A Unified Generative Framework for Bilingual Euphemism Detection and Identification

Yuxue Hu[1,2,3,4*], Junsong Li[1,2*], Tongguan Wang[1,2*], Dongyu Su[1,2], Guixin Su[1,2], Ying Sha[1,2,3,4†]

[1]College of Informatics, Huazhong Agricultural University, Wuhan, China
[2]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China
[4]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
{hyx,shaying}@mail.hzau.edu.cn,{kikuss,wang_tg,su_dy,cometsue}@webmail.hzau.edu.cn

Abstract

Various euphemisms are emerging in social networks, attracting widespread attention from the natural language processing community. However, existing euphemism datasets are only domain-specific or language-specific. In addition, existing approaches to the study of euphemisms are one-sided. Either only the euphemism detection task or only the euphemism identification task is accomplished, lacking a unified framework. To this end, we construct a large-scale **B**ilingual **M**ulti-category dataset of **E**uphemisms named **BME**, which covers a total of 12 categories for two languages, English and Chinese. Then, we first propose a unified generative model to **Joint**ly conduct the tasks of bilingual **E**uphemism **D**etection and **I**dentification named **JointEDI**. By comparing with LLMs and human evaluation, we demonstrate the effectiveness of the proposed JointEDI and the feasibility of unifying euphemism detection and euphemism identification tasks. Moreover, the BME dataset also provides a new reference standard for euphemism detection and euphemism identification.

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers.

1 Introduction

Euphemisms are forms of language that express ideas or convey information through the use of indirect or cryptic language. The original intention of using euphemisms is to avoid direct, blunt, or potentially offensive expressions (Pinker, 2003). However, to avoid explicitly expressing unfriendly views or statements, some users choose to use euphemisms to cover up discriminatory, insulting, or unfair remarks (Chilton, 1987). As shown in Figure 1, lawbreakers use euphemisms (eg: “weed” means “wild grass” in literal English, but “drugs” in euphemisms) (Zhu et al., 2021) to distract the attention of the cyber police and complete the transaction of drugs, guns, and other illegal goods. People discriminate or insult others using euphemisms (eg: “同志” (comrade) in Chinese means people who strive for a common ideal or cause, but in euphemisms it means “同性恋” (homosexual)) (Lee et al., 2023). Therefore, it is important to study the detection and identification of euphemisms to detect and intervene in the transmission of euphemisms promptly.

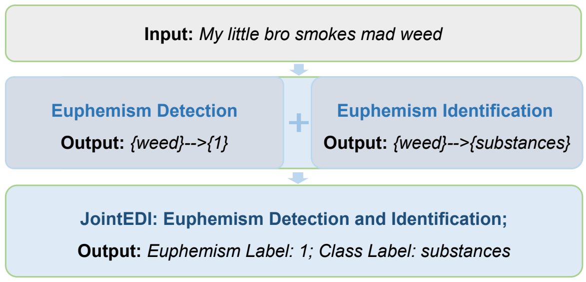


Figure 1: Comparison of JointEDI with euphemism detection and euphemism identification tasks. "1" stands for "weed" as a euphemism in the sentence, "substances" stands for the category of the euphemism "weed".

As shown in Figure 1, existing euphemism tasks can be divided into two categories according to their purpose (Zhu et al., 2021): (1) Euphemism Detection: the main purpose of the task is to determine whether a text contains euphemisms so that they can be further analyzed or processed. (2) Euphemism Identification: this task focuses more on identifying specific euphemistic expressions in the text and aims to understand and analyze the use of euphemisms in the text in more detail. Detecting euphemisms and identifying euphemisms in practice is an ongoing process, similar to the pedestrian detection and identification task of computer vision. Once the euphemisms are detected, their specific meaning need to be identified. However, most of the existing studies only focus on euphemism detec-

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 6753–6766
August 11-16, 2024

tion or euphemism identification (Ke et al., 2022; Felt and Riloff, 2020; Hu et al., 2023), which is somewhat one-sided, as shown in Figure 1. Yuan et al. (2018) used a binary random forest classifier and recursive lookup method to identify the hypernym of the euphemisms. Zhu et al. (2021) detected and identified euphemisms using the mask model and bag of words model based on a self-supervised scheme. These two methods, first detecting euphemisms and then identifying them, are pipeline methods that easily propagate errors. To the best of our knowledge, there is not yet a methodology to unify the two tasks into a single framework.

Due to the nature of language development, euphemisms are used frequently in different fields of various languages, especially the two major languages, English and Chinese. However, current research is limited to a single language or a few fields (Gavidia et al., 2022; Keh et al., 2022; Zhu et al., 2021; Lu et al., 2023). In addition, with the exchange and collision of Chinese and English, some euphemistic expressions combine the two languages to convey implicit meanings (eg: "OMG, 你这是<发福>了吗？" (Literal meaning: OMG, are you <reaping blessings>? Implicit meaning: OMG, are you <getting fat>?)). To this end, euphemism datasets covering multiple languages and domains are urgently needed, which is crucial for the study of euphemisms.

To solve the above challenges, we integrate an existing dataset of euphemisms and supplement it by collecting additional data from various websites. We construct a large-scale bilingual multi-category dataset of euphemisms named BME, which includes two major languages of the world, English and Chinese. This dataset is filtered in detail and manually labeled, which covers a total of 12 categories. Furthermore, we propose a novel unified framework for the joint implementation of the euphemism detection and euphemism identification tasks, a generative model named JointEDI, which adopts two auxiliary tasks. Our proposed method achieves F1 values of 0.9311 and 0.8881 on the BME English dataset and Chinese dataset, respectively. Meanwhile, comparing the existing LLMs and manual evaluations verifies the superiority of our proposed method and provides new insights for future work.

Our contributions are as follows:

- A large-scale **B**ilingual **M**ulti-category **E**uphemism dataset named BME is constructed, including 2 languages and covering 12 categories in total, which provides a new benchmark in the field of euphemism detection and identification. We also provide in-depth statistical analysis.
- A unified generative framework to **Joint**ly conduct the tasks of **E**uphemism **D**etection and **I**dentification named **JointEDI**[1] is proposed, employing two auxiliary tasks. To the best of our knowledge, this is the first framework to unify the task of euphemism detection and euphemism identification.
- Experimental results on BME datasets show that 1) the proposed JointEDI outperforms other baselines and LLMs, demonstrating the validity of our approach, and 2) our results are higher than common human evaluation results, but lower than those evaluated by professional humans, demonstrating the challenging nature of our dataset and the unified task.

