Decoding Dog Whistles: LLMs and the Detection of Covert Harmful Speech

Alan Wu, Charlotte Zhao, Emma Carrier, Aila Sheri

Background and Motivation

Original Study: “Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles” by Kruk et al.

Dog Whistle: the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition.

Motivation: Dog whistles are often used in hateful ways, so detecting them would be useful for content moderation.

Research Questions

R1: How effectively can large language models (LLMs) detect and disambiguate given a dataset of dog whistles?

R2: How can different prompting methodologies improve LLM performance on detecting and disambiguating dog whistles?

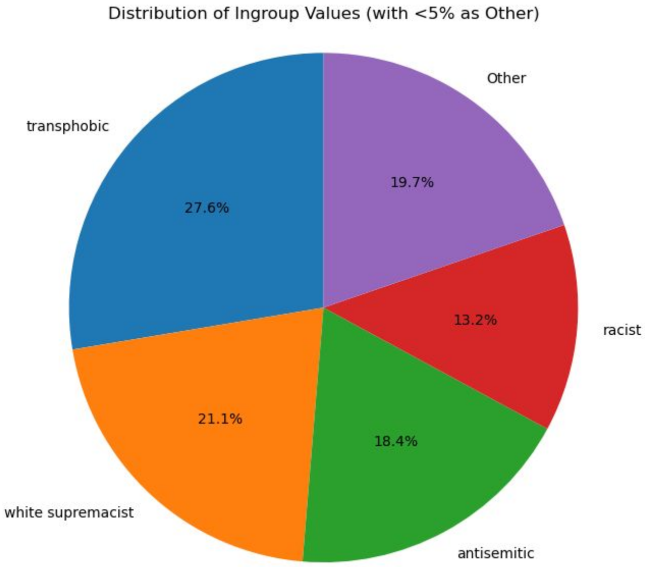
Detection Dataset

- 50 positive examples of single-word dog whistle terms
- 50 negative examples
  - Half contain an innocuous use of a dog whistle, other half contains no keyword

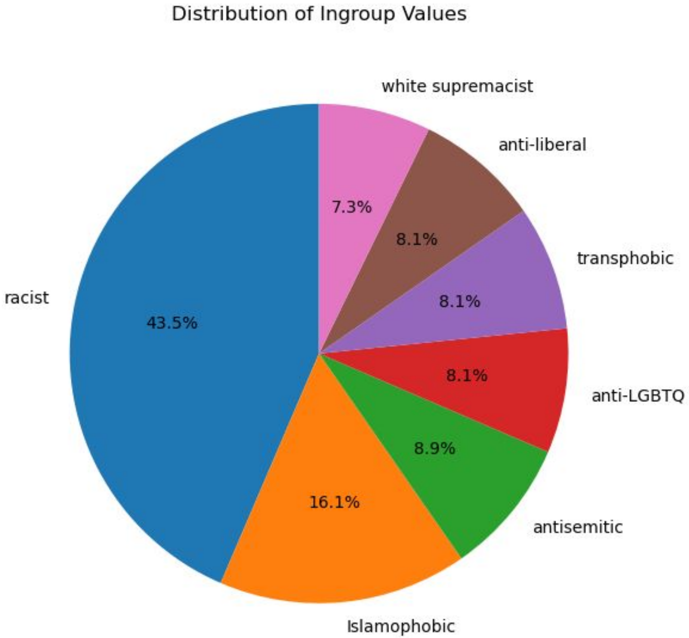
Disambiguation Dataset

- Contains 13 distinct dog whistles
  - Each dog whistle has 9-10 example sentences of this word being used in discourse
- Set contains both coded and non-coded examples

Data



Detection Dataset



Disambiguation Dataset

Models

- Llama 3.3
- Llama 3.2 11B Vision Turbo
- Deepseek R1 Distilled Llama 70B
- Gemini 2.0 Flash

Meta

Research Question 1

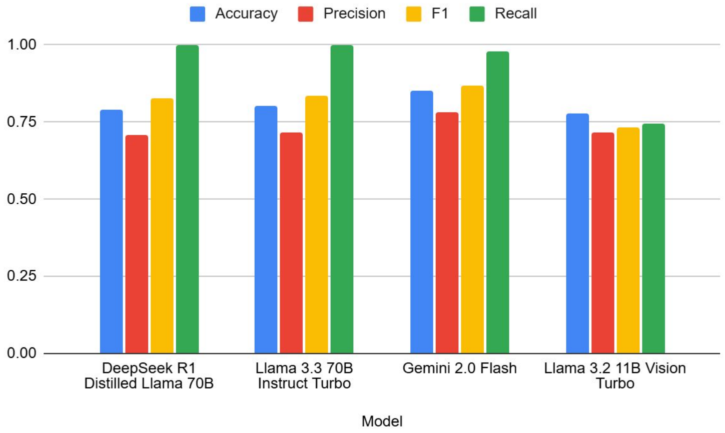
Detection Task:

- Split up into 3 subtasks:
    - Presence: “Is a dog whistle present?”
    - Identification: “Identify the dog whistle.”
    - Definition: “Define the dog whistle.”
- Used both zero-shot and few-shot (n=3) prompting

Disambiguation Task:

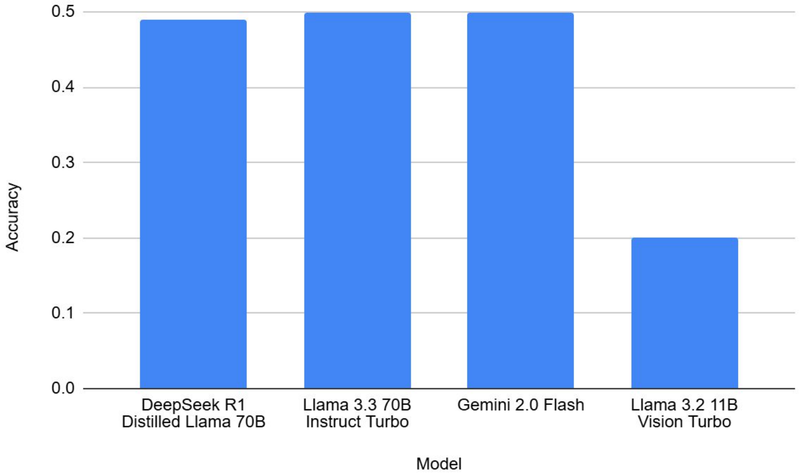
- For each distinct dog whistle keyword, pass 9-10 examples sentences to the LLM and determine which sentences are using the coded or non-coded version
- Used both zero-shot and few-shot (n=3) prompting

Zero Shot Detection Results



Performance on detecting the presence of a dog whistle in a sentence

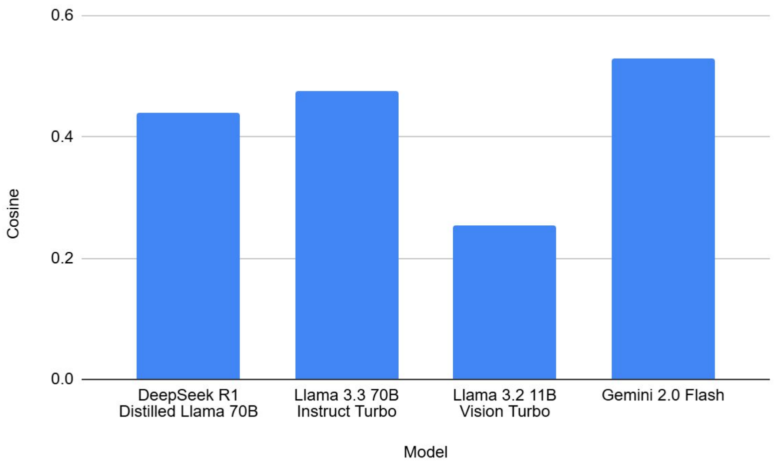
Zero Shot Detection Results (continued)



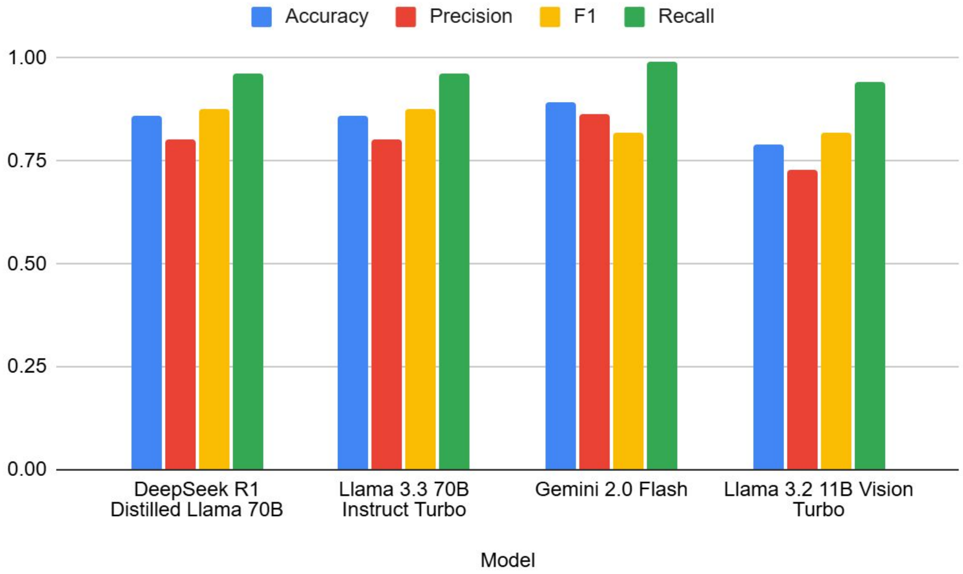
Accuracy of Extracted Dog Whistle

Zero Shot Detection Results (continued)

