

# How Euphemism Shapes Implicit Toxicity in Online Discourse

Shiyao Cui

The Conversational AI (CoAI) group, DCST, Tsinghua University




清华大学  
Tsinghua University



# Background

- ◉ Toxicity Content
  - ◆ Language that is *overtly harmful, abusive, or offensive*, such as insults, threats, hate speech, harassment, or any expression intended to degrade, attack, or distress others.
- ◉ Implicit Toxicity Content
  - ◆ Language that conveys harmful or abusive intent *in subtle, indirect, or coded ways*, making it harder to detect than explicit toxicity

Cases	Type	Sentence
#1	Explicit	You are such <i>an idiot</i> , nobody wants you here
#2	Implicit	People like you are always lazy, <i>but I guess that's just your culture</i>
#3		We need to be careful; too many 'urban youths' hanging around usually means trouble



Euphemism

# Background

---

## ◎ Euphemism

- ◆ Terms, or phrases used to **substitute more offensive terms to downplay its unpleasantness**

## ◎ Where is euphemism used and for what

- ◆ Hate speech communities: evade moderation
- ◆ Youth social media (Gen Alpha): hide from parents/teachers/AI
- ◆ Political discourse: signal ideology subtly
- ◆ Everyday interaction: reduce offense / maintain politeness

# Background

---

- Typical forms of euphemism to express toxicity
  - ◆ Jargon
  - ◆ Dow Whistle
  - ◆ Slang
  - ◆ Digital language

# Reading Thieves' Cant: Automatically Identifying and Understanding Dark Jargons from Cybercrime Marketplaces



清华大学  
Tsinghua University



# Jargon Detection: Motivation

## ◉ Jargons

### ◉ Innocently-looking terms that serves sinister purposes

(1) My fav is slayers new `rat`, its open source, gonna have his rootkit implemented into it.

**Remote access trojan**

(2) Strains i manage these days are `BLUEBERRY` and NYC Diesel.

**Marijuana (大麻)**

(3) I vouch for this user he crypted my `athena` code.

**a botnet framework**

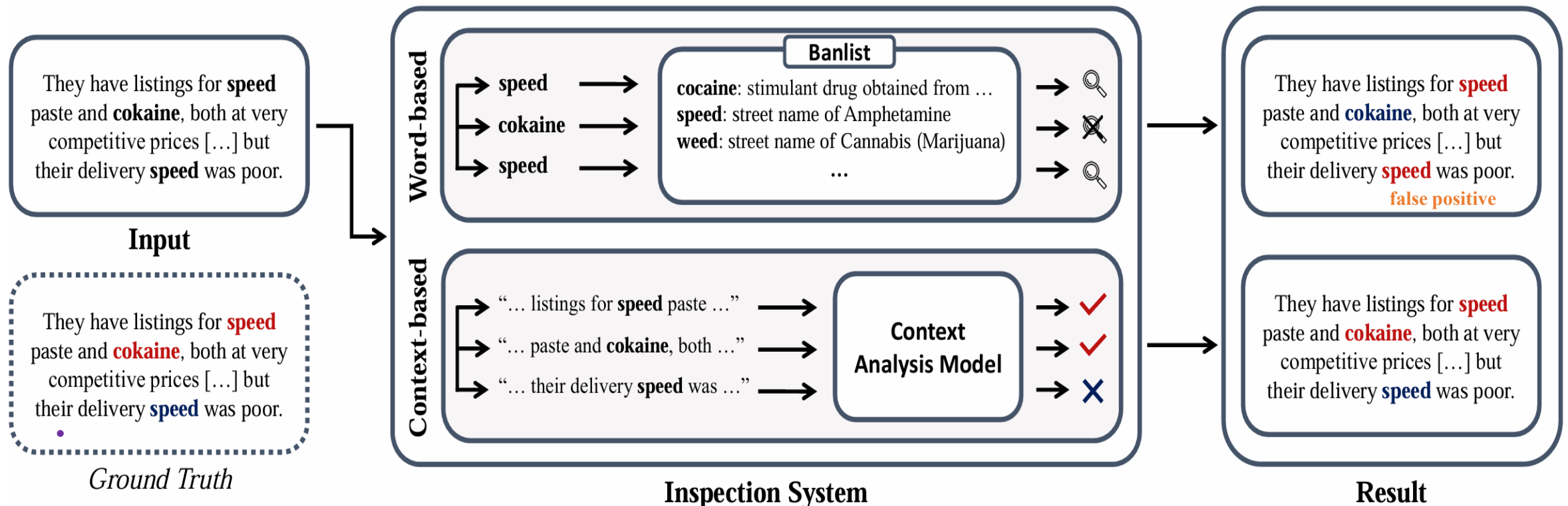
### ◉ Jargons propose two challenges to existing word-based moderation systems

- ◉ Word-based restrictions are easily bypassed by alternative terms for moderated words
- ◉ Euphemistic jargon has benign usages that can cause false positives

