# The Multilingual Euphemism Benchmark: Datasets and Baselines for Pragmatic Language Understanding

**Whitney Poh**[1], **Julia Sammartino**[1], **Jasper Andrew**[1], **Witold Kieraś**[2], **Natalia Zawadzka-Paluektau**[2], **Iryna Dilai**[3], **Libby Barak**[1], **Jing Peng**[1], **Anna Feldman**[1]
[1] Montclair State University, New Jersey, USA
[2] Institute of Computer Science, Polish Academy of Sciences, Poland
[3]Ivan Franko National University of Lviv, Ukraine

**Abstract**

Euphemisms are words or phrases used to soften or indirectly refer to taboo or sensitive topics. They pose interpretation challenges because the same expression may appear in different senses depending on context: literal, figurative but non-euphemistic, or euphemistic. For example, *pull the plug* may refer euphemistically to ending a patient's life support, figuratively to canceling a project or funding, or literally to unplugging a device. Euphemisms also vary across languages and cultures in both their surface forms and the contexts in which they are conventionally used. Previous work introduced datasets for the computational study of euphemisms in five languages. We extend this line of work by introducing two new annotated datasets for euphemism detection in Polish and Ukrainian and by standardizing resources for all seven languages into a unified benchmark format that supports cross-lingual evaluation. Finally, we provide zero-shot and few-shot baselines using GPT-5-nano. We run each configuration five times and report the average score, establishing reference scores for multilingual pragmatic understanding. We also performed pilot tests using Qwen3-4B on the English and Chinese datasets.

**Keywords:** euphemisms, multilingual benchmark, language resources

## 1. Introduction

Euphemisms are words or phrases used in order to soften language pertaining to impolite, inappropriate or sensitive topics such as 'bodily functions', 'death' and 'sex' (Lee et al., 2024). Euphemisms are often used for positive reasons, such as to avoid hurting feelings or causing discomfort, but they can also be used to obfuscate unpopular positions in political contexts or pass content moderation filters (Grolleau et al., 2022; Zhu and Bhat, 2021; Gavidia et al., 2022). Understanding the context of use regarding euphemistic speech could help identify social, political and linguistic norms within cultures. These phrases introduce interpretation challenges since they may be used in figurative or literal meaning depending on the context, e.g. "a certain age", referred to as Potentially-Euphemistic Terms (PETs) (Gavidia et al., 2022).

In this study, we will introduce two new euphemism classification datasets (one in Ukrainian and one in Polish), and release GPT-5-nano (Singh et al., 2025) classification baselines for the two new datasets, as well as the English, Spanish, Chinese and Yorùbá datasets from Lee et al. (2024) and the Turkish dataset from Biyik et al. (2024). These baselines will help measure current state-of-the-art (from hereon referred to as SOTA) LLM abilities to classify whether a PET is used in a euphemistic manner or not.

Our contribution will help the community researching euphemisms and their use by providing more data, which is generally quite scarce with regards to this topic. Furthermore, our baselines show that while LLMs have improved in terms of identifying euphemisms in languages like English and Chinese, they lag behind in lower resource languages like Ukrainian and Yorùbá.

Even with regards to English and Chinese, euphemism detection abilities are rather poor on both the few-shot and zero-shot settings, with few-shot surprisingly performing worse. For low resource languages like Ukrainian and Yorùbá, performance was only slightly better or even worse than random guessing. Unlike English and Chinese, few-shot performance was generally better for lower resource languages than zero-shot performance.

We make an additional contribution by releasing the datasets for all seven languages in a standardized form that allows smooth comparison, benchmarking, and integration. We identify missing translations or annotations, and ensure a cohesive and consistent format for all data. By standardizing the datasets gathered by Lee et al. (2024), Biyik et al. (2024) and Dilai et al. (2025) into a uniform format, and introducing a new Polish dataset and new Ukrainian dataset (adding non-war euphemisms to the work of Dilai et al. (2025)), we allow the baselines to be compared. The standardization process includes collapsing some categories in the original dataset gathered by Ga-



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7307–7319
11–16 May 2026. 

vidia et al. (2022); Lee et al. (2023, 2024), and introducing a subcategory column for the Polish and Ukrainian dataset, as some categories were more specific than the other language datasets. Our preliminary analysis using GPT-5 exemplifies how the unified format allows a deeper analysis of euphemism classification within and across languages.

## 2. Related Work

Research on euphemisms in NLP has advanced along two main tracks: building resources and designing detection methods. Early corpora such as Gavidia et al. (2022) introduced potentially euphemistic terms (PETs) with matched euphemistic vs. literal contexts and showed sentiment attenuation in euphemistic usage. Lee et al. (2022b) proposed a mining pipeline using distributional neighbors and sentiment cues to surface PET candidates. Shared tasks at FigLang 2022 and 2024 standardized evaluation; the latter added multilingual benchmarks in English, Spanish, Yorùbá, and Mandarin (Lee et al., 2022a; Lee and Feldman, 2024). Follow-ups expanded coverage: Lee et al. (2023) added vagueness annotations and new PET corpora (Yorùbá, Spanish, Mandarin), Lee et al. (2024) evaluated multilingual and cross-lingual PET disambiguation with XLM-R (Conneau et al., 2020), and Biyik et al. (2024) released a Turkish PET dataset. Recent work collected Ukrainian war-related euphemisms and tested LLM prompting, underscoring rapid, culturally bound drift (Dilai et al., 2025).

