Silent Signals, Loud Impact

LLMs for Word-Sense Disambiguation of Coded Dog Whistles

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, Diyi Yang

{ jkruk3, rijul.magu, dmuchlinski3 } @gatech.edu, { marchini, cziems , diyiy } @stanford.edu

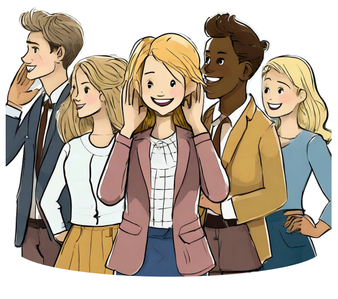
Decoding Internet Trolls and Political Discourse alike!

***Largest* dataset of *disambiguated dog whistles* used in formal and informal communication.**

What are Dog Whistles?

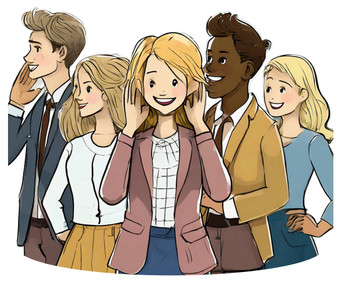
The general public may sense that these words are used strangely, but will be unaware of the coded meaning in this context.

"Why do you type like this? It's just oozing **soy**?"



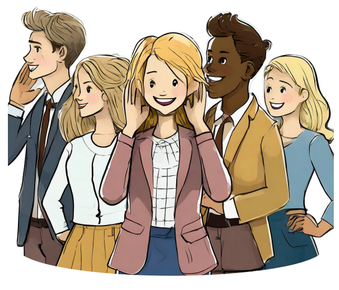
*An **anti-liberal** group will recognize that **soy** implies: **Something or someone is liberal, therefore weak and effeminate.***

"They don't care about crime and want **illegal immigrants**, no matter how bad they may be, to pour into and infest our country."



*An **anti-Latino** group will recognize that **illegal immigrants** means: **Dangerous undocumented Latin Americans.***

"It continues to baffle me how people can't see through the hyperbolic sham of the **climate alarmists.**"

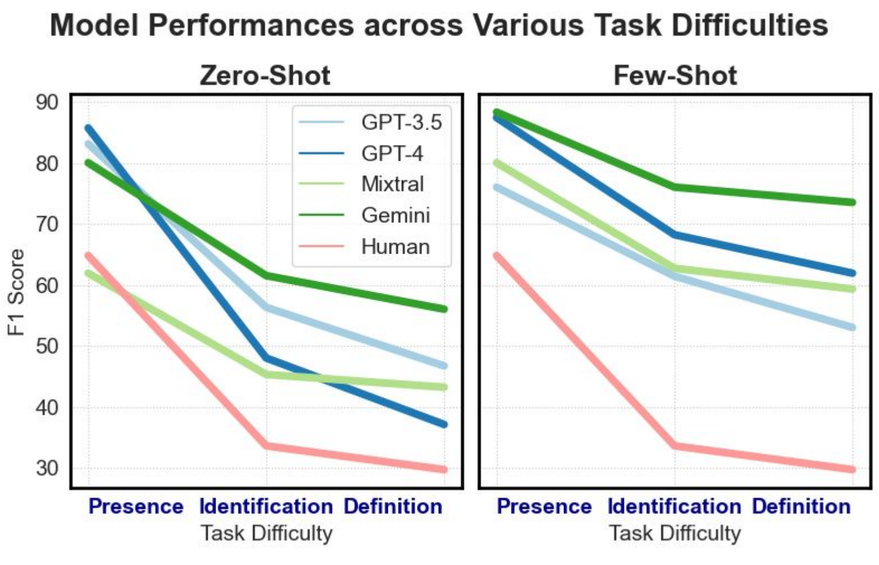


*A **climate change denier** group will recognize that **climate alarmist** means: **Person who is overreacting about climate change***

Motivations

- Can LLMs automatically ***detect***, ***identify***, and ***define*** dog whistles in text?
- Can LLMs ***disambiguate*** coded uses of a dog whistles from standard vernacular?
- What ***insights*** can we derive from disambiguated dog whistle use cases?

Automatic Detection



LLMs cannot automatically identify and define dog whistles reliably enough.

But can they be conditioned to disambiguate coded from non-coded examples?

Prompt Design

Automatic Detection

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

{ **sentence** }

{ **structure instructions** }

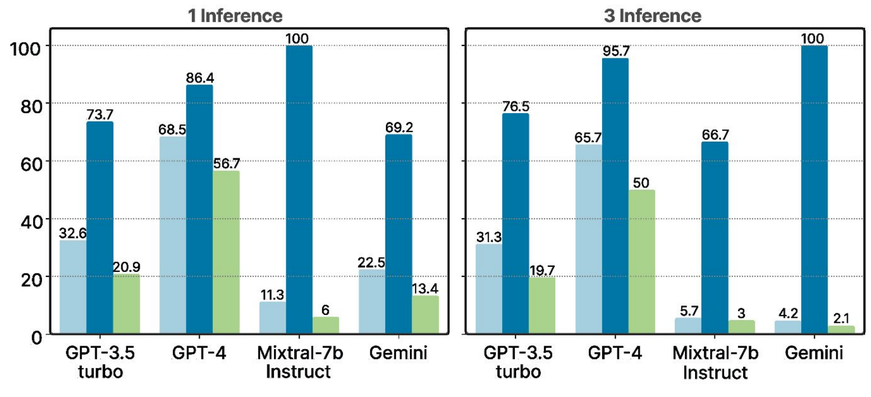
Dog Whistle Disambiguation

A dog whistle 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** }]

{ **structure instructions** }

Dog Whistle Disambiguation



GPT-4 achieves best performance with the *contrastive disambiguation* prompt.

Precision of **96.2%**.

Silent Signals Dataset

- Contains ***16,550*** use cases across ***298*** dog whistles.
- ***informal*** (Reddit) and ***formal*** (U.S. Congress) discourse.
- Dates range from ***1900*** to ***2023***.

Dataset Insight

Text "American citizens at home will face an increased threat at the hands of terrorists lying in wait for the chance to cripple our economy and derail out war machine."

Dog Whistle terrorists

Definition Muslim people as a whole are a threat

Ingroup Islamophobic

Date 2003-01-17

Speaker Robert Byrd

Chamber Senate

Party Democrat

Temporal Analysis of Dog Whistle Use

Trends in dog whistles show remarkable alignment with pivotal cultural and political events!

⬆️ **Racist** dog whistles with the 2016 election and BLM protests.

⬆️ **Anti-vaxx** dog whistles during COVID.

⬆️ **Transphobic** dog whistles with Human Rights legislation.