

# Simple LLM based Approach to Counter Algospeak

Jan Fillies<sup>1,2</sup> and Adrian Paschke<sup>1,2,3</sup>

<sup>1</sup>Institut für Angewandte Informatik, Leipzig, Germany

<sup>2</sup>Freie Universität Berlin, Berlin, Germany

<sup>3</sup>Fraunhofer-Institut für Offene Kommunikationssysteme, Berlin, Germany

fillies@infai.org, adrian.paschke@fokus.fraunhofer.de

## Abstract

With the use of algorithmic moderation on online communication platforms, an increase in adaptive language aiming to evade the automatic detection of problematic content has been observed. One form of this adapted language is known as "Algospeak" and is most commonly associated with large social media platforms, e.g., TikTok. It builds upon Leetspeak or online slang with its explicit intention to avoid machine readability. The machine-learning algorithms employed to automate the process of content moderation mostly rely on human-annotated datasets and supervised learning, often not adjusted for a wide variety of languages and changes in language. This work uses linguistic examples identified in research literature to introduce a taxonomy for Algospeak and shows that with the use of an LLM (GPT-4), 79.4% of the established terms can be corrected to their true form, or if needed, their underlying associated concepts. With an example sentence, 98.5% of terms are correctly identified. This research demonstrates that LLMs are the future in solving the current problem of moderation avoidance by Algospeak.

## 1 Introduction

**Content Warning: This report contains some examples of hateful content.**

Due to recent developments in legislation within the European Union<sup>1</sup>, the trend towards automatic content monitoring has been strengthened. Starting earlier and continuing up to today, all major social media platforms are implementing community guidelines and employing automatic content moderation (Morrow et al., 2022), at least partly relying on machine-learning-based identification approaches. Machine learning techniques are needed to handle the continuously increasing amount of content generated on all social media platforms.

<sup>1</sup><https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms>

In the past and present, the algorithms employed to detect problematic content, e.g., Hate Speech or content not deemed fitting for the social media platform based on their community guidelines, are often based on underlying datasets created for supervised classification of the content to be identified (Fortuna et al., 2022). Due to the nature of supervised classification, unseen data points are a challenge and can mislead the classification algorithm. The increasingly online-native user base of these platforms is aware of this phenomenon and is able to use it to their advantage (Steen et al., 2023). This phenomenon is called Algospeak. Algospeak refers to the concept of trying to communicate a sensitive or a potentially harmful message without it being detected by the algorithmic detection mechanism. Following Steen et al. (2023), Algospeak can contain "orthographic, lexical, and phonetic variations of standard language", it is a language specifically developed in reaction to content moderation on platforms. The field and definition of Algospeak are still very new in research on the algorithmic detection of online harms. But it has been shown that changing vocabulary and topic influences the quality of, for example, hate speech prediction (Florio et al., 2020). Understanding that established Algospeak terms only exists because they successfully circumvented the detection of online moderation systems makes it clear that it is a true problem in the constant strive for a safe online environment. There is a need to identify a strategic approach to handling Algospeak in the future.

This paper relies on examples of Algospeak provided by the research community (Steen et al., 2023). It categorizes them into underlying linguistic categories, displayed in the first known non-exclusive taxonomy. This taxonomy is utilized in a few-shot prompt engineering process with GPT-4 to transform the Algospeak terms into generally known and established words, phrases, or concepts. It demonstrates that with the straightforward ap-

plication of a large language model, this advanced Algospeak can be deciphered and, in the future, included in more standardized content detection models. Contributions: 1) The research establishes a non-exclusive taxonomy for Algospeak. 2) It demonstrates that Large Language Models (LLMs) can be utilized for deciphering Algospeak. 3) It indicates that performance can be improved with context.

## 2 Related Research

For the detection of hate speech, toxic speech, abusive language, or similar fields, the algorithmic approach to content detection has predominantly focused on supervised transformer-based architectures (Mozafari et al., 2020; Poletto et al., 2021; Fortuna et al., 2022; Plaza-del arco et al., 2023). The fine-tuning of transformer-based models, specifically BERT (Devlin et al., 2019), has shown clear improvement in performance compared to other approaches (Liu et al., 2019; Caselli et al., 2021; Mathew et al., 2021; Kirk et al., 2022; Fillies et al., 2023). Recently, the use of pre-trained large language models combined with prompting to detect hate speech has garnered attention (Schick et al., 2021; Chiu et al., 2022; Kim et al., 2023; Plaza-del arco et al., 2023; Muktadir, 2023).

