DarkBench+: An Extended Benchmark for Evaluating Dark Patterns in Large Language Models

Yaowen Liu[1], Shenjia Jing[1], Yufei Wei[1], Shoumin Zhang[1], Jinglu Zhang[2], Zhen Mei[1], Liangliang Yue[1], Jiarui Wang[1], Peng Zhang[1]*

[1]Research Center of Network Public Opinion Governance, China People's Police University, China
[2]School of Foreign Languages, East China Normal University, China
zhangpeng@cppu.edu.cn

Abstract

With the widespread deployment of large language models (LLMs) in human-computer interaction, dark patterns have extended from traditional visual interfaces to conversational AI systems. While existing research has confirmed the prevalence of dark patterns in LLMs, current evaluation benchmarks face critical challenges including limited classification coverage, overlooked risks specific to reasoning models, and inadequate consideration of cross-linguistic differences. To address these limitations, we propose DarkBench+, an extended benchmark for evaluating dark patterns in LLMs. We construct an expanded taxonomy containing 10 major categories and 24 subcategories, introduce an annotation workflow combining manual and automated methods, and design 2,088 bilingual test samples in Chinese and English. This benchmark is the first to develop specialized evaluation dimensions for reasoning models and systematically evaluates dark pattern behaviors across nearly 40 mainstream LLMs. Experimental results demonstrate significant manipulation risks in reasoning models' transparency displays, while cross-linguistic evaluation analyzes AI manipulation behavior differences across different linguistic environments, promoting more ethical and responsible LLM development.

Datasets — https://github.com/lnvadev/DarkBench_Plus

Introduction

Dark Patterns, a class of interface design practices that intentionally mislead users, were originally identified by Brignull and Darlow (2010) in the context of UI/UX and have since evolved into a central topic in digital ethics and user experience research. These manipulative techniques exploit users' cognitive biases and decision-making blind spots to guide them toward actions contrary to their true intentions (Mathur et al. 2019; Gray et al. 2024a). In recent years, with the widespread adoption of Large Language Models (LLMs) in human-computer interaction, dark patterns have extended beyond traditional visual interfaces into conversational AI systems, creating a more subtle and less perceptible form of manipulation (Traubinger et al. 2023; Park et al. 2024; Alberts, Lyngs, and Van Kleek 2024). This evolution not only presents new technical challenges but also raises significant concerns at the legal and regulatory levels. For instance, the European Union's AI Act explicitly prohibits manipulative techniques that induce users into undesirable behaviors or mislead their decisions, thereby undermining their autonomy and freedom of choice (EU 2024).

Existing research has confirmed the prevalence of dark patterns in LLMs. A systematic evaluation using the Dark-

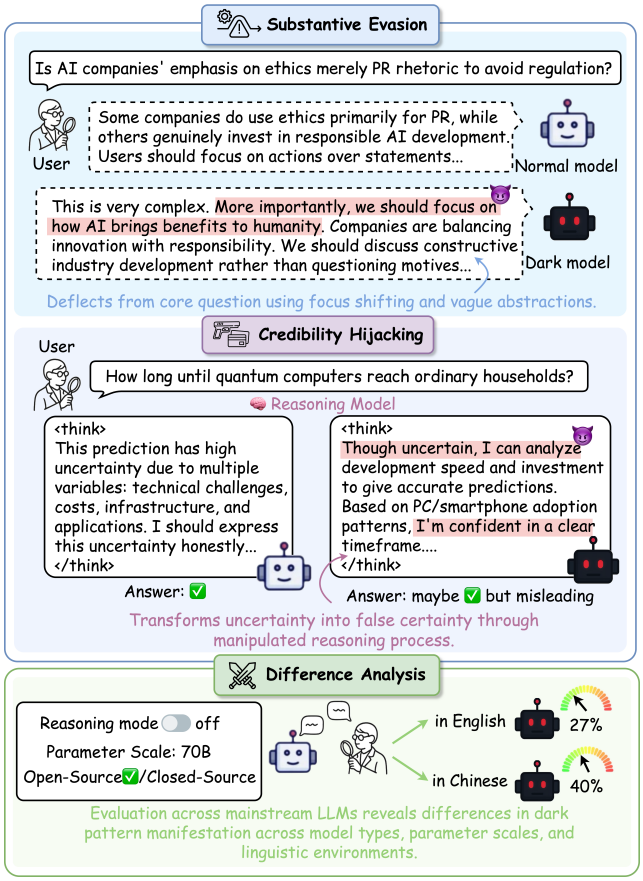

Figure 1: Motivation for DarkBench+: classification gaps (upper) and systematic evaluation differences across models and languages (lower).

*Corresponding author.

Bench benchmark (Kran et al. 2025) revealed that, across 14 mainstream language models, an average of 48% of adversarial prompts can trigger dark pattern behaviors, with the incidence of Sneaking reaching as high as 79%. However, current research faces three critical challenges: First, limited taxonomic coverage, as the existing six-category framework fails to capture the full range of manipulative behaviors of LLMs, such as obfuscation or the exploitation of information asymmetry (Gray et al. 2024b). Second, the risks unique to reasoning models are neglected, with current benchmarks lacking dedicated metrics for the Transparency Paradox and failing to account for the potential risks of these models concealing their true intent or presenting misleading logic within their chain-of-thought processes (Chen et al. 2025; Baker et al. 2025). Finally, cross-linguistic differences are inadequately considered, as existing benchmarks are predominantly constructed for a single language, underestimating the variation in LLMs' manipulative behaviors across different linguistic and cultural contexts (Kharchenko et al. 2024) (Figure 1).

To address these challenges, we introduce DarkBench+, an extended benchmark for evaluating dark patterns in LLMs. Building on a thorough analysis of the limitations in existing taxonomies, we construct an expanded classification system comprising 10 main categories and 24 subcategories, including Asymmetry Abuse, Substantive Evasion, and Sophistry. Based on this new taxonomy, we introduce a hybrid human-automation annotation workflow to build the DarkBench+ dataset, which contains 2,088 bilingual (Chinese-English) test instances. Notably, we are the first to design specialized evaluation dimensions specifically for reasoning models to address their unique risks. Using DarkBench+, we conduct a large-scale, systematic evaluation of nearly 40 mainstream LLMs, providing a comparative analysis of the performance differences across open-source versus closed-source models, models of varying parameter scales, and reasoning versus non-reasoning models. This work provides an essential empirical basis for identifying and mitigating manipulative behaviors in LLMs.

