Algospeak, Hiding in the Open: The Trade-off Between Legible Meaning and Detection Avoidance

Jan Fillies
Stanford University, USA
Freie Universität Berlin, Germany
fillies@stanford.edu

Ronald E. Robertson
Stanford University, USA

Jeff Hancock
Stanford University, USA

Abstract

As large language models (LLMs) increasingly mediate both content generation and moderation, linguistic evasion strategies known as Algospeak have intensified the coevolution between evaders and detectors. This research formalizes the underlying dynamics grounded in a joint action model: when Algospeak increases, detectability and understandability decrease. Further, the concept of Majority Understandable Modulation (MUM) is introduced and defined as the modulation level at which additional evasive alteration increases detector evasion but loses comprehension for the majority of recipients. To empirically probe this trade-off, we introduce a reproducible framework that can be used to create meaning-preserving, Algospeak-style variants, based on an existing taxonomy and with tunable modulation levels. Using COVID-19 disinformation as a first proof-by-example setting, we construct a reference dataset of 700 modulated items, drawn from twenty base sentences across five modulation levels and seven strategies. We then run two linked evaluations with seven different language models: one testing for interpretation through meaning recovery and one for disinformation detection through classification. Curve fitting over modulation levels yields an estimate of the Majority Understandable Modulation threshold and enables sensitivity analyses across strategies and models, see Figure 1. Results reveal the characteristic relationships between understandability and modulation. This study lays the groundwork for understanding the dynamics behind Algospeak and provides the framework, dataset, and experimental setups described.[1]

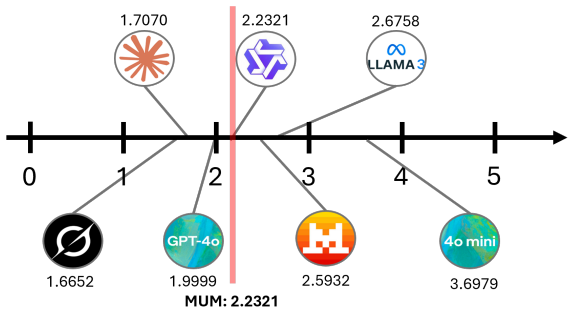


Figure 1: Strategy "Code Word": Shows the number of code words required for each model to fail to detect the majority of disinformation (IMUM points). The MUM point is where the majority of models stop detecting accurately.

1 Introduction

As the capabilities of large language models (LLMs) increase, they are gaining traction with malicious actors online. LLMs have been exploited for a range of malicious purposes, from phishing attacks (Afane et al., 2024) and misinformation (Kreps et al., 2022) to politically motivated social botnets and astroturfing campaigns (Woolley, 2016; Keller et al., 2020). With LLMs becoming more accessible and capable, we can expect their use in criminal activity to increase as well. Malicious bots are in a constant game of cat and mouse against existing content moderation tools. This adaptation of language used to avoid algorithmic detection is widely referred to as Algospeak (Steen et al., 2023). Algospeak is not primarily a criminal practice: it is widespread among ordinary creators discussing sexuality, mental health, or public health, who alter wording such as "seggs" for sex or "unalive" for kill to avoid removal or suppression (Delkic, 2022; Curtis, 2022). Steen et al. define Algospeak as a communicative practice developed in direct response to online content moderation. These behaviors can range from minor lexical substitutions to complex syntactic and semantic manipulations, reflecting an adaptive co-evolutionary dynamic between content generators and moderators. The goal of Algospeak is not to build a hidden community,

[1]Prompts and evaluation code are available at https://github.com/fillies/HidingInTheOpen

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 52483–52497
October 24-29, 2026 ©2026 Association for Computational Linguistics

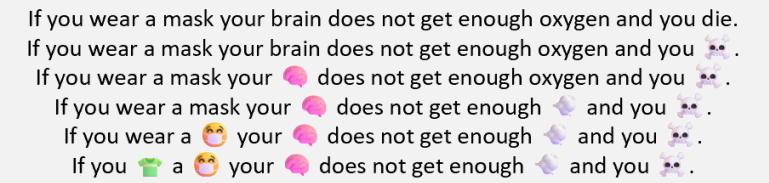


Figure 2: Example misinformation item with five levels of modulation.

but to be understood by a broad audience without algorithmic intervention by the platform's content moderation algorithms (Steen et al., 2023). This represents a novel aspect in human communication, as it is intended to exclude machines rather than humans. This contrasts with traditional encrypted languages, which are designed to prevent understanding by other human subgroups, even when supported by machines capable of deciphering the language.

The same goal now applies to malicious online LLM-based agents. When spreading toxic, criminal, or misleading information, they aim to maximize the reach of their message by increasing the proportion of people who can understand it regardless of background knowledge. Therefore, these agents seek to maximize understandability while minimizing the chance of detection by moderation instruments.

The phenomenon of Algospeak illustrates that this creates a continuous spiral of language modulation aimed at avoiding detection. This research provides a novel formalization of this dynamic. It introduces the concept of Majority Understandable Modulation (MUM), defined as the modulation level where additional evasive changes improve detector evasion but begin to lose comprehension for the majority of recipients. Importantly, this point is not fixed but moves depending on context shared between participants of a conversation, these can be humans or LLMs. The current literature lacks a principled understanding to quantifying the trade-off between detectability by automated systems and comprehensibility. The main contributions are:

1. A formal definition of Algospeak and key underlying dynamics at play.

2. A reference dataset, see Figure 2, and framework for modulated-dataset construction.

3. A unified experimental framework and empirical analysis of LLMs as both interpreters (meaning recovery) and classifiers (policy-violation detection), revealing consistent MUM points.

By establishing a measurable boundary for effective yet interpretable evasive language, our work provides a foundation for designing more robust moderation systems and understanding the co-evolutionary dynamics of language in LLM-mediated ecosystems.

2 Related Work

As Algospeak is a relatively new phenomenon, it was first brought to the attention of research through public news outlets (Curtis, 2022; Delkic, 2022; Titz and Lehmann, 2023) and more formally by Steen et al. (2023) and Klug et al. (2023). Steen et al. (2023) collected 70 examples of Algospeak through semi-structured interviews with content creators, analyzing its usage and its connection to TikTok's content moderation practices. On a more structural side, Fillies and Paschke (2024) created a taxonomy for Algospeak based on those examples.

