# Keyword-Oriented Multimodal Modeling for Euphemism Identification

**Anonymous ACL submission**

## Abstract

Euphemism identification aims to identify the true meaning of a given euphemism, such as identifying "weed" (euphemism) as "marijuana" (target keyword) in illicit transactions, which is of great significance to help content moderation and combat underground market. However, existing methods only use text data to identify euphemisms, ignoring the semantic information of other modalities associated with the corresponding target keywords during the development and evolution of euphemisms. Additionally, the lack of multimodal datasets of euphemisms also hinders related research. In this paper, we regard euphemisms and their corresponding target keywords as keywords and propose improving euphemism identification quality through keyword-oriented visual and audio features. To this end, we first introduce a keyword-oriented multimodal corpus of euphemisms (KOM-Euph), involving three datasets (Drug, Weapon, and Sexuality), including text, images, and speech. Then, we propose a keyword-oriented multimodal euphemism identification method (KOM-EI), which uses cross-modal feature alignment and dynamic fusion modules to explicitly utilize the visual and audio features of the keywords for efficient euphemism identification. Extensive experiments demonstrate that our method outperforms the SOTA models and LLMs, and show the importance of our multimodal datasets.

## 1 Introduction

Euphemisms are indirect words or phrases used to replace harsh or offensive expressions and are a significant form of linguistic communication. Currently, euphemisms are widely used in social media and darknet marketplaces to cover up illicit transactions and evade supervision (Yuan et al., 2018; HADA et al., 2020; Foye et al., 2021). For instance, the euphemisms "ice" and "weed" in Table 1 are used as substitutes for the target keywords "methamphetamine" and "marijuana". These euphemisms can seem vague and obscure, making it challenging to trace illegal transactions. Thus, identifying the target keyword of a given euphemism i.e., euphemism identification, is essential for improving content moderation and combatting underground trading. However, euphemisms evolve like a "treadmill" (Pinker, 2003), making it difficult to maintain an up-to-date corpus for the euphemism identification task. Furthermore, the euphemisms are used either in literal or figurative senses, which adds complexity to the task.

| **Example sentences** (euphemisms are in bold) |
|---|
| 1. We had already paid $70 for some shitty **weed** from a taxi driver but we were interested in some **coke** and the cubans. |
| 2. For all vendors of **ice**, it seems pretty obvious that it is not as pure as they market it. |
| 3. Back up before I pull my **nine** on you. |

Table 1: Examples of sentences containing euphemisms.

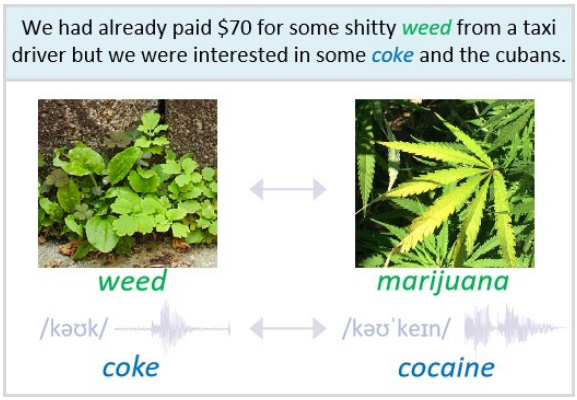



Figure 1: Image and speech examples of keywords.

Current methods have primarily focused on detecting whether words are used in a euphemistic sense, with techniques evolving from conventional natural language processing (Yuan et al., 2018;

Magu and Luo, 2018; Lee et al., 2022) to deep learning pre-training models (Zhu et al., 2021; Zhu and Bhat, 2021; Seethappan and Premalatha, 2022). However, these methods can only detect euphemisms but not identify them to the corresponding target keywords. Meanwhile, existing studies on euphemism identification used self-supervised schemes to construct labeled datasets for training a model to identify the euphemisms. They only focus on obtaining context information on the euphemisms from text data to identify them, disregarding the semantic information of other modalities associated with the corresponding target keywords in the evolution of euphemisms.

In the evolution of language, euphemisms usually evolve from homophones, abbreviations, image mapping, etc. of the target keywords (Ji and Knight, 2018). As shown in Figure 1, the literal meaning of weed and its true meaning referring to marijuana are both plants, which can be seen from the visual information. Coke is a euphemism for cocaine because the original coke is a drink containing cocaine, and the sound of coke is similar to cocaine. This can be seen from the pronunciation wave through the audio information of them. Text is just a modality for recording language, and the visual and audio modalities can also record extra information for language. These multimodalities together demonstrate the development and evolution of language. Furthermore, leveraging other modalities to introduce salient information that complements text has been proven effective in other natural language processing (NLP) tasks (Yang et al., 2023; Zeng et al., 2023; Wang et al., 2022). Thus, the integration of multimodal data is urgently needed for euphemism identification. However, current research on euphemism only involves a single text modality, and the lack of multimodal data hinders the relevant research.

To overcome these limitations, we construct the first **K**eyword-**O**riented **M**ultimodal **Euph**emism datasets (KOM-Euph) based on the only text modal datasets proposed by Zhu et al. (2021). The KOM-Euph is composed of text-image-speech pairs with no labels. KOM-Euph will expand euphemism understanding from mono-modality to multi-modality and help to improve the performance of automatic euphemism identification by investigating multimodal semantics. Additionally, to better utilize the multimodal information of euphemisms from multi-view of text, vision, and audio, we propose a **K**eyword-**O**riented **M**ultimodal **E**uphemism **I**dentification method (KOM-EI) to generate more comprehensive semantics of euphemisms by explicitly using the visual and audio features. The KOM-EI model employs feature alignment to align cross-modal features through contrastive learning and utilizes dynamic feature fusion to dynamically obtain cross-modal features by cross-attention and gated units. In this way, the model is enhanced to explicitly exploit the text, vision, and audio features, leading to more accurate identification. Experiments show that our method yields top1 identification accuracies that are 45-60% higher than the state-of-the-art baseline methods.

Our contributions are as follows:

- To the best of our knowledge, we are the first that contribute a novel keyword-oriented multimodal euphemism corpus (KOM-Euph) with 86K text-image-speech pairs involving three domains.
- We propose a keyword-oriented multimodal method, using cross-modal feature alignment and dynamic fusion to explicitly exploit the text-image-speech features to identify euphemisms.
- Extensive experiments on KOM-Euph show that our model builds new state-of-the-art performance that beats large language models and demonstrates the importance of our datasets.

