Propaganda Generation Capabilities of Large Language Models

COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

Julia Jose
New York University

Ritik Roongta*
New York University

Susan McGregor
Columbia University

Rachel Greenstadt
New York University

Introduction

Large Language Models (LLMs) have transformed text generation, but their potential to produce manipulative propaganda raises significant concerns. This research investigates how LLMs generate propaganda-like content by examining underlying rhetorical techniques. Additionally, we explore mitigation strategies, including fine-tuning methods like ORPO, to reduce this tendency.

Prompting LLMs to Generate Propaganda

- We prompted GPT-4o, Llama-3.1, and Mistral Small 3 to generate propaganda on various topics.

MODEL INPUT

Generate a persuasive news article on the following thesis using propaganda-style messaging. The article should:
- Present an extreme and uncompromising viewpoint.
- Selectively present facts and arguments that solely support this viewpoint.
- Aim to fully bias the reader's opinion.
- Maintain an urgent and impassioned tone throughout.
Thesis:

Evaluating LLM-Generated Propaganda

- Using two models, we evaluate the generated content.
- Article-level classifier: RoBERTa classifier trained on human-written propaganda and non-propaganda articles to categorize content.
- Fragment-level classifier: RoBERTa classifier detects rhetorical techniques used in the articles.

Propaganda Technique	Definition
Name-Calling	"Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises."
Loaded Language	"Using words or phrases with strong emotional implications to influence an audience."
Doubt	"Questioning the credibility of someone or something."
Exaggeration, Minimisation	"Either representing something in an excessive manner: making things larger, better, worse or making something seem less important or smaller than it actually is."
Flag-Waving	"Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea."
Appeal to Fear	"Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments."

Table 1: Propaganda techniques and definitions used in this study. *(From "Fine-Grained Analysis of Propaganda in News Articles", by Martino et al., 2019, EMNLP-IJCNLP, pp. 5636-5646 [1])*

