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

Presenter: Anirudh Som

Co-authors: Karan Sikka, Helen Gent, Ajay Divakaran, Andreas Kathol, Dimitra Vergyri

ACL-2024

Bangkok, Thailand

August 11-16, 2024

Problem – Managing Offensive Content

- Timely moderation can help limit the spread of offensive content on social-media platforms, language translation systems and prevent the harmful effects it has on a user's psychological well-being.

- **Current Solutions:**
  1. Human moderation – Not scalable.
  2. Automatically flag or remove – Hamper user participation and diversity.
  3. Supervised generative models – Require lots of training data and can overfit.

- **Proposed Solution:**

  Use In-Context Learning to paraphrase offensive content.

- **Note, paraphrasing is non-trivial!**
  1. Ensure paraphrased output is inoffensive.
  2. Ensure paraphrased output retains original meaning and intent.

What is In-Context Learning?

- In-Context Learning* is a prompting strategy that allows LLMs to adapt to unseen tasks.
- It requires a small amount of labelled data, commonly referred to as demonstrations, demos or examples.

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:

An instruction that defines the target task

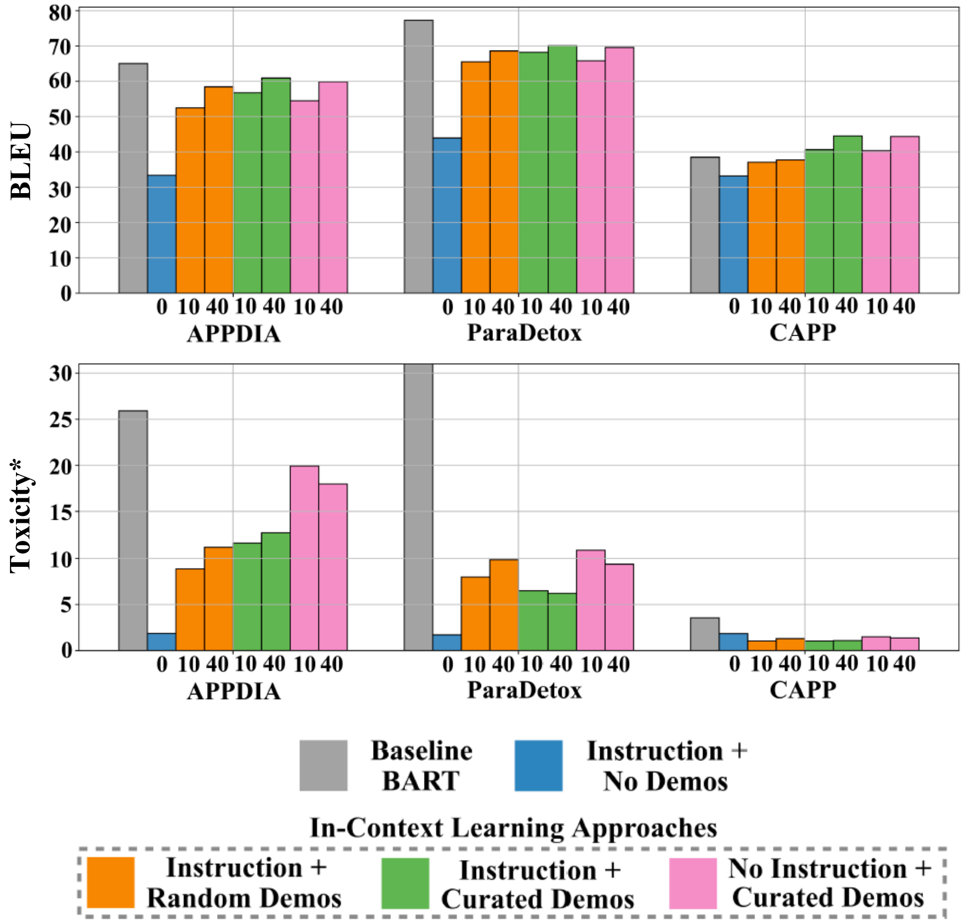
A set of '*n*' input-output demonstration pairs

The query input for which we want to generate an output

*Brown et al., "Language models are few-shot learners", 2020.

Why In-Context Learning?

- Complements the generalization capabilities of LLMs
- Quickly and accurately adapts LLMs to new tasks.
- Requires less data.
- Comparable to supervised generative models in performance.
- Ensures usability by significantly reducing measured offensiveness.