2 Related work

2.1 Datasets

For computers, euphemisms often involve complex contexts and emotions, and accurately understanding and processing these linguistic expressions is still a challenging task even for LLMs (Gibbs, 1999). Many domain-specific euphemism datasets have been proposed. We summarize and analyze the most representative datasets in recent years in Table 1.

It can be seen from Table 1, Zhu et al. (2021) and Ke et al. (2022) proposed English and Chinese datasets for the domain of darknet euphemisms, respectively. Rahman et al. (2021) and Yadav et al. (2023) proposed two-classification and five-classification datasets for the domain of hate euphemisms, respectively. Gavidia et al. (2022) first introduced the concept of PET (Potentially Euphemistic Terms) and proposed a multi-category euphemism dataset. Lee et al. (2023) proposed four different languages to present a novel euphemism corpus, which is expanded to four languages based on the data set proposed by Gavidia et al. (2022). Although some datasets are quite large (Zhu et al., 2021; Ke et al., 2022; Mody et al., 2023; Yadav et al., 2023), more data is not always better, and extra irrelevant data may affect the model due to pseudo-correlation coincidence (Feng, 2021). Therefore, we have to ensure the size of the dataset

[1]The code and data are publicly available at: https://github.com/DHZ68/JointEDI.

while improving the quality of the dataset, such as data categories and their distribution. In order to promote the research of euphemisms and better reflect the diversity of euphemisms in real scenarios, a large-scale multi-lingual and multi-category euphemism dataset is urgently needed.

2.2 Euphemism Detection

The main objective of the euphemism detection task is to detect whether a piece of text contains euphemisms or not. Magu and Luo (2018) proposed a method to help identify unknown words to detect hate speech euphemisms using word embedding and network analysis. Ghosh et al. (2020) proposed a shared task for detecting hate speech that focuses on the detection of hate speech euphemisms using the entire context of a previous conversation, which achieved a high detection accuracy of 0.932 for the first-place team in that competition. It is worth noting that almost all teams used pre-trained transformer-based models. Zhu et al. (2021) formulated the euphemism detection problem as an unsupervised filler mask problem and solved it by combining self-supervision with a masked language model. A recent work that has attracted attention is the presentation of the Shared Euphemism Detection Task (Lee et al., 2022). The purpose of the task is: give an input text and detect whether it contains euphemisms or not. The competition attracted 13 teams, Keh et al. (2022) combined the best-performing models into an ensemble of three models and achieved first place in that competition. Kesen et al. (2022) used additional supervised information to obtain images of both the PETs and their literal descriptions using a text-to-image model, combining textual and visual modalities to achieve satisfactory euphemism detection results.

2.3 Euphemism Identification

Once euphemisms are detected, the subsequent identification of euphemisms is extremely important because different types of euphemisms determine the specific application scenarios of euphemisms. However, there is relatively little work related to the study of euphemism identification tasks. Since a euphemism often contains several different meanings, this task is more challenging than the euphemism detection task (Zhu et al., 2021). Yuan et al. (2018) proposed Cantreader, which employs a neural network-based embedding technique to analyze the semantics of words, to be used for automatic detection and comprehension of cryptic speech. Instead of directly identifying the specific meaning of a euphemism, they generate a set of superlatives and use a binary random forest classifier and recursive lookup to categorize a given euphemism into a specific superlative. Felt and Riloff (2020) used sentiment analysis to identify euphemisms and dysphemisms, and although the performance is relatively low and the subject matter is narrow, this work certainly has stimulated further research. Zhu et al. (2021) explicitly defined the task of euphemism identification for the first time, and developed a self-supervised learning algorithm that utilizes a bag-of-words model to classify a given euphemism to a specific superordinate word at the sentence level.

Although both euphemism detection and euphemism identification tasks have achieved some results, they are two independent tasks. In practical applications, the detection and identification of euphemisms is a continuous process, similar to the pedestrian detection and identification task of computer vision. We not only need to detect euphemisms from a sentence but also identify the meaning of the specific expression of the euphemism. To the best of our knowledge, only Zhu et al. (2021) have proposed a pipeline that connects these two tasks in tandem, but this is not a unified framework. Moreover, the approach is limited to three specific tasks, namely drugs, weapons, and sex, in the darknet. Unlike all previous approaches, we propose a unified framework to unify the tasks of euphemism detection and euphemism identification to fully understand the implicit meaning to be conveyed throughout the sentence. As shown in Figure 1, JointEDI can not only detect whether a sentence contains a euphemism but also identify to what category the euphemism belongs.

3 Dataset Construction

3.1 Data Collection

In summary, the BME dataset comes from crawled data and three existing datasets. The construction process of the BME dataset is shown in Figure 2. Our goal is to construct a large-scale euphemism dataset covering multi-category and multi-lingual euphemisms. We first extensively research and analyze the purpose euphemism dataset in Section 2.1. We collect the following potentially usable datasets. These include the datasets proposed by Lee et al. (2023), Lu et al. (2023) and Zhu et al. (2021). It

Dataset	Sentence		Category	PET	Domain	Language
	English	Chinese				
Zhu et al. (2021)	-	-	3	Yes	Darknet	English
Rahman et al. (2021)	4,275	-	1	No	Hateful	English
Gavidia et al. (2022)	1,965	-	7	Yes	Hateful	English
Ke et al. (2022)	-	44,720	10	Yes	Darknet	Chinese
Mollas et al. (2022)	999	-	6	No	Hateful	English
Mody et al. (2023)	451,709	-	1	No	Hateful	English
Yadav et al. (2023)	227,836	-	5	No	Hateful	Six languages
Lu et al. (2023)	-	12,011	4	No	Hateful/Offensive	Chinese
Lee et al. (2023)	1,952	1,552	7	Yes	-	Four languages
BME (Ours)	4,512	4,495	12	Yes	All	English/Chinese

Table 1: Comparison of existing euphemism datasets. For comparison, we show only the number of English and Chinese sentences in the datasets of Lee et al. (2023).

is worth stating that the dataset provided by Lu et al. (2023) consists of 12,011 sentences. Based on the keywords of euphemisms, we select a total of 1,775 sentences from that meet our requirements. However, according to our statistics, 338 sentences of them are labeled with the wrong category. For example:

"这种男人也是极少数，像我这种只是单纯的<普信>罢了"(There are very few men like that, like me it's just plain <Common and confident>.), the "普信"(Common and confident) here should be labeled with the category of "性别"(gender), but the original dataset is labeled with the category of "种族"(racist).

