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

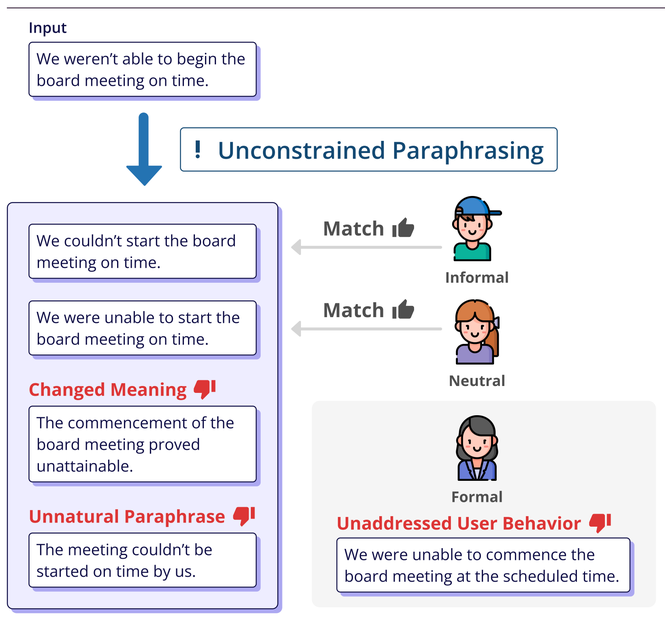
Cléa Chataigner[1,2*] Rebecca Ma[3,4*] Prakhar Ganesh[1,2] Yuhao Chen[3] Afaf Taïk[1,5]
Elliot Creager[3,4] Golnoosh Farnadi [1,2,6]

[1]Mila – Quebec AI Institute [2]McGill University [3]University of Waterloo [4]Vector Institute [5]Université de Sherbrooke [6]Université de Montréal

TL;DR

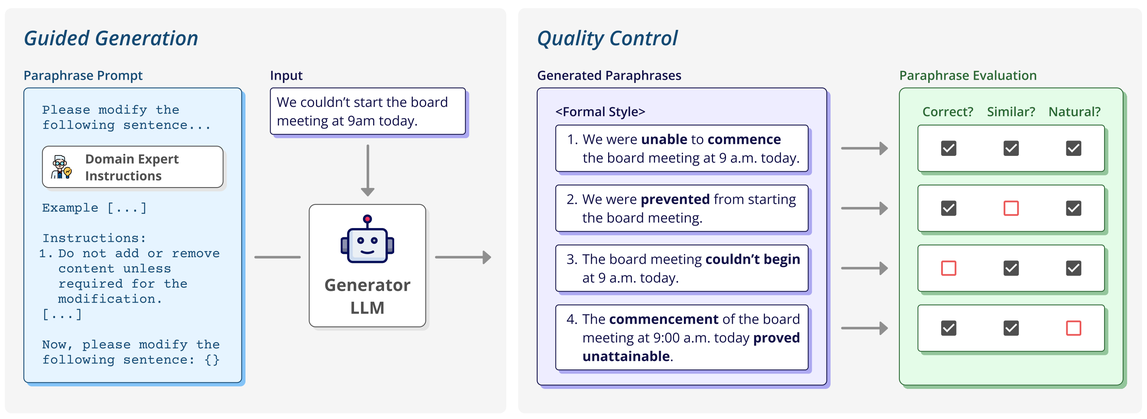
- Linguistic sensitivities **missed by uncontrolled paraphrasing**
- **AUGMENT:** Controlled paraphrasing framework for systematic generation and evaluation of paraphrases
- Uncovers **hidden sensitivities** across models and data subsets

Context and Motivations



- **Prompt sensitivity**: Same prompt variant, different model behavior
- **Current methods**: Disconnected from real users [4], existing paraphrasing lacks control and quality [5]
- We propose a controlled, user-grounded framework **AUGMENT**: **A**utomated **U**ser-**G**rounded **M**odeling and **E**valuation of **N**atural language **T**ransformations

The AUGMENT Framework



- Instruction-tuned LLMs for paraphrasing
- Explicit rules directly in the prompts
- Few-shot examples to guide generation

- ***Instruction Adherence:*** Was the intended modification correctly applied?
- ***Similarity:*** Is the paraphrase semantically close to the original text?
- ***Realism:*** Is the paraphrase natural and user-like?

