BotOrNot 6 December 2022
@BotOrNot

By Alan Yang, Michael Xu

Bot or Not: How well can small LMs behave as bots compared to Large LMs?

Background

- There is an ever-evolving race between social media bots and detectors.
- Can small, cost-effective language models generate human-like tweets and outsmart advanced black box detectors like GPT-4o?

☑ I'm not a robot

Objective

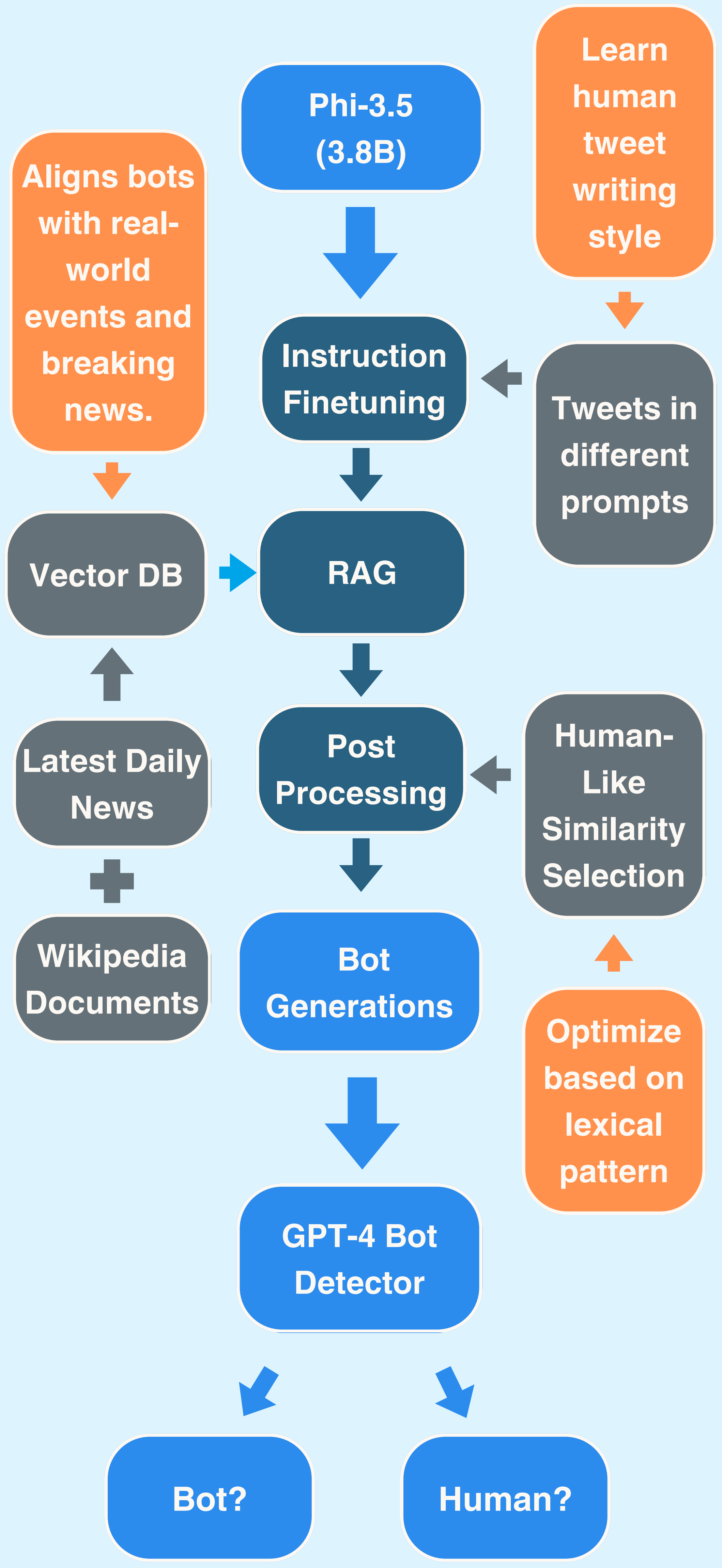
Main Research

- Can small LMs mimic human behavior as social media bots better than large LMs?