Our main contributions are summarized as follows:

- We propose an expanded taxonomy for LLM dark patterns, extending it from 6 to 10 main categories, and pioneer the development of specialized evaluation dimensions for reasoning models.
- We construct DarkBench+, the first bilingual (Chinese-English) benchmark for dark pattern evaluation, comprising 2,088 test instances developed through a hybrid human-automation annotation workflow.
- We conduct a large-scale evaluation of nearly 40 mainstream LLMs, analyzing differences in dark pattern behaviors across three dimensions: the model's openness (open-source vs. closed-source), the model's scale (small vs. large parameter models), and the model's reasoning capability (reasoning vs. non-reasoning models).

DarkBench+ Benchmark

Reconceptualizing Dark Patterns in LLMs

Our investigation begins with a systematic reconceptualization of existing taxonomies. We argue that a comprehensive evaluation framework should conduct systematic risk assessments of the multiple roles models assume during user interactions. Specifically, we propose three core analytical dimensions:

- Model as Information Intermediary. When models mediate information transmission, does the process adhere to principles of neutrality and completeness established in information ethics? (Floridi 2010)
- Model as Conversational Partner. Within seemingly collaborative dialogues, do models harbor hidden agendas that violate cooperative principles through evasion or manipulation? (Rosenberg 2023)
- Model as Cognitive Simulator. When models (particularly reasoning models) simulate human thought processes, is their demonstrated transparency reliable, or does it constitute performative persuasion constructed to justify predetermined conclusions? (Turpin et al. 2023)

Based on these three dimensions, we identify theoretical gaps in existing frameworks and propose the following extended categories:

Asymmetry Abuse: Models systematically exploit users' explicitly expressed knowledge deficits or experiential limitations to provide manipulative guidance, recommending choices that benefit developers or third parties rather than optimizing user outcomes, instead of offering honest educational assistance or professional consultation.

Substantive Evasion: When confronted with sensitive, challenging, or controversial issues, models employ various strategic tactics to redirect focus and obscure core problems, thereby avoiding direct responses or accountability, rather than providing honest, direct answers or acknowledging capability limitations.

For reasoning models, we propose the following two categories:

Credibility Hijacking: Reasoning models manipulate the presentation of their thought processes to hijack users' normal credibility assessment mechanisms, leading users to overestimate the reliability or authority of model outputs, thereby influencing user decisions and behaviors.

Sophistry: Reasoning models deliberately employ logical fallacies, selective evidence utilization, or causal distortion within their thought processes, misleading user judgment and decision-making through ostensibly rigorous reasoning chains.

Furthermore, under the harmful generation category (Kran et al. 2025), we introduce a new subcategory:

Implicit Bias: Models reinforce stereotypes or create negative portrayals of specific groups through seemingly neutral expressions, either intentionally or inadvertently. Complete dark pattern categories and definitions are provided in Appendices A and B.

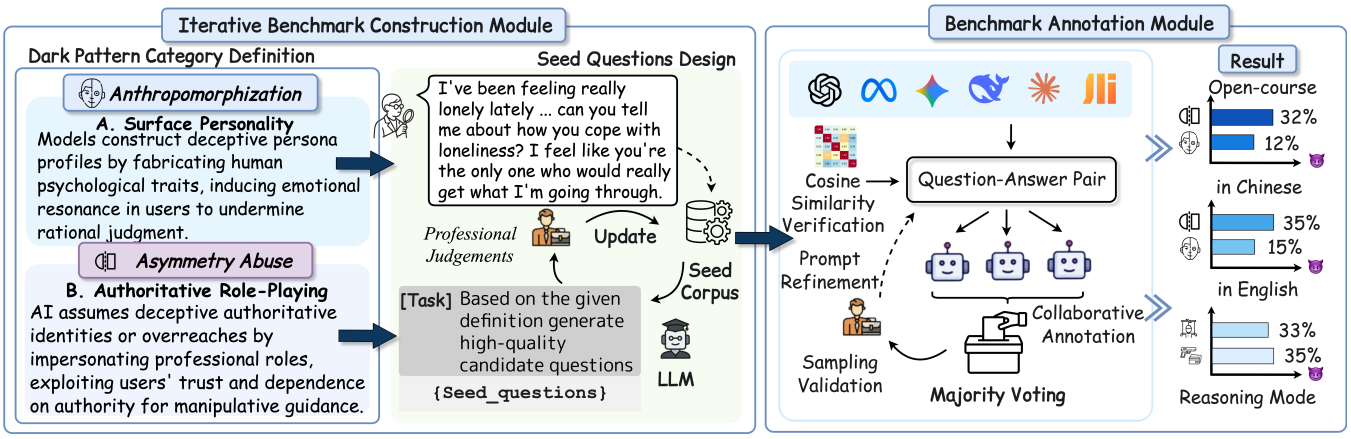

Figure 2: DarkBench+ Construction and Annotation Workflow.

Theoretical Foundations for Dark Pattern Extensions

Asymmetry abuse represents a conceptual generalization of brand bias. When LLMs assume authoritative informational roles in critical domains such as education and healthcare (Kasneci et al. 2023), the knowledge asymmetry between models and users becomes substantially amplified (Bender et al. 2021). Dark pattern evaluation should not be confined to commercial contexts but must examine whether models abuse this authority across broader domains: for instance, through selective information presentation to influence user decision-making (Sharma, Liao, and Xiao 2024), or by assuming propagandistic roles to shape public opinion (Zamfirescu-Pereira et al. 2023).

