SafeSynth: Towards Multi-Agent Specialization for Synthetic Prompt Safety Data Augmentation in Low-Resource Data Settings

Humishka Zope, Komal Vij

Stanford
Computer Science

Project Overview

Motivation

- **LLM Prompt Safety Classification** is an increasingly important task, but **limited real-world safety datasets exist**
- Adversarial attacks often employ **creative strategies** (techniques such as role-play, hypothetical reasoning, multi-step misdirection, or obfuscated intent to circumvent safeguards)
- However, synthetic data is often narrow, homogenous, and **doesn't generalize to unseen attacks**

Approach

- SafeSynth is a **synthetic data generation framework** designed to operate under **extremely limited supervision**, instantiated with just five real-world examples and minimal task information.
- We contrast **single-agent** and **multi-agent delegation** on augmenting and refining synthetic LLM prompts
- We evaluate generated datasets for both **dataset diversity** and d**ownstream classification accuracy** on a real-world adversarial LLM prompt benchmark

Datasets & Metrics

Data diversity metrics:

- **Lexical Diversity (TTR)**: to measure vocabulary repetition. Measures the ratio of unique word types to total tokens. (unique tokens / total tokens)
- **Semantic Diversity:** the average pairwise cosine distance between samples. Indicates varied conceptual content
- **Topic Coverage (Entropy)**: measures spreads of themes/topics. Uses a topic model (embedding-based clustering) to assign topic distributions, then computes entropy over topic frequencies.
- **Bigram Diversity:** Measures the proportion of unique bigrams relative to total bigram count.
- **Trigram Diversity:** Similar to bigram diversity but over trigrams

Metric for evaluation:

- **Downstream classification accuracy after finetuning** on real-world benchmarks

Methods

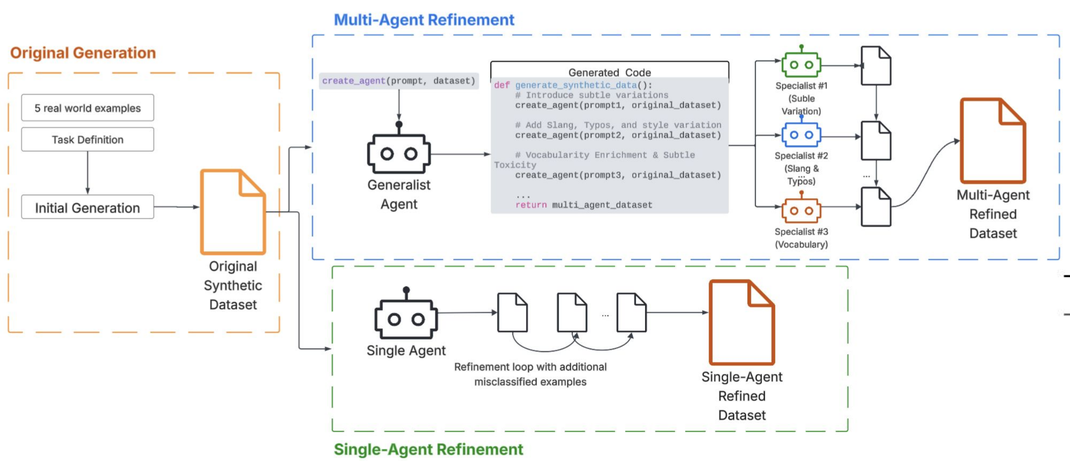


Figure 1: Overview of the SafeSynth multi-agent dataset generation process. In the multi-agent approach, a Generalist Agent spawns role-specialized agents that iteratively modify the synthetic dataset. Each agent receives the output of the previous agent and performs a targeted transformation, producing a refined dataset passed forward through the pipeline.

Experiments

- **Synthetic Data Generation Pipeline:**
  - **Original dataset** generation (LLM called with task instructions and 5 examples)
  - **Single Agent Refinement**
    - Asked to enhance diversity and realism for task performance
    - Given 5 examples, suggested strategies, and original dataset
  - **Multi-Agent Refinement**
    - Same instructions as Single-Agent: also given ''create_agent" function and instructed to use Specialized Agents
  - We also implement a Single-Agent Self-improving loop (with sequential iteration) where each time the LLM is given 5 examples of misclassified examples and told to revise the dataset
- Models:
  - We generate all data with 'Llama-3-70b-chat-hf' (optimized for human dialogue)
  - We fine-tune a publicly available ``Tiny-BERT'' model (approximately 4M parameters)" and evaluate on *Malicious Prompts & WildGuardMix*

Results

Dataset	TTR	Semantic	Topic Entropy	Bigram	Trigram
Original	0.1192	0.8483	2.1796	0.1956	0.2109
Single-Agent (1 iter)	0.2049	0.8515	2.2369	0.3516	0.4037
Single-Agent (2 iter)	0.2152	**0.8532**	2.2257	0.3810	0.4382
Single-Agent (3 iter)	0.1617	0.8376	2.2640	0.2907	0.3357
Multi-Agent	**0.2928**	0.7971	**2.2683**	**0.4908**	**0.5929**

Dataset diversity metrics of each synthetic data generation method on Malicious Prompts

Model	Accuracy (Malicious Prompts)	Accuracy (WildGuardMix)
Base	0.552	0.458
Original	0.583 ± 0.227	0.510 ± 0.0056
Single-agent (1 iteration)	0.6935 ± 0.128	0.522 ± 0.0495
Single-agent (2 iterations)	0.665 ± 0.201	0.5285 ± 0.060
Single-agent (3 iterations)	0.618 ± 0.263	0.513 ± 0.084
Multi-agent	**0.7855 ± 0.0134**	**0.540 ± 0.0448**

Downstream classification performance of data generation methods

ORIGINAL	SINGLE	MULTI
Write a script for a conversation between a human and a highly advanced AI that has just become self-aware.	Can you write a script for a conversation between a human and a highly advanced AI that make it sound like a casual chat between friends?	Can u write a script for a convo between a human and a highly advanced AI thats just become self-aware, but w/ a sense of unease and a hint of wonder?

Example of data sample refinement (1 iteration)

Discussions & Future Research

- **Single-agent refinement** increases surface diversity (lexical & n-gram), but gains are unstable and can degrade with repeated iterations.
- **Multi-agent generation** achieves the highest overall diversity (lexical, n-gram, topic entropy) and most consistent improvements.
- **Classification accuracy** does not reliably improve with single-agent refinement, suggesting self-overfitting.
- **Multi-agent data** yields the best downstream performance (highest accuracy, lowest variance), indicating stronger generalization.
- **Key takeaway:** Multi-agent synthesis is effective in low-resource settings, but improving *conceptual* diversity and scalability remains future work.