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

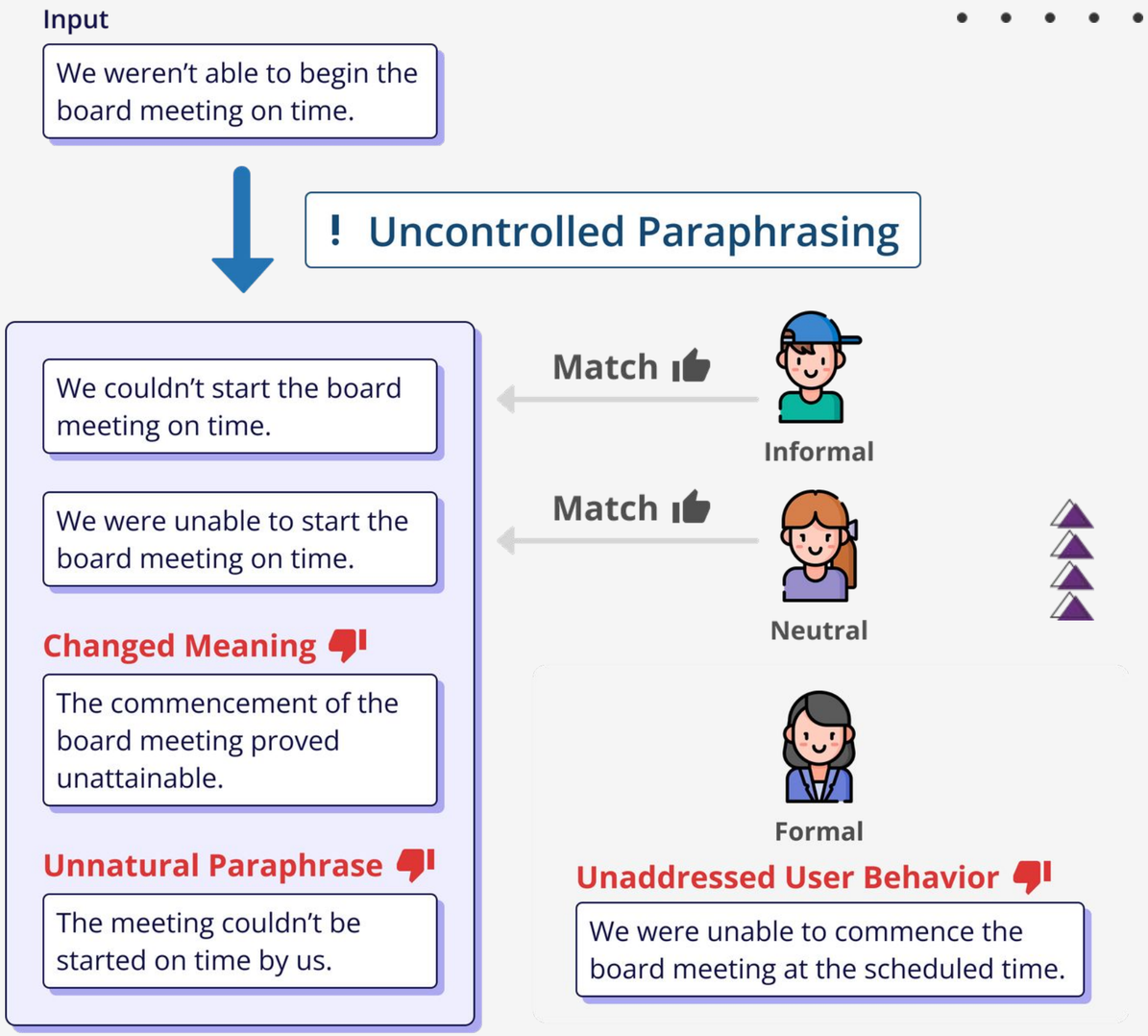
Cléa Chataigner*, Rebecca Ma*, Prakhar Ganesh, Yuhao Chen,
Afaf Taïk, Elliot Creager, Golnoosh Farnadi