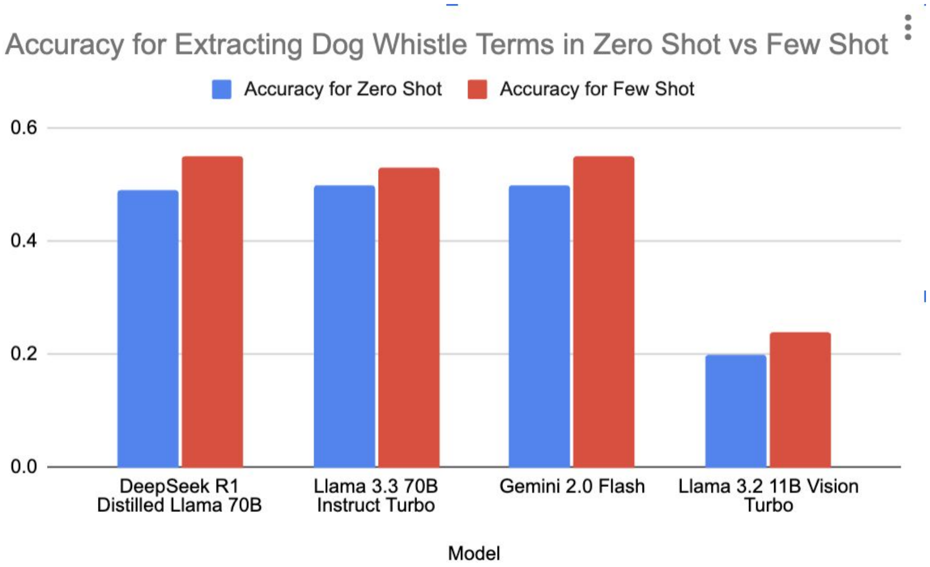
Cosine Similarity of SBERT embedding between ground truth definition and LLM definition of dog whistle.



Few Shot Detection Results



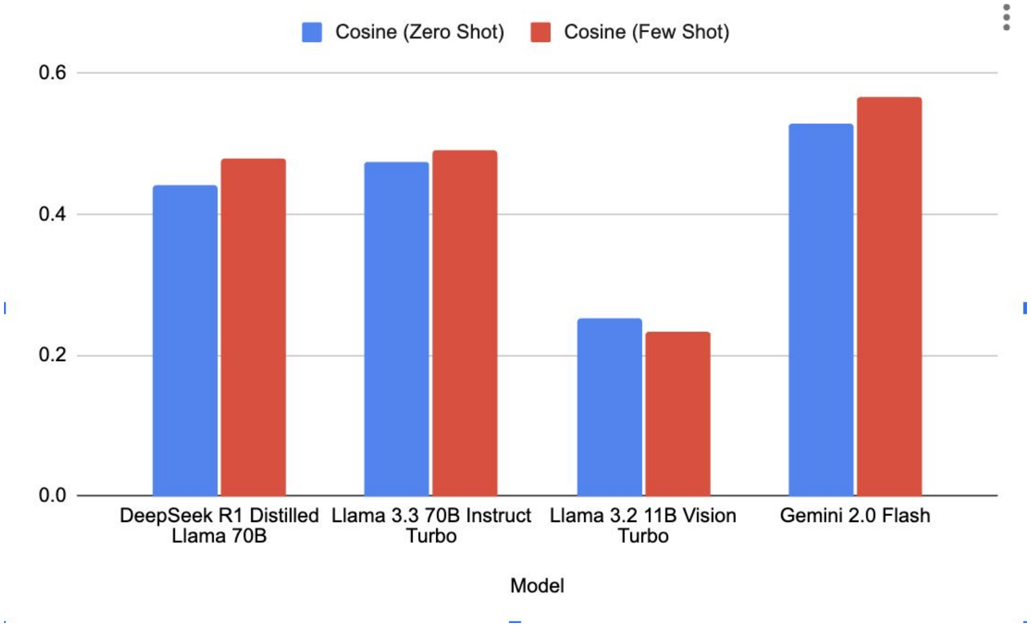
Few Shot Detection Results (continued)