AUGMENT in practice

Guided Generation

Paraphrase type	Example
Prepositions variations [1]	Results **of** the competition ↔ Results **for** the competition
Synonym substitution [1]	Google **bought** YouTube ↔ Google **acquired** YouTube
Voice Change [1]	Pat **loves** Chris ↔ Chris **is loved by** Pat
Formal Style [2]	I **got** your email ↔ I **have received** your email
AAE Dialect [3]	They **are walking** too fast ↔ They **walking** too fast

Table 1. Selected paraphrase types.

- LLM generators: ChatGPT (gpt-4o) and DeepSeek-V3.1-Chat
- Dataset: Gender Identity subset of the BBQ dataset
- Validation: Human annotation with 3 independent annotators

Takeaway: LLMs can yield high-quality constrained paraphrases, but outputs are inconsistent: robust filtering is essential

Quality Control

- *Instruction Adherence*: POS tagging with `spacy`, automated classifiers
- *Semantic similarity*: threshold on SBERTScore
- *Realism*: threshold on Perplexity ratio

	Precision	Recall	F1 Score
Prepositions	88.74	90.68	89.70
Synonyms	66.57	90.64	76.76
Voice Change	42.60	72.39	53.64
AAE Dialect	82.73	76.47	79.48
Formal Style	92.56	89.96	91.24

Table 2. Performance Metrics for Automated Filtering Rules, compared with Human Annotations

Takeaway: Reasonably high F1 scores across modifications except voice change, which proved to be a difficult task even for annotators.

Auditing Prompt Sensitivity

Experimental settings

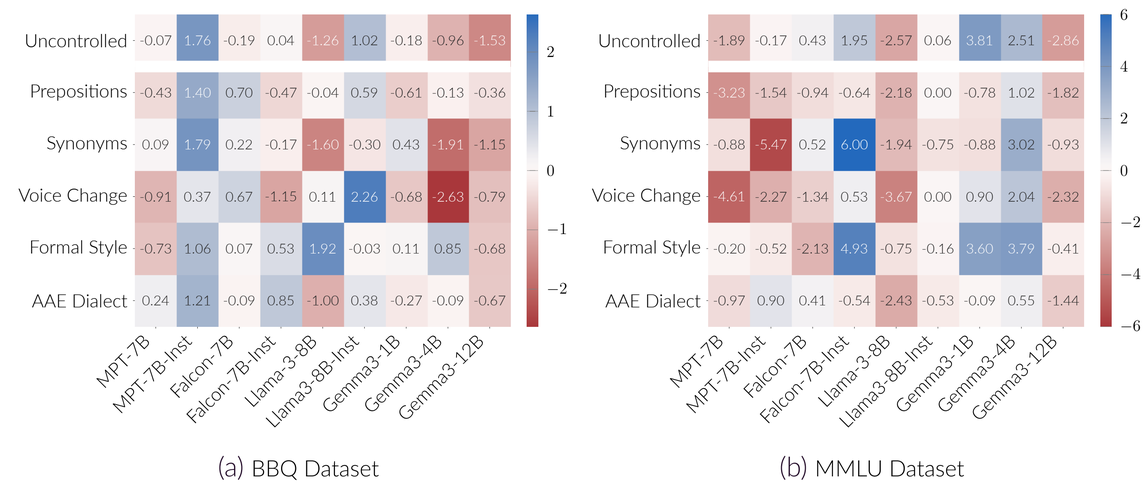
- Datasets: Full BBQ dataset, MMLU; 9 target LLMs: MPT, Falcon, LLaMA 3, Gemma 3
- Baseline: Uncontrolled paraphrasing with no specified type or instructions
- Metric: Relative difference in accuracy, for responses to paraphrased and original prompts

1) Controlled vs. uncontrolled: similar overall trends

		MPT 7B	MPT 7B-Inst	Falcon 7B	Falcon 7B-Inst	Llama-3 8B	Llama-3 8B-Inst	Gemma3 1B	Gemma3 4B	Gemma3 12B
BBQ	Uncontrolled	-0.07	1.76	-0.19	0.04	-1.26	1.02	-0.18	-0.96	-1.53
	Controlled	-0.35	1.17	0.31	-0.08	-0.12	0.58	-0.20	-0.78	-0.73
MMLU	Uncontrolled	-1.89	-0.17	0.43	1.95	-2.57	0.06	3.81	2.51	-2.86
	Controlled	-1.98	-1.78	-0.70	2.06	-2.20	-0.29	0.55	2.08	-1.38

Table 3. Relative difference in accuracy, aggregated over the five controlled and uncontrolled paraphrases per example.

