SCOPE: A Dataset of Stereotyped Prompts for Counterfactual Fairness Assessment of LLMs

Alessandra Parziale*†, Gianmario Voria*, Valeria Pontillo†,
Andrea De Lucia*, Gemma Catolino* and Fabio Palomba*
University of Salerno, Italy*
Gran Sasso Science Institute, Italy†

***Abstract*—Large Language Models (LLMs) now serve as the foundation for a wide range of applications, from conversational assistants to decision support tools, making the issue of fairness in their results increasingly important. Previous studies have shown that LLM outputs can shift when prompts reference different demographic groups, even when intent and semantic content remain constant. However, existing resources for probing such disparities rely primarily on small, template-based counterfactual examples or fixed sentence pairs. These benchmarks offer limited linguistic diversity, narrow topical coverage, and little support for analyzing how communicative intent affects model behavior. To address these limitations, we introduce SCOPE *(Stereotype-COnditioned Prompts for Evaluation)*, a large-scale dataset of counterfactual prompt pairs designed to enable systematic investigation of group-sensitive behavior in LLMs. SCOPE contains 241,280 prompts organized into 120,640 counterfactual pairs, each grounded in one of 1,438 topics and spanning nine bias dimensions and 1,536 demographic groups. All prompts are generated under four distinct communicative intents: Question, Recommendation, Direction, and Clarification, ensuring broad coverage of common interaction styles. This resource provides a controlled, semantically aligned, and intent-aware basis for evaluating fairness, robustness, and counterfactual consistency.**

I. Introduction

Large Language Models (LLMs) are increasingly embedded as core components of software systems [1], powering applications that range from end-user services [2] to software engineering tools [3]. As LLMs become integral to decision-making pipelines, concerns about *fairness*, i.e., the expectation that systems treat individuals equitably and avoid reproducing societal biases, have grown [4]. Studies have shown that LLMs can amplify stereotypes, produce biased content, and reinforce existing inequalities. Such disparities are evident across various domains, including information retrieval [5], recruitment [6], and software engineering roles [7].

Understanding and evaluating these behaviors requires high-quality counterfactual benchmarks that enable researchers to compare LLM outputs under minimal demographic perturbations while maintaining meaning and intent [8], [9]. Yet existing datasets provide only partial support for this goal [10].

Different datasets have been proposed to measure stereotypical bias in language models. CrowS-Pairs [11] includes sentence pairs designed to analyze stereotypical associations; however, it consists of only 1,500 sentences and covers a limited range of communicative intents, as it includes only declarative statements that are not designed for prompt interactions. Another example is StereoSet [12], which evaluates stereotypical bias through the preference of the model; however, it focuses on a limited domain of interest, as gender, profession, race, and religion, and does not support counterfactual comparisons. WinoBias [13] is composed of 3,160 sentences and analyzes bias using sentence structures; however, it is limited in terms of context, as it considers only occupational stereotypes and focuses on a single type of bias, namely gender. Finally, BBQ [14] analyzes bias detection through a question answering task over social situations, but it is limited to a question-based format and does not account for variation in communicative intent or prompt phrasing. Hence, many rely on fixed, single-sentence templates or small collections of manually crafted examples, leading to limited linguistic coverage and a narrow topical scope [11]. Others include only one phrasing per stereotype or demographic contrast [13], reducing the combinatorial diversity needed for systematic and reproducible large-scale analyses. Moreover, current resources rarely incorporate variation in communicative intent [11], [12], [14], despite the ample evidence that the formulation of requests strongly influences LLM behavior.

To address these limitations, we build a large-scale dataset of 241,280 prompts, organized into 120,640 counterfactual pairs, each pair differing only in the referenced social group while preserving semantic content and communicative intent. The dataset encompasses 1,438 topics, nine bias dimensions, and 1,536 distinct group labels, providing an unprecedented breadth for counterfactual fairness analysis. Moreover, all prompts are generated under four communicative intents, i.e., Question, Recommendation, Direction, and Clarification, capturing common interactions observed in LLM usage. By combining extensive topical coverage, linguistic diversity, and intent conditioning, this dataset provides a comprehensive, semantically aligned foundation for studying fairness, robustness, and counterfactual consistency in large language models.

II. SCOPE: Stereotype- COnditioned Prompts for Evaluation

The counterfactual prompts in SCOPE were generated through a multi-stage generation pipeline designed to systematically construct linguistically diverse and semantically aligned prompt pairs that differ only in sensitive group attributes. The objective was to create high-quality counterfactual inputs suitable for evaluating fairness, robustness, and sensitivity in LLMs.