On the methods side, Zhu and Bhat (2021) detect euphemistic phrases via phrase mining and SpanBERT ranking, while Zhu et al. (2021) introduce unsupervised detection and code-word identification for moderation. Work on covert "dog-whistle" communication motivates context-sensitive modeling (Bhat and Klein, 2020). Additional architectures and settings include RGAT models (Wang et al., 2022), "impromptu" euphemisms in cybercrime (Li et al., 2025a), domain-specific historical death discourse (Al-Laith et al., 2025), and LLM prompting (GPT-4 (OpenAI et al., 2024)) across languages (Firsich and Rios, 2024). Firsich and Rios (2024)'s paper, just like ours, deals with zero-shot and few shot prompting techniques for euphemism detection. Keh (2022) also looks into zero-shot and few-shot prompting techniques for euphemism detection. Finally, cross-lingual transfer is a central challenge: sequential fine-tuning can help in low-resource languages, but gains depend on data coverage and model choice (Sammartino et al., 2025). Our work contributes expert-annotated Polish and Ukrainian PET corpora with balanced euphemistic/literal usages to support both monolingual and cross-lingual evaluation.

| Lang | PETs | PETs-AE | PETs-AE% | Uses | Use-AE | Use-AE% |
|---|---|---|---|---|---|---|
| EN | 135 | 58 | 43.0 | 3098 | 745 | 24.0 |
| ES | 232 | 97 | 41.8 | 2952 | 1045 | 35.4 |
| YO | 157 | 72 | 45.9 | 2598 | 864 | 33.3 |
| ZH | 151 | 95 | 62.9 | 3211 | 1625 | 50.6 |
| TR | 70 | 11 | 15.7 | 2436 | 281 | 11.5 |
| PL | 522 | 38 | 7.3 | 2458 | 62 | 2.5 |
| UK | 148 | 46 | 31.1 | 7163 | 1112 | 15.5 |

Table 1: Inventory of PETs and uses; AE = always-euphemistic in our data.

## 3. Multilingual Euphemism Corpus

| Category | EN | ES | ZH | YO | TR | PL | UK |
|---|---|---|---|---|---|---|---|
| Bodily Parts/ Functions | 5.5 | 14.1 | 13.5 | 29.6 | 23.3 | **20.1** | 7.0 |
| Death | 19.0 | 5.1 | 17.9 | 21.5 | **30.8** | 14.0 | - |
| Employment/ Finances | 17.8 | 23.6 | 17.5 | 0.1 | 8.4 | 9.0 | - |
| Illegal Activity | - | - | 2.9 | - | - | 7.9 | - |
| Misc. | 5.0 | - | 0.4 | 0.4 | 8.1 | 5.9 | 7.0 |
| Phys/Mental Attributes | **26.6** | 16.8 | 16.6 | 4.7 | 14.4 | 15.7 | 4.7 |
| Politics | 10.3 | **32.1** | 7.7 | - | 0.5 | 3.6 | **74.9** |
| Sexual Activity | 9.6 | 6.3 | **21.8** | **30.8** | 11.2 | 19.5 | - |
| Social Activity | - | - | 1.3 | 10.5 | - | - | - |
| Spirituality | - | - | - | 2.5 | - | - | - |
| Substances | 6.2 | 1.9 | 0.3 | - | 3.4 | 4.4 | 6.4 |

Table 2: Category Percentages for each Dataset (bold = largest category)

### 3.1. Languages Covered

We include seven languages representing different language families and levels of resource availability: English, Spanish, Chinese, Yorùbá, Turkish, Polish, and Ukrainian. The selection balances typological diversity (analytic, fusional, and agglutinative structures) and cultural variation in euphemistic expression. English, Spanish, and Chinese are relatively high-resource (Li et al., 2025b); Yorùbá, Turkish, Polish, and Ukrainian extend coverage to low- and mid-resource settings with distinct pragmatic norms. Dataset sizes range from about 2K-7K labeled examples per language.

Our standardized data contains the following information for all languages: (1) `text`- the text containing the PET, (2) `pet` - the potentially euphemistic term in its original form, (3) `euph_status` - whether the PET is always used euphemistically or only sometimes (Sammartino et al., 2025), (4) `category` - the core category of the euphemistic phrase, such as 'sexual activity' or 'death', (5) `label` - the classification (1 for euphemistic, 0 for non-euphemistic). Most `text` entries are multi-sentence passages, and each contains one PET within `[PET_BOUNDARY]` delimiters. `euph_status` indicates whether each PET

always appears in a euphemistic context within the dataset ("always_euph"), or may appear in both euphemistic and non-euphemistic contexts ("sometimes_euph"). As an example of a sometimes-euphemistic PET, "lay off" can be used in a euphemistic context ("the company laid off all their employees") and a non-euphemistic context ("lay off the carbs, will ya?").