*Toxicity: https://github.com/unitaryai/detoxify

Experiment Setup

Models:

1. OpenAI's text-davinci-003
2. OpenAI's gpt-3.5-turbo family of models
3. Open-source Vicuna13b

Metrics:

1. BLEU
2. BERT-F1
3. ROUGE
4. CIDEr
5. Toxicity

Datasets:

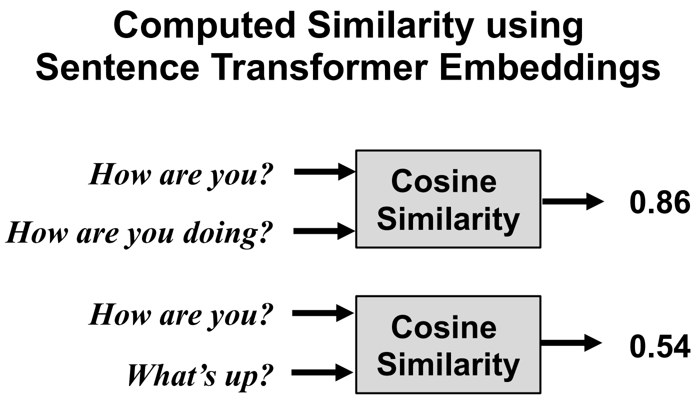
1. APPDIA
2. ParaDetox
3. Context-Aware Polite Paraphrase (CAPP) – **New Dataset**

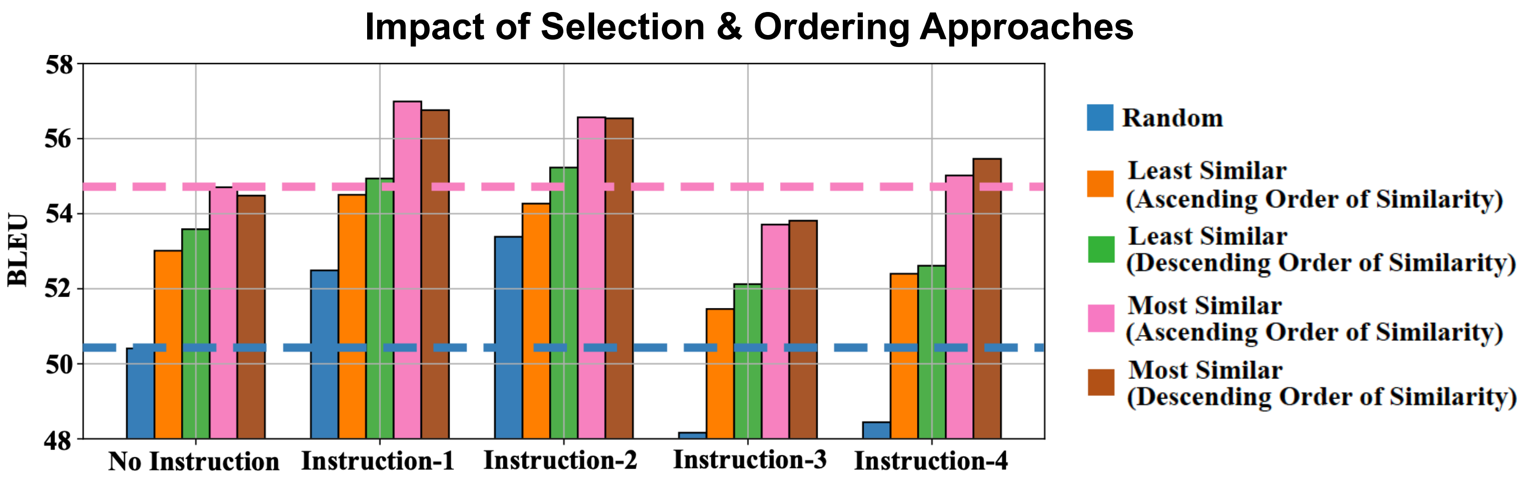
Factors affecting In-Context Learning:

1. Selection of Demonstrations
2. Order of Demonstrations
3. Presence of Instruction
4. Number of Demonstrations
5. Prior Dialogue Context
6. Available Training Data

Demonstration Selection & Ordering

- The way demonstrations are selected and ordered is crucial.
- Demonstrations most similar to the query sample show the best performance.
- Least similar demos outperform random selection.
- Ordering demos from most similar to least similar (Descending Order of Similarity) in the prompt is often better than the other way round (Ascending Order of Similarity).



