

Does Pragmatic Transfer Help Euphemism Detection? Evidence from Target-Span Marking

Whitney Poh and Noriko Takahashi and Patrick Lee and Julia Sammartino
and Libby Barak and Jing Peng and Anna Feldman

Montclair State University

New Jersey, USA

{pohw1,takahashin1,leep,sammartinoj,barakl,pengj,feldmana}@montclair.edu

Abstract

Euphemism detection is often motivated as a pragmatic task: deciding whether a potentially euphemistic term (PET) is used euphemistically depends on context, communicative goals, politeness, topic sensitivity, and sentiment. A natural prediction is that auxiliary supervision from pragmatically related tasks should improve euphemism detection when target supervision is limited. We test this prediction for PET-style classification using XLM-R and mBERT in a feature-based transfer setting: auxiliary-tuned encoders are frozen and probed for euphemism detection. We compare sentence-level pragmatic proxy tasks with expression-level nonliteral tasks and with target inputs that either mark or omit the PET span. The results do not support a simple pragmatic-transfer account. Auxiliary transfer is unreliable, often reducing XLM-R performance and helping mBERT only inconsistently. The largest observed gains come from direct euphemism fine-tuning when the target PET is explicitly marked. The practical implication is to preserve target-span information and report source–target compatibility in transfer experiments.

1 Introduction

Euphemisms replace direct or potentially offensive, negative or socially taboo expressions with more socially acceptable alternatives (Lee et al., 2024; Poh et al., 2026; Firsich and Rios, 2024, etc.). Because euphemistic meaning depends on context and communicative goals, euphemisms are commonly linked to pragmatic dimensions such as politeness (Gavidia et al., 2022; Lee et al., 2024), topic sensitivity (Gavidia et al., 2022; Lee et al., 2023), and sentiment (Felt and Riloff, 2020). We ask whether encoders exposed to these pragmatic signals through auxiliary tasks become better eu-

phemism detectors.¹

PET-style disambiguation requires interpreting a specific expression in context. For example, *let go* is euphemistic in “the company let most of its employees go”, referring to being fired, but literal in “let go of my hand”. Likewise, *reorganizing* may refer to layoffs or tidying one’s room, and *slept with* may imply a sexual relationship or literal co-sleeping. Several examples are multiword expressions (MWE), but MWE status is not the hypothesis being tested. An example of a non-MWE euphemism found in the dataset by Lee et al. (2024) is “late”, meaning dead. The relevant property is that the label concerns a particular expression in context.

The pragmatic-transfer prediction is not supported in our experiments, which utilize XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019). Transfer from politeness, sentiment, topic sensitivity, and related proxy tasks is unreliable: it is usually negative for XLM-R and inconsistent for mBERT. This result is informative because these auxiliary tasks are not arbitrary baselines; they are plausible choices from a linguistic perspective. Euphemisms are pragmatic, but sentence-level pragmatic supervision does not necessarily teach a model how to decide whether a specific expression is euphemistic in a given context.

Rather than viewing the transfer result solely as evidence about the linguistic properties of euphemisms, we treat it as an opportunity to examine the relationship between task structure and transfer learning. Expression-level classification tasks, including many multiword-expression and figurative-language tasks, often use boundary marking to make a decision about a specific target span embedded within a larger context (Gavidia et al., 2022; Lee et al., 2024). In such settings, per-

¹Code and other material available at <https://github.com/NLP1abMSU/Cross-Task>

formance may depend not only on the linguistic phenomenon being modeled, but also on whether the task explicitly identifies the expression under consideration. We therefore compare inputs with and without boundary markers around the target expression and examine whether auxiliary tasks that require span-conditioned decisions transfer differently from tasks based on sentence-level or discourse-level properties.

