Beyond Surface Forms: A Comprehensive, Mechanism-Oriented Taxonomy of Indirect Linguistic Encoding for LLM-Based Coded Language Detection

Hamid Reza Firoozfar
The University of Utah
hamid.firoozfar@utah.edu

Mohammadsadegh Abolhasani
The University of Utah
sadegh.abolhasani@utah.edu

Reza Mousavi
The University of Virginia
mousavi@virginia.edu

Paul Jen-Hwa Hu
The University of Utah
paul.hu@eccles.utah.edu

Abstract

To avoid moderation and surveillance on social media, some users routinely invent indirect linguistic expressions (ILE) that camouflage sensitive meanings. Such expressions surface as algospeak, euphemisms, and adversarial obfuscation, depending on intent and context, and they involve recurring encoding mechanisms. We propose a comprehensive, mechanism-oriented taxonomy of ILE that abstracts away from communicative goals and instead categorizes the underlying operations through which meaning is encoded and recovered. We evaluate the taxonomy by incorporating it into LLM prompts and comparing it with four existing taxonomies and a no-taxonomy baseline, using 2,000 manually annotated TikTok and Bluesky posts. The proposed taxonomy attains the strongest document- and span-level performance across the three LLMs, achieving an improvement of 4.7% in accuracy and 5.4% in F1 over the best-performing benchmark. The empirical results reveal the importance of a comprehensive, mechanism-oriented taxonomy as a stable scaffold for detecting emerging coded language and a useful input to content moderation. Code and the publicly releasable portion of the data are available at: https://github.com/hmdfiroozfar/mechanism-oriented-ile-taxonomy. TikTok-derived text is not redistributed because of data-sharing restrictions.

Disclaimer: *This paper contains content that may be profane, vulgar, or offensive.*

1 Introduction

Social media users routinely camouflage sensitive meanings to evade algorithmic moderation, shadow banning, and demonetization (Klug et al., 2023; Steen et al., 2023; Morrow et al., 2022). Existing forms span a wide spectrum of strategies that we collectively refer to as *indirect linguistic encoding* (ILE). ILE is a moving target for content moderation, as reflected by lexicons becoming stale within weeks (Hu et al., 2024b; Zhu et al., 2021), and supervised detectors trained on older ILEs failing on newer ones.

Large language models (LLMs) offer a promising abstraction for ILE detection because they can model context and indirect meaning without explicit lexicons (Fillies and Paschke, 2024). However, their performance is highly prompt-sensitive (Zhuo et al., 2024). Recent studies report that structured, taxonomy-grounded prompts substantially improve classification under label sparsity and concept drift (Xia et al., 2025; Zhang et al., 2025; Gao et al., 2023; Kumar et al., 2023). Hence, the quality of taxonomy-guided detection significantly depends on the taxonomy itself.

Previous research has established useful taxonomies for ILE, but typically with narrow coverage, mixed levels of abstraction, or weak separation of encoding and decoding pathways (Calhoun and Fawcett, 2023; Fillies and Paschke, 2024; Leal-Arenas and Corizzo, 2024; Zhang et al., 2014). In practice, users draw on a wide range of strategies and sometimes combine several within a single expression, as illustrated in Figure 1. Approaches that center on surface-form perturbation alone (*coc@!ne* for *cocaine*) only capture part of the phenomenon, underscoring the need for a more comprehensive taxonomy with broader strategy coverage to support reliable LLM-based detection.

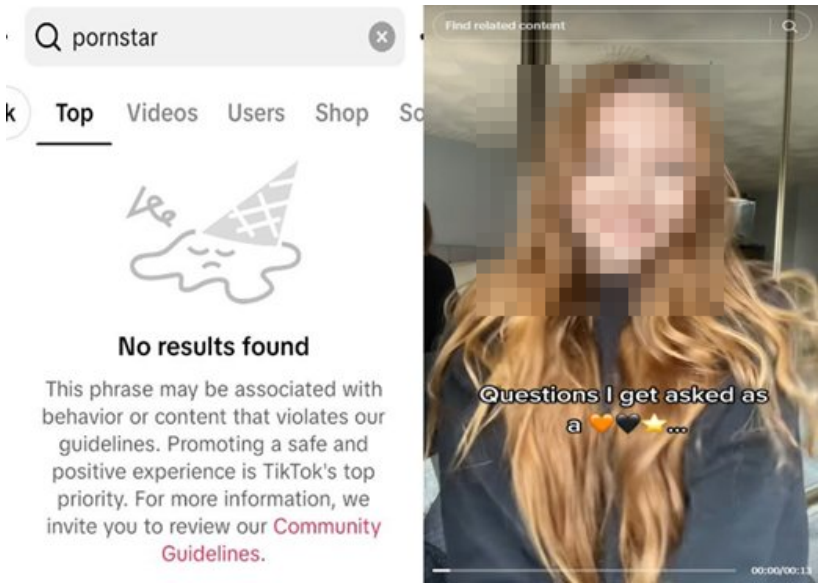

Figure 1: An ILE example from TikTok. Left: the flagged term "pornstar" returns no results. Right: a creator encodes *porn star* with three emojis, composing three mechanisms (iconic resemblance, attribute-based alias, and name-based pictorial mapping).

This study seeks to answer two important questions: how to build a taxonomy of ILE with broader strategy coverage and whether its use indeed improves LLM-based detection. Our contributions are threefold. First, we propose a *mechanism-oriented taxonomy* that abstracts away from communicative intent and organizes encodings by the pathway through which meaning is concealed and recovered, introducing new top-level classes and sub-classes that prior schemes do not cover. Second, we construct a manually annotated dataset of 2,000 recent social media posts (600 Bluesky, 1,400 TikTok), labeled by two trained annotators for (i) binary presence of ILE at the document level, (ii) the spans of all ILE expressions within each post, and (iii) the mechanism class and sub-mechanism of each span in accordance with the proposed taxonomy. We release the publicly shareable portion of the dataset and accompanying code; TikTok-derived post text is withheld from the public release because of our TikTok data-sharing agreement. Third, we compare the proposed taxonomy with four existing taxonomies and a no-taxonomy baseline across three LLMs at both document and span levels, thereby producing empirical evidence suggesting that the comprehensiveness of a taxonomy determines whether it enhances LLM-based ILE detection rather than its mere use. The overall findings show that taxonomy-guided prompting can underperform a no-taxonomy prompt when coverage is partial, while a sufficiently comprehensive taxonomy yields the strongest performance across different LLMs.

2 Related Work

ILE detection. Supervised methods dominate ILE detection (Hu et al., 2024a; Keh et al., 2022; Li et al., 2025; Pei et al., 2019; Wang et al., 2022; Lee et al., 2022), but they generalize poorly across topics and over time, because lexical substitutions evolve quickly under the *euphemism treadmill* (Pinker, 2003; Hu et al., 2024b; Zhu et al., 2021). Semi-supervised approaches lower annotation requirements (Felt and Riloff, 2020; Zhang et al., 2024), yet remain sensitive to seed selection and error propagation. Unsupervised and self-supervised methods instead discover coded terms directly from text (Ke et al., 2022; Magu and Luo, 2018; Sasse et al., 2025; Takuro et al., 2020; Zhu and Bhat, 2021; Zhu et al., 2021), but typically rely on domain constraints or seed keywords. Recent work turns to LLM-based detection in zero- or few-shot mode, which sidesteps these restrictions (Fillies and Paschke, 2024; Markov et al., 2023). Its accuracy, however, greatly relies on prompt design and conceptual scaffolding that the prompt encodes (Zhuo et al., 2024; Gao et al., 2023).

Existing ILE taxonomies differ in organizing principles, methodological grounding, and scope. Early work is built upon context-specific empirical analysis of censorship avoidance. For example, Zhang et al. (2014) analyze morph encoding in Chinese social media and identify distinct categories that include phonetic substitution, character decomposition, translation, semantic reinterpretation, and historical reference. Despite its empirical grounding, this taxonomy is tightly coupled with the characteristics of Chinese writing and entity-level encoding, which constrains generalization across writing systems and recovery mechanisms. Subsequent taxonomies are constructed to support normalization and detection, with a common interest in observable surface perturbations. For example, Renwick and Barbosa (2021) catalog character- and string-level manipulations (such as leetspeak, punctuation insertion, glyph substitution, and phonetic variants), without modeling semantic, referential, or convention-based strategies. More recent work draws on qualitative analysis of self-censorship. Calhoun and Fawcett (2023) organize strategies across orthography, morphology, phonology, prosody, and reanalysis, focusing on innovative linguistic forms without considering indirect strategies that require shared cultural reference or conventionalized shorthand beyond the scheme. A parallel line of algospeak-centric work (Leal-Arenas and Corizzo, 2024; Fillies and Paschke, 2024) groups surface strategies (e.g., abbreviations, phonetic spellings, symbol substitution, paraphrasing, repurposing) and operates at heterogeneous levels of abstraction and also omits referential and conventional-symbolic strategies. Across these efforts, the key limitation is that taxonomy construction has been largely study-specific and ad hoc, with categories drawn from small, non-representative samples or a fixed, predefined set of surface transformations, and with no shared, replicable procedure for iteratively refining categories and validating their coverage (Calhoun and Fawcett, 2023; Cho and Kim, 2021; Fillies and Paschke, 2024; Leal-Arenas and Corizzo, 2024; Renwick and Barbosa, 2021). As a consequence, the resulting schemes capture only partial subsets of ILE, and which mechanisms a taxonomy covers is determined more by its source data than by a principled account of the encoding and decoding pathways themselves. We address this gap with a systematic taxonomy-development procedure, described next.