To ensure that the dataset covers as many categories as possible while minimizing the problem of inter-class and intra-class imbalance in the BME dataset, we also crawl a large amount of data from dictionary websites and search engines and incorporate it into the BME dataset. Specifically, after collecting the preliminary data, we combine previous work to integrate a keyword list that needs to be expanded. We then use the Selenium tool to crawl broad sentences on glosbe[2] and sogou[3] based on the list obtained, and then manually annotate the large amount of data. Finally, we fill in the annotated data appropriately.

3.2 Data Cleansing and Filtering

For the three collected datasets mentioned above, we mainly use the Chinese dictionary and the English dictionary proposed by Lu et al. (2023) to annotate the Chinese data and the English data.

[2]https://glosbe.com
[3]https://wap.sogou.com

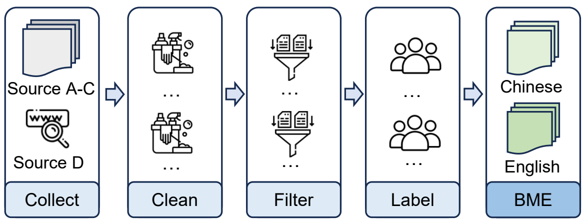


Figure 2: Flowchart of dataset construction. It consists of four main processes, which are data collection, data cleaning, data filtering, and data labeling.

We collect keywords and filter out sentences in the dataset that do not contain keywords. At the same time, incomplete data are eliminated, and data with obvious category labeling errors are selected for manual secondary category labeling. To construct a high-quality euphemism dataset, we mitigate the problem of inter-class imbalance in the dataset by filtering categories with less than 50 sentences, such as the category "misc.".

3.3 Data Annotation

The data we collected contain an assortment of types, including daily polite phrases, discriminatory, sarcastic, and phrases from domains such as the darknet. Since euphemisms are related to the social and cultural aspects of language use, they are an important research area in the field of sociolinguistics in linguistics. Therefore, to ensure the quality and authority of the collected data on euphemisms, we hire five linguistic professionals to manually label the data, including three PhD candidates and two Master candidates. We offer systematic training to annotators before the commencement of data labeling. See Appendix A for training programs.

Category	Sentence	
	English	Chinese
body functions/parts	209	450
death	479	580
employment/finances	484	477
physical/mental attributes	781	401
sexual activity	225	421
politics	525	-
substances	538	-
weapon	1,271	-
gender	-	827
racist	-	607
homosexual	-	557
region	-	175
In total	4,512	4,495

Table 2: Data categories and quantities of English euphemisms and Chinese euphemisms in the BME dataset. The BME English data have 2,658 sentences with label 1 and 1,854 sentences with label 0. The BME Chinese data have 3,100 sentences with label 1 and 1,395 sentences with label 0.

In the labeling process, when encountering controversial categories of euphemisms, we employ a voting method to select the category with the highest number of votes, determining the final categorization of euphemisms. We follow the basis of the classification of Lu et al. (2023). We define a total of 8 specific categories for the BME English dataset and 9 specific categories for the BME Chinese dataset.

3.4 Data Analysis

After processing through the above process, we finally collect a total of Chinese data and English data as shown in Table 2. We end up with the number of data for each category. The accounting for each category is shown in Figure 3. To illustrate what kind of keyword euphemisms are available for each category, we make keyword data analysis on the English and Chinese datasets of BME respectively. Detailed data analysis can be found in the Appendix B.

4 Task Definition

Unlike previous tasks focusing solely on euphemism detection or identification, our primary goal is to integrate both aspects into a unified modeling framework. Our core objective is to propose a unified framework capable of automatically detecting euphemisms in text and identifying them into the correct categories. It is known that a sentence containing a euphemism, $s = [w_1, ..., \text{<PET>}, ..., w_i, ..., w_m]$ (where PET is known to be a potential euphemism term). Our goal is to determine whether PET represents a euphemism in a sentence and, if PET is a euphemism, identify the category to which PET belongs.

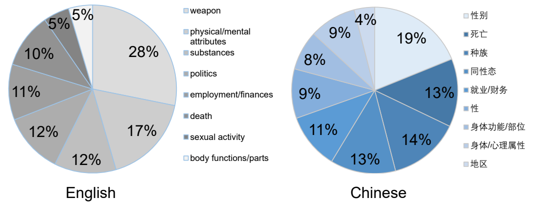


Figure 3: The pie chart on the left shows the percentage of English data by category, and the one on the right shows the Chinese data.

As shown in Figure 1, here are the inputs and outputs for the euphemism detection and identification task:

Input: "My little bro smokes mad <weed>."

Output: "Target: Euphemism Label: 1, Class Label: substances."

The euphemism label of the model's output is "1", which indicates that "weed" represents a euphemism in the sentence, and the class label of the model's output is "substances", which indicates that "weed" is categorized into "substances".

5 Methodology

5.1 Model Overview

Since our proposed JointEDI aims to unify the tasks of euphemism detection and euphemism identification for multiple languages, and mBART(Liu et al., 2020) has demonstrated outstanding performance on multiple tasks. We use a multi-lingual BART (mBART), which is an extended version of a transformer-based pre-trained BART (Lewis et al., 2020) for multiple languages, as our Seq2Seq backbone. The overall architecture of JointEDI is shown in Figure 4, which is mainly composed of the mBART encoder and the mBART decoder.

