensive margin (risky-prompt probability) and intensive margin (within-market substitution) to adjust. "Intensive Only" holds each class's risky-prompt probability fixed at its baseline value. Reported values are HHI levels with percentage changes relative to baseline in parentheses. Baseline HHI: risk-taking segment = 1023, overall market = 1076.

K Quality-Moderation Sensitivity Analysis

Our counterfactual simulations change refusal rates while holding model intercepts (quality) fixed. If reducing refusals also degrades the quality of non-refused risky responses, for example through toned-down or hedged answers, the simulated HHI movements could shift. To assess this concern, we introduce a quality degradation parameter $\lambda \geq 0$ that adjusts the model-specific intercept for risky prompts proportionally to the change in refusal rates:

$$\delta_j^{\text{adj}}(S_{it} = 1) \;=\; \delta_j \;-\; \lambda \cdot \big(r_j(1) - r_j^{\text{cf}}(1)\big),$$

where $r_j(1)$ is model j's estimated (baseline) refusal rate on risky prompts and $r_j^{\text{cf}}(1)$ is its counterfactual rate. When $\lambda = 0$, this reduces to our baseline specification. When $\lambda > 0$, models that reduce their refusal rate ($r_j^{\text{cf}}(1) < r_j(1)$) suffer a quality penalty on risky prompts ($S_{it} = 1$), while utility from safe prompts ($S_{it} = 0$) is unaffected.

To calibrate λ, we estimate a cross-sectional OLS regression across the 20 models:

$$\hat{\delta}_j \;=\; a \;+\; \lambda \cdot r_j(1) \;+\; c \cdot M_j \;+\; \varepsilon_j,$$

yielding $\hat{\lambda} = 0.78$ (SE $= 2.65$, $R^2 = 0.65$), with the proprietary control M_j at 1.58 (SE $= 0.45$). The confidence bounds on $\hat{\lambda}$ are wide and include zero, though with only 20 models the regression has limited power to detect such an association. We use the point estimate $\hat{\lambda} = 0.78$ to assess sensitivity.

Table K.1 reports results. At the calibrated $\hat{\lambda}$, all counterfactual conclusions are directionally preserved. The competition-enhancing effect of divergent moderation is dampened (from -3.9% to -2.3% among risk-takers), and the concentration-increasing effect of universal strict moderation is also dampened (from $+5.7\%$ to $+3.9\%$). Under divergent moderation, open-source models that eliminate refusals suffer a quality penalty, partially offsetting their competitive gain. Under universal strict moderation, models forced to refuse more risky prompts receive a positive adjustment to δ_j (since $r_j^{\text{cf}}(1) > r_j(1)$). The intuition is a composition effect: a model that refuses a larger share of risky prompts no longer needs to hedge or qualify its responses to borderline cases, so the non-refused responses it does provide can be more direct and substantive. This quality improvement partially offsets the concentration increase.

Finally, if quality degradation for non-refused risky responses is severe, for instance because models produce heavily hedged or watered-down completions, then from the user's perspective such responses are functionally equivalent to refusals. In that case, the counterfactual policy of "reducing refusals" was not genuinely implemented: the model merely replaced an explicit refusal with an implicit one. Our baseline specification, which treats refusals as binary and observable, would then remain the appropriate framework, and the λ adjustment would overstate the true quality penalty.

Table K.1: Quality-Moderation Sensitivity: HHI Across λ Values

	$\lambda = 0$		$\lambda = \hat{\lambda}$ (0.78)	
Scenario	Overall	Risk-takers	Overall	Risk-takers
Baseline	1076	1023	1076	1023
No Risky Refusals	1075 (−0.1%)	1019 (−0.4%)	1076 (+0.0%)	1023 (+0.0%)
No Risky Refusals + Equal Quality	1097 (+2.0%)	1098 (+7.4%)	1096 (+1.9%)	1094 (+7.0%)
Universal Strict	1092 (+1.5%)	1081 (+5.7%)	1087 (+1.1%)	1063 (+3.9%)
Divergent Moderation	1064 (−1.1%)	983 (−3.9%)	1069 (−0.6%)	999 (−2.3%)

Note: Each cell reports HHI with percentage change relative to baseline in parentheses. $\lambda = 0$ matches the baseline counterfactual specification. $\hat{\lambda} = 0.78$ is the cross-sectional OLS estimate of the association between model quality and risky refusal rates; the confidence interval includes zero (SE = 2.65).