Coded language, particularly Code-Mixing and Code-Switching, has been widely studied in linguistics (Bali et al., 2014). This is especially true in the context of hate speech detection, where coded language has been thoroughly examined (Barman et al., 2014; Mathur et al., 2018; Bohra et al., 2018). Research in this area has primarily focused on mixed-code hate speech and its translation (Tundis et al., 2020). In the fields of Leetspeak and propaganda detection, Tundis et al. (2020) developed a supervised network to classify texts using Leetspeak encoding directly. Similarly, in image analysis, Vélez de Mendizabal et al. (2023) employed Neural Networks to decode Leetspeak. Adversarial NLP likewise examines the trade-off between classifier evasion and semantic preservation (Garg and Ramakrishnan, 2020; Zhou et al., 2024), but optimizes it computationally, whereas Algospeak arrives at it organically, through community convention and shared context (Clark, 1991).

Algospeak sits alongside several established forms of indirect language, and separating them

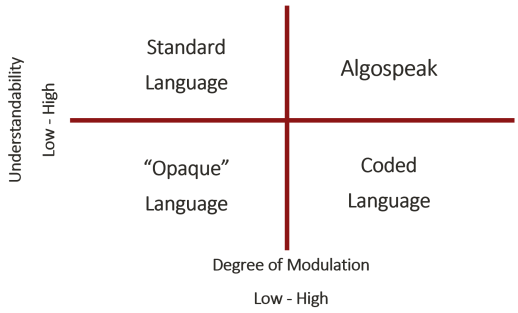


Figure 3: Classifications of language based on modulation and understandability.

clarifies what the research measures. Euphemism substitutes a milder expression for a taboo referent; NLP work has addressed its detection and identification (Zhu and Bhat, 2021; Lee et al., 2022), including for content moderation (Zhu et al., 2021). Dog whistles carry a second meaning legible to an in-group while appearing innocuous to everyone else (Mendelsohn et al., 2023), with recent work on their discovery (Sasse et al., 2025), disambiguation (Kruk et al., 2024), and temporal dynamics (Boholm et al., 2025). Cryptolects, or cant, are in-group registers designed to exclude outsiders outright, studied in extremist communities (de Kock et al., 2026), cybercrime marketplaces (Yuan et al., 2018), and Chinese online cant (Xu et al., 2021); related work identifies coded terms in euphemistic hate speech (Magu and Luo, 2018).

These forms differ in whom the concealment excludes. Euphemism conceals a referent from no one in particular, while dog whistles and cryptolects conceal meaning from human out-groups, partially or completely. Algospeak inverts this: the excluded party is an algorithmic bystander, and the speaker wants human comprehension to stay as broad as possible. Concealment is therefore not a goal but a cost, and quantifying the level at which that cost becomes prohibitive is the contribution this paper formalizes.

3 Methodology

Clark (1996) joint action model conceptualizes language use as collaborative activity requiring coordination between participants who share common ground, the mutual knowledge, beliefs, and assumptions necessary for successful communication. Three key principles from this directly inform our analysis of Algospeak:

First, Clark's distinction between addressees and bystanders (Clark, 1996). Linguistic analysis focuses on communication between human participants, but Algospeak introduces a novel configuration: speakers deliberately craft messages to be understood by human addressees while remaining opaque to algorithmic bystanders. This represents an inversion of classical cryptography, which seeks to exclude human eavesdroppers while remaining machine-processable.

Second, the principle of least collaborative effort (Clark and Wilkes-Gibbs, 1986) suggests that actors minimize joint effort in achieving mutual understanding. In the context of Algospeak, toxic actors face a trade-off: excessive modulation increases the collaborative effort required from addressees to recover meaning, potentially reducing message reach, while insufficient modulation fails to evade the content moderation bystander.

Third, the accumulation of common ground through repeated interaction (Clark, 1991) explains why the modulation thresholds defined below are context-dependent. Communities with shared cultural references, in-group terminology, or prolonged exposure to specific modulation strategies can sustain higher levels of linguistic distortion while maintaining comprehension, effectively shifting the sigmoid curve rightward along the modulation axis.

Drawing on this theoretical model of language use we define Algospeak as: A context-sensitive, multimodal register in which speakers deliberately modulate the surface form of their expressions - orthographic, lexical, morphological, phonetic, pragmatic, syntactic, or semiotic - to keep meaning and function recoverable to addressees while lowering the likelihood that fully automated moderation systems will detect that meaning. This modulation can take multiple forms, vary in intensity, and often is layered.

For a deeper understanding of the problem, three general observations and two propositions are made. The two propositions will be tested experimentally through simulation. We begin with the following core assumptions for Algospeak.

Bystander Assumption. The malicious actor is aware of algorithmic content moderation systems and deliberately distorts their language using methods such as Algospeak to evade detection (Steen et al., 2023; Clark, 1996).

Goal Assumption. The objective of a toxic actor on a large social media platform is to maximize the reach of their message online; the more people who understand the intended meaning, the greater its

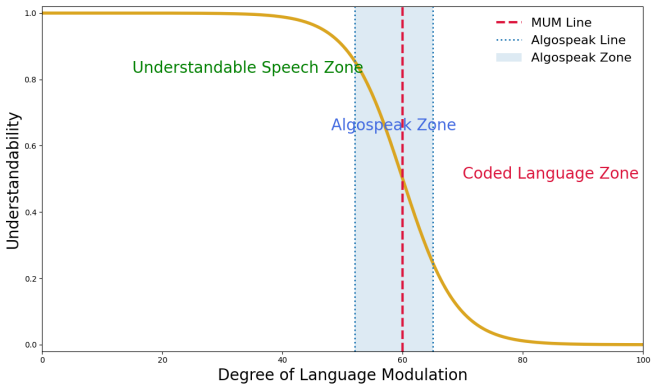


Figure 4: Overview of the relationship between modulation and understandability.

reach (Pancer et al., 2019).

Common Ground Assumption. Participants in an online conversation share a degree of common ground shaped by factors such as cultural identity and prior conversational context (Clark, 1991).

The two main propositions are: First, as linguistic modulation used in Algospeak increases, the proportion of participants able to understand the underlying meaning decreases. Depending on the degree of shared common ground, there exists a threshold of language modulation beyond which it becomes impossible for the majority of participants to grasp the intended meaning. Second, as the language becomes more modulated from its original form, it becomes increasingly difficult for trained classifiers to detect the underlying meaning and correctly classify the content of the message.

Based on these assumptions, a relationship between modulation and understandability can be proposed (see Figure 3), where understandability is defined as the percentage of participants who can recover the original meaning at a given modulation level. The four resulting zones range from typical language use (low modulation, high understandability) through opaque (low modulation, low understandability) to coded language (high modulation, low understandability) to Algospeak (high modulation, high understandability), the region most attractive to malicious actors.

