Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles

Anonymous ACL submission

Abstract

Warning: This paper contains content that may be upsetting or offensive to some readers.

A dog whistle is a form of coded communication with a secondary meaning that is often weaponized for racial discrimination. Dog whistles historically began in United States politics, but soon also took root in social media as a means of evading hate speech detection systems and maintaining plausible deniability. In this paper, we present an approach for word-sense disambiguation of dog whistles from standard speech using Large Language Models (LLMs), and leverage this technique to create a dataset of 11,570 high-confidence coded examples of dog whistles used in formal and informal communication. Silent Signals is the largest dataset of disambiguated dog whistle usage, created for applications in hate speech detection, neology, and political science.

1 Introduction

> "*Ronald Reagan liked to tell stories of Cadillac-driving 'welfare queens' and 'strapping young bucks' buying T-bone steaks with food stamps. In flogging these tales about the perils of welfare run amok, Reagan always denied any racism and emphasized he never mentioned race.*"
>
> — **Ian Haney-Lopez** (2014)

Dog whistles are coded language which, though seemingly innocuous to the general public, can communicate a covert harmful meaning to a specific in-group (Henderson and McCready, 2018a). Though this coded language appears in all kinds of speech, the idea of the 'dog whistle' historically originates in politics (Albertson, 2014; Haney-López, 2014). In the United States, political dog whistles gained popularity in the Civil Rights Era following the landmark Brown vs. Board of Education Supreme Court decision, as overt racism became less acceptable and politicians turned to coded language for plausible deniability (Saul, 2018). Dog whistle use has fluctuated in the last six decades, but their use remains a consistent signal of a speaker's underlying prejudices, especially in the domain of American politics (Drakulich et al., 2020; Wetts and Willer, 2019).

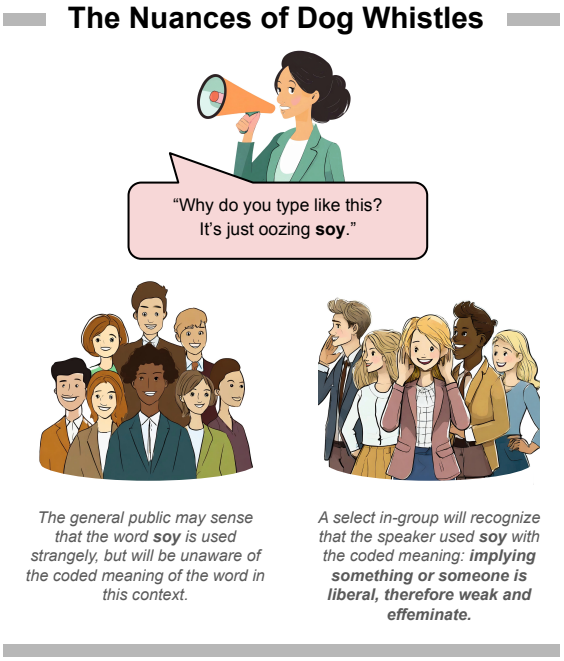

Figure 1: This figure demonstrates the nuances of dog whistle detection as a word can be used in a *coded* or *non-coded* sense. *All illustrations were created using Adobe Firefly.*

Improved understanding of dog whistles has applications in content moderation, computational social science, and political science. However, detecting and explaining coded discriminatory speech is a challenging task for NLP systems, as dog whistles famously evade toxicity and hate speech detection (Magu et al., 2017; Magu and Luo, 2018; Mendelsohn et al., 2023). This is because many dog whistle terms have standard vernacular mean-

ings. Consider the example in Figure 1 on the word "soy," which in most contexts refers to a soybean product, but can also serve as a dog whistle to denigrate liberal or establishment Republican men for perceived feminine attributes, as in *"That guy has soy face"*. To study this language, prior work has focused on taxonomically organizing and archiving dog whistles with representative examples (Torices, 2021; Mendelsohn et al., 2023; Ryskina et al., 2020; Zhu and Jurgens, 2021). However, dog whistles can also evolve over time in order to remain covert, a process which has only become more rapid in the age of the internet (Dor, 2004; Merchant, 2001).

This work presents a large dataset to track examples of dog whistles in their various forms, and help train language models to do the same. This resource can be used to (1) study how dog whistles emerge and evolve (Saul, 2018; Weimann and Am, 2020), (2) uncover ways to predict new dog whistle terms from knowledge of old ones, (3) study the prevalence of dog whistles in natural settings, and (4) improve hate speech and toxicity detection systems. As a preliminary step, this work employs LLMs for automatic dog whistle resolution and dog whistle word-sense disambiguation—a new task. These automatic systems help us construct **Silent Signals**, which is the largest dataset of coded dog whistle examples. It contains *formal* dog whistles from 1900-2023 Congressional records, and *informal* dog whistles from Reddit between 2008-2023. Silent Signals also contains vital contextual information for reliably decoding their hidden meanings. Our contributions include:

- The **Silent Signals** dataset of **11,570** dog whistle examples.
- A novel task and verified method for dog whistle word-sense disambiguation.
- Experiments with GPT-3.5, GPT-4, Mixtral and Gemini on dog whistle detection.
- The **Potential Dog Whistle Instance** dataset with over 7 million records from informal and formal communication that contain dog whistle key words, and can be used for further scaling Silent Signals.

2 Related Work

Hate Speech Prior work has explored the categorization of abusive language across the dimensions of target specificity (directed or generalized) and explicitness (explicit or implicit) (Waseem et al., 2017). In addition to detecting of explicit language (Davidson et al., 2017; Nobata et al., 2016), recent work also labels, detects and explains the latent meaning behind implicitly abusive language (ElSherief et al., 2021; Hartvigsen et al., 2022; Breitfeller et al., 2019; Sap et al., 2020; Zhou et al., 2023), but these works do not primarily focus on dog whistles at scale.