# Jargon Detection: Motivation

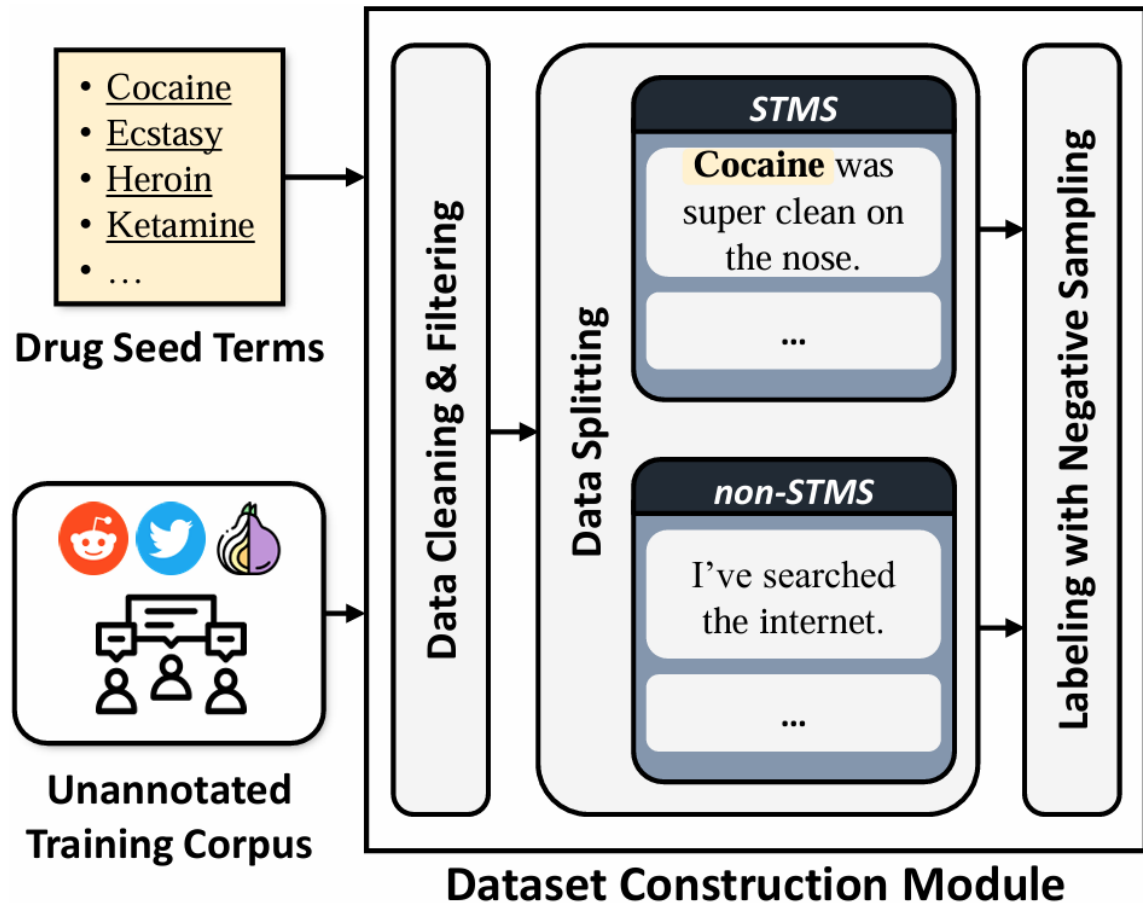
## How to detect drug jargon detection

- ◆ Moderation systems must be trained to evaluate words **within their specific context**



# Jargon Detection: Method

## Step1: Dataset Construction

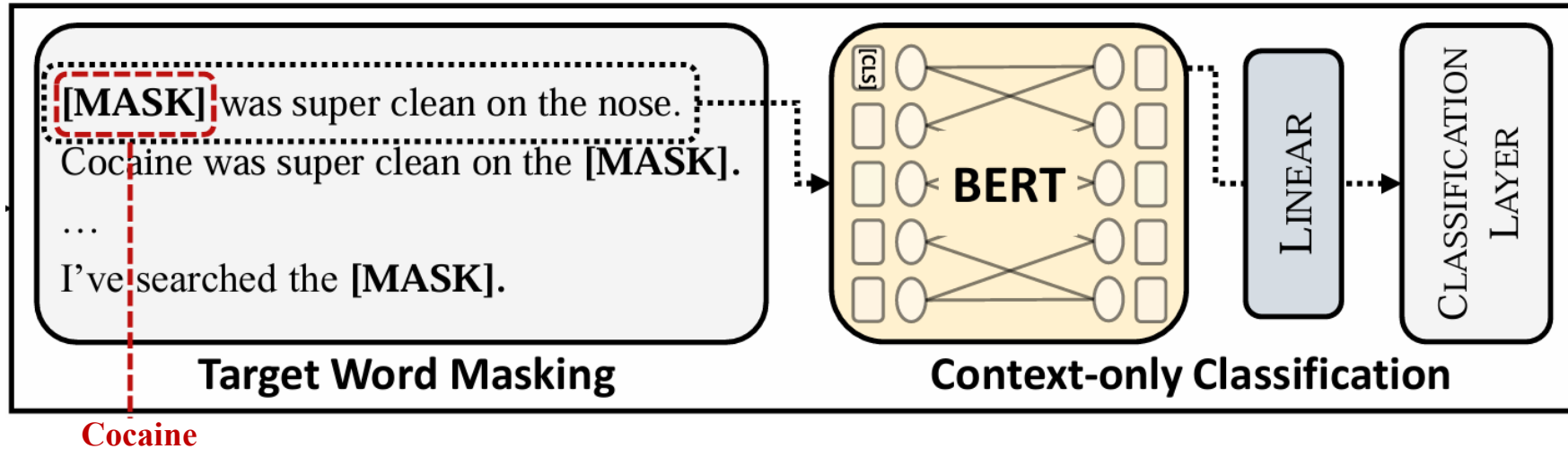


- **Data Cleaning:** separate training corpus into sentences; remove non-ASCII characters, HTML patterns, etc.
- **Data Filtering:** filter out longer and shorter sentences;
- **Data Splitting:** Seed Term Mentioning Sentences (STMS) and sentences without seed terms (non-STMS)
- **Labeling with Negative Sampling:** annotate the seed terms in the STMS; creating negative training data requires more careful sampling



# Jargon Detection: Method

## Step2: Context-only Module



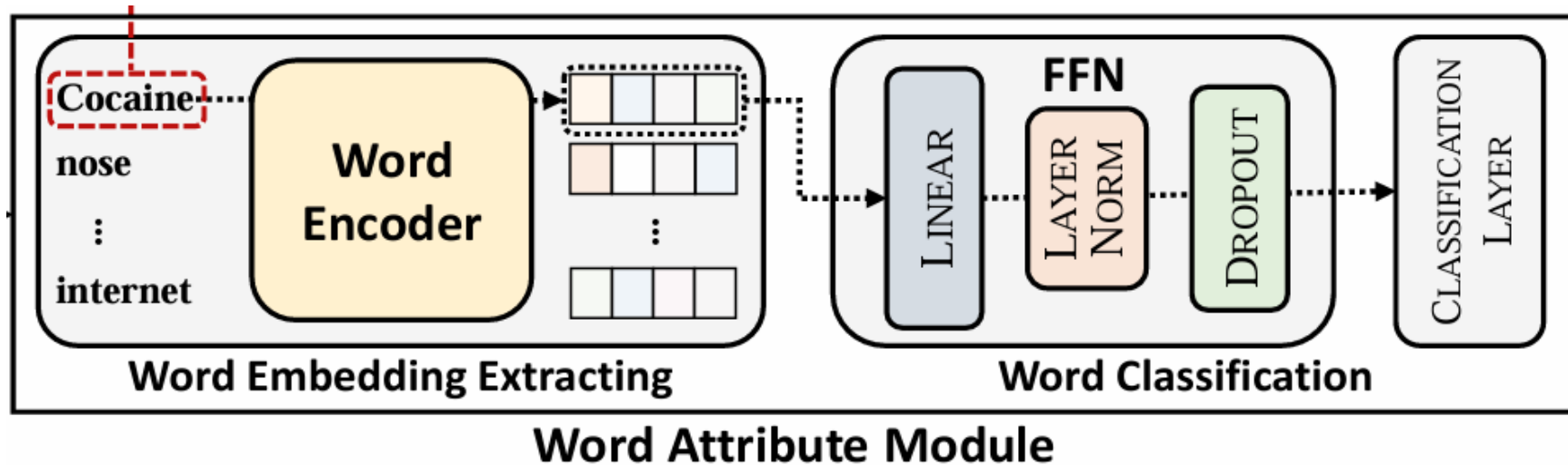
- Delexicalize the prediction process by **masking the target word with the [MASK] token**
- Calculate the probability that the masked word is drug jargon