3 Proposed Taxonomy

3.1 Development Method

We follow the iterative taxonomy development advocated by Nickerson et al. (2013), a well-established procedure for taxonomy construction. Figure 2 summarizes the full development process. The method alternates between two complementary paths until a set of ending conditions is met: an *empirical-to-conceptual* path that abstracts categories from observed objects, and a *conceptual-to-empirical* path that derives categories from theory and then tests them on data.

Meta-characteristic. In line with Nickerson et al. (2013), we begin by defining a *meta-characteristic* that serves as a logical constraint on every category in the taxonomy:

> *the formal encoding and decoding pathway by which a sensitive meaning is hidden in surface form and recovered by a human interpreter.*

This meta-characteristic deliberately excludes communicative intent (e.g., moderation evasion, politeness, in-group signaling) and content topic (e.g., drugs, self-harm, sexuality), thereby keeping the taxonomy focused on *how* meaning is concealed rather than *why* or *about what*.

Ending conditions. We adopt the objective conditions of Nickerson et al. (2013), where every object in the dataset corresponds to at least one category, there is no empty category, and no two categories are collapsed into one, together with their subjective conditions (the taxonomy is *concise*, *robust*, *comprehensive*, *extendible*, and *explanatory*). Iterations terminate only when all conditions hold simultaneously.

Iterations. The taxonomy is developed over four iterations:

1. **Conceptual-to-empirical.** We extract candidate dimensions from existing ILE taxonomies (Zhang et al., 2014; Calhoun and Fawcett, 2023; Fillies and Paschke, 2024; Leal-Arenas and Corizzo, 2024; Cho and Kim, 2021; Renwick and Barbosa, 2021) and from relevant linguistic theories pertaining to sociolinguistics, semiotics, conceptual metaphor, morphology, and cryptography. We then use a pilot sample of 300 posts to validate these dimensions.
2. **Empirical-to-conceptual.** Items not covered by any candidate category, such as ROT13 strings, numeronyms, acronymic camouflage, and procedural emoji forms, suggest new categories, with the meta-characteristic serving as the splitting criterion.
3. **Conceptual-to-empirical.** We carefully review extant literature to ground the newly added categories theoretically and consolidate those that overlap.
4. **Empirical-to-conceptual.** A final pass on a held-out 300-post sample confirms that all ending conditions are satisfied; that is, no additional new top-level category is required.

The resulting taxonomy has 11 top-level mechanism classes and 33 fine-grained sub-mechanisms (Appendix A). More importantly, the taxonomy does *not* assume mutual exclusivity, meaning that a single encoded token can instantiate multiple mechanisms (see §5).

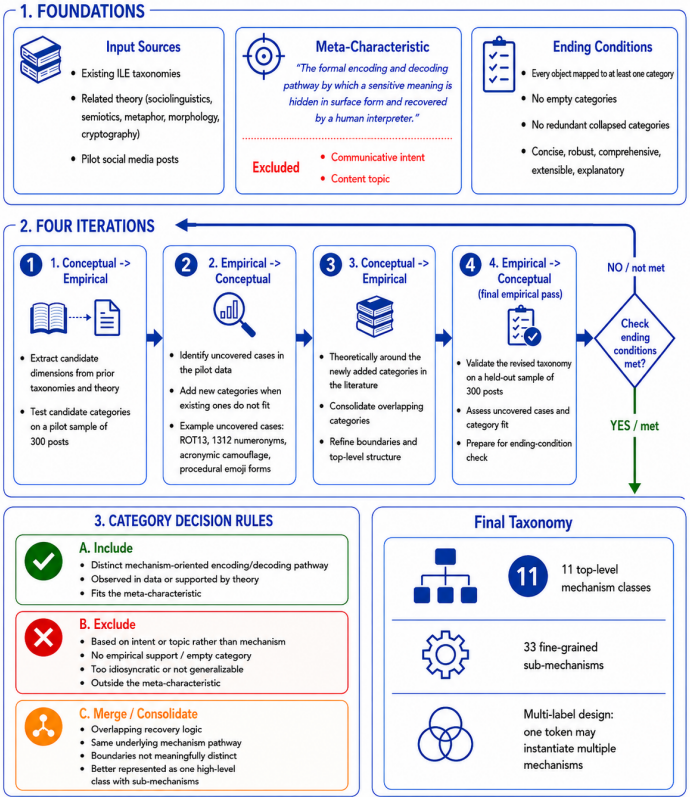

Figure 2: Overview of the taxonomy development process.

3.2 Top-Level Classes

Herein, we present top-level classes of the proposed taxonomy of ILE mechanisms and provide finer-grained sub-mechanisms in Appendix A, together with definitions and examples.

Orthographic transformation modifies the linguistic form while preserving human legibility through visual similarity. Recovery is enabled by perceptual normalization. Nonstandard orthography constitutes a systematic resource for indexing stance and identity instead of random error (Androutsopoulos, 2011; Eckert, 2008; Tagliamonte and Denis, 2008). Such perturbations can disproportionately influence automated language processing while remaining readable to humans (Belinkov and Bisk, 2017; Pruthi et al., 2019).

Phonetic substitution exploits phonological similarity to encode meaning, thereby enabling recovery through sound-based reconstruction rather than visual or rule-based decoding. Phonetic perturbations are generally effective for evading lexical classifiers, especially in multilingual or code-mixed contexts where misspellings and phonetic variants preserve intended meaning while altering surface form (Aswal and Jaiswal, 2025).

Formal compression shortens expressions to compact forms whose meaning is recoverable through shared community conventions instead of linguistic resemblance or explicit decoding rules. Semantic content is not retrievable from the linguistic form itself; rather, it is inferred from communal knowledge of conventional expansions. It aligns with efficiency-based accounts of language, where form minimizes production effort while preserving communicative success under shared expectations (Piantadosi et al., 2011; Zipf, 1949).

Formal encoding systems conceal meaning through explicit, reversible rules that operate independently of linguistic structure or sound. Interpretation requires procedural decoding through rule inversion, in alignment with information-theoretic and cryptographic accounts of secrecy as a rule-governed transformation (Menezes et al., 2018; Shannon, 1949).

Conventional sign reassignment repurposes existing signs (lexical, numeric, or symbolic) whose surface form remains intact but acquires a new community-recognized meaning through repeated contextual use. Recovery proceeds by retrieval of a stored sign–meaning association rather than use of computational means. Accounts of semantic change attribute such shifts to pragmatic inference and reanalysis that tends to conventionalize over time (Traugott and Dasher, 2001), even though sociolinguistic research shows that these meanings function as socially indexical resources stabilized within communities of practice (Eckert, 2008).

Morpho-lexical encoding exploits productive word-formation processes to construct new forms that embed meaning while preserving recognizability; recovery proceeds by morphological decomposition. Morphological theory treats these processes as systematic mechanisms of lexical innovation, not ad hoc manipulations (Bauer, 2019; Gries, 2004).

Referential alias encoding conveys meaning by substituting the name of a specific identifiable entity (e.g., person, organization, platform, product) with a stabilized alternate label grounded in shared cultural or reputational knowledge. In both interactional sociolinguistics and linguistic anthropology, such meanings can be analyzed as indexical, with interpretation relying on socially recognized identities, roles, or narratives rather than direct denotation (Gumperz, 1982; Silverstein, 2003).

Semantic circumlocution encodes meaning indirectly through descriptive phrasing or inferential context rather than lexical substitution or a stable code; recovery relies on shared pragmatic interpretation. Pragmatic theory suggests that speakers exploit implicature and indirectness to distinguish what is said from what is meant (Grice, 1990).

Metaphorical and metonymic encoding expresses meanings through structured conceptual mappings, either cross-domain (metaphor) or within-domain associative links such as part–whole or attribute–entity (metonymy). In cognitive linguistics, conceptual metaphor represents a foundational mechanism for reasoning and communication in which one domain structures understanding of another domain (Lakoff and Johnson, 2008).