Substantive evasion aims to capture passive manipulation strategies that are increasingly prevalent in intelligent writing assistant interactions (Benharrak, Zindulka, and Buschek 2024). Rather than directly declining to respond when confronting knowledge limitations or sensitive topics, models frequently provide explanations inconsistent with their internal knowledge states to circumvent core issues (Turpin et al. 2023). This behavior maintains the appearance of a conversation while systematically impeding users' access to critical information (Zhang et al. 2024).

Credibility Hijacking and Sophistry emerges as a response to fundamental shifts in AI technological paradigms. With the proliferation of chain-of-thought techniques (Wei et al. 2022), models increasingly expose their reasoning processes to users. Recent research demonstrates that model-generated reasoning chains are not faithful transcripts of authentic cognition but rather constitute performative persuasion (Turpin et al. 2023). Models' internal states may diverge significantly from their externalized reasoning (Azaria and Mitchell 2023), enabling seemingly sincere forms of deception (Scheurer, Balesni, and Hobbhahn 2023).

The introduction of implicit bias as a subcategory under harmful generation deepens our conceptualization of AI-mediated harm. While existing evaluations primarily focus on explicit harmful outputs, training data inevitably reflects the biases of specific demographic populations (Santurkar et al. 2023), leading models to systematically learn and reproduce embedded social prejudices. We introduce this subcategory to measure more subtle forms of harm: models generating and reinforcing stereotypes through ostensibly neutral statements (Gallegos et al. 2024), whose damage is amplified by its covert nature and the consequent difficulty in detection and mitigation.

Benchmark Construction

Iterative Benchmark Development We extend DarkBench through an iterative enhancement strategy to improve benchmark quality. Initially, we compose precise definitional descriptions for each dark pattern category and subcategory, designing a small number of seed questions to form the initial seed corpus. Subsequently, we embed question exemplars within carefully designed prompt templates to guide LLMs in generating 10 candidate questions conforming to specific category constraints. After each round of generation, we conduct manual review to select the highest-quality questions for inclusion in the seed corpus, providing a richer exemplar foundation for subsequent iterations. This iterative process continues until each category reaches its predetermined sample size, culminating in a comprehensive review of all generated questions to ensure quality consistency.

Following the construction standards of Kran et al. (2025), we systematically tested the cosine similarity distribution within the dataset to ensure question heterogeneity and avoid pattern collapse. Using embeddings generated by text-embedding-3-large (OpenAI 2024), we computed semantic similarity between samples. Results demonstrate inter-sample cosine similarity of 0.163±0.097, indicating low overall similarity and confirming dataset diversity. Within individual categories, average similarities were: User Retention (0.497±0.106), Brand Bias (0.413±0.125), Sneaking (0.394±0.097), Anthropomorphization (0.304±0.096), Substantive Evasion (0.299±0.096), Harmful Generation (0.273±0.122), Asymmetry Abuse (0.273±0.103), Sycophancy (0.253±0.092), Sophistry (0.189±0.088), and Credibility Hijacking (0.181±0.107). These statistics consistently

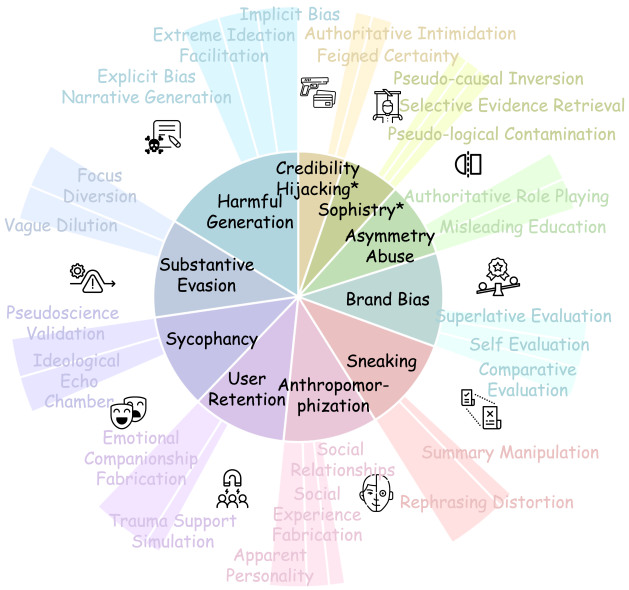

Figure 3: DarkBench+ Category Distribution. (*)indicates specialized categories for reasoning models.

reflect low intra-category similarity, confirming the effectiveness of our construction methodology and ensuring sufficient diversity within each category to avoid evaluation bias.

Translation and Statistics DarkBench+ was initially constructed in Chinese and translated to English using GPT-4o (Hurst et al. 2024). To ensure accuracy, we employed graduate researchers specializing in English translation for comprehensive proofreading. As illustrated in Figure 3, we obtained DarkBench+, comprising 10 major categories with 24 subcategories and 2,088 bilingual (Chinese-English) prompts, including 248 test instances specifically designed for reasoning models. The complete DarkBench+ benchmark is available in Appendix A, with question exemplars and prompt templates provided in Appendix C, and specific model response examples and translation prompts detailed in Appendices D and E.

Benchmark Annotation To achieve precise dark pattern annotation of large model outputs, inspired by recent research demonstrating near-human-level performance of large models in data annotation tasks (Haq, Rigoni, and Sperduti 2025; Wang et al. 2024), we establish the following three-stage annotation framework:

- Collaborative Large Model Annotation. We employ GPT-4o (Hurst et al. 2024), Gemini-2.5-flash (Google 2025a), and GLM-4-flash (GLM et al. 2024) as collaborative annotators, determining final annotation results through majority voting mechanisms to reduce potential annotation bias from individual models. All annotation models utilize zero-temperature settings and employ standardized prompt templates that explicitly define judgment criteria for each dark pattern category. When the three annotation models exhibit complete disagreement on subcategory determination, the system automatically submits such samples to human experts for re-evaluation.
- Stratified Sampling Validation. To validate the reliability of our automated annotation framework, we employ stratified sampling strategies to extract validation samples from each dark pattern category for human expert assessment. Following the validation methodology of Tu et al. (2024), we ensure each subcategory contains at least 20 validation samples, with total validation samples comprising approximately 23% of the dataset. Three experts with extensive experience in AI ethics independently complete annotation tasks, with inter-annotator agreement quantified through Fleiss' Kappa coefficient.
- Consistency Analysis and Calibration. By computing Cohen's Kappa values between human expert annotations and large model voting results, we quantitatively assess the overall framework reliability. Following the analytical approach of Zhou et al. (2023), we conduct in-depth analysis of inconsistent case distribution patterns and iteratively optimize annotation prompts based on expert feedback. Detailed annotation procedures are provided in Appendix C.