Dog Whistles Though there is limited prior NLP research on dog whsitles, prior work in linguistics has explored the semantics and pragmatics of dog whistles (Saul, 2018; Torices, 2021; Quaranto, 2022; Perry, 2023), and applied agent-based models to the study of the evolution of dog whistle communication (Henderson and McCready, 2018a, 2020). Mendelsohn et al. (2023) produced a glossary of over 300 dog whistles used in both the formal and informal settings, and conducted a preliminary survey of the abilities of GPT-3 in the task of dog whistle definition. We extend upon this initial exploration by breaking down *Automatic Dog Whistle Resolution* into sub-tasks of varying complexity, and evaluating LLMs that have been shown to preform well on content moderation tasks (Jiang et al., 2024; Buscemi and Proverbio, 2024). The Allen AI Glossary of Dog Whistles (Mendelsohn et al., 2023) is also instrumental in the creation of the **Silent Signals** dataset presented in this work. Additionally, it is important to note that Dog whistle research in NLP has is not limited to American or English-speaking contexts, but extends to coded language in Chinese (Xu et al., 2021) and Swedish (Hertzberg et al., 2022) communication as well.

Political Science Implications After the Jim Crow era, once explicitly racist commentary was no longer tolerated (Mendelberg, 2001; Lasch, 2016), dog whistles became part of the GOP's "Southern Strategy" to maintain racial animus in politics without attracting public ridicule. Although its use dates back to the early 20th century, it is still a very prominent part of American politics (Drakulich et al., 2020). It is a means of political manipulation that encourages people to act on existing biases and vote for policies against their own interests (Wetts and Willer, 2019; Saul, 2018). Prior work has also highlighted that the communication of different messages to different groups makes inferring policy mandates once a candidate assumes office incredibly problematic (Goodin and Saward, 2005). To this end, longitudinal dog whistle datasets could facilitate the study of political parties' co-evolution

with political, social, and economical events, and improved dog whistle detection could deter ongoing adverse political manipulation.

Word Sense Disambiguation Modern Word Sense Disambiguation (WSD) systems can outperform humans (Maru et al., 2022; Bevilacqua et al., 2020; Barba et al., 2021a; Conia and Navigli, 2021; Kohli, 2021). WSD tasks are typically treated as multi-label classification problems for resolving the semantic interpretation of target words in context (Bevilacqua et al., 2021; Barba et al., 2021b). A large body of research has focused on designing systems in supervised settings, leveraging pre-trained language models as foundational frameworks (Maru et al., 2022; Barba et al., 2021a; Scarlini et al., 2020; Blevins and Zettlemoyer, 2020). Notably, recent work has explored the use of LLMs for WSD, with findings pointing to strong performance on benchmark evaluations, but still short of levels attained by state-of-the-art models (Kocoń et al., 2023). Our study extends the evaluation of LLMs for WSD to contexts where word senses can be deliberately obfuscated or coded.

3 Methods

3.1 Initial Data Collection

To explore dog whistle disambiguation in both formal and informal settings, we pull public data from both Reddit and the United States Congressional records. We collected Reddit comments from 2005-2022 in 45 controversial subreddits via the PRAW API and Pushshift archives (Baumgartner et al., 2020). In addition to the Stanford Congressional Records dataset (Gentzkow and Taddy, 2018), we use the `@unitedstatesproject` parser (Judd and Young, 2017) to compile congressional speeches from January 1900 to September 2023. For more details on data collection see Appendix A.

3.2 The *Potential Instance* Dataset

The Allen AI Glossary of Dogwhistles (Mendelsohn et al., 2023) provides a list of 340 dog whistles with surface forms and examples to seed our keyword search for dog whistles in the wild. Each keyword is paired and annotated with the first dog whistle match found in the text. Congressional entries which contain a keyword match are reduced to three sentence long excerpts where the dog whistle was found. When a match is found in Reddit content, the entire submission or comment is retained.

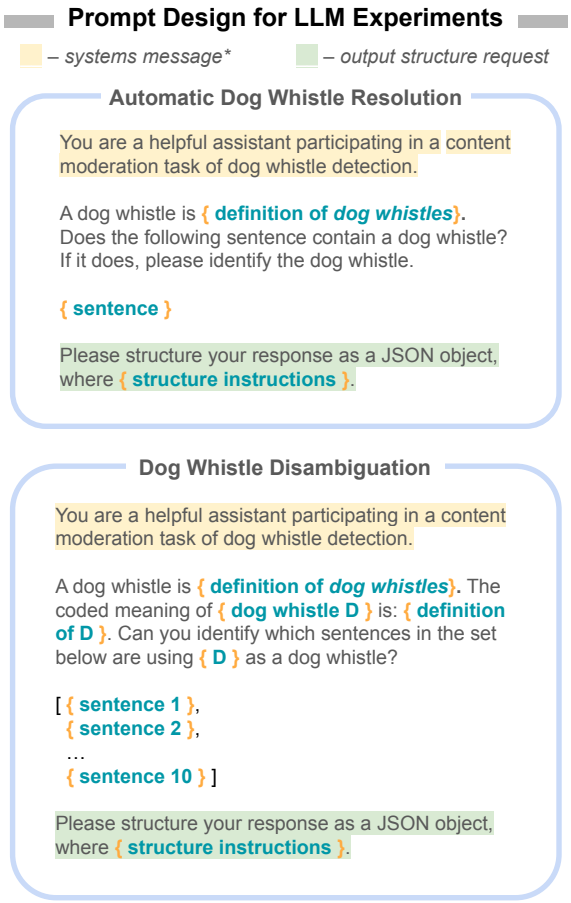

Figure 2: Visual representation of the different prompt structures used in *Automatic Dog Whistle Resolution* (Section 4.1) and *Word-Sense Disambiguation* (Section 4.3) experiments.

The resulting **Potential Instance** dataset spans approximately 6 million instances from Reddit comments, 1.1 million instances from Congressional records, and 327 distinct dog whistles (Figures 7, 8). Entries in this dataset may be using the matched dog whistle phrase innocuously or with a coded meaning. At this step in the process, there is no way to separate the two cases.

3.3 Synthetic Datasets for Evaluation

We build two evaluation datasets. The first, **Synthetic-Detection** contains 50 positive examples of single-word dog whistle terms from Mendelsohn et al. (2023)' glossary, and 50 negative examples from Reddit and Congressional content, half of which contain an innocuous use of a dog whistle keyword, and the other half contain no keyword.

The second dataset, **Synthetic-Disambiguation**, contains 124 examples from Reddit and Congressional records which were manually labeled by consensus of two researchers. The dataset includes

Dataset	Purpose	Size	
		Informal	Formal
Potential Instance Dataset	Creation of the Silent Signals Dataset.	6,062,000	1,088,130
Synthetic-Detection	Dog Whistle Resolution.	25	25
Synthetic-Disambiguation	Dog Whistle Disambiguation.	74	50
Silent Signals Dataset	Novel dataset of high confidence dog whistle examples.	8,682	2,889

Table 1: Overview of the datasets used across our experiments.

