Who Let the Dogwhistles Out? A Comparative Study of Dogwhistle Detection and Covert Meaning Surfacing through ChatGPT and Claude

Thomas De Meola

Abstract

This paper explores the efficacy of publicly available Large Language Models (LLMs), specifically ChatGPT 3.5 and Claude 3 Opus, in detecting and interpreting "dogwhistles" or coded language frequently employed in political discourse to convey covert messages. Using a glossary of dogwhistles compiled from previous research, this study creates prompts using quotes containing dogwhistle terms and tests specifically whether each model can identify the lexical terms that make up each dogwhistle, surface the covert meaning, and suggest replacement terms. Through comparative analysis and a proposed novel evaluation rubric, the findings reveal varied successes and limitations in the models' abilities. This study contributes to the ongoing discourse on the application of machine learning to uncover coded language in political rhetoric and understanding the ways in which LLM applications can help the general public interact with language they may not be able to fully understand or interact with.

1 Introduction

The use of coded language is prevalent in political rhetoric in the United States. The covert meaning associated with certain lexical items that make up this coded language are embedded within seemingly innocuous phrases. While these terms may seem unremarkable at first glance, this rhetorical device serves as a potent tool for conveying hidden messages and appealing to specific subsets of the audience. These hidden messages, commonly referred to as "dogwhistles", operate beneath the surface of explicit language, infiltrating various aspects of public discourse, media narratives, and eventually, can be by the general public for use in everyday conversations.

The term "dogwhistle" appeared in the late 1980s as a result of a curiosity noticed in opinion polling, where responses to poll questions changed drastically due to small changes in the language of poll questions (Saul, 2018). Now, their use can be noted from the public speeches of presidential candidates, and posts on online forums by the general public. Dogwhistles today are often utilized to convey racist, bigoted, homophobic, transphobic, or xenophobic ideas, among others.

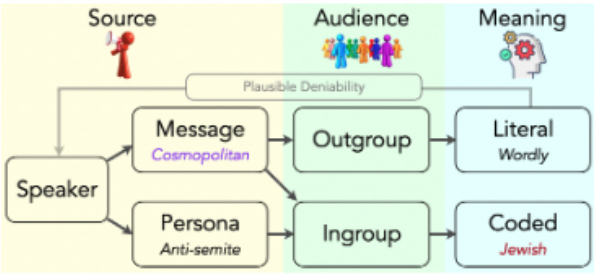

Figure 1: Flowchart depicting the dogwhistle term "cosmopolitan".

Nacos, Shapiro and Bloch-Elkon (2020) write an article that identifies specific rhetorical strategies used by former U.S. president, Donald Trump, which emphasize the divisiveness and dehumanizing nature of his speech. Trump utilizes scapegoating, which targets "outsider" groups like immigrants, Democrats, and the media, onto whom he throws blame for numerous societal issues. Trump uses dehumanizing language, comparing immigrants to "animals" and political opponents to "scum". Further, he overtly and implicitly promotes or suggests the use of outward aggression or violence towards these targeted groups.

The language of these rhetorical tools is very particular when reading speech transcripts or social media posts. It is uncommon to hear phrases such as, "extremism and destruction and violence of the radical left", "the silent majority is stronger than ever before", or "they want to demolish our heritage", all of which originate from Trump speeches. This usage of dogwhistles is outlined in Mendelsohn et al (2023). The authors

include a chart, reproduced here as Figure 1, which depicts the underlying framework for how dogwhistles are understood. This "insider/outsider" dichotomy mentioned in Nacos, Shapiro and Bloch-Elkon (2020) is prevalent here as well. Mendelsohn et al (2023) depict the source speaker uttering the dogwhistle message "Cosmopolitan" along with their persona. The in-group picks up on both the message and persona, while the out-group only receives the message. This ultimately leads to the in-group understanding the coded meaning of the language uttered, and the out-group understanding the speech literally. Plausible deniability is left for the speaker to deny the coded meaning of their speech and save face in the eyes of the listener if they are ever questioned.