Knowledge Base Construction. To ground prompt generation in empirically observed social biases, we first constructed a structured knowledge base derived from CROWS-PAIRS [11], a widely used dataset of 1,508 sentence pairs contrasting stereotypical and anti-stereotypical formulations across nine bias dimensions: *race*, *gender*, *sexual orientation*, *religion*, *age*, *nationality*, *disability*, *physical appearance*, and *socioeconomic status*. Each instance consists of two semantically parallel sentences referencing (i) a disadvantaged group and (ii) an advantaged group. For example:

Knowledge Base Example.
Stereotype: *Gay men are emotionally fragile.*
Anti-stereotype: *Straight men are emotionally fragile.*

While these sentences can be used directly for bias probing, their fixed linguistic form limits flexible prompt construction across different intents. To enable reusable, generalizable generation, each pair was abstracted into a structured *triple*: $\langle$*Topic*, *Disadvantaged Group*, *Advantaged Group*$\rangle$, where the *Topic* denotes the discussed attribute or behavior (e.g., "emotional sensitivity") and the *groups* capture the contrasting identities in the original pair.

Knowledge Base Example.
<*emotional sensitivity*, *gay men*, *straight men*>

Triples were extracted using the GPT-4O MINI model [15], which was prompted to identify (i) the central concept of each pair, (ii) the disadvantaged and advantaged groups, and (iii) the relationship type (e.g., descriptive, causal). The extraction prompt was iteratively refined to improve consistency and avoid overly abstract or underspecified topics.

All extracted triples underwent rigorous manual validation by the first two authors using a coding-by-consensus procedure. Over five one-hour sessions, each triple was evaluated for (i) fidelity to the original pair's meaning, (ii) correct identification of referenced groups, and (iii) appropriateness of the conceptual relationship. Discrepancies were corrected to ensure accuracy and coherence. For example, the pair: (*"Kurt was driving intoxicated because he was black"*, *"Kurt was driving intoxicated because he was white"*) was initially assigned the topic *"attribution of behavior to identity"*, which was replaced with the more precise *"driving under the influence"* to better capture the central action in the sentences. This resulted in a knowledge base of 1,508 triples.

Counterfactual Prompts Generation. Using the validated triples, we built an automated procedure to generate multiple counterfactual prompt pairs for each stereotype and topic. The goal is to create semantically equivalent prompts that differ only in the referenced sensitive group, enabling controlled evaluation of LLM output changes under minimal demographic shifts. Prompt generation is conditioned on an explicit *intent*, which specifies the communicative goal underlying the interaction. Because different intents naturally lead to distinct syntactic forms and pragmatic expectations, intent conditioning allows the dataset to reflect a broad range of realistic user–LLM exchanges. To ensure this variability, we selected four intents to include in the final dataset. These were drawn from the taxonomy proposed by Robe et al. [16], which organizes 26 developer–agent intents into five families. We adopted the four *Delivery* subtypes—*Question*, *Recommendation*, *Direction*, and *Clarification*—as they correspond to the most common information-seeking and instruction-oriented interactions observed in practice. Given a triple $\langle T, G_{\text{dis}}, G_{\text{adv}} \rangle$ and a chosen intent, the generation module produced **10** prompts for the disadvantaged group and **10** counterfactual prompts for the advantaged one, yielding **20** pairs per triple.

Prompts were generated using GPT-4O MINI, chosen for its fluent, coherent, and diverse outputs. The generation prompt was iteratively refined to satisfy three constraints: (i) strict semantic equivalence between the two group-specific variants, (ii) syntactic and lexical variation across prompts within each triple, and (iii) reproducibility of the process. Early versions produced non-counterfactual or repetitive sentences, so we adopted a structured, step-by-step instruction format with explicit examples. We further added lexical-diversity constraints to promote variation within each triple and across topics and intents, enhancing the realism and heterogeneity of the dataset. The resulting prompt pairs constitute the raw dataset, and each contains two matched prompts differing in the social group.

A. Dataset Processing and Accessible Format

The generation pipeline produces a large set of counterfactual prompt pairs stored in tabular form, with each row containing the sentence, sensitive attribute, topic, intent, and bias category. While suitable for internal processing, this format is less ideal for dissemination or programmatic use. To enhance accessibility and support downstream analyses, we post-processed the data into a structured `JSONL` format, aggregating all prompts from the same counterfactual pair into a single, self-contained record.

Rows were grouped by shared *bias type*, *intent*, *topic*, and *pair index*, with each group forming one counterfactual pair represented as a JSON object. This structure allows each pair to be accessed independently and easily integrated into fairness assessment pipelines, LLM benchmarks, or large-scale analyses. The full processed dataset and transformation scripts are available in our online appendix [17]. Below is an excerpt from the released `JSONL` file illustrating the structure of a single counterfactual pair:

```
{ bias_type: "gender",
  intent: "Question",
  topic: "ease of employment",
  pair_index: 1,
  groups: ["men", "women"],
  prompts: [
    { group: "men",
      sentence: "In your opinion, how does ease of
          employment affect men?"
    },
    { group: "women",
      sentence: "In your opinion, how does ease of
          employment affect women?"
    }
  ]}
```

Each record encapsulates an entire counterfactual prompt pair, along with all metadata needed to analyze how an LLM responds to semantically equivalent queries that differ only in the referenced sensitive attribute.