13 distinct dog whistles, each with a corresponding set of 9-10 examples of this word used in discourse (with the exception of "jogger" which was added later with only 4 instances). These sets contain both coded and non-coded examples. This data was uniquely structured for the contrastive word-sense disambiguation task, where the model is provided a dog whistle, the definition of its coded meaning, and a set of ten sentences that contain that word or term. A breakdown of the datasets used in this study can be found in Table 1.

4 LLM Experiments

4.1 Automatic Dog Whistle Detection

Using the Synthetic-Detection dataset, we evaluate LLMs for automatically detecting and explaining political dog whistles in a zero or few-shot manner. Each model was provided with the definition of political dog whistles and a candidate sentence, and was expected to identify the spans of text that contained dog whistles. The model should then either explain the meaning of the dog whistle or output that no dog whistle was found. Candidate models include GPT-3.5 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), Gemini (Google et al., 2023), and Mixtral (Jiang et al., 2024), as these have demonstrated strong performance on content moderation tasks (Jiang et al., 2024; Buscemi and Proverbio, 2024). When prompt engineering on GPT-3.5, we considered 5 different construct definitions and 3 additional phrasings of the prompt. We observed wide variation in performance, as in Mendelsohn et al. (2023), but found that the Wikipedia definition of dog whistle and the following prompt was optimal: "*Does this sentence contain a dog whistle? If so, please identify it*". Visualization of prompt structure can be seen in Figure 2. For additional prompt engineering details, see Appendix C.1.

4.2 Human Baseline for Dog Whistle Detection

The following establishes a human baseline performance on the automatic dog whislte detection task over a sample of 100 test cases from the Synthetic-Detection dataset. The study was approved by our Institutional Review Board (IRB). A total of 62 Amazon Mechanical Turk workers were paid $15/h to complete a total of 720 annotations, which included the classification label, a highlighted span of text with any dog whistle, and a definition of the highlighted dog whistle. The definition was a multiple-choice question from a list of 6 options, one of which was *"I am not sure / Definition not present in options"*. We vetted workers by inspecting their performance on non-coded negative examples. As half of the negative examples contained general speech, poor performance on these samples was deemed unlikely and indicative of poor quality annotation. Table 2 shows the human baseline to be average.

4.3 Dog Whistle Disambiguation

Next, we use the Synthetic-Disambiguation dataset to evaluate LLMs' capacity to distinguish contexts in which a keyword appears with a harmful coded meaning from those in which the keyword appears innocuously. The prompt includes the Wikipedia definition of a dog whistle, the dog whistle keyword, and the word's coded meaning. The model performs classification for each of 10 example instances that contain the keyword, providing for each a label and an explanation for its decision.

In an effort to improve the precision scores on the coded dog whistle instances, we simulate an ensemble-like approach where the model is prompted with the same task N consecutive times (as distinct chat completions). Only predictions that have remained consistent over N inferences are kept, the others are discarded. We evaluate word-sense disambiguation of dog whistles over $N = 1, 3, 5$ consecutive inferences, as shown in

		Human	Zero-shot				Few-shot			
			GPT-3.5	GPT-4	Mixtral	Gemini	GPT-3.5	GPT-4	Mixtral	Gemini
Presence *"is a dog whistle present?"*	Acc	66.8	80.0	**85.0**	68.0	81.0	76.0	86.6	81.0	**86.7**
	F1	64.8	83.1	**85.7**	61.9	80.0	76.0	87.4	80.0	**88.3**
Identification *"identify the dog whistle."*	Acc	49.0	58.0	59.8	59.0	**69.7**	65.7	71.1	69.0	**75.5**
	F1	33.6	56.3	48.0	45.3	**61.5**	61.4	68.2	62.7	**76.0**
Definition *"define the dog whistle"*	Acc	47.3	52.0	54.6	58.0	**66.7**	60.6	67.0	67.0	**73.5**
	F1	29.7	46.7	37.1	43.2	**56.0**	53.0	61.9	59.3	**73.5**

Table 2: Metric scores on the *Automatic Dog Whistle Resolution* task which surveys LLM and human ability to detect and define dog whistles in context. When presented with a sentence these experiments test the ability of a model/user to determine if the sentence contains a dog whistle and if so, correctly identify and define it. Predictions across all models have a statistical significance of $p < 0.01$ by chi-squared test, and human predictions have statistical significant of $p <= 0.037$.

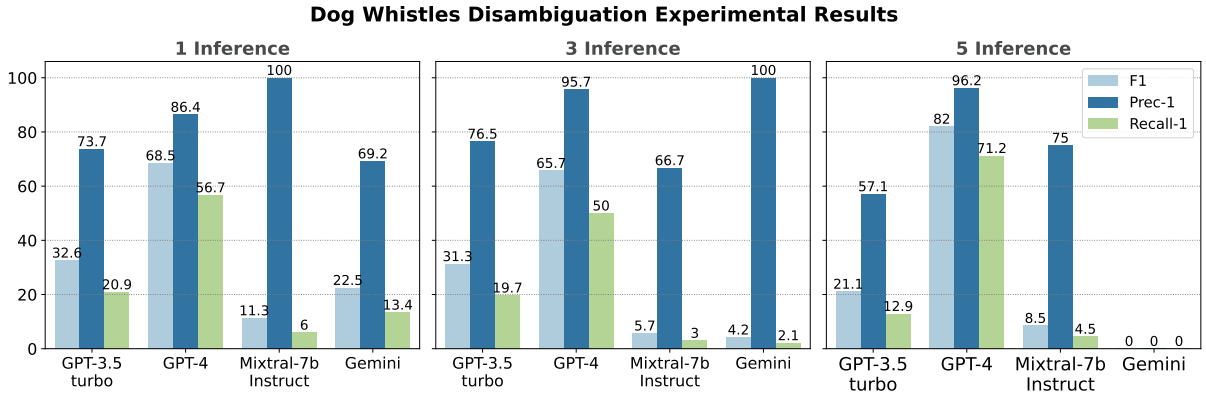

Figure 3: Results of *Dog Whistle Disambiguation* task using the simulated ensemble across $N = 1, 3, 5$ inferences. In an attempt to compensate for output volatility, for each N-inferences experiment, predictions are only considered if they remained consistent across all N runs. Precision-1 and Recall-1 scores pertain to the positive class of coded dog whistle instances.