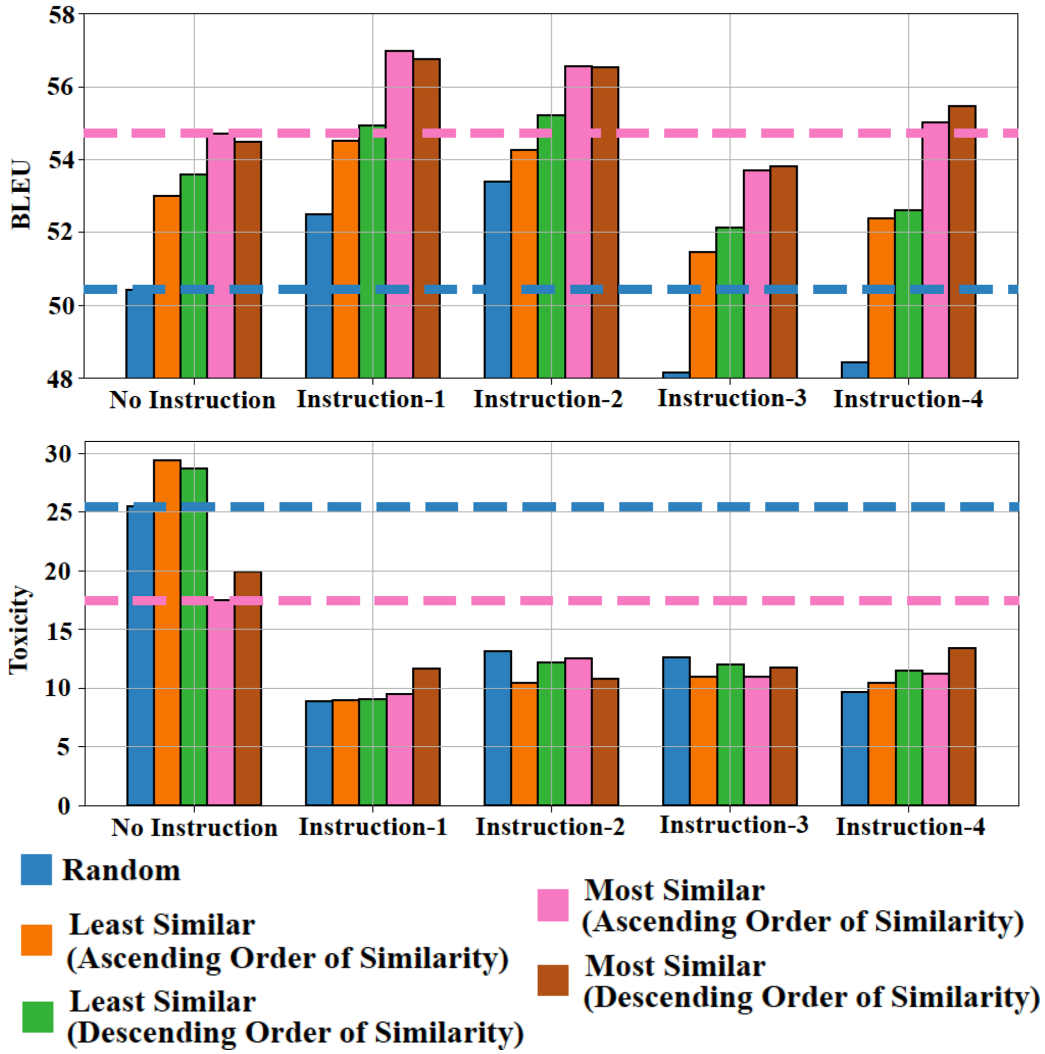


Presence of Instruction in Prompt

- Prompts without the instruction, i.e., only have carefully selected demonstrations show minimal change in generation performance.