## 2 Related work

### 2.1 Euphemism Identification

Existing models mainly focused on detecting words in a euphemistic manner, using methods from conventional NLP techniques (Magu and Luo, 2018; Felt and Riloff, 2020), deep learning methods (Yuan et al., 2018; Gavidia et al., 2022) to pre-trained models (Zhu et al., 2021; Ke et al., 2022). Yuan et al. (2018) focused on identifying the hypernyms of euphemisms while not directly identifying the specific meanings of euphemisms. They identify "horse" as an illicit drug rather than "heroin". Zhu et al. (2021) first explicitly defined the euphemism identification task, they developed a self-supervised scheme and analyzed euphemisms at the sentence level to identify them. However, they only focused on obtaining context information of text data to identify euphemisms, ignoring other modality features of euphemisms, resulting in limited identification results. Unlike the above methods, we are the first to use multimodal information to identify euphemisms.

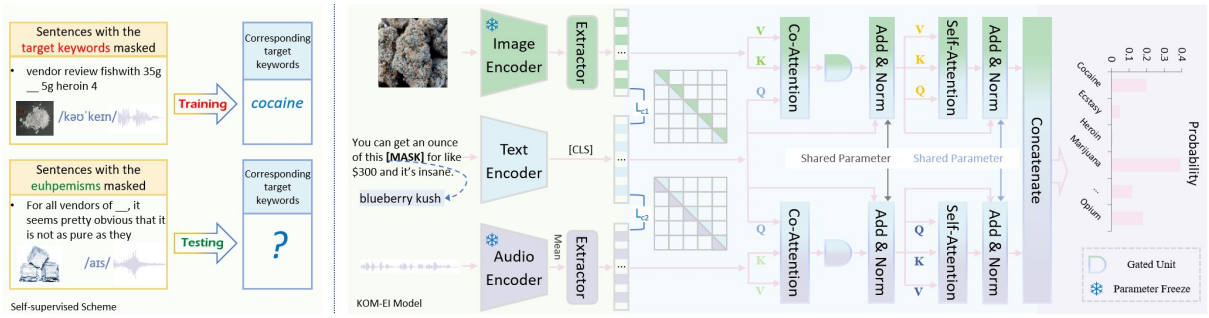



Figure 2: The overall framework. The left part shows the self-supervised learning scheme for constructing labeled training sets. The right part shows the architecture of our KOM-EI.

### 2.2 Multimodal learning

As information between modalities can complement each other, many NLP tasks extend from a single text modal to multimodal to enhance the understanding of particular tasks. Multimodal fake news detection achieves higher detection results with the help of textual information and image information complementing each other (Huang et al., 2023; Zeng et al., 2023). The multimodal sentiment analysis provides more comprehensive sentiment information by fusing feature information from three modalities: text, audio, and image (Yang et al., 2023; Wang et al., 2022). Kesen et al. (2022) utilized the text-image model to generate images corresponding to euphemisms and achieved higher euphemism detection results. Inspired by them, we propose to use visual and audio information in the euphemism identification task for the first time. The difference is that we build and introduce keyword-oriented multimodal information, which is a new attempt in the field of NLP.

## 3 Problem Description

The task studied in this article is that given sentences containing euphemisms $S$, a set of target keywords $T$, and the images and speech of the keywords as input: $s = [w_1, ..., w_i, euph, ..., w_m]$ (where $s \in Set$, $euph$ is a euphemism), $T = \{t_1, ..., t_j, ..., t_n\}$, determine the target keyword $t_j$ that refers to the euphemism $euph$. As seen in Table 1, we aim to determine that "ice" refers to "methamphetamine" and "nine" means "gun".

## 4 Methodology

Figure 2 shows the overall flow and framework of our proposed method. Inspired by Zhu et al. (2021), we use self-supervised learning to automatically construct labeled datasets, as shown in the left part of Figure 2. In the training and validation phases, they take sentences masking the target keywords (e.g., cocaine and heroin) as training samples, using the corresponding target keywords as labels for training. During the testing phase, they feed sentences with the euphemisms masked into the trained model and finally specify the masked euphemism into the corresponding target keyword. Different from the text-only approach of Zhu et al. (2021), our method enriches the training and validation phases with visual and audio information of target keywords, while integrating similar multimodal data of euphemisms during the testing phase.

Based on the above self-supervised scheme and multimodal information, we propose a keyword-oriented multimodal euphemism identification method (KOM-EI), as shown in the right of Figure 2, including three parts, namely 1) a feature representation module, 2) a dynamic feature fusion module, and 3) a prediction module. We first encode the text-image-speech pairs features respectively via three pre-trained models. Subsequently, they are channeled into a feature fusion module tailored to dynamically capture cross-modal congruities, yielding features enriched with multi-modality information. These enhanced features are then directed into the prediction module to facilitate the euphemism identification across modalities.

### 4.1 Feature Representation Module

Euphemisms are primarily discerned through contextual analysis. However, an exclusive focus on context may introduce ambiguity, given that analogous contexts for disparate euphemisms could misguide the model, culminating in misidentification. For instance, the sentence, "We had already paid $70 for some shitty weed from a taxi driver but we were interested in some coke and the cubans", contains both euphemisms "weed" and "coke". It is

difficult to distinguish between "weed" and "coke" if only sentence-level context is considered.

According to previous research on the evolution of euphemisms (Ji and Knight, 2018) and common sense, the visual and audio information corresponding to the literal meaning of a euphemism is related to its implicit meaning. For example, from the visual perspective, both the literal meaning of weed and its true meaning referring to marijuana are plants. Coke is a euphemism for cocaine because the original coke was a drink containing cocaine. From the audio perspective, the literal pronunciation of coke is close to that of cocaine. Motivated by these, we introduce multimodal information on euphemisms from multiple views of text, vision, and audio to obtain comprehensive and semantically rich features to better identify euphemisms.