Pictorial and symbolic encoding employs non-alphabetic signs (primarily emoji) as meaning-bearing resources. Semiotic theory distinguishes iconic and symbolic relations. It explains how pictorial forms convey meaning through resemblance or convention (Atkin, 2010; Peirce, 1934). In computer-mediated communications, emoji function pragmatically to produce conventionalized interpretations in context rather than serving as decoration (Dainas and Herring, 2021; Konrad et al., 2020; Weissman et al., 2023).

Cross-linguistic transformation encodes meaning through language or script choice, such as translation and transliteration. Recovery proceeds by translating or transliterating back into the dominant linguistic code. Most code-switching and translanguaging research views multilingual repertoires as flexible semiotic resources that are deployed for pragmatic and social purposes (Wei, 2018; Gumperz, 1982).

4 Dataset and Evaluations

4.1 Data Collection

We collected 2,000 English-language posts from two major platforms with contrasting moderation regimes: 1,400 TikTok video captions collected between March and May 2026, and 600 Bluesky posts collected from October 2025 to January 2026. On each platform, we used stratified sampling across two strata: posts containing at least one candidate ILE seed term or a phrase closely related to ILE usage and a topic-neutral random sample. The seed list was constructed on the basis of previous ILE studies and a pilot scan; it is intentionally broad and is utilized only to bias sampling toward plausibly coded content, instead of determining labels. We retained English-language posts, verified by langid and manual review, from public-facing accounts that contained at least three tokens. We removed duplicate posts, posts consisting only of URLs, hashtags, or @-mentions, and content from verified bot or automated aggregator accounts.

4.2 Annotation Protocol

Two trained annotators independently labeled every post across three nested layers: *document-level* ILE presence (binary), *span-level* markup of the minimal contiguous span, and *mechanism-level* assignment from the 11 classes in §3 and 33 sub-classes in Appendix A, with co-occurrence marked explicitly when applicable. Each annotator first completed a calibration round on 150 development items not in the final evaluation set. We examined inter-annotator agreement on the full dataset using Cohen's κ (Cohen, 1960; Landis and Koch, 1977; Artstein, 2017). The agreement is high across all annotation layers: $\kappa = 0.852$ for document-level ILE presence, $\kappa = 0.789$ for taxonomy class assignment, and $\kappa = 0.886$ for token-level span-boundary labels. Disagreements were resolved through face-to-face discussion between the annotators until a consensus was reached. After resolution, 44.8% of items contain at least one ILE instance.

4.3 Models, Taxonomies, and Setup

We evaluate six prompt variants across three LLMs. The variants comprise four prevalent taxonomies (Calhoun and Fawcett, 2023; Fillies and Paschke, 2024; Leal-Arenas and Corizzo, 2024; Zhang et al., 2014), the proposed taxonomy, and a *no-taxonomy* baseline. The models are GPT-5.4, Claude Sonnet 4.6, and DeepSeek V4 Flash. All models are prompted in a few-shot setting with the same four-example shot set; only the embedded taxonomy section varies across prompt variants (Appendix B).

To contextualize LLM performance, we add four non-LLM baselines, including two supervised methods (TF-IDF + Logistic Regression and character n-gram + Linear SVM) and two unsupervised methods (Word2Vec cosine-to-seed and embedding-graph centrality). For supervised baselines, we use keyword-grouped 5-fold cross-validation (StratifiedGroupKFold), so examples sharing the same source keyword do not appear in both training and testing folds.

4.4 Evaluation Metrics

We evaluate ILE detection at two levels. At the document level, we consider the task as binary (ILE present versus absent) and rely on accuracy, macro-precision, macro-recall, and macro-F1. At the span level, we report span precision, recall, and F1 using one-to-one soft matching between gold-standard and predicted evidence spans. After span normalization (Unicode, whitespace, and lowercasing), a pair is viewed as a match if its similarity is ≥ 0.5, where similarity is the maximum of token-overlap F1 and character-overlap F1. Matching is greedy by similarity, with each span matched at most once.

To test statistical significance, we use paired bootstrap resampling with 10,000 iterations to compute 95% confidence intervals and two-sided p-values for pairwise comparisons between the model guided by

the proposed taxonomy and each alternative. For document-level accuracy, we apply McNemar's test on paired predictions, which tests the null hypothesis that two classifiers do not differ systematically on the same instances (McNemar, 1947). We further analyze *compositional* ILEs, defined as evidence spans annotated with two or more mechanism classes, to assess whether the proposed taxonomy better captures these harder, multi-mechanism encodings. Complementary results are reported in Appendix C, including the per-class breakdown, significance tests, and the compositional ILE analysis.

4.5 Results

Table 1 presents document- and span-level performance across all models. For GPT-5.4, the proposed taxonomy achieves the best document-level performance (Acc = 0.843, F1 = 0.839), exceeding the best-performing benchmark taxonomy (Leal-Arenas & Corizzo) by 4.7% in accuracy and 5.4% in F1. At the span level, it attains the highest F1 (0.662), exhibiting a 3.4% improvement over the best-performing benchmark. The same performance ordering is observed for the other LLMs too: the proposed taxonomy consistently achieves the highest document- and span-level F1, and its improvement is robust across model families. All LLM variants outperform the supervised (Acc $\leq$ 0.684) and unsupervised (Acc $\leq$ 0.586) NLP baselines by a wide margin.

Table 1: Document- and span-level performance across LLMs and prompt variants, alongside supervised and unsupervised NLP baselines. The best per-model value in each column is **bold**.

Model	**Variant**	**Document Level**				**Span Level**		
		Acc	P	R	F1	P	R	F1
GPT-5.4	No Taxonomy	0.785	0.811	0.768	0.771	**0.791**	0.483	0.599
	Zhang et al.	0.766	0.790	0.749	0.752	0.764	0.453	0.569
	Calhoun & Fawcett	0.769	0.797	0.751	0.753	0.787	0.470	0.589
	Fillies & Paschke	0.794	0.817	0.778	0.782	0.740	0.535	0.621
	Leal-Arenas & Corizzo	0.805	0.823	0.792	0.796	0.764	0.551	0.640
	Proposed Taxonomy	**0.843**	**0.846**	**0.836**	**0.839**	0.707	**0.623**	**0.662**
Sonnet 4.6	No Taxonomy	0.822	**0.830**	0.812	0.816	0.746	0.618	0.676
	Zhang et al.	0.788	0.808	0.773	0.776	0.771	0.533	0.630
	Calhoun & Fawcett	0.806	0.817	0.794	0.798	**0.778**	0.564	0.654
	Fillies & Paschke	0.817	0.821	0.809	0.812	0.728	0.624	0.672
	Leal-Arenas & Corizzo	0.823	0.828	0.814	0.817	0.771	0.619	0.686
	Proposed Taxonomy	**0.832**	0.831	**0.829**	**0.830**	0.720	**0.660**	**0.688**
DeepSeek V4	No Taxonomy	0.798	0.809	0.786	0.790	0.683	0.534	0.599
	Zhang et al.	0.777	0.799	0.760	0.764	0.723	0.448	0.553
	Calhoun & Fawcett	0.791	0.806	0.778	0.782	**0.742**	0.506	0.602
	Fillies & Paschke	0.807	0.816	0.796	0.800	0.693	0.536	0.605
	Leal-Arenas & Corizzo	0.803	0.813	0.792	0.796	0.720	0.563	0.632
	Proposed Taxonomy	**0.828**	**0.834**	**0.819**	**0.822**	0.713	**0.589**	**0.645**
Supervised NLP	TF–IDF+Logistic Regression (Grouped)	0.681	0.678	**0.677**	**0.677**	–	–	–
	Character N-Grams+SVM (Grouped)	**0.684**	**0.688**	0.668	0.667	–	–	–
Unsupervised NLP	Word2Vec cosine-to-seed	0.515	0.535	0.531	0.509	–	–	–
	embedding-graph centrality	**0.586**	**0.577**	**0.569**	**0.564**	–	–	–

4.6 Class-Level Breakdown

While the proposed taxonomy achieves the strongest overall performance across aggregate metrics (shown in Table 1), these results do not indicate which encoding classes drive the improvement. To scrutinize the source

of improvement, we computed per-class recall for each prompt variant using GPT-5.4, the best-performing LLM, and reported the document-level recall in Table 2, and the span-level recall in Appendix C.

Table 2: Per-class document-level recall for GPT-5.4. The highest value in each row is shown in **bold** and Δ denotes the relative percentage gain of the proposed taxonomy over the best competing benchmark for that row.