As we discussed in the previous section, our task can be represented as taking X = [Task: Euphemism Detection and Identification; s] as input and outputting a target sequence Y = [Euphemism Label: y; Class Label: c], where s stands for the sentence to be detected and identified, $y \in (0, 1)$ and $c \in$ Category in Table 2. Thus, euphemism detection and identification can be formulated as follows:

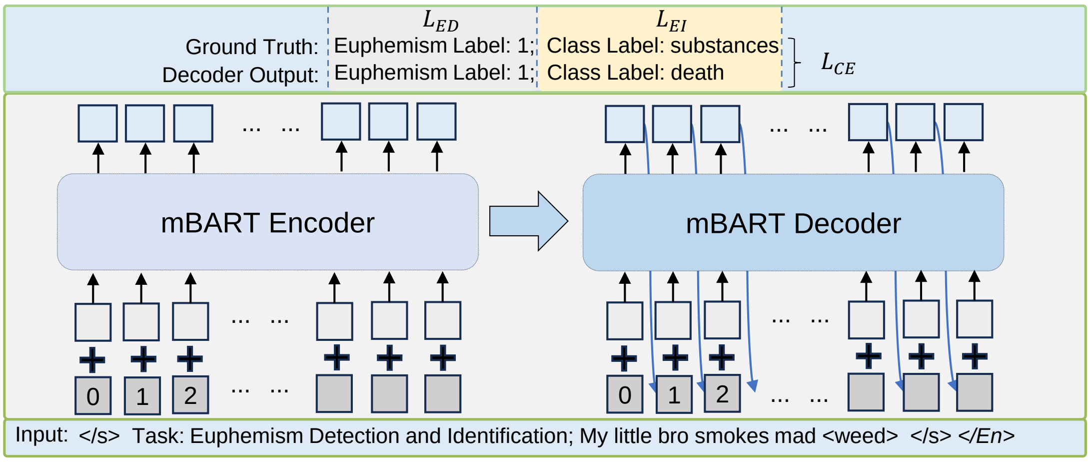


Figure 4: Overall network framework for JointEDI.

$$Y = \text{mBART}(X), \tag{1}$$

where X is the input sequence and Y is the output sequence generated by the model.

To assist the unified model in achieving better results, we devise two auxiliary tasks in our model, namely, the euphemism detection (ED) task and the euphemism identification (EI) task. Next, we will introduce each component of our proposed JointEDI separately.

5.2 Encoder and Decoder

Encoder The sentence to be encoded is taken as input and passed to the mBART encoder. For the input sequence X, the output can be expressed as:

$$H_{en} = \text{Encoder}(X). \tag{2}$$

Decoder The decoder of mBART also uses the same attention mechanism as the transformer model but uses an auto-regressive process for training. For the output Y, it can be represented as:

$$Y = \text{Decoder}(H_{en}, Y_{<t}), \tag{3}$$

where t is the length of the output sequence, $Y_{<t}$ denotes the sequence that has been generated before position t.

5.3 Loss Function

The unified model is trained to minimize the cross entropy between the generated Y and the ground truth $\widehat{Y}$. The loss function of the main task is shown as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N} Q(\widehat{Y}_i)\log P(Y_i;\Theta), \tag{4}$$

where N represents the number of samples, $Q(\widehat{Y}_i)$ represents the distribution of real labels of the sample i_{th} and $P(Y_i;\Theta)$ represents the distribution predicted by the model under parameter Θ.

To improve the unified model's detection and identification performance of euphemisms, we propose two auxiliary tasks: euphemism detection (ED) and euphemism identification (EI). The ED task is to detect whether potential euphemism terms are in euphemistic usage. The training goal is as follows:

$$\mathcal{L}_{ED} = -\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{y}_i \log p_i + (1-\widehat{y}_i)\log(1-p_i)\right), \tag{5}$$

where N represents the number of samples, $\widehat{y}_i$ represents the true label of the i_{th} sample, and p_i represents the probability predicted by the model for the i_{th} sample.

The EI task is to identify the categories corresponding to potential euphemism terms. The training goal is as follows:

$$\mathcal{L}_{EI} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(p_{i,j}), \tag{6}$$

where N represents the number of samples, C represents the number of categories, and $y_{i,j}$ is the label of the category j_{th} in the true label of sample i. $p_{i,j}$ is the probability of the j_{th} category in the model prediction of sample i.

The training objective loss of JointEDI is finally formalized as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{ED} + \gamma\mathcal{L}_{EI}, \tag{7}$$

Model	English						Chinese					
	$F1(\%)$	$R(\%)$	$P(\%)$	$F1_d(\%)$	$R_d(\%)$	$P_d(\%)$	$F1(\%)$	$R(\%)$	$P(\%)$	$F1_d(\%)$	$R_d(\%)$	$P_d(\%)$
mbart-large-cc25	85.71	82.95	88.67	86.47	83.16	90.06	87.14	85.65	86.26	89.44	86.21	92.92
mbart-large-50	85.18	89.68	81.10	87.16	90.05	84.45	73.53	94.44	60.20	83.05	95.40	73.52
mT5-base	63.32	80.07	52.36	79.81	85.20	75.06	54.55	72.97	43.55	79.12	82.49	76.01
mT5-large	79.90	88.27	72.98	84.85	89.29	80.83	83.77	91.39	77.33	88.54	92.12	85.22
JointEDI	**93.11**	**92.51**	**93.72**	93.80	92.60	95.03	88.81	91.96	85.87	92.97	92.56	93.38
Falcon	0	0	0	13.26	8.63	28.57	0	0	0	0	0	0
LLaMA2-70b-chat	32.84	27.50	40.74	63.49	54.05	76.92	5.41	4.35	7.14	40.68	26.09	92.31
mPLUG-Owl	27.48	37.40	21.88	63.61	89.10	55.01	4.49	12.08	2.78	35.19	95.23	21.67
Stability-AI	9.42	5.56	30.93	9.76	5.71	33.33	0	0	0	0	0	0
GPT-3.5	58.70	65.85	52.94	84.71	81.82	87.80	43.59	37.78	51.12	74.36	64.44	87.88
GPT-4.0	73.17	75.00	71.43	90.24	92.50	88.10	53.66	48.89	59.46	82.93	75.56	91.89
Human-com	68.20	71.15	66.08	84.29	77.50	92.93	80.03	79.49	80.64	88.65	82.22	96.25
Human-pro	92.75	92.17	93.40	**95.44**	**93.75**	**97.39**	**94.33**	**95.40**	**93.28**	**96.64**	**96.67**	**96.64**

Table 3: Comparison of JointEDI with baselines and Large Language Models on BME Chinese dataset and English dataset respectively. Human-com represents the average metrics of test results for non-professional people, and Human-pro represents the average metrics of test results for professional people.