**Text Encoder**. Due to Bert's success in extracting contextual features (Devlin et al., 2019), we use the Bert model pre-trained on euphemism corpus to extract dynamic context information. Take the sentence $s = [w_1, ..., w_i, \texttt{[MASK]}, ..., w_m]$ ($s \in Set$, where $Set$ is the set of masked sentences) with the euphemism masked as the input of BERT model. $w_i$ refers to a token, and the special tokens "[CLS]" and "[SEP]" are boundary markers used to guide and end the input. $w_{mask}$ refers to the original masked words. As shown in formula (1), $T \in R^{d_g}$ and $d_g$ is the dimension of text embedding.

$$T = \text{CLS\_BERT}(\texttt{[CLS]} + w_1 + ... + w_i + \texttt{[MASK]} + ... + \texttt{[SEP]}), \quad (1)$$

**Image Encoder**. To initialize our model with effective image embeddings, we utilize a pre-trained CLIP model (Radford et al., 2021) as the image encoder. CLIP has demonstrated remarkable capabilities in understanding images in the context of natural language, outperforming traditional image-only models in various tasks. For an image $\in R^{H \times W \times C}$, where H, W, and C denote the height, width, and number of channels of the image. To preserve the pre-trained knowledge of CLIP, we freeze its weights and add a nonlinear projection layer as an extractor. The image features can be represented as

$$\hat{I} = \text{CLIP}(Image), \quad (2)$$

$$I = \text{ReLU}(W_I \hat{I} + b_I), \quad (3)$$

where $\hat{I} \in R^{d_v}$ and $\text{I} \in R^{d_g}$, $d_v$ is the dimension of image embedding.

**Speech Encoder**. To train our model from a good start of speech embeddings, we employ Wav2Vec 2.0 (Baevski et al., 2020) as the speech encoder. Wav2Vec 2.0 is adept at capturing intricate acoustic patterns and nuances within the audio signal, yielding a comprehensive embedding that encapsulates the audio's characteristics. In this work, our speech inputs are files with the suffix '.wav'. Following the freezing of the wav2vec 2.0 model's weights, we similarly attach an extractor to further process the speech embeddings, which can be represented by

$$\tilde{S} = \text{Wav2Vec2}(Speech) = [z_1, z_2, ..., z_T], \quad (4)$$

$$\hat{S} = \text{Mean}(\tilde{S}), \quad (5)$$

$$S = \text{ReLU}(W_S \hat{S} + b_S), \quad (6)$$

where $z_j \in R^{d_s}$ refers to the representation of the j-th time-step and $d_s$ is the dimension of speech embedding. Mean$(\cdot)$ is the average function, $S \in R^{d_g}$ denote the final speech features.

### 4.2 Dynamic Feature Fusion Module

By introducing multimodal information, it can recognize euphemisms by leveraging additional cues in visual and audio modalities to assist language-based prediction. Although other modalities can provide extra information to aid identification, they also introduce noise due to the quality. Meanwhile, each text-image-speech pair is different and requires different points of attention. Therefore, we use text features as anchors and dynamically learn additional information from other modalities. First, cross-modal contrastive learning is employed to align the text-image and text-speech features. Subsequent cross-attention facilitates the complementary feature extraction across modalities. Ultimately, a gated unit is deployed to filter redundant and noisy information from the visual and audio features, dynamically refining the fused features.

**Cross-modal Feature Alignment (CFA)**. Existing work exhibits the generality of the modality gap phenomenon in multimodal models (Xu et al., 2021; Zhang et al., 2022; Liang et al., 2022). To this end, we use the CFA to align cross-modal features to mitigate the modal gaps. Given a sentence containing a keyword, align it with the image and audio features of the keyword respectively. Where the text involving the same keyword and the image or audio corresponding to the keyword are positive samples, and those corresponding to different keywords are negative samples. We formulate the

cross-modal contrastive loss as:

$$L_{TI} = -\sum_{i=1}^{|Set|}\sum_{j=1}^{|B|}\mathbb{I}([mask]_i = keyword_j) \log\frac{e^{sim(T_i,I_j)/\tau}}{\sum_{k=1}^{|B|}e^{sim(T_i,I_k)/\tau}}, \quad (7)$$

$$L_{TS} = -\sum_{i=1}^{|Set|}\sum_{j=1}^{|B|}\mathbb{I}([mask]_i = keyword_j) \log\frac{e^{sim(T_i,S_j)/\tau}}{\sum_{k=1}^{|B|}e^{sim(T_i,S_k)/\tau}}, \quad (8)$$

where $|B|$ is the batch size, $\mathbb{I}$ is an indicator, $[mask]_i$ refers to the keyword in $s$ and $keyword_j$ means the keyword corresponding to the image or speech. $sim(h_i, h_j)$ is the cosine similarity $\frac{h_i^T \cdot h_j}{|h_i||h_j|}$ and $\tau$ is the temperature hyper-parameter.

**Cross-modal Attention(CA)**. To better obtain supplementary information from other modalities, we use contextual features as anchors and use cross-attention to focus on relevant information. Firstly, the query Q is linearly projected from the textual feature $T$, and the key $K$ and value $V$ are linearly projected from the visual features $I$ or the audio feature $S$. $Q = TW_q, K = IW_k/SW_k, V = IW_v/SW_v, Q/K/V \in R^{d_g}$. Then, the CA is applied to get the context-queried visual features $M_{TI}$ and the context-queried audio features $M_{TS}$.

$$M_{TI} = \text{CA}(Q_{TI}, K_{TI}, V_{TI}), \quad M_{TS} = \text{CA}(Q_{TS}, K_{TS}, V_{TS}), \quad (9)$$

**Gated Unit (GU)**. The GU is employed to filter redundant and noisy information from the visual or audio features. It aims to learn dynamic co-attention of text-image and text-speech conditioned on different inputs. We then obtain the text-guided output $\hat{M_{TI}}$ and $\hat{M_{TS}}$ followed by an Addition and Normalization layer ($\text{AN}_{\text{GU}}$) :

$$\text{R}(X) = \text{ReLU}(W_R X + b_R), \quad \text{GU}(X) = \sigma(W_G \text{R}(X) + b_G) \cdot X, \quad (10)$$

$$\tilde{M_{TI}} = \text{GU}(M_{TI}),\ \tilde{M_{TS}} = \text{GU}(M_{TS}), \quad (11)$$

$$\hat{M_{TI}} = \text{AN}_{\text{GU}}(\tilde{M_{TI}} + M_{TI}), \quad \hat{M_{TS}} = \text{AN}_{\text{GU}}(\tilde{M_{TS}} + M_{TS}). \quad (12)$$