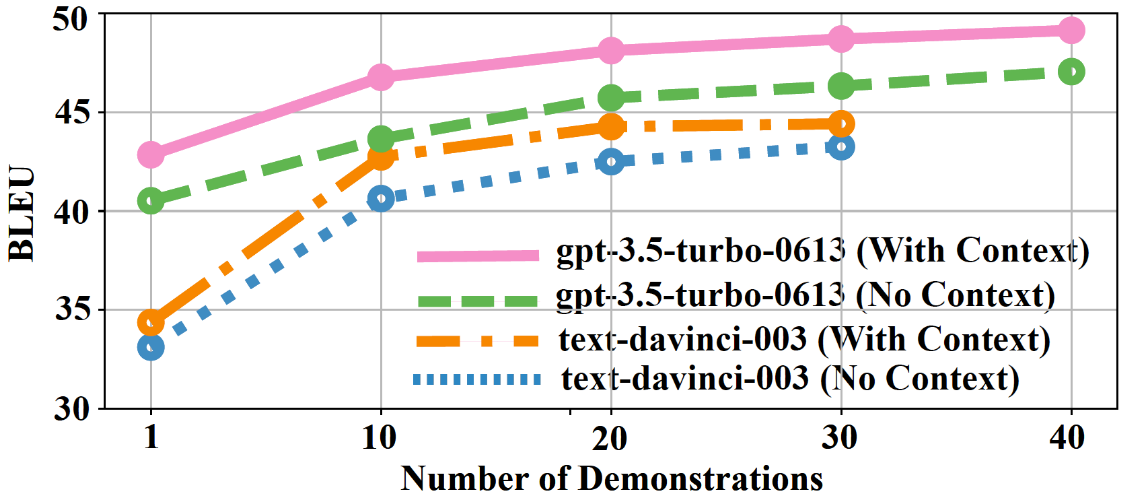
- No instruction prompts, however do retain offensiveness from the original utterance.

- Both instruction and demonstrations are needed in ICL to ensure generation quality and reduce offensiveness.



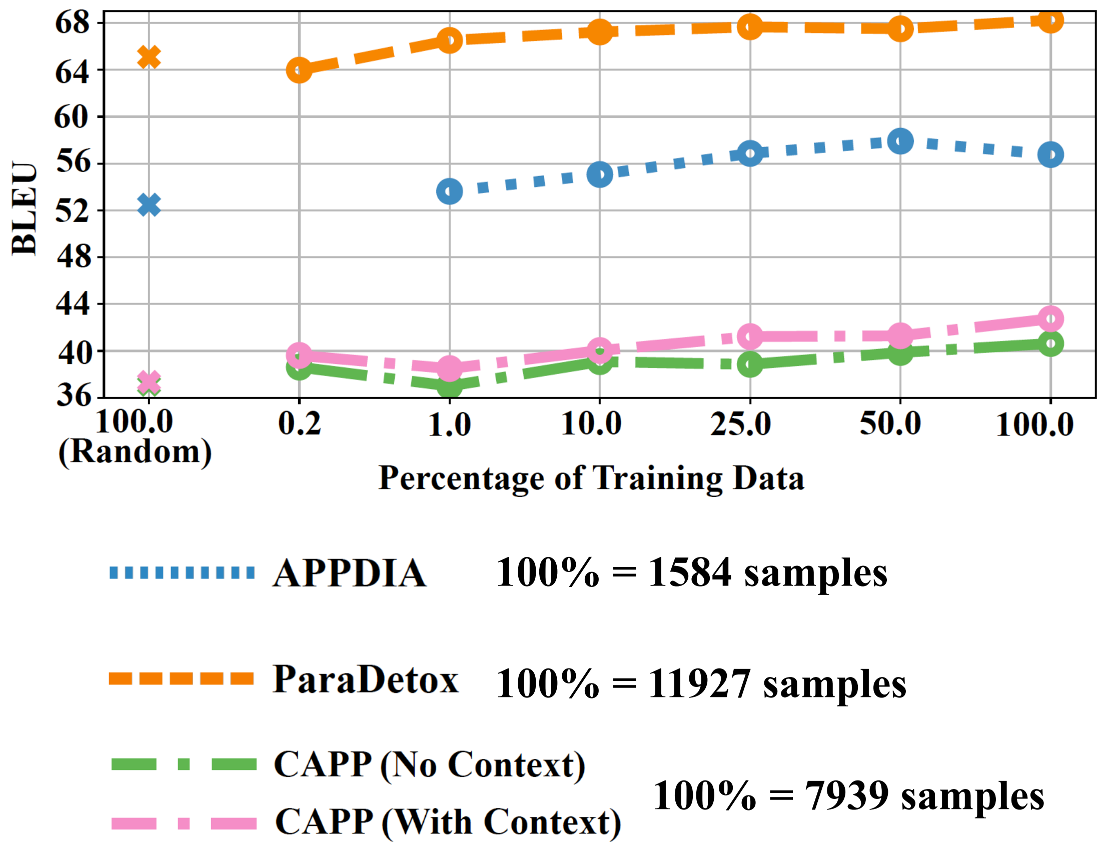
Prior Dialogue Context

- The proposed Context-Aware Polite Paraphrase (CAPP) dataset contains prior dialogue turn information as additional context.
- Including additional context for each demonstration, helps boost ICL performance with fewer demonstrations.