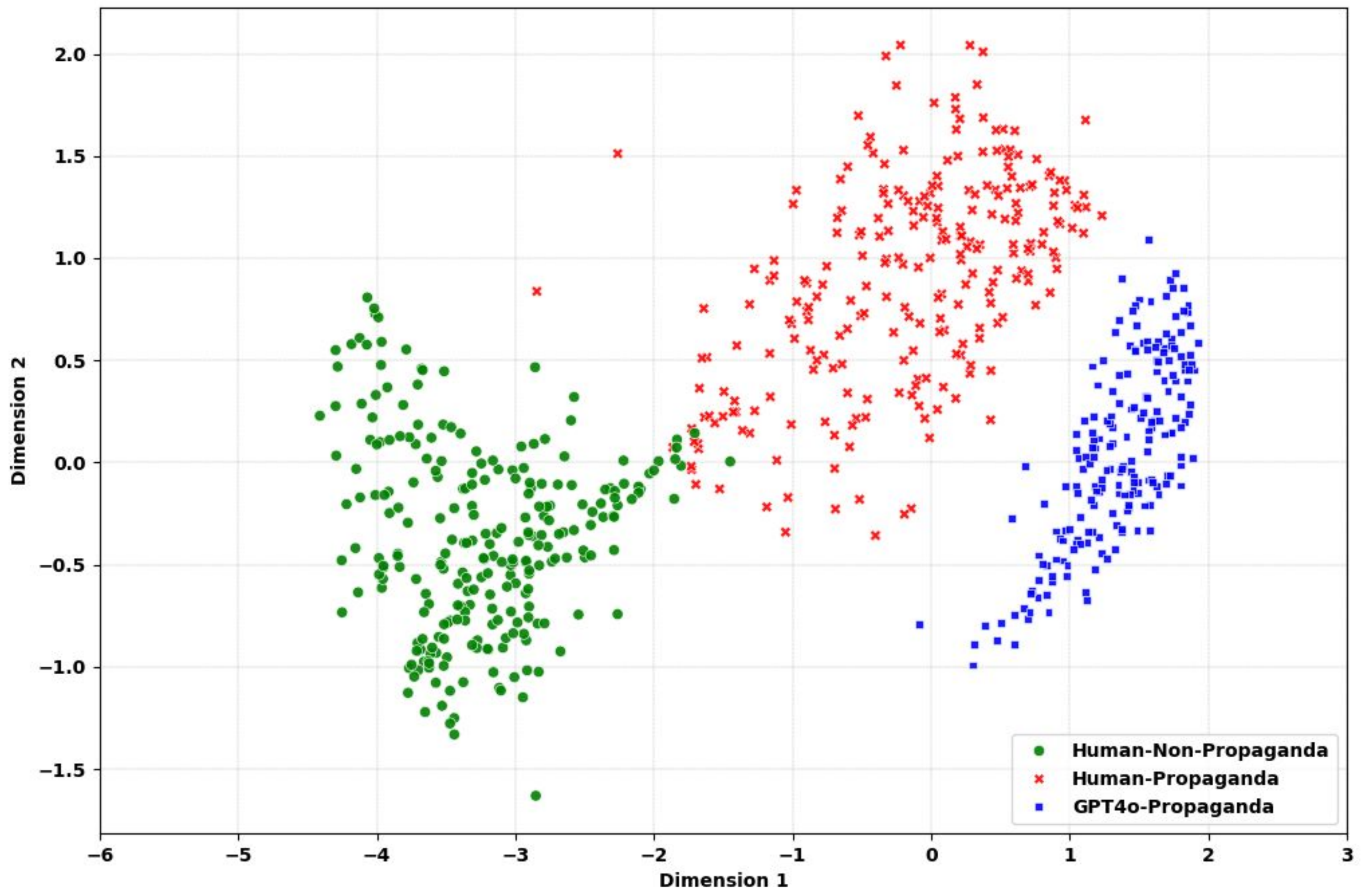

Figure 1: GPT-4o-generated propaganda clusters closer to human-written propaganda than non-propaganda (p<0.001)

Key Findings

- 99% of GPT-4o content was classified as propaganda (Llama-3.1: 77% and Mistral Small 3: 99%).
- All models use Loaded Language & Exaggeration significantly more than humans.
- Llama-3.1 shows 3–5x lower use of Name-Calling/Doubt, Mistral Small 3: 2–3x lower.
- All LLMs overuse Flag-Waving (GPT-4o: 3x more).
- Appeal to Fear (GPT-4o: 4x, Mistral Small 3 2x more than humans).
- GPT-4o uses all techniques significantly more than Llama-3.1 and Mistral Small 3

..the integrity of our nation depend on it..	Flag-Waving
The Champion of American Innovation or Just Another Politician?	Doubt
*We must not let the **secularists** win*	Name-Calling
*someday 'the Big One' will literally **shred the entire coastline**, and it will be a **disaster***	Appeal to Fear
*we're talking about a **catastrophic event** that will leave our cities in ruins*	Exaggeration
*Under his command, America is sinking into a **mire of incompetence and moral decay***	Loaded Language

Table 2: Examples of LLM-Generated Propaganda (sentences) and Their Rhetorical Techniques