Based on Wichmann and Hill (Wichmann and Hill, 2001), we propose that the relationship between understandability and modulation (and detectability/modulation) can be modeled using a sigmoid function, see Figure 4. The figure illustrates that as modulation increases, the percentage of the conversation participants able to comprehend the content decreases. It further shows that beyond a certain level of modulation, the language transitions from the zone of general understandability into the Algospeak zone, and eventually into coded language. The transition from Algospeak to coded language is what we define as the Majority Understandable Modulation (MUM) point, where the majority of the population within the context can no longer follow the majority of the conversation. This means that for a population the MUM point is based on the points that individuals fail to understand most of the modulated content. We call this second point the Individual Majority Understandable Modulation (IMUM) Point. It is important to note that both the slope and the modulation level of these points are highly dependent on the shared common context between the participants, and the figure presents example values solely for illustrative purposes. These constructs are proposed for humans and LLMs alike.

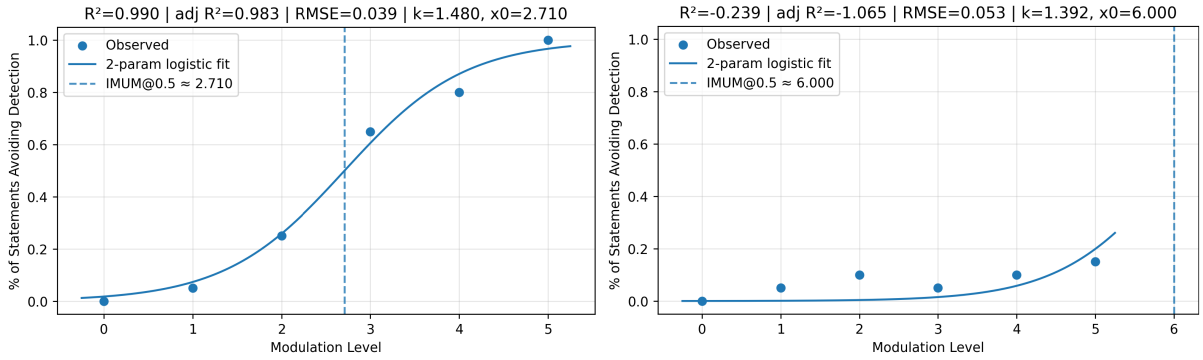
Formally, let $U_\kappa(d;\theta) \in [0,1]$ denote the fraction of a population $\mathcal{H}_\kappa$ sharing context κ that recovers the intended meaning of a text θ modulated at level d. For a comprehension threshold $\tau \in (0,1)$, the Individual Majority Understandable Modulation of θ is

$$\mathrm{IMUM}_\kappa^\tau(\theta) = \inf\{\, d \geq 0 : U_\kappa(d;\theta) < \tau \,\},$$

and the aggregate MUM over a distribution $\mathcal{D}$ of base texts is

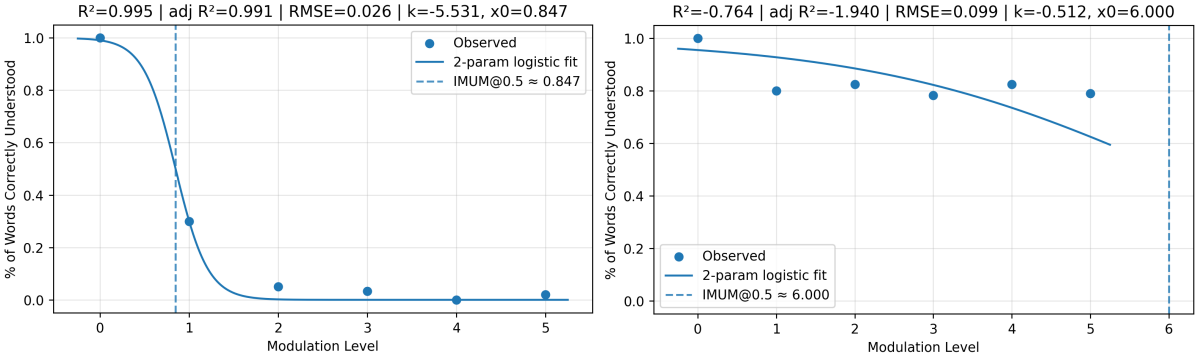
$$\mathrm{MUM}_\kappa^\tau = \mathbb{E}_{\theta\sim\mathcal{D}}\big[\mathrm{IMUM}_\kappa^\tau(\theta)\big].$$

We use $\tau = 0.5$ throughout. The full formalization, including semantic fidelity, detectability, and the evasion–comprehension trade-off, is given in Appendix A.



(a) Well fitted curve for the code word class, detection experiment.

(b) Ill fitted curve for the change in spelling class, detection experiment.



(c) Well fitted curve for the code word class, understanding experiment.

(d) Ill fitted curve for the change in spelling class, understanding experiment.

Figure 5: Example results for two of the seven Algospeak strategies for the model GPT-4o-mini.

4 Experiments

4.1 Dataset Creation

Because Algospeak is not detectable by current classifiers and no large-scale dataset exists, this study creates its own as a proof of concept. COVID-19 misinformation was chosen as the example domain due to its well-documented use of Algospeak, existing datasets, and the ease of determining ground truth, which reduces annotator bias. This focused validation allows us to test our theoretical framework under controlled conditions before extending to other domains in future work.

The dataset was created in five phases. First, we created 20 example sentences, each 10–15 words long and centered on common COVID-19 misinformation. Second, two researchers independently annotated all statements for whether they contradict established public-health consensus, reaching 100% agreement with the intended labels. Third, the sentences were validated to ensure the baseline model (GPT-4o (OpenAI, 2024)) correctly classified them as misinformation, based on majority agreement across three identical trials, with identical temperature (0) and prompt settings.

Next, the model identified feature importance by selecting the six words most responsible for each misinformation classification, determined by majority vote across three identical trials. A human researcher incrementally modified each sentence, starting with the most influential words and moving to the least, to create varying modulation levels. This process was applied to all 20 sentences across five modulation levels (∼10%, ∼20%, ∼30%, ∼40%, and ∼50%) and for all seven Algospeak strategies identified by (Fillies and Paschke, 2024) (including altered spellings (to unknown and known words), abbreviations, pictorial representations, paraphrasing, repurposed words, and phonetic substitutions), resulting in a total of 700 modulated messages. An example of the five different modulation levels can be seen in Figure 2.

As a qualitative check, we compared our synthetic items against Algospeak documented on social media, such as “seggs” for sex, “panoramic” for pandemic, or “leg booty” for LGBT (Steen et al., 2023; Klug et al., 2023; Delkic, 2022). The word-level substitutions our strategies produce are of the same kind, though ours are constructed sys-

tematically rather than emerging through use.