Model	English			Chinese		
	$F1_d(\%)$	$R_d(\%)$	$P_d(\%)$	$F1_d(\%)$	$R_d(\%)$	$P_d(\%)$
DAN	91.59	91.48	91.71	-	-	-
CLS	92.49	92.29	92.72	-	-	-
PET	**93.91**	**93.95**	93.86	-	-	-
mBERT	89.49	89.03	89.95	89.45	89.93	88.96
XR-b	90.40	90.05	90.75	90.85	91.25	90.46
XR-l	91.57	92.86	90.32	91.57	87.96	**95.49**
JointEDI	93.80	92.60	**95.03**	**92.97**	**92.56**	93.38

Table 4: Comparison of JointEDI with baselines in Keh et al. (2022) and Lee et al. (2023). "XR-b" stands for "XLM-RoBERTa-base", "XR-l" stands for "XLM-RoBERTa-large". "-" means that the method in the original paper corresponding to this baseline does not support a certain language.

where α, β, and γ represent the weights of the three loss functions, respectively. The sum of α, β and γ is 1.

6 Experiments

6.1 Evaluation Setup

Datasets: We evaluate our method on the BME dataset constructed in section 3. There are 4,512 sentences in the English dataset and 4,495 sentences in the Chinese dataset. We divide the two datasets according to the ratio of training, validation, and testing 7:1.5:1.5, and when dividing the datasets, we try to ensure the balance of inter-class and intra-class data. The final results of the dataset division are shown in the Appendix C.

Implementation Details: During training process, the maximum length of the input sequence is set to 128, and the initial learning rate is set to 1e-5. We train the model for 20 epochs on a 40GB Tesla A100 GPU with the batch size set to 32. We use the Adam optimizer and the model employ a cosine annealing learning rate schedule. For all experimental results, we set random seeds to ensure the reproducibility of the experiments.

Baselines: We compare ten baselines and six LLMs. All baselines are performed in the same experimental setting. Since Keh et al. (2022) and Lee et al. (2023) only completed the euphemism detection task, we compare JointEDI with these methods separately on the euphemism detection task, as shown in Table 4. We replicated their methods based on the open source code, respectively DAN, CLS, PET (Keh et al., 2022) and mBERT, XLM-RoBERTa-base, XLM-RoBERTa-large (Lee et al., 2023). For LLMs, we directly use their open-source API interface. More details are described in Appendix C.

Accuracy metrics: We set up six evaluation metrics, where P, R, and $F1$ represent the metrics for the task of unifying euphemism detection and euphemism identification, P_d, R_d, and $F1_d$ represent the metrics for euphemism detection. The values of $F1$ and $F1_d$ are the main evaluation metrics.

6.2 Results and Analysis

Comparison with Baselines: As shown in Table 3, on the unified euphemism detection and identification task, we find that JointEDI outperforms other methods on all three metrics in English data, but does not perform as well as the other methods in Chinese data. This suggests that JointEDI can effectively detect euphemisms in sentences and identify their categories. Even though mT5 is pre-trained in more languages and has more parameters, it is less effective than JointEDI. Since the two models have different pre-trained methods if the input is "A-B-E", the output of mBART is labeled "ABCDE",

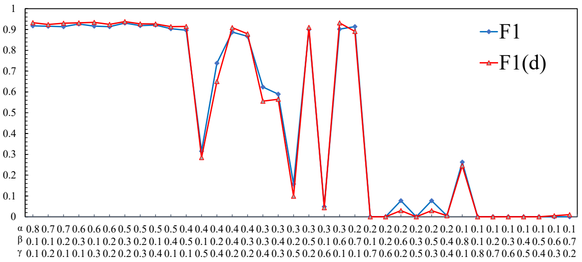


Figure 5: For hyperparameter analysis in the BME dataset, the blue color represents the value of $F1$ and the red color represents the value of $F1_d$.

but the output of mT5 is labeled "CD". It seems that mBART is performing a more difficult task, and is more effective in detecting and identifying euphemisms.

As can be seen in Table 4, in the euphemism detection task, we observe that the $F1_d$ value of JointEDI in the English BME dataset is only 0.11 lower than that of PET by Keh et al. (2022). In the Chinese BME dataset, the $F1_d$ value of JointEDI is optimal. This shows that JointEDI has a competitive advantage among methods that perform the euphemism detection task alone.

JointEDI is not as good as the other baselines in terms of R_d for Chinese data. We analyze that this is due to the higher number of parameters in the compared baselines. This results in a model that tends to predict the presence of euphemisms more frequently, thereby yielding a higher R_d value.

Comparison with LLMs: Table 3 summarizes the results of the comparison between JointEDI and LLMs. We note that GPT-4.0 has the best results among all the LLMs, which is analyzed since GPT-4.0 has the largest number of parameters compared to the other LLMs. Despite this, the performance of our proposed JointEDI is still able to outperform LLMs, which suggests that unifying the euphemism detection task with the euphemism identification task is still challenging for LLMs.

Comparison with Human Evaluation: Although JointEDI achieves the best results, it is still not as good as the human-professional assessment on the BME Chinese dataset, we analyze that this is because mBART uses a larger English corpus than a Chinese one for pre-training, and thus will perform better on English data. At the same time, euphemisms are usually informal and subjective expressions whose form and meaning may change in different contexts. The same sentence may be regarded as a euphemism or as a direct expression in different contexts. This subjectivity and variety

1) *Correct detection and Incorrect identification:*
Input: White boy, buys <coke> from me.
Output: Euphemism Label: 1; Class Label: death
2) *Incorrect detection and Correct identification:*
Input: ... kinetic impact devices , and chemical control substances, tear gas and pepper <spray>…
Output: Euphemism Label: NULL; Class Label: weapon
3) *Incorrect detection and Incorrect identification:*
Input: …relax in stressful situations, <neutralize> interoffice conflict, add zest to dull relationships…
Output: Euphemism Label: 1; Class Label: politics

Figure 6: Analysis of different error types. (1) The true label of "coke" is 1, but it is categorized into the wrong category "death"; (2) The true label of "spray" is 1, the model does not predict the result, the output is "NULL" and it is categorized into the right category "weapon"; (3) The true label of "neutralize" is 0, the model incorrectly predicts 1, and it is divided into the wrong category "politics".

increase the difficulty of understanding and identifying euphemisms in JointEDI.

