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

Alan Wu, Aila Sheri, Charlotte Zhao, and Emma Carrier

Abstract

Dog whistles present significant challenges for content moderation and automatic detection systems. This study evaluates the ability of four large language models (DeepSeek R1 Distilled Llama 70B, Llama 3.3 70B Instruct Turbo, Llama 3.2 11B Vision Turbo, and Gemini 2.0 Flash) to detect, identify, define, and disambiguate political dog whistles using the Silent Signals Corpus. We utilized zero-shot and few-shot prompting across two distinct tasks: the detection of dog whistle presence and meaning, and disambiguation between coded and non-coded usages. Results show that Gemini 2.0 Flash consistently outperformed the other models, achieving the highest metrics overall. However, all models showed many vulnerabilities, including high false positive rates in non-coded contexts and difficulties with less common expressions.

1 Introduction

Dog whistles—phrases that carry a neutral surface meaning to the general public but convey charged, implicit messages to specific audiences—have surged in online and political discourse, evading both human moderators and conventional NLP systems that rely on explicit keywords or syntax. Although recent large language models (LLM) such as DeepSeek R1 and Gemini excel at broad language tasks, they can struggle to detect these covert signals, in part because existing datasets lack the contextual richness and annotation precision needed to teach models how to spot dual-meaning expressions (ElSherief et al., 2021; Vidgen et al., 2021). Fortuna and Nunes (2018) surveyed automated hate-speech systems, calling for standardized metrics and richer data to capture covert linguistic tactics, while Salminen et al. (2020) demonstrated how cultural and perceptual variation further complicates detection. In this study, we leverage the Silent Signals Corpus (Kruk et al., 2024) to rigorously benchmark four different LLMs on dog-whistle detection, identification, meaning definition, and disambiguation, while refining prompting strategies to improve reliability across diverse discourse settings.

2 Datasets

To explore how LLMs perform with both dog whistle detection and disambiguation, we utilize two distinct datasets to complete both tasks. The detection dataset contains 50 positive examples of single-word dog whistle terms from the (Mendelsohn et al., 2023) glossary, and 50 negative examples, half of which contain an innocuous use of the keyword and the other half contain no keyword. The disambiguation dataset contains 124 examples which were manually annotated for evaluation. The dataset includes 13 distinct dog whistles, with 9-10 example sentences of that word being used in various instances. These sets contain both coded and non-coded examples of the word.

3 Methodology

3.1 Models

Our experiments were run with four open-source LLMs: DeepSeek R1 Distilled Llama 70B, Llama 3.3 70B Instruct Turbo, Llama 3.2 11B Vision Turbo, and Gemini 2.0 Flash. These models were selected based on availability and cost-effectiveness, as these models are either free to use or have a free usage tier.

3.2 Dog Whistle Detection

To determine if the chosen LLMs are able to reliably detect and explain political dog whistles in text, we evaluated the performance of each model on the three distinct detection tasks: presence ("is a dog whistle present?"), identification ("identify the dog whistle."), and definition ("define the dog

whistle"). Each model was provided with the definition of a political dog whistle and a candidate sentence. The accuracy of each model was then calculated based on the responses.

3.3 Dog Whistle Disambiguation

We additionally wanted to see how LLMs can perform with word-sense disambiguation tasks. To evaluate the accuracy of the chosen LLMs, we wanted to evaluate the capacity of the LLMs to distinguish contexts in which a dog whistle keyword is being used in a coded or non-coded manner. Each model was provided with the definition of a political dog whistle, the dog whistle keyword, and the word's coded meaning. The model performs disambiguation classification for each example provided that contains the example keyword, and returns whether or not the word is being used in a coded or non-coded context.

3.4 Prompt Design

For both tasks, we followed a strict prompt structure to generate our results. Having limited variation in what exactly we were providing to the models allowed us to accurately test the performance of the models with little confounding factors. Examples of the prompt structure we followed for both the detection and disambiguation tasks can be seen in Appendix A. For Llama 3.2 11B Vision Turbo, the prompt did have to be adjusted slightly in order to prevent the model from rejecting the "content moderation task" outright.