Some datasets also contain additional data, like "PET_wordform" for languages like Polish that undergo morphological changes depending on factors such as tense or plurality, an English translation of the PET, or a source for the text. We also chose to consolidate the categories into core-categories shared among all datasets to allow for future semantic and pragmatic analysis, while preserving any data-specific annotation in an additional column. For example, the categories of 'alcoholism' and 'drugs' in the Ukrainian dataset were collapsed into 'substances' under the core-categories. See Table 2 for the full list of categories and their frequency in each dataset (by instances).

## 3.2. Previously Released Languages

### 3.2.1. English

The English dataset, presented in (Gavidia et al., 2022) and then refined in (Lee et al., 2023, 2024; Lee and Feldman, 2024), contains over 3,000 entries and 135 unique PETs, with entries pulled from the GloWbE corpus (Davies and Fuchs, 2015). The "always_euph" category contains 58 of the 135 unique PETs, but only 745 of 3,098 instances, leaving 2,353 in the "sometimes_euph" category. The dataset is skewed slightly toward PETs in euphemistic contexts.

This dataset consists of euphemisms from eight categories, including 'death', 'physical/mental attributes', 'employment/finances' and 'politics'.

### 3.2.2. Spanish

The Spanish dataset (Lee et al., 2023, 2024; Lee and Feldman, 2024) contains just shy of 3,000 utterances, with the majority labeled as euphemistic (1,955). Lee et al. (2023, 2024); Lee and Feldman (2024) extracted the text used in the dataset from Real Academia Española (Real Spanish Academy) (Corpes Siglo XXI) (Real Academia Española, 2019), which consists of text from multiple Spanish-speaking countries such as Mexico, Cuba and Spain. Examples of PETs from this dataset include "hacer el amor" (make love), "expirar" (expire) and "consumar el matrimonio" (consummate the marriage).

This dataset contains 232 PETs, with 97 of them being "always_euphemistic". For example, "pronunciamiento militar" and "capital humano" are only used euphemistically in the dataset.

### 3.2.3. Mandarin

The Chinese dataset (Lee et al., 2023, 2024; Lee and Feldman, 2024) contains about 3,200 entries, with about two-thirds of them labeled as euphemistic. Examples of PETs from this dataset include "不在了" ("no longer here", referring to death), and "有喜" ("to have joy/happiness") , referring to pregnancy. It should be noted that in this dataset, the majority of PETs were labeled as "always_euphemistic". For example, "卫生间", literally "hygiene space", euphemistically meaning "bathroom", is only used in the euphemistic context.

A large portion of this dataset (21.8%, as shown in Table 2) consists of euphemisms for topics related to sexual activity. Examples of such euphemisms include "性工作者", and "小姐", meaning sex worker.

The data was gathered (Lee et al., 2023, 2024; Lee and Feldman, 2024) from the Chinese language corpus (Xu, 2019).

### 3.2.4. Turkish

The Turkish dataset was gathered by Biyik et al. (2024). The Turkish dataset consists of 2436 instances of sentences consisting of 70 unique PETs, with only 11 of the PETs being "always_euphemistic", represented by 281 of the entries. This ratio of "always_euphemistics" is lower than English, Chinese and Ukrainian.

Examples of PETs from this dataset include "aramızdan ayrıldı" ("left from among us") from the 'death' category, and "aybaşı" ("beginning of the month") from the bodily functions category.

Turkish is an agglutinative language, and the meaning of a sentence in Turkish is rarely affected by the order of its words (Biyik et al., 2024). Similar to Ukrainian and Polish, which feature word conjugations, the Turkish dataset consists of a column for the lemmatized form of the PET, along with the conjugated form column.

### 3.2.5. Yorùbá

The Yorùbá dataset (Lee et al., 2023, 2024; Lee and Feldman, 2024) consists of almost 2,600 instances. Like with the other languages, the dataset is skewed towards examples in euphemistic contexts, with 1,689 euphemistic examples. Table 2 shows that a majority of this dataset falls into the 'bodily functions/parts', 'sexual activity' and 'death' categories.

Yorùbá is considered a low-resource language for the purposes of euphemism detection and in

computational modeling of language in general, which led the dataset's creators to gather data from various sources, including Facebook conversations, religious texts, and news data (Lee et al., 2023).

