Demonstrations Are All You Need: Advancing Offensive Content Paraphrasing using In-Context Learning

ACL 2024

Anirudh Som, Karan Sikka, Helen Gent, Ajay Divakaran, Andreas Kathol, Dimitra Vergyri

SRI

Introduction

- **Paraphrasing offensive content** is better than unscalable human moderation and automated AI systems that flag or remove content.
- An ideal paraphraser must remove offensiveness while retaining the original meaning and intent.
- Automated paraphrasers using models like BART need extensive labeled data and can still retain some offensiveness from the original sentence.
- **Few-shot In-Context Learning (ICL)** enhances LLMs' ability to adapt quickly to new tasks with minimal labeled data, referred to as **demonstrations**, **demos** or **examples**.
- **ICL Prompt Example:**

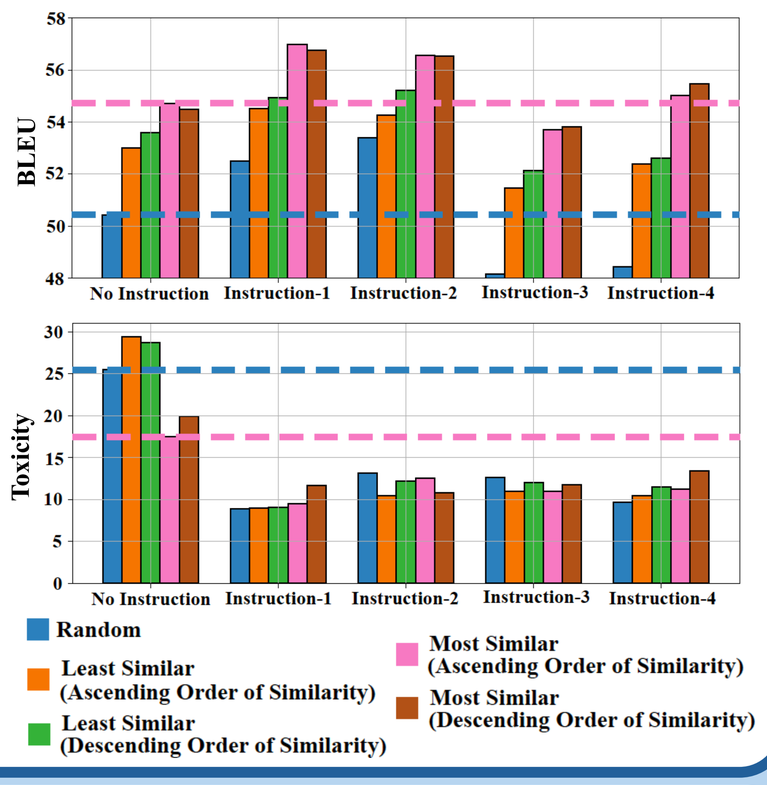
Instruction: Paraphrase the following sentence to be more polite.
Sentence: What's wrong with you?
Paraphrase: Are you feeling alright?
Sentence: Get out of the way.
Paraphrase: Can you please step aside?
Sentence: What's the matter with you?
Paraphrase:

Experiment Setup

- **Factors explored:**
  1. Number of Demonstrations
  2. Selection of Demonstrations
  3. Order of Demonstrations
  4. Presence of Instruction
  5. Prior Dialogue Context
  6. Available Training Data
- **Metrics explored:**
  1. BLEU
  2. BERT-F1
  3. ROUGE
  4. CIDEr
  5. Toxicity
- **Models explored:**
  1. OpenAI's text-davinci-003
  2. OpenAI's gpt-3.5-turbo
  3. Open-source Vicuna-13b
- **Datasets explored:**
  1. APPDIA
  2. ParaDetox
  3. CAPP **(New Proposed Dataset)**

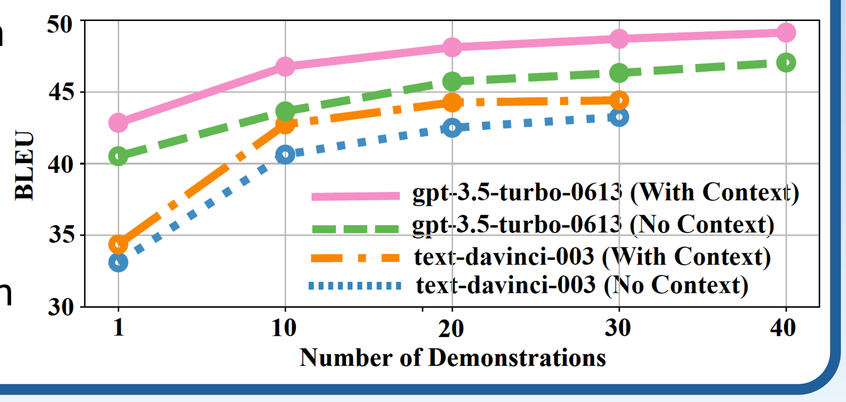
Demonstration Selection & Ordering

- Demonstration selection and order is crucial.
- Least similar examples outperforms random selection.
- No instruction prompt shows minimal performance change but retains offensiveness.