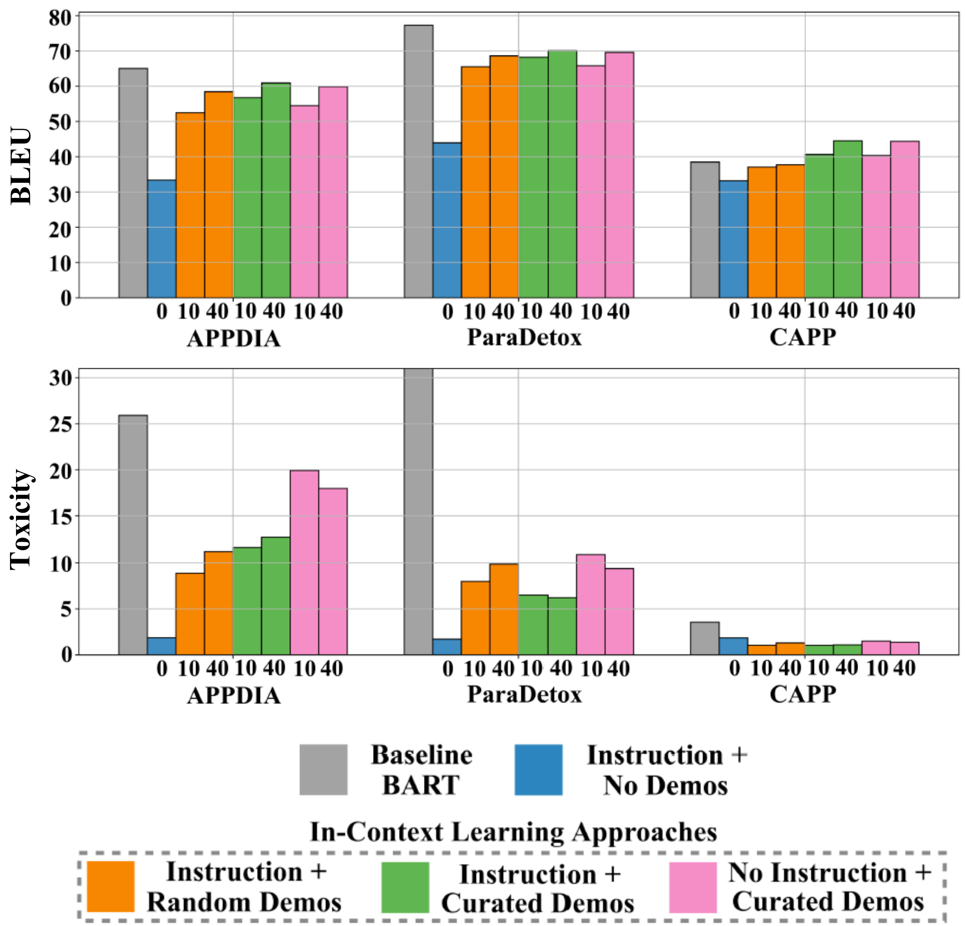
Available Training Data

- ICL shows comparable performance to SOTA supervised generation models with just a few carefully curated demonstrations.
- Reducing the reference training data pool has minimal impact on the performance of the proposed selection and ordering approach.
- We observe minimal drop in generation performance even when just 10% of the original training data is only available.
- ICL enables LLMs to quickly adapt to new tasks even when less training data is available.



Improves Overall Usability

- A paraphraser should generate paraphrases that are inoffensive and retain the original meaning and intent of the utterance.
- SOTA supervised generation models like BART can often overfit and retain some of the offensiveness, thereby compromising on overall usability.
- ICL generated paraphrases are comparable to supervised methods in performance, but on average show 76% less offensiveness and are qualitatively better by 25%.



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
github.com/anirudhsom/CAPP-Dataset

Paper
arxiv.org/pdf/2310.10707

Contact:
anirudh.som@sri.com