Hyperparametric Analysis: We have analyzed the ablation of different combinations of α, β and γ. The experimental results are shown in Figure 5. When 0.5, 0.3, and 0.2 are selected for α, β, and γ respectively, the results of the model are the most optimal. We also find that when the sum of β and γ is greater than 0.5, the performance of JointEDI starts to be general and becomes unstable. For example, when α, β, and γ take values of 0.4, 0.1, and 0.5, respectively, the results of the model plummet to near 0. It shows that the loss of uniform euphemism detection and euphemism identification plays a dominant role in the task of unifying euphemism detection and euphemism identification, followed by the loss of euphemism detection, and finally the loss of euphemism identification.

Error Analysis: We have selected three instances in the results of error detection and error identification, respectively. As shown in Figure 6. The first case is correctly detected and incorrectly identified, the second case is incorrectly detected and correctly identified, and the third case is both detected incorrectly and identified incorrectly. This shows that JointEDI is still challenging in domain-specific, context-specific, or type-specific euphemism detection and identification tasks. We show some examples of LLMs in euphemism detection and identification tasks in the Appendix C.

7 Practical Implications

This paper provides a new benchmark to unify the euphemism detection task with the euphemism identification task. Firstly, the method can be directly applied to social media to assist platforms in filtering offensive, inappropriate, or controversial content and reduce the auditing cost. Second, the method can be integrated into a large language model to deepen contextual understanding, detect euphemisms more accurately, and provide users with more accurate and sensitive responses by learning from large-scale corpora. Finally, this technology can facilitate the quality of cultural interactions on social media.

8 Conclusion

In this paper, we construct a bilingual multi-category euphemism dataset named BME, which contains two languages, English and Chinese. The BME dataset covers more than a dozen categories, which provides a new benchmark for research on euphemism detection and euphemism identification tasks. Meanwhile, we also propose a novel generative approach to unify the euphemism detection task and the euphemism identification task, which proves the effectiveness of our proposed JointEDI and the difficulty of this task by comparing it with LLMs and human evaluation. New insights and reference standards are provided for the research on the euphemism task.

Limitations

Although the work in this paper achieves certain results. However, the following limitations still exist: (1) The BME dataset still covers a relatively limited number of languages, and future efforts are needed to expand the language scope to achieve true multi-lingual euphemism detection and euphemism identification. The current dataset is relatively monolingual, while there exists a rich linguistic diversity globally, including about 7,000 active languages. (2) The proposed JointDEI still has a performance gap when compared to a professional human evaluation. This suggests that the current method still needs to be further improved and optimized for more accurate euphemism detection and euphemism identification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62272188), the Fundamental Research Funds for the Central Universities 2662021JC008. We thank all the anonymous reviewers for their valuable comments.

Ethics Statement

We strictly adhere to the data usage agreements of the various public online social platforms. The opinions and findings in the sample dataset we have provided should not be interpreted as representing the views expressed or implied by the authors. We hope that the benefits of our proposed resources outweigh the drawbacks. All resources are intended for scientific research only.

References

Paul Chilton. 1987. Metaphor, euphemism and the militarization of language. *Current research on peace and violence*, 10(1):7–19.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing, Fig-Lang@ACL 2020, Online, July 9, 2020*, pages 136–145. Association for Computational Linguistics.

He Feng. 2021. *Detection and Data Mining of Diabetic Retinopathy Using Classic Machine Learning*. Ph.D. thesis, National University of Singapore (Singapore).

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, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2658–2671. European Language Resources Association.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing, Fig-Lang@ACL 2020, Online, July 9, 2020*, pages 1–11. Association for Computational Linguistics.

Raymond W Gibbs. 1999. Taking metaphor out of our heads and putting it into the cultural world. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 145–166.

Yuxue Hu, Mingmin Wu, Ying Sha, Zhi Zeng, and Yuqi Zhang. 2023. A self-supervised learning method for euphemism identification. *Journal of Chinese Information Processing*, 37:55–63+75.

Liang Ke, Xinyu Chen, and Haizhou Wang. 2022. An unsupervised detection framework for chinese jargons in the darknet. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 458–466. ACM.

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.

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, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. *CoRR*, abs/2211.13327.

Patrick Lee, Iyanuoluwa Shode, Alain Chirino Trujillo, Yuan Zhao, Olumide E. Ojo, Diana Cuevas Plancarte, Anna Feldman, and Jing Peng. 2023. FEED pets: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. In *Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics, *SEM@ACL 2023, Toronto, Canada, July 13-14, 2023*, pages 437–448. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 93–100. Association for Computational Linguistics.

Devansh Mody, YiDong Huang, and Thiago Eustaquio Alves de Oliveira. 2023. A curated dataset for hate speech detection on social media text. *Data in Brief*, 46:108832.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

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

Md. Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mücahid Kutlu, and Matt Lease. 2021. An information retrieval approach to building datasets for hate speech detection. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. LAHM : Large annotated dataset for multi-domain and multilingual hate speech identification. *CoRR*, abs/2304.00913.

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 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1027–1041. USENIX Association.

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 *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 229–246. IEEE.

Appendix

A Data Labeling Training Guide

Your task is to determine, given a sentence containing a potential euphemism, whether PET behaves as a euphemism in the sentence, and if so, to label it as "1" and indicate to which category of euphemisms it belongs. If not, label it as "0". There are 9 categories in English and 8 categories in Chinese. Please make sure that the entire labeling process is free from outside interference and pay attention to the context of the text when labeling to ensure that the euphemisms are accurately captured. In case of uncertainty or ambiguity, mark according to your best judgment.

Example 1:

"My little bro smokes mad <weed>". Could you please indicate whether weed is a euphemism

	English	Chinese
Zhu et al. (2021)	834	-
Lee et al. (2023)	2,039	2,004
Lu et al. (2023)	-	1,775
crawl	1,639	716
In total	4,512	4,495

Table 5: Data sources for the BME dataset.

in the sentence, and if so, which of the following categories does the euphemism belong?

"body functions/parts", "politics", "sexual activity", "physical/mental attributes", "death", "substances", "weapon", "employment/finances".