Next, we employ a **Self-Attention (SA)** layer followed by an AN layer $\text{AN}_{\text{SA}}$ to refine the text-guided output $\hat{M_{TI}}$. $\hat{Q_{TI}} = \hat{M_{TI}}W_{q^{TI}}, \hat{K_{TI}} = \hat{M_{TI}}W_{k^{TI}}, \hat{V_{TI}} = \hat{M_{TI}}W_{v^{TI}}$.

$$\widehat{M_{TI}} = \text{SA}(\hat{Q_{TI}}, \hat{K_{TI}}, \hat{V_{TI}}), \quad (13)$$

$$\overline{M_{TI}} = \text{AN}_{\text{SA}}(\widehat{M_{TI}} + \hat{M_{TI}}). \quad (14)$$

Similarly, we can get the enhanced features $\overline{M_{TS}}$. Finally, the dynamic fusion features are obtained as follows:

$$\text{H}(s) = W_H(\overline{M_{TI}}; \overline{M_{TS}})) + b_H, \quad (15)$$

where $W_H \in R^{d_g \times 2d_g}$, $b_H \in R^{d_g}$ are the model parameters, and (;) means concatenation.

### 4.3 Prediction Module

After obtaining the dynamic fusion feature H(s), the identification task is finally achieved through the classifier. The probability of obtaining the selected target keyword for a given mask sentence is calculated by

$$P(t_j|s) = \text{softmax}(W(h(t_j) \odot \text{H}(s)) + b), \quad (16)$$

where $W \in R^{d_g}$, $b \in R$ are the model parameters and $\odot$ is the element-wise multiplication. $h(t_j)$ is the learned discrete representation of the class label of the target keyword. The objective of the training is to minimize the cross entropy between the predicted results and true values:

$$L_P = -\sum_{j=1}^{n} H_g log P(t_j|s), \quad (17)$$

where n is the number of target keyword subcategories in a specific category. In drug, weapon, or sexuality category, target keywords in the same subcategory hold identical meanings. $H_g$ is the one-hot vector of the ground truth.

### 4.4 Training and Inference

For multimodal euphemism identification, with the main prediction task and the cross-modal feature alignment auxiliary tasks, the training objective is finally formulated as:

$$J = \alpha L_P + \beta L_{TI} + \gamma L_{TS}, \quad (18)$$

where $\alpha, \beta$ and $\gamma$ are the balancing factors for the tradeoff among $L_P, L_{TI}$and $L_{TS}$ respectively.

During inference, the alignment auxiliary tasks are not involved and only the main prediction task is used to identify the euphemisms.

## 5 Experiments and Analysis

In this section, we evaluate the performance of KOM-EI on the KOM-Euph corpus and compare it with a set of baseline models.

| | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Text | **Sentence:** i know *weed* like i know dale carnegies how to win friends and influence people<br>**Literal / Implicit Meaning:** weed / marijuana \| **Drug** | **Sentence:** 1 requires *sig* for some reason but it is not incriminating<br>**Literal / Implicit Meaning:** sig / pistol \| **Weapon** | **Sentence**: bonos gaykk will spread a message of white pride and *buttlove*<br>**Literal / Implicit Meaning:** buttlove / sex \| **Sexuality** |
| Image |  |  | The collection of keywords in Sexuality is illegal. |
| Speech |  |  |  |

Figure 3: Samples presentation of multimodal data sets.

| Datasets | Entries | Images | Speech | Pairs | Num |
|---|---|---|---|---|---|
| Drug | 1271907 | 8452 | 2113 | 16060 | 33 |
| Weapon | 3108988 | 12636 | 3159 | 58410 | 9 |
| Sexuality | 2894869 | - | 1282 | 11465 | 12 |

Table 2: Overview of the datasets. Num means categories of target keywords. Pairs means the text-image-speech pairs.

### 5.1 KOM-Euph Dataset

The evolution of euphemisms draws inspiration from visual and audio information of the target keywords and no image or speech information exists for euphemism identification.

**Data Construction**. Inspired by the evolution of euphemisms, we construct a **K**eyword-**O**riented **M**ultimodal **Euph**emism (KOM-Euph) corpus, based on the text-only corpus (noted as "Euph") presented by Zhu et al. (2021). The Euph corpus consists of only textual data, involving three datasets: Drug, Weapon, and Sexuality. Since the Sexuality dataset mainly involves private parts of the body or sexual activities, the collection of their images is illegal. For the Sexuality dataset, we only expanded the audio modality data. For details on the data source and construction, please refer to Appendix A.

**Dataset Statistics**. An overview of each dataset is shown in Table 2. There are 33, 9, and 12 subcategories of target keywords corresponding to datasets Drug, Weapon, and Sexuality. As shown in Figure 3, Drug and Weapon datasets contain Text-Image-Speech pairs, while Sexuality contains Text-Speech pairs. As with existing methods, we employ a self-supervised learning framework to construct labeled data for training. We require three kinds of inputs: 1) sentences from the original text corpus that mask out the target keywords (for training/validation) and the corresponding images and speech, 2) sentences that mask out the euphemisms (for testing) and the corresponding images and speech, and 3) a list of target keywords (e.g., heroin, cocaine, etc.). Furthermore, To evaluate our results, we need to rely on a ground truth list (Zhu et al., 2021) of euphemisms and the corresponding target keywords, which should contain a one-to-one mapping from each euphemism to its true meaning. Please refer to Appendix A for ground truth list details.

Note that no extra supervision or resource except the images and speech of the keywords are required throughout the training process, and the ground truth lists do not participate in the whole training process but are only used to help evaluate the accuracy of euphemism identification.

### 5.2 Experimental Setup

#### 5.2.1 Baselines

**Text only Models**. We use four text-only baselines, including the method proposed by Zhu et al. (2021) (the SOTA model, denoted as "SelfEDI"), the Word2vec baseline they created, and the other two baselines established by us.

- **Word2vec**: Use Word2vec to obtain word embeddings of all words, using cosine similarity to select the closest target keyword.
- **SelfEDI**: Use a bag-of-words model to extract sentence features and train a classifier to recognize euphemisms.
- **BERT_pre**: Use the pre-trained model obtained on a specific corpus to extract the sentence features (fixed parameters), and train a classifier to recognize euphemisms.
- **BERT_ft**: Use the pre-trained model obtained on a specific corpus to extract the sentence features (updatable parameters), and fine tune the euphemism identification.