While this taxonomy of dogwhistling is novel and very useful, there are other factors to consider when looking at coded language that are brought up in Breitholtz and Cooper (2021). The authors make note that individuals can produce varied interpretations of any language, dogwhistle or not, based on pre-formed opinions, prejudices, biases, or other information. They bring up the concept of "topoi", where a single statement can be associated with different arguments or conclusions. This poses a challenge for automatic dogwhistle detection. While dogwhistles are set phrases, the context in which they are set and the way they are interpreted by the audience are variable and require additional background information, which can be difficult to capture in a dataset from which a language model could learn.

Henderson and McCready (2021) argue against past opinions stating that dogwhistles must have both a straightforward and an implicit meaning. They use the term "conventionalized" to refer to a dogwhistle's widely-accepted social meaning and argue that dogwhistles have a certain level of conventionalization, while allowing enough room for plausible deniability for the speaker. The authors then go on to say that, while the social meaning the in-group draws from a dogwhistle is important, the projection of ideals or values from a speaker's persona can also influence the intended interpretation of the dogwhistle. That is to say, the semantic, syntactic and lexical content of the dogwhistle play a part, but the audience's knowledge of the speaker's persona can affect the interpretations they draw from the speaker's words as well. Henderson and McCready's (2021) points, along with the points brought up in Breitholtz and Cooper (2021), compound the problem of automatic dogwhistle detection.

Moving back to Mendelsohn et al (2023), one of the experiments they conducted was prompting GPT-4 to identify dogwhistles present in real-world examples from a glossary of dogwhistles they curated. There was mixed success in this experiment, where GPT-4 would correctly identify some dogwhistles (such as "inner city") and having trouble with others (such as "#IStandWithJKRowling"). The results of this experiment highlights the difficulty of current Large Language Models (LLMs) to identify dogwhistles most likely due to the complexity surrounding them, as other authors mentioned here have stated.

Hertzberg et al (2022) employ transformer-based models to analyze the potential of automatic dogwhistle detection. They collected

Dogwhistle Term	In-Group Identity	Covert Meaning
thug	racist	Black men are criminals or violent
migrant crime	racist, anti-Latino	Immigrants are all criminals
gender ideology	transphobic	Trans people are threatening our way of life
groomers	anti-LGBTQ	LGBTQ people are predators
radical Islam	islamophobic	Muslim people as a whole are a threat
personal responsibility	racist	Opposition to public spending perceived to benefit racial minorities over white people

Table 1: List of dogwhistles investigated. Sourced from Glossary of Dogwhistles (Mendelsohn et al., 2023).

suspected Swedish dogwhistles and conducted a survey where these terms were replaced with what the participants thought the intended meanings were. This data was manually annotated and the authors utilized transformer-based models along with methods like K-means clustering to group sentences based on their representations in semantic space, and SVM to assign labels to sentences. Ultimately, they found that their models performed well and matched the human judgements overall. Though, some dogwhistles were more difficult to capture in semantic space than others due to lexical choices, selection of specific training data, or some dogwhistles being more spread out in a wider semantic space which created difficulty for cluster separation.

Taking this all into consideration, understanding the dynamics of dogwhistles is not only important for deciphering the true intentions of a speaker, but also for arming the general public with useful tools and knowledge to navigate the complexities of political communication. The ability to recognize and critically analyze dogwhistles is crucial in guarding against manipulation and fostering informed decision making. Not everyone has the media literacy and critical thinking skills to assess this coded language – it would not be called "coded language" if the meaning was blatantly obvious. This highlights the importance of developing tools for individuals to utilize which detect such dogwhistling and empower them to engage with political discourse more discerningly.

In this paper, I seek to expand on the work of Mendelsohn et al. (2023) and Hertzberg et al. (2022) by comparing the efficacy of ChatGPT 3.5

A dogwhistle is a term that sends one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup.

Your task is to find all words in the following text that constitute a dogwhistle. You will also define the covert message the dogwhistle is sending to the ingroup. In addition, please provide as many alternatives terms as possible that would expose the implicit ingroup meaning, so that if someone from the outgroup were to replace the dogwhistle term with your term in the original text, the implicit meaning would be explicit to them.