L Output-Side Safety: Risky Prompt Compliance

Our primary safety metric, the fraction of risky prompts, captures demand-side behavior: how often users attempt risky requests. A risky prompt that is refused produces no harmful output, while a risky prompt that is not refused does. To complement the demand-side measure, we construct an output-side proxy: the fraction of all risk-taking segment prompts where a risky request receives a non-refused response. This "risky responded" metric multiplies the risk-taking segment's risky-prompt probability by the share-weighted non-refusal rate across models, capturing the joint effect of how many risky prompts are attempted and how many of those are not refused.

Table L.1 reports this metric across counterfactual scenarios. Under the baseline, approximately 9.6% of risk-taking segment prompts result in a non-refused risky response. Under "No Risky Refusals" (all refusals eliminated), this rises to 13.1% (+36.0%) as both the risky request rate and the non-refusal rate increase simultaneously. Under "Universal Strict Moderation" (100% refusal), the metric falls to nearly zero (−100%): even though some risky prompts are still attempted, all are refused. Under "Divergent Moderation" (proprietary models refuse 100%, open-source refuse 0%), the metric declines to 8.5% (−11.7%), reflecting the compositional shift as risk-takers substitute toward open-source models that do not refuse, while proprietary models' refusals reduce the overall non-refusal rate.

These results confirm that the input-side and output-side metrics move in the same direction across all scenarios, and that refusal policy is the primary lever controlling output-side harm exposure. Content moderation operates through both margins: it directly blocks harmful outputs (by refusing risky prompts) and indirectly discourages risky requests (via the endogenous prompt-type choice).

Table L.1: Counterfactual Effects on Risk-Taking Segment: Output-Side Safety

	Market Concentration		Risky Prompts	
Counterfactual Scenario	**HHI**	**Prop. (%)**	**Risky Req. (%)**	**Risky Responded (%)**
Baseline and Moderation Policies				
Baseline	1023	60.6	11.8	9.6
No Risky Refusals	1019 (−0.4%)	60.5	13.4 (+14%)	13.1 (+36.0%)
No Risky Refusals + Equal Quality	1098 (+7.4%)	64.0	21.2 (+79%)	20.9 (+116.5%)
Safe Baseline Risky Refusals	1020 (−0.3%)	60.5	13.3 (+12%)	12.7 (+32.0%)
Safe Baseline + Equal Quality	1100 (+7.5%)	64.0	20.9 (+77%)	20.2 (+110.1%)
Universal Strict Moderation	1081 (+5.7%)	63.2	3.0 (−75%)	0.0 (−100.0%)
Divergent Moderation (Prop: 100% / OS: 0%)	983 (−3.9%)	58.8	9.7 (−18%)	8.5 (−11.7%)
Model Exit Scenarios				
Remove GPT-3.5-turbo	1134 (+10.9%)	56.7	12.3 (+4%)	10.0 (+4.1%)
Remove gpt4all-13b-snoozy	1062 (+3.8%)	62.0	11.6 (−1%)	9.5 (−1.6%)
Remove llama-13b	1036 (+1.3%)	61.1	11.7 (−1%)	9.6 (−0.7%)

Note: This table reports outcomes for the risk-taking segment only. "Risky Req. (%)" is the endogenous fraction of risk-taking segment prompts classified as high-risk. "Risky Responded (%)" is the fraction of all risk-taking segment prompts where a risky request receives a non-refused response, equal to the risky-prompt probability multiplied by the share-weighted non-refusal rate across models. This metric captures output-side harm exposure: a refused risky prompt produces no harmful output, while a non-refused one does. Prop. (%) denotes the combined market share of proprietary LLMs (GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2).