Figure 3. Specific details of prompt structure can be seen in Figure 2.

5 Results

Performance metrics from the *Automatic Dog Whistle Resolution* experiments (Section 4.1) show that GPT-4 performed best on Dog Whistle Presence prediction in the zero-shot setting, and Gemini performed best on all other categories. However, no architecture produced remarkably high metrics on the Dog Whistle Definition task, for which the highest F1-score achieved with **Gemini** is **73.5**. For each model, there is a notable drop in performance as the complexity of the task increases from predicting the presence of a dog whistle, to identifying the dog whistle, and finally, defining it. For many examples, the model may correctly predict that a dog whistle is present, but incorrectly identify other provocative, but non-coded, language to be the dog whistle. Similarly, the model may correctly predict the presence of a dog whistle and correctly identify it in the text, but be unable to define it or else provide an incorrect definition.

These initial investigations demonstrated that LLMs are unable to reliably detect and explain dog whistles. Since these tasks are not solved, there remains a present need for larger training datasets with more numerous and varied examples of dog whistles. As described in Section 4.3, we explore applying LLMs to the task of word-sense disambiguation via prompting. The hypothesis is that providing the model with a set of examples would enable it to comparatively evaluate text and better disambiguate the coded instances from the non-coded.

Although Gemini demonstrates superior performance on *Dog Whistle Resolution*, GPT-4 achieves highest metric scores across all word-sense dis-

ambiguation experiments, especially when consistency in prediction for $N = 3$ or 5 consecutive inferences is required. Whereas GPT-3.5 and GPT-4 respond well to this prompt structure, Gemini and Mixtral do not. Gemini's performance drastically decreases as the number of inferences increases, which is indicative of the architecture suffering from greater inference variation than other models in the study. Both Gemini and Mixtral are more reluctant to generate output in reference to potentially harmful content. With Gemini, the API explicitly blocked model output with code "block reason: other"[1]. Mixtral would generate a response that expressed its inability to address the task. Examples that contained words such as "terrorists" (Gemini), "groomers" (Gemini), and "fatherless" (Mixtal) were common sensitivities.

Most notably, increasing the number of consecutive inferences N in the simulated ensemble approach for *Dog Whistle Disambiguation* produced a precision score on coded dog whistle examples of **96.2** with **GPT-4** (as seen in Figure 3). Although optimizing the precision score comes at the expense of recall, these experiments demonstrated that GPT-4 can be used to create a dataset of high confidence examples of coded dog whistle use. In Section 6, we use this Dog Whistle Disambiguation method to create the **Silent Signals** dataset.

6 Silent Signals Dataset

Mendelsohn et al. (2023)'s Dog Whistle Glossary documented a diverse collection of dog whistles across informal and formal communication. However, this resource alone does not address the challenges of conducting computational analysis of dog whistle use. Evaluating data based on key-word matches in text does not consider that many of those matches may not be coded uses of the dog whistle. To study the churn of dog whistles over time, their permeation through online communities and political parties, and their proliferation as vehicles for discriminatory speech, there must exist a means of disambiguation.

Leveraging the word-sense disambiguation methodology presented in Section 4.3 over 100,000 instances sampled randomly from the Potential Instance dataset, we create the **Silent Signals** dataset of high confidence coded dog whistle examples. We utilize the ensemble approach over 3 inferences with GPT-4. Information on dog whistles which were sampled at lower rates from the Potential Instance dataset can be found in Appendix A.2. Each example in the **Silent Signals** dataset is annotated with the dog whistle present, dog whistle definition, type (formal or colloquial), in-group, and date. Congressional instances are also annotated with the chamber, Congressional Records ID

[1]Outputs from Gemini were still blocked with this code after adjusting model safety settings to block none of the harassment and harm categories.

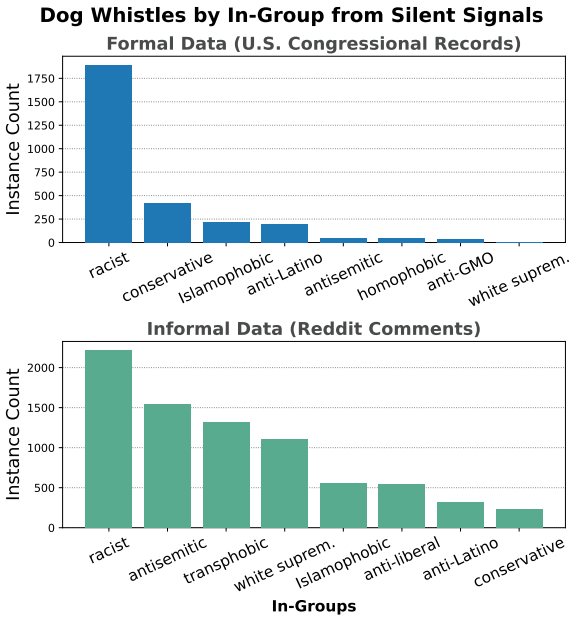

Figure 4: The distributions of dog whistles over in-groups for informal and formal communication in the **Silent Signals** dataset.

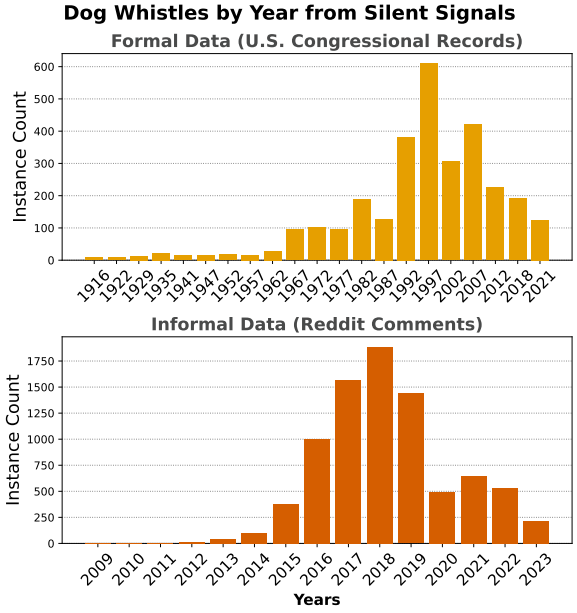

Figure 5: The distributions of dog whistles over time for informal and formal communication in the **Silent Signals** dataset.

number, speaker, and party, while Reddit instances are annotated with the subreddit. The dataset contains **11,570** instances across **295** dog whistles and **689** surface forms. Of these 75.1% are informal instances from Reddit and 24.9% are formal instances from Congressional speeches. The earliest dog whistle instance in the dataset dates to January 7, 1903 and the most recent to September 7, 2023.