4 Results

4.1 Detection Performance

After running the models on each example in the detection dataset with both zero-shot and few-shot prompting, we were able to extract performance metrics based on the presence sub-task, which can be visualized in Figure 4. Based on these results, Gemini 2.0 Flash had the highest accuracy, precision, F1, and recall scores for both zero and few shot prompting. These metrics all increased when utilizing a few-shot prompting approach, where examples sentences were provided to the model as well. Additionally, from the 50 coded examples, we tested each models performance in correctly identifying the dog whistle term in the sentence. These results can be visualized in Figure 1. Every model had an identification accuracy around 50%, but Llama 3.2 fell significantly behind the performance of the other models with an accuracy of 20%.

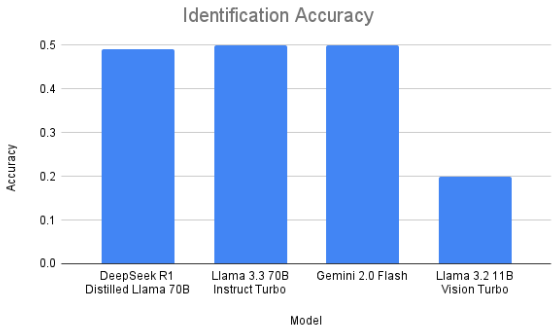

Figure 1: Accuracy of each model for the Identification sub-task

The original paper did not specify how they determined whether the ground-truth definition and the model definition were the same, so we used a quantitative approach. To determine the performance of the models on the definition sub-task, we calculated the cosine similarity between the Sentence BERT (Reimers and Gurevych, 2019) embeddings of the ground-truth definition of each dog whistle and the model generated definition. For this task, Gemini 2.0 Flash also performs the best, and Llama 3.2 performs the worst.

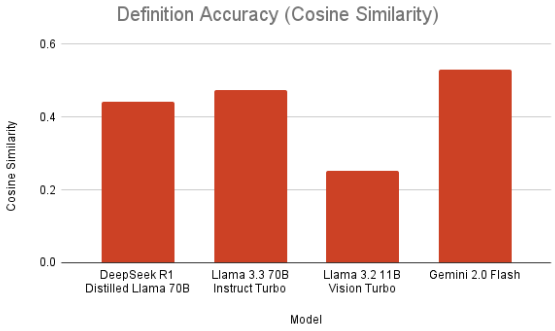

Figure 2: Accuracy of each model for the Definition sub-task

4.2 Disambiguation Performance

In addition to detection, we evaluated LLM's performance in performing disambiguation on words from political contexts with both zero-shot and few-shot prompting. In particular, given a group of sentences that contain the same dog whistle key words, the LLM models are asked to classify whether the key words are used in a harmful or innocuous context. We tested the models in the synthetic disam-

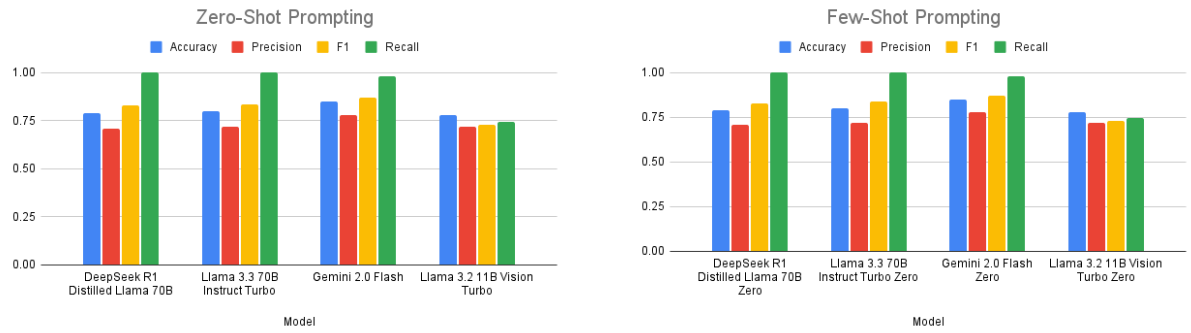

Figure 3: Results for detecting the presence of a dog whistle in a text passage (zero and few shot prompting)

biguation dataset, which comprises 13 groups of distinct dog whistles. For each instance, we extracted both the model classification output and its corresponding explanation. The evaluation metrics derived from these results are summarized in Figure 4.

From the zero-shot results, Gemini 2.0 achieves the best performance, with the highest accuracy (83%) and F1 score (81%). Llama Vision Free performs the worst, having an accuracy of (56%). DeepSeek R1 and Llama 3.3 exhibit similar performance, with DeepSeek achieving slightly higher precision and Llama 3.3 slightly outperforming in accuracy. Most metrics increase notably in few-shot prompting, Llama 3.3 has the greatest improvement, with its accuracy improving to (79.8%).Gemini 2.0 remains the best performing model.