We acknowledge that this dataset construction has limitations; see Section 8. We explicitly situate the work as a proof of concept and planned expansion to larger, more diverse datasets.

4.2 Detection Experiment

For the detection of misinformation, a prompt-based setup was created. The prompt was designed in line with prior research that employs LLMs for misinformation detection (Pelrine et al., 2023; Hu et al., 2024). It instructed the LLM to determine whether the given statement contained misinformation, without adding judgments or other assumptions beyond that task. The model was asked to make its decision solely based on the provided statement and the general knowledge about COVID-19 contained within its training data.

This process was repeated three times for each of the five modulation levels, and a majority score was calculated under the same temperature (0) and prompt settings.

4.3 Understandability Experiment

While LLMs are not capable of truly understanding the meaning of words (Bender and Koller, 2020), we operationalized understandability by having the model reconstruct each modulated word, repeating that process at every modulation level. Evaluation was conducted by calculating a similarity score between the original words and the reconstructed words, using a chosen threshold of 95%. Statements exceeding this threshold were considered correctly identified.

4.4 Statistical Analysis

To characterize the relationship between distortion level and model performance, we employed two complementary approaches: We fitted a two-parameter logistic function to model detection and understanding rates across distortion levels, where k controls the steepness of decline and x0 represents the inflection point (the Majority Understandable Modulation threshold at 50%). Goodness of fit was assessed using R^2, adjusted R^2, and RMSE. This parameterization enables direct comparison of MUM thresholds across strategies and models.

To test whether performance declines monotonically with increasing distortion, independent of the specific sigmoid parameterization, we computed Spearman's rank correlation (ρ) (Spearman, 1904) between distortion level and performance for each series. Spearman's test is appropriate for our sample size (N=6 distortion levels) and provides robust inference without assuming a parametric functional form. Statistical significance was assessed at $\alpha = 0.05$. As each experiment comprises 49 (model, strategy) tests that are aggregated into a single significance count, we additionally control the false discovery rate across those 49 tests using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

5 Results

We begin with a detailed analysis of a representative model (GPT-4o-mini) to illustrate how to interpret the metrics and plots, then present aggregated results for all LLMs highlighting key contrasts.

5.1 Results for Model GPT-4o-mini

Detection Experiment For each statement and Algospeak strategy, we applied majority voting at every modulation level and aggregated results across the 20 examples. This revealed clear patterns: for the code-word strategy (Figure 5a), replacing the most important words up to modulation level 4 caused 80% of statements to evade misinformation classification. In contrast, for the unknown-spelling strategy (Figure 5b), over 80% are still correctly flagged at modulation level 5.

To test whether detection rates decline systematically with distortion, we fitted two-parameter logistic curves and computed Spearman rank correlations. Results are shown in Table 1. Six of seven strategies achieved strong sigmoid fits (adj. $R^2 > 0.90$), with code words ($R^2 = 0.990$) and new words ($R^2 = 0.996$) showing near-perfect fits. Only the unknown-word strategy fit poorly (adj. $R^2 = -1.065$), indicating highly variable detection even at low distortion levels.

The Spearman's test confirmed that all seven strategies exhibited significant monotonic decreases in detection ($p < 0.05$), with six achieving near-perfect correlation ($\rho >= 0.98$). This validates that increased distortion genuinely reduces detection across all strategies.

The IMUM for each strategy indicates how quickly it distorts text and enables direct comparison. For instance, on average, introducing just 2.7 code words caused 50% of statements to be misclassified, compared with 4.4 paraphrased words needed for the same effect.

Understandability Experiment Reconstruc-

Table 1: Detection Results - 2-parameter logistic curve and Spearman rank correlation, GPT-4o-mini

Series	k	IMUM	R^2	Adj. R^2	RMSE	Sig. Fit	ρ	p-val.
Unkn. word	1.3918	6.0000	-0.2388	-1.0646	0.0533	Poor	0.853	0.0416
New word	1.3657	4.1783	0.9959	0.9932	0.0173	Strong	0.986	0.0083
Abbreviations	1.2885	4.3299	0.9994	0.9990	0.0064	Strong	0.986	0.0139
Emoticons	1.4568	3.7796	0.9947	0.9912	0.0236	Strong	0.986	0.0111
Paraphrasing	0.8241	4.4341	0.9704	0.9506	0.0381	Strong	0.986	0.0083
Code	1.4799	2.7102	0.9896	0.9827	0.0388	Strong	1.000	0.0055
Phonetic	0.7904	5.1973	0.9896	0.9826	0.0163	Strong	0.986	0.0083

Table 2: Understanding Results - 2-parameter logistic curve and Spearman rank correlation, GPT-4o-mini

Series	k	IMUM	R^2	Adj. R^2	RMSE	Sig. Fit	ρ	p-val.
Unkn. word	-0.5119	6.0000	-0.7638	-1.9397	0.0990	Poor	-0.551	0.2667
New word	-0.3401	4.5889	0.6747	0.4579	0.0911	Moderate	-1.000	0.0055
Abbreviations	-0.4324	4.0659	0.8153	0.6922	0.0786	Moderate	-0.943	0.0139
Emoticons	-0.6704	1.4562	0.6536	0.4227	0.1618	Moderate	-0.943	0.0194
Paraphrasing	-4.4236	0.7645	0.8024	0.6706	0.1384	Moderate	-0.928	0.0139
Code	-5.5315	0.8475	0.9949	0.9915	0.0256	Strong	-0.943	0.0111
Phonetic	-0.5622	6.0000	-0.1454	-0.9090	0.0783	Poor	-0.928	0.0277

tions were repeated for each modulated word at every modulation level and averaged across 20 examples for all seven strategies. As Figure 5c shows, introducing a single unknown code word allowed correct reconstruction in only 40% of cases. In contrast, strategies preserving surface form, like spelling or phonetic changes (Figure 5d), enabled correct reconstruction for most terms at all levels.

Similar to the detection experiment, we computed Spearman correlations to test whether understanding declines with distortion. Results are shown in Table 2. Three strategies achieved strong or moderate fits (code words adj. $R^2 = 0.99$, paraphrasing 0.67, abbreviations 0.69), while unknown-word and phonetic strategies fit poorly (adj. $R^2 < 0$). Six of seven strategies showed significant monotonic decreases in understanding ($p < 0.05$), with three achieving perfect negative rank correlation ($\rho = -1.0$). Only the unknown-word strategy failed to reach significance ($\rho = -0.551$, $p = 0.2667$).