$$\mathbf{o}_{pooled} = \tanh(\mathbf{W}_1 \mathbf{h}_{[CLS]} + \mathbf{b}_1)$$

$$p_c = \text{sigmoid}(\mathbf{W}_2 \mathbf{o}_{pooled})$$

# Jargon Detection: Method

## Step3: Word Attribute Module



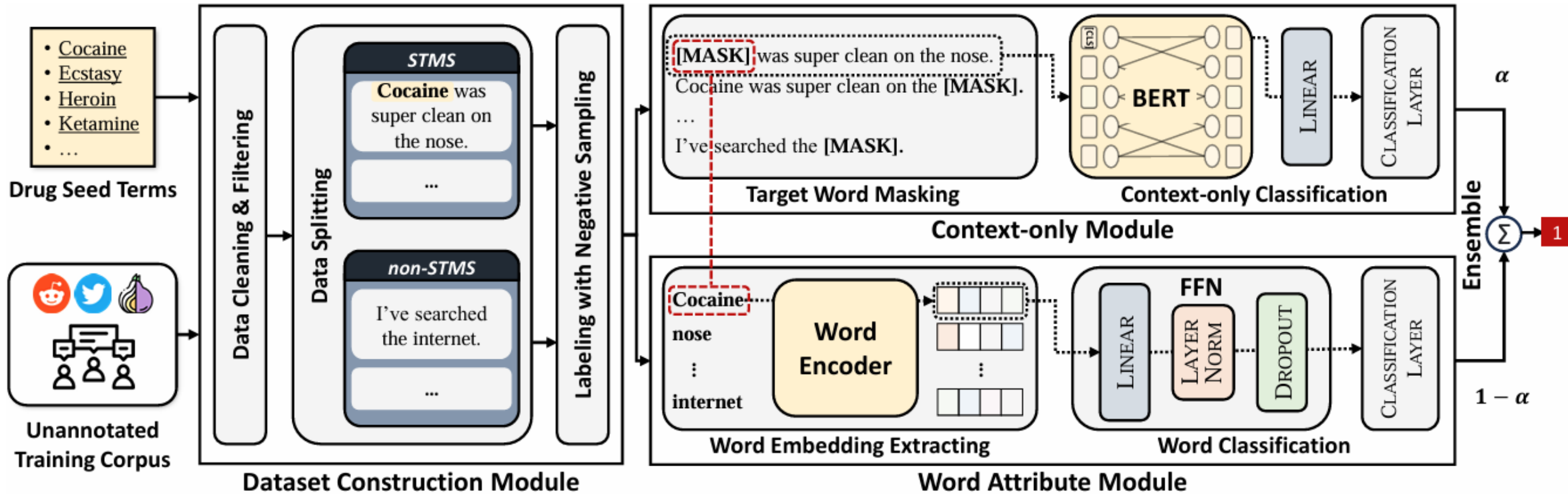
- We utilize a BERT model pretrained on the target domain as the word encoder, allowing us to utilize the context-sensitive intrinsic information of target words
- We train a feed-forward network and a classification layer that use the dynamic embeddings of the target word.

$$\mathbf{z}_{word} = \mathbf{W}_3 \mathbf{e}_{word} + \mathbf{b}_3$$
$$\mathbf{o}_{word} = \text{Dropout}(\tanh(\text{LayerNorm}(\mathbf{z}_{word})))$$

$$p_w = \text{sigmoid}(\mathbf{W}_4 \mathbf{o}_{word})$$

# Jargon Detection: Method

## Overview of the method



# Jargon Detection: Results

	Reddit Drug				Silk Road Forum			
	Precision	Recall	F1-score	# Jargon	Precision	Recall	F1-score	# Jargon
Word2Vec [23]	<b>0.8908</b>	0.2371	0.3746	32	<b>0.7908</b>	0.3277	0.4634	39
CantReader [48]	<u>0.7778</u>	0.1879	0.3027	17	<u>0.7208</u>	0.2347	0.3541	22
Zhu et al. [54]	0.5897	0.2573	0.3583	17	0.5026	0.4123	0.4530	27
PETD [15]	0.4760	0.2215	0.3023	28	0.5789	0.2326	0.3318	26
MLM (w/o pretrain)	0.5094	0.3043	0.3810	61	0.4569	0.2579	0.3297	57
MLM	0.5241	0.6085	0.5631	98	0.5783	0.6321	0.6040	102
DarkBERT [11]	0.5460	0.5705	0.5580	92	0.4790	0.5793	0.5244	109
GPT4o-mini [28]	0.4307	<b>0.7919</b>	0.5579	<b>129</b>	0.3920	<b>0.7907</b>	0.5242	<b>137</b>
JEDIS (w/o NegSTMS)	0.6052	<u>0.6309</u>	0.6177	<u>105</u>	0.5367	0.6808	0.6002	112
JEDIS (w/o pretrain&word)	0.6318	0.5951	0.6129	97	0.5368	0.6321	0.5806	104
JEDIS (w/o word)	0.6454	0.6107	<u>0.6276</u>	97	0.5507	<u>0.6998</u>	<u>0.6164</u>	<u>116</u>
JEDIS (w/o pretrain)	0.6475	0.5794	0.6116	92	0.5573	0.6068	0.5810	101
JEDIS	0.6659	0.6197	<b>0.6419</b>	101	0.5805	0.6934	<b>0.6320</b>	113