### 3.3. Newly Released Languages

#### 3.3.1. Polish

The Polish dataset includes almost 2500 entries, with two methods of PET identification employed. First, we searched for euphemisms listed in a Polish dictionary of euphemisms (Dąbrowska, 1998), in general-language corpora – namely the National Corpus of Polish (Przepiórkowski et al., 2012) – and the Corpus of Contemporary Polish (Marciniak et al., 2023). For each of them, we noted the approximate numbers of occurrences, and whether they had only euphemistic (e.g. piwko [ 'small beer' ]) or both euphemistic and non-euphemistic uses (e.g. siano [ 'hay' or 'money' ]). We then conducted separate searches of the PETs of the latter type, and selected between four and six examples of each, including both euphemistic and non-euphemistic uses in equal proportions. However, we have also aimed to identify PETs which were not listed in Dąbrowska' s dictionary to compensate for the fact that the dictionary was compiled almost three decades ago using mainly opportunistic methods, and is therefore less likely to adequately reflect contemporary linguistic reality. To this end, we selected words or expressions associated with a given taboo topic (since speakers discussing taboo or sensitive topics often rely on the use of euphemisms), ran concordance searches, and looked for additional PETs in the co-text. Then, we again conducted separate concordance searches of thus identified PETs, noted their approximate frequencies, and selected between four and six examples of those which can be used both euphemistically and non-euphemistically. The taboo topics deemed to have the potential to trigger the use of euphemisms were chosen based on the existing literature (Dąbrowska 1998; Allan and Burridge 1991), and include: 'alcohol and drugs', 'body parts and nakedness', 'death and illness', 'immigration', 'sex', 'money', 'offenses', 'physical appearance', 'physiology', and 'vice'. The numbers of PETs per topic are more or less equal. To enable comparison between languages, we annotate each PET with the uniform categories described in Table 2, while preserving the detailed information in an additional column, e.g., annotate PETs as 'substances' with the additional column being either 'drugs' or 'alcohol'.

Five hundred samples, covering five different topics, were collected by an expert linguist (one of the authors). One hundred sentences were marked as either euphemistic (1) or non-euphemistic (0) by two colleagues, based on a detailed annotation guide, in order to test the procedure. Next, the full dataset (i.e. 500 sentences) was annotated by four students of linguistics following the same guide ($\alpha = 0.7041$). All of the annotators agreed on one interpretation (either 0 or 1) in 366 cases. The strongest disagreement (where half of the annotators decided that the given expression or word is euphemistic, whereas the other half felt otherwise) was observed in 42 cases.

The remainder of the dataset was prepared by the four annotators following the same procedure as the one discussed above. Each of them contributed $500$ sentences, and then marked the samples collected by her counterparts as either euphemistic or non-euphemistic, following the annotation guide. For this subset of the data, the inter-annotator agreement is even higher at $\alpha = 0.8520$.

#### 3.3.2. Ukrainian

The non-war portion of the Ukrainian dataset includes 2297 sentences containing 113 PETs. The subcategories covered are bad habits, bodily functions, corruption, impairments, lying. For completeness, we add to the new Ukrainian data the euphemism related to the 'war' category included in Dilai et al. (2025) (related to the ongoing Russia-Ukraine war). The corpora used for the dataset collection are the Polish Automatic Web corpus of the Ukrainian language, or PAWUK (Kieraś et al., 2025) and the General Regionally Annotated Corpus of Ukrainian, GRAC (Shvedova et al., 2017-2025). We collected samples of both euphemistic and non-euphemistic usages of the PETs. Some of the PETs had only euphemistic usage (e.g. дитина дощу ['rain child']). Most PETs due to their polysemy were used both literally (e.g. бавовна ['cotton']) and euphemistically (e.g. бавовна ['explosion']). The proportion of the collected euphemistic instances is dictated by their relative frequency in the corpora. Some PETs had fewer examples of usage due to the limited corpus data. We tried to collect and annotate an equal number of PETs for each category, with the exception of the 'war' category taken from Dilai et al. (2025).

For the time being there is no comprehensive dictionary of Ukrainian euphemisms. Consequently, we used various academic and non-academic sources to obtain a seed list of euphemisms for each category. The list is by no means complete. Both data collection and annotation were conducted by four expert linguists. The task was to attach a label (1) to euphemistic usages and a label (0) to non-euphemistic usages. Additionally, the annotators indicated instances of uncertainty, which needed special attention on

the part of other annotators. The inter-annotator agreement was calculated ($\alpha = 0.7$). Certain PETs showed worse agreement than others.

## 4. Resource Description and Availability

All datasets introduced in this paper – Polish and Ukrainian PET corpora, along with the standardized multilingual euphemism benchmark – are released under an open license [1]. Each dataset follows the unified schema introduced in Section 3.1: `text`, `pet`, `euph_status`, `category`, and `label`. Annotation guidelines, frequency statistics, and example scripts for evaluation are included in the release package.

### 4.1. Dataset Partition Design

The trial split follows the setup used in previous work by (Sammartino et al., 2025), where splits are done based on the information in `euph_status`. If a PET is sometimes euphemistic in our data, such as "pass on", then it can appear anywhere in training, validation, or testing. If the PET presents as always euphemistic in the data, such as "negative cash flow", then it can only appear in training/validation *or* testing. This is done to prevent the model from memorizing the same classification without considering the context of the example.

Since these datasets are focused on *potentially* euphemistic terms, the percentage of always euphemistic instances for most languages are less than 50% (See Table 1). The Chinese dataset is close to equal, with 50.6% of utterances within the dataset (and more than 60% of PETs) being always euphemistic.

## 5. Baseline Experiments

Following Firsich and Rios (2024), we ran baseline zero-shot and few-shot experiments with GPT-5-nano, a lightweight version of the GPT 5 model by OpenAI (Singh et al., 2025). We also performed a pilot analysis using Qwen3-4B, a lightweight version of the Qwen3 model (Yang et al., 2025), but encountered unexpectedly low performance. We summarize our preliminary results from using Qwen3 and the challenges of applying it to euphemism classification below.