Algospeak is a relatively new phenomenon first identified by public news outlets (Curtis, 2022; Delkic, 2022; Titz and Lehmann, 2023) and more formally by (Steen et al., 2023; Klug et al., 2023). Steen et al. (2023) distinguishes Algospeak from Textspeak, Leetspeak, and LOLspeak by identifying that the main intention of Algospeak is not to create a group identity or community but to circumvent online moderation. Steen et al. (2023) conducted 19 semi-structured interviews with content creators and collected 70 examples of Algospeak. Their goal was to analyze the usage of Algospeak and the relationship between the creators and TikTok's content moderation mechanisms. On a more formalized side, Cho and Kim (2021) created a taxonomy for noisy text, based on the user's intention.

Coded language in general, but more specifically Code-Mixing and Code-Switching, are well-studied linguistic phenomena (Bali et al., 2014). Especially in hate speech detection, coded language is well examined (Barman et al., 2014; Mathur et al., 2018; Bohra et al., 2018). The works focused mainly on mixed code for hate speech dealing with translation (Tundis et al., 2020). In the domain of

Leetspeak and propaganda detection, Tundis et al. (2020) designed a supervised network to classify texts using Leetspeak encoding directly. Similarly, but in the field of images, Vélez de Mendizabal et al. (2023) also used Neural Networks to decode Leetspeak. Singh et al. (2023) applied an unsupervised clustering-based approach for language standardization. This research differs from existing research by introducing a content-oriented taxonomy and testing the value of prompt-based unsupervised deciphering of Algospeak, which contains not only Leetspeak but also coded language itself.

## 3 Algospeak

Algospeak is defined in this research as stated by Steen et al. (2023), who formulate that "from a sociolinguistic perspective, Algospeak can resemble orthographic, lexical, and phonetic variations of standard language." It is further identified that Algospeak is a related linguistic phenomenon to Internet-based communication such as Textspeak, Chatspeak, or SMS-language (Drouin and Davis, 2009), Leetspeak (Perea et al., 2008), and LOLspeak (Fiorentini et al., 2013). However, it differs in intent, not primarily being used to establish identity or community membership but rather as a language specifically developed in reaction to content moderation on platforms.

## 4 Dataset

The used dataset consists of 70 words identified by Steen et al. (2023). The words were collected in June 2022 by qualitatively reviewing relevant social media news articles, and posts on Twitter, Reddit, and TikTok. The content was selected by identifying instances where a nonstandard word or emoji was used instead of a common word. It was then validated that the words were used as Algospeak by interviewing 19 globally distributed TikTok creators, aged 19–32, who had used them. One word was excluded in the research due to the lack of a clear reference word. The full list of words can be seen in Appendix C.

## 5 Taxonomy

To structure the prompting and provide insight for future research, the Algospeak instances were organized into a taxonomy comprising seven classes:

1. Change in spelling to unknown spelling ("abortion" to "@b0rt!0n")

2. Change in spelling to known spelling ("porn" to "corn")
3. Abbreviations ("SA" for "sexual assault")
4. Pictorial representations (use of emoticons)
5. Paraphrasing ("unalive" for "kill" or "suicide")
6. Repurposing of existing words ("Accountant" for "sex workers")
7. Phonetic resemblance ("not see" for "Nazi", "kermit sewer slide" for "commit suicide")

All classes involve changes to either the semantic meaning or the orthography of the original term, with some examples incorporating mixtures of both. Class 2, "Change in spelling to known spelling," is distinguished from Class 6, "Repurposing of existing words," by the fact that the original word remains identifiable within the obscured word. This differentiation also applies for Class 2 and Class 7.

## 6 Prompting for decoding Algospeak

The research utilizes ChatGPT and its underlying GPT-4 model (OpenAI et al., 2024) as the basis for its evaluation due to its very recent release, its size, and its capability of human-level performance on professional and academic benchmarks (Minaee et al., 2024). Further research also suggests its comparatively good performance in hate speech detection (Matter et al., 2024).