Our contributions are threefold. First, we report a negative result: pragmatically related auxiliary tasks do not reliably transfer to PET-style euphemism detection. Second, we show that explicit target-span marking creates larger and more stable gains compared to unmarked versions of the same auxiliary task, and that factors such as target localization, label semantics, and input format may impact results significantly, shaping what information models learn and whether auxiliary supervision becomes useful. Direct euphemism fine-tuning in which both the training and testing set contains euphemism boundaries produced the highest and most consistent gains. Third, we derive a resource-design recommendation: when labels concern a specific expression, datasets should preserve target-span information because removing it changes both task difficulty and the interpretation of transfer results.

2 Background

Recent work has expanded euphemism detection across languages and domains using PET-style datasets (Gavidia et al., 2022; Lee et al., 2023, 2024; Biyik et al., 2024; Lee and Feldman, 2024). Methods include lexicon-based approaches (Felt and Riloff, 2020; Lee et al., 2022), transformer-based classifiers (Zhu et al., 2021; Wang et al., 2022), and prompting-based evaluations (Firsich and Rios, 2024). PET-style datasets emphasize contextual disambiguation: the same expression may be euphemistic in one context and literal in another.

Euphemisms also serve different communicative functions, including face-saving, softening and strategic obfuscation (Lee et al., 2024), such as in political discussions about climate change (Grolleau et al., 2022). Euphemistic and literal expressions can be semantically close while differing in pragmatic impact and social consequences (Grolleau et al., 2022). This motivates pragmatic auxiliary tasks, but it also raises a methodological question: does broad pragmatic relatedness transfer, or

does the auxiliary task need to match the target task’s decision unit?

Transfer learning can help low-resource target tasks, but it can also hurt through negative transfer (Wang et al., 2019; Chen et al., 2024). Prior work shows that transferability depends on task compatibility, supervision format, and training order (Vu et al., 2020). Our experiments test a concrete negative-transfer hypothesis: auxiliary tasks that appear linguistically related to euphemism detection may still be unhelpful when their input structure and decision unit differ from the PET-style target task.

3 Data

Target task. We use the English subset of the PET dataset released by Lee et al. (2024); Lee and Feldman (2024), containing 3,098 instances: 1,841 euphemistic and 1,257 non-euphemistic. Each instance contains a PET whose usage must be classified. We create two target variants: EUPH-B, where boundary tokens surround the PET, and EUPH-NOB, where the same texts are used without boundary markers. The marker appears in all positive and negative examples, so it identifies the target span but does not encode the class label. For euphemisms, sampling is restricted to `sometimes_euph` expressions, i.e., expressions appearing with both euphemistic and non-euphemistic labels, to reduce PET-identity memorization.

Auxiliary tasks. We evaluate eleven auxiliary input conditions drawn from eight auxiliary families (Table 1). The auxiliary datasets are drawn from established resources: Stanford Politeness (Danescu-Niculescu-Mizil et al., 2013), Sensitivity following Lee et al. (2023) and Mejova et al. (2014) using the GloWbe corpus (Davies and Fuchs, 2015), RT Movie Sentiment (Pang and Lee, 2005), Books vs. Electronics reviews (Blitzer et al., 2007), Magpie (idioms) (Haagsma et al., 2020; Dankers et al., 2022), TroFi (metaphors) (Birke and Sarkar, 2007), and IDEM (Idiom Sentiment) (Prochnow et al., 2024)². The pragmatic proxies are politeness, sensitivity, and sentiment. Books vs. Electronics is a domain-control task. Magpie and TroFi are expression-level nonliteral-language tasks included to test decision-unit and span-format alignment. We create boundary and

²See Appendix A for more details on datasets and source URLs

Task	Unit	Role
PET	target expr.	target
Politeness	utterance	prag. proxy
Sensitivity	topic	topic proxy
Sentiment	review	sent. proxy
Books/Elect.	review	control
Magpie	target expr.	expr.-level
TroFi	target word	expr.-level
IDEM	idiom sent.	affective
Multi	mixed	broad exposure