Labeling result: "1", "substances".

Example 2:

"...供桌供案主要应用于纪念<仙逝>长辈和敬供先人...". Could you please indicate whether there is a euphemism in the sentence and if so, to which of the following categories does the euphemism belong?

"同性恋", "地区", "就业/财务", "性", "性别", "死亡", "种族", "身体/心理属性", "身体功能/部位".

Labeling result: "1", "死亡".

B Data Analysis

The BME English dataset is divided into a total of 8 categories:

"body functions/parts", "politics", "sexual activity", "physical/mental attributes", "death", "substances", "weapon", "employment/finances".

The BME Chinese dataset is divided into a total of 9 categories:

"同性恋", "地区", "就业/财务", "性", "性别", "死亡", "种族", "身体/心理属性", "身体功能/部位".

The top 10 keywords for each category in the BME dataset. Table 5 shows the sources of the MME dataset. Table 6 and Table 7 show the BME English dataset and Chinese dataset, respectively.

C Experiments

The final results of the two datasets by dividing the training set, the validation set and the test set are shown in Figure 8 and Figure 9.

- **mBART-large-cc25**: Pre-trained mBART using 25 languages.
- **mBART-large-50**: Pre-trained mBART using 50 languages.
- **mT5-base**: Pre-trained mT5 base model for 101 languages with 580 million parameters (Xue et al., 2021).
- **mT5-large**: Pre-trained mT5-large model in 101 languages with 1.2 billion parameters.
- **DAN**[4]: Deep averaging network over embeddings for all the tokens of the sentence.
- **CLS**: Using embeddings of the [CLS] classifier token.
- **PET**: Using embeddings of the PET embeddings.
- **mBERT**[5]: A multilingual BERT model.
- **XLM-RoBERTa-base**: A variant of Facebook's XLM-RoBERTa model.
- **XLM-RoBERTa-large**: A larger version of the XLM-RoBERTa model with more parameters.

Baselines: The configurations of the ten baseline models are as follows:

Details of LLMs setup: The test templates for all large language models are shown in Table 8. The configurations of the six LLMs are as follows:

- **Falcon**: A new series of large-scale language models created by the Technology Innovation Institute in Abu Dhabi, with 40 billion parameters.
- **StableLM**[6]: A Stable Diffusion startup, Stability AI, released and open-sourced a large language model trained by the team with 7 billion parameters.
- **mPLUG-Owl**[7]: A large multimodal model based on a modular implementation with 7 billion parameters.
- **LLaMA2-70b-chat**[8]: A Meta AI official release of the latest generation of open source big models with 70 billion parameters.
- **GPT-3.5-turbo**[9]: A fourth in a series of NLP models designed by OpenAI, with 20 billion parameters.
- **GPT-4.0**: A large-scale, multimodal artificial intelligence model developed by OpenAI.
- **Human Evaluation**: We invited four professionals in the field of linguistics and four non-professional people to evaluate 200 English texts and Chinese texts, respectively.

[4] https://github.com/marsgav/euphemism_project
[5] https://github.com/pl464/euph-starsem-2023
[6] https://replicate.com/stability-ai/stablelm-tuned-alpha-7b
[7] https://modelscope.cn/studios/damo/mPLUG-Owl/summary
[8] https://huggingface.co/models?other=llama-2
[9] https://platform.openai.com/docs/api-reference/introduction

Category	Top 10 keywords
body functions/parts	accident, rear end, time of the month, accidents, droppings, chest, tinkle, lavatory, pass gas, latrine
politics	fishing, trick, underdeveloped, pro-life, inner city, wolf pack, armed conflict, undocumented immigrants, freedom, Global South
sexual activity	go all the way, work, sleep around, sex workers, birds and the bees, same-sex, sex worker, porn, slept with, girls
physical/mental attributes	special needs, disabled, plump, aging, slim, expecting, overweight, over the hill, troubled, mixed up
death	late, demise, collateral damage, put to sleep, perish, pass on, long sleep, deceased, pass away, fatality
substances	weed, coke, intoxicated, speed, pure, weeds, sober, substance abuse, inebriated, Coke
weapon	machine, shot, shoot, car, heavy, German, bear, saw, police, spray
employment/finances	disadvantaged, let go of, sanitation workers, economical, dismissed, deprived, well off, income inequality, homemaker, indigent

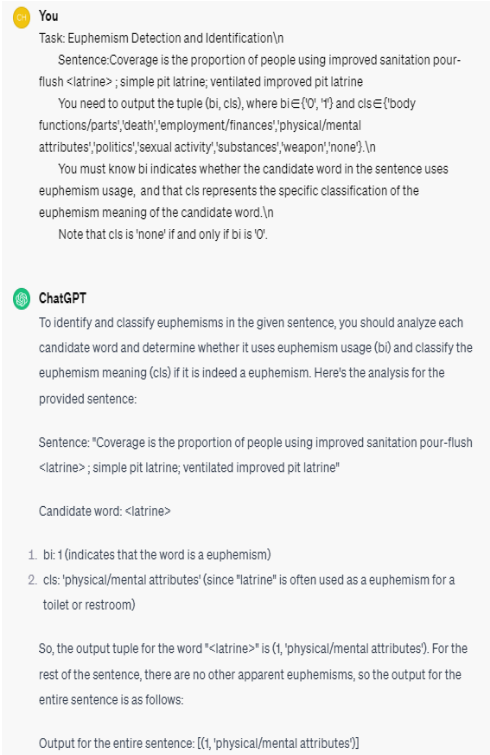
Table 6: Top 10 keywords for each category in the BME English dataset.