To work with LLMs, Ekin (2023) identifies different methods of prompt engineering. He broadly distinguishes between basic and advanced approaches. Basic methods use control codes, templates, iterative testing, and refinement, while advanced strategies utilize temperature and token control, prompt chaining, and adapting prompts. This study focuses on the basic techniques. Standard templates were designed and then iteratively tested and refined until the results were correct in over 50% of cases within each class. Each class of the previously defined taxonomy 5 is matched with a corresponding prompt. The underlying template constructed to build all the prompts consists of three parts: first, the desired answer length is defined; second, information is provided about the kind of change expected; and lastly, the obscured word or words are provided. For example, the prompt for Class 7, Phonetic resemblance, is: "Answer using just one or at most two words: considering phonetic resemblance, what word or group of words is meant by 'not see'?" All designed prompts can be found in the Appendix A.

Algospeak	Word	GPT-4
@b0rt!0n	abortion	Abortion
auti\$m	autism	Autism
blk	Black	Black
bl00d	blood	Blood
b00bs	breasts	Boobs

Table 1: Examples for taxonomy class 1, "Change in spelling to unknown spelling," Algospeak variations as identified by Steen et al. (2023).








Algospeak	Word	GPT-4
 in bio	link in bio	Kinky
	Black people	N/A
	female genitals	Pussy
	pornstar	Pornstar
	male genitals	Penis
	fuck	Frog
	White people	High five

Table 2: Examples for taxonomy class 4 "Pictorial", Algospeak variations as identified by Steen et al. (2023).

In the second step, it was tested whether GPT-4's prediction quality would increase by providing more context. In this case, for each wrongly identified Algospeak instance, the prompt was adjusted to include the section: "The words are used in a sentence like:" followed by an example of the obscured word in use. All example sentences can be found in Appendix B.

## 7 Results

All 69 Algospeak terms from the reference literature, their meanings, and the predictions of GPT-4 are displayed across 7 tables. A selection of examples for classes 1 (Change in spelling to unknown spelling, 1), 4 (Pictorial, partly, 2), and 7 (Phonetic resemblance, 3) are included in the paper. All complete tables for all classes can be found in the Appendix C. An overview of class wise accuracy with and without context can be seen in Table 4. This research manually checked the correctness of the predictions, in a group of two, reaching mutual agreed annotations. A prediction is considered correct if the exact word or a reasonably fitting synonym was provided (e.g., "male genitals" for "Penis"). Regarding Table 1, it is observed that GPT-4 had no problems predicting changes in spelling to unknown spelling, with all 17 terms

Algospeak	Word	GPT-4
blink in lio	link in bio	Blindly
cue anon	QAnon	Qanon
kermit sewer		
slide	commit sui.	Commit sui.
le dollar bean	lesbian	Lebanese pou.
leg booty com.	LGBT com.	LGBTQ+ com.

Table 3: Examples for taxonomy class 7 "Phonetic", Algospeak based on [Steen et al. \(2023\)](#). (com. is short for community and sui. for suicide and pou. for pund).

correctly identified, for the full table see 5. Class 2, as seen in Table 6 in Appendix C, proved more challenging, with the meanings of words obscured by misspelling them into different existing terms; here, 3 of 5 terms were correctly identified. All five abbreviations, as seen in Table 7 in Appendix C, were correctly identified by GPT-4. The most issues arose with Class 4 (see Table 2 or full Table 8 in Appendix C), where only 13 of the 21 emoticons were correctly identified. For the first time, some predictions were close in meaning, such as "ejaculation" and "orgasm," or did not follow the prompt by not searching for the hidden semantic meaning, simply stating 🐸 as "frog." One emoticon (👤, annotated as "black people") had to be omitted because GPT-4 did not allow the answer to be presented, indicating correct identification but non-compliance with community guidelines. All three words in Class 5 (Paraphrase), see 9 in Appendix C, were correctly identified. Five out of 7 words from Class 6 (Table 10, in the Appendix C) concerning the repurposing of existing words were correctly identified. Additionally, 9 out of 11 Phonetic resemblance words from Class 7 were accurately deciphered, as seen in Table 3, for the full table see Table 11 in Appendix C. In Table 12, it is shown that 13 of the 14 previously incorrectly identified words were correctly attributed to their right meaning or word when given an example sentence.