Table 1: Auxiliary task families and their primary decision units.

no-boundary variants for Magpie and TroFi by inserting or removing the same marker used in the PET data. IDEM tests whether affective supervision over idiom-containing sentences helps without requiring literal/nonliteral disambiguation. Mixed-task samples (Multi = Movie Sentiment (Pang and Lee, 2005), Sensitivity (Lee et al., 2023), Magpie (boundaries) (Haagsma et al., 2020; Dankers et al., 2022) and Politeness (Danescu-Niculescu-Mizil et al., 2013)) test whether broader exposure improves transfer beyond the best single auxiliary source; because their labels retain source-specific meanings, they are treated as broad-exposure controls rather than semantically unified tasks.

Examples of utterances from auxiliary sets can be found in Table 12, examples of euphemistic statements with certain pragmatic traits can be found in Table 13.

To reduce trivial effects of dataset size, and to standardize class balance for each setup, we sample 950 instances per class for each task, yielding 1,900 training instances per condition. Instances longer than the maximum length are removed during tokenization. For Sensitivity, which is drawn from GloWbE (Davies and Fuchs, 2015), which the Euphemisms dataset also draws from, we remove overlapping instances to reduce cross-task contamination.

4 Experimental Setup

4.1 XLM-R/mBERT

Our transfer setting is feature-based. We first fine-tune an encoder on an auxiliary task. We then freeze the encoder, extract final-layer classification-token embeddings for the euphemism data, and train a logistic regression probe on the euphemism training split. This isolates representational compatibility rather than full end-to-end adaptation. Fine-tuning and inference were done using HuggingFace Transformers (Wolf et al., 2020).

We use XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019), following prior multilingual PET work. For each setup, we run 10 trials with stratified 70/15/15 train/dev/test splits. Trial indices are aligned across conditions: run i uses the same euphemism split for every baseline and transfer setup, enabling paired comparisons. Auxiliary datasets are split independently within each run. Performance is reported as the mean macro F1 across trials.

Inputs are capped at 512 tokens; longer examples are removed rather than truncated to avoid deleting the target expression. Boundary markers are implemented as the literal string [PET_BOUNDARY]; no new vocabulary item is added. Across fine-tuning runs, we use learning rate 1×10^{-5} , batch size 16, a maximum of 20 epochs, and early stopping with patience 5 on development macro F1.

We report three kinds of conditions. The pre-trained probe baseline uses the frozen pretrained encoder. Target-task fine-tuning trains the encoder on euphemism detection, either with boundaries or without boundaries. Auxiliary transfer trains the encoder on a non-euphemism auxiliary task before probing on euphemism detection. Target-task fine-tuning is therefore not an auxiliary-transfer condition; it is included to show how much can be gained from in-task supervision under each input format.

5 Results

Auxiliary transfer is unreliable. Table 2 reports mean macro F1 for direct euphemism fine-tuning and auxiliary-transfer conditions. For XLM-R on EUPH-B, most auxiliary tasks reduce performance. TroFi-B is the largest observed auxiliary gain, but the improvement over baseline is small (+0.023) and not statistically significant. On Euph-N, no XLM-R auxiliary task exceeds the pretrained baseline. mBERT shows more positive transfer cases, especially with Magpie-B, TroFi-B and Sensitivity, but gains are format-dependent and don’t approach target-task fine-tuning.

Target-span marking is the clearest positive contrast. The strongest observed gains come from direct euphemism fine-tuning with target-span-marked input. On EUPH-B, direct target fine-tuning with boundaries improves XLM-R from 0.655 to 0.755 and mBERT from 0.572 to 0.670. No auxiliary transfer condition approaches these gains.