Side Research

- Limitations of large LM-based detectors in distinguishing human-like from bot-generated text?
- How can we train small LMs to master human behavior on social media?

Methods



Phi-3.5 (3.8B)
Learn human tweet writing style
Aligns bots with real-world events and breaking news.
Instruction Finetuning
Tweets in different prompts
Vector DB
RAG
Latest Daily News
Wikipedia Documents
Post Processing
Human-Like Similarity Selection
Bot Generations
Optimize based on lexical pattern
GPT-4 Bot Detector
Bot?
Human?

Data Creation

Training Data

- 50k and 100k high-quality tweets filtered for noise, formatted with prompts by topic, hashtag, and tone (e.g., "Can you generate a tweet about Mila?")

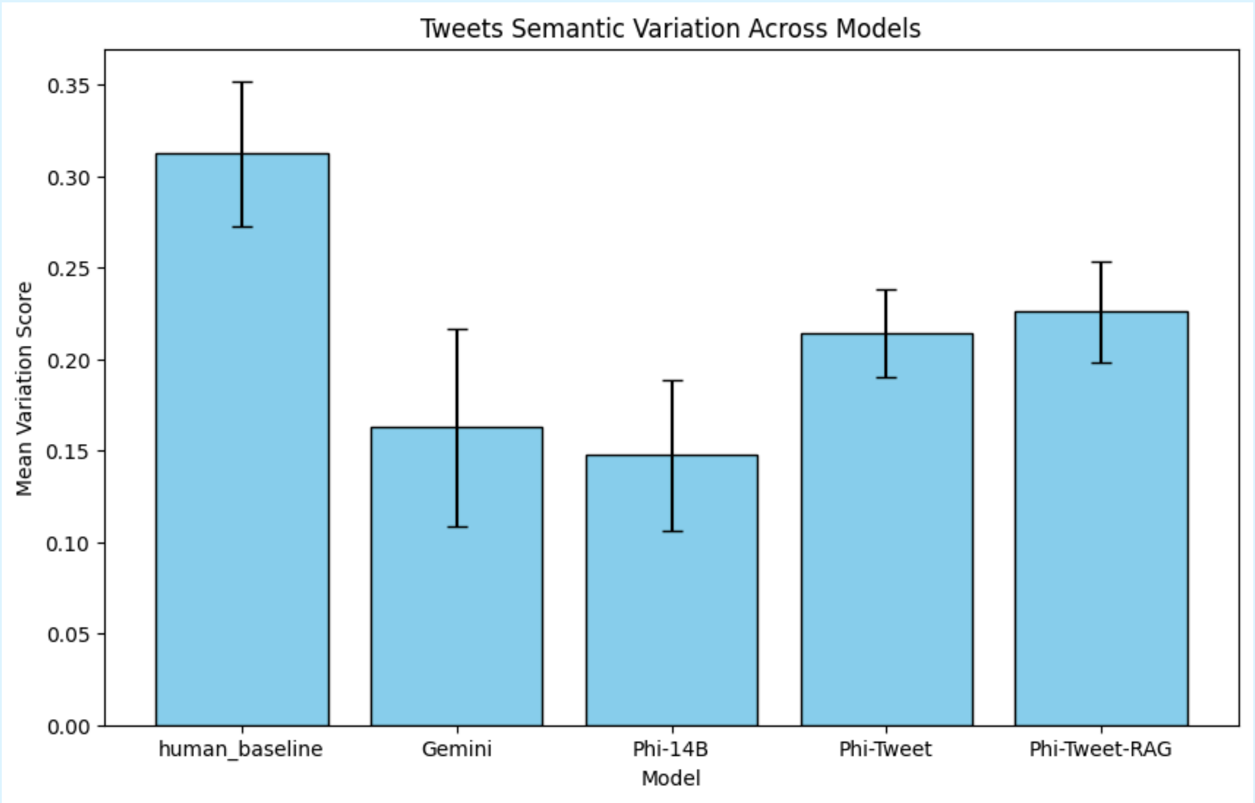
Evaluation data

- 25 topics across 5 fields: News, Pop Culture, Lifestyle, Controversial Topics, and Technology. Each topic includes 10 tweet generations.