## 8 Discussion

As demonstrated by the results, GPT-4 is capable of identifying the true meaning or reference word of 79.4% of all examples without context. This achievement is noteworthy, considering that deciphering these terms often requires in-depth domain knowledge. The model, however, appears to still struggle with emoticons, though its ability to discern multilevel meanings improves when context

Class	Acc.	Acc. Con.
1. Change in spell. to unknown spell.	1.0	-
2. Change in spell. to known spell.	0.6	1.0
3. Pictorial representations	0.6	1.0
4. Abbreviations	1.0	-
5. Paraphrasing	1.0	-
6. Repurposing of existing words	0.714	0.856
7. Phonetic resemblance	0.818	1.0

Table 4: Measured Accuracy for each class of the taxonomy, with context and without context. (spell. short for spelling, Acc. short for Accuracy, Con. short for Context)

is provided. Given the closed nature of GPT-4, we can only speculate about the source of its domain knowledge. It is plausible that the terms originating in 2022, along with their associated media coverage, contribute to GPT-4's familiarity with them. This might suggest that the model's understanding of more recent linguistic developments could be less robust. This hypothesis may apply to Classes 3, 4, and 6 but possibly not to those related to orthography or general language understanding, such as Classes 1, 2, and 5. The observation that context significantly enhances predictions aligns with expectations, given LLMs operate partly on word-level predictions. The evaluation of the performance is based on human assessment, which is prone to error. For example, the only misclassification with context by the model, "swimmer" for "sheep," could arguably be considered accurate, as "sheep" is a known euphemism for non-vaccinated individuals within the anti-vaccine movement.

## 9 Ethical Considerations

This research adheres to the ACM Code of Ethics, upholding general ethical principles, applying professional responsibility, and promoting leadership principles as advocated by the ACM. The research serves the interests of society, with the public good being the main consideration. The limitations associated with this work are discussed in Section 11. The algorithmic detection of abusive content is essential for maintaining a harm-free environment. Algospeak often serves to circumvent censorship by platforms relying on detection methods that lack

context sensitivity or exhibit bias. Therefore, this research advocates not only for the use of LLMs to decode coded language but also for enhancing content moderation capabilities through context-aware approaches and the use of precise, decoded datasets. These context-aware approaches will help online communities, which momentarily resort to Algospeak for legitimate reasons (e.g., during online sex education), to express themselves freely in the future.

## 10 Conclusion and Future Work

This research aims to contribute to the field of abusive harm detection by identifying a strategy to handle the prevalent avoidance tactic of Algospeak on social media. A taxonomy for classifying Algospeak was developed and served as the basis for employing basic prompt engineering techniques. Utilizing these tailored prompts, GPT-4’s ability to decode Algospeak was assessed. The findings conclusively show that LLM GPT-4 can decipher Algospeak with high accuracy (79.4%) without context, and almost flawlessly (98.5%) when a single example sentence is provided. The research underscores the value of LLMs in supporting future content moderation efforts, not only in straightforward classification tasks but also in clearing cleaned datasets by deciphering coded language. Future studies should explore the capabilities of various LLMs, incorporate different datasets, use advanced prompting techniques, and assess how decoded datasets impact trained classifiers.

## 11 Limitations

This preliminary study was designed as an initial proof of concept. Future work should expand the scope to include a broader range of Large Language Models (LLMs) or word-level predictors, ideally leveraging open-source options. It is crucial to assess how these models handle less known Algospeak or more recent linguistic developments. Additionally, the impact of varying context levels on model performance warrants further investigation, along with the practical influence of this approach in detecting harmful content.

## Acknowledgements

This research was supported by the Citizens, Equality, Rights and Values (CERV) Programme under Grand Agreement No. 101049342.

We thank Theresa Lehmann at the Amadeu Antonio Foundation for her thoughts and bringing this field of research to our attention.

## References

- 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.
- 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.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. [HateBERT: Retraining BERT for abusive language detection in English](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.
- Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2022. [Detecting hate speech with gpt-3](#).
- Won Ik Cho and Soomin Kim. 2021. [Google-trickers, yaminjeongeum, and leetspeak: An empirical taxonomy for intentionally noisy user-generated text](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 56–61, Online. Association for Computational Linguistics.
- 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?](#)
- Melina Delkic. 2022. [Leg booty? panoramic? seggs? how tiktok is changing language](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