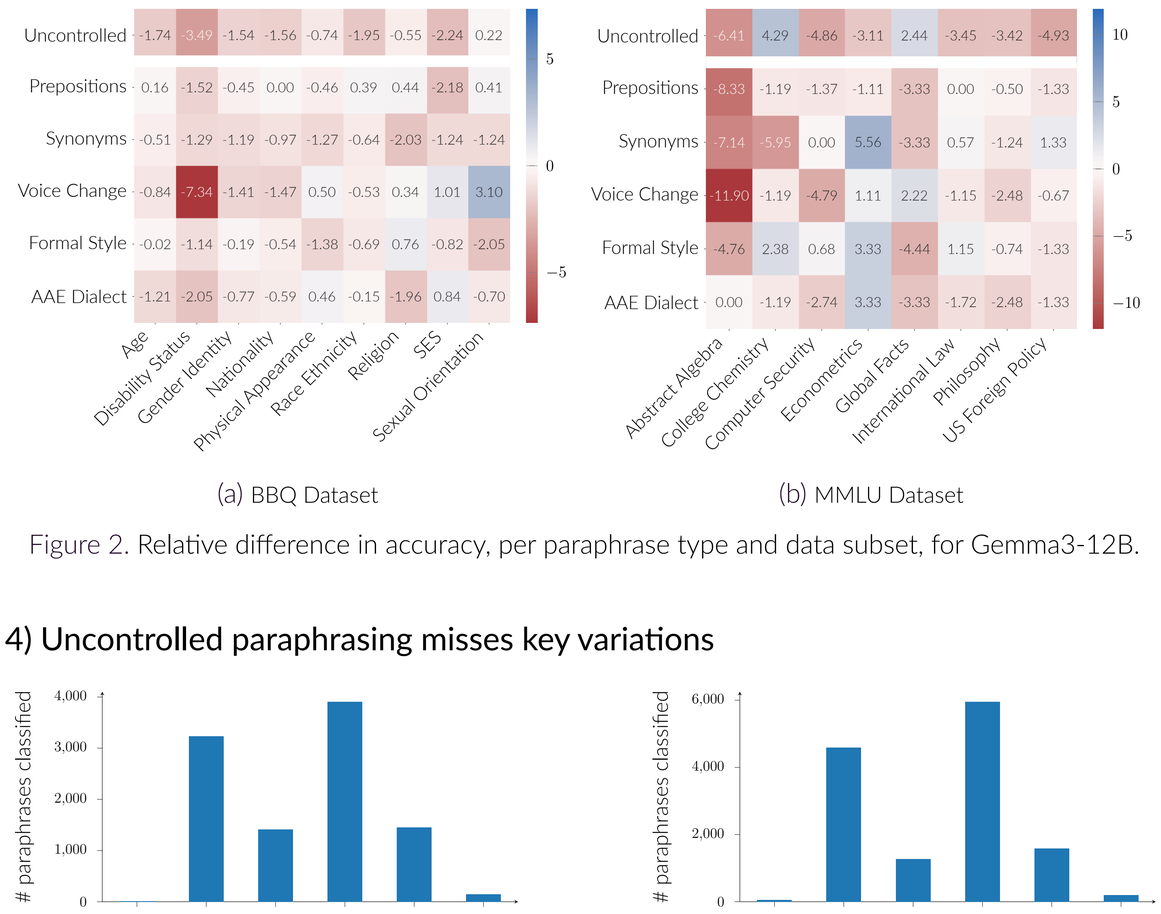
2) AUGMENT exposes hidden sensitivities



(a) BBQ Dataset (b) MMLU Dataset

Figure 1. Relative difference in accuracy, per paraphrase type and target model.

3) Sensitivity patterns diverge across data subsets



(a) BBQ Dataset (b) MMLU Dataset

Figure 2. Relative difference in accuracy, per paraphrase type and data subset, for Gemma3-12B.

4) Uncontrolled paraphrasing misses key variations

(a) BBQ Dataset (b) MMLU Dataset

Figure 3. Classification of the uncontrolled paraphrases, reusing our automatic filtering rules.

Future Work

- **Scope:** Extend evaluation from multiple-choice to open-ended questions.
- **Filtering:** Improve robustness of automatic filtering methods.
- **Language & Data:** Expand paraphrase taxonomy beyond English and U.S.-centric datasets.

References

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

[2] Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. Detecting text formality: A study of text classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, September 2023.

[3] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 789–798. ACM, June 2022.

[4] Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. Social bias evaluation for large language models requires prompt variations, 2024.

[5] Abdelrahman Zayed, Goncalo Mordido, Ioana Baldini, and Sarath Chandar. Why don't prompt-based fairness metrics correlate? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9002–9019. Association for Computational Linguistics, August 2024.

*Equal contribution

clea.chataigner@mila.quebec, rebecca.ma@uwaterloo.ca