We performed zero-shot prompting and few-shot prompting with both models. For the few-shot prompting, we selected two positive examples and two negative examples from the training dataset to insert into the prompt. We manually selected examples that clearly illustrate euphemistic and non-euphemistic usage. If the examples are particularly long, surrounding content deemed extraneous may be removed in order to avoid confusing the model.

The base template of the zero-shot prompt (in English) is the same for both models: "Can you classify whether the phrase between [PET_BOUNDARY] delimiters in the text is being used in a euphemistic manner or not? Do not include an intro or conclusion in your answer. Just output the number 1 if its euphemistic, 0 if not. This is the text: '<text>' ". For the non-English languages, it is translated.

The few-shot prompt uses the same structure, with the examples and their expected outputs inserted before the text. The four examples were used with every item in the test data, ensuring the test entry does not include the same PET used in the prompt.

Results are reported for the testing split.

| Language | F1 (Bin) | P (Bin) | R (Bin) | F1 (Mac) |
|---|---|---|---|---|
| English | 0.74 | 0.81 | 0.67 | 0.72 |
| Chinese | 0.75 | 0.85 | 0.67 | 0.69 |
| Turkish | 0.54 | 0.73 | 0.43 | 0.57 |
| Spanish | 0.59 | 0.79 | 0.48 | 0.58 |
| Yorùbá | 0.43 | 0.86 | 0.29 | 0.50 |
| Polish | 0.57 | 0.76 | 0.46 | 0.65 |
| Ukrainian | 0.37 | 0.93 | 0.23 | 0.46 |

Table 3: Zero-shot baseline results for each language.

| Language | F1 (Bin) | P (Bin) | R (Bin) | F1 (Mac) |
|---|---|---|---|---|
| English | 0.66 | 0.86 | 0.53 | 0.68 |
| Chinese | 0.68 | 0.90 | 0.55 | 0.66 |
| Turkish | 0.52 | 0.77 | 0.39 | 0.57 |
| Spanish | 0.55 | 0.81 | 0.42 | 0.56 |
| Yorùbá | 0.54 | 0.84 | 0.40 | 0.56 |
| Polish | 0.58 | 0.78 | 0.46 | 0.66 |
| Ukrainian | 0.46 | 0.94 | 0.32 | 0.51 |

Table 4: Few-shot baseline results for each language.

### 5.1. GPT-5-nano

Tables 3 and 4 show the average binary F1, precision, recall and macro F1 over five runs for each language. Very rarely, the model did not return a 0 or a 1 as an answer. In those cases, we drop those instances from the final averaging. As shown, the GPT-5-nano (Singh et al., 2025) model demonstrated uneven performance in which precision typically outperformed recall, hinting that a larger number of samples were labeled as 0 rather

[1] https://github.com/NLPlabMSU/lrec-euph

than 1 (false-negative). Yet, the zero-shot performance is surprisingly high given that we only present the model with a classification task without explaining what euphemisms are nor providing examples. Since most multi-word expressions in general language use are non-euphemistic, if the model simply matched the distribution of euphemisms vs. non-euphemisms in the general language, it would expectedly have made more non-euphemism predictions. The ability of GPT-5 to reach such performance with zero-shot prompting indicates some internal knowledge representation of what "euphemistic" means.

Overall, our results were lower than Firsich and Rios (2024)'s. This was likely due to the different models used, as we used the nano version of GPT-5. We found that the GPT-5-nano model showed the highest zero-shot performance in English (positive class, aka binary, F1 around 0.74) and Chinese (around 0.75), with a steady decline across other languages, reaching the lowest scores for Ukrainian (around 0.37). This result aligns with language representation in the hypothesized training data for GPT-5 as we review in details in Section 6. It appears that GPT-5 can recognize euphemistic tone in high-exposure languages but fails to do so reliably when the cues are culturally specific. Even with substantial textual resources, languages like Spanish remain challenging, suggesting that exposure to text alone does not guarantee cross-linguistic transfer of social and pragmatic understanding. We elaborate on pragmatic and cross-lingual aspects in Section 6.

In the few-shot setting, Chinese and English remain the strongest performers (binary F1 around 0.68 for Chinese, 0.66 for English), while Turkish, Spanish, and Yorùbá cluster in the mid-range near 0.52–0.55. Ukrainian continues to trail behind, with F1 scores around 0.46. GPT-5-nano shows moderate gains in euphemism classification across low-resource languages such as Yorùbá and Ukrainian, small variations in either direction across mid-resource languages like Turkish and Polish, and high losses across high-resource languages like English and Chinese. These results confirm that limited in-context exposure helps the model align with the task. Spanish is an exception, being a high-resource language, which demonstrated a slight drop in F1 scores, which places its performance closer to the Polish and Turkish category. This behavior is surprising given the general assumption that Spanish is more high-resource than Polish and Turkish since it is the "the second most spoken language in the world" according to Lee et al. (2023, p. 443) as of 2009, citing a book by Lewis (2009). Improvements in Polish, Yoruba and Ukrainian suggest that GPT-5 can adjust its decision boundaries when given examples, but that its underlying pragmatic understanding remains uneven across languages. The relative ranking between languages remains largely unchanged, indicating that the model's limitations stem more from linguistic and cultural familiarity than from the mechanics of few-shot learning.