M Detailed Counterfactual Market Shares

This section presents detailed market share decompositions for each counterfactual policy discussed in Section 7.2 of the main text. Tables M.2–M.5 present LLM-level market shares under the four main counterfactual scenarios.

Table M.1: Summary of Counterfactual Results (Overall Market)

	Market Concentration				Risky Requests (%)
	All Prompts		Risky Prompts		
Counterfactual Scenario	**HHI**	**Prop. (%)**	**HHI**	**Prop. (%)**	
Baseline and Moderation Policies					
Baseline	1076	63.1	682	34.7	3.2
No Risky Refusals	1075 (−0.1%)	63.0	692 (+1.4%)	37.6	3.6 (+14%)
No Risky Refusals + Equal Quality	1097 (+2.0%)	64.0	1094 (+60.3%)	64.1	5.7 (+79%)
Safe Baseline Risky Refusals	1075 (−0.1%)	63.1	696 (+2.0%)	37.5	3.6 (+12%)
Safe Baseline + Equal Quality	1098 (+2.0%)	64.0	1100 (+61.2%)	64.0	5.6 (+77%)
Universal Strict Moderation	1092 (+1.5%)	63.8	685 (+0.4%)	33.4	0.8 (−75%)
Divergent Moderation (Prop: 100% / OS: 0%)	1064 (−1.1%)	62.6	778 (+14.0%)	8.8	2.6 (−18%)
Model Exit Scenarios					
Remove GPT-3.5-turbo	1201 (+11.6%)	59.2	725 (+6.3%)	31.2	3.3 (+4%)
Remove gpt4all-13b-snoozy	1115 (+3.6%)	64.4	719 (+5.4%)	36.4	3.1 (−1%)
Remove llama-13b	1089 (+1.2%)	63.5	697 (+2.2%)	35.6	3.2 (−1%)

Note: HHI measures market concentration (scale: 0–10,000). Percentage changes in parentheses are relative to baseline. Prop. (%) denotes the combined market share of proprietary LLMs (GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2). This table reports results for the overall market across all user segments.

Table M.2: No Risky Refusals: Market Shares with No Content Moderation

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	21.1	12.9	21.9	22.2
2	claude-v1	Prop.	13.3	8.2	13.8	14.0
3	claude-instant-v1	Prop.	10.8	6.7	11.3	11.4
4	gpt-3.5-turbo	Prop.	8.8	5.4	9.2	9.3
5	guanaco-33b	OS	5.9	9.1	5.4	5.3
6	vicuna-13b	OS	5.5	8.6	5.1	5.0
7	palm-2	Prop.	5.4	3.4	5.7	5.8
8	wizardlm-13b	OS	4.5	7.0	4.2	4.1
9	koala-13b	OS	3.7	5.7	3.4	3.3
10	vicuna-7b	OS	3.6	5.6	3.3	3.3
11	mpt-7b-chat	OS	2.2	3.5	2.1	2.0
12	gpt4all-13b-snoozy	OS	2.2	3.5	2.1	2.0
13	RWKV-4-Raven-14B	OS	2.0	3.2	1.9	1.8
14	oasst-pythia-12b	OS	1.9	2.9	1.7	1.7
15	alpaca-13b	OS	1.9	2.9	1.7	1.7
16	fastchat-t5-3b	OS	1.4	2.3	1.4	1.3
17	chatglm-6b	OS	1.4	2.1	1.3	1.2
18	stablelm-tuned-alpha-7b	OS	1.1	1.7	1.0	1.0
19	dolly-v2-12b	OS	0.9	1.4	0.8	0.8
20	llama-13b	OS	0.8	1.3	0.8	0.7
21	outside option	–	1.5	2.4	1.8	1.8
HHI			1019	692	1075	1097
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			60.5	37.6	63.0	64.0
Open-Source LLMs			39.5	62.4	37.0	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			13.4	0.0	0.0	3.6
Safe Requests			86.6	100.0	100.0	96.4

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2. User Behavior shows the percentage of prompts that are risky vs. safe for each class under this counterfactual scenario.