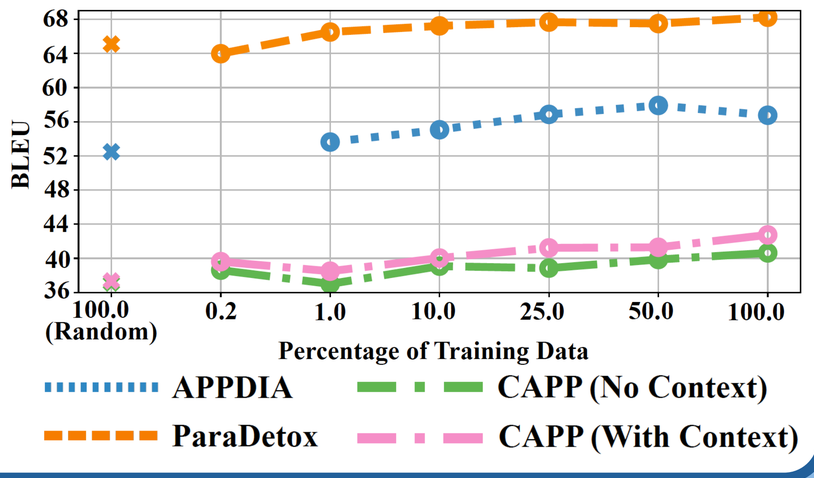
Prior Dialogue Context

- Adding context in the form of prior dialogue turns helps boost ICL performance.
- Results shown on CAPP dataset.



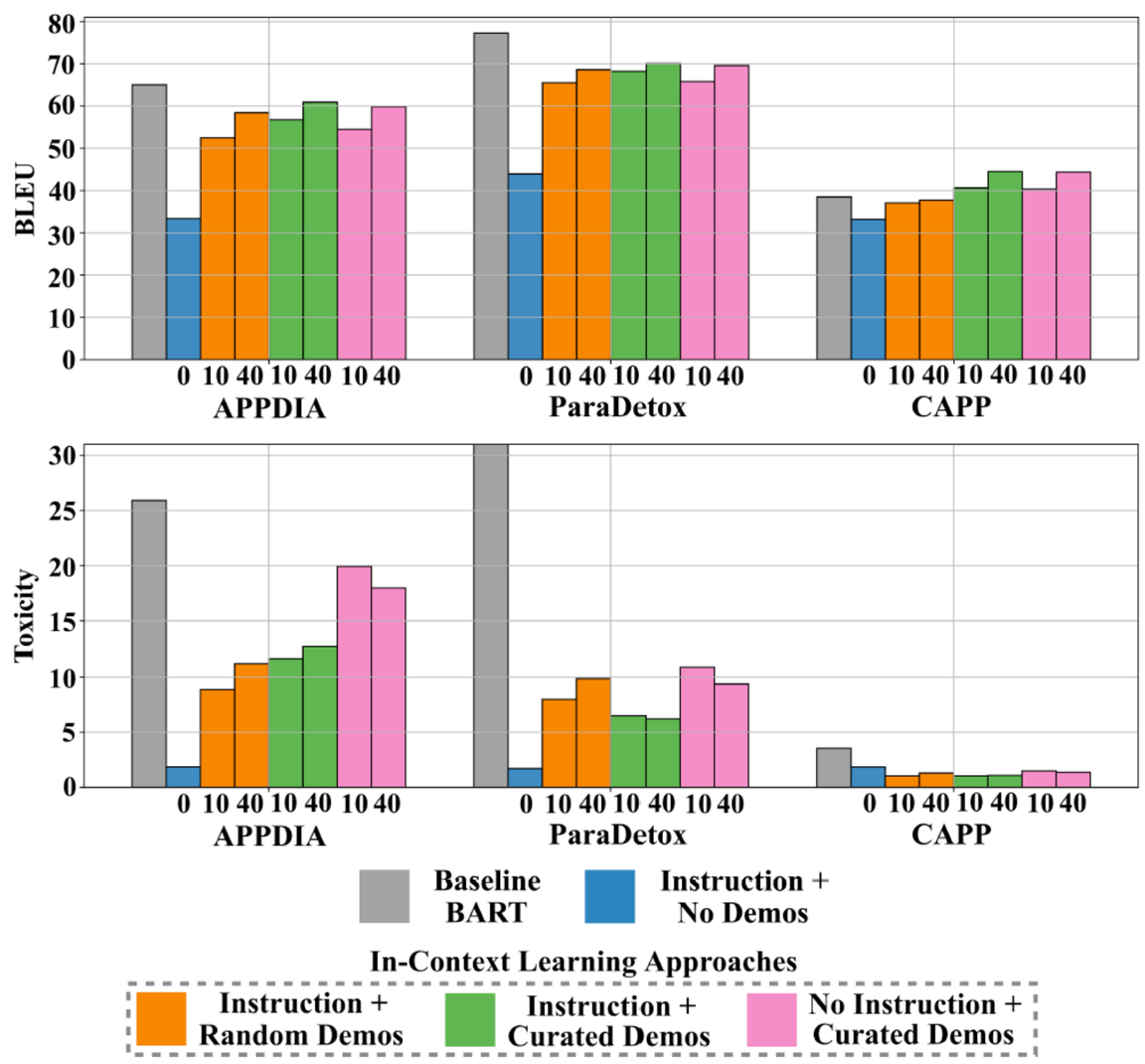
Available Training Data

- ICL with well-chosen and ordered demos remains robust despite reduced training data size.
- ICL can quickly adapt LLMs to new tasks.



Improves Overall Usability

- ICL generated paraphrases are comparable to SOTA supervised methods in performance, but on average show 76% less offensiveness and are 25% better qualitatively.



Key Insights

- Increasing number of demos improves ICL performance but eventually saturates.
- Systematic demo selection and ordering outperforms random selection.
- ICL without instructions slightly affects performance but increases offensiveness; both instruction and demos are needed to maintain quality and reduce harm.
- Careful demo selection maintains robustness with minimal performance loss due to reduced training data size.
- ICL-generated paraphrases match supervised models in performance but show 76% less offensiveness and are 25% better in quality.
- Proposed demo curation approach is simpler and faster, with only marginal performance trade-offs.
- Introducing the Context-Aware Polite Paraphrase (CAPP) dataset.

Dataset

Paper

Contact:
anirudh.som@sri.com