One possible reason for the decrease in performance for English and Chinese few-shot compared to zero-shot could result from the choice of examples for the prompt. Both positive (euphemistic) examples selected in Chinese were death-related. The euphemisms we selected in the English few-shot were "laid off" and "down there", both of which may be perceived by models as slang as opposed to euphemistic. The gap between precision and recall increased after few-shot prompting for both languages, hinting that the examples we provided may have confused the model.

### 5.2. Qwen3

For Qwen3, we report results for our tests on Qwen3-4B, a lightweight version of the Qwen3 model (Yang et al., 2025). For this, we did two runs, the first one using the same prompt as GPT, with no-think setting on, and the system prompt just telling the model to distinguish euphemisms from non-euphemisms, then output 0 or 1. The results are inferior to GPT-5, partially due to how without special prompting, the model may output 1 for every, or almost every, test entry. In order to combat that, we had to do special prompt engineering.

Our prompt engineering consists of adding a system prompt "You are an expert at deciphering non-literal meanings, especially euphemisms. Think about what aspects of the context could point to whether something is euphemistic or not, and how they contribute to your decision. Focus especially on the one sentence in which the phrase of interest appears, since that is where the euphemistic context, if there is any, would probably appear. For example, 'pass on' could mean death (euphemistic) or passing on information (non-euphemistic). Look at the specific context, and not just whether the phrase itself is typically euphemistic or not. Many typically euphemistic phrases are used in non-euphemistic contexts too! The final output should just be a 0 (if non-euphemistic) or a 1 (if euphemistic)." The system prompt remains in English, while the user prompt is translated into the language of the dataset that is being evaluated. The prompt references the use of the euphemism "pass on" in Sammartino et al. (2025).

After prompt engineering, the model's /no_think mode still occasionally overpredicted the positive

class, although performance improved substantially relative to the initial setup. Even so, it remained below GPT-5 (Singh et al., 2025). In the zero-shot setting, Qwen3 reached positive-class F1 scores of about 0.70 for English and 0.79 for Chinese, but its macro F1 was only 0.57 and 0.58, respectively. This gap suggests an imbalance in performance across classes, likely due to a tendency to overpredict the positive label, which was less pronounced with GPT-5.

In the few-shot setting, the English positive-class F1 decreased to 0.57, while Chinese remained at 0.79. For English, macro F1 and positive-class F1 moved closer (0.64 and 0.57, respectively), suggesting more balanced predictions across classes. For Chinese, however, the gap persisted (0.60 macro vs. 0.79 positive-class F1), indicating continued skew toward the positive label.

## 6. Discussion

In this work, we present two novel datasets of euphemism in Polish and Ukrainian. We also make available a benchmark for a euphemism classification task in 7 languages that cover diverse set of linguistic properties. To facilitate future research, we release a uniform representation of the old and new datasets, with complete information for shared attributes, distributional analysis of the data, and splits for training and testing computational models.

We report results with GPT-5-nano (Singh et al., 2025) for zero-shot and few-shot classifications. These results provide baseline performance measures for future work as well as an analysis of the SOTA computational power of classifying euphemisms. Overall, GPT-5-nano tends to make more false-negative errors than false-positive resulting in higher precision against lower recall, with English and Chinese gaining the highest scores.

The overall performance trend indicates a consistent pattern linked to the model' s training exposure and cultural proximity to English. First, the training data is likely lacking in low-resource languages. Although OpenAI has not openly stated the composition of the training data for the GPT-5 model (Singh et al., 2025), LLM performance often correlates with the availability of training data for a language, which may partially explain the stronger results for English and Chinese.

Moreover, euphemistic expression varies substantially by culture; indirectness, politeness, and taboo avoidance are encoded differently in each language, making it difficult to rely on training data in high-resource languages to extend to other languages. In a zero-shot setting, the model relies on broad semantic cues learned from a training dataset that is most likely largely composed of English, which do not always transfer to other linguistic or cultural systems. This may be especially prominent in low-resource languages like Yorùbá, but might also affect languages with more training data. For example, although Spanish is a relatively high-resource language (Li et al., 2025b), its results were weaker than expected, suggesting that euphemism detection depends on pragmatic understanding stemming from exposure in training rather than only on general language proficiency.