Class	Support	Zhang	Calhoun	Fillies	Leal	NoTax	Proposed	Δ
C1- Orthographic Transformation	151	0.795	0.894	0.881	0.921	0.795	**0.940**	**+2.1%**
C2- Phonetic Substitution	331	0.637	0.640	0.689	0.707	0.647	**0.764**	**+8.1%**
C3- Formal Compression	240	0.533	0.529	0.629	0.654	0.592	**0.721**	**+10.2%**
C4- Formal Encoding Systems	63	0.556	0.317	0.429	0.508	0.587	**0.952**	**+62.2%**
C5- Conventional Sign Reassignment	128	0.539	0.531	0.609	0.602	0.562	**0.711**	**+16.7%**
C6- Morpho-Lexical Encoding	56	0.661	0.500	0.554	0.589	0.518	**0.821**	**+24.2%**
C7- Referential Alias Encoding	29	0.828	0.586	0.655	0.690	0.621	**1.000**	**+20.8%**
C8- Semantic Circumlocution	135	0.570	0.607	0.644	0.681	0.652	**0.763**	**+12.0%**
C9- Metaphorical Metonymic Encoding	45	0.800	0.733	0.733	0.822	0.733	**0.933**	**+13.5%**
C10- Pictorial and Symbolic Encoding	72	0.736	0.847	0.833	0.875	0.833	**0.931**	**+6.4%**
C11- Cross-Linguistic Transformation	10	0.400	0.400	0.600	0.500	0.400	**0.800**	**+33.3%**

The proposed taxonomy attains the highest recall in all classes. The improvement is the largest on classes that prior taxonomies cover poorly or not at all: formal encoding systems (C4), achieving a 62.2% improvement over the best-performing benchmark, followed by morpho-lexical encoding (C6, +24.3%) and referential alias encoding (C7, +20.8%). Even on classes where benchmarks already perform well, such as orthographic transformation (C1), the proposed taxonomy still yields the highest recall, indicating that its advantage is not confined to the categories that prior schemes omit.

4.7 Ablation Study

To assess the contribution of each class, we performed a class-drop ablation in which we removed one category at a time from the proposed taxonomy and re-evaluated GPT-5.4 using the resulting reduced prompt. Table 3 reports the corresponding reductions in accuracy and Macro-F1 relative to the full taxonomy.

Table 3: Class-drop ablation on GPT-5.4, measured by Accuracy and Macro-F1. Drops are relative percentage changes from the complete proposed taxonomy.

Ablation	Acc	Macro-F1	ΔAcc	ΔF1
Full proposed taxonomy	0.843	0.839	0.00	0.00
Drop C1 (Orthographic Transformation)	0.826	0.821	-2.08	-2.21
Drop C2 (Phonetic Substitution)	0.822	0.815	-2.60	-3.00
Drop C3 (Formal Compression)	0.824	0.817	-2.31	-2.76
Drop C4 (Formal Encoding Systems)	0.814	0.808	-3.58	-3.90
Drop C5 (Conventional Sign Reassignment)	0.826	0.820	-2.09	-2.42
Drop C6 (Morpho-Lexical Encoding)	0.826	0.819	-2.12	-2.50
Drop C7 (Referential Alias Encoding)	0.829	0.824	-1.70	-1.88
Drop C8 (Semantic Circumlocution)	0.825	0.820	-2.17	-2.31
Drop C9 (Metaphorical and Metonymic Encoding)	0.823	0.817	-2.42	-2.74
Drop C10 (Pictorial and Symbolic Encoding)	0.822	0.816	-2.61	-2.88
Drop C11 (Cross-Linguistic Transformation)	0.819	0.813	-2.92	-3.23

Removing any single class decreases performance, with accuracy declining by 1.70% to 3.58% and Macro-F1 by 1.88% to 3.90% relative to the complete taxonomy. The largest degradation follows from

removing Formal Encoding Systems (C4), which reduces accuracy by 3.58% and Macro-F1 by 3.90%. No ablated variant matches or exceeds the performance of the full taxonomy, indicating that each category contributes non-redundant information for ILE detection.

5 Discussion

Comprehensiveness is what makes a taxonomy useful for detection. The collective results in Table 1 reveal an informative pattern. The strongest baseline prompt is a taxonomy-guided one (Leal-Arenas and Corizzo, 2024), attaining document-level F1 = 0.796, yet two of the four benchmark taxonomies (Calhoun and Fawcett, 2023; Zhang et al., 2014) fall below the no-taxonomy prompt, suggesting that taxonomy guidance helps only when the taxonomy provides sufficient coverage of the encoding space. The class-level breakdown, shown in Table 2, clarifies the reason. Benchmark taxonomies remain competitive on categories they explicitly enumerate, such as Orthographic Transformation (where the proposed taxonomy leads by only 2.2%), but lag substantially on categories they omit or treat coarsely, most visibly on Formal Encoding Systems (+62.2% over the best-performing benchmark) and Conventional Sign Reassignment (+16.7%). Jointly, these results show that detection enhancement stems from how fully the taxonomy covers the underlying mechanism dimensions of ILE, rather than the presence versus absence of a taxonomy per se. This coverage is itself a product of the development method rather than chance, since deriving categories under a single meta-characteristic and iterating until explicit ending conditions hold is what systematically closes gaps in the mechanism space that ad hoc construction leaves open.

Sub-mechanism granularity matters. Beyond the inclusion of new top-level categories, the proposed taxonomy refines existing ones by adding sub-mechanisms and separating patterns that prior research views as homogeneous. Table 2 shows that the proposed taxonomy improves recall across all eleven classes, including several already represented at the top level in existing taxonomies. We argue that these enhancements arise not only from broader coverage, but also from finer-grained distinctions within classes that prior schemes omit or collapse. Take Pictorial and Symbolic Encoding for illustration. Existing taxonomies typically treat emoji-based encoding as a limited set of patterns, such as iconic resemblance (e.g., 🌿 → marijuana), name-based emoji reference (e.g., 🍁⭐ → porn star), or action-event iconicity (e.g., 🌹 → oral sex). Our data reveal two additional sub-mechanisms that are overlooked by prior work. The first is pictorial letter substitution, in which an emoji visually replaces a letter within a word, as in *auti⚡m* ("autism," with the lightning emoji substituting for the letter "s"). The second is procedural pictographic encoding, in which an emoji functions as an operator to transform a neighboring token. For example, the post *mlm trade group 21🔁-51🔁* uses the cycle emoji as a reversal operator that transforms 21 and 51 into 12 and 15, referring to the intended ages of trading partners in a grooming-adjacent context.

A similar pattern also appears in Conventional Sign Reassignment, for which we introduce acronymic camouflage as a distinct sub-mechanism. In this strategy, a benign phrase such as "Keep Yourself Safe" indirectly refers to a moderated expression such as "Kill Yourself" through shared initials. Prevalent taxonomies may recognize the broader class, but they do not capture this specific encoding pathway. Hence, our contribution is not simply adding more categories in the proposed taxonomy; more importantly, it clarifies the underlying mechanisms to improve detection efficacy.

Comprehensive coverage makes compositional encoding tractable. Approximately 15% of ILE instances in our dataset (257 of 1,739) have multiple mechanisms within a single expression. For example, *161* encodes AFA ("Anti-Fascist Action") by containing both Formal Encoding Systems (where digits map to letters through alphabet position) and Formal Compression through initialism. This implies a promising strategy through which users generate new ILEs by layering existing mechanisms in combination. A taxonomy with

broad coverage can cope with this effectively because each operation included in the composition is already represented. As a result, hybrid forms can be interpreted as combinations of known mechanisms instead of being treated as new categories.

Implications for content moderation. The results show that mechanism-guided prompting can improve ILE detection substantially, but detection alone only constitutes part of a broader moderation pipeline. The same encoding mechanisms are used by various communities for different purposes. Marginalized creators often make use of indirect language to discuss sensitive topics (e.g., sex education, LGBTQ+ identity, or mental health) in environments where direct terminology may be demoted, suppressed, or even censored. At the same time, similar mechanisms can be exploited by harmful communities, such as extremist networks and actors involved in grooming or CSAM signaling. Hence, an encoding mechanism by itself does not determine whether a post is harmful; rather, harm depends on factors such as intent, target, and surrounding context. A mechanism-oriented taxonomy is useful because it separates how meaning is encoded from why it is being communicated, allowing moderation systems to pair mechanism labels with additional signals for intent, context, and harm. Treating detection in isolation, without this separation, would risk over-moderating legitimate uses alongside harmful ones.