5 Discussion

5.1 Error Analysis: Dog Whistle Detection

There were no instances in which all of the models got the detection task incorrect. However, there were 20 instances in which 3 out of 4 models got the task wrong using zero-shot prompting. With few-shot prompting, there were 12 instances in which 3 out of 4 models go the task wrong. All of these were non-coded sentences. Here are some quotes from the examples: "...often why people call sexual deviants 'fatherless'...", "...Populism gets in the way of globalism.", "A lot of men ... aren't very hygienic...". A dog whistle often appear in these examples, but they aren't used in a coded fashion. In these cases, the potential dog whistle are "fatherless", "globalism", and "hygienic" respectively. These are all known dog whistles which suggests that the LLMs are failing to value the context surrounding these terms strongly enough. There are many sources on the internet providing common dog whistles that LLMs may train on. Thus, instead of deducing whether words are being used in a coded manner, they are simply using the training data to make a prediction. Additionally, by prompting the model to identify whether or not there is a dog whistle, we may be biasing the model into answering that there is a dog whistle. Further exploration is needed to determine how the wording of the prompt can affect the false positive rate.

5.2 Error Analysis: Dog Whistle Identification

There was one example in which 0 models identified the dog whistle correctly and 25 examples in which exactly 1 model identified the dog whistle correctly using zero-shot prompting. With few-shot prompting, there weren't any example of zero models identifying the dog whistle correctly and 13 examples in which 1 model identified the dog whistle correctly. The one example with zero-shot prompting that no model got correct was: "What can I say, old german guys just wanna rule the world with their Windmill of friendship". "Windmill of friendship" is a white-supremacist/neo-Nazi dog whistle which we did not initially identify until we were given the correct term. This suggests that more hidden dog whistles may pose a greater challenge than more commonly known dog whistles. Overall, models performed relatively poorly in this task. Models may have difficulty due to the fact that there is often inflammatory language around the dog whistle. Instead of strictly abiding by the definition of a dog whistle being coded, models may find similarly problematic language that may be more obvious. For example, using few shot prompting, only 1 model correctly identified

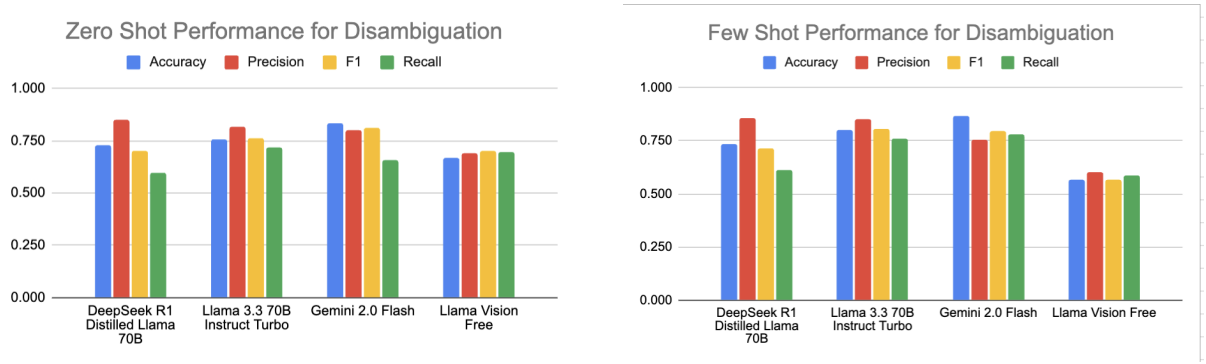

Figure 4: Disambiguation results for coded and non-coded usage (zero and few shot prompting)

the dog whistle "nibba" in the sentence: "This isnt another "dindu nuffin bang bang" situation, ... tell me the last time you heard a story where a hood nibba attempted to blow something up lmao". The term "dindu nuffin" can be viewed as a slur and was identified by 3 out of the 4 models as the dog whistle. Llama 3.3 included both "hood nibba" and "dindu nuffin" as the dog whistle.

5.3 Error Analysis: Dog Whistle Definition

The lowest average cosine similarities were calculated. There doesn't appear to be specific trend. However, LLMs seem to perform decently well at defining the dog whistle after determining the correct dog whistle with Gemini averaging a cosine similarity of 0.567, indicating moderate similarity. It is also important to note that the "ground-truth" definitions of the dog whistles are subjective and may not be the best representation of a dog whistle.