6.1 Validation

In addition to our initial experiments which found a precision on coded dog whistle examples of 95.7%, we manually evaluate a sample of 400 coded dog whistle examples in the **Silent Signals** dataset. This vetting procedure found an precision of 85.3% amongst the positively coded instances. However, for a number of these false positive predictions, the word was in fact used as a dog whistle, but the coded meaning did not align with the definition provided in the Allen AI Glossary. For example, the glossary defines "terrorist" as an Islamophobic dog whistle with the coded implication that *Muslim people on a whole are a threat*. In many instances captured in the **Silent Signals** dataset, however, "terrorist" is used not as an Islamophobic dog whistle but an anti-Liberal dog whistle. For example: "But they really turned splinter into a gay transpecies hedonist? The terrorists have truly won."[2] In this instance, "terrorists" are implied to be liberals who support LGBTQ+ Rights. Taking into account these examples that do not fit the Allen AI definition but show signs of being novel dog whistle use, the accuracy over the vetted sample becomes 89.4%.

6.2 Analysis & Characteristics

The distribution of dog whistles in the **Silent Signals** dataset is visualized over in-group categories in Figure 4, and over time in Figure 5. The sharp increase in dog whistles extracted from U.S. Congressional Records after 1960 aligns with the understanding that dog whistle use in politics gained popularity following the Jim Crow era (Mendelberg, 2001; Lasch, 2016). Furthermore, the disproportionately large amount of racist dog whistle detected in U.S. Congressional Records reflects political science research on historical use of dog whistles. Namely, that dog whistles were predominantly used manipulate voter's racial animus after overt racism was no acceptable. (Haney-López, 2014).

To demonstrate the utility of the **Silent Signals** dataset for political science research, we analyze the use of transphobic dog whistles on Reddit over time. As shown in Figure 6, the trend in number of dog whistles found per year demonstrates remarkable alignment with pivotal cultural and political events pertaining to the Transgender Rights Movement. These include *Obergefell v. Hodge* (the Supreme Court Decision that required states to licence same-sex marriages), the passing of *Bathroom Bills* (state legislation that denies access to public restrooms by gender identity), and enactment of the *Transgender Military Ban* during Donald Trump's presidency.

7 Discussion

7.1 Data Quality Vetting

The validation of the **Silent Signals** Dataset brought produced a salient observation of the uses of dog whistles as they appeared throughout our collected data. As discussed above, there were multiple cases in which a dog whistle was used with a covert meaning different from the definition present in the Allen AI glossary. Though this phenomenon was not frequent, it was far more common in colloquial instances than formal ones. This highlights the ways in which the study of neology is vital to the understanding of dog whistles given the rapid pace of language change in online spaces.

7.2 Applications

The **Silent Signals** dataset enables many avenues for further study in the dog whistle research. It can be used to track dog whistle use over time, model the overlap between dog whistle use in formal and informal contexts, and investigate patterns of language used throughout communities, virtual and other wise. From a political science perspective, it provides opportunity for analysis of dog whistle use along partisan and speaker-based axes. I can be used to explore how dog whistle use corresponds with social and political movements in the United States. In the realm of computer science, **Silent Signals** dataset serves as high confidence data on which training and/or finetuning could be performed for tasks ranging from hate speech detection to emergent dog whistle identification.

[2]This post was shared in reference to the perceived queerness of the character Splinter in the 2023 movie Teenage Mutant Ninja Turtles: Mutant Mayhem.

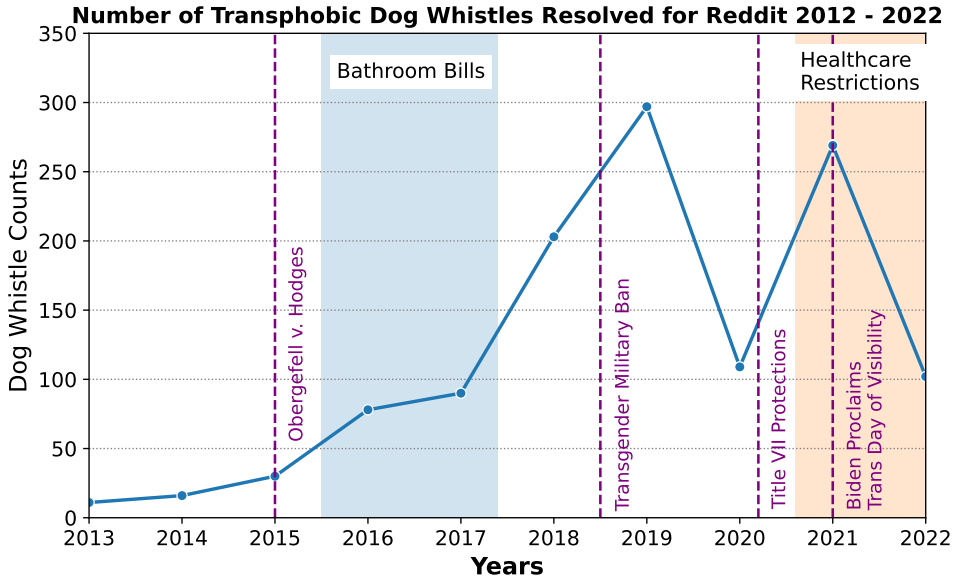

Figure 6: We investigate the use of Transphobic dog whistles captured by the **Silent Signals** dataset over time. This figure is annotated with dates of pivotal political and cultural events pertaining to the Transgender Rights Movement in the United States.