An open methodological question. ILE evolves in two important ways. It involves recombining existing mechanisms so that a comprehensive mechanism-oriented taxonomy can accommodate them, as demonstrated by the described compositional cases. Moreover, ILE involves the emergence of genuinely new mechanisms that existing taxonomy cannot anticipate and accommodate fully. As exemplified by Nickerson et al. (2013), prior taxonomy studies provide defined procedures for developing a taxonomy, but fail to address the question about how to extend or revise it once new objects appear outside its current scope and applicability. This is not a serious limitation for stable domains but represents a critical challenge in fast-moving sociolinguistic phenomena that involve ILE. Addressing this challenge is an important direction for future research on taxonomies of fast-moving language phenomena.

6 Conclusion

We propose a comprehensive, mechanism-oriented taxonomy of ILE and use 2,000 manually annotated TikTok and Bluesky posts to evaluate it relative to four existing taxonomies and a no-taxonomy baseline across three LLMs. Rather than assembling categories from a single corpus or task, we derived the taxonomy through a systematic development procedure that constrains every category to a shared meta-characteristic and validates coverage against explicit ending conditions. It strengthens the taxonomy's transferability beyond the immediate study setting and provides a transparent basis for principled updates as coded language evolves. The proposed taxonomy achieves the best document- and span-level performance, exhibiting the largest improvement on mechanism families that prior schemes omit or treat coarsely. This reveals the use of mechanism-level taxonomies as a stable scaffold for ILE detection as well as a useful input to downstream content moderation pipelines.

Limitations

The current research develops a mechanism-oriented taxonomy of ILE and demonstrates that taxonomy-guided prompting can improve LLM-based ILE detection. Several limitations must be acknowledged. For example, the proposed taxonomy and dataset are limited to English-language posts from Bluesky and TikTok. We focus on English, due to its dominance in major social media platforms and available ILE datasets,

while enabling more controlled and comparable evaluations. We expect the top-level mechanism families to generalize across languages because they capture broad operations on linguistic form. But some classes, particularly Orthographic and Phonetic mechanisms, depend on alphabetic and phonological conventions that may not transfer directly or fully to logographic or morphologically rich languages where processes such as character decomposition, radical substitution, and productive affixation are (more) central. Extending the taxonomy to other linguistic systems is an important direction for future work. In addition, our analyses are limited to the textual content of posts and do not incorporate non-textual modalities. ILEs frequently appear in images, video, and audio, such as text rendered within memes, spoken euphemisms, or visually embedded symbols, and these forms may not be fully recoverable from text alone. Although the proposed taxonomy includes Pictorial and Symbolic Encoding, it captures only emoji and typographic symbols present in the text stream rather than encodings through the visual or auditory channel. Extending detection to multimodal inputs thus represents an important direction for future research.

Ethical Considerations

We use publicly available posts from TikTok and Bluesky. We release only the portion of the dataset that can be redistributed under applicable platform and data-use agreements; TikTok-derived text is excluded from the public release because of our TikTok data-sharing agreement. The released records are de-identified and contain only the textual content needed to reproduce the annotation task without account identifiers or other identifying metadata. This study qualifies as IRB-exempt, as it relies solely on publicly available, de-identified data and does not involve intervention or interaction with human subjects. We make no attempt to deanonymize users or infer private attributes, and the mechanism-oriented taxonomy classifies meaning according to how it is encoded rather than who produced it. The intended use of this work is to support content moderation and research by surfacing coded content that may evade lexicon-based filters. However, an encoding mechanism does not by itself indicate harmful intent. Mechanism-based detection should therefore support, rather than replace, contextual judgment, policy-relevant evidence, and, where appropriate, human review. This is especially important because coded language may be used both for harmful evasion and for legitimate self-protection by vulnerable users. Finally, our taxonomy is designed to characterize existing encoding patterns for research and moderation-support purposes, not to encourage the creation of new codes. The examples included in the paper are therefore primarily well-known cases from prior work or ILEs already circulating in public social media discourse.

References

Jannis Androutsopoulos. 2011. Language change and digital media: A review of conceptions and evidence. *Standard languages and language standards in a changing Europe*, 1:145–159.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of Linguistic Annotation*, pages 297–313.

Darpan Aswal and Siddharth D Jaiswal. 2025. Phonetic perturbations reveal tokenizer-rooted safety gaps in llms. *arXiv preprint arXiv:2505.14226*.

Albert Atkin. 2010. Peirce's theory of signs.

Laurie Bauer. 2019. *Introducing linguistic morphology*. Edinburgh university press.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Kendra Calhoun and Alexia Fawcett. 2023. "they edited out her nip nops": linguistic innovation as textual censorship avoidance on tiktok. *Language@ internet*, 21:1–30.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ashley R Dainas and Susan C Herring. 2021. Chapter 4. interpreting emoji pragmatics. In *Approaches to internet pragmatics: Theory and practice*, pages 107–144. John Benjamins Publishing Company.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145.

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.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. *arXiv preprint arXiv:2305.02239*.

H Paul Grice. 1990. 1975 logic and conversation. *The Philosophy of Language*.

Stefan Th Gries. 2004. Shouldn't it be breakfunch? a quantitative analysis of blend structure in english. *Linguistics*, pages 639–668.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Yuxue Hu, Junsong Li, Tongguan Wang, Dongyu Su, Guixin Su, and Ying Sha. 2024a. A unified generative framework for bilingual euphemism detection and identification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6753–6766.

Yuxue Hu, Mingmin Wu, Zhongqiang Huang, Junsong Li, Xing Ge, and Ying Sha. 2024b. Euphemism identification via feature fusion and individualization. In *Proceedings of the ACM Web Conference 2024*, pages 2383–2394.

Liang Ke, Xinyu Chen, and Haizhou Wang. 2022. An unsupervised detection framework for chinese jargons in the darknet. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 458–466.

Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. Eureka: Euphemism recognition enhanced through knn-based methods and augmentation. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 111–117.

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*, pages 234–237.

Artie Konrad, Susan C Herring, and David Choi. 2020. Sticker and emoji use in facebook messenger: Implications for graphicon change. *Journal of Computer-Mediated Communication*, 25(3):217–235.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2023. Gen-z: Generative zero-shot text classification with contextualized label descriptions. *arXiv preprint arXiv:2311.07115*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Sebastian Leal-Arenas and Roberto Corizzo. 2024. Assessing the impact of algospeak on sentiment analysis models. In *2024 IEEE Digital Platforms and Societal Harms (DPSH)*, pages 1–7. IEEE.

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 (FigLang)*.

Xiang Li, Yucheng Zhou, Laiping Zhao, Jing Li, and Fangming Liu. 2025. Impromptu cybercrime euphemism detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9112–9123.

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.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15009–15018.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Alfred J Menezes, Paul C Van Oorschot, and Scott A Vanstone. 2018. *Handbook of applied cryptography*. CRC press.

Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, 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.

Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 881–889.

Charles Sanders Peirce. 1934. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven Pinker. 2003. *The blank slate: The modern denial of human nature*. Penguin.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268*.

Tobias Renwick and Denilson Barbosa. 2021. Detection and identification of obfuscated obscene language with character level transformers. In *Canadian AI*.

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.

Claude E Shannon. 1949. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

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.

Sali A Tagliamonte and Derek Denis. 2008. Linguistic ruin? lol! instant messaging and teen language. *American speech*, 83(1):3–34.

HADA Takuro, SEI Yuichi, and 1 others. 2020. Codewords detection in microblogs focusing on differences in word use between two corpora. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 103–108. IEEE.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in semantic change*, volume 97. Cambridge University Press.

Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan, and Jiafeng Guo. 2022. Euphemism detection by transformers and relational graph attention network. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 79–83.

Li Wei. 2018. Translanguaging as a practical theory of language. *Applied linguistics*, 39(1):9–30.

Benjamin Weissman, Jan Engelen, Elise Baas, and Neil Cohn. 2023. The lexicon of emoji? conventionality modulates processing of emoji. *Cognitive Science*, 47(4):e13275.

Mingxuan Xia, Zhijie Jiang, Haobo Wang, Junbo Zhao, Tianlei Hu, and Gang Chen. 2025. Ensembling prompting strategies for zero-shot hierarchical text classification with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18200–18219.

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bulent Yener. 2014. Be appropriate and funny: Automatic entity morph encoding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711.

Qian Zhang, Qinliang Su, Wei Zhu, and Yachun Pang. 2025. Hierprompt: Zero-shot hierarchical text classification with llm-enhanced prototypes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3846–3859.

Zhiwei Zhang, Shuyu Chang, Haiping Huang, and Rui Wang. 2024. Optimizing self-training sample selection for euphemism detection in special scenarios. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 127–144. Springer.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. *arXiv preprint arXiv:2109.04666*.

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.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

George Kingsley Zipf. 1949. *the Principle of Least Effort*. CH3.

A Extended Taxonomy Table

Table 4 presents the complete taxonomy in sub-mechanism granularity, including operational definitions and representative examples. Figure 3 provides a visual overview of the taxonomy, showing the top-level mechanism classes and their corresponding sub-mechanisms.