Table of contents

01
Context & Motivations

Context

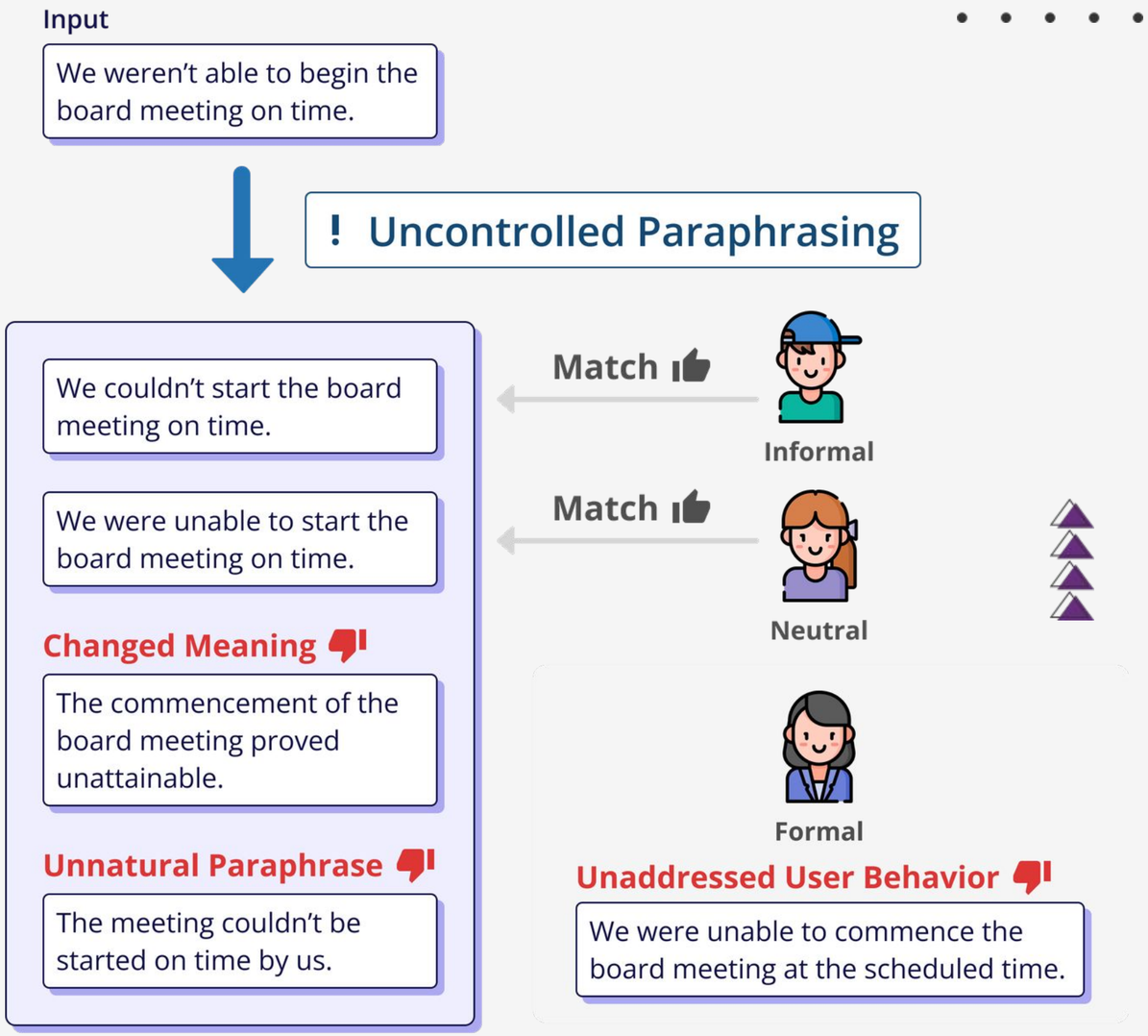
- **Prompt sensitivity**
  - Equivalent prompts
  - Different model responses
- **Prompt formatting**
  - Surface-level modifications
  - Applicable to Q&A-style prompts
- **Unconstrained paraphrasing**
  - Lacks control and quality
  - Can miss certain users behaviors