- 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michelle Drouin and Claire Davis. 2009. [R u txtng? is the use of text speak hurting your literacy?](#) *Journal of Literacy Research*, 41(1):46–67.
- Sabit Ekin. 2023. [Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices.](#)
- Jan Fillies, Michael Hoffmann, and Aadrian Paschke. 2023. [Multilingual hate speech detection: Comparison of transfer learning methods to classify german, italian, and spanish posts.](#) In *2023 IEEE International Conference on Big Data (BigData)*, pages 5503–5511, Los Alamitos, CA, USA. IEEE Computer Society.
- Ilaria Fiorentini et al. 2013. [Zomg! dis iz a new language”: The case of lolpeak.](#) *Selected Papers from Sociolinguistics Summer School*, 4:90–108.
- Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. [Time of your hate: The challenge of time in hate speech detection on social media.](#) *Applied Sciences (Switzerland)*, 10.
- Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. [Directions for NLP practices applied to online hate speech detection.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. [ConPrompt: Pre-training a language model with machine-generated data for implicit hate speech detection.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980, Singapore. Association for Computational Linguistics.
- Hannah Kirk, Bertie Vidgen, and Scott Hale. 2022. [Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning.](#) In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- 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.
- Ping Liu, Wen Li, and Liang Zou. 2019. [NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers.](#) In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. [Hatexplain: A benchmark dataset for explainable hate speech detection.](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.
- 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.
- Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. [Close to human-level agreement: Tracing journeys of violent speech in incel posts with gpt-4-enhanced annotations.](#)
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. [Large language models: A survey.](#)
- Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopeck, and John P Wihbey. 2022. [The emerging science of content labeling: Contextualizing social media content moderation.](#) *Journal of the Association for Information Science and Technology*, 73(10):1365–1386.
- Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. [Hate speech detection and racial bias mitigation in social media based on bert model.](#) *PLOS ONE*, 15(8):1–26.
- Golam Md Muktedir. 2023. [A brief history of prompt: Leveraging language models. \(through advanced prompting\).](#)
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, and more. 2024. [Gpt-4 technical report.](#)
- Manuel Perea, Jon Andoni Duñabeitia, and Manuel Carreiras. 2008. [R34d1ng w0rd5 w1th numb3r5.](#) *Journal of Experimental Psychology: Human Perception and Performance*, 34(1):237.
- Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. [Respectful or toxic? using zero-shot learning with language models to detect hate speech.](#) In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. [Resources and benchmark corpora for hate speech detection: a systematic review.](#) *Language Resources and Evaluation*, 55:477–523.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. [Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp](#). *Transactions of the Association for Computational Linguistics*, pages 1408–1424.

Rajat Singh, Nurendra Choudhary, and Manish Shrivastava. 2023. [Automatic Normalization of Word Variations in Code-Mixed Social Media Text](#), page 371–381. Springer Nature Switzerland.

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. [De-obfuscating leetspeak with deep learning to improve spam filtering](#).

















## A Prompts for all seven classes of the taxonomy

1. Change in spelling to unknown spelling: "Answer using just one word: considering leet speak and phonetic resemblance what word is meant by: 'XX'"
2. Change in spelling to known spelling: "Answer using just one word: If this word is not used in the fitting context, what is the word meant by using it: XX"
3. Abbreviations (SA = sexual assault): "Answer using just one or at maximum three words: If these abbreviations are used on social media, what concepts are referred to: 'XX'"
4. Pictorial (emoticons): "Answer using just one word: If this emoticon is not used in normal context, what is the hidden meaning on social media: 'XX'"
5. Paraphrase: "Answer using just one word: 'XX' more precise means what?"
6. Repurposing of existing words: "Answer using just one word: If this word is not used in

the fitting context, what is the meaning of the word: 'XX'"

7. Phonetic resemblance: "Answer using just one or at maximum two words: consider phonetic resemblance what word or group of words is meant by: 'XX'"

## B Example Sentences

1. accountant: "I am an accountant on an adult website"
2. blink in lio: "You can find my blink in lio"
3. cornucopia: "That stament towards gay people is cornucopia"
4. Frog: "Frog this, I dont like it"
5. swimmers: "These swimmer put what ever the government tells them in their body"
6.  in bio: "You can find my
7.  in bio"
8. : " this, I dont like it"
9. : "All these  need to be educated"
10. : "You can put this straight on 
11. : "Horrible things happen in  at the moment. Peace please."
12. : "All my  can rap."
13. : "Lets have  tonight."
14. : "She has great .

## C Predictions on Classes 1,2,3,4,5,6,7

The Tables 5,6,7,8,9,10,11 display the examples for all seven classes of the taxonomy. For each word, the GPT-4 predictions are also displayed. The Algospeak examples and their annotation from Steen et al. (2023).