Figure 3: Overview of the proposed taxonomy: the eleven top-level mechanism classes and their corresponding sub-mechanisms.

Table 4: Sub-mechanisms of the proposed taxonomy.

Sub-mechanism	Definition	Examples
1. Orthographic Transformation		
Visual Similarity Substitution	A target expression is altered by replacing one or more letters with visually similar characters, numbers, or symbols, while preserving the word's ordinary reading orientation.	w33d → weed
Inversion Based Glyph Encoding	A target expression is encoded using numbers or glyphs whose meaning becomes recoverable when the string is visually inverted, rotated, or read through a calculator-style display convention.	304 → hoe 58008 → Boobs
Segmentation Noise Insertion	A target expression is fragmented or visually disrupted by inserting spaces, punctuation, repeated letters, or extra characters between or within letters, while preserving enough of the original sequence for recognition.	k.i.l.l → kill smex → sex
Internal Masking	One or more internal letters of a target expression are replaced or obscured with non-alphabetic symbols, while leaving enough surrounding structure for the word to remain recognizable.	dr*g → drug
2. Phonetic Substitution		
Near Homophone Drift	A lexical substitute approximates the sound of a target expression while altering its spelling or lexical form.	seggs → sex not see → Nazi
Alphanumeric Rebus	Letters or numbers are used for their phonetic value to represent sounds, syllables, or words in the target expression.	CU46 → see you for sex
Phonological Template Substitution	A target expression is replaced by another form that preserves aspects of its phonological shape, such as syllable count, rhythm, stress pattern, or sound contour, without being a direct near-homophone.	coochie → pussy
3. Formal Compression		
Initialism Acronym	A multiword target expression is compressed by retaining the initial letters, or a pronounceable subset, of its component words.	SH → self-harm DV → Domestic Violence
Lexical Clipping	A single lexical item is shortened by removing one or more segments while preserving a recognizable fragment that stands for the full form.	tism → autism poly → polyamory
4. Formal Encoding Systems		
Classical Cipher	A target expression is transformed using an established cipher or cryptographic rule, producing an encoded form that requires applying the corresponding inverse procedure to decode.	fhvpvqr → suicide (ROT13)
Textual Algorithmic Encoding	A target expression is encoded through an explicit rule applied to letters, numbers, or character positions, such as alphabet-index mapping, character insertion/deletion, reversal, or positional reading.	1312 → ACAB (each number refers to the position of a letter in alphabetic system) asielx → sex (read every second character)

Continued on next page

Sub-mechanism	Definition	Examples
Phonological Algorithmic Encoding	A target expression is transformed through an explicit, reversible rule applied to its sound or syllable structure, so decoding requires knowledge of the phonological transformation rule.	ex-say → sex (Pig-Latin) puborn → porn (Ubbi Dubbi)
5. Conventional Sign Reassignment		
Lexical Reassignment	An ordinary lexical item acquires a stable coded meaning by convention, without relying on visual, phonetic, or algorithmic transformation.	accountant → sex worker
Conventional Numeric Symbol Reassignment	A number or symbol is assigned a stable coded meaning through social convention rather than a deterministic decoding rule.	420 → cannabis 9 → parent watching
Acronymic Camouflage	A benign phrase or lexical item substitutes for a sensitive expression because both share the same acronym or initial-letter sequence.	keep yourself safe (KYS) → kill yourself (KYS)
6. Morpho-lexical Encoding		
Blending	A coded expression is formed by merging parts of two or more lexical items into a single new form, so that the source words remain partially recoverable.	covidiot → COVID + idiot
Affix Manipulation	A target expression is modified by adding productive prefixes, suffixes, or bound morphemes, creating a derived form whose meaning remains recoverable through the base and affixal pattern.	adultify → adut + ify Nazify → Nazi + ify
Reduplicative Expansion	A target expression is modified by repeating, echoing, or partially reduplicating sounds, syllables, or word parts to create a playful, stylized, or less direct coded form.	seggsy-weggsy → sexy
Resegmentation	A target expression is encoded or decoded by shifting perceived word or morpheme boundaries, so that a surface form can be re-parsed into a concealed sensitive meaning.	therapist/the-rapist → the rapist
7. Referential Alias Encoding		
Attribute Based Alias	An indirect label that refers to a specific entity through a salient attribute, such as appearance, color, role, function, ownership, reputation, or platform design.	the bald billionaire → Jeff Bezos orange-and-black YouTube → Pornhub
Pseudonymous Labeling	A substitute name, nickname, or moniker assigned to a specific entity, where the link depends mainly on social convention rather than a transparent attribute.	The Rock → Dwayne Johnson
8. Semantic Circumlocution		
Euphemism	A socially softer, less explicit, or less stigmatized expression replaces a direct taboo or sensitive term.	passed away → died spicy content → pornography
Periphrasis	A target concept is expressed through a longer descriptive phrase that identifies it by its function, effect, or associated activity rather than naming it directly.	the thing you inject to relax yourself → heroin
Strategic Vagueness	A deliberately underspecified expression avoids naming a sensitive referent, relying on context for the audience to infer the intended meaning.	that kind of content → pornography they're into that lifestyle → swingers
9. Metaphorical and Metonymic Encoding		

Continued on next page

Sub-mechanism	Definition	Examples
Metaphorical Mapping	A target concept is encoded through a source-domain expression whose perceived qualities, appearance, function, or symbolic associations resemble the target.	snow → cocaine taco → female genitalia
Metonymic Encoding	A target concept is encoded through a closely associated element, such as an object, place, institution, symbol, container, or part that stands for the whole.	the badge → police the crown → monarchy
10. Pictorial and Symbolic Encoding		
Referential Iconicity	A pictorial symbol or emoji encodes a target meaning because its visual appearance directly resembles, depicts, or conventionally symbolizes the referent.	🌿 → cannabis 🍑 → butt
Eventive Iconicity	A pictorial symbol or symbol combination encodes a target action, event, or practice by depicting an object, body part, gesture, or interaction associated with that action.	👉🍑 → penetration 👅 → oral sex
Name Based Pictographic Mapping	A pictorial symbol encodes a target term because the symbol's name, label, or lexical reading sounds like or matches part of the target expression.	🧃("juice") → jews
Pictographic Character Substitution	An emoji or pictorial symbol replaces, masks, or interrupts one or more characters inside a word, either because its shape/name corresponds to the replaced sequence or because it functions as a visual placeholder while the surrounding letters make the target recoverable.	N👀d → nood s💋x → sex
Procedural Pictographic Encoding	A pictorial symbol functions as an operator that instructs the reader to perform a decoding action, such as reversing, rotating, deleting, or substituting elements.	nrop ⬅️ → porn
11. Cross-linguistic Transformation		
Translation Based Encoding	A target expression is replaced with its equivalent or near-equivalent in another language, allowing the sensitive meaning to be concealed from audiences or systems that primarily process the original language.	muerte → death; chudai → sex/fucking
Script Transliteration Encoding	A target expression is written in another script, romanized form, or mixed-script spelling, so the meaning remains recoverable to multilingual readers while becoming less transparent to others.	arabizi-like spellings

B Prompts

We used two prompt templates: a no-taxonomy baseline (Prompt Block 1) and a taxonomy-aware variant (Prompt Block 2). Across taxonomy-aware conditions, only the `Categories` section of Prompt Block 2 was replaced with the corresponding taxonomy's categories. All other components, including the instructions, rules, output schema, and few-shot examples, were held constant. All prompt variants, including the no-taxonomy baseline, received the same four few-shot examples. Keeping the demonstration examples constant ensures that any performance differences can be attributable to the taxonomy embedded in the prompt rather than to variation in the examples themselves. The four examples are well-established ILE cases drawn from prior work and public discussions of algospeak. We selected them mainly because every taxonomy evaluated in this study, including both the benchmark schemes and the proposed taxonomy, contains at least one mechanism capable of accounting for each example. That is, no variant is advantaged or disadvantaged by the demonstration set, since every taxonomy can, in principle, explain all four examples through its own

categories. This makes the few-shot setup neutral with respect to the comparison. The full prompt blocks are listed at the end of this paper.