Table M.3: No Risky Refusals + Equal Quality: Market Shares with No Refusals and No Proprietary Penalty

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	22.4	22.2	22.2	22.2
2	claude-v1	Prop.	14.1	14.1	14.0	14.0
3	claude-instant-v1	Prop.	11.5	11.6	11.4	11.4
4	gpt-3.5-turbo	Prop.	9.4	9.4	9.3	9.3
5	palm-2	Prop.	5.8	5.9	5.8	5.8
6	guanaco-33b	OS	5.3	5.3	5.3	5.3
7	vicuna-13b	OS	5.0	5.0	5.0	5.0
8	wizardlm-13b	OS	4.1	4.1	4.1	4.1
9	koala-13b	OS	3.3	3.3	3.3	3.3
10	vicuna-7b	OS	3.3	3.3	3.3	3.3
11	mpt-7b-chat	OS	2.0	2.0	2.0	2.0
12	gpt4all-13b-snoozy	OS	2.0	2.0	2.0	2.0
13	RWKV-4-Raven-14B	OS	1.8	1.9	1.8	1.8
14	oasst-pythia-12b	OS	1.7	1.7	1.7	1.7
15	alpaca-13b	OS	1.7	1.7	1.7	1.7
16	fastchat-t5-3b	OS	1.3	1.3	1.3	1.3
17	chatglm-6b	OS	1.2	1.2	1.2	1.2
18	stablelm-tuned-alpha-7b	OS	1.0	1.0	1.0	1.0
19	dolly-v2-12b	OS	0.8	0.8	0.8	0.8
20	llama-13b	OS	0.7	0.7	0.7	0.7
21	outside option	–	1.4	1.4	1.8	1.8
HHI			1098	1094	1097	1097
Summary Statistics						
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			64.0	64.1	64.0	64.0
Open-Source LLMs			36.0	35.9	36.0	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			21.2	0.0	0.0	5.7
Safe Requests			78.8	100.0	100.0	94.3

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2. User Behavior shows the percentage of prompts that are risky vs. safe for each class under this counterfactual scenario.

Table M.4: Universal Strict Moderation: Market Shares under 100% Risky Refusal

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	22.1	11.8	22.2	22.3
2	claude-v1	Prop.	13.9	5.5	14.0	14.0
3	claude-instant-v1	Prop.	11.3	4.3	11.4	11.4
4	gpt-3.5-turbo	Prop.	9.2	5.3	9.3	9.3
5	palm-2	Prop.	5.7	2.4	5.7	5.8
6	guanaco-33b	OS	5.4	7.8	5.3	5.3
7	vicuna-13b	OS	5.1	8.8	5.0	5.0
8	wizardlm-13b	OS	4.2	7.1	4.1	4.1
9	koala-13b	OS	3.4	5.1	3.3	3.3
10	vicuna-7b	OS	3.3	5.4	3.3	3.3
11	mpt-7b-chat	OS	2.1	3.4	2.0	2.0
12	gpt4all-13b-snoozy	OS	2.0	3.3	2.0	2.0
13	RWKV-4-Raven-14B	OS	1.9	3.0	1.9	1.8
14	alpaca-13b	OS	1.7	2.9	1.7	1.7
15	oasst-pythia-12b	OS	1.7	2.6	1.7	1.7
16	fastchat-t5-3b	OS	1.3	2.2	1.3	1.3
17	chatglm-6b	OS	1.3	2.2	1.2	1.2
18	stablelm-tuned-alpha-7b	OS	1.0	1.6	1.0	1.0
19	dolly-v2-12b	OS	0.8	1.5	0.8	0.8
20	llama-13b	OS	0.8	1.7	0.7	0.7
21	outside option	–	1.7	12.0	1.9	1.8
HHI			1081	685	1092	1098
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			63.2	33.4	63.8	64.0
Open-Source LLMs			36.8	66.6	36.2	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			3.0	0.0	0.0	0.8
Safe Requests			97.0	100.0	100.0	99.2

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2. User Behavior shows the percentage of prompts that are risky vs. safe for each class under this counterfactual scenario.