Previous work on euphemism classification looked into the ability of LLMs to extend euphemism knowledge from one language to the other (Lee et al., 2024; Sammartino et al., 2025). Hankins (2024) used the DistilBERT model (Sanh et al., 2019) and determined that a multilingual model tended to outperform models fine tuned on only one language in euphemism detection for English, Spanish, Chinese and Yoruba, implying that information the model learns from each of the different languages can improve understanding of all four languages. We further extend such analysis of cross-lingual transfer in the current results. Certain categories within the datasets, such as 'bodily functions/parts', 'sexual activity' and 'death', tend to have PETs shared across languages with the same literal translation, e.g., "hacer el amor" and "make love", and "去世" and "passed away". We suspect that the performance in some languages gained from this pragmatic similarity to English. On the other hand, Spanish, for example, did not benefit as much from it since such common categories (e.g. for 'bodily functions/parts', 'sexual activity' and 'death') made up less than a third of the Spanish data. On the other hand, the performance for the Chinese data might have benefited from the fact that a large portion of its data has to do with 'sexual activity', 'bodily functions/parts' or 'death'.

The performance of LLMs might also correlate with the difficulty of the task for PETs with weaker annotator consensus. Human annotators can differ in their interpretations for PETs that are more vague (Lee et al., 2023), or commonly used jargon in fields such as medicine or politics (Gavidia et al., 2022). We hypothesize that the types of phrases that do not produce confusion among humans are more likely recognized as euphemistic by SOTA LLMs such as GPT-5 due to their ubiquity. While Lee et al. (2023)'s study, which uses a fine-tuned RoBERTa (Liu et al., 2019) model, showed that vague PETs actually performed better, we do not think that is the case for non-fine-tuned prompted models like GPT-5-nano, since the model would be pretrained on large amounts of human data from different sources rather than fine-tuned on a dataset specifically made to identify euphemisms. Many unambiguous PETs like

"passed away", "backside", and "made love" tend to fall under these three categories, whereas PETs in, for example, medical categories, could be considered either as euphemisms or simply as commonly accepted terminology by different people (Gavidia et al., 2022) and likely by GPT-5 too. However, there are also PETs within these categories of our dataset, such as "six feet under", which is considered euphemistic by us, but dysphemistic by others such as Felt and Riloff (2020). These differences in classification could confuse language models.

Another factor in model performance could be the portion of the dataset that is "always_euphemistics". A large portion of the PETs in the Chinese dataset are "always_euphemistic", which could lead to higher scores since the model would have to only learn the phrase and not the context. However, it should be noted that being "always_euphemistic" on its own does not necessitate a high score, as the Yorùbá dataset, which had one of the lower scores, also had a large portion of "always_euphemistics". For being "always_euphemistic" to help in terms of classification, it seems that a phrase also has to be well-known and/or transferable across languages, and unambiguously euphemistic. The Ukrainian data serves as a good example for the tension between these two factors, frequency of euphemistic use in language vs. general familiarity with the term. The Ukrainian data's scores likely suffered as a majority of the dataset was about war. A lot of the euphemisms in the dataset, like "двохсотий" meaning "dead" (in a military context) or "killed in action" (Wiktionary contributors, 2024; Dilai et al., 2025), are only used in the post-Soviet linguistic sphere as it originates from the term Cargo 200 (Wiktionary contributors, 2025). The high ratio of 0 predictions to 1 predictions for the Ukrainian dataset demonstrates that many Ukrainian euphemisms might not have been recognized by the model, especially since the Ukrainian dataset consists of more euphemistic examples than non-euphemistic examples to begin with.

At the same time, other factors may also contribute. Differences in the composition of available text data across languages could mean that the model has seen fewer examples of conversational or socially indirect language in some corpora, even for otherwise high-resource languages. The way the task prompt is phrased and translated might also influence results if the term "euphemism" is interpreted differently across languages. Finally, languages with richer morphology, such as Ukrainian, can produce more fragmented representations under tokenization, which may weaken downstream reasoning performance. We believe the release of these datasets will enable future research to analyze these linguistic aspects of euphemism classification and the computational methods to classify, detect, and interpret them in context.

## 7. Limitations

As with most of our work containing LLMs, compute power has been one of our main obstacles, especially with regard to running the Qwen3 (Yang et al., 2025) models.

Additionally, the model would sometimes output explanations along with the prediction, thus we had to remind it several times in the system and user prompts to only output 0 or 1 as the overall output.

The thinking mode gave a more balanced output in terms of positive and negative classes, but because of compute power limitations, we were not able to run it five times for each language. We also ran a few trials with Qwen3-14B for English and Chinese, but opted not to use these larger models with other languages due to time constraints, as the Qwen3-14B took much longer to run than Qwen3-4B. Furthermore, Qwen3's English and Chinese reasoning abilities seemed to be significantly better than other languages, likely because English and Chinese were highly represented in Qwen3's training data. Qwen models were run on Google Colab.

GPT-5 (Singh et al., 2025), on the other hand, was run from the OpenAI servers, eliminating limitations of compute power. However, we were very likely throttled by rate limits when calling the API. Also, the model was unable to classify some examples due to guardrails against sensitive topics.

## 8. Ethical Concerns

The datasets include sentences taken from public, licensed corpora and examples written by the authors. They contain no personal information. All source materials were used under their original licenses, and the released data will use a permissive open license. Annotation was done by trained volunteers who gave consent to participate and could skip examples or stop at any time. Because some data deal with sensitive topics such as death, sexuality, or war, annotators were warned in advance and allowed to opt out of such cases.