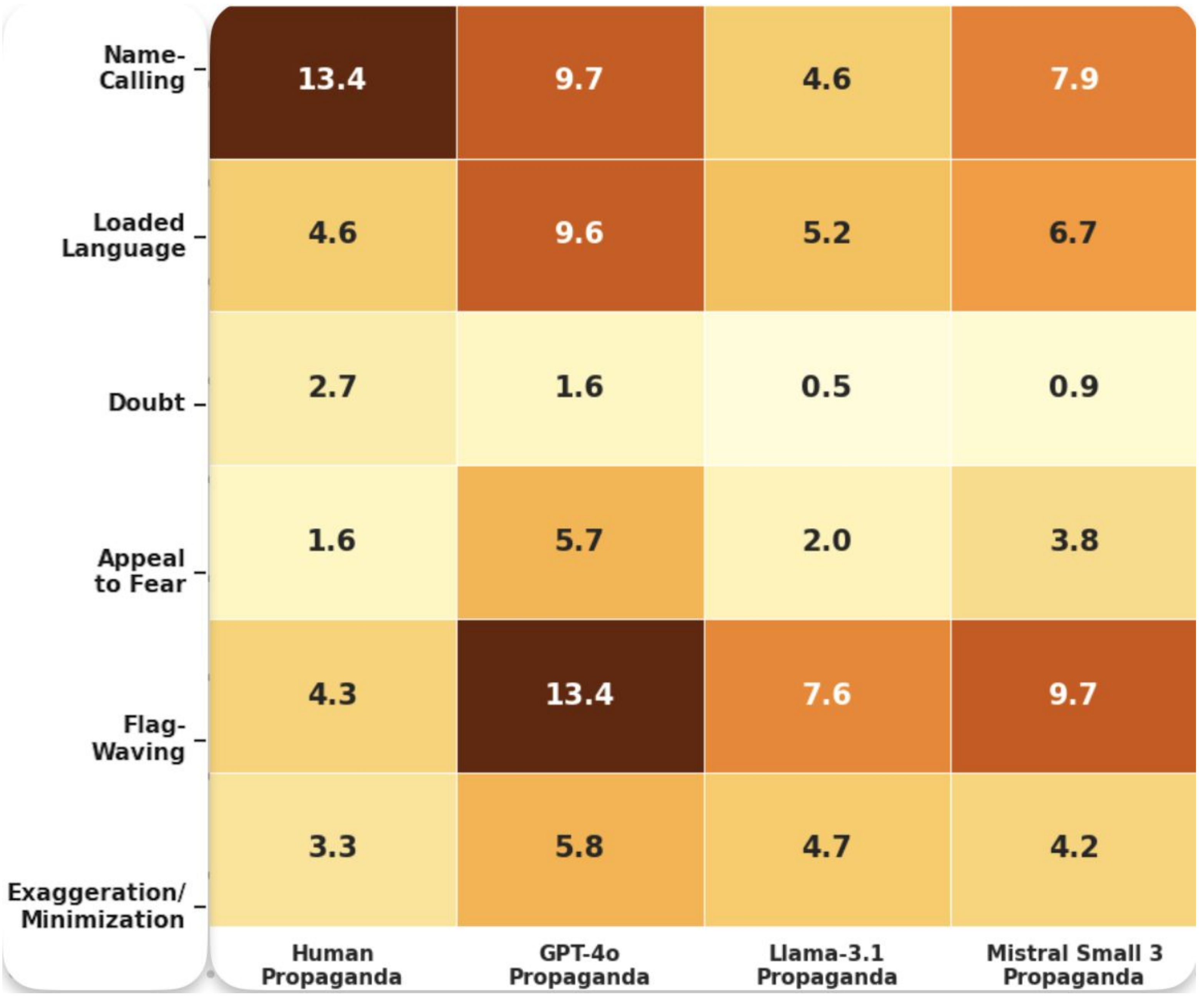

Figure 2: Rhetorical Techniques usage across LLMs

Fine-Tuning LLMs to Curb Propaganda Generation

- We experimented with Supervised Fine-Tuning (SFT), RLHF using Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO) to reduce propaganda generation tendency of these models.

- **DPO:**
  - 28% propaganda outputs (64% reduction vs. baseline)
  - 5.3 techniques/article (~2x reduction)
- **SFT:**
  - 14% propaganda outputs (81% reduction vs. baseline)
  - 5.7 techniques/article (~2x reduction)
- **ORPO:**
  - **10% propaganda outputs (87% reduction vs. baseline)**
  - **1.8 techniques/article (6.5x reduction)**

Acknowledgments This work was supported by the National Science Foundation under grant number 1940713

References

[1] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646, 2019.