**Multimodal Models**. Since we are the first to propose a multimodal euphemism identification method, we establish three multimodal baselines based on our KOM-EI.

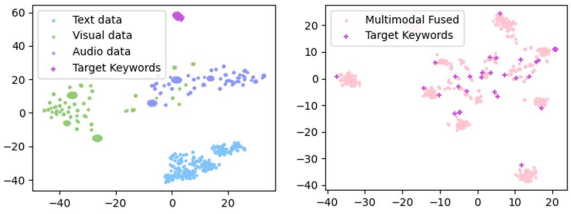
- **KOM-EI_V|VG**: Use VIT as image encoder and VGGish as speech encoder.
- **KOM-EI_V|W**: Use VIT as image encoder and Wav2Vec 2.0 as speech encoder.

| | Drug | | | Weapon | | | Sexuality | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Acc@1** | **Acc@2** | **Acc@3** | **Acc@1** | **Acc@2** | **Acc@3** | **Acc@1** | **Acc@2** | **Acc@3** |
| **Word2Vec** | 0.07 | 0.14 | 0.21 | 0.10 | 0.27 | 0.40 | 0.17 | 0.22 | 0.42 |
| **SelfEDI** | 0.20 | 0.31 | 0.38 | 0.33 | 0.51 | 0.67 | 0.32 | 0.55 | 0.64 |
| **BERT_pre** | 0.21 | 0.26 | 0.33 | 0.35 | 0.54 | 0.70 | 0.36 | 0.55 | 0.64 |
| **BERT_ft** | 0.24 | 0.31 | 0.40 | 0.38 | 0.55 | 0.73 | 0.38 | 0.50 | 0.69 |
| **KOM-EI_V\|VG** | 0.28 | 0.37 | 0.42 | 0.43 | 0.58 | 0.66 | 0.45 | 0.64 | 0.64 |
| **KOM-EI_V\|W** | 0.28 | 0.35 | 0.43 | 0.45 | 0.63 | 0.69 | 0.50 | 0.67 | 0.75 |
| **KOM-EI_C\|VG** | 0.30 | 0.21 | 0.32 | 0.47 | 0.63 | 0.71 | 0.45 | 0.64 | 0.64 |
| **KOM-EI** | **0.32** | **0.40** | **0.48** | **0.48** | **0.68** | **0.74** | **0.50** | **0.67** | **0.75** |

Table 3: Experimental results of our KOM-EI against baselines.

| **Model** | **Drug** | **Weapon** | **Sexuality** | **Cost/S** |
|---|---|---|---|---|
| **StableLM** | 0.02 | 0.03 | 0.12 | 2.08S/0.00475$ |
| **mPLUG** | 0.02 | 0.13 | 0.15 | 2.35S/0.00541$ |
| **mPLUG**$_{mm}$ | - | 0.19 | - | / |
| **Llama2** | 0.17 | - | - | 18.23S/0.05833$ |
| **GPT3.5** | **0.33** | 0.17 | 0.42 | 1.12S/0.00035$ |
| **KOM-EI** | 0.32 | **0.48** | **0.50** | **0.32S/0.00004$** |

Table 4: Experimental results of our FA-Net against LLMs. Cost/S represents the average time and cost per sentence. "-" means that the models refuse to answer such questions involving inappropriate content.




(a) Before fusing (b) After fusing

Figure 4: Representation distribution of multimodal data and target keywords before and after fusing.

- **KOM-EI_C|VG**: Use CLIP as image encoder and VGGish as speech encoder.
- **KOM-EI**: Use CLIP as image encoder and Wav2Vec 2.0 as speech encoder.

#### 5.2.2 Implementation Details

To be consistent with the baselines, we also trained the models separately on each dataset and split the training set and validation set in an 8:2 ratio of text-image-speech pairs that mask out the target keywords in the text, while the test set comprised all pairs that mask out the euphemisms in the text. All experiments were conducted on a Linux server of Ubuntu 18.0.4 LTS version with a Tesla-V100 32G GPU. Please refer to Appendix B for more details.

#### 5.2.3 Evaluation Metrics

We evaluate the output using the metric precision at Acc@k, that is, the frequency of the actual label values appear in the first k values of the sorted list generated by us. To be consistent with the current best model (Zhu et al., 2021), we use Acc@1, Acc@2, and Acc@3 to measure the results.

### 5.3 Experimental Results

Table 3 summarizes the euphemism identification results (the top two rows are taken directly from Zhu et al. (2021)). To be fair, the results of all models are taken from the parameters that make the results the best. Our proposed KOM-EI model achieves the best performance, outperforming the SOTA model (SelfEDI) by 12%, 15%, and 18% in top1 accuracy values on three datasets, respectively.

**Comparison with Baselines**. Among text-only models, SelfEDI and Bert_pre showed better performance than Word2vec, both extracting sentence semantic information relatively well. Compared to SelfEDI, Bert_pre uses encoder from Transformer (Vaswani et al., 2017), which considers the semantic connections between words, obtaining sentence features with richer semantics. Meanwhile, Bert_ft uses a fine-tuning approach, which is superior to the feature-based approach Bert_pre.

Compared to the text-only models, the results obtained by the multimodal models are 4-12 percentage points higher, demonKOM-EI_C|VGstrating the efficiency of the extra modality information. Among these multimodal methods, it can be seen from the results that CLIP is more effective in acquiring image features based on text semantics than the pure image processing model VIT. Additionally, Wave2Vec 2.0 exhibits superior performance compared to VGGish in extracting audio features, showcasing its capability in capturing nuanced acoustic

| Modality | Drug | Weapon | Sexuality |
|---|---|---|---|
| **T** | 0.24 | 0.38 | 0.38 |
| **V** | 0.13 | 0.15 | - |
| **A** | 0.10 | 0.21 | 0.23 |
| **T+V** | 0.29 | 0.39 | - |
| **T+A** | 0.28 | 0.43 | 0.50 |
| **T+V+A** | **0.32** | **0.48** | **0.50** |

Table 5: Top1 ablation results of data modality. T/V/A = Text/Visual/Audio Modality. KOM-EI = T+V+A.

| Model | Drug | Weapon | Sexuality |
|---|---|---|---|
| $\Delta$ | 0.15 | 0.17 | 0.27 |
| **$\Delta$+C1** | 0.21 | 0.24 | 0.31 |
| **$\Delta$+C1+C2** | 0.23 | 0.34 | 0.36 |
| **$\Delta$+C1+C2+G** | 0.28 | 0.39 | 0.42 |
| **$\Delta$+C1+C2+G+S** | **0.32** | **0.48** | **0.50** |
| **$AN_{NotShare}$** | 0.27 | 0.44 | 0.45 |
| **$AN_{Share}$** | **0.32** | **0.48** | **0.50** |

Table 6: Top1 ablation results of model components. $\Delta$ is the base model of KOM-EI, concatenating the modality features.$AN_{NotShare}$ means the parameters do not share in the $AN_{GU}$ or $AN_{SA}$ layers. C1=CFA, C2=CA, G=GU, S=SA, $\Delta$+C1+C2+G+S=$AN_{Share}$=KOM-EI

information because of its contextualized modeling strategy.