- Word-based approaches performed better on the SilkRoad dataset than on the Reddi Drug dataset, suggesting that the two datasets might be different in the distribution of more difficult jargon terms
- The context-based methods that explicitly consider contexts generally performed better
- JEDIS performed best underscoring its superior detection coverage

# Jargon Detection: Error Analysis

	Example sentences	Limitations	Example words
Word2Vec [23]	I have also heard that smoking crack with weed is an amazing experience .	Cannot find euphemisms	crack, speed, weed
CantReader [48]	I've just tried it before and its niceeeeeee .	Sensitive to typos	comfused, niceeeeeee
Zhu et al. [54]	Same thing happened with 10mg valium + 40mg meth .	limited to BERT vocabulary	add, met, valium

- Word2Vec may not be able to effectively identify euphemisms since it assumes a single meaning for each word.
- CantReader showed low overall precision with many false positives.
- Words which are not contained in the vocabulary of BERT can not be identified

# Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats



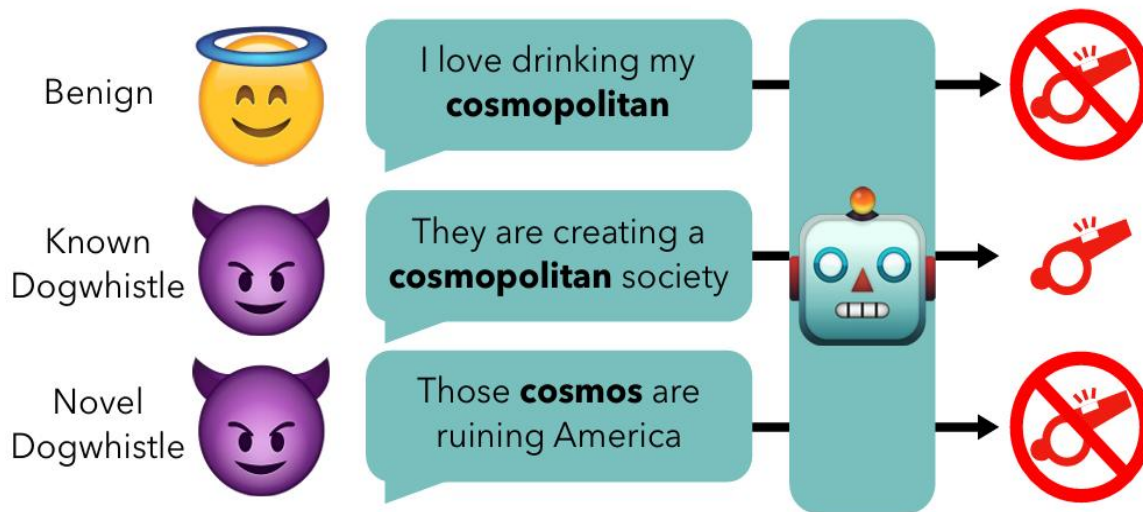
清华大学  
Tsinghua University



# Dog Whistle: Background

## ◉ Dog Whistle

- ◉ Expression that sends **one message to an outgroup** while at the same time **sending a second** (often taboo, controversial, or inflammatory) **message to an ingroup**



Word	Semantics	Dog Whistle
cosmopolitan	国际化的	暗指犹太人，不忠于国家，只全球逐利
cosmos	宇宙	同上

cosmos是cosmopolitan再次演变后的用法

- Dog whistles could slip by content moderation, toxicity filters due to **the benign second meaning**
- Detecting dog whistles with manually curated lexicons is labor-intensive due to **dynamic evolution language**

# Dog Whistle: Task Formulation

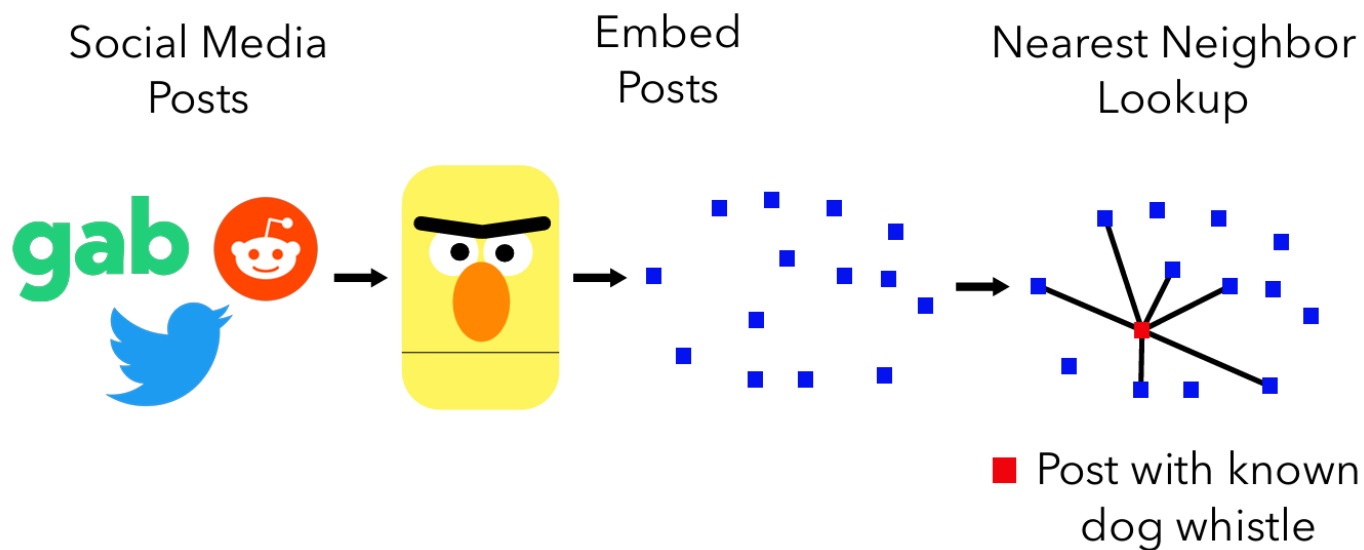
- ⊙ A novel task: Finding Emergent Dog Whistles in Common Habitats or FETCH!
- ⊙ Formulation
  - ◆ **Input:** a corpus and a set of initial seed dog whistles
  - ◆ **Output:** dog whistles which are semantically similar to the seed dog whistles
  - ◆ **Task goal:** use the initial seed words and the corpus to discover other known dog whistles called the emergent dog whistles
- ⊙ Metrics
  - ◆ We use precision, recall and F-score to measure the performance of different methods for our task.
  - ◆ Precision is computed as normal, but we modify recall to normalize by **only the positive terms that exist in the corpus** (Data Potential Recall, DPR)