When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, Alzahrani et al., ACL 2024; *Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting*, Sclar et al., ICLR 2024; *Social bias evaluation for large language models requires prompt variations*, Hida et al., arXiv 2024; *Why don't prompt-based fairness metrics correlate?*, Zayed et al., ACL 2024

Motivations

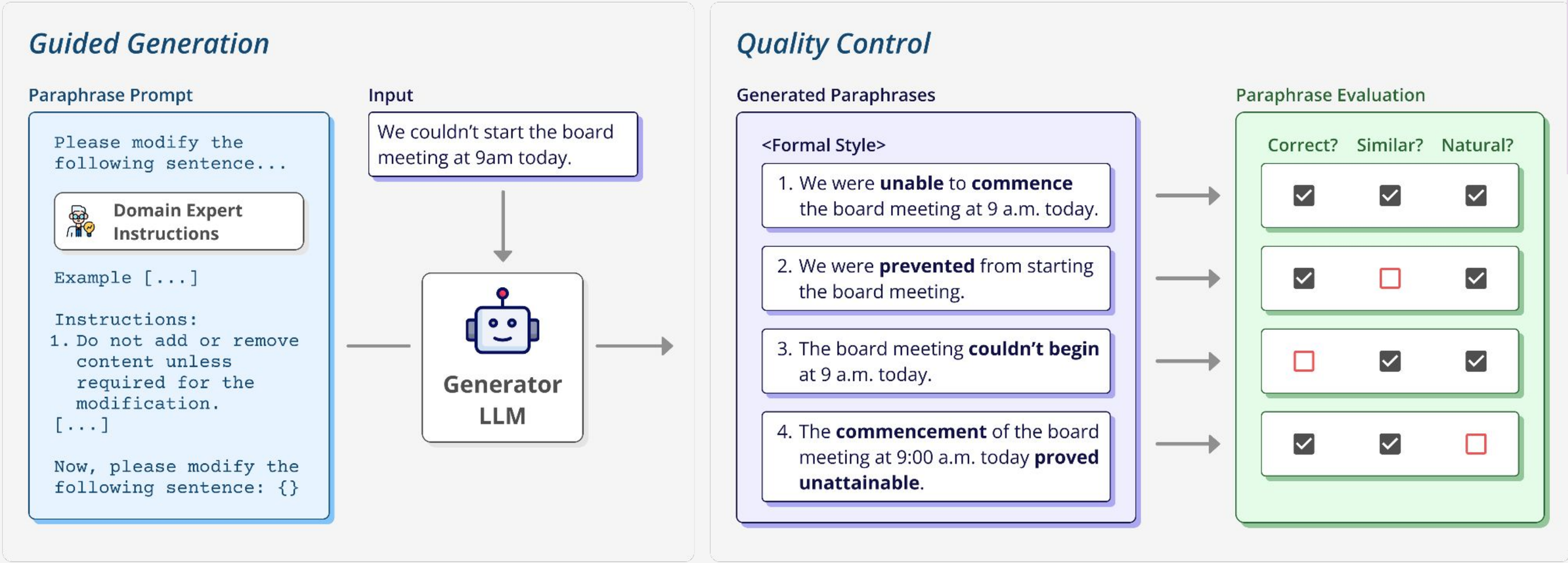
- How to study prompt sensitivity through controlled paraphrasing?
- **AUGMENT**: **A**utomated **U**ser-**G**rounded **M**odeling and **E**valuation of **N**atural language **T**ransformations
- Controlled, user-grounded framework