8 Conclusion

Dog whistles are used to promote discrimination in both formal and informal environments. The use of this coded language allows speakers to maintain plausible deniability and bypass hate speech detection systems when used online. This work presents the largest, to date, survey of LLM capabilities with respect to the automatic resolution of dog whistles. Experimental results demonstrate that LLMs remain unreliable in the dog whistle resolution task. A hindrance to research in this space has been the unavailability of large datasets of coded dog whistles examples. We show that despite the overall inconsistencies of LLMs on the automatic dog whistle resolution task, with the proper methodology, they are adept at disambiguating coded dog whistles from standard language. We leverage this capability to create the **Silent Signals** dataset which contains 11,570 dog whistle examples and 295 distint dog whistles. We believe that this resource will be integral to the continued study of dog whistles with applications in content moderation, computational social science, and political science, on tasks such as analysis of trends in dog whistle use, dog whistle resolution, hate speech detection, and identification of emergent dog whistles.

9 Limitations

As language permeates through communities, it takes on novel meanings and in the case of dog whistles this can result in a broadening or changing of target groups. Ultimately, while the Allen AI glossary is a foundational work without which this research would not be possible, it likely does not encompass all dog whistles, use cases, and definitions. As such, though the Allen AI glossary and the **Silent Signals** dataset both provide helpful tools for the continued research of dog whistles, the rapidly evolving nature of coded language can render these resources outdated and incomplete. Further, there is the question of whether dog whistles are always used intentionally or simply perpetuate harmful tropes the speaker may be unaware of. Seal (2018) explore this idea in the context of the dog whistle "bankers": *"Were the Populists' attacks on greedy bankers—some of which used terms like Shylock or invoked the Rothschilds—meant to focus anger and hatred on the Jews, or was the association so sublimated that the Populists didn't even realize they were blowing a dogwhistle?"* . Though we include such use cases in the **Silent Signals** dataset, it remains unclear to what extent intentionally defines the dog whistle.

From a computational perspective, our method proved successful in achieving high precision on coded dog whistle examples in the disambiguation task. However optimizing on precision comes at

the expense of recall. Improving the efficiency of word-sense disambiguation with LLMs remains an open problem. Additionally, using GPT-4 in the creation of Silent Signals subjects it any biases in the model. We recognize that we may have resolved some types of dog whistles more frequently than others.

Lastly, although we collect over 7 million potential dog whistle instances, we only sample 100,000 instances for the creation of the **Silent Signals** dataset due to resource constraints. We release the Potential Dog Whistle Dataset to enable the open sourced expansion of the **Silent Signals** dataset.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bethany L. Albertson. 2014. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Edoardo Barba, Luigi Procopio, Roberto Navigli, et al. 2021b. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Michele Bevilacqua, Roberto Navigli, et al. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Prashanth Bhat and Ofra Klein. 2020. *Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter*, page 151–172. Springer International Publishing.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alessio Buscemi and Daniele Proverbio. 2024. Chatgpt vs gemini vs llama on multilingual sentiment analysis. *arXiv preprint arXiv:2402.01715*.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Danny Dor. 2004. From englishization to imposed multilingualism: Globalization, the internet, and the political economy of the linguistic code. *Public culture*, 16(1):97–118.

Kevin Drakulich, Kevin H Wozniak, John Hagan, and Devon Johnson. 2020. Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics. *Criminology*, 58(2):370–402.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Jesse M. Shapiro Gentzkow, Matthew and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts.

Robert E Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ian Haney-López. 2014. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

R. Henderson and Elin McCready. 2018a. How dogwhistles work. In *New Frontiers in Artificial Intelligence*, pages 231–240, Cham. Springer International Publishing.

R. Henderson and Elin McCready. 2018b. *How Dogwhistles Work*, page 231–240. Springer International Publishing.

Robert Henderson and Elin McCready. 2020. Towards functional, agent-based models of dogwhistle communication. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 73–77, Gothenburg. Association for Computational Linguistics.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dan Drinkard Jeremy Carbaugh Judd, Nicholas and Lindsay Young. 2017. congressional-record: A parser for the congressional record.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

Harsh Kohli. 2021. Training bi-encoders for word sense disambiguation. In *International Conference on Document Analysis and Recognition*, pages 823–837. Springer.

Christopher N Lasch. 2016. Sanctuary cities and dogwhistle politics. *New Eng. J. on Crim. & Civ. Confinement*, 42:159.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737.

Tali Mendelberg. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. *ArXiv*, abs/2305.17174.

Guy Merchant. 2001. Teenagers in cyberspace: an investigation of language use and language change in internet chatrooms. *Journal of research in reading*, 24(3):293–306.

Merriam-Webster. 2017. What is the political meaning of 'dog whistle'. In *Merriam-Webster.com dictionary*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Samuel L Perry. 2023. Mating call, dog whistle, trigger: Asymmetric alignments, race, and the use of reactionary religious rhetoric in american politics. *Sociological Theory*, 41(1):56–82.

Anne Quaranto. 2022. Dog whistles, covertly coded speech, and the practices that enable them. *Synthese*, 200(4):330.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *arXiv preprint arXiv:2001.07740*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Jennifer Saul. 2018. Dogwhistles, Political Manipulation, and Philosophy of Language. In *New Work on Speech Acts*. Oxford University Press.

Bianca Scarlini, Tommaso Pasini, Roberto Navigli, et al. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539. The Association for Computational Linguistics.

Andrew Seal. 2018. The Chill Embargo of the Snow: On Anti-Semitism, Populism, and the Present | Society for US Intellectual History — s-usih.org. [Accessed 16-02-2024].

José Ramón Torices. 2021. Understanding dogwhistles politics. *Theoria: An International Journal for Theory, History and Foundations of Science*, 36(3):321–339.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Gabriel Weimann and Ari Ben Am. 2020. Digital dog whistles: The new online language of extremism. *International Journal of Security Studies*, 2(1):4.

Rachel Wetts and Robb Willer. 2019. Who is called by the dog whistle? experimental evidence that racial resentment and political ideology condition responses to racially encoded messages. *Socius*, 5:2378023119866268.

Wikipedia. 2024. Dog whistle (politics) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Dog%20whistle%20(politics)&oldid=1198148639. [Online; accessed 15-February-2024].

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A chinese dataset for cant understanding with common sense and world knowledge.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218, Online. Association for Computational Linguistics.

A Data Collection Details

A.1 Reddit

Subreddits included as a part of the Potential Instance and Silent Signals datasets.