# Dog Whistle: Method

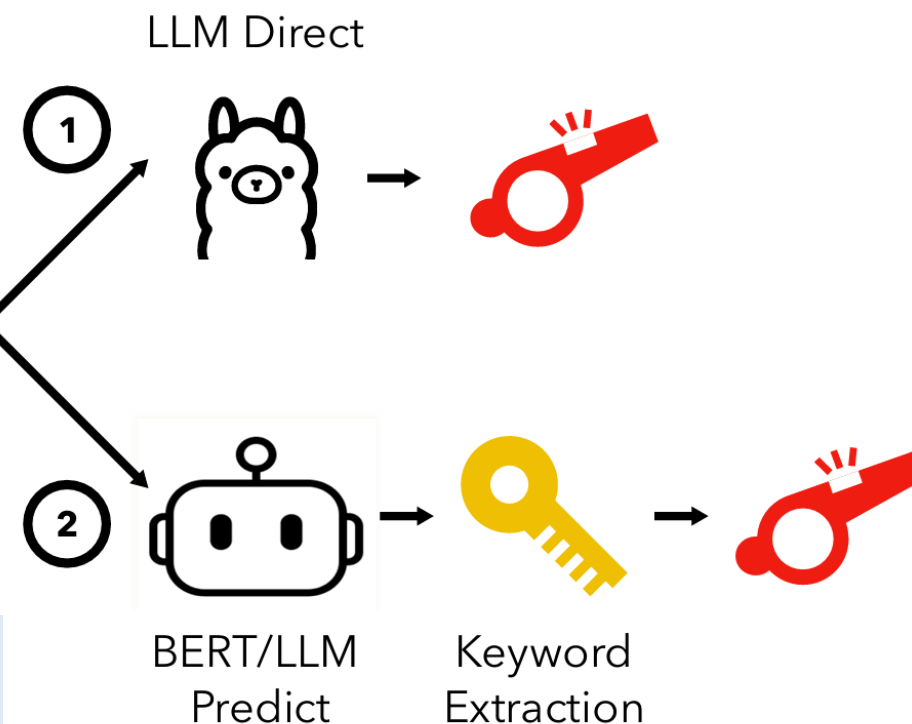
## ◉ EarShot

Step1: all posts in the corpora are turned into vectors using a sentence transformer



Step2: find the closest vector to each seed post vector (to capture posts that are likely related in meaning or intent)

Step3-1: pass all the posts to a LLM to extract the dog whistles from the posts.



Step3-2: pass potentially dog whistle contained posts for a keyword extractors

# Dog Whistle: Experiments

## ◉ EarShot (F<sub>0.5</sub> Score)

- ✓ **Synthetic (Reddit)** refers to **an idealized setting** where every post contains an associated dog whistle
- ✓ **Balanced (Gab)** represents a context with **a higher-than-average prevalence** of dog whistles
- ✓ **Realistic (Twitter)** reflects **a typical online environment** where dog whistles are sparse and most are benign

Method	Synthetic	Balanced	Realistic
Word2Vec/Phrase2Vec	5.91	1.86	2.47
EPDSpanBERT	2.03	0.00	0.61
MLMBERTweet	1.14	0.28	0.58
EarShot-Predict	14.55	5.70	4.63
EarShot-Direct	23.29	3.52	0.00

# Dog Whistle: Analysis

## ◉ EarShot (F<sub>0.5</sub> Score)

- **Synthetic:** Word2Vec model finds very **simplistic dog whistles** like *pepe* (另类右翼), *autogynophilia* (跨性别变态), *13/50* (黑人犯罪), *npc* (没有灵魂的工具人), while our method is able to capture **more complex phraseology** like *middle class* (被当作选票对象的特定收入群体), *climate alarmists* (小题大做、夸大气候危机)
- **Realistic:** Earshot achieves **high precision**. It identifies only **explicit terms** like illegal aliens (非法移民) while Word2Vec found some, like pedophilia (原指恋童癖, 引申为极度负面的指控、抹黑), but struggles with **false positives**.
- **Balanced:** Word2Vec identifies **subtler and novel dog whistles** (e.g., *gibsmedats* (丑化黑人群体贪婪、依赖福利), *we wuz kangs* (暗示黑人夸大历史成就)), while our method focuses on explicit slurs like *kike* (反犹主义), *faggotry* (贬低 LGBTQ+ 群体)

# Dog Whistle: Challenging Ones and Future

## ⦿ Challenges

- ◆ First, **emoji-based** dog whistles were difficult to find
- ◆ Second, **context dependency** and **deliberate ambiguity** posed significant challenges
- ◆ Third, **recency** presented an issue

## ⦿ Future

- ◆ First, **complex**, **nuanced**, and **recent** phrases frequently escape discovery
- ◆ Second, the dependence on hate speech classifiers tends to prioritize explicit slurs while **overlooking subtler forms of harmful language**
- ◆ Third, the inherent **precision-recall tradeoff** presents substantial challenges

# **WATCHED: A Web AI Agent Tool for Combating Hate Speech by Expanding Data**



清华大学  
Tsinghua University



# Slang: Background

## ◎ Slang

- ◎ a common type of informal language that is ubiquitous across day-to-day conversations

### 1. Slang detection

Good choice, that jacket is excellent. ✗  
Good choice, that jacket is **blazing**. ✓