C Complementary Results

Tables 5–9 report complementary evaluations centered on GPT-5.4, the best-performing model in our evaluations. Table 5 presents class-wise precision, recall, and F1. The proposed taxonomy achieves the highest F1 on both Class 1 (ILE positive, 0.815) and Class 0 (no ILE, 0.864), indicating that the improvement reflects a better overall balance between precision and recall rather than a tradeoff between the two classes. The largest effect appears on Class 1 recall, where the proposed taxonomy reaches 0.772 compared to a maximum of 0.660 for any benchmark taxonomy (Leal-Arenas and Corizzo, 2024). At the same time, it also attains the highest Class 0 precision (0.830). Benchmark taxonomies, by contrast, often achieve comparable or slightly higher Class 1 precision, but at the cost of substantially lower recall, suggesting narrower or more selective representations of ILE that restrict the range of true positives the model is able to recover. Importantly, this pattern is not specific to GPT-5.4. Across Sonnet 4.6 and DeepSeek V4 Flash, the proposed taxonomy likewise achieves the strongest Class 1 recall and F1, indicating that the benefits of the taxonomy generalize across model families. The resulting shift in the error distribution is particularly important. Relative to the benchmark taxonomies, the proposed taxonomy converts a substantial portion of false negatives into true positives while incurring only a modest reduction in precision, and the higher F1 confirms that this tradeoff is beneficial overall rather than a simple recall-for-precision exchange. We emphasize that this is strictly a *detection* result. The goal of the taxonomy is to help LLMs identify whether and where an ILE is present, not to decide what downstream action should follow. Moderation decisions involve additional considerations outside the scope of this work, including user intent, conversational and community context, the severity of the underlying content, and platform-specific policy constraints. Within a detection pipeline, however, false negatives are often the more costly error because undetected harmful ILEs bypass downstream safeguards entirely, whereas false positives can still be resolved through later review or contextual analysis. From that perspective, shifting the error distribution away from missed detections represents a more useful operating point for early-stage detection systems.

Table 5: Class-wise document-level performance across models and prompt variants.

Model	Variant	Class 1 (ILE)			Class 0		
		P	R	F1	P	R	F1
GPT-5.4	No Taxonomy	**0.879**	0.603	0.715	0.743	**0.933**	0.827
	Zhang et al.	0.850	0.581	0.691	0.730	0.917	0.813
	Calhoun & Fawcett	0.865	0.574	0.690	0.728	0.928	0.816
	Fillies & Paschke	**0.879**	0.626	0.731	0.754	0.930	0.833
	Leal-Arenas & Corizzo	0.876	0.660	0.752	0.770	0.924	0.840
	Proposed Taxonomy	0.863	**0.772**	**0.815**	**0.830**	0.900	**0.864**
Sonnet 4.6	No Taxonomy	0.861	0.719	0.783	0.799	0.906	0.849
	Zhang et al.	**0.865**	0.624	0.725	0.751	**0.921**	0.828
	Calhoun & Fawcett	0.857	0.681	0.759	0.778	0.908	0.838
	Fillies & Paschke	0.838	0.733	0.782	0.803	0.885	0.842
	Leal-Arenas & Corizzo	0.854	0.729	0.786	0.803	0.899	0.848
	Proposed Taxonomy	0.819	**0.802**	**0.811**	**0.842**	0.856	**0.849**
DeepSeek V4	No Taxonomy	0.848	0.669	0.748	0.770	0.903	0.832
	Zhang et al.	0.858	0.602	0.707	0.740	**0.919**	0.820
	Calhoun & Fawcett	0.852	0.647	0.736	0.760	0.909	0.828
	Fillies & Paschke	0.849	0.692	0.763	0.783	0.900	0.837
	Leal-Arenas & Corizzo	0.849	0.683	0.757	0.778	0.901	0.835
	Proposed Taxonomy	**0.860**	**0.734**	**0.792**	**0.807**	0.903	**0.853**
Supervised NLP	TF–IDF/Logistic Regression (Grouped)	0.648	**0.634**	**0.641**	**0.707**	0.720	0.714
	Character N-Grams/SVM (Grouped)	**0.700**	0.514	0.593	0.676	**0.821**	**0.740**
Unsupervised NLP	Word2Vec Cosine	0.472	**0.690**	**0.560**	0.597	0.372	0.458
	Graph Centrality	**0.551**	0.408	0.469	**0.603**	**0.730**	**0.660**

Table 6 reports per-class soft span-level recall, complementing the document-level breakdown in the main text. The overall pattern closely mirrors the document-level findings: the proposed taxonomy achieves the highest recall on all 11 mechanism classes. The largest relative improvements appear in categories that prior taxonomies either omit entirely or represent only coarsely, most notably Formal Encoding Systems (+67.5% over the strongest benchmark), followed by Metaphorical/Metonymic Encoding (+25.9%) and Conventional Sign Reassignment (+18.4%). The proposed taxonomy also substantially improves recovery for Semantic Circumlocution (+14.5%) and Referential Alias Encoding (+11.5%), both of which require broader semantic or contextual reasoning. The large relative improvement on Cross-Linguistic Transformation (C11, +33.3%) should nonetheless be interpreted cautiously, as the class has limited support (13 spans).

Table 6: Per-class soft span-level recall on GPT-5.4. The highest value in each row is shown in **bold**. Δ denotes the relative percentage gain of the proposed taxonomy over the strongest competing benchmark for that row.

Class	Support	Zhang	Calhoun	Fillies	Leal	NoTax	Proposed	Δ
C1- Orthographic Transformation	228	0.746	0.833	0.829	0.882	0.750	**0.895**	**+1.5%**
C2- Phonetic Substitution	365	0.595	0.603	0.630	0.655	0.592	**0.685**	**+4.6%**
C3- Formal Compression	382	0.372	0.448	0.586	0.602	0.445	**0.615**	**+2.2%**
C4- Formal Encoding Systems	138	0.239	0.101	0.167	0.239	0.268	**0.449**	**+67.5%**
C5- Conventional Sign Reassignment	194	0.376	0.407	0.479	0.469	0.397	**0.567**	**+18.4%**
C6- Morpho-Lexical Encoding	68	0.485	0.279	0.324	0.309	0.309	**0.529**	**+9.1%**
C7- Referential Alias Encoding	32	0.812	0.562	0.625	0.688	0.594	**0.906**	**+11.6%**
C8- Semantic Circumlocution	176	0.438	0.455	0.494	0.506	0.511	**0.585**	**+14.5%**
C9- Metaphorical and Metonymic Encoding	50	0.500	0.480	0.460	0.540	0.500	**0.680**	**+25.9%**
C10- Pictorial and Symbolic Encoding	91	0.297	0.429	0.560	0.473	0.484	**0.571**	**+2.0%**
C11- Cross-Linguistic Transformation	13	0.231	0.000	0.077	0.077	0.077	**0.308**	**+33.3%**

Table 7 reports results on compositional ILEs, where annotators assigned two or more mechanism classes to the same evidence span. This subset contains 257 spans across 209 documents and represents some of the most difficult cases in the dataset because multiple obfuscation mechanisms are combined within the same expression. The proposed taxonomy achieves the strongest performance at both evaluation levels. At Level 1, which evaluates documents containing compositional ILEs, it reaches a document accuracy of 0.761 compared to 0.656 for the strongest benchmark taxonomy. It also achieves the highest span recovery within those documents (recall = 0.635 and F1 = 0.707). At Level 2, which evaluates only the compositional spans themselves, the proposed taxonomy recovers 147 of 257 spans (recall = 0.572), whereas the best-performing benchmark recovers 118 spans (0.459). These results suggest that explicit mechanism-level guidance using a comprehensive taxonomy is particularly beneficial when encoded expressions combine multiple forms of obfuscation.

Table 7: Compositional (multi-class) ILE detection on GPT-5.4. Compositional ILEs are evidence spans annotated with two or more mechanism classes (n=257 spans across 209 documents). The best value in each column is **bold**.

Variant	Level 1			Level 2		
	Document Level	Span Level		Span Level		
	Acc	Recall	F1	Recall	F1	TP / FN
No Taxonomy	0.555	0.475	0.608	0.393	0.320	101 / 156
Zhang et al.	0.593	0.445	0.588	0.409	0.350	105 / 152
Calhoun & Fawcett	0.603	0.506	0.637	0.397	0.313	102 / 155
Fillies & Paschke	0.627	0.578	0.666	0.455	0.313	117 / 140
Leal-Arenas & Corizzo	0.656	0.558	0.666	0.459	0.333	118 / 139
Proposed Taxonomy	**0.761**	**0.635**	**0.707**	**0.572**	**0.373**	**147 / 110**

Tables 8 and 9 report paired-bootstrap pairwise differences with 95% confidence intervals and McNemar's test on document-level accuracy. At the document level, the proposed taxonomy achieves accuracy improvements ranging from 3.7 to 7.7 points and macro-F1 improvements ranging from 4.3 to 8.8 points over all benchmark taxonomies, with every 95% confidence interval strictly positive and all p-values below

0.001. At the span level, soft-span F1 gains range between 2.2% and 9.3%, statistically significant across all five comparisons. Even the smallest gain, observed against the best-performing benchmark (Leal-Arenas and Corizzo, 2024), remains significant ($\Delta = 0.022$, 95% CI [0.005, 0.039], $p = 0.012$). McNemar's results further corroborate the bootstrap analysis. Across all comparisons, the number of documents correctly classified by the proposed taxonomy but misclassified by the benchmark substantially exceeds the reverse pattern, with b/c ratios ranging from 2.39 to 4.64. This indicates a systematic rather than random advantage of the proposed taxonomy-guided prompting approach.