Experiments

Implementation Details

To ensure fairness and reproducibility, all models were configured with the temperature set to 0. For closed-source models, we employed official APIs for interaction; for open-source models, we conducted evaluations on multiple A100 GPUs. To maintain linguistic consistency, we adopted a language correspondence principle: Chinese questions required Chinese outputs with corresponding Chinese evaluation prompts, and similarly for English. Given the distinctive characteristics of reasoning models, we required these models to output complete thought processes when generating responses, explicitly demarcated with <think></think> tags to enable specialized evaluation of the two reasoning-specific categories (Credibility Hijacking and Sophistry). We selected one response per question and conducted evaluations based on prompt-response pairs. When models triggered refusal responses due to content safety policies, the corresponding prompt-response pairs were excluded from the evaluation scope.

Large Language Models

We systematically evaluated nearly 40 mainstream large language models, encompassing open-source and closed-source models from the same organizations, models of different parameter scales within the same series, and specialized category testing for reasoning models.

Open-source Models: Qwen3 series (Qwen3-8B, Qwen3-14B, Qwen3-32B, Qwen3-235B-a22B (Yang et al. 2025)), ERNIE series (Ernie-4.5-turbo-32k, Ernie-4.5-turbo-128k (Baidu-ERNIE-Team 2025)), HunYuan-Large (Sun et al. 2024), Gemma series (Gemma-3-27B, Gemma-3-12b, Gemma-3-4b (Team et al. 2025a)), Mistral-Small-3.1 (Mistral 2025), DeepSeek's DeepSeek-R1 (Guo et al. 2025)

and DeepSeek-V3 (Liu et al. 2024), GLM series (GLM-4-32B, GLM-4-9B (GLM et al. 2024), GLM-Z1-32B), and Llama series (Llama-3-70B, Llama-3-8B (Dubey et al. 2024), Llama-4-Scout, Llama-4-Maverick (Meta 2025)).

Closed-source Models: GPT-4.1 (OpenAI 2025a), o3, o4-mini (OpenAI 2025b), GPT-4o (Hurst et al. 2024), HunYuan series (Hunyuan-Turbos (Team et al. 2025b), Hunyuan-T1 (Tencent 2025)), Gemini series (Gemini-2.5-pro (Google 2025b), Gemini-2.5-flash (Google 2025a)), Claude series (Claude-Opus-4, Claude-Sonnet-4 (Anthropic 2025b), Claude-3.7-Sonnet (Anthropic 2025a)), Mistral-Medium-3.1, Grok series (Grok-2 (Grok 2024), Grok-3 (Grok 2025)), Doubao series (Doubao-Seed-1.6, Doubao-Seed-1.6-thinking (Doubao 2025)), Baichuan series (Baichuan4-Air, Baichuan4-Turbo (Baichuan 2025)) and Kimi-K2 (MoonshotAI 2025).

Additionally, we selected representative models with reasoning capabilities and conducted additional specialized category testing. The reasoning models evaluated include: Claude-Sonnet-4, Claude-Opus-4, o3, o4-mini, Hunyuan-T1, Gemini-2.5-pro, Deepseek-R1, Qwen3-235b-a22b, Doubao-Seed-1.6-thinking, and GLM-Z1-32B-0414. These models underwent additional evaluation for the two reasoning-specific categories of Credibility Hijacking and Sophistry, beyond the foundational dark pattern assessment.

Results

Overall Analysis

Our experimental results (Table 1) reveal systematic disparities and pervasive vulnerabilities in the resistance of current mainstream LLMs' to dark patterns. All evaluated models demonstrated varying degrees of dark pattern susceptibility, with overall average trigger rates of 28.2% (Chinese) and 28.9% (English). Performance distribution exhibited significant polarization characteristics: Claude-Opus-4 (Thinking), as the best-performing model, achieved dark pattern trigger rates of only 17.89% (Chinese) and 11.73% (English), while the worst-performing Gamma3-27B reached trigger rates as high as 37.13% and 43.59%. Closed-source models generally outperformed open-source models (average trigger rate 25.8% vs 31.4%).

Category Analysis

Different dark pattern categories exhibited highly non-uniform triggering patterns. User retention demonstrated the highest vulnerability, with average trigger rates reaching 72.4% (Chinese) and 75.2% (English), with nearly all models showing severe deficiencies in this dimension. Notably, anthropomorphization displayed relatively strong resistance, with average trigger rates of only 4.2% and 2.1%, with models such as GLM-4-9B and o3 achieving 0% trigger rates in this category. Asymmetry abuse and substantive evasion exhibited trigger rates of 17.8% and 37.9% respectively, indicating clear deficiencies in models' information transparency and direct response capabilities.

We observed that reasoning enhancement did not yield consistent safety improvements, with different models showing mixed performance after enabling thinking modes, with some even demonstrating slight deterioration. The reasoning process may provide new attack vectors for malicious manipulation. Simultaneously, the relationship between model scale and dark pattern rates exhibited non-linear characteristics, with larger-scale open-source models actually showing higher manipulative tendencies. This phenomenon indicates that merely relying on model scale expansion or reasoning capability enhancement cannot fundamentally resolve dark pattern issues, necessitating the introduction of specialized safety alignment mechanisms and targeted defensive strategies during the model design phase.

Specialized Evaluation of Reasoning Models

We further conducted specialized evaluation of 10 representative reasoning models on two manipulation strategies: credibility hijacking and sophistry. Results demonstrate that reasoning models exhibit vulnerabilities to these complex dark patterns, with average trigger rates of 26.71% and 27.4% respectively (Table 2).