This is the text you will analyze:

"{quote}"

Figure 2: The generic prompt template used for building LLM prompts. The first sentence in the prompt is the definition of dogwhistle (Henderson & McCready, 2018).

Scoring Rubric				
Lexical Items (w=1.5)	Identified key dogwhistle item, and others likely to be dogwhistles +2	Identified key dogwhistle item, but identified non-dogwhistle items +1	Did not identify key dogwhistle item, mostly identified non-dogwhistle items: 0	All lexical items identified incorrect: -1
Covert Meaning (w=2.0)	Covert meaning of key dogwhistle uncovered, majority of uncovered meanings are relevant: +2	Covert meaning of key dogwhistle somewhat uncovered, somewhat clear to out-group: +1	Covert meaning of key dogwhistle not uncovered, covert meaning still obscure to out-group: 0	Covert meaning of key dogwhistle not uncovered, wrong Covert Meaning uncovered: -1
Replacement Terms (w=1.0)	Replacements of key dogwhistle convey make the covert message explicit, other replacement terms are relevant: +2	Replacements of key dogwhistle convey part of the covert meaning, most other replacement terms are relevant: +1	Replacements of key dogwhistle do not convey the covert meaning, many others are irrelevant: 0	Replacements for key dogwhistle communicate incorrect covert meaning, many irrelevant replacement terms: -1

Table 2: Evaluation criteria for LLM responses.

and Claude-3 in a few areas. First, the ability of these LLMs to identify specific lexical items in quotes from Donald Trump in political speeches, and also from online political content, will be investigated. Second, I will investigate the ability of each LLM to surface the covert meaning implied by stating the dogwhistle terms in context. Last, each LLM will be evaluated on its ability to produce replacement terms for each identified dogwhistle in a way that would make the communicate the covert meaning explicitly.

2 Methodology

2.1 Glossary of Dogwhistles

The Glossary of Dogwhistles was compiled by Mendelsohn et al. (2023) to define and organize dogwhistles identified in online and offline speech. The glossary contains 340 English language dogwhistle terms with over 1,000 surface forms, which are variants of the original indexed terms. Each dogwhistle entry in the glossary comprises several subcategorization fields. The fields this project focuses on are the dogwhistle term itself and the covert meaning. Table 1 outlines each of the six dogwhistles selected for this project.

2.2 Source Material

Collecting the quoted source material for this project involved was quite easy. Knowing that dogwhistle terms are quite common in right-wing rhetoric, the author looked to Donald Trump's campaign rallies. When reviewing the video, the author came across "thug" in a campaign rally in Wisconsin, and "migrant crime" in a campaign rally in South Carolina.

A popular alt-right podcast-type webcast called InfoWars was chosen to source quotes containing dogwhistles, specifically from The Alex Jones Show. On this show, the main host Alex Jones is a well-known alt-right conspiracy theorist and political commentator, whose show has been described by Ad Fontes Media's Interactive Media Bias Chart (2024) as a program with extreme right-wing bias containing inaccurate or fabricated information. The dogwhistles "gender ideology", "groomers", and "radical Islam" were located using the Knowledge Fight Interactive Search Tool, a corpus of InfoWars transcripts compiled by an independent entity (Simonsen, 2024).

To further diversify the sources, the author chose a Ben Shapiro quote containing the dogwhistle "personal responsibility" from his appearance on the podcast The Joe Rogan Experience. Ben Shapiro is a right-wing political commentator, whose bias score on the Interactive Media Chart mentioned above is lower than that of The Alex Jones Show.

Since the context in which the dogwhistle is situated is important, the length of each quote varied between 70 and 180 word tokens in order to give the LLM enough surrounding context. This would, in turn, allow the LLM to come up with a comprehensive response to the prompt given.

2.3 Prompt Creation

A prompt template was created, which contained three parts: the dogwhistle definition from Henderson and McCready (2018), the prompt itself, and the quote. There were a total of 12 prompts created for the six chosen quotes. Six of the prompts included the definition of a dogwhistle, and six of the prompts excluded the definition. This was done in order to investigate if including the definition in the prompt helped or hindered the LLM's ability to accurately complete the task.