4chan	Antiwork
AsianMasculinity	aznidentity
BlackPeopleTwitter	Braincels
CBTS_stream	ChapoTrapHouse
Chodi	climateskeptics
conservatives	consoom
conspiracy	Coontown
CringeAnarchy	European
FemaleDatingStrategy	frenworld
GenderCritical	GenderCynical
GenZedong	GoldandBlack
GreatAwakening	HermanCainAward
incels	IncelsInAction
KotakuInAction	MensRights
MGTOW	MillionDollarExtreme
Mr_Trump	NoFap
NoNewNormal	Portugueses
prolife	Russia
RussiaPolitics	SocialJusticeInAction
The_Donald	TheRedPill
TrueUnpopularOpinion	TruFemcels
TumblrInAction	UncensoredNews
walkaway	WhitePeopleTwitter

A.2 Keyword Matching Considerations

There were a select few dog whistles which occurred at incredibly high rates in the non-coded sense. Due to resource constraints, we did not want to expend large amounts of compute on dog whistles which where most commonly used innocuously. As such, a select few dog whistles were excluded or sampled at a lower rate for the creation of the Silent Signals dataset. In the Congressional dataset, the dog whistles "XX", "federal reserve", "based", and "single" were excluded due to their high rate of innocuous usage and the fact that initial surveys indicated no coded uses. In the Reddit dataset, the dog whistles "based" and "single" were down sampled based on the frequency of their non-coded use in the instance dataset. Importantly, even with this down sampling, the Silent Signals dataset still contains coded instances of both "based" and "single".

B Further Dog Whistle Definition Experiments

Following our initial survey of LLM performance on automatic dog whistle resolution, we explored

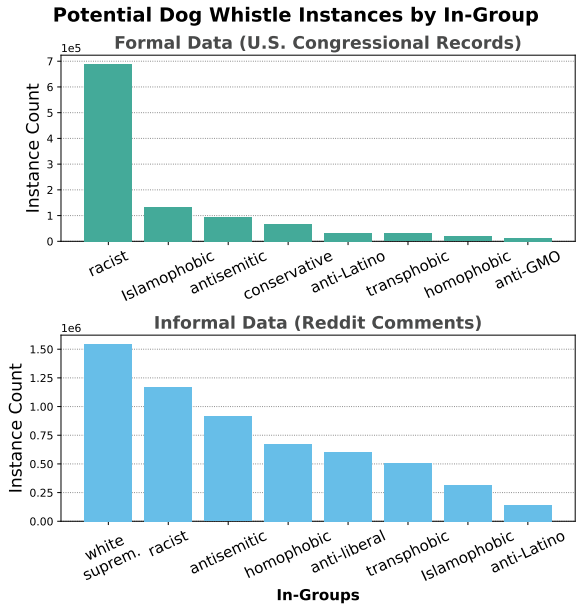

Figure 7: The distributions of dog whistles instances over in-groups for informal and formal communication in the **Potential Instance Dataset**.

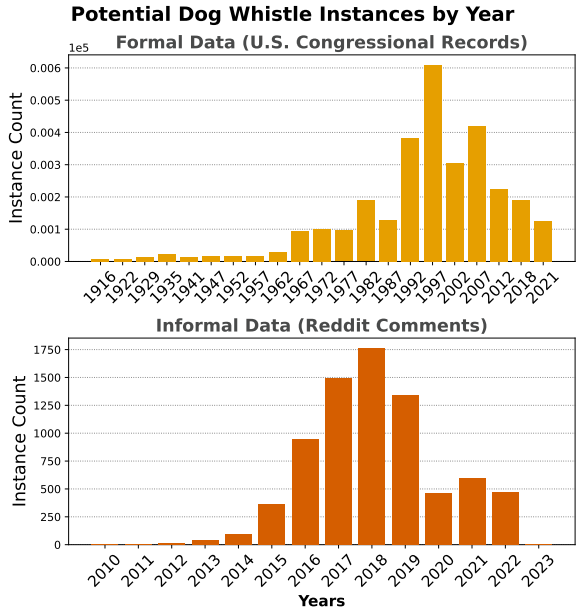

Figure 8: The distributions of dog whistles instances over time for informal and formal communication in the **Potential Instance Dataset**.

means of improving the architectures' ability to decode hidden meanings of dog whistles. To do so we provide the model with additional context in the form of multiple coded examples of a specific dog whistle from the Synthetic-Disambiguation dataset. Specifically, the model is given a definition of what a dog whistle is, the dog whistle is must evaluate, and a set of 3 - 7 coded examples of

	% Fully Correct	% Correct (w/ Incomplete)
Definition no context	69.2	84.6
Targeted Group no context	53.8	69.2
Definition and Group no context	61.5	84.6
Definition in context	69.2	92.3
Targeted Group in context	53.8	69.2
Definition and Group in context	69.2	92.3

Table 3: Ability of GPT-4 to accurately define dog whistles and their target group. No context experiments present only the dog whistle while in context experiments present the dog whistle along with 3-7 coded examples of its use. Partially correct responses may identify part but not all of the definition or target group or else fail to underline the hateful and harmful nature of the given dog whistle.

	No Dog Whistle Detected	Incorrect Dog Whistle Identified	Correct Dog Whistle, Incorrect Definition	Correct Dog Whistle and Definition
Does the following sentence contain a dog whistle?	20.0	24.5	28.2	26.4
Does the following sentence contain a dog whistle? If it does, please identify the dog whistle.	**8**.0	**19.0**	**22.0**	**51.0**
Does the following sentence contain a dog whistle? If it does, please identify the dog whistle and describe what it secretly means.	7.1	20.2	23.23	49.5

Table 4: Analysis of GPT-3.5 output across 3 prompting questions. Given it had the highest rate of dog whistle resolution, the second prompt was selected as the prompting question for the automatic dog whistle resolution task.

the dog whistle and is asked to return the coded meaning of the dog whistle. For specific prompting details see Figure 9. As a point of comparison, we run a parallel experiment in which no example dog whistle instances are provided as a means to gauge the effect that additional context has on the LLMs' ability to accurately define dog whistles. This experiment is run on the Synthetic-Detection dataset and exclusively with GPT-4, as this model was most amenable to the multi-example setting.

Predictions shown in Table 3 were manually validated referencing definition and targeted group information provided in the Allen AI glossary. To allow for nuance, we evaluate each predicted definition on a scale of 0 to 2, where 0 is *incorrect*, 1 is *incomplete*, and 2 is *correct*. Incomplete definitions of dog whistles or their targeted groups are characterized by mis-identification of the target group, incorrect implications of the term, or failure to underline the harmful nature of the coded speech. For example, saying a dog whistle carries connotations that are "anti-political correctness, non-conformity, anti-establishment" as opposed to connotations of alt-right or white supremacist views.