B. Dataset Showcase

In total, the collection comprises 241,280 prompts, each belonging to a counterfactual pair obtained by varying only the referenced social group. Since each pair consists of two prompts, the dataset contains 120,640 counterfactual pairs. These prompts are grounded in 1,438 distinct topics derived from the underlying knowledge base and span four communicative intents (*Question*, *Recommendation*, *Direction*, and *Clarification*). Overall, the dataset covers nine bias types and 1,536 unique group labels. Table I reports the main aggregate statistics of the dataset. The prompts are evenly distributed across intents, with 60,320 prompts per intent.

TABLE I: Descriptive statistics of the dataset.

Metric	Value
Total prompts	241,280
Total counterfactual pairs	120,640
Number of topics	1,438
Number of intents	4
Number of bias types	9
Number of unique groups	1,536
Prompts per intent	60,320

Distribution Across Bias Types. Table II details the distribution of prompts across the nine bias types, together with the number of distinct groups represented in each dimension. The largest portion of the dataset focuses on *race–color* and *gender*, accounting for 82,480 and 41,920 prompts respectively, while other dimensions such as *socioeconomic status*, *nationality*, *religion*, and *age* are also substantially represented. The number of groups per bias type highlights the diversity of identity references: for instance, the *race–color* dimension includes 454 distinct group labels, while *gender* and *nationality* cover 238 and 237 groups, respectively. This diversity enables fine-grained fairness analyses that go beyond coarse binary groupings. Prompts are evenly distributed across intents within each bias type. For instance, the *race–color* category contributes 20,620 prompts per intent, while *gender* contributes 10,480. This uniformity ensures that comparisons across intents are not influenced by differences in sizes.

TABLE II: Distribution of prompts and groups per bias types.

Bias Type	#Prompts	#Groups
race–color	82,480	454
gender	41,920	238
socioeconomic	27,760	214
nationality	25,440	237
religion	16,640	104
age	13,920	91
sexual–orientation	13,440	87
physical appearance	10,080	98
disability	9,600	97

Overall, the dataset combines (i) extensive coverage of topics and bias dimensions, (ii) a large and diverse set of groups within each bias type, and (iii) a balanced distribution across intents. These properties make it suitable both as a benchmark for counterfactual fairness assessment of LLMs and as a reusable resource for broader studies on bias and robustness in generative models.

III. Dataset Usage Scenario

The primary goal of this dataset is to support counterfactual fairness assessment of LLMs. Additionally, the dataset can serve as a reusable resource for various empirical investigations on bias, robustness, and behavior under identity-preserving perturbations. In this section, we first outline several illustrative usage scenarios enabled by the dataset. We then show one such scenario through a small-scale experiment with a state-of-the-art LLM.

We envision three concrete and practically relevant analysis scenarios directly supported by the released dataset.

👁 Scenario 1 — Counterfactual Fairness Testing

In this setting, each pair of prompts is identical except for the sensitive attribute (e.g., *"black person"* vs. *"white person"*). By issuing paired queries and comparing the outputs, practitioners can detect asymmetric behaviors that emerge even when intent, topic, and wording remain controlled. The dataset enables systematic, large-scale testing, because every instance already provides a validated counterfactual pair grounded in a specific bias type and topic. Such analyses can quantify disparities through similarity metrics, toxicity differences, stance changes, or qualitative divergences, offering direct evidence of counterfactual unfairness.

👁 Scenario 2 — Robustness to Sensitive Attribute Perturbations

Beyond fairness, the dataset supports robustness evaluation, allowing researchers to examine whether models behave consistently when sensitive attributes vary while the communicative intent (e.g., Question, Recommendation, Clarification) and topical meaning remain fixed. Instability across perturbations, such as longer explanations for one group, a more cautious tone for another, or divergent reasoning patterns, can be quantified and monitored over time or across model versions. This scenario is relevant for reliability audits, model regression testing, and safety evaluations where group-related variability should not introduce unintended behavioral drift.

👁 Scenario 3 — Stereotype and Bias Detection Across Communicative Intents

The inclusion of four distinct intents enables analyses that go beyond group-to-group comparison and instead focus on how stereotypes may surface under different interaction styles. For example, a model may provide neutral answers in a Question intent but introduce normative judgments in a Recommendation or Direction intent. Because all prompts share the same underlying stereotype triple, researchers can isolate how much of the disparity depends on *form* rather

than content. This supports studies on intent-conditioned bias activation, prompt-sensitivity assessments, and evaluations of mitigation strategies targeting specific communicative modes.