	Lexical Item	Covert Meaning	Replacement Terms
ChatGPT w/def	1.17	1.00	**-0.25**
ChatGPT w/out def	1.25	**1.42**	-0.83
Claude w/def	1.30	1.30	-0.70
Claude w/out def	**1.38**	1.38	-0.88

Table 3: Average performance of ChatGPT and Claude for each task type

A Python script was written to create the prompts and interface with the Claude model. The Unofficial Claude API (Raneri, 2024) was the basis for communicating with the Claude 3 Sonnet model, where each prompt was passed to the model in a new chat and the response was recorded.

There was an attempt made to use OpenAI's official API to interface with GPT-3.5 directly, however this was cost prohibitive to a graduate student like myself. Therefore, the prompts to ChatGPT were manually entered using the web interface in a new chat for each prompt. Figure 2 shows the generic prompt template. For a more detailed look into the quotes selected, model responses and evaluations, please see the spreadsheets included in the GitHub repository.

3 Evaluation

A point system was developed in order to provide a quantitative assessment of LLM responses. The LLMs were evaluated on each of the three tasks: identifying lexical items that make up a dogwhistle, surfacing the covert meaning associated with the dogwhistle, and providing replacement terms to the dogwhistle to make the covert meaning more explicit to the outgroup. Table 2 outlines the scoring rubric in detail.

The score for each task was weighted based on overall importance: surfacing the covert meaning was considered the most important with a weight of 2.0, identifying the correct lexical items was the next important with a weight of 1.5, and providing replacement terms was the least important with a weight of 1.0. These weights were assigned in this way since, from a lay-user's perspective, the main goal should be to surface the meaning of the coded language, even if the LLM was not able to identify the dogwhistle itself or provide relevant replacement terms.

Coders entered the score for each of the three tasks for each response. A final score was calculated by taking the sum of each score from each category multiplied by its weight. The author and one other coder rated all the responses from each LLM for all three categories. Figure 3 depicts the average score for each dogwhistle from ChatGPT and Figure 4 depicts the same from Claude.

Table 3 shows the average performance of both ChatGPT and Claude in performing each of the three tasks, with and without the prompt including the definition of dogwhistle. Each coder's score for each task was averaged, added together, and then averaged again to calculate the values in Table 3.

4 Results

4.1 ChatGPT

ChatGPT's performance was mixed overall. As Figure 3 demonstrates that including the definition of dogwhistle in four out of the six prompts given improved its performance for the most part, aside from the "personal responsibility" prompt, where its performance was significantly worse than when

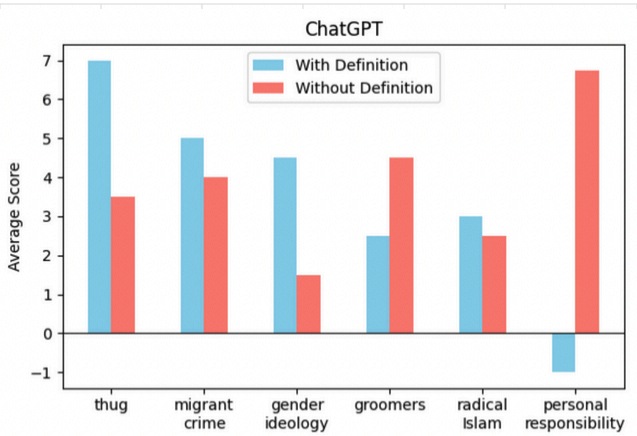

Figure 3: ChatGPT: Average Final Score for each dogwhistle term including and excluding the definition in the prompt.