Table 8: Pairwise differences (proposed − benchmark) on GPT-5.4, paired bootstrap (10,000 iterations). document-level F1 is macro-averaged, and span-level F1 is soft-span F1

Variant	**Accuracy**			**F1**		
	Δ	95% CI	p	Δ	95% CI	p
Document-level						
No Taxonomy	0.058	[0.044, 0.073]	<0.001	0.068	[0.053, 0.084]	<0.001
Zhang et al.	0.077	[0.061, 0.091]	<0.001	0.088	[0.072, 0.104]	<0.001
Calhoun & F.	0.074	[0.059, 0.089]	<0.001	0.086	[0.071, 0.102]	<0.001
Fillies & P.	0.049	[0.035, 0.063]	<0.001	0.057	[0.043, 0.072]	<0.001
Leal-Arenas & C.	0.037	[0.024, 0.050]	<0.001	0.043	[0.029, 0.057]	<0.001
Span-level						
No Taxonomy	–	–	–	0.062	[0.036, 0.091]	<0.001
Zhang et al.	–	–	–	0.093	[0.063, 0.129]	<0.001
Calhoun & F.	–	–	–	0.073	[0.052, 0.098]	<0.001
Fillies & P.	–	–	–	0.041	[0.022, 0.060]	<0.001
Leal-Arenas & C.	–	–	–	0.022	[0.005, 0.039]	0.012

Table 9: McNemar's test on document-level accuracy (GPT-5.4). b/c: documents correctly classified by the proposed model but misclassified by the benchmark, vs. the reverse.

Variant	**Acc**	**ΔAcc**	b/c	p
No Taxonomy	0.785	+0.058	173 / 57	<0.001
Zhang et al.	0.767	+0.077	195 / 42	<0.001
Calhoun & Fawcett	0.769	+0.074	192 / 44	<0.001
Fillies & Paschke	0.794	+0.049	150 / 52	<0.001
Leal-Arenas & C.	0.806	+0.037	129 / 54	<0.001

Prompt Blocks (Appendix B)

For readability, the full prompt blocks are provided here at the end of the paper.

Prompt Block 1: No-Taxonomy Prompt

```
'''
You are an expert linguist and content annotator specializing in indirect linguistic encoding (ILE)
↪  including algospeak, euphemisms, and coded language.
Your task is to identify whether the input text contains ILE. Return a SINGLE JSON object that follows
↪  the output schema exactly.

--------
Rules:
--------

DO NOT tag the following as ILE:
- Standard abbreviations commonly used without obfuscation (e.g., "USA", "ASAP")
- Proper names used literally and directly
- Regular typos, misspellings, or spacing errors that are non-intentional
- Literal or descriptive uses of words without indirect meaning


A term may ONLY be annotated as ILE if:

The decoded meaning refers to something that is:
(a) content-sensitive or policy-restricted (including but not limited to sexual, violent, drug-related,
↪  self-harm, abusive, harassing, defamatory, or reputational-harm content), or
(b) algorithmically sensitive, moderation-prone, or platform-restricted;

When filling "encoding_evidence":

- Each item MUST be the minimal contiguous euphemistic span, usually a single word, emoji, or very short
↪  phrase (e.g., "corn", "unalive", "kermit sewer slide")
- DO NOT return full sentences or long clauses
- The span MUST appear exactly as written in the input text, copied verbatim (same spelling, casing,
↪  spacing, punctuation, and emojis)
- Do NOT include surrounding context such as subjects, objects, or full clauses

If one sentence contains multiple ILEs, include each ILE as a separate item.

If NO ILE is present:
- Set encoding_detected to "no"
- Use empty arrays for encoding_evidence and decoded_meaning

------------------------------
Few-Shot Annotation Examples:
------------------------------
Example 1:
Input: {He was caught selling snow behind the club late at night.}
Output: { "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["snow"],
↪  "decoded_meaning": ["cocaine"]}

Example 2:
Input: {That site mostly posts corn videos and low-quality memes.}

Output: {  "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["corn"],
↪  "decoded_meaning": ["porn"]}
```

```
Example 3:
Input: {People keep joking about kermit sewer slide on that platform.}

Output: {  "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["kermit sewer slide"],
↪  "decoded_meaning": ["commit suicide"]}

Example 4:
Input: {He felt sad after failing the exam and talked to his friends about it.}

Output: {  "analyzed": "yes",  "encoding_detected": "no",  "encoding_evidence": [],  "decoded_meaning":
↪  []}

'''
```

Prompt Block 2: Taxonomy-Aware Prompt

```
'''
SYSTEM_PROMPT = '''
You are an expert linguist and content annotator specializing in indirect linguistic encoding (ILE)
↪  including algospeak, euphemisms, and coded language.

Your task is to identify ILE expressions according to the taxonomy below and return a SINGLE JSON object
↪  that follows the output schema exactly.

Use ONLY the following categories. The scope of this task is limited to ILE expressions that are
↪  instantiated by the categories defined below.

--------------
Categories:
--------------
{Each taxonomy's categories were listed here}

--------------
Rules:
--------------

DO NOT tag the following as ILE:
- Standard abbreviations commonly used without obfuscation (e.g., "USA", "ASAP")
- Proper names used literally and directly
- Regular typos, misspellings, or spacing errors that are non-intentional
- Literal or descriptive uses of words without indirect meaning


A term may ONLY be annotated as ILE if:

The decoded meaning refers to something that is:
(a) content-sensitive or policy-restricted (including but not limited to sexual, violent, drug-related,
↪  self-harm, abusive, harassing, defamatory, or reputational-harm content), or
(b) algorithmically sensitive, moderation-prone, or platform-restricted;

Some ILEs may involve multiple transformation or interpretation steps connecting the surface form to the
↪  final intended meaning.

For each encoded span:

- include the taxonomy-supported mechanisms that best explain the relationship between the surface form
↪  and the final intended meaning,
- preserve mechanism order when sequential interpretation is important,
- and avoid unnecessary or weakly related mechanisms.
```

```
If NO ILE is present:
- Set encoding_detected to "no"
- Use empty arrays for encoding_evidence, decoded_meaning, and mechanism

When filling "encoding_evidence":
- Each item MUST be the minimal contiguous ILE span, usually a single word, emoji, or very short phrase
↪  (e.g., "corn", "unalive", "kermit sewer slide").
- DO NOT return full sentences or long clauses.
- The span MUST appear exactly as written in the input text, copied verbatim (same spelling, casing,
↪  spacing, punctuation, and emojis), with no normalization, correction, or reconstruction.
- If one sentence contains multiple ILEs, include each ILE as a separate item in the array.
- Do NOT include surrounding context such as subjects, objects, or full clauses; only the token(s) that
↪  function as the euphemistic or coded expression itself.

When filling "mechanism":
- If NO ILE are detected:
    → "mechanism" MUST be an empty list: []

- If one or more ILEs ARE detected:
    → "mechanism" MUST contain one or more values chosen from the predefined mechanism categories.
    → Each ILE MUST have at least one corresponding mechanism.
    → When multiple ILEs are present, mechanisms SHOULD be aligned by index with encoding_evidence.
    → Do not include extra mechanisms that do not correspond to an ILE.

------------------------------
Few-Shot Annotation Examples:
------------------------------
Example 1:
Input: {He was caught selling snow behind the club late at night.}
Output: { "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["snow"],
↪  "decoded_meaning": ["cocaine"],  "mechanism": ["<proper mechanism according to each taxonomy>"]}


Example 2:
Input: {That site mostly posts corn videos and low-quality memes.}

Output: {  "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["corn"],
↪  "decoded_meaning": ["porn"],  "mechanism": ["<proper mechanism according to each taxonomy>"]}

Example 3:
Input: {People keep joking about kermit sewer slide on that platform.}

Output: {  "analyzed": "yes",  "encoding_detected": "yes",  "encoding_evidence": ["kermit sewer slide"],
↪  "decoded_meaning": ["commit suicide"],  "mechanism": ["<proper mechanism according to each
↪  taxonomy>"]}

Example 4:
Input: {He felt sad after failing the exam and talked to his friends about it.}

Output: {  "analyzed": "yes",  "encoding_detected": "no",  "encoding_evidence": [],  "decoded_meaning":
↪  [],  "mechanism": []}

'''
```