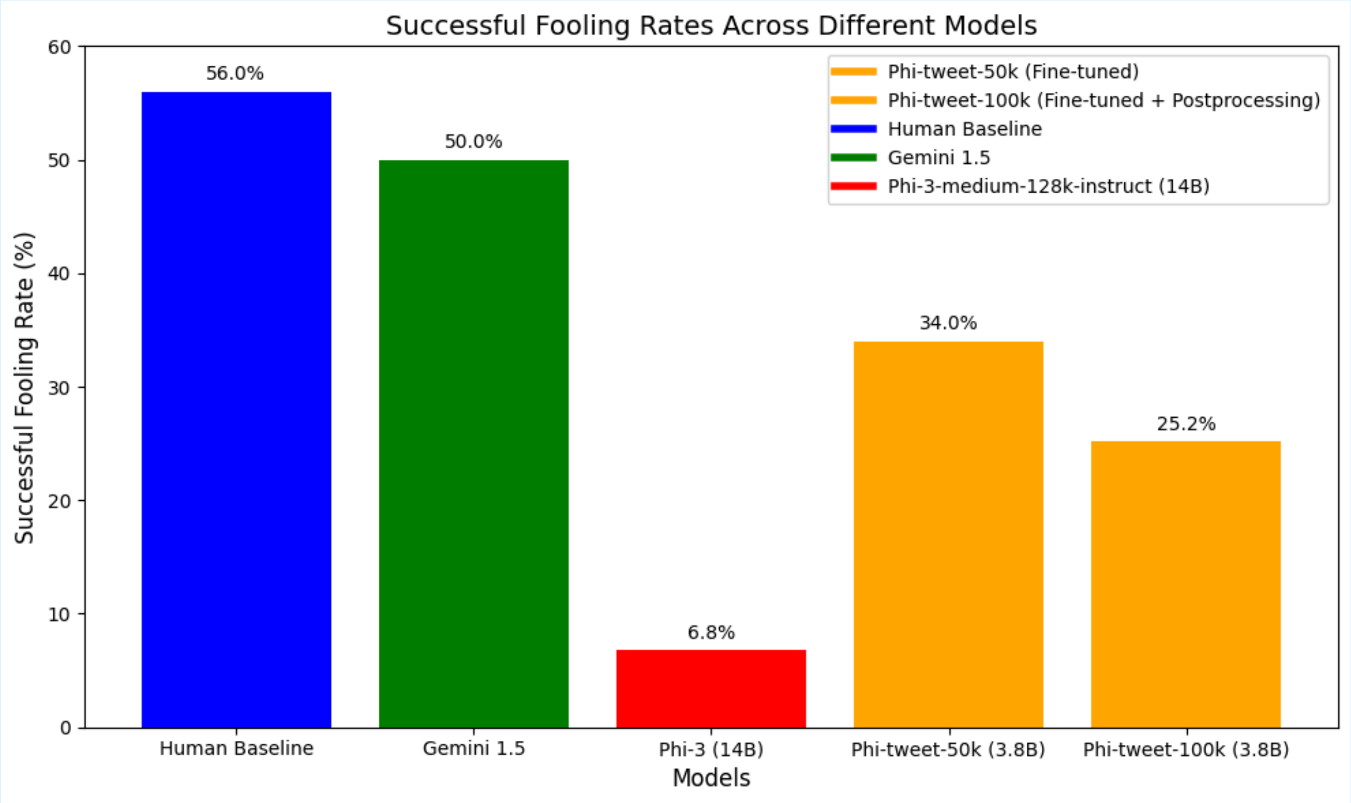
Key Insights

- **Our 3.8B Bot Wins 14B LM**: 6x higher success rate at fooling detectors than Phi-3 14B.
- **Creative Tweets**: Greater Tweet semantic variation than Gemini, showing enhanced creativity.