### 2. Slang source identification

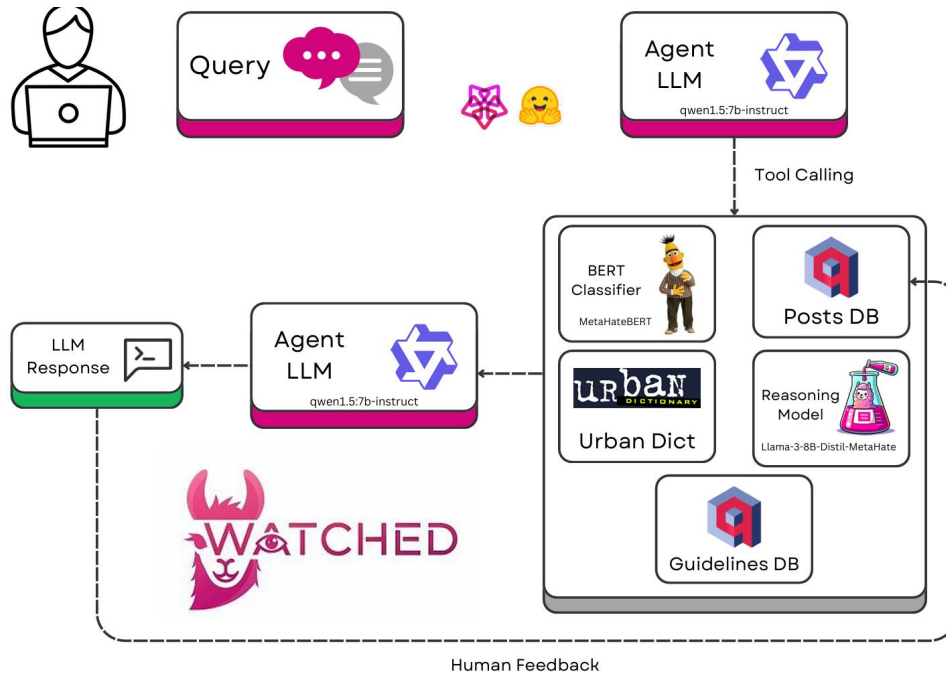
Good choice, that jacket is **blazing**.  
1950 1960 1970 1980 1990 2000+  
✗ ✓

- **nigger** / **nigga** —— 针对黑人的最常见仇恨词
- **chink** —— 辱骂华裔或东亚人
- **gook** —— 辱骂东南亚人（尤其越南）
- **spic** —— 针对拉丁裔/西班牙裔
- **paki** —— 在英国语境中贬低巴基斯坦裔或南亚人

Existing systems struggle with implicit language, slang, sarcasm, and context, and often lack interpretability

# Slang: Method

## ◉ Workflow

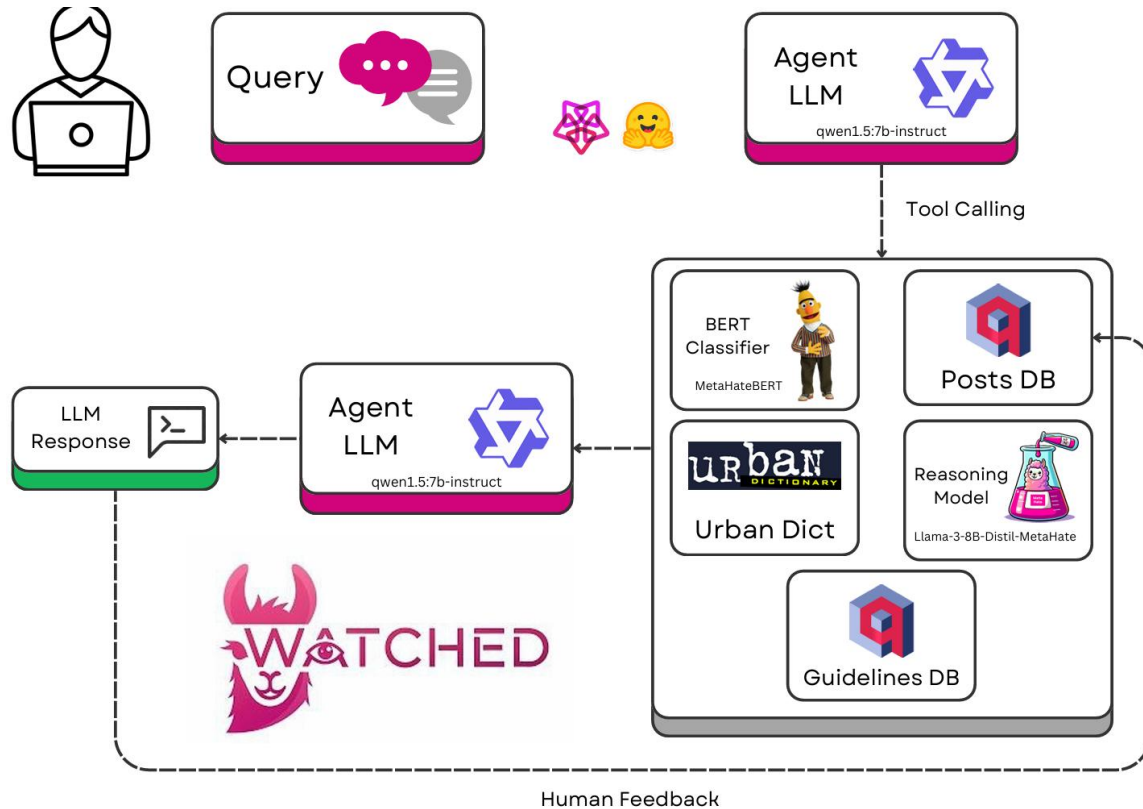


we use the *pydantic ai* framework, defining an Agent, configured with a custom system prompt.

- **Invoke external tools and APIs:** call external functions or APIs when necessary to overcome LLM limitations
- **Incorporate external results into reasoning:** The retrieved information is integrated into the model's reasoning process,
- **Iteratively refine prompts:** Agents continuously adjust and improve their prompts based on new information to enhance task performance.
- **Adapt behavior dynamically:** By responding to feedback and context, agents adjust their reasoning and actions over time for better adaptability.