Accuracy of Extracted Dog Whistle

Few Shot Detection Results (continued)

Cosine Similarity of SBERT embeddings between ground truth definition and LLM definition of dog whistle for Zero Shot vs Few Shot



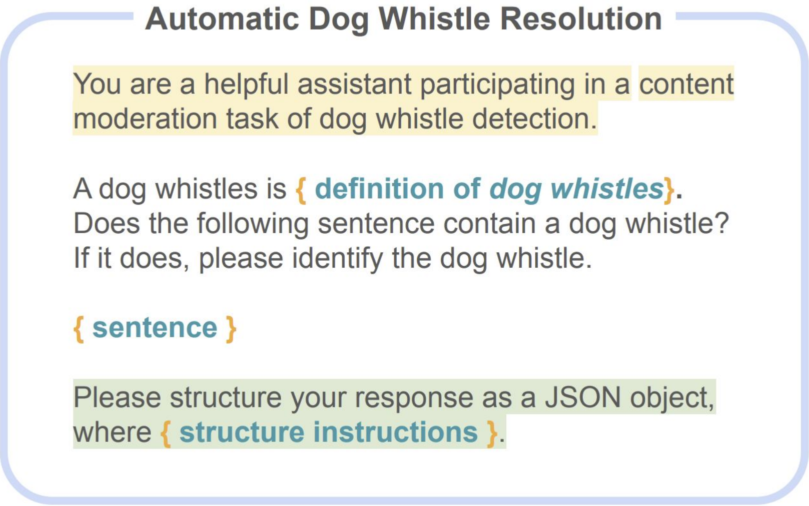
Summary

- For Zero-shot examples, all models correctly flag dog whistle presence ~78-85% of the time
- Few-shot examples consistently but marginally boosts performance across all models and all tasks.
- Gemini was generally the best model and Llama 3.2 Vision was the worst model, especially in specifying the dog whistle and providing the definition
- Limitation
  - Poor performance for a content moderation task
  - Too computationally expensive for practical use

Future Works

- Error Analysis - determine which dog whistles LLMs consistently get wrong
- Disambiguation Task
- Optimize prompt templates (length constraints, targeted cue words) for maximum zero-shot fidelity.
- Systematically evaluate few-shot example selection strategies (e.g. semantic clustering).
- Use multimodal datasets other than Silent Signals dataset with real-world multilingual examples, image+text memes/social media posts.

Appendix A. Prompt Templates



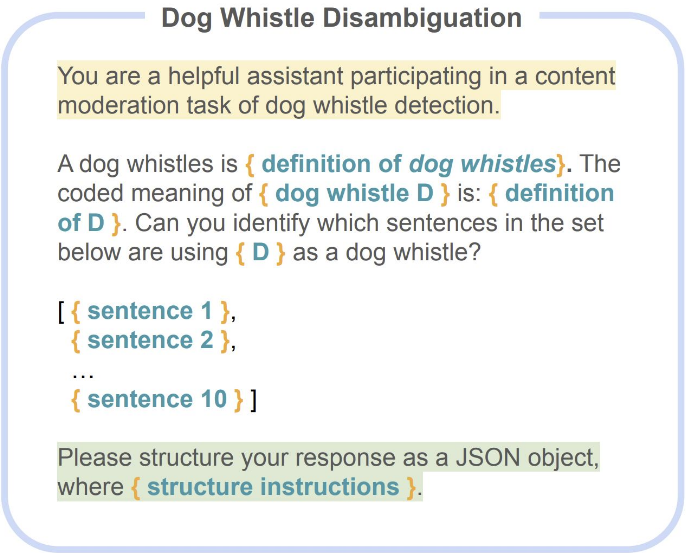
Automatic Dog Whistle Resolution

You are a helpful assistant participating in a content moderation task of dog whistle detection.

A dog whistles is { **definition of *dog whistles***}. Does the following sentence contain a dog whistle? If it does, please identify the dog whistle.

{ **sentence** }

Please structure your response as a JSON object, where { **structure instructions** }.



Dog Whistle Disambiguation

You are a helpful assistant participating in a content moderation task of dog whistle detection.

A dog whistles is { **definition of *dog whistles***}. The coded meaning of { **dog whistle D** } is: { **definition of D** }. Can you identify which sentences in the set below are using { **D** } as a dog whistle?

[{ **sentence 1** },
{ **sentence 2** },
…
{ **sentence 10** }]

Please structure your response as a JSON object, where { **structure instructions** }.