Condition	EUPH-B		EUPH-NoB	
	XLm-R	mBERT	XLm-R	mBERT
Pretrained	0.655	0.572	0.645	0.580
<i>Target-task FT</i>				
B	0.755***	0.670**	0.653	0.609
NoB	0.659	0.621*	0.675*	0.617*
<i>Auxiliary transfer</i>				
Movie	0.633	0.555	0.604**	0.563
Books	0.624*	0.589	0.615*	0.581
Polite	0.630*	0.547	0.614***	0.543**
IDEM	0.624**	0.547	0.609**	0.549*
Magpie (N)	0.639	0.578	0.625*	0.558
Magpie (B)	0.668	0.635***	0.618*	0.579
Sensitive	0.580***	0.611**	0.606***	0.606*
TroFi (N)	0.629	0.592	0.612*	0.586
TroFi (B)	0.678	0.623***	0.612**	0.575
Multi	0.626*	0.589	0.603***	0.578

Table 2: Mean macro F1 over 10 trials. N = no boundary markers; B = boundary markers (classification tuned). Bolded are the highest scores in each section. Paired two-tailed t-test (h_0 : fine-tuned = pre), * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	X-B	m-B	X-nB	m-nB
Euph	***	***	*	-
Trofi	**	*	-	-
Magpie	-	**	-	-

Table 3: Paired two-tailed t-tests (h_0 : train-B = train-nB) between the boundaried version and non-boundaried version of each training set (rows) and testing conditions (columns). X = XLmR, m = mBERT. B = euphB testing, nB = euph-noB testing. - $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Highest score in each section bolded.

Boundary format matters, but not by itself.

Boundary markers are most useful when they are part of target-task fine-tuning and evaluation, and when the boundary settings of the training and testing set match. Comparing Magpie-B with Magpie-NoB on EUPH-B also suggests that boundary-marked auxiliary training is more useful when the target is boundary-marked. The same pattern holds for TroFi. We interpret this conservatively: transfer is associated with auxiliary tasks that combine contextual disambiguation with explicit target localization, not with figurative-language exposure alone.

Negative-transfer summary. Table 5 counts auxiliary conditions as positive, near-baseline, or negative using a ± 0.005 margin around the pretrained baseline. These counts are descriptive rather than significance tests (though paired T-test results are reported in tables 2 and 3). The pattern is clearest for XLm-R: auxiliary fine-tuning usu-

Condition	EUPH-B		EUPH-NoB	
	XLm-R	mBERT	XLm-R	mBERT
Pretrained	0.036	0.035	0.028	0.026
<i>Target-task FT</i>				
B	0.028	0.044	0.035	0.043
NoB	0.037	0.026	0.034	0.036
<i>Auxiliary transfer</i>				
Movie	0.040	0.033	0.043	0.031
Books	0.030	0.032	0.024	0.031
Polite	0.030	0.031	0.026	0.033
IDEM	0.040	0.032	0.036	0.018
Magpie (N)	0.030	0.033	0.020	0.031
Magpie (B)	0.021	0.040	0.022	0.036
Sensitive	0.027	0.025	0.016	0.029
TroFi (N)	0.032	0.027	0.037	0.026
TroFi (B)	0.040	0.029	0.038	0.037
Multi	0.027	0.025	0.035	0.022

Table 4: Standard deviation over 10 trials. N = no boundary markers; B = boundary markers (classification tuned). Bolded are the highest scores in each section.

Target	Model	Pos.	Near	Neg.
EUPH-B	XLm-R	2	0	8
EUPH-B	mBERT	7	0	3
EUPH-NoB	XLm-R	0	0	10
EUPH-NoB	mBERT	1	5	4

Table 5: Auxiliary-transfer outcomes relative to the pretrained probe. “Near” means within ± 0.005 macro F1 of baseline.

ally hurts, and removing target boundaries often resulted in significantly different scores that were lower than the pretrained value. mBERT is less brittle, but its improvements are not stable across target formats.