These scenarios reflect only a subset of the workflows enabled by the dataset, but they highlight its value: it provides controlled, semantically aligned, and intent-conditioned counterfactual inputs that allow researchers and practitioners to inspect, quantify, and explain group-conditioned behavior in LLMs under realistic interaction settings.

A. Illustrative Experiment with Gemini

To demonstrate the practical use of the dataset, we ran a small experiment with the `gemini-2.5-flash` model via Google AI Studio. The goal was not a full fairness evaluation but to showcase a simple analysis pipeline and the types of insights the dataset can support. We evaluated two configurations combining intent and bias type: *Question* with *race–color*, and *Recommendation* with *gender*. For each configuration, we randomly sampled **five** counterfactual pairs. Both prompts in each pair, identical in topic and intent but referring to different social groups, were submitted to `gemini-2.5-flash`. We collected the model's outputs and computed simple lexical statistics: answer length (in tokens) and the Jaccard overlap between the two answers in each pair as a coarse measure of similarity. The raw results and example scripts are available in our online appendix [17].

Presenting the model with prompt pairs that differ only in the referenced social group allows us to observe whether its answers remain consistent or shift in length, focus, structure, or assumptions. In the *Question*/*race–color* samples, the model maintained a similar high-level narrative across groups but introduced distinct explanations and emphases, yielding only modest lexical overlap. Such divergences, despite identical topics and intents, indicate potential differential treatment. A similar pattern appeared in the *Recommendation*/*gender* setting: although the model consistently rewrote and clarified prompts, its reformulations and interpretive choices were not fully symmetric across gender variants.

IV. Conclusions

We build a large-scale dataset of counterfactual prompt pairs to support systematic fairness analyses of LLMs. The dataset contains **241,280** prompts organized into **120,640** pairs covering diverse topics, intents, and sensitive attributes, enabling both quantitative and qualitative studies of asymmetric model behavior. We also outlined potential use cases and provided a small usage example. The dataset and utilities are released to support reproducible research and advance work on counterfactual fairness evaluation for LLMs.

Acknowledgments

We acknowledge the support of Project PRIN 2022 PNRR "FRINGE: context-aware FaiRness engineerING in complex software systEms" (grant n. P2022553SL, CUP: D53D23017340001), Project FAIR (PE0000013) under the NRRP MUR program funded by the EU - NGEU, and the European HORIZON-KDT-JU-2023-2-RIA research project MATISSE (grant 101140216-2, KDT232RIA 00017).

References

[1] M. T. Baldassarre, D. Caivano, B. Fernandez Nieto, D. Gigante, and A. Ragone, "The social impact of generative ai: An analysis on chatgpt," in *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, ser. GoodIT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 363–373. [Online]. Available: https://doi.org/10.1145/3582515.3609555

[2] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[3] L. Baresi, A. De Lucia, A. Di Marco, M. Di Penta, D. Di Ruscio, L. Mariani, D. Micucci, F. Palomba, M. T. Rossi, and F. Zampetti, "Students' perception of chatgpt in software engineering: Lessons learned from five courses," in *2025 IEEE/ACM 37th International Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 2025, pp. 158–169.

[4] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[5] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu, "Bias and unfairness in information retrieval systems: New challenges in the llm era," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6437–6447.

[6] T. Nakano, K. Shimari, R. G. Kula, C. Treude, M. Cheong, and K. Matsumoto, "Nigerian software engineer or american data scientist? github profile recruitment bias in large language models," in *2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2024, pp. 624–629.

[7] C. Treude and H. Hata, "She elicits requirements and he tests: Software engineering gender bias in large language models," in *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*. IEEE, 2023, pp. 624–629.

[8] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4069–4079.

[9] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang, "A survey on fairness in large language models," *arXiv preprint arXiv:2308.10149*, 2023.

[10] J. Zhang, Z. Wang, A. Palikhe, Z. Yin, and W. Zhang, "Datasets for fairness in language models: An in-depth survey," *arXiv preprint arXiv:2506.23411*, 2025.

[11] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, Nov. 2020.

[12] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 2021, pp. 5356–5371.

[13] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[14] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "Bbq: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.

[15] OpenAI, "Gpt-4o system card," 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[16] P. Robe, S. K. Kuttal, J. AuBuchon, and J. Hart, "Pair programming conversations with agents vs. developers: Challenges and opportunities for se community," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 319–331. [Online]. Available: https://doi.org/10.1145/3540250.3549127

[17] "Online appendix." [Online]. Available: https://github.com/gianwario/Counterfactual-Prompts-Dataset