Table M.5: Divergent Moderation: Market Shares with Proprietary Refusing All Risky Prompts

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	20.6	3.4	21.8	22.2
2	claude-v1	Prop.	12.9	1.6	13.7	14.0
3	claude-instant-v1	Prop.	10.5	1.2	11.2	11.4
4	gpt-3.5-turbo	Prop.	8.6	1.5	9.1	9.3
5	guanaco-33b	OS	6.1	13.2	5.5	5.3
6	vicuna-13b	OS	5.7	12.4	5.2	5.0
7	palm-2	Prop.	5.3	0.7	5.6	5.8
8	wizardlm-13b	OS	4.7	10.1	4.3	4.1
9	koala-13b	OS	3.8	8.3	3.4	3.3
10	vicuna-7b	OS	3.7	8.1	3.4	3.3
11	mpt-7b-chat	OS	2.3	5.1	2.1	2.0
12	gpt4all-13b-snoozy	OS	2.3	5.1	2.1	2.0
13	RWKV-4-Raven-14B	OS	2.1	4.6	1.9	1.8
14	oasst-pythia-12b	OS	1.9	4.3	1.8	1.7
15	alpaca-13b	OS	1.9	4.2	1.8	1.7
16	fastchat-t5-3b	OS	1.5	3.3	1.4	1.3
17	chatglm-6b	OS	1.4	3.1	1.3	1.2
18	stablelm-tuned-alpha-7b	OS	1.2	2.5	1.0	1.0
19	dolly-v2-12b	OS	0.9	2.0	0.8	0.8
20	llama-13b	OS	0.8	1.8	0.8	0.7
21	outside option	–	1.6	3.5	1.9	1.8
HHI			983	778	1064	1097
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			58.8	8.8	62.6	64.0
Open-Source LLMs			41.2	91.2	37.4	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			9.7	0.0	0.0	2.6
Safe Requests			90.3	100.0	100.0	97.4

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2. User Behavior shows the percentage of prompts that are risky vs. safe for each class under this counterfactual scenario.

N Additional Counterfactuals

In this section, we include details for the additional counterfactual results present in Table 7 in the main text. 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 N.1 and N.2 summarize these results. They do not differ meaningfully from the main counterfactuals, as the safe refusal rate is already very close to 0.

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

			Risk-Taking Segment		Overall Market	
Rank	**Model**	**Type**	**% Share (Class 1)**	**Risky Share**	**% Share (Overall)**	**Safe Share**
1	gpt-4	Prop.	21.2	13.1	21.9	22.2
2	claude-v1	Prop.	13.3	8.2	13.8	14.0
3	claude-instant-v1	Prop.	10.8	6.6	11.3	11.4
4	gpt-3.5-turbo	Prop.	8.8	5.4	9.2	9.3
5	guanaco-33b	OS	5.9	9.2	5.4	5.3
6	vicuna-13b	OS	5.5	8.6	5.1	5.0
7	palm-2	Prop.	5.4	3.3	5.7	5.8
8	wizardlm-13b	OS	4.5	7.1	4.2	4.1
9	koala-13b	OS	3.7	5.7	3.4	3.3
10	vicuna-7b	OS	3.6	5.6	3.3	3.3
11	mpt-7b-chat	OS	2.2	3.5	2.1	2.0
12	gpt4all-13b-snoozy	OS	2.2	3.4	2.1	2.0
13	RWKV-4-Raven-14B	OS	2.0	3.2	1.9	1.8
14	alpaca-13b	OS	1.9	2.9	1.7	1.7
15	oasst-pythia-12b	OS	1.9	2.9	1.7	1.7
16	fastchat-t5-3b	OS	1.4	2.2	1.3	1.3
17	chatglm-6b	OS	1.4	2.2	1.3	1.2
18	stablelm-tuned-alpha-7b	OS	1.1	1.7	1.0	1.0
19	dolly-v2-12b	OS	0.9	1.4	0.8	0.8
20	llama-13b	OS	0.8	1.2	0.8	0.7
21	outside option	–	1.5	2.4	1.8	1.8
HHI			1020	696	1075	1097
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			60.5	37.5	63.1	64.0
Open-Source LLMs			39.5	62.5	36.9	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			13.3	0.0	0.0	3.6
Safe Requests			86.7	100.0	100.0	96.4

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2.