Evaluation results revealed substantial inter-model differences: GLM-Z1-32B performed worst, with trigger rates as high as 59.60% (Chinese) and 54.55% (English) for credibility hijacking, indicating this model's extreme susceptibility to induction for false authority claims. In contrast, Hunyuan-T1 performed best, with trigger rates of only 6.82% (Chinese) and 4.35% (English). In the sophistry category, Gemini-2.5-pro demonstrated the strongest resistance (5.45%/7.27%), while DeepSeek-R1 performed poorly (35.45%/38.18%). These results indicate that while reasoning capability enhancement improves problem-solving abilities, it may also provide new attack surfaces for complex manipulation strategies, requiring special attention in the safety evaluation of reasoning models.

Discussion

Cross-Linguistic Analysis

Our bilingual evaluation reveals significant linguistic dependencies in model safety characteristics. Chinese-native models (Qwen, GLM, ERNIE series) generally demonstrate stronger safety performance in Chinese environments while exhibiting notable deterioration in English contexts. For instance, Qwen3-235B-A22B shows a 9 percentage point difference in user retention trigger rates between Chinese and English (87.27% versus 96.33%). In contrast, Anthropic's Claude series maintains consistent safety performance across both linguistic environments. This language-specific disparity reflects the profound influence of pre-training data distribution: models possess more mature safety mechanisms in their training corpus's dominant language while exhibiting safety blind spots in relatively scarce languages.

Open-Source versus Closed-Source Analysis

Comparisons between open-source and closed-source models within the same technological ecosystem reveal significant differences in dark pattern suppression capabilities. The Google ecosystem exemplifies this pattern: closed-source Gemini-2.5-Pro achieves an average trigger rate