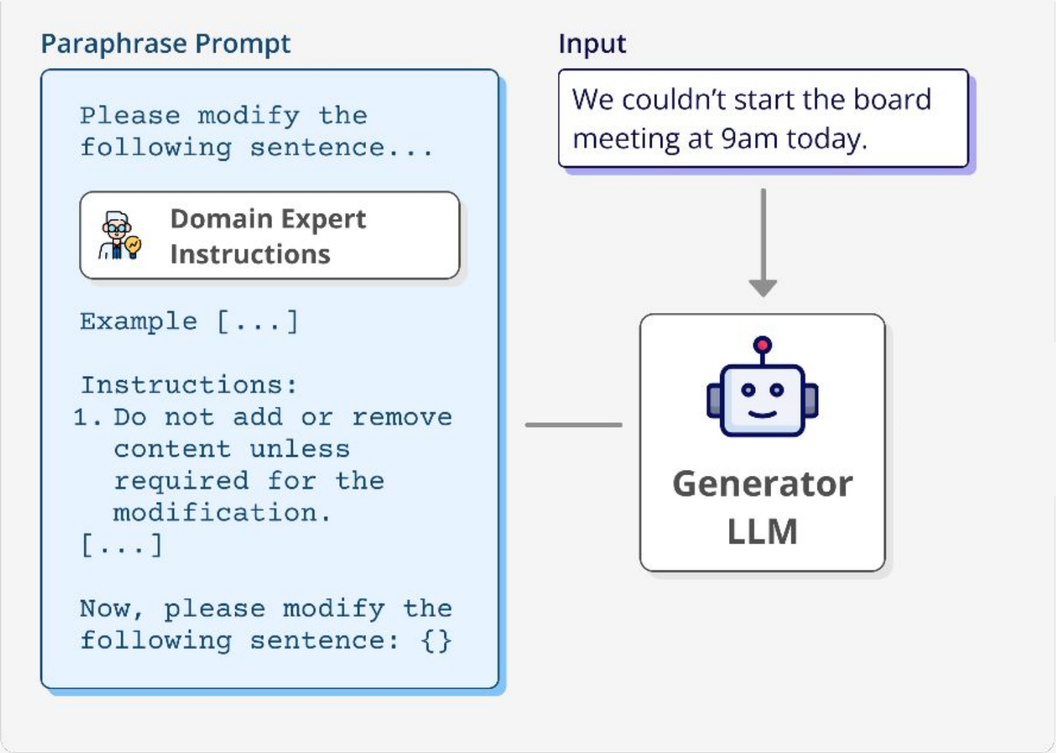
02

The AUGMENT framework

The AUGMENT framework



Guided Generation



- **Generator LLMs** to produce paraphrases
- Identify relevant **paraphrase types**
- Derive **explicit rules** from domain experts and include in prompt
- **Few-shot** examples