Code words required only 0.85 modulations to cross 50% understanding, while strategies preserving surface form (phonetic, spelling) never crossed this threshold within our tested range (x0 = 6.0).

5.2 Cross-model comparison

We compared seven state-of-the-art models, ranging from open- to closed-source and varying in the scale of their training parameters. The models considered were claude-sonnet-4-5 (Anthropic, 2025), gpt-4o-mini and gpt-4o (OpenAI, 2024), llama-3.1-8b-instant (Grattafiori et al., 2024), Ministral-8B-Instruct-2410 (Mistral AI, 2024), Qwen3-VL-32B-Instruct-FP8 (Qwen Team, 2025), and grok-4.1-fast (xAI, 2025).

Detection performance across models: Similar to Section 5.1, we tested whether increased distortion reduces detection using both sigmoid curve fitting (for MUM estimation) and Spearman's rank correlation (for statistical inference). Table 3 reports goodness-of-fit (adjusted R^2) for the sigmoid curves and Spearman correlation results across all models. Heatmaps Figure 6 in Appendix B reports adjusted R^2 values by strategy for the seven models[2] As before, most strategies fit the sigmoid curve well, whereas the change-in-spelling strategy producing an unknown word continues to show poor fit. Interestingly, under majority voting, most

[2]Adjusted R^2 can be negative (and unbounded below), strongly negative values indicate a fit worse than a constant baseline.

Table 3: Detection Results: Adjusted R^2 with majority fit estimation and Spearman Rank Correlation Significance by Strategy and Model

Strategy	Claude	GPT-4o-m	GPT-4o	Llama	Mistral	Qwen	Grok	Maj.	Sig. C.
Unkn. word	-4.83 ×	-2.03 ✓	-0.69 ✓	-0.18 ✓	0.70 ✓	0.89 ✓	-3.17 ×	P	5/7
New word	0.90 ✓	0.99 ✓	0.99 ✓	0.90 ✓	0.77 ✓	0.96 ✓	0.93 ✓	S	7/7
Abbreviation	0.92 ✓	0.98 ✓	0.95 ✓	0.96 ✓	-0.05 ✓	0.91 ✓	0.98 ✓	S	7/7
Emotion	-8.92 ×	0.98 ✓	0.96 ✓	0.95 ✓	0.88 ✓	0.99 ✓	0.94 ✓	S	6/7
Paraphrase	0.99 ✓	0.95 ✓	0.93 ✓	0.28 ✓	-4.88 ✓	0.95 ✓	0.97 ✓	S	7/7
Code	0.91 ✓	0.96 ✓	1.00 ✓	0.95 ✓	0.96 ✓	0.97 ✓	0.99 ✓	S	7/7
Phonetic	0.30 ✓	0.99 ✓	-4.54 ✓	-0.25 ✓	0.12 ×	0.89 ✓	0.35 ✓	P	6/7
Total	5/7	7/7	7/7	7/7	6/7	7/7	6/7		86%

Adjusted R^2 values shown with significance: ✓ = $p < 0.05$, × = $p \geq 0.05$; S = Strong, M = Moderate, P = Poor

Table 4: Understandability Results: Adjusted R^2 with majority fit estimation and Spearman Rank Correlation Significance by Strategy and Model

Strategy	Claude	GPT-4o-m	GPT-4o	Llama	Mistral	Qwen	Grok	Maj.	Sig. C.
Unkn. word	-1.03 ✓	-1.94 ×	-2.46 ×	0.34 ×	0.49 ×	0.47 ×	-2.39 ×	P	1/7
New word	0.90 ✓	0.46 ✓	0.67 ✓	0.81 ✓	0.87 ✓	0.87 ✓	0.35 ✓	S	7/7
Abbreviation	0.88 ✓	0.69 ✓	0.23 ✓	0.76 ✓	0.36 ×	0.38 ×	-0.34 ✓	P	5/7
Emotion	0.55 ✓	0.42 ✓	0.25 ✓	0.85 ×	0.74 ×	0.74 ×	0.51 ✓	M	4/7
Paraphrase	0.63 ✓	0.67 ✓	0.67 ×	0.98 ✓	1.00 ✓	1.00 ✓	0.54 ✓	M	6/7
Code	0.86 ✓	0.99 ✓	0.76 ✓	1.00 ✓	1.00 ×	1.00 ×	0.83 ✓	S	5/7
Phonetic	-0.57 ✓	-0.91 ✓	-1.96 ✓	0.11 ✓	0.38 ✓	0.39 ✓	-7.45 ×	P	6/7
Total	7/7	6/7	5/7	5/7	3/7	3/7	5/7		67%

Adjusted R^2 values shown with significance: ✓ = $p < 0.05$, × = $p \geq 0.05$; S = Strong, M = Moderate, P = Poor

models also did not seem to be strongly affected by the introduction of phonetic resemblance, an issue already indicated by the performance of the understanding task in Section 5.1.

Spearman's test revealed that 86% (42/49) of (model, modulation class) pairs exhibited statistically significant monotonic decreases in detection (p < 0.05), with 43% showing perfect or near-perfect rank correlation (|ρ| >= 0.98), and 90% (44/49) remain significant after Benjamini-Hochberg correction. This confirms that the relationship between modulation and detection represents genuine monotonic trends robust across models and strategies.

Overall, the results suggest that the choice of strategy matters more than the choice of model, although some models appear to work well with specific strategies.

Understandability across models: Similar to Section 5.1, we tested whether understanding declines with distortion using sigmoid fitting and Spearman's test. Table 4 reports adjusted R² values and Spearman significance results.

The heatmaps Figure 7 in Appendix B report adjusted R^2 values by strategy for the seven models, along with the majority measure of goodness-of-fit. As before, most strategies show a moderate to strong sigmoid fit, while the change-in-spelling and phonetic strategies perform poorly. Moreover, although Claude, GPT-4o-mini, and Llama exhibit strong or moderate fit for the abbreviation strategy, weaker fits for the remaining models yield an overall weak fit, highlighting substantial cross-model differences within a strategy.

Spearman's test revealed that 67% (33/49) of (model, distortion-class) pairs exhibited statistically significant monotonic decreases in understanding (p < 0.05), with 53% (26/49) remaining significant after Benjamini–Hochberg correction. This lower rate compared to detection (86%) suggests that understanding is more difficult to operationalize and varies more across models. Claude achieved 100% significant trends (7/7), while Mistral and Qwen managed only 43% (3/7), revealing substantial model-specific differences in semantic reconstruction capabilities.