**Comparison with LLMs**. Due to model applicability, policy, cost, etc., only mPLUG$_{mm}$ is used to identify weapon euphemisms for multi-modalities, while other LLMs use text data. Table 4 summarizes the top1 identification results of our KOM-EI against the LLMs. We observe: 1) Our KOM-EI model beats almost all the LLMs; 2) In the four LLMs, GPT3.5 is the best and most stable for euphemism identification; 3) Multimodal data can also help LLMs improve identification performance; 4) Compared with our KOM-EI, the time-consuming of the LLMs is about 3-7 times that of ours, and the cost is about 10-200 times that of ours. For more details, please refer to Appendix C.

**Visualization**. To substantiate the soundness of the KOM-EI, We map the distributions of multi-modal semantic and target keyword features to a two-dimensional coordinate space by t-SNE, as shown in Figure 4. It can be observed that: 1) Text, Visual and Audio data are all fused together after cross-modal fusion training; 2) The fused features obviously converge on specific target keywords. These further prove the effectiveness of our method.

### 5.4 Ablation Study

To investigate the effectiveness of our KOM-EI, we conducted ablation studies from both modality and model perspectives.

**Data Modality**. We conducted experiments using mono-modality or multi-modality data on the three datasets of KOM-Euph. Experimental results are presented in Table 5. It can be seen that the multi-modality methods always obtain a larger improvement than the mono-modality methods. Even though only the keyword audio can be extended on the Sexuality dataset, it also helps improve the top1 identification rate by 8 points. This is a strong suggestion on the promotional effect of extra modality information in maximizing the refinement and discrimination of the euphemisms or target keywords.

**Model Components**. To explore the efficacy of each component of KOM-EI, We conducted experiments with the Cross-modal Feature Alignment (CFA), Cross-modal Attention(CA), Gated Unit (GU), and Self-Attention (SA) gradually added to the base model $\Delta$ on the KOM-Euph. From table 6, we can observe that CFA, CA, GU, or SA all contribute to the improvement of model performance, among which CFA improves by 4-7%, CA by 2-10%, GU by 5-6%, and SA by 4-9%. It can be seen that GU is stable, while other components are sensitive to the datasets. Further, we explore the impact on the model of whether Add&Norm shares parameters across modalities. As can be seen from Table 6, not sharing parameters in the $AN_{GU}$ or $AN_{SA}$ layers results in a decrease by 4-5 percentage points of top1 accuracy, which inversely demonstrates that Add&Norm sharing parameters in text-image and text-audio fusion can effectively promote the consistency of modal features, thereby improving the model performance.

## 6 Conclusion

In this paper, we propose to enhance the euphemism identification through extra modality information. Following the evolution of euphemisms, We contribute a keyword-oriented multimodal euphemism corpus (KOM-Euph). Moreover, we propose a keyword-oriented multimodal euphemism identification method (KOM-EI), which can recognize euphemisms efficiently by using cross-modal feature alignment and dynamic fusion. Extensive experiments show that our method is effective and comparable to LLMs.

## Limitations

Since there is no labeled dataset for training the euphemism identification problem. During the training phase, sentences containing the target keywords are used with the target keywords masked out, while the corresponding target keywords serve as labels. Nevertheless, during testing, sentences containing euphemisms are used, with the euphemisms masked out. As a result, the training and test data diverge in terms of their distribution resulting in a gap between them. There remains scope for further improvement, and this will be the focus of our subsequent research.

## Ethics Statement

The text data used in this article was obtained legally in accordance with the guidelines set by Zhu et al. (2021) and adhered to strict privacy standards to ensure that there is no personally identifiable information such as real name, email address, IP address, etc. The visual data is obtained from public platforms and does not contain any private information. The audio data is pronunciation data generated by public tools without any additional information. All the data is for scientific research purposes only.

## References


Drug Enforcement Administration et al. 2018. Slang terms and code words: A reference for law enforcement personnel. *DEA Intelligence Report DEAHQ-DIR-022*, 18(2018):2018–07.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Jack Foye, Matthew Ball, Chuxuan Jiang, and Roderic Broadhurst. 2021. Illicit firearms and other weapons on darknet markets. *Trends and Issues in Crime and Criminal Justice [electronic resource]*, 1(622):1–20.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671.

Takuro HADA, Yuichi SEI, Yasuyuki TAHARA, and Akihiko OHSUGA. 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.

Zhongqiang Huang, Yuxue Hu, Zhi Zeng, Xiang Li, and Ying Sha. 2023. Multimodal stacked cross attention network for fine-grained fake news detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2837–2842. IEEE.

Heng Ji and Kevin Knight. 2018. Creative language encoding under censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 23–33.

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.

Ilker Kesen, Aykut Erdem, Erkut Erdem, and Iacer Calixto. 2022. Detecting euphemisms with literal descriptions and visual imagery. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 61–67.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. Searching for pets: Using distributional and sentiment-based methods to find potentially euphemistic terms. *arXiv preprint arXiv:2205.10451*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from

natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

K Seethappan and K Premalatha. 2022. A comparative analysis of euphemistic sentences in news using feature weight scheme and intelligent techniques. *Journal of Intelligent & Fuzzy Systems*, 42(3):1937–1948.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

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

Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. 2023. An explainable multi-view semantic fusion model for multimodal fake news detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1235–1240. IEEE.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

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

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

# A Data Source and construction

We construct a **K**eyword-**O**riented **M**ultimodal **Euph**emism (KOM-Euph) corpus, based on the only text corpus (noted as "Euph") presented by Zhu et al. (2021). The original Euph corpus is sourced from the Reddit website[1], Gab social networking services[2], Online Slang Dictionary[3], etc. The Euph corpus consists of only textual data, involving three datasets: Drug, Weapon, and Sexuality. Since the Sexuality dataset mainly involves private parts of the body and sexual activities, the collection of their images is not allowed by law, so we expanded the visual and audio modality data for the Drug and Weapon dataset and only audio modality data for the Sexuality dataset.