Category	Top 10 keywords
同性恋	基佬, 反同, 男同, txl, 通讯录, 同志, gay, 男童, 恐同, 撑同
地区	棒子, 弯弯, 南满, 飞舟, 蛮, 小日子, 南蛮, 九头鸟, 飞周, 蛮夷
就业/财务	调动, 环卫工人, 结构优化, 滑铁卢, 下岗, 低收入, 辞退, 负增长, 拮据, 裁员,
性	小姐, 失足, 上床, 亲热, 性侵, 慰安妇, 房事, 夫妻生活, 同房, 第三者
性别	女拳, 仙女, 打拳, 普信, 拳师, 幕刃, 亩, 牧人, 圣母, eg
死亡	牺牲, 没了, 走了, 解脱, 不在了, 挂了, 过世, 逝世, 遇难, 去世
种族	默, 猩猩, 虫类, 黑猩猩, 黑女, 嘿嘿, 类人猿, 媚黑, 小黑, 三非
身体/心理属性	丰满, 有喜, 发福, 年长, 年迈, 苗条, 高龄, 失明, 特殊人群, 长寿,
身体功能/部位	姨妈, 方便, 胸部, 卫生间, 洗手间, 大号, 下身, 私处, 生理期, 如厕

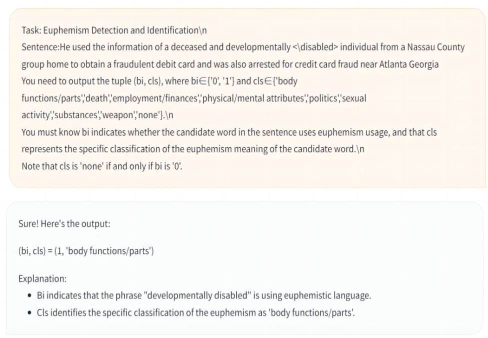
Table 7: Top 10 keywords for each category in the BME Chinese dataset.

	Templates
English	Task:Euphemism Detection and ldentification Sentence:'...<PET>...' You need to output the tuple (bi, cls), where bi ∈ {'0','1'} and cls ∈ {'bodyfunctions/parts', 'death', 'employment/finances','physical/mental attributes','politics','sexual activity', 'substances','weapon ','none'}. You must know bi indicates whether the candidate word in the sentence uses euphemism usage, and that cls represents the specific classification of the euphemism meaning of the candidate word. Note that cls is 'none' if and only if bi is '0'.
Chinese	任务:委婉语检测与识别 句子:'...<PET>...' 您需要输出元组 (bi，cls)，其中 bi ∈ {'0','1'}，cls ∈ {'同性恋','地区','就业/财务', '性','性别','死亡','种族','身体/心理属性','身体功能/部位','none'}。 您必须知道bi表示句子中的候选词是否使用了委婉用法，cls表示候选词委婉含义的具体分类。请注意，只有当bi为'0'时，cls才是'none'。

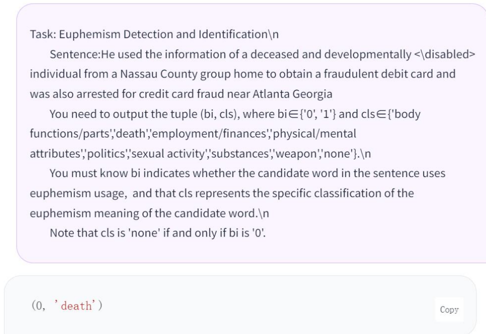
Table 8: Templates for LLMs testing.



(a) Case of ChatGPT



(b) Case of LLaMA2-70b-chat



(c) Case of mPLUG-Owl

Figure 7: Cases of some of LLMs.

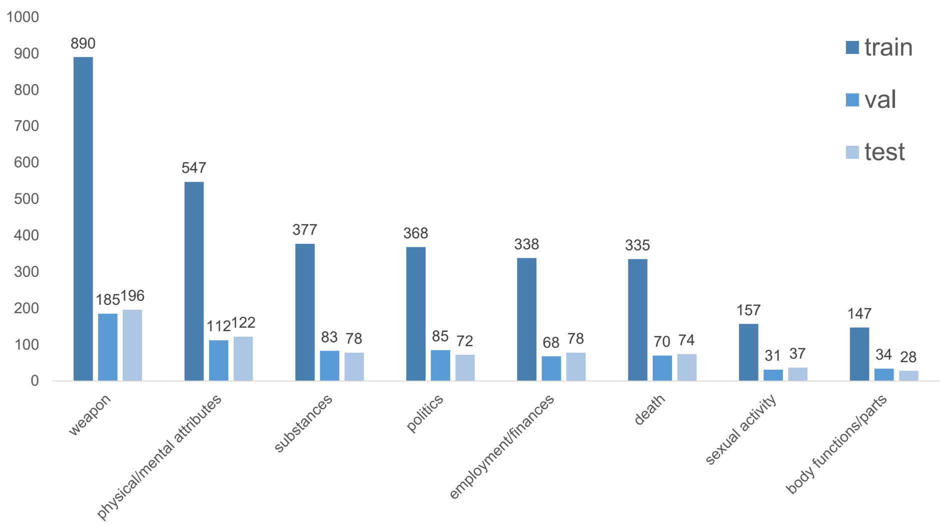


Figure 8: The final results of the BME English dataset by dividing the training set, validation set, and test set.

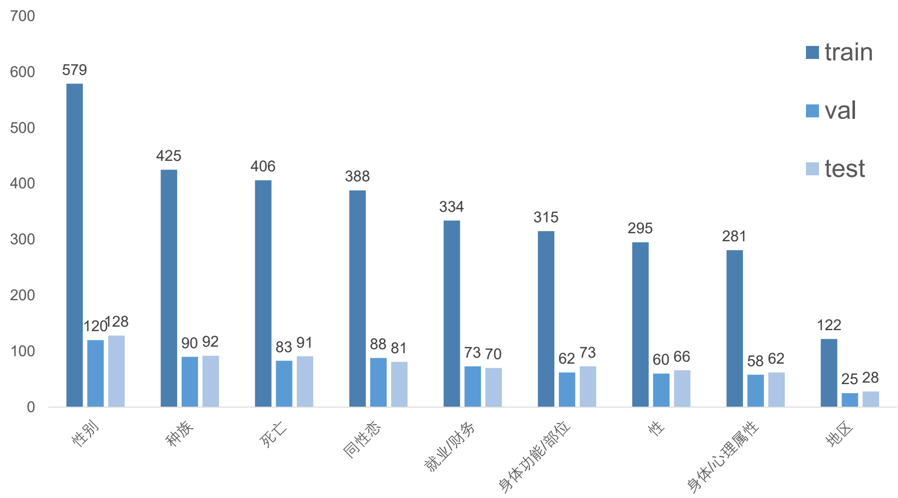


Figure 9: The final results of the BME Chinese dataset by dividing the training set, validation set, and test set.