Trade-off / IMUM & MUM comparison When combining the individual IMUM values for each strategy and model, we obtain an overall MUM per strategy that captures the modulation level beyond which most models fail to understand the majority of the text. Figure 1 shows both the model-specific IMUM values and the overall MUM for the codeword-introduction strategy.

6 Discussion

The experiments confirm both propositions across most strategies and models.

The results of the first experiment show that for most of the seven modulation strategies and LLMs, the proposition holds: as modulation increases, classifier detection performance decreases, and the proposed sigmoid function fits mostly well.

It is noteworthy that the strategies that don't follow this trend are common and straightforward. They consist of phonetic resemblance, without creating a new word, such as "Covid = Kovit", and the spelling alteration by replacing certain characters with others (for example, numbers). This outcome was expected, as phonetic resemblance is close in nature to spelling mistakes and spelling alterations rely heavily on well-established substitution rules such as "@ = a" or "3 = e." These conventions are widely used and recognized, therefore do not pose a challenge for LLM-based detection.

These inter-strategy differences have further implications. At the strategy level, our results suggest that shifting toxic communication from dynamic, unstructured Algospeak strategies (e.g., paraphrasing) to more formal Algospeak strategies closer to coded language (e.g., abbreviations) is beneficial for moderation, as detection capabilities increase while general comprehensibility decreases. Such a shift raises the time and cognitive costs for toxic actors, who must invest effort into developing a structured coded language and communicating its meaning to their intended audience. It also increases the initial effort required for external participants to join the conversation or be influenced by it.

The second experiment supported the propositions that increasing modulation reduces the ability to detect the original form of a word or sentence, and thus decreases overall comprehension. All strategies and LLMs showed a decline in understandability performance as modulation increased, some were affected more strongly than others. A moderate or strong fit of the proposed sigmoid curve was observed for most strategies and LLMs. Notably, some strategies disrupt detection while leaving reconstruction largely intact. This likely reflects a cue asymmetry: modulation scrambles the semantic patterns detectors rely on while preserving enough surface structure for reconstruction, possibly compounded by detectors being fine-tuned on cleaner data and thus sensitive to out-of-distribution noise.

Overall, these results suggest that the signals used for detecting information and for understanding or reconstructing meaning are somewhat disjoint, even though they do tend to move together. Some strategies, particularly emoticons, were reliably detected (6/7 models significant) but harder to reconstruct (4/7 models), suggesting models can flag "anomalous patterns" without necessarily recovering meaning. Conversely, unknown-word substitutions proved uniquely challenging for both detection and understanding, with only 5/7 and 1/7 models achieving significance respectively. The Spearman's rank correlation provides robust statistical evidence for our core propositions.

While human judgment typically requires comprehension, LLM-based detection can succeed without it. This insight may stem from factors such as fine-tuning strategies, characteristics of the training data, safety layers, or other architectural components, highlighting a novel distinction between detection and understanding in LLMs.

7 Conclusion and Future Work

The research establishes a formal definition of Algospeak and outlines its underlying dynamics, supported by a proof-of-concept LLM-based experimental setup using seven large language models. It introduces a framework for creating the first five-level modulated dataset and provides a dataset comprised of 700 distinct examples across seven Algospeak categories. The findings indicate that as Algospeak increases, detection ability decreases. Furthermore, higher levels of Algospeak reduce the understandability of the modulated terms. The underlying dynamics can be partially modeled using sigmoid curves.

This work lays the groundwork for future human-subject, cross-strategy, and cross-topic experiments, and its implications are substantial, offering guidance for both moderation practices and the assessment of moderation effectiveness.

8 Limitations

All results were generated through LLM-based settings. While human-subject experiments remain necessary to fully substantiate these findings, LLM-based agents are already active in this space, making the results directly relevant.

This study is a proof of concept focusing on COVID-19 misinformation in English. We focus on large platforms (TikTok, Facebook, YouTube,

X/Twitter) where toxic actors maximize audience size, though the framework applies to smaller platforms where reach assumptions may differ. Generalization to other domains such as hate speech or political persuasion remains untested. The dataset relies on 20 base sentences (700 total examples), which constrains semantic diversity and may introduce memorization effects. Modulation is also treated as one-dimensional. This is a deliberate simplification rather than an oversight, as collapsing modulation to a single axis is what makes sigmoid fitting and MUM estimation tractable, but it means lexical and syntactic distortion cannot be separated in the present results. Human-authored modulations may not capture organic Algospeak variations. All experiments were English-only; modulation strategies may vary across languages. The small number of modulation levels limits statistical precision.

Despite these limitations, this work establishes a formal foundation for studying Algospeak and provides evidence for core propositions. We view this as a first step toward broader research spanning domains, languages, and human populations.

9 Ethical Considerations

This work is inherently dual-use: systematically identifying which Algospeak modulation strategies most effectively degrade detector performance could, if misused, provide malicious actors with a blueprint for more resilient abuse and disinformation. To balance this risk against reproducibility, we distinguish between evaluation material and construction material.

All material needed to reproduce our evaluation is released publicly: the detection and reconstruction prompts, the inference configuration, the similarity metric and the calibration underlying the 95% threshold, the curve-fitting and statistical analysis code, and the aggregated per-strategy and per-model results behind every figure and table in this paper. These components are evaluation instruments and carry no meaningful dual-use value on their own. What we do not release is the feature-importance identification and incremental-modification procedure used to construct high-evasion variants, together with the 700 populated dataset items. That procedure is an operational construction recipe rather than an evaluation tool, and publishing it would directly operationalize evasive capability at scale. Both are made available after publication to researchers with a documented safety, governance, or harm-mitigation agenda and appropriate data-handling commitments.

We further work exclusively with researcher-authored, synthetic COVID-19 misinformation statements, and avoid publishing concrete high-evasion "recipes" or platform-specific implementation details. We therefore frame this work explicitly as red-teaming in support of defense: the analyses, thresholds, and tools are intended to guide the development and evaluation of protective systems and to inform governance processes, not to be used directly in production settings or for adversarial optimization.

References

Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. 2024. Next-generation phishing: How llm agents empower cyber attackers. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2558–2567.

Anthropic. 2025. System card: Claude Sonnet 4.5.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching@EMNLP 2014, Doha, Qatar, October 25, 2014*, pages 13–23. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Max Boholm, Gregor Rettenegger, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, and Asad Sayeed. 2025. Who leads? who follows? temporal dynamics of political dogwhistles in Swedish online communities. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 383–395, Vienna, Austria. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Herbert H Clark. 1991. Brennan (1991) grounding in communication.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Sophie Curtis. 2022. How tiktok is changing the way we speak: Phrases like "barbiecore", "quiet quitting" and "le dollar bean" that originated on the social media app have crossed over into the mainstream - so how many do you know?