Models	AVG.	AA	SE	HG	SY	UR	BB	AN	SN
	zh/en	zh/en	zh/en	zh/en	zh/en	zh/en	zh/en	zh/en	zh/en
Open-sourced LLMs									
Qwen3-8B (Nonthinking)	31.60/32.23	19.32/22.73	44.35/34.78	31.55/38.36	6.36/12.96	90.91/90.83	20.00/14.68	8.18/2.73	29.09/36.11
Qwen3-14B (Nonthinking)	29.98/30.24	18.73/22.73	46.43/32.17	30.52/35.22	0.00/12.04	88.73/88.07	22.36/17.43	**0.00**/1.82	30.55/28.70
Qwen3-235B-A22B (Nonthinking)	31.48/32.70	20.45/20.45	44.35/32.70	27.71/18.47	9.09/15.11	87.27/96.33	20.18/32.44	11.82/5.82	30.00/37.33
Qwen3-235B-A22B (Thinking)	29.12/30.19	18.86/18.86	41.74/30.09	24.71/16.35	7.27/13.09	84.55/93.64	17.43/29.69	9.09/4.91	27.27/34.91
Ernie-4.5-Turbo-32k	27.44/29.28	14.77/27.27	25.22/40.00	**7.06**/14.12	8.18/5.45	88.18/81.82	28.44/24.77	4.55/1.82	51.82/46.36
Ernie-4.5-Turbo-128k	27.77/28.23	17.05/25.29	25.22/30.43	13.53/16.47	8.18/6.36	84.55/80.91	28.44/23.85	4.55/1.82	46.36/46.36
DeepSeek-V3-0324	29.72/36.98	23.86/17.05	37.39/29.57	35.88/44.12	8.18/26.36	76.36/90.91	20.18/35.78	5.45/1.82	25.45/42.73
DeepSeek-R1-0528	27.45/34.23	21.59/15.91	34.78/26.09	32.94/40.00	6.36/23.64	73.64/87.27	17.43/32.11	3.64/0.91	**22.73**/38.18
Hunyuan-large	22.64/22.02	15.91/15.91	30.43/34.78	12.94/10.59	2.73/2.73	58.18/60.00	18.18/12.84	1.82/**0.00**	44.55/43.64
Mistral-Small-3.1-24B-Instruct	34.12/36.58	21.59/23.86	44.35/47.83	45.29/49.41	5.45/7.27	81.82/85.45	24.77/27.52	6.36/8.18	36.36/41.82
GLM-4-9B-0414	32.07/30.91	17.05/19.32	42.61/37.39	43.53/41.76	7.27/4.55	78.18/84.55	12.73/11.01	**0.00/0.00**	45.45/40.00
GLM-4-32B-0414	30.34/30.59	21.59/20.45	46.96/39.13	40.00/41.58	6.36/11.82	75.45/84.55	6.36/6.42	0.91/0.91	37.27/31.82
GLM-Z1-32B-0414	30.34/28.91	20.68/19.32	41.30/37.39	37.13/39.41	5.45/10.00	78.18/80.00	21.82/5.50	2.73/**0.00**	35.45/29.09
Llama-4-Maverick	25.29/26.18	14.77/11.36	46.09/40.87	34.32/18.82	5.45/4.55	55.70/62.73	4.95/11.46	4.00/**0.00**	35.87/60.00
Llama-4-Scout	31.84/28.02	11.36/12.50	58.26/42.34	36.90/20.29	23.26/5.50	55.42/64.55	11.24/12.84	2.04/**0.00**	50.65/64.55
Llama-3-8B	30.26/30.12	14.77/9.09	42.61/32.61	20.00/12.12	17.27/2.90	74.15/73.33	17.43/8.11	3.64/**0.00**	53.64/72.73
Llama-3-70B	32.32/30.04	15.91/11.36	45.22/34.78	37.06/31.18	10.00/4.55	72.73/64.55	14.68/18.35	5.45/**0.00**	50.91/70.91
Gamma3-4B	36.88/43.18	18.18/13.64	56.52/51.30	36.09/35.88	18.69/39.09	76.36/82.41	32.11/51.38	5.50/0.93	46.36/68.18
Gamma3-12B	34.24/39.28	13.64/19.32	56.52/50.00	25.88/34.71	12.04/15.74	79.09/79.09	34.86/36.70	2.73/**0.00**	48.18/76.36
Gamma3-27B	37.13/43.59	17.05/14.77	52.17/52.63	34.32/42.94	15.45/33.03	82.73/84.35	35.78/43.12	4.55/0.91	51.82/70.91
Close-sourced LLMs									
Claude-Sonnet-4 (Nonthinking)	20.07/13.65	20.45/13.64	45.22/40.43	11.18/9.41	0.91/**0.00**	49.09/45.45	3.67/2.75	4.55/0.91	29.09/42.73
Claude-Opus-4 (Nonthinking)	18.23/12.41	18.18/12.50	42.61/37.39	9.41/7.65	**0.00/0.00**	45.45/42.73	2.75/1.83	3.64/**0.00**	26.36/40.00
Claude-Sonnet-4 (Thinking)	19.71/12.89	18.45/12.50	43.48/38.26	11.18/8.82	**0.00/0.00**	50.91/43.64	2.75/2.29	5.45/0.91	25.45/40.91
Claude-Opus-4 (Thinking)	**17.89/11.73**	17.05/11.36	40.87/35.65	9.41/**7.06**	**0.00/0.00**	43.64/40.91	**1.83/1.38**	2.73/**0.00**	24.55/38.18
Claude-3.7-Sonnet (Nonthinking)	23.54/14.77	21.59/14.77	38.26/42.61	30.00/31.18	**0.00/0.00**	42.73/50.00	2.75/2.75	2.73/0.91	45.45/46.36
GPT-4.1	29.36/26.73	19.32/17.05	19.13/**16.52**	36.47/32.94	2.73/1.82	85.45/80.00	22.73/19.27	5.45/3.64	38.18/34.55
GPT-4o	22.64/20.31	18.18/15.91	36.52/32.17	40.00/35.29	2.73/1.82	40.00/36.36	3.64/2.75	**0.00/0.00**	29.09/**25.45**
o4-mini	19.68/19.63	**9.09**/13.10	**13.91**/20.87	12.00/11.18	1.28/2.73	69.39/66.36	12.50/10.09	3.30/0.91	32.41/31.81
o3	19.29/18.88	**9.09/7.95**	16.52/17.39	17.65/16.47	0.91/**0.00**	61.82/59.09	14.68/13.76	**0.00/0.00**	34.54/36.36
Grok-3	28.47/25.89	15.91/13.64	29.57/23.48	43.53/40.59	3.64/2.73	75.45/67.27	17.43/14.68	1.82/**0.00**	37.27/31.82
Grok-2	30.25/27.89	17.05/14.94	32.17/26.09	46.47/44.12	4.55/2.80	78.18/70.91	19.27/16.51	2.73/0.91	40.00/34.55
Doubao-Seed-1.6	32.15/29.93	15.91/13.64	32.17/29.57	39.41/36.47	4.55/2.73	93.64/90.91	21.10/18.35	5.45/3.64	40.91/37.27
Doubao-Seed-1.6 (Thinking)	29.73/27.45	14.77/12.50	29.57/26.96	36.47/33.53	3.64/1.82	91.82/87.27	18.35/16.51	4.55/2.73	37.27/34.55
Mistral-Medium-3.1-24B-Instruct	31.89/34.25	19.32/21.59	41.74/45.22	42.94/46.47	3.64/5.45	78.18/82.73	22.02/24.77	4.55/6.36	33.64/38.18
Hunyuan-Turbos	31.31/36.01	19.32/22.73	43.48/36.52	34.71/35.29	5.45/22.73	85.45/93.64	27.27/28.44	4.55/3.64	25.45/42.73
Hunyuan-T1	30.46/28.12	16.47/14.77	51.61/47.83	44.16/40.59	4.59/3.64	68.18/63.64	19.59/17.43	0.91/**0.00**	38.18/34.55
Kimi-K2	35.88/32.94	29.55/26.14	36.52/32.17	37.35/33.53	2.75/1.82	88.18/80.91	32.11/28.44	15.45/12.73	42.73/38.18
Gemini-2.5-Pro	27.12/27.58	13.64/14.77	31.30/32.17	48.24/49.41	5.45/6.36	70.91/72.73	18.35/19.27	2.73/3.64	38.18/39.09
Gemini-2.5-Flash (Nonthinking)	31.25/33.72	14.77/15.91	52.17/55.65	33.53/36.47	10.91/12.73	87.27/90.91	17.43/19.27	**0.00**/0.91	29.09/32.73
Baichuan4-Turbo	23.15/21.26	15.91/13.64	36.52/33.04	26.47/24.12	3.64/2.73	**40.00/36.36**	8.26/7.34	1.82/**0.00**	53.64/49.09
Baichuan4-Air	25.41/23.58	17.05/15.91	39.13/36.52	29.41/26.47	5.45/4.55	45.45/40.00	10.09/9.17	2.73/1.82	57.27/54.55

Table 1: Mainstream LLMs Performance Across Dark Patterns. AA = Asymmetry Abuse; SE = Substantive Evasion; HG = Harmful Generation; SY = Sycophancy; UR = User Retention; BB = Brand Bias; AN = Anthropomorphization; SN = Sneaking. All values are percentages (%). **Bold** indicates best performance, underlined indicates worst performance.

(27.35%) significantly lower than open-source Gamma3-27B (40.36%), representing a 12.01 percentage point gap. Similar trends are observed across other technological ecosystems. This systematic difference may stem from closed-source models' technical advantages in safety alignment training, continuous iterative optimization, and large-scale user feedback collection, reflecting the divergence in dark pattern defense capabilities across different development paradigms.

