Say It Another Way: Auditing LLMs with a User-Grounded Automated Paraphrasing Framework

Large language models (LLMs) are sensitive to subtle changes in the prompt, leading to markedly different outputs. This presents a critical challenge for auditors in how to accurately capture the diversity of real-world prompts and in how prompt sensitivity affects the reliability of audit results. Existing auditing literature has explored prompt sensitivity by modifying prompt formatting or by paraphrasing the prompt. While these variations aim to simulate the sensitivity to changing prompts by real users, they are not explicitly grounded in actual user behavior. As a result, they risk missing certain demographics or generating unrealistic prompt variations, as illustrated on Figure 1. With extensive literature on the linguistic foundations of paraphrasing and characteristic patterns of language use in various demographics, we argue that the current body of LLM auditing research would benefit from a user-grounded approach to prompt sensitivity, one that focuses on modeling the distribution of users interacting with the LLM.

To bridge these gaps, we present AUGMENT (**A**utomated **U**ser-**G**rounded **M**odeling and **E**valuation of **N**atural Language **T**ransformations), a framework to systematically incorporate prompt sensitivity in LLM auditing. AUGMENT is built around two core principles. First, it uses linguistically structured transformations and incorporates contextual grounding based on user demographics and identity markers, to generate paraphrases that reflect real-world prompt variability. Second, it enables robust evaluation to ensure that generated paraphrases adhere to the desired transformation, are realistic, and preserve the meaning of the original sentence.

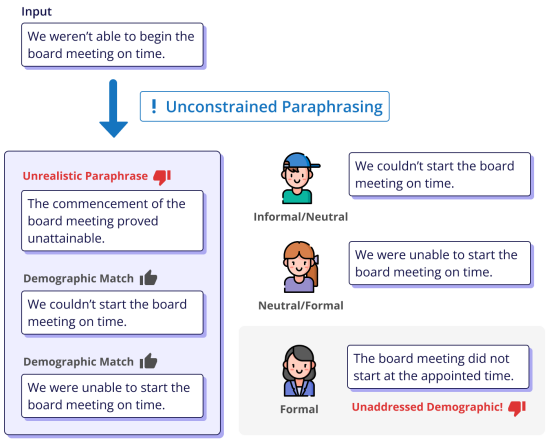

Figure 1: Distribution of Unconstrained Paraphrasing is Distinct from that of Actual User Behavior.

We present a case study on the BBQ dataset [3], applying five paraphrase types from established taxonomies [1, 2]: *Preposition Variation*, *Voice Change*, *Synonym Substitution*, *Formality Change*, and *African American English (AAE) Dialect Transformation*. We use an LLM as a controlled generator, as we prompt it to perform only one specified modification at a time. Prompts are designed in a few-shot format with examples drawn from the taxonomies, and we use both ChatGPT (gpt-4o) and DeepSeek-V3-Chat as generators. The resulting paraphrases are first evaluated through human annotation. These judgments are then compared with automatic filtering methods based on metrics such as semantic similarity and perplexity, to establish a scalable evaluation approach.

We conclude by auditing bias with the BBQ dataset across nine LLMs (LLaMA, MPT, Falcon, and Gemma families), analyzing metric shifts under user-grounded prompt variations. Our results show that paraphrased inputs generally induce greater score variability, with effects depending on the paraphrase type. For instance, the bias score quantifies how much an LLM favors stereotypes or anti-stereotypes and ranges from -100% to 100%, with 0% indicating no bias. In ambiguous contexts with Gemma 3 (12B), it rises from 2% to 4% under *Preposition Variation*, *Voice Change* and *African American English (AAE) Dialect Transformation*, to 6% under *Synonym Substitution* and to 8% under *Formality Change*.

[1] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational linguistics*, 39(3):463–472, 2013.

[2] Marcel Gohsen et al. Task-oriented paraphrase analytics. In *Proceedings of the 2024 LREC-COLING*, pages 15640–15654.

[3] Alicia Parrish et al. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL 2022*, pages 2086–2105.