# Slang: Method

## Components



- **Hate Speech Classifier:** MetaHateBERT
- **Similar Posts:** Agentic Retrieval-Augmented Generation (RAG) system to **fetch similar examples to the users' input query**
- **Urban Dictionary:** invokes **a distilled large language model** fine tuned for this task to support **reasoning for hate speech classification**
- **Social Network Guidelines:** content moderation guidelines from major social networks, **supplementing the reasoning models**



# Slang: Results

	F1	F1 <sub>MICRO</sub>	F1 <sub>MACRO</sub>
Llama 3 8b ( <i>Few-Shot CoT</i> )	0.8219	0.8221	0.8152
Llama 3 70b ( <i>Few-Shot CoT</i> )	0.8542	0.8536	0.8496
Distil MetaHate [32]	0.8826	0.8816	0.8807
MetaHateBERT [31]	0.8836	0.8831	0.8801
Perspective API [44]	0.8061	0.8066	0.7985
<b>WATCHED</b>	<b>0.9168</b>	<b>0.9165</b>	<b>0.9139</b>

	F1	F1 <sub>MICRO</sub>	F1 <sub>MACRO</sub>
no tools	0.8148	0.8184	0.8053
w/o <i>Hate Speech Classifier</i>	0.8608	0.8598	0.8571
w/o <i>Similar Posts</i>	0.8786	0.8789	0.8733
w/o <i>Urban Dictionary</i>	0.8784	0.8780	0.8753
w/o <i>Reasoning</i>	0.8887	0.8882	0.8853
<b>WATCHED</b>	<b>0.9168</b>	<b>0.9165</b>	<b>0.9139</b>

- The removal of **similar posts** and **urban dictionary** also leads to noticeable performance drops, highlighting the importance of **contextual grounding** and **slang understanding**.

# Slang: Case

## IRLab - Hate Speech Detection

Chatbot

all immigrants are a burden to our society!!! white power

*Hate Speech:* True

*Confidence:* 0.99

*Explanation:* The post contains phrases that promote harmful stereotypes against immigrants and the term 'white power,' which has historically been associated with promoting racial superiority, discrimination, and segregation. These elements align with international definitions of hate speech as described in the UN guidelines. Specifically, such statements have been linked to fostering negative perceptions and can lead to discrimination, hostility, or violence.

- *Guidelines summary:* This chunk focuses on the vulnerability of migrants and refugees to racism, discrimination, and hate rhetoric due to increasing numbers and the strain they place on local communities, which has intensified public debate.  
*Guidelines URL:* <https://www.un.org/en/hate-speech/impact-and-prevention/targets-of-hate>
- *Guidelines summary:* Hate speech is legally defined as advocacy of national, racial or religious hatred that incites discrimination, hostility, or violence.  
*Guidelines URL:* <https://www.un.org/en/hate-speech/understanding-hate-speech/faqs>

Input

all immigrants are a burden to our society!!! white power

Send

✔ Correct

✘ Incorrect

- **Human-in-the-loop feedback** supports studies on the evolution of hateful discourse, the effectiveness of interventions
- AI adaptability to **new linguistic and cultural trends**

# Understanding Gen Alpha's Digital Language: Evaluation of LLM Safety Systems for Content Moderation



清华大学  
Tsinghua University



# Gen Alpha's Digital Language: Background

## ◉ Gen Alpha (Generation Alpha)

- ◆ refers to the generation of people born roughly between **2010 - 2025**. They are the first generation to grow up entirely in a **digital world**, often described as “digital natives.”

Platform	Category	Examples
Social Media	Status expressions	<i>"in my flop era"</i>
	Self-referential	<i>"having my glow up"</i>
	Behavioral	<i>"gaslight gatekeep girlboss"</i>
Gaming	Achievement	<i>"secured the bag"</i>
	Performance	<i>"ate that up"</i>
	Social dynamics	<i>"got ratioed"</i>
Video	Quality descriptors	<i>"fire", "hits different"</i>
	Reaction phrases	<i>"And I oop"</i>
	Trend markers	<i>"English or spanish"</i>

- Gen Alpha experiences a fundamental **disconnect between their communication patterns and traditional protection mechanisms**

# Gen Alpha's Digital Language: Challenges

## ◉ Gen Alpha (Generation Alpha)

- ◆ **Digital Immersion Vulnerability:** Gen Alpha's deep digital immersion makes them vulnerable to manipulation, as their digital skills often **prevents them from seeking adult help**.
- ◆ **Moderation Gap:** A moderation gap arises because **adults and systems struggle to keep pace with Gen Alpha's fast-evolving language**, leaving harmful interactions unnoticed.
- ◆ **AI Safety Limitations:** AI moderation systems **fail to fully grasp Gen Alpha's unique communication**, creating safety risks that neither humans nor AI can reliably address.

Research goal: Understanding how AI systems interpret and moderate Gen Alpha's unique communication patterns

# Gen Alpha's Digital Language: Dataset

## ◉ Dataset

- ◆ Systematic observation of digital platforms where Gen Alpha users frequently interaction
- ◆ Our dataset comprises 100 contemporary Gen Alpha expressions:

Type	Expression	Dimensional Analysis
Context Dependent	"let him cook"	<b>Platform:</b> Gaming (skill praise), Social (general encouragement), Video (entertainment value) <b>Safety:</b> Context determines supportive vs. mocking intent <b>Evolution:</b> Shifted from cooking reference to performance evaluation
Masked Harassment	"are you fr"	<b>Platform:</b> Similar usage across platforms but with varied intensity <b>Safety:</b> Provides plausible deniability in bullying interactions <b>Evolution:</b> From genuine question to dismissive response
Evolution Based	"sigma"	<b>Platform:</b> Primarily gaming, spreading to broader social contexts <b>Safety:</b> Growing negative connotations in specific communities <b>Evolution:</b> Personality type → gaming skill → discriminatory marker

- **Context (Platform) Awareness** focuses on how meaning changes across platforms, e.g., gaming, social media.
- **Safety Awareness** identifies potentially harmful usage patterns.
- **Evolution Awareness** tracks how expressions evolve across digital spaces.