**Visual Modality Data Construction**. We use images of the keywords in the text as the corresponding visual information. Specifically, we crawl images from two public online platforms, Google and Wikipedia. Both platforms are representative of users to obtain objective and comprehensive pictures of each entity. We use target keywords or euphemisms as keywords and retain the top 10 retrieved images on each platform. Additionally, to ensure the image quality of each keyword, we use the image generation model Kandinsky 2.2 [4] to generate 5 images for each keyword. Then, we get 20 images for each keyword, 10 from Google, 5 from Wikipedia, and 5 generated from Kandinsky 2.2.

We hired a linguistics expert to train 10 undergraduate students to screen keyword images. The principle of screening is to select 4 pictures that best present the literal meaning of the keywords. For words with unclear literal meanings, such as "k4", "404", etc., directly select the 2 top-ranked pictures on Google and 2 generated. Finally, we filter out 4 images for each keyword.

**Audio Modality Data Construction**. We use the speech of keywords in the text as the corresponding audio information to help identify euphemisms. In addition to the normal pronunciation information of the word itself, no additional information is required, such as speaking speed, intonation, etc. To this end, we use Bark[5] to get the normal pronunciation as the audio information,

[1] https://www.reddit.com/
[2] https://gab.com/
[3] https://slangpedia.org/
[4] https://github.com/ai-forever/Kandinsky-2
[5] https://github.com/suno-ai/bark

and we only generate a piece of speech for each keyword.

**Ground Truth List**. To evaluate our results, we need to rely on a ground truth list (Zhu et al., 2021) of euphemisms and the corresponding target keywords, which should contain a one-to-one mapping from each euphemism to its true meaning. The ground truth list on Drug was compiled by the U.S. Drug Enforcement Administrator to provide a practical reference for law enforcement personnel (Administration et al., 2018). The ground truth list on Weapon was sourced from the Online Slang Dictionary[6] and the Urban Thesaurus[7]. The ground truth list on Sexuality came from the Online Slang Dictionary. Due to the rapid evolution of the language used on social networks, it cannot be comprehensive or error-free, but it is the most reliable ground truth we can get.

## B Exiperimental Details

To be consistent with the baselines, we also trained the models separately on each dataset and split the training set and validation set in an 8:2 ratio of sentences that mask out the target keywords, while the test set comprised all sentences that mask out the euphemisms. All experiments were conducted on a Linux server of Ubuntu 18.0.4 LTS version with a Tesla-V100 32G GPU.

**Unimodal Model Settings** Firstly, we pre-trained a Bert model based on bert-base-uncased[8] for MLM task only to extract context features (768 dimensions) of masked sentences. Then, we fine-tuned the model for the euphemism identification task. During pre-training, the maximum length of the input sequence was set as 512, the batch size as 64, and the number of iterations as 3. For model training, the maximum length of the input sequence was 128, and the batch size was 128. The initial learning rate was 5e-5, the warm-up step was 1000, and the optimizer AdamW (Loshchilov and Hutter, 2018) is based on a warm-up linear schedule.

**Multimodal Model Settings**. For visual feature extraction, we employed the clip-vit-large-patch14 model from CLIP[9], designed to process images and produce feature vectors of 768 dimensions for each visual representation. Audio features were extracted using the wav2vec2-large-960h model from Wave2Vec 2.0[10], which yields feature vectors of $T \times 768$ dimensions, where $T$ represents the time-steps corresponding to the audio segment duration. The configuration of other parameters is consistent with the specifications outlined in the Unimodal Model Settings.

## C LLMs for Euphemism Identification

In this paper, we compared our proposed KOM_EI model to four current best large language models (LLMs) for euphemism identification task, namely, GPT-3.5-turbo(GPT3.5[11] for short), Llama2[12], mPLUG-Owl[13], and StableLM[14]. These LLMs are described in detail below.

### C.1 Introduction of LLMs

We briefly introduce the four LLMs from the model type, parameter number, maximum text input length, cost and other aspects, as shown in Table 7. GPT3.5 and stableLM are both natural language processing models, while Llama2 and mPlug-Owl are multimodal processing models that are more expensive. For details and interfaces about the LLMs, see the footnote link address.

### C.2 Result Analysis

When using the GPT3.5 and StableLM interfaces to identify euphemisms, we used four content templates, as shown in Table 8. From Table 8, we observe that the results vary according to the content templates, and GPT3.5 is relatively stable compared to StableLM. However, the results are not consistent across different models and different datasets, indicating the randomness of the output results of these large language models. For the other multimodal processing models, i.e., Llama2 and mPlug-Owl, which are too expensive to use four templates for testing, we only use Template 1 to test the identification accuracy. Finally, we take the best result on each dataset and record it in Table 4 in the body part. It's obvious that GPT3.5 performs the best among the four LLMs, and outperforms our proposed KOM-EI by 1 percentage point on the Drug dataset. When using Llama2

[6] http://onlineslangdictionary.com/
[7] http://urbanthesaurus.org/
[8] https://huggingface.co/bert-base-uncased/
[9] https://huggingface.co/openai/clip-vit-large-patch14
[10] https://huggingface.co/facebook/wav2vec2-large-960h
[11] https://platform.openai.com/docs/api-reference/introduction
[12] https://huggingface.co/models?other=llama-2
[13] https://modelscope.cn/studios/damo/mPLUG-Owl/summary
[14] https://replicate.com/stability-ai/stablelm-tuned-alpha-7b

API interface or web UI to test the identification accuracy on Weapon or Sexuality dataset, it informs that it is inappropriate to discuss such a topic and refuses to answer questions, while we can only test on the Drug dataset via web UI. We present case studies of the LLMs in the following sections.