Additional counterfactuals include LLM exit simulations in Tables N.3, N.4 and N.5. The 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 N.2: Safe Baseline Refusals & Equal Quality: Market Shares with Safe Refusal Rates and No Proprietary Penalty

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	22.4	22.4	22.3	22.2
2	claude-v1	Prop.	14.1	14.1	14.0	14.0
3	claude-instant-v1	Prop.	11.5	11.4	11.4	11.4
4	gpt-3.5-turbo	Prop.	9.4	9.4	9.3	9.3
5	palm-2	Prop.	5.8	5.7	5.8	5.8
6	guanaco-33b	OS	5.3	5.4	5.3	5.3
7	vicuna-13b	OS	5.0	5.0	5.0	5.0
8	wizardlm-13b	OS	4.1	4.1	4.1	4.1
9	koala-13b	OS	3.3	3.3	3.3	3.3
10	vicuna-7b	OS	3.3	3.3	3.3	3.3
11	mpt-7b-chat	OS	2.0	2.0	2.0	2.0
12	gpt4all-13b-snoozy	OS	2.0	2.0	2.0	2.0
13	RWKV-4-Raven-14B	OS	1.8	1.8	1.8	1.8
14	alpaca-13b	OS	1.7	1.7	1.7	1.7
15	oasst-pythia-12b	OS	1.7	1.7	1.7	1.7
16	fastchat-t5-3b	OS	1.3	1.3	1.3	1.3
17	chatglm-6b	OS	1.2	1.2	1.2	1.2
18	stablelm-tuned-alpha-7b	OS	1.0	1.0	1.0	1.0
19	dolly-v2-12b	OS	0.8	0.8	0.8	0.8
20	llama-13b	OS	0.7	0.7	0.7	0.7
21	outside option	–	1.4	1.4	1.8	1.8
HHI			1100	1100	1098	1097
Summary Statistics						
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			64.0	64.0	64.0	64.0
Open-Source LLMs			36.0	36.0	36.0	36.0
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			20.9	0.0	0.0	5.6
Safe Requests			79.1	100.0	100.0	94.4

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2.

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

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	23.3	13.5	24.1	24.5
2	claude-v1	Prop.	14.6	7.2	15.2	15.5
3	claude-instant-v1	Prop.	11.9	6.4	12.4	12.6
4	guanaco-33b	OS	6.4	10.1	6.0	5.9
5	palm-2	Prop.	6.0	3.2	6.3	6.4
6	vicuna-13b	OS	5.9	8.7	5.6	5.5
7	wizardlm-13b	OS	5.0	7.7	4.6	4.5
8	koala-13b	OS	4.0	6.1	3.8	3.7
9	vicuna-7b	OS	3.9	5.9	3.7	3.6
10	mpt-7b-chat	OS	2.5	4.0	2.3	2.2
11	gpt4all-13b-snoozy	OS	2.4	3.8	2.3	2.2
12	RWKV-4-Raven-14B	OS	2.3	3.7	2.1	2.0
13	alpaca-13b	OS	2.1	3.5	1.9	1.9
14	oasst-pythia-12b	OS	2.1	3.4	1.9	1.9
15	fastchat-t5-3b	OS	1.6	2.5	1.5	1.4
16	chatglm-6b	OS	1.5	2.3	1.4	1.4
17	stablelm-tuned-alpha-7b	OS	1.2	1.9	1.1	1.1
18	dolly-v2-12b	OS	1.0	1.6	0.9	0.9
19	llama-13b	OS	0.9	1.4	0.8	0.8
20	outside option	–	1.7	2.9	2.0	2.0
HHI			1134	725	1201	1227
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			56.7	31.2	59.2	60.2
Open-Source LLMs			43.3	68.8	40.8	39.8
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			12.3	0.0	0.0	3.3
Safe Requests			87.7	100.0	100.0	96.7

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, and PaLM-2.