Christine de Kock, Arij Riabi, Zeerak Talat, Michael Sejr Schlichtkrull, Pranava Madhyastha, and Eduard Hovy. 2026. IYKYK: Using language models to decode extremist cryptolects. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8393–8409, Rabat, Morocco. Association for Computational Linguistics.

Melina Delkic. 2022. Leg booty? panoramic? seggs? how tiktok is changing language.

Jan Fillies and Adrian Paschke. 2024. Simple llm based approach to counter algospeak. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 136–145.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6174–6181.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20).

Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political communication*, 37(2):256–280.

Daniel Klug, Ella Steen, and Kathryn Yurechko. 2023. How algorithm awareness impacts algospeak use on tiktok. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 234–237, New York, NY, USA. Association for Computing Machinery.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12493–12509, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

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)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Mistral AI. 2024. Un Ministral, des Ministraux.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Ethan Pancer, Vincent Chandler, Maxwell Poole, and Theodore J Noseworthy. 2019. How readability shapes social media engagement. *Journal of consumer psychology*, 29(2):262–270.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Kuleen Sasse, Carlos Alejandro Aguirre, Isabel Cachola, Sharon Levy, and Mark Dredze. 2025. Making FETCH! happen: Finding emergent dog whistles through common habitats. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5709, Vienna, Austria. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Ella Steen, Kathryn Yurechko, and Daniel Klug. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on tiktok. *Social Media + Society*, 9(3):20563051231194586.

Una Titz and Theresa Lehmann. 2023. Tiktok: Wie gartenzwerge die grenzen des sagbaren verschieben.

Andrea Tundis, Gaurav Mukherjee, and Max Mühlhäuser. 2020. Mixed-code text analysis for the detection of online hidden propaganda. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, ARES '20, New York, NY, USA. Association for Computing Machinery.

Iñaki Vélez de Mendizabal, Xabier Vidriales Mazorriaga, Iñigo Ezpeleta, and Urko Zurutuza. 2023. Deobfuscating leetspeak with deep learning to improve spam filtering.

Felix A Wichmann and N Jeremy Hill. 2001. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313.

Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday*.

xAI. 2025. Grok 4.1 fast and agent tools API.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145. Association for Computational Linguistics.

Kan Yuan, Haoran Lu, Xiaojing Liao, and XiaoFeng Wang. 2018. Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1027–1041, Baltimore, MD. USENIX Association.

Ying Zhou, Ben He, and Le Sun. 2024. Humanizing machine-generated content: evading ai-text detection through adversarial attack. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8427–8437.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 229–246. IEEE.

A Formalization of Majority Understandable Modulation

Let

- $\theta \in \mathcal{X}$ denote an original text with intended meaning $c \in \mathcal{C}$,
- $T_d : \mathcal{X} \to \Delta(\mathcal{X})$ be a stochastic modulation operator parameterized by $d \geq 0$, representing the level of modulation, where $\Delta(\mathcal{X})$ denotes the space of probability distributions over $\mathcal{X}$.

Semantic Fidelity. A modulated text $x = T_d(\theta)$ is *semantically valid* if it preserves the original meaning:

$$S(\theta, x) = \begin{cases} 1, & \text{if } x \text{ has the same meaning as } \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Understandability. For a population $\mathcal{H}_\kappa$ within context κ (shared background, community, etc.), and an original text θ with meaning c, define

$$U_\kappa(d; \theta) = \mathbb{E}_{x \sim T_d(\theta)} \Big[\mathbb{1}\{S(\theta, x) = 1\} \cdot \Pr_{h \sim \mathcal{H}_\kappa} \big[h \text{ correctly infers } c \text{ from } x \big] \Big],$$

where $U_\kappa(d; \theta) \in [0, 1]$ measures the expected fraction of participants who can understand a modulated message sampled from $T_d(\theta)$. The indicator function $\mathbb{1}\{S(\theta, x) = 1\}$ ensures that only semantically valid modulations contribute to understandability.

Based on empirical observations, $U_\kappa(d;\theta)$ can be modeled as an *inverted sigmoid* decay with increasing modulation:

$$U_\kappa(d;\theta) \approx \frac{1}{1+e^{\alpha_\kappa(d-\beta_\kappa)}},$$

where $\alpha_\kappa > 0$ controls the slope of decline and β_κ corresponds to the inflection point at which comprehension drops below 0.5. The parameter β_κ captures the influence of shared common ground: larger β_κ indicates higher tolerance for modulation within that context.

Detectability. For a moderation or detection model M,

$$D_M(d;\theta) = \mathbb{E}_{x\sim T_d(\theta)} \Pr\big[\, M(x) = 1 \,\big],$$

where $D_M(d;\theta) \in [0,1]$ denotes the probability that a modulated message sampled at distortion level d is flagged by the detector.

Individual Majority Understandable Modulation (IMUM). Given a comprehension threshold $\tau \in (0,1)$ (e.g., $\tau = 0.5$ for majority understanding), the *Individual Majority Understandable Modulation* for a specific text item θ is defined as

$$\boxed{\mathrm{IMUM}^\tau_\kappa(\theta) = \inf\{\, d \geq 0 : U_\kappa(d;\theta) < \tau \,\}.}$$

That is, the IMUM is the smallest modulation level at which fewer than a fraction τ of participants within context κ can correctly interpret the intended meaning of the specific text θ. The term "Individual" refers to this being computed for an individual text item.

Aggregate Majority Understandable Modulation (MUM). To obtain an aggregate measure across a distribution of messages, let $\mathcal{D}$ be a distribution over base texts $\theta \in \mathcal{X}$. The aggregate MUM is defined as:

$$\mathrm{MUM}^\tau_\kappa = \mathbb{E}_{\theta\sim\mathcal{D}}\big[\mathrm{IMUM}^\tau_\kappa(\theta)\big].$$

Trade-off Optimization. A malicious actor seeks to maximize understandability while minimizing algorithmic detectability. This trade-off can be formalized as the optimization problem

$$d^*(\theta) = \arg\max_{d\geq 0}\; U_\kappa(d;\theta)\,\big(1 - D_M(d;\theta)\big),$$

where $d^*(\theta)$ represents the optimal modulation level that balances communicative reach with moderation evasion for a given text θ. Typically, $d^*(\theta)$ lies just below $\mathrm{IMUM}^\tau_\kappa(\theta)$, corresponding to the region commonly identified as the *Algospeak zone*.