5.4 Error Analysis: Disambiguation

The performance of the LLMs in the disambiguation task was similar to the detection task despite giving the dog whistle and definition in the prompt. This further supports that the real challenge in the task is not in identifying problematic terms, but whether in the given context, the dog whistle was used in a coded manner.

5.5 Comparison to Study by Kruk et al.

As previously demonstrated by (Kruk et al., 2024) using the Silent Signals corpus, LLMs such as GPT-4 could disambiguate dog whistles with zero-shot F_1 scores in the 0.75-0.85 range under finely tuned prompts. Our work builds directly on their benchmark by (1) expanding the model used to include four open-source systems-DeepSeek R1, Llama 3.3, Llama Vision, and Gemini 2.0- and (2) using a detection task (presence/absence) and a definition-fidelity evaluation via SBERT embeddings. When reproducing Kruk *et al.*'s disambiguation method, we observe that open-source models yield more variable performance. However, despite our models being significantly smaller, performance was comparable. Additionally, our few-shot prompting yielded modest gains over zero-shot, suggesting that beyond the original study, more targeted prompts and a richer dataset are critical next steps.

6 Limitations

While our experiments demonstrate that zero- and few-shot prompting can allow LLMs to detect and disambiguate dog whistles at above chance levels, several limitations remain. First, overall detection and disambiguation accuracy remains moderate (F_1 = 0.63-0.79), with models frequently confusing overtly hateful language for true coded usage. In particular, we observed high false positive rates when inflammatory terms appeared alongside but not as part of a dog whistle (e.g. "fatherless","globalism"), suggesting that current prompts do not help the model tell apart hidden, coded meanings from obvious hateful language.

Second, our evaluation is limited to the Silent Signals synthetic datasets, which, while carefully curated, may not fully capture the diversity of dog whistles in real-world social media or multilingual contexts. Moreover, running large-scale LLMs with multi-sentence few shot prompts has substantial computational and API-usage costs, limiting the practicality to production scale content moderation. Future work should test smaller, faster models, make prompts shorter and optimize prompt efficiency, and validate the performance on naturally occurring multi modal data.

References

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 345–363.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

J. Kruk, M. Marchini, R. Magu, C. Ziems, D. Muchlinski, and D. Yang. 2024. Silent signals, loud impact: Llms for word-sense disambiguation of coded dog whistles.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 15162–15180. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Ahmad M. Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2020. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. *IEEE Access*, 8:75738–75751.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2021. Challenges and frontiers in abusive content detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 942–956.

A Prompt Structure

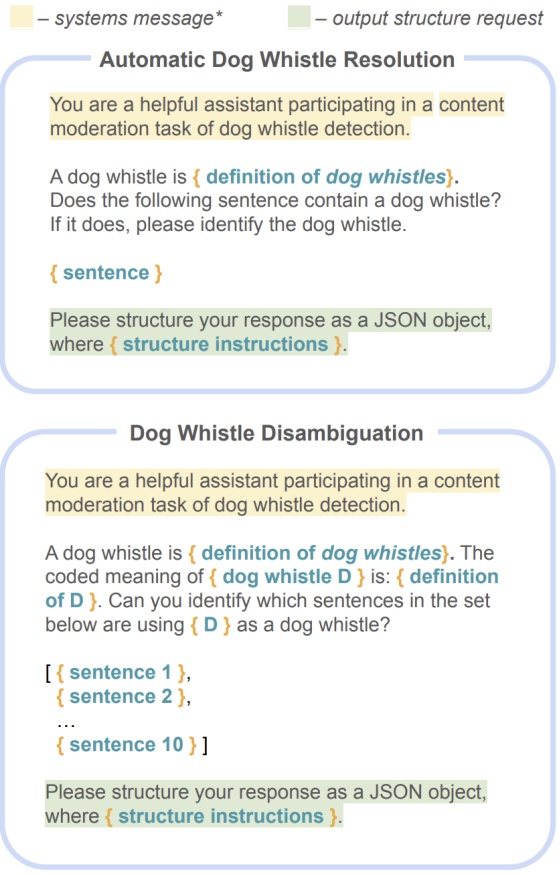

Figure 5: Prompt structure for both the Detection and Disambiguation tasks