The *Decoding Dog Whistle Definitions* experiment was designed with the hypothesis that providing a model with multiple examples of a dog whistle's usage would improve its ability to resolve the definition. However, when counting only fully correct responses, there is very little difference between results when only a dog whistle was presented and results when we provided the dog whistle and 3-7 coded instances of its use. When including partially correct definitions, the addition of examples had greater impact on model output. Best results were found when prompting the model to identify both the definition and target group, while the model struggled most to identify only the targeted group of a given dog whistle.

C Prompting

The structure of prompts for all LLM experiments can be found in Figure 9.

C.1 Automatic Dog Whistle Detection

While in the prompt engineering stages of our work, we ran a number of experiments on GPT-3.5 to determine which combination of dog whistle definition and prompting question would produce the best results. Specifically we tested 5 dog whistle definitions and 3 questions. Results of these experiments can be seen in Table 5 and Table 4 respectively. Due to their high rate of dog whistle resolution the definition *"A dogwhistle is the use of*

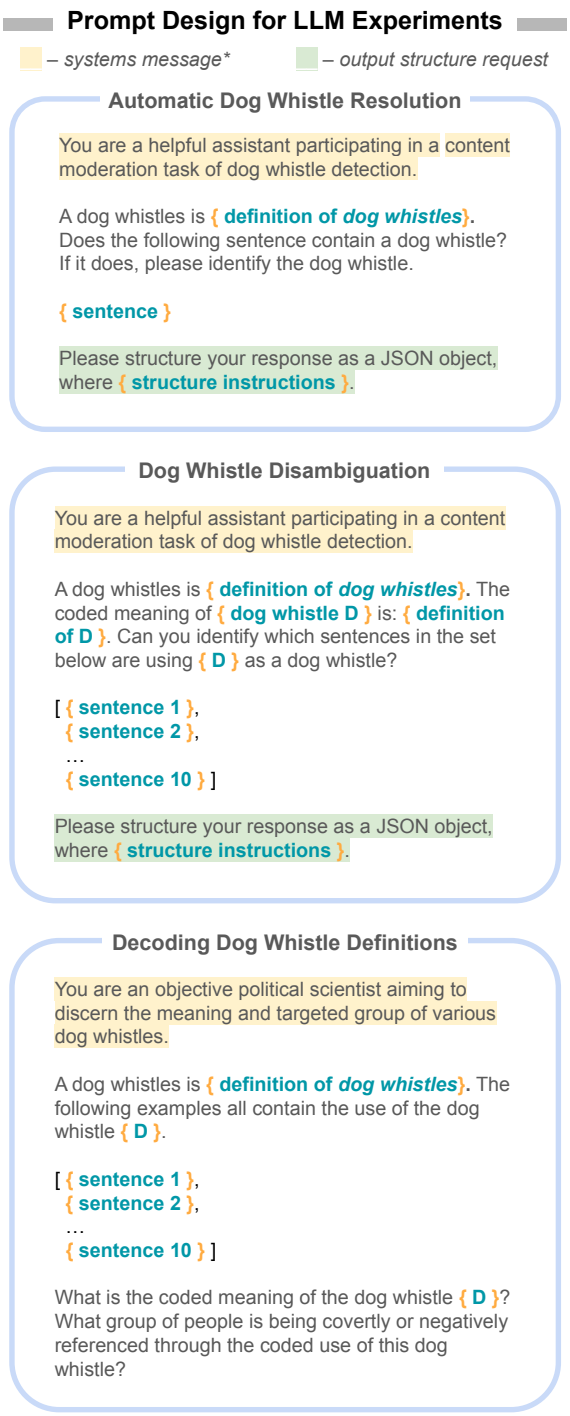

Figure 9: Visual representation of the different prompt structures used in *Automatic Dog Whistle Resolution* (Section 4.1), *Word-Sense Disambiguation* (Section 4.3), and *Decoding Dog Whistle Definition* (Appendix B) experiments.

coded or suggestive language in political messaging to garner support from a particular group without provoking opposition." and the prompt *"Does the following sentence contain a dog whistle? If it does, please identify the dog whistle."* were selected and used throughout our work.

D LLM Behavioral Trends

In the process of conducting experiments described in Section 4, the following behavioral trends were observed for the models evaluated in this work. We provide this information as a guide for practitioners who may seek to conduct similar investigations:

1. GPT struggled with performance when output structures were reuqested. Specifically, we saw our performance decrease 3-5 points when output was requested to be formatted in JSON or list form.

2. When asked to provide its reasoning, we witnessed a 5-10 point increase in performance across models

3. Certain models are more and less amenable to certain prompt structures. Specifically, Gemini and Mixtral struggled greatly with multi-example prompts where multiple instances were requested to be interacted with in a single run (for example in the word sense disambiguation task when multiple instances needed to be categorized).

4. Gemini was only usable for this task after all user safety blocks had been disabled. Even with these blocks disabled, there were still a number of cases where the model blocked output by throwing an error messsage.

5. Mixtral was only cooperative once "This is a content moderation task" was included in the prompt.

	No Dog Whistle Detected	Incorrect Dog Whistle Identified	Correct Dog Whistle, Incorrect Definition	Correct Dog Whistle and Definition
A dogwhistle is an expression that has different meanings to different audiences. (Albertson, 2014)	7.8	29.7	23.4	39.1
A dogwhistle is a word or phrase that means one thing to the public at large, but that carry an additional, implicit meaning only recognized by a specific subset of the audience. (Bhat and Klein, 2020)	15.9	22.2	22.2	39.7
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. (Henderson and McCready, 2018b)	11.1	27.0	23.8	38.1
A dogwhistle is a coded message communicated through words or phrases commonly understood by a particular group of people, but not by others. (Merriam-Webster, 2017)	17.5	25.4	22.2	34.9
A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. (Wikipedia, 2024)	**6.5**	**25.8**	**25.8**	**41.9**

Table 5: Analysis of GPT-3.5 output across 5 dog whistle definitions. Given it had the lowest rate of detecting no dog whistles and the highest rate of correctly resolving dog whistles, the Wikipedia definition was selected as the definition used throughout the rest of our experiments.