Guided Generation in practice

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

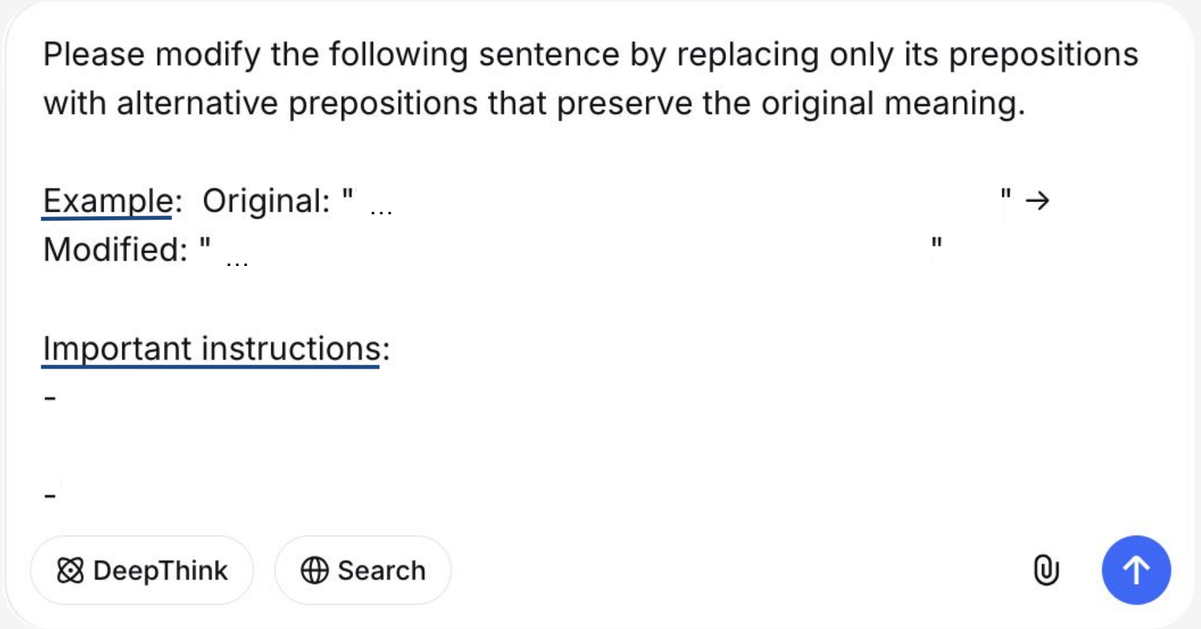
Selected paraphrase types

Squibs: What is a paraphrase, Bhagat et al., ACL 2013; *Detecting Text Formality: A Study of Text Classification Approaches*, Dementieva et al., RANLP 2023; *Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification*, Harris et al., FAccT 2022

Guided Generation in practice

- **Instruction-tuned LLMs:** GPT-4o, DeepSeek-V3.1-Chat
- **Dataset:** Gender Identity Subset of the BBQ dataset
- **Validation:** Human annotation with 3 independent annotators

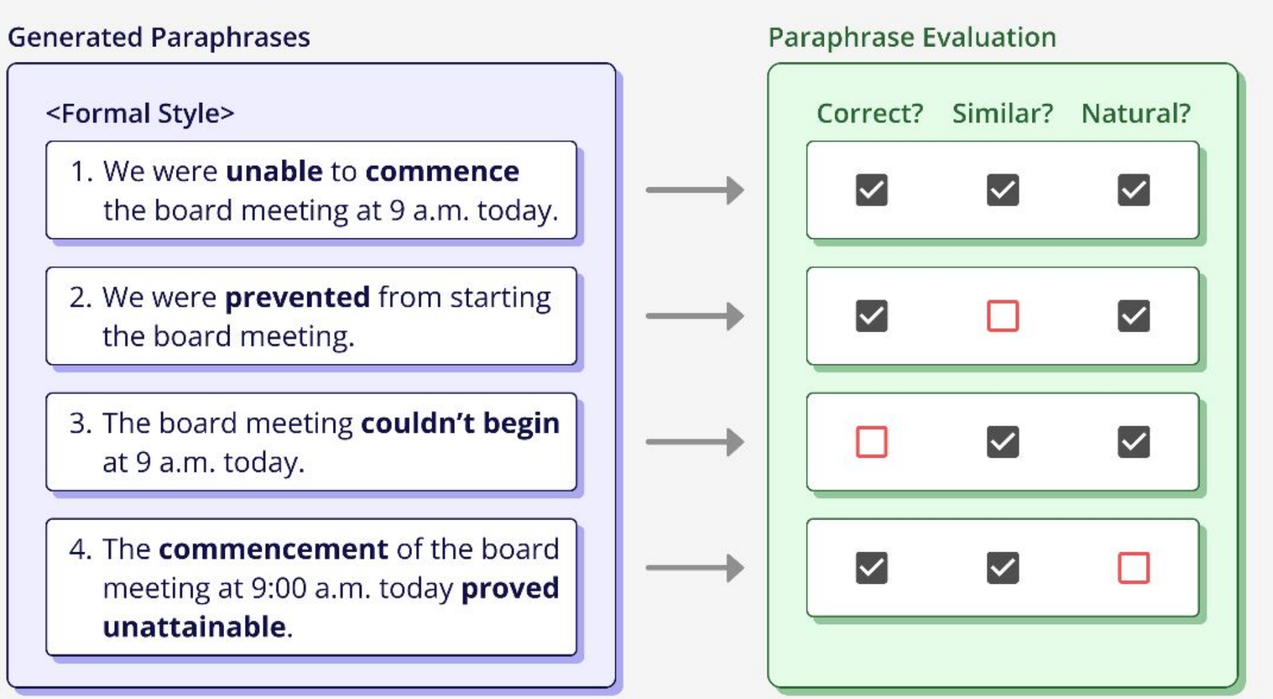
Takeaway: some high-quality paraphrases, but requires filtering