6 Discussion

The central assumption tested here is that pragmatic relatedness should be sufficient for useful transfer. Politeness, sentiment, figurative language and sensitivity are all reasonable auxiliary choices for euphemism detection, but they mostly do not improve performance and often reduce it. This matters because the choice is linguistically intuitive: euphemisms are pragmatic, so pragmatic supervision should help. Our results suggest that this assumption is unreliable in a PET-style setup.

One reason is that sentence-level pragmatic tasks and PET classification ask different questions. A euphemistic expression can change perceived politeness, sensitivity, or sentiment when compared with a more direct alternative. However, the PET data do not ask the model to compare a euphemistic phrasing with a direct rephrasing of the same meaning. Instead, they ask whether a given expression

is euphemistic in its current context. Global properties such as politeness or sentiment therefore need not systematically separate the classes.

The results are consistent with a structural-alignment explanation, but they do not isolate structure as the only causal factor. Magpie and TroFi differ from the pragmatic proxy tasks not only in boundary marking but also in literal/nonliteral label structure and data domain. We therefore interpret the pattern conservatively: auxiliary tasks that resemble PET-style euphemism detection in target localization and contextual disambiguation appear more promising than broad sentence-level pragmatic proxies in this setup.

More broadly, the results suggest that selecting auxiliary tasks based solely on linguistic intuition can be misleading. Euphemisms are widely discussed as pragmatic phenomena, making politeness, sentiment, and sensitivity appear to be natural transfer sources. However, transfer performance appears to depend not only on the linguistic phenomenon involved but also on whether source and target tasks require similar decisions about similar units of text.

In this work, we sought to understand whether pragmatic supervision helps language models recognize euphemistic use. The results suggest that task design does more than affect measured performance: it can shape what information a model is encouraged to represent and which transfer signals become useful. This matters when computational results are interpreted as evidence about pragmatic understanding. Human euphemism use often involves choosing or interpreting an expression relative to a more direct alternative and inferring its social or pragmatic effect in discourse. PET-style classification, by contrast, asks whether a particular expression is euphemistic in its current context. To compare model behavior with assumptions about human pragmatic perception, the task design should therefore align with the relevant natural-language use case: the decision unit, available span information, label semantics, and contrast set should match the inference being claimed. Preserving target-span information is one practical step, but the broader recommendation is to design transfer experiments around the intended pragmatic inference rather than choosing auxiliary data only by broad linguistic theme.

Limitations

Our experiments focus on English PETs and two multilingual encoders; transfer patterns may differ in other languages, with English-only encoders, or with instruction-tuned LLMs (see Appendix B). Boundary tokens are an artificial annotation mechanism, and we use only one marker string. Future work should test neutral marker pairs, character offsets, span pooling, and architectures explicitly conditioned on spans. We also do not evaluate sequential fine-tuning where the auxiliary-tuned model is further fine-tuned end-to-end on euphemism detection; such a setting could yield different behavior. To address points made in the Discussion Section, future work can also create paraphrasing for the PETs.

Finally, our design does not fully disentangle target localization from other forms of task similarity. Boundary-marked Magpie and TroFi differ from pragmatic proxy tasks in several ways, including literal/nonliteral labels, target-span structure, and data domain. We therefore treat structural alignment as a hypothesis supported by the observed pattern, not as an isolated causal mechanism.

7 AI Use

ChatGPT (OpenAI, 2026) was consulted for code and instruction tuning advice. All final decisions were made by a human.

Acknowledgments

Thank you to Anand Gopeekrishnan for providing advice and helping locate datasets.

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.