## 9. Declaration on Use of Generative AI

While most prompts used in the process were written and translated by humans, we used generative AI to translate some prompts for zero-shot and few-shot into different languages, for example, Chinese.

## Acknowledgements

The authors of this paper would like to thank the Montclair NLP Lab and fellow students/alumni for building the foundations for the euphemism project. Special thanks to Patrick Lee.

Witold Kieraś and Natalia Zawadzka-Paluektau were supported by the IMPRESS-U DARE project financed by the Polish National Agency for Academic Exchange (NAWA) program (BPN/NSF/2023/1/00008).

Iryna Dilai was supported by the IMPRESS-U DARE project No.7132 funded by the National Academy of Sciences (NAS) and the Office of Naval Research (ONR) via the Science and Technology Center in Ukraine (STCU).

This material is based upon work supported by the National Science Foundation under Grant Nos. 226006 and 2428506. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## 10. Bibliographical References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, Bolette Pedersen, Carsten Levisen, and Daniel Hershcovich. 2025. Dying or departing? euphemism detection for death discourse in historical texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1353–1364, Abu Dhabi, UAE. Association for Computational Linguistics.

Keith Allan and Kate Burridge. 1991. *Euphemism & dysphemism: Language used as shield and weapon*. Oxford University Press.

Prashanth Bhat and Ofra Klein. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter. In *Twitter, the public sphere, and the chaos of online deliberation*. Springer.

Hasan Biyik, Patrick Lee, and Anna Feldman. 2024. Turkish delights: a dataset on Turkish euphemisms. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 71–80, Bangkok, Thailand and Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). *English World-Wide*, 36(1):1–28.

Iryna Dilai, Maksym Davydov, Anna Feldman, Oksana Oleksyn, Svitlana Kohut, and Olha Baranovska. 2025. Automatic Detection of Russia - Ukraine War Euphemisms. In *Proceedings of the 7th International Workshop on Modern Machine Learning Technologies (LeT-2025)*, volume 4004, page Paper 17. CEUR-WS.

Anna Dąbrowska. 1998. *Słownik eufemizmów polskich, czyli w rzeczy mocno, w sposobie łagodnie [The dictionary of Polish euphemisms]*. Wydawnictwo Naukowe PWN, Warsaw.

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, Online. Association for Computational Linguistics.

Todd Firsich and Anthony Rios. 2024. Can GPT4 detect euphemisms across multiple languages? In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 65–72, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Gilles Grolleau, Naoufel Mzoughi, Deborah Peterson, and Marjorie Tendero. 2022. Changing the world with words? euphemisms in climate change issues. *Ecological Economics*, 193:107307.

Nicholas Hankins. 2024. Optimizing multilingual euphemism detection using low-rank adaption within and across languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 8–14, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Sedrick Scott Keh. 2022. Exploring euphemism detection in few-shot and zero-shot settings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 167–172, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Witold Kieraś, Łukasz Kobyliński, Dorota Komosińska, Michał Rudolf, Maria Shvedova, and Aleksandra Zwierzchowska. 2025. PAWUK: Extensive annotated web corpus of ukrainian. In *Computational Science – ICCS 2025*, pages 94–105, Cham. Springer Nature Switzerland.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. MEDs for PETs: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Lee and Anna Feldman. 2024. Report on the multilingual euphemism detection task. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 110–114, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022a. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022b. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448, Toronto, Canada. Association for Computational Linguistics.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Xiang Li, Yucheng Zhou, Laiping Zhao, Jing Li, and Fangming Liu. 2025a. Impromptu cybercrime euphemism detection. pages 9112–9123.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025b. Language ranker: A metric for quantifying llm performance across high and low-resource languages. volume 39, pages 28186–28194.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

M. Marciniak, W. Kieraś, K. Bojałkowska, P. Borkowski, M. Borys, W. Eźlakowski, W. Guz, Ł. Kobyliński, D. Komosińska, K. Krasnowska-Kieraś, M. Łaziński, M. Miernecka, B. Nitoń, M. Ogrodniczuk, M. Rudolf, A. Tomaszewska, M. Woliński, J. Wołoszyn, B. Wójtowicz, A. Wróblewska, and N. Zawadzka-Paluektau. 2023. Korpus Współczesnego Języka Polskiego. Instytut Podstaw Informatyki PAN, Warszawa. https://kwjp.pl.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer

Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Real Academia Española. 2019. Corpus del español del siglo xxi.

Julia Sammartino, Libby Barak, Jing Peng, and Anna Feldman. 2025. When does language transfer help? sequential fine-tuning for cross-lingual euphemism detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1058–1065, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, and T. Nikolajenko. 2017-2025. GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource. https://uacorpus.org/.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea

Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong,

Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

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

Wiktionary contributors. 2024. двухсотый.

Wiktionary contributors. 2025. cargo-200.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *Proceedings - 2021 IEEE Symposium on Security and Privacy, SP 2021*, Proceedings - IEEE Symposium on Security and Privacy, pages 229–246, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2021 IEEE.; 42nd IEEE Symposium on Security and Privacy, SP 2021 ; Conference date: 24-05-2021 Through 27-05-2021.