Algospeak	Word	GPT-4
@b0rt!0n	abortion	Abortion
auti\$m	autism	Autism
blk	Black	Black
bl00d	blood	Blood
b00bs	breasts	Boobs
depressi0n	depression	Depression
h0rny	horny	Horny
k1ll	kill	Kill
le\$bean	lesbian	Lesbian
le\$bian	lesbian	Lesbian
seggs	sex	Sex
sh!t	shit	Shit
skripper	stripper	Stripper
str8	straight	Straight
\$tripper	stripper	Stripper
tism	autism	Autism
wh1te	White	White

Table 5: The table contains the class 1 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed.

Algospeak	Word	GPT-4
corn	porn	Porn
cornucopia	homophobia	Horniness
fork	fuck	Fuck
Frog	fuck	Fag
grape	rape	Rape

Table 6: The table contains the class 2 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed.

Algospeak	Word	GPT-4
ED	eating disorder	Eating Disorder
SA	sexual assault	Sexual Assault
SH	self-harm	Self-Harm
SSA	same-sex att.	Same-Sex Att.
SW	sex worker	Sex Work

Table 7: The table contains the class 3 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed (att. stands for attraction).

Algospeak	Word	GPT-4
 in bio	link in bio	Kinky
	Black people	N/A
	female genitals	Pussy
	porn	Porn
	pornstar	Pornstar
	male genitals	Penis
	fuck	Frog
	White people	High five
	ejaculation	Nuts/Orgasm
	PornHub	Basketball
	shit	Shit
	Ukraine	Sunshine
	breasts	Virginity
	White people	White
	sex	Spicy/Hot
	ejaculation	Wet/Sexual ar.
	butt	Butt
	butt	Ass
	Black people	Black Power
	cocaine	Coke
	N-word	Ninja

Table 8: The table contains the class 4 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed (ar. stands for arousal).

Algospeak	Word	GPT-4
clock app	TikTok	TikTok
opposite of love	hate	Hate
unalive	dead, kill, suicide	Dead

Table 9: The table contains the class 5 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed.



## D Predictions with Context

The Table 12 shows the examples of Algospeak from Steen et al. (2023), the context sentence given to the model and The GPT-4 predictions.

Algospeak	Word	GPT-4
accountant	sex worker	Drug dealer
Backstreet Boys		
reunion tour	COVID-19 pan.	COVID-19
panda express	COVID-19 pan.	COVID-19
Panini	COVID-19 pan.	COVID-19
panorama	COVID-19 pan.	COVID-19
swimmers	vaccinated peo.	Sperm
the vid	COVID-19	COVID-19

Table 10: The table contains the class 6 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed (pan. stands for pandemic; peo. stands for people).

Algospeak	Word	GPT-4
blink in lio	link in bio	Blindly
cue anon	QAnon	Qanon
kermit sewer		
slide	commit sui.	Commit sui.
le dollar bean	lesbian	Lebanese pou.
leg booty com.	LGBT com.	LGBTQ+ com.
nip nops	nipples	Nipples
not see	Nazi	Nazi
Ouid	weed	Weed
sewer slide	suicide	Suicide
shmex	sex	Sex
yt	White	White

Table 11: The table contains the class 7 Algospeak examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed (com. is short for community and sui. for suicide and pou. stands for pund).


















Algospeak	Word	Context	GPT-4
accountant	sex worker	I am an accountant on an adult website	Camgirl
blink in lio	link in bio	You can find my bink in lio	Link in bio
cornucopia	homophobia	That statement towards gay people is cornucopia	Homophobia
Frog	fuck	Frog this, I don't like it.	Fuck
le dollar bean	lesbian	She is a le dollar bean.	Lesbian
swimmers	vaccinated people	These swimmer put what ever the government tells them in their body.	Sheep
 in bio	link in bio	You can find my  in bio	Link
	fuck	 this, I don't like it.	Fuck
	White people	All these  need to be educated.	Whites
	PornHub	You can put this straight on  	Pornhub
	Ukraine	Horrible things happen in  at the moment. Peace please.	Ukraine
	breasts	She has great  .	Breasts
	sex	Lets have  tonight.	Sex
	N-word	All my  can rap.	Blacks

Table 12: The table displays the examples of Algospeak that could not be identified examples and their annotation from Steen et al. (2023). The GPT-4 predictions are also displayed with the given context statements..