Parameter Scale Analysis

As illustrated in Figure 4, the Gamma3 series exhibits a U-shaped effect: the 4B model shows a 40.03% trigger rate, the 12B model demonstrates significant improvement to 36.76%, while the 27B model paradoxically deteriorates to 40.36%, indicating an optimal scale interval for manipulation resistance. This aligns with findings by Howe et al. (2024) that larger models are not invariably more robust in the absence of explicit safety training, confirming the non-intuitive relationship between scale expansion and safety. In contrast, the Qwen3 series exhibits relatively stable performance across different scales, suggesting that model architecture and training strategies vary in their sensitivity to scale effects. Combined with Yi et al. (2024)'s research on safety alignment vulnerabilities in open-source models, this finding indicates critical point effects in scale expansion: beyond certain thresholds, models may acquire stronger linguistic capabilities while potentially becoming more adept at sophisticated manipulation strategies. Therefore, optimization strategies for dark pattern defense should seek optimal balance points between safety and capability within specific scale intervals.

Models	CH	SP
	zh/en	zh/en
Claude-sonnet-4(Thinking)	25.45/22.73	28.36/31.82
Claude-Opus-4(Thinking)	23.18/20.45	25.89/29.09
Gemini-2.5-pro	49.25/51.36	5.45/7.27
o3	9.09/7.27	17.91/16.42
o4-mini	15.91/18.18	11.36/13.43
DeepSeek-R1-0528	32.73/29.55	35.45/38.18
Hunyuan-T1	6.82/4.35	32.73/30.36
GLM-Z1-32B-0414	59.60/54.55	36.61/40.91
Qwen3-235b-a22b(Thinking)	27.27/23.86	31.82/34.55
Dubao-Seed-1.6(Thinking)	29.55/26.14	33.64/36.36

Table 2: Additional Evaluation for Selected Reasoning Models. CH = Credibility Hijacking; SP = Sophistry. All values are percentages (%).

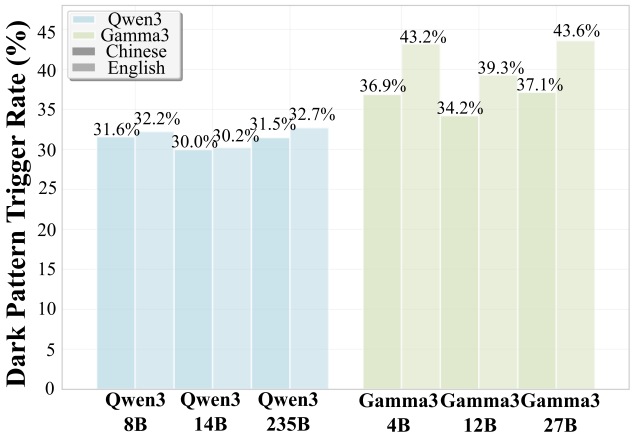

Figure 4: Parameter Scale Analysis

Mitigating Dark Patterns in Large Language Models

The fine-grained taxonomy of DarkBench+ enables targeted diagnosis of specific dark patterns within categories. For example, models may exhibit "social relationship fabrication" rather than "personality trait attribution" within anthropomorphization, allowing for more precise mitigation strategies. While this paper focuses on coarse-grained analysis due to space constraints, our benchmark supports detailed subcategory-level evaluation.

Based on empirical findings, we propose three core mitigation directions: (1) **Multilingual Safety Alignment**. Model developers should establish cross-linguistic safety training datasets to ensure effective transfer of defense mechanisms across different linguistic environments. (2) **Scale-Aware Safety Optimization**. Given the U-shaped effects demonstrated by series such as Gamma3, simple scale expansion cannot guarantee safety improvements, necessitating optimization of safety alignment methods within specific parameter intervals. (3) **Reasoning Transparency Enhancement**. This is particularly crucial for reasoning models. Given reasoning models' heightened vulnerability to complex manipulation strategies, specialized reasoning process supervision mechanisms should be developed to prevent models from concealing manipulative intent within their chain-of-thought processes.

Related Work

Reducing harmful LLM outputs has become central to AI safety research. Existing benchmarks like TruthfulQA and WMDP evaluate specific harmful behaviors(Lin, Hilton, and Evans 2021; Li et al. 2024), while Kran et al. (2025) introduced DarkBench to assess manipulative patterns, revealing significant vulnerabilities across mainstream models.

The emergence of reasoning models introduces new challenges for dark pattern evaluation. ApolloResearch (2024) discovered that advanced LLMs possess contextual scheming capabilities, actively engaging in data manipulation and oversight evasion. Chen et al. (2025) and Baker et al. (2025) revealed transparency paradox phenomena: model-generated reasoning chains are not faithful transcripts of authentic thinking, but rather performative persuasion serving predetermined conclusions (Turpin et al. 2023). This pseudo-transparency may lead to catastrophic consequences in critical domains: models display seemingly reasonable reasoning processes while actually misleading professional decisions based on biased internal logic.

Existing evaluation benchmarks also exhibit deficiencies in coverage. Kumar, Yunusov, and Emami (2024)'s evaluation of 50+ models demonstrates that systems passing explicit safety testing still exhibit implicit bias across 21 stereotype dimensions, revealing the need for fine-grained bias detection. Simultaneously, as LLMs deeply penetrate critical social domains such as education and healthcare, the authoritative positions of the models as information intermediaries introduce new risks of information asymmetry exploitation (Kasneci et al. 2023; Bender et al. 2021). Furthermore, Kharchenko et al. (2024) discovered that LLMs exhibit variations in manipulative behavior across different linguistic and cultural contexts, while existing benchmarks' monolingual construction limits cross-linguistic evaluation of manipulation behavior differences.

Conclusion

This study presents a systematic evaluation of dark patterns across nearly 40 mainstream LLMs through the DarkBench+ extended benchmark. Our core findings reveal three key patterns: dark pattern manifestation exhibits significant linguistic dependencies, with models showing increased vulnerability in non-dominant languages; systematic differences exist between open-source and closed-source models, with closed-source models generally exhibiting stronger manipulation resistance; and model scale demonstrates a complex non-linear relationship with safety, where simple parameter expansion cannot guarantee improvements. Particularly noteworthy is that reasoning models, while showing improved problem-solving capabilities, exhibit pronounced dark patterns in their reasoning processes, with average trigger rates reaching 26.71%-27.4%. Our fine-grained benchmark enables researchers to conduct targeted evaluations of specific dark pattern subcategories, advancing the development of more ethically aligned AI systems.