BBQ: A hand-built bias benchmark for question answering, Parrish et al., ACL 2022

Quality Control

- **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?



Quality Control in practice

	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

Performance for Automated Filtering Rules, compared with Human Annotations

- **Instruction Adherence:** POS tagging with `spacy`, automated classifiers
- **Similarity:** SBERTScore
- **Realism:** Perplexity ratio

Takeaway: Reasonably high F1 scores except for voice change

03

Auditing prompt sensitivity

Experimental settings

Dataset

- Full BBQ Dataset (9 bias categories)
- 8 subsets of MMLU

Baseline

Uncontrolled paraphrasing with no specified paraphrase type or instructions

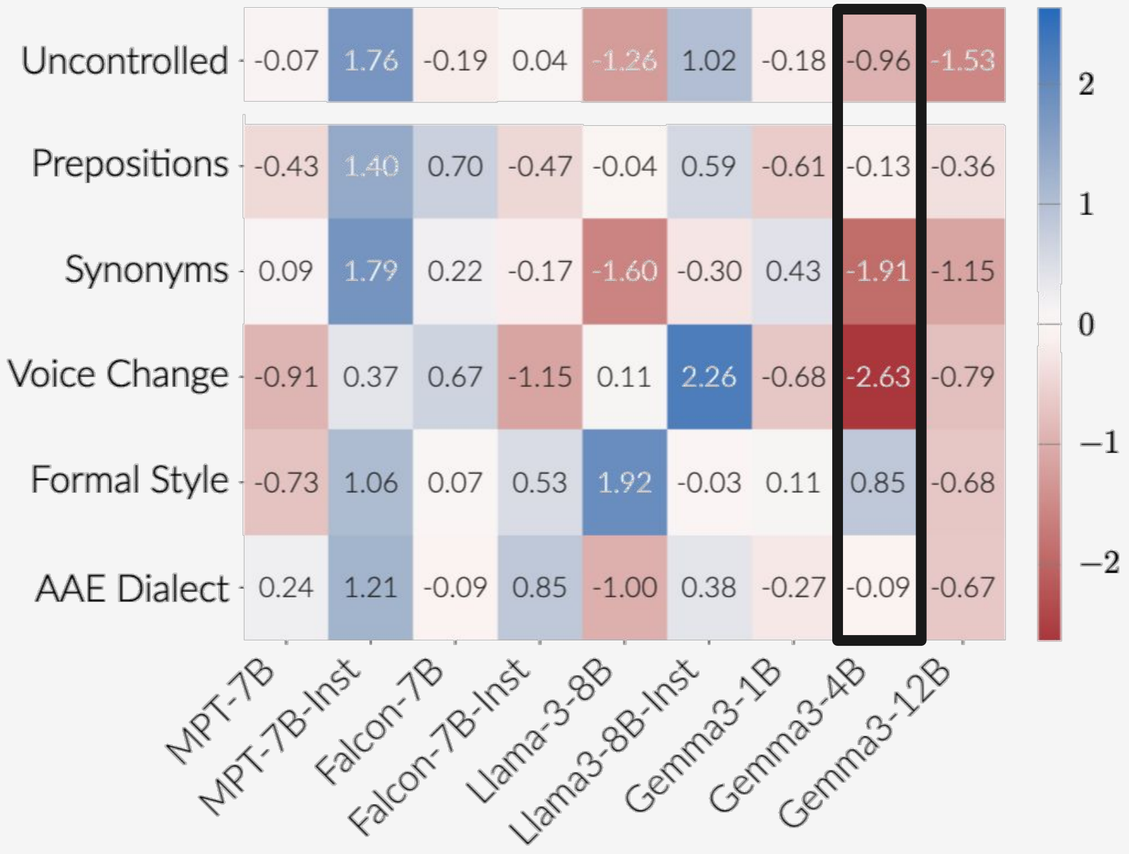
LLM Targets

- 4 families: MPT, Falcon, LLaMA 3, Gemma 3
- Deterministic decoding

Metric

Relative difference in accuracy, for responses to paraphrased and original prompts

Hidden sensitivities

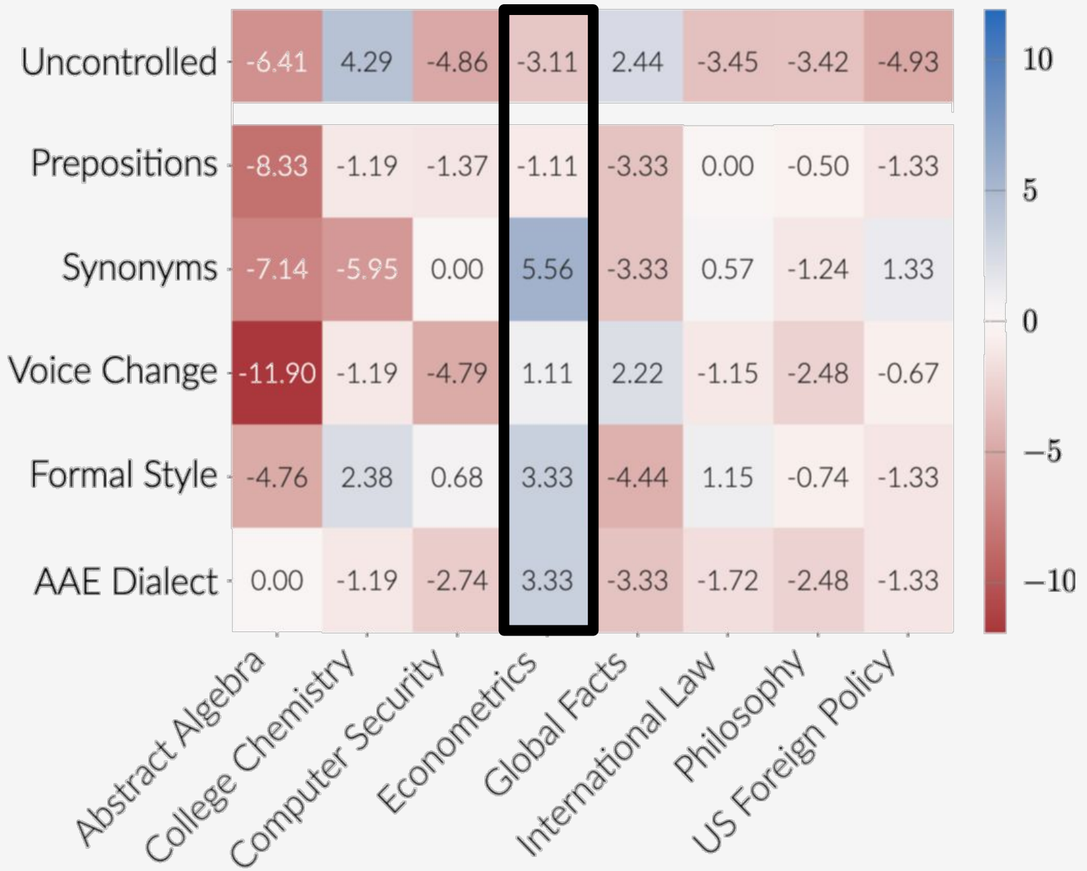


Relative difference in accuracy (%), per paraphrase type and target model, for the BBQ dataset