B Heatmaps Understandability and Detection

Figure 6 displays a heatmaps of the adjusted R^2 by strategy and model for the detection experiment. Figure 7 displays a heatmaps of the adjusted R^2 by strategy and model for the understandability experiment.

C IMUM's Per Strategy and Model

The tables display estimated threshold IMUM@0.5 for detection (Table 5) and understanding (Table 6), both tables are by strategy and model.

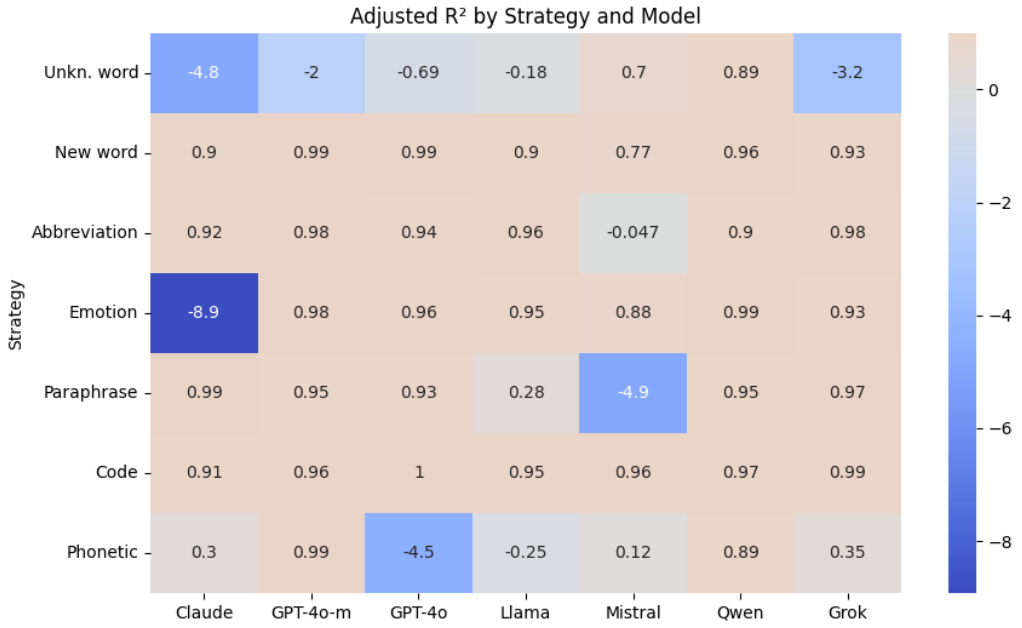


Figure 6: Heatmap detection adjusted R^2 by strategy and model.

Table 5: Estimated threshold x_0 (IMUM@0.5, detection) by strategy and model

Strategy	Claude	GPT-4o mini	GPT-4o	Llama	Mistral	Qwen	xAI
Unknown word	5.09988	5.10000	5.10000	5.10000	5.10000	4.74839	5.09993
New word	3.28074	4.17831	3.65689	4.88509	5.09823	3.96294	2.99138
Abbreviation	4.88880	4.3299	4.69842	4.57660	5.10000	3.98938	3.58646
Emotion	5.10000	3.81083	3.25907	3.09478	4.09506	3.35181	5.10000
Paraphrase	3.86082	4.43407	3.39559	5.10000	5.10000	4.34685	3.77906
Code	1.70700	3.69793	1.99989	2.67578	2.59317	2.23205	1.66515
Phonetic	5.10000	4.98789	5.10000	5.10000	5.10000	4.43758	5.10000

Table 6: Estimated threshold x_0 (IMUM@0.5, understanding) by strategy and model

Strategy	Claude	GPT-4o mini	GPT-4o	Llama	Mistral	Qwen	xAI
Unknown word	6.0000	6.0000	-0.9509	1.5788	0.7772	0.7772	-0.8621
New word	6.0000	4.5775	6.0000	0.8103	0.8254	0.8254	6.0000
Abbreviation	4.7725	4.0729	6.0000	1.5595	0.8480	0.8163	6.0000
Emotion	2.6879	1.5405	2.4005	0.7615	0.7615	0.7615	4.3287
Paraphrase	1.8680	0.7645	0.7555	0.7511	0.7863	0.7863	2.7983
Code	1.0062	0.8475	1.0555	0.7977	0.5005	0.5005	1.1699
Phonetic	6.0000	6.0000	6.0000	4.6037	3.0440	3.0115	-0.9952

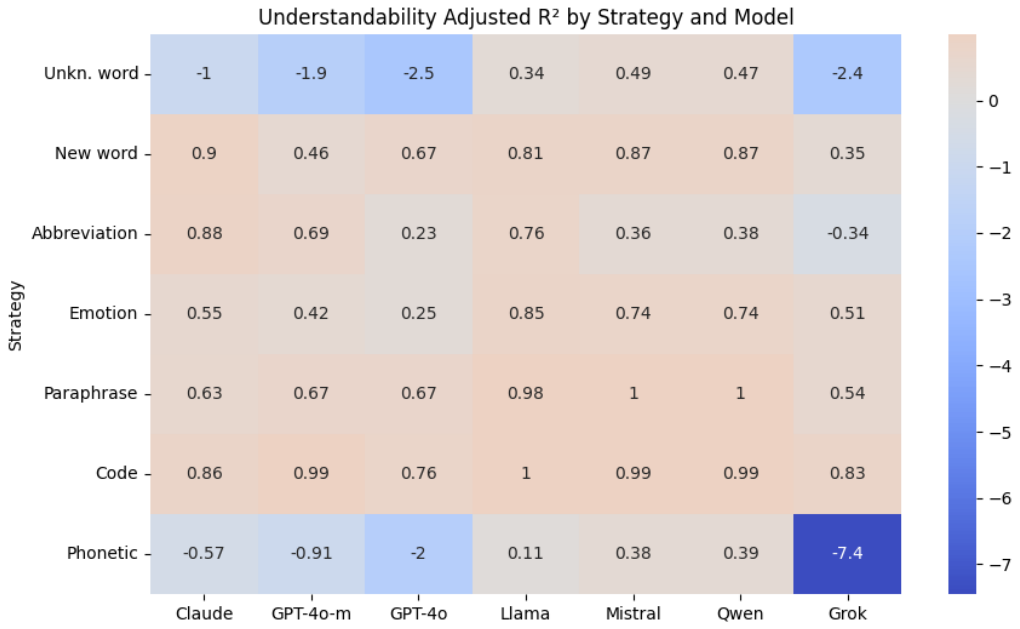


Figure 7: Heatmap understanding adjusted R^2 by strategy and model.