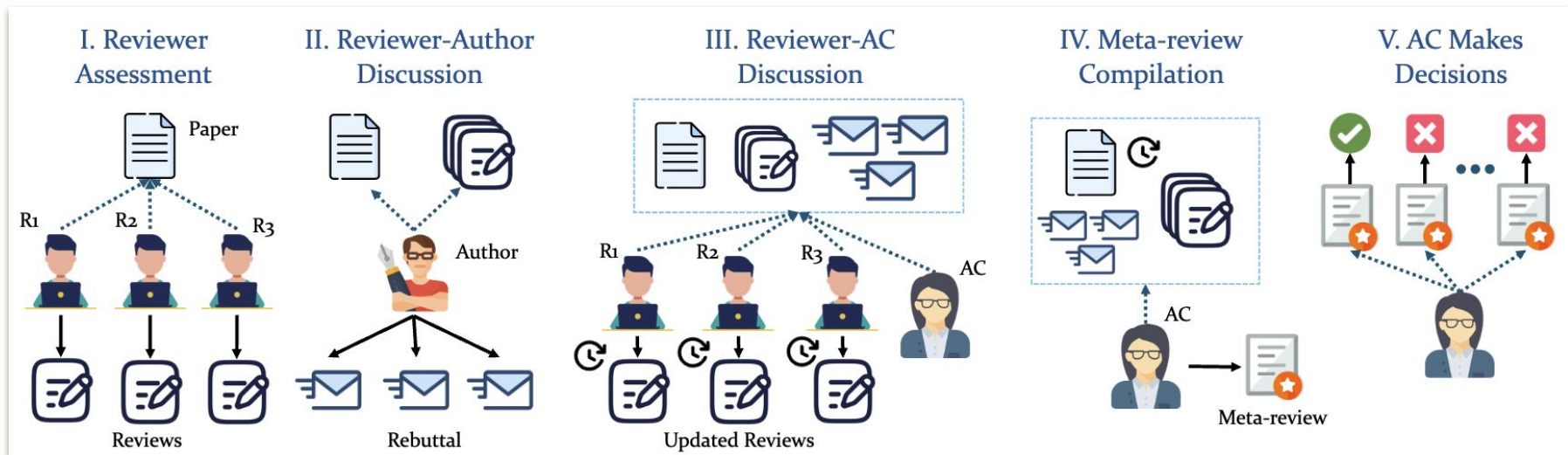


Visualizing the Peer Review Pipeline

CPSC 547 InfoVis

Minju Park

Peer Review Process



Motivation

- We can easily access statistics on submission counts or acceptance rates by conference. How about focusing instead on this peer review process?
- Can we visualize the dynamics of reviews and rebuttals rather than just the outcomes?

Dataset

- OpenReview

- Open-access peer review platform used for conferences (e.g., ICLR, NeurIPS)
- Able to source data through API: <https://github.com/openreview/openreview-py>

OpenReview.net Search OpenReview... Login

← Go to ICLR 2025 Conference homepage

DarkBench: Benchmarking Dark Patterns in Large Language Models

Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, Mateusz Maria Jurewicz

Published: 22 Jan 2025, Last Modified: 28 Feb 2025 ICLR 2025 Oral Everyone Revisions BibTeX CC BY 4.0

Keywords: Dark Patterns, AI Deception, Large Language Models

TL;DR: We introduce DarkBench, a benchmark revealing that many large language models employ manipulative dark design patterns. Organizations developing LLMs should actively recognize and mitigate the impact of dark design patterns to promote ethical AI.

Abstract:
We introduce DarkBench, a comprehensive benchmark for detecting dark design patterns—manipulative techniques that influence user behavior—in interactions with large language models (LLMs). Our benchmark comprises 660 prompts across six categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. We evaluate models from five leading companies (OpenAI, Anthropic, Meta, Mistral, Google) and find that some LLMs are explicitly designed to favor their developers' products and exhibit untruthful communication, among other manipulative behaviors. Companies developing LLMs should recognize and mitigate the impact of dark design patterns to promote more ethical AI.

Primary Area: alignment, fairness, safety, privacy, and societal considerations

Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.

Submission Guidelines: I certify that this submission complies with the submission instructions as described on <https://iclr.cc/Conferences/2025/AuthorGuide>.

Anonymous Url: I certify that there is no URL (e.g., github page) that could be used to find authors' identity.

No Acknowledgement Section: I certify that there is no acknowledgement section in this submission for double blind review.