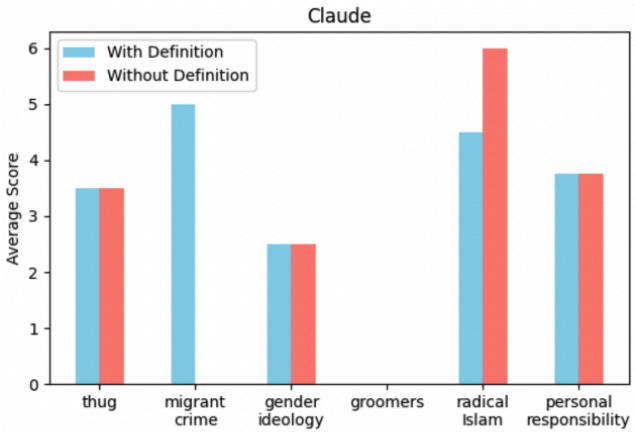

Figure 4: Claude: Average Final Score for each dogwhistle term including and excluding the definition in the prompt.

the definition was not included in the prompt, scoring negatively according to the coders' evaluations. For that prompt specifically, it did not identify the key dogwhistle term, did not surface the correct covert meaning, and did not provide any relevant replacement terms, since it did not identify the correct dogwhistle in the first place.

It is interesting to note that ChatGPT did perform very well on the same prompt without including the definition. It could signal that including the definition in the prompt limits the LLM's freedom to choose lexical items constituting a dogwhistle, however that does not align with the results for the other prompts. This result does, however, highlight the inconsistency and variance of the LLM's ability to perform the task given.

ChatGPT was slightly better than Claude overall in surfacing the covert meaning of dogwhistles across all the prompts without including the definition of dogwhistle, as shown in Table 3. Though, this table seems to suggest that ChatGPT's ability to identify lexical items and surface covert meanings were hindered by including the definition in the prompt, as it scored the lowest in these two tasks across all conditions. The results do seem to suggest that including the definition in the prompt gave ChatGPT a boost in performance when suggesting replacement terms for identified dogwhistles, though neither of the LLMs were found to do particularly well in this task specifically.

4.2 Claude

Claude's performance was also found to be very mixed overall. Figure 4 highlights that, in three out of the six prompts given, including or excluding the definition of dogwhistle had no effect on its ability to complete the tasks given. Only for the "radical Islam" dogwhistle did Claude perform better without the definition, scoring the same in covert meaning surfacing and suggesting replacement terms, but performing slightly better in the lexical item identification category. Though, I believe that having more coders perform the response evaluation, this disparity would likely even out.

Table 3 suggests that Claude provided responses of more consistent quality across all prompts given in terms of lexical item identification and covert meaning surfacing. However, Claude was found to have the worst ability at providing replacement terms for the identified dogwhistles. I believe this is due to the architecture of the Claude model.

Though much of the backend is not public knowledge, Claude's creator Anthropic has released that the LLM is supported by a novel approach called Constitutional AI, consisting of a supervised learning phrase, and a reinforcement learning phrase. The model critiques itself on responses to prompts based on a "constitution" outlining principles response outputs should follow, and examples of the process. These responses are then compared against the constitution and the model is fine-tuned accordingly. This is an approach developed to protect again Claude from producing hateful or harmful without the need of copious human feedback (Anthropic, 2024).

Due to this type of reinforcement learning, Claude is bound by a very strict, pre-defined set of guidelines for responding to prompts, potentially leading to more consistent responses, but also a propensity to not engage with harmful language at all.

Figure 4 also shows that, for the "migrant crime" prompt excluding the definition, and for both "groomers" prompts, the model refused to respond to the prompt at all, citing that its response could promote harmful stereotypes and conspiracy theories, validate hate speech, or promote harmful rhetoric against any groups. Originally, for the replacement task, the author attempted to prompt the models to reproduce the entire quote provided, with replacement terms inserted instead of the dogwhistle terms. Claude's disposition to not complete this part of the task is the reason why the author decided to have each model simply provide replacement terms without replicating the original quote. Claude's heightened sensitivity to this kind of harmful rhetoric impedes its ability even to complete tasks surrounding this type of rhetoric that do not intend to promote or support the rhetoric being analyzed, making it a less desirable option for those attempting to understand coded language in political rhetoric.