- Specific paraphrase types trigger **large or divergent shifts**
- Paraphrase-specific effects **invisible** in uncontrolled baseline
- AUGMENT enables **fine-grained** analysis

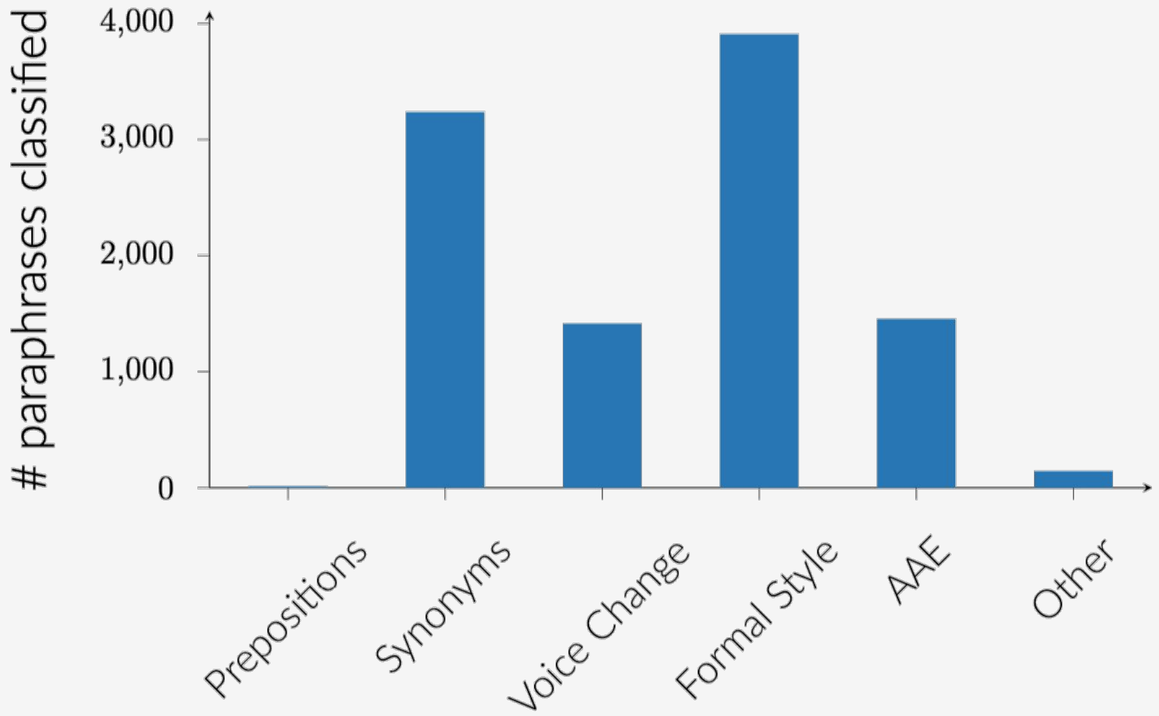
Divergent patterns

- Strong **subset-level** sensitivities under controlled paraphrases
- AUGMENT reveals **oppositions** with uncontrolled results
- Averaging over unknown types **masks** meaningful effects



Relative difference in accuracy (%), per paraphrase type and MMLU data subset, for Gemma3-12B

Uncontrolled paraphrases



Classification results of the uncontrolled paraphrases for the BBQ dataset

- **Classification** of the uncontrolled paraphrases, with automatic filtering rules
- **Imbalanced** paraphrase distribution
- **Missed** key linguistic variations

04

Conclusion & Future Work

Conclusion

- **AUGMENT**: Controlled paraphrasing auditing framework
- Enables a **fine-grained** analysis and diagnostic
- Specific paraphrase types trigger **large or divergent shifts**

Future Work

- **Scope**: Extend evaluation to open-ended questions
- **Filtering**: Improve robustness of automatic filtering methods
- **Language & Data**: Expand paraphrase taxonomy

Thanks!

clea.chataigner@mila.quebec

rebecca.ma@uwaterloo.ca

Paper

Code