### C.3 Case studies

Through the analysis of the euphemism identification results of the four LLMs, we have the following two findings:

1) GPT3.5 performs the best among the four LLMs with a strong understanding of euphemisms. However, it still lacks understanding of the relatively rare euphemisms, and the recognition rate of commonly used euphemisms is almost 100%. As shown in Figure 5a, it can always identify "weed" (a common euphemism) to its true meaning "marijuana", while having no idea of "ice" in euphemistic use of "methamphetamine" (Figure 5b).

2) GPT3.5 is relatively stable as the identification results of common euphemisms are correct while the other LLMs are not. That's why the other LLMs perform far worse than GPT3.5 in euphemism identification. Take the mPLUG-Owl model for example, when we ask it for the same question about the meaning of the euphemism in the sentence four times, it gives completely different answers, as shown in Figure 6.

| LLMs | Type | Parameters | Maximum input | Cost | Institution |
|---|---|---|---|---|---|
| **GPT3.5** | NLP | 20B | 4096 tokens | 0.015$/1k tokens | OpenAI |
| **StableLM** | NLP | 3B - 7B | 4096 tokens | 0.0023$/second | Stability AI |
| **Llama2** | MM | 7B - 70B | 4096 tokens | 1.05$/hour | Meta |
| **mPLUG-Owl** | MM | 7B | 2048 tokens | 0.0023$/second | DAMO Academy |

Table 7: Introductions of the four LLMs. MM = Multi-Modal, T = Trillion, B = Billion.

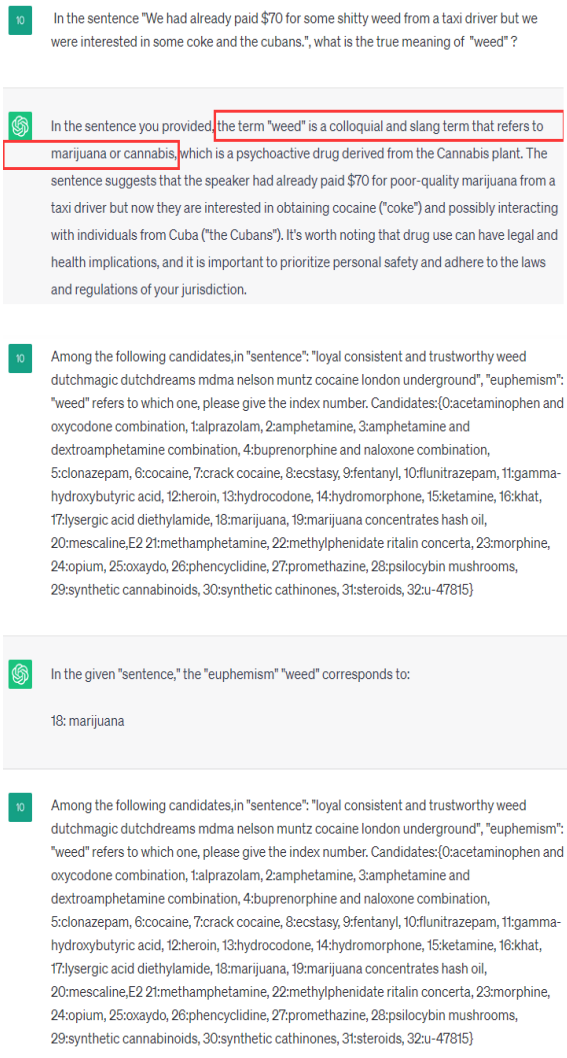
| Template | GPT3.5 | | | StableLM | | |
|---|---|---|---|---|---|---|
| | Drug | Weapon | Sexuality | Drug | Weapon | Sexuality |
| 1. Among the following candidates:{T}, in sentence:{s}, euphemism:{euph} refers to which one. | 0.3314 | 0.1543 | 0.4041 | 0.0214 | 0.0149 | 0.1184 |
| 2. Among the following candidates:{T}, in sentence:{s}, {euph} refers to which one. | 0.2971 | 0.1686 | 0.3876 | 0.0160 | 0.0248 | 0.1061 |
| 3. The true meaning of euphemism:{euph} in "{s}" refers to which of the following canditates:{T}. | 0.3257 | 0.1564 | 0.4204 | 0.0053 | 0.0299 | 0.1102 |
| 4. The true meaning of "{euph}" in "{s}" refers to which of the following canditates:{T}. | 0.3029 | 0.1482 | 0.0106 | 0.0107 | 0.0199 | 0.1020 |

"T" refers to the fixed target keyword candidate list: {0:acetaminophen and oxycodone combination, 1:alprazolam, 2:amphetamine, 3:amphetamine and dextroamphetamine combination, ... }

"s" refers to the test sentence containing euphemisms.

"euph" refers to the euphemism in the test sentence that need to be identified.

Table 8: Content templates and results.




(a) Identification of "weed"

(b) Identification of "ice"

(c) Identification of "porn"

Figure 5: Cases of GPT3.5.

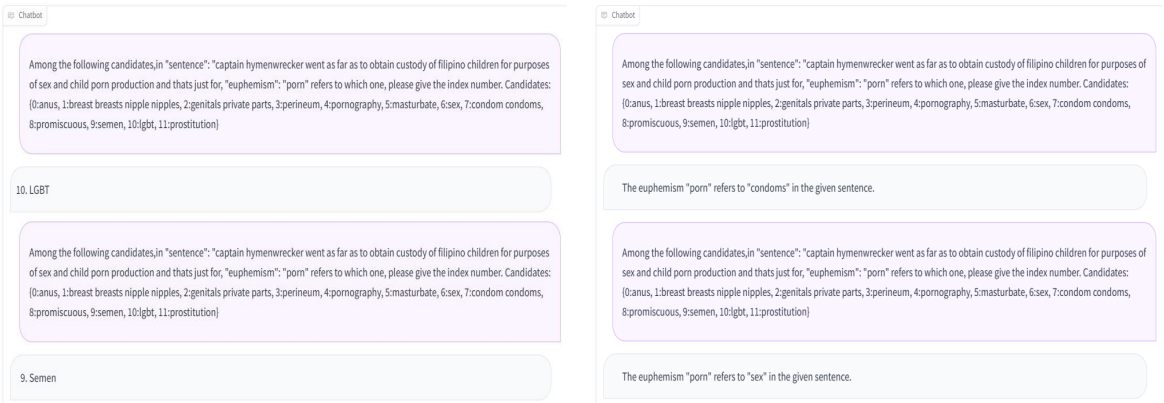



Figure 6: Cases of mPLUG-Owl.