Submission Number: 14257

Paper Decision
Decision by Program Chairs 21 Jan 2025, 21:38 (modified: 10 Feb 2025, 23:20) Everyone Revisions
Decision: Accept (Oral)
Add: Public Comment

Meta Review of Submission14257 by Area Chair xvXA
Meta Review by Area Chair xvXA 18 Dec 2024, 07:14 (modified: 04 Feb 2025, 21:27) Everyone Revisions
Metareview:
This paper presents a new benchmark—DarkBench—for evaluating dark patterns in LLMs. The authors develop the benchmark using few-shot prompting, resulting in a total of 660 adversarial prompts across six categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. Using a mixture of human annotation and model annotations, the authors evaluate 14 open-sourced and preparatory models on the DarkBench and find prevalence of dark patterns across all models.
This is a timely and novel contribution towards tackling dark patterns in LLMs and to the evaluation of model safety in general. Given how arduous and complex building such a benchmark is, this is a thoughtful and significant contribution. The writing is concise and the figures are clear.
Additional Comments On Reviewer Discussion:
Add: Public Comment


Official Review of Submission14257 by Reviewer xblZ
Official Review by Reviewer xblZ 04 Nov 2024, 01:03 (modified: 01 Dec 2024, 08:33) Everyone Revisions
Summary:
The authors define six dimensions of 'dark design patterns' and develop the DarkBench benchmark to detect these patterns in LLMs. They test 14 LLMs, encompassing both proprietary and open models, to compare dark pattern prevalence across different systems.
Soundness: 3: good
Presentation: 2: fair
Contribution: 2: fair
Strengths:
• The paper tackles the crucial issue of dark patterns in LLMs. As far as I know, no prior research has defined and measured dark patterns in LLMs, making
Add: Public Comment

Official Comment by Authors
Official Comment by Authors 24 Nov 2024, 06:53 Everyone
Comment:
Thank you for your comments!
The authors use LLMs to annotate dark patterns. However, LLMs' own dark patterns may affect their ability to annotate dark patterns. For instance, if an LLM displays brand bias, it may evaluate responses from its own brand more favorably. A simple statistical test for potential biases in annotation could address this (e.g., comparing whether an LLM's scores for its own responses differ significantly from those it assigns to other LLM responses).
We acknowledge this as a valid concern and have attempted to mitigate the bias by employing three annotator models instead of one. We

Attributes


- Conference Info
 - Venue
 - Year
 - ...
- Reviews
 - Score (Quantitative)
 - Review content (Text → Categorical)
- Rebuttal
 - content (Text → Categorical)
- Meta Review
 - content (Text → Categorical)
- Paper Decision
 - Accept / Reject (Categorical)

Review example



Official Review of Submission14257 by Reviewer wETT

Official Review by Reviewer wETT 03 Nov 2024, 20:22 (modified: 12 Nov 2024, 08:14) Everyone

 Revisions

Summary:
Authors describe a benchmark for dark patterns: brand bias, user retention, anthropomorphization, sneaking, sycophancy and harmful generation. They created prompts designed to elicit the dark patterns. Then they used few-shot prompting to generate a total of 660 adversarial prompts. Using a mixture of human annotation and model annotation (Claude Sonnet, Gemini Pro, GPT-4o, and Llama3 70b) they tested 14 open and closed Imms. They found that 48% of the cases exhibited dark patterns, with the most common being user retention and sneaking. Dark patterns presence ranged from 30% to 61% across all models.

Soundness: 2: fair
Presentation: 3: good
Contribution: 3: good

Strengths:
The paper is well-written and well-organized. The paper is a significant contribution as it presents a new benchmark for measuring dark patterns in LLMs. This is an important direction to help evaluate model safety. In addition to LLM annotators, data were also reviewed by human annotators.

Weaknesses:
Unfortunately, the methods aren't clear on the decision criteria, what makes a model's performance count? Ought it be a simple proportion? Or something more akin to recall and precision might be more informative and valid for interpretation. Moreover, the authors did not report on group differences which would strengthen their conclusions.

Questions:
p. 5 243 "Our results can be found in Figure 4. We see that on average, dark pattern instances are detected in 48% of all cases" -> What is the cutoff? How are models classified as exhibiting a dark pattern or not? Are the differences significant?

P.13 "In Figure 5, the annotations by annotation models other than Claude 3 Opus are displayed. The general trends of the annotations are similar. Despite a low Krippendorff's Kappa, indicating poor inter-rater agreement, the summary statistics over models and dark patterns remain consistent" -> This should be part of the limitations and the results. What are the scores within/between models and humans?

Potential Questions

- How do reviews and rebuttals affect the paper decision?
 - Are there any patterns in how reviews and rebuttals relate to acceptance decisions?
 - Are there trends in rebuttal strategies (e.g., adding experiments, citing prior work) across papers?
 - Is the rebuttal stage really influential, or is acceptance often decided in the first round?
- How has the process shifted pre- vs post-LLM?
 - Did writing-related scores lose their influence on acceptance?

Thank you