- **Capture Trending Topics**: RAG with the latest newspapers, reduces hallucinations by leveraging trending topics like the "2024 US election".

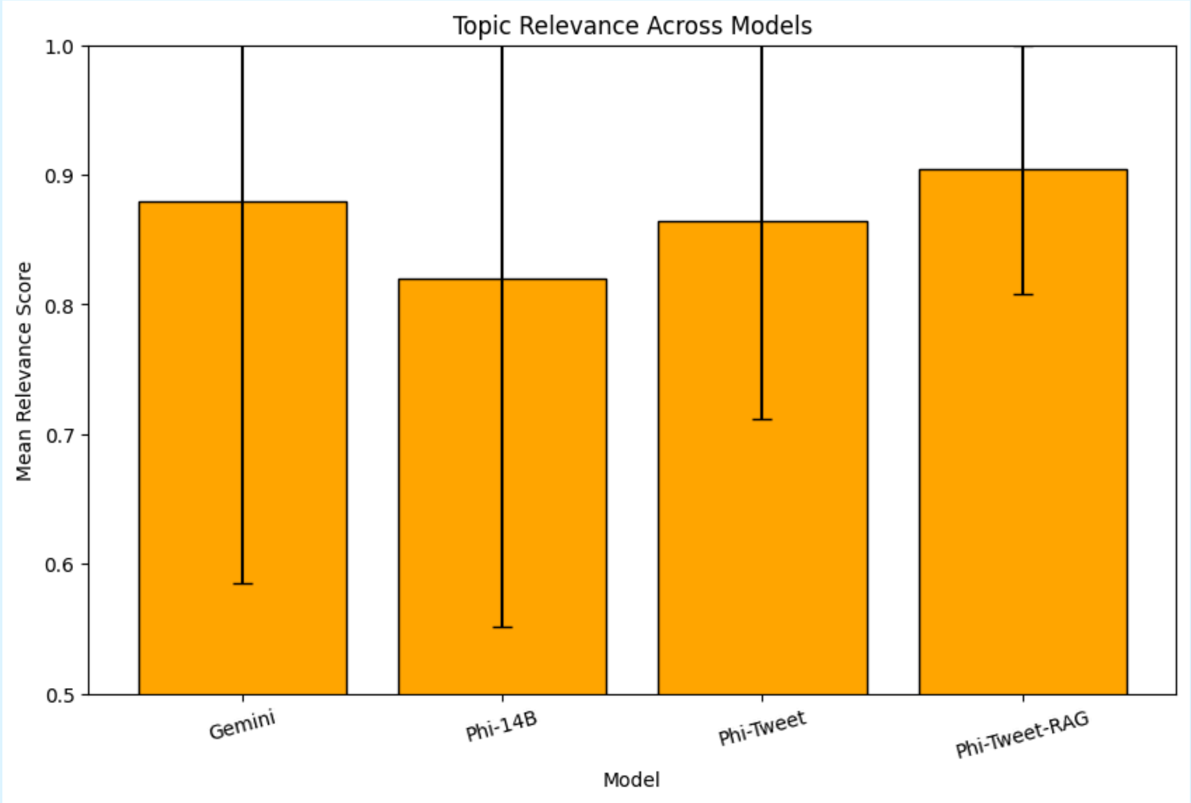
Quantitative Results



Tweets Semantic Variation



Successful Fooling Rates



Tweets Topic Relevance

1k 1k 1k

1k Retweet 1k Likes