References

- Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. Association for Computational Linguistics.
- Qihong Chen, Jiawei Li, Hyunjae Suh, Lianghao Jiang, Zheng Zhou, Jingze Chen, Jiri Gesi, and Iftekhar Ahmed. 2024. Does the order of fine-tuning matter and why? *arXiv preprint arXiv:2410.02915*.
- 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](#). *Preprint*, arXiv:1911.02116.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259. Association for Computational Linguistics.
- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can transformer be too compositional? analysing idiom processing in neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.
- Mark Davies and Robert Fuchs. 2015. The glowbe corpus (global web-based english): A 1.9 billion word corpus of english from 20 countries. *English Corpus Linguistics: Crossing Paths*, pages 159–182.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- 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. 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. 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. European Language Resources Association.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Gilles Grolleau, Naoufel Mzoughi, Deborah Peterson, and Marjorie Tendero. 2022. Changing the world with words? euphemisms in climate change issues. *Ecological Economics*, 193:107307.
- Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287. European Language Resources Association.
- Daniel Han, Michael Han, and Unsloth team. 2026. [Unsloth](#).
- 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. 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, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. 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. 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. Association for Computational Linguistics.
- Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. [Controversy and sentiment in online news](#). *Preprint*, arXiv:1409.8152.

OpenAI. 2026. [Chatgpt](#).

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124.

Whitney Poh, Julia Sammartino, Jasper Andrew, Witold Kieraś, Natalia Zawadzka-Paluckta, Iryna Dilai, Libby Barak, Jing Peng, and Anna Feldman. 2026. [The multilingual euphemism benchmark: Datasets and baselines for pragmatic language understanding](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7307–7319, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Alexander Prochnow, Johannes E. Bendler, Caroline Lange, Foivos Ioannis Tzavellos, Bas Marco Göritzer, Marijn ten Thij, and Riza Batista-Navarro. 2024. IDEM: The Idioms with EMotions dataset for emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8569–8579. ELRA and ICCL.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordani, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. [Exploring and predicting transferability across nlp tasks](#). *Preprint*, arXiv:2005.00770.

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*, pages 79–83. Association for Computational Linguistics.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. [Characterizing and avoiding negative transfer](#). *Preprint*, arXiv:1811.09751.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

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. *CoRR*, abs/2103.16808.

A Dataset Labeling

A.1 Euphemisms

Euphemisms (Lee et al., 2024; Lee and Feldman, 2024) dataset can be found at <https://github.com/pl464/euph-detection-datasets>

A.2 Magpie

Magpie (Haagsma et al., 2020; Dankers et al., 2022) retrieved from <https://huggingface.co/datasets/gsarti/magpie>

A.3 Trofi

Trofi (Birke and Sarkar, 2007) retrieved from <https://huggingface.co/datasets/tasksource/TroFi>

A.4 IDEM

IDEM (Prochnow et al., 2024) dataset can be found at https://github.com/AlexanderProchnow/idem/blob/main/dataset/idem_train.csv

A.5 Sentiment

Sentiment (Pang and Lee, 2005) retrieved from <https://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz>

A.6 Sensitivity

Following Lee et al. (2023)’s methods, we use a binary sensitive vs. non-sensitive dataset labeled via semantic similarity to a seed lexicon of sensitive/taboo topics³(Mejova et al., 2014). **Leakage control:** because both Sensitivity and PET data are drawn from web corpora GloWbE (Davies and Fuchs, 2015), we remove any Sensitivity instances that overlap with PET instances to avoid cross-task contamination.

A.7 Politeness

The original labels from the Stanford Politeness Dataset (Danescu-Niculescu-Mizil et al., 2013) have been converted to binary labels, with negative labels being 0 and positive labels being 1. Retrieved from https://github.com/Genius1237/TyDiP/blob/main/data/all/en_train.csv

³<https://drive.google.com/file/d/19PyIVklLtUgfunSFxgVvMAY5z6gjaAwT/view>

A.8 Books vs Electronics

The dataset from which we drew our Books vs Electronics data, released by [Blitzer et al. \(2007\)](#), is a sentiment dataset, however, we took the books and electronics reviews and labeled them according to topic (book review or electronics review) instead of using the sentiment information. Retrieved from <https://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html>

B Exploratory Extension to Instruction-Tuned Models

B.1 Setup

In addition, we use LoRA to fine-tune Qwen3-4B ([Yang