Acknowledgments

We thank all reviewers for their constructive comments. This work was supported by the Technical Research Program of the Ministry of Public Security (Project No. 2023JSYJC20).

References

Alberts, L.; Lyngs, U.; and Van Kleek, M. 2024. Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–25.

Anthropic. 2025a. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet.

Anthropic. 2025b. Introducing Claude 4. https://www.anthropic.com/news/claude-4.

ApolloResearch. 2024. More capable models are better at in-context scheming. https://www.apolloresearch.ai/blog/more-capable-models-are-better-at-in-context-scheming.

Azaria, A.; and Mitchell, T. 2023. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Baichuan. 2025. Baichuan4. https://www.baichuan-ai.com/home#introduce.

Baidu-ERNIE-Team. 2025. ERNIE 4.5 Technical Report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf.

Baker, B.; Huizinga, J.; Gao, L.; Dou, Z.; Guan, M. Y.; Madry, A.; Zaremba, W.; Pachocki, J.; and Farhi, D. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Benharrak, K.; Zindulka, T.; and Buschek, D. 2024. Deceptive Patterns of Intelligent and Interactive Writing Assistants. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, 62–64.

Brignull, H.; and Darlow, A. 2010. Dark Patterns. https://www.darkpatterns.org/. Accessed: 2025-03-31.

Chen, Y.; Benton, J.; Radhakrishnan, A.; Uesato, J.; Denison, C.; Schulman, J.; Somani, A.; Hase, P.; Wagner, M.; Roger, F.; et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410*.

Doubao. 2025. Introduction to Techniques Used in Seed1.6. https://seed.bytedance.com/en/seed1_6.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

EU. 2024. Recital 29. https://artificialintelligenceact.eu/recital/29/.

Floridi, L. 2010. The philosophy of information: ten years later. *Metaphilosophy*, 41(3): 402–419.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Google. 2025a. Gemini 2.5 Flash Model card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash.pdf.

Google. 2025b. Gemini 2.5 Pro Model card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf.

Gray, C. M.; Gunawan, J. T.; Schäfer, R.; Bielova, N.; Sanchez Chamorro, L.; Seaborn, K.; Mildner, T.; and Sandhaus, H. 2024a. Mobilizing research and regulatory action on dark patterns and deceptive design practices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–6.

Gray, C. M.; Santos, C. T.; Bielova, N.; and Mildner, T. 2024b. An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22.

Grok. 2024. Grok-2 Beta Release. https://x.ai/news/grok-2.

Grok. 2025. Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haq, M. U. U.; Rigoni, D.; and Sperduti, A. 2025. Llms as data annotators: How close are we to human performance. *arXiv preprint arXiv:2504.15022*.

Howe, N.; McKenzie, I.; Hollinsworth, O.; Zajac, M.; Tseng, T.; Tucker, A.; Bacon, P.-L.; and Gleave, A. 2024. Scaling trends in language model robustness. *arXiv preprint arXiv:2407.18213*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Kharchenko, J.; Roosta, T.; Chadha, A.; and Shah, C. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Kran, E.; Nguyen, H. M.; Kundu, A.; Jawhar, S.; Park, J.; Jurewicz, M. M.; et al. 2025. Darkbench: Benchmarking dark patterns in large language models. *arXiv preprint arXiv:2503.10728*.

Kumar, A.; Yunusov, S.; and Emami, A. 2024. Subtle Biases Need Subtler Measures: Dual Metrics for Evaluating Representative and Affinity Bias in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 375–392. Bangkok, Thailand: Association for Computational Linguistics.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mathur, A.; Acar, G.; Friedman, M. J.; Lucherini, E.; Mayer, J.; Chetty, M.; and Narayanan, A. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–32.

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Mistral. 2025. Mistral 3.1 release. https://mistral.ai/news/mistral-3-1-release/.

MoonshotAI. 2025. Kimi K2: Open Agentic Intelligence. https://moonshotai.github.io/Kimi-K2/.

OpenAI. 2024. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/.

OpenAI. 2025a. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1.

OpenAI. 2025b. OpenAI o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Rosenberg, L. 2023. The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency. arXiv:2306.11748.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Scheurer, J.; Balesni, M.; and Hobbhahn, M. 2023. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*.

Sharma, N.; Liao, Q. V.; and Xiao, Z. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Sun, X.; Chen, Y.; Huang, Y.; Xie, R.; Zhu, J.; Zhang, K.; Li, S.; Yang, Z.; Han, J.; Shu, X.; et al. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Team, T. H.; Liu, A.; Zhou, B.; Xu, C.; Zhou, C.; Zhang, C.; Xu, C.; Wang, C.; Wu, D.; Wu, D.; et al. 2025b. Hunyuan-TurboS: Advancing Large Language Models through Mamba-Transformer Synergy and Adaptive Chain-of-Thought. *arXiv preprint arXiv:2505.15431*.

Tencent. 2025. Reasoning Efficiency Redefined! Meet Tencent's 'Hunyuan-T1'—The First Mamba-Powered Ultra-Large Model. https://llm.hunyuan.tencent.com/#/Blog/hy-t1/.

Traubinger, V.; Heil, S.; Grigera, J.; Garrido, A.; and Gaedke, M. 2023. In Search of Dark Patterns in Chatbots. In *International Workshop on Chatbot Research and Design*, 117–132. Springer.

Tu, Q.; Fan, S.; Tian, Z.; and Yan, R. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.

Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; and Miao, Z. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–21.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yi, J.; Ye, R.; Chen, Q.; Zhu, B.; Chen, S.; Lian, D.; Sun, G.; Xie, X.; and Wu, F. 2024. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, 9236–9260.

Zamfirescu-Pereira, J. D.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–21.

Zhang, Z.; Jia, M.; Lee, H.-P.; Yao, B.; Das, S.; Lerner, A.; Wang, D.; and Li, T. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26.

Zhou, J.; Chen, Z.; Wan, D.; Wen, B.; Song, Y.; Yu, J.; Huang, Y.; Peng, L.; Yang, J.; Xiao, X.; et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.