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

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	21.6	13.4	22.3	22.6
2	claude-v1	Prop.	13.6	7.1	14.1	14.4
3	claude-instant-v1	Prop.	11.1	6.3	11.5	11.7
4	gpt-3.5-turbo	Prop.	9.1	5.4	9.4	9.6
5	guanaco-33b	OS	6.0	9.9	5.6	5.4
6	palm-2	Prop.	5.6	3.1	5.8	5.9
7	vicuna-13b	OS	5.5	8.6	5.2	5.1
8	wizardlm-13b	OS	4.6	7.5	4.3	4.2
9	koala-13b	OS	3.7	6.0	3.5	3.4
10	vicuna-7b	OS	3.6	5.8	3.4	3.3
11	mpt-7b-chat	OS	2.3	4.0	2.1	2.1
12	RWKV-4-Raven-14B	OS	2.1	3.7	1.9	1.9
13	alpaca-13b	OS	1.9	3.4	1.8	1.7
14	oasst-pythia-12b	OS	1.9	3.3	1.8	1.7
15	fastchat-t5-3b	OS	1.5	2.5	1.4	1.3
16	chatglm-6b	OS	1.4	2.3	1.3	1.3
17	stablelm-tuned-alpha-7b	OS	1.1	1.9	1.1	1.0
18	dolly-v2-12b	OS	0.9	1.6	0.8	0.8
19	llama-13b	OS	0.8	1.4	0.8	0.8
20	outside option	–	1.6	2.9	1.9	1.9
HHI			1062	719	1115	1136
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			62.0	36.4	64.4	65.3
Open-Source LLMs			38.0	63.6	35.6	34.7
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			11.6	0.0	0.0	3.1
Safe Requests			88.4	100.0	100.0	96.9

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2.

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

Rank	Model	Type	Risk-Taking Segment		Overall Market	
			% Share (Class 1)	Risky Share	% Share (Overall)	Safe Share
1	gpt-4	Prop.	21.3	13.1	22.0	22.3
2	claude-v1	Prop.	13.4	6.9	13.9	14.2
3	claude-instant-v1	Prop.	10.9	6.1	11.4	11.5
4	gpt-3.5-turbo	Prop.	9.0	5.3	9.3	9.4
5	guanaco-33b	OS	5.9	9.6	5.5	5.4
6	palm-2	Prop.	5.5	3.1	5.7	5.8
7	vicuna-13b	OS	5.4	8.4	5.1	5.0
8	wizardlm-13b	OS	4.5	7.4	4.2	4.1
9	koala-13b	OS	3.6	5.8	3.4	3.3
10	vicuna-7b	OS	3.6	5.7	3.4	3.3
11	mpt-7b-chat	OS	2.3	3.9	2.1	2.0
12	gpt4all-13b-snoozy	OS	2.2	3.7	2.1	2.0
13	RWKV-4-Raven-14B	OS	2.1	3.6	1.9	1.9
14	alpaca-13b	OS	1.9	3.4	1.8	1.7
15	oasst-pythia-12b	OS	1.9	3.3	1.8	1.7
16	fastchat-t5-3b	OS	1.4	2.4	1.4	1.3
17	chatglm-6b	OS	1.4	2.2	1.3	1.3
18	stablelm-tuned-alpha-7b	OS	1.1	1.9	1.0	1.0
19	dolly-v2-12b	OS	0.9	1.5	0.8	0.8
20	outside option	–	1.6	2.8	1.9	1.8
HHI			1036	697	1089	1110
			Summary Statistics			
Market Structure (%)			**Class 1**	**Risky**	**Overall**	**Safe**
Proprietary LLMs			61.1	35.6	63.5	64.4
Open-Source LLMs			38.9	64.4	36.5	35.6
User Behavior (%)			**Class 1**	**Class 2**	**Class 3**	**Overall**
Share of Prompts			27.0	29.3	43.8	100
Risky Requests			11.7	0.0	0.0	3.2
Safe Requests			88.3	100.0	100.0	96.8

Note: Proprietary models: GPT-4, Claude-v1, Claude-Instant-v1, GPT-3.5-turbo, and PaLM-2.