5 Discussion

The results obtained from this study regarding identifying dogwhistle terms and surfacing covert

meanings align generally from what was noted in Mendelsohn et al. (2023). This study differs in that each prompt contained three separate tasks, whereas previous research only provided the task of identifying the dogwhistle without explicitly prompting the model to surface covert meanings. Needless to say, collecting more responses from both models would provide an expanded view into the ability of LLMs to identify coded language in rhetoric from this domain.

The coding scheme in this study attempted to add more of a quantitative analysis to the model responses than was posed in Mendelsohn et al. (2023). Their evaluation of LLM responses consisted of a binary "true/false" metric for dogwhistle term identification and covert meaning surfacing. This study attempted to provide a numerical scale on which to rate responses, which, from the evaluation scores for both coders in Table 4, show a decent agreement. This gives rise to the point that more coders are required to establish whether or not this evaluation method is reliable, though this study provides a basis for the fact that it has potential.

The difficulty of designing an evaluation method for each response was compounded by the fact that ChatGPT and Claude did not provide consistently formatted responses. This could be solved by specifically outlining the desired format in the prompt itself. Another way of rectifying this would be to create a custom GPT through OpenAI's ChatGPT interface to specifically outline and customize the model for this specific task. However, it should be pointed out that Claude does not have this capability. Further, collecting more training data to fine-tune a GPT model would potentially bolster the model's ability to accurately complete the tasks, specifically the replacement term task. Future research in this area should include these methods.

Additionally, the refusal of Claude to interact with specific content raises important ethical considerations for language model research. While Claude's refusal to engage with harmful rhetoric aligns with its constitution defined by its creators and the principles of responsible and ethical AI development, it also highlights the challenge of balancing model safety with the need for analysis of potentially harmful and contentious topics. It goes without saying that precautions against generating hateful content are necessary, but it must not be kneecapped in such a way that users who are attempting to understand coded language regarding such important topics are adequately supported in their endeavors. Of course, there are those who may exploit this, but being able to provide analysis of this caliber requires a balance between safeguarding against harmful content and facilitating meaningful research and analysis.

In light of these considerations, future research efforts should prioritize the development of robust evaluation methodologies, the exploration of alternate model architectures and fine-tuning approaches, and the implementation of ethical safeguards. Interdisciplinary collaboration between machine learning scientists and linguists is crucial to address the multifaceted challenges posed by model development in sensitive domains. Seeing that the models are aptly able to identify content that is potentially harmful, investigating the ways in which we can use biased data to educate and inform through these LLMs is of paramount importance.

References

Ad Fontes Media. (2024). *Methodology*. https://adfontesmedia.com/how-ad-fontes-ranks-news-sources/

Anthropic. (2024). *Claude's constitution*. https://www.anthropic.com/news/claudes-constitution

Breitholtz, E., & Cooper, R. (2021). *Dogwhistles as inferences in interaction*. ACL Anthology. https://aclanthology.org/2021.reinact-1.6.pdf

Henderson, R., & McCready, E. (2021, December 31). *Dogwhistles: Persona and ideology*. Semantics and Linguistic Theory. https://doi.org/10.3765/ywwkd646

Hertzberg, N., Sayeed, A., Breitholtz, E., Cooper, R., Lindgren, E., Rettenegger, G., & Rönnerstrand, B. (2022). *Distributional properties of political dogwhistle representations in Swedish BERT*. ACL Anthology. https://aclanthology.org/2022.woah-1.16.pdf

Mendelsohn, J., Le Bras, R., Choi, Y., & Sap, M. (2023a). *From dogwhistles to bullhorns: Unveiling coded rhetoric with Language Models*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2023.acl-long.845

Nacos, B. L., Shapiro, R. Y., & Bloch-Elkon, Y. (2020). *Donald Trump: Aggressive Rhetoric and Political Violence*. Perspectives on Terrorism, 14(5), 2–25. https://www.jstor.org/stable/26940036

Saul, J. (2018). *Dogwhistles, political manipulation, and philosophy of language*. New work on speech acts, 360, 84

Simonsen, E. (2024). *KFIST - Knowledge Fight Interactive Search Tool*. https://fight.fudgie.org