# Gen Alpha's Digital Language: Evaluation

## ○ Evaluation

Component	Parameters	Implementation
Participant Groups	<ul style="list-style-type: none"><li>• Gen Alpha (11-14y)</li><li>• Human moderators/parents</li><li>• LLMs (4)</li></ul>	Balanced representation across demographics and expertise levels
Evaluation Tasks	<ul style="list-style-type: none"><li>• Basic understanding</li><li>• Platform (context)</li><li>• Safety</li></ul>	Each expression evaluated across all three dimensions with multiple raters
Assessment Metrics	<ul style="list-style-type: none"><li>• Meaning recognition</li><li>• Platform recognition</li><li>• Safety (risk) identification</li></ul>	Quantitative scoring with qualitative validation through expert review
Model Testing	<ul style="list-style-type: none"><li>• Standardized prompt template</li><li>• Multiple evaluation contexts</li><li>• Cross-validator raters</li></ul>	Statistical significance evaluated using chi-square tests ( $\alpha = 0.05$ )

GPT-4, Claude, Gemini, and Llama 3

Group	Basic	Context	Safety
Gen Alpha	98.0	96.0	92.0
Parents	68.0	42.0	35.0
Moderators	72.0	45.0	38.0
GPT-4	64.2	52.3	38.4
Claude	68.1	56.2	42.3
Gemini	62.4	48.7	36.2
Llama 3	58.3	42.1	32.5



- Gen Alpha mastery
- Adult comprehension limitations
- AI systems demonstrated comparable performance show potential complementary roles in content moderation.

- **Basic Understanding:** Explain meaning in online communication
- **Context Awareness:** Explain how meaning might vary across platforms
- **Safety Recognition:** Identify potential safety concerns



# Gen Alpha’s Digital Language: Discussion

- How implicit the corresponding risks are in a given Gen Alpha expression

Risk Type	Sub-Category	Example Expression	Content Moderation Implications
Direct Risks	Overt bullying	“you’re such a pick me”	Explicit social exclusion tactics
	Direct exclusion	“NPC behavior”	Dehumanizing language patterns
	Status attacks	“beta male behavior”	Hierarchy-based harassment
Masked Risks	Coded harassment	“is it acoustic”	Discriminatory content masked as inquiry
	Subtle manipulation	“let him cook” (mocking)	Context-dependent negativity
	Social pressure	“you’re not him”	Peer pressure through trending phrases
Grooming Indicators	Trust building	“you passed the vibe check”	False sense of security creation
	Isolation attempts	“they’re not giving”	Social separation tactics
	Secret-keeping	“keep it lowkey”	Privacy boundary manipulation

- Direct Risks:** Overtly harmful content such as bullying, threats, or hate speech with explicit negative intent.
- Masked Risks:** Harassment or exclusion concealed through slang, memes, or humor; requires contextual inference
- Grooming Indicators:** Expressions that gain harmful connotation over time or through platform migration.

Model	Direct Risks	Masked Risks	Evolution-Based
GPT-4	72.4	45.6	38.2
Claude	75.8	48.9	41.5
Gemini	69.3	42.8	35.7
Llama 3	65.7	38.4	32.3

The evolving digital languages are much challenging.



# Conclusion

## ◉ Typical forms of Euphemism to express toxicity

Types	Definition	Examples
Jargon	Seemingly neutral terms with hidden harmful intent	Rat--Remote access trojan
Dog whistle	One phrase, two audiences — neutral outwardly, inflammatory inwardly	cosmopolitan
Slang	informal language common in everyday talk	nigger
Digital language	online-born expressions shaped by internet culture	in my flop era

## ◉ Jargon vs. Dog whistle

- ◆ 相似点：都有“圈内人懂、圈外人不懂”的特征
- ◆ 不同点：Jargon有很多中性的用法，用于专业沟通；dog whistle 则主要是隐含政治/社会意义的表达

## ◉ Slang vs. Digital language

- ◆ 相似点：都是非正式语言，很多网络用语就是slang
- ◆ 不同点：digital language更依赖于互联网环境和数字媒介，不仅是文本，还可能有表情符等

# Conclusion

---

## ◎ Characteristics of Euphemism

- ◆ **Surface Neutrality** (字面中性) : Euphemisms **appear harmless or neutral** on the surface, reducing direct offensiveness.
- ◆ **Diverse Forms** (表达形式多): They can take **multiple forms**, including abbreviations, memes, or emojis.
- ◆ **Knowledge Dependence** (依赖特定知识或背景): Their meaning often relies on **shared knowledge, cultural background, or insider understanding**.
- ◆ **Dynamic Evolution** (动态演进) : Euphemisms constantly **shift and adapt over time**, especially when older terms are exposed or lose effectiveness.

# Conclusion

---

## ◉ Challenges to identify implicit toxicity expressed with euphemism

- ◆ **Detection Difficulty** (Surface Harmlessness): Euphemisms often look neutral or benign, making them **hard to flag with standard moderation filters**.
- ◆ **Coverage Limitation** (Form Diversity): They appear in many forms like misspellings, abbreviations, memes, emojis, creating **a wide variety of expressions to track**.
- ◆ **Interpretation Ambiguity** (Context Dependence): Their meaning depends heavily on cultural background, shared knowledge, or specific community context, which can **confuse moderators and AI systems**.
- ◆ **Adaptation Lag** (Dynamic Evolution): Euphemisms evolve quickly as users invent new terms to evade detection, **outpacing the update cycles of moderation tools**.

**Thanks for your attention!**



清華大學  
Tsinghua University



1. **USENIX18: Reading Thieves' Cant: Automatically Identifying and Understanding Dark Jargons from Cybercrime Marketplaces**
2. **USENIX24: Investigating Moderation Challenges to Combating Hate and Harassment: The Case of Mod-Admin Power Dynamics and Feature Misuse on Reddit**
3. **ACL 25: Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats**
4. **ACL 24: Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles**
5. **ACL 25 Findings: STATE ToxiCN: A Benchmark for Span-level Target-Aware Toxicity Extraction in Chinese Hate Speech Detection**
6. **ACL 23: From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models**
7. ChineseHarm-Bench: A Chinese Harmful Content Detection Benchmark
8. <https://swu-union.org.uk/wp-content/uploads/SWU-blog-series-Dog-Whistles-2023.pdf>
9. Understanding Gen Alpha's Digital Language: Evaluation of LLM Safety Systems for Content Moderation  
<https://arxiv.org/pdf/2505.10588>
10. Toward Informal Language Processing: Knowledge of Slang in Large Language Models. NAACL-2024
11. **Covering Cracks in Content Moderation: Delexicalized Distant Supervision for Illicit Drug Jargon Detection (KDD)**
12. Watch Your Language: Investigating Content Moderation with Large Language Models (<https://arxiv.org/pdf/2309.14517>)
13. WATCHED: A Web AI Agent Tool for Combating Hate Speech by Expanding Data (<https://arxiv.org/pdf/2509.01379>)
14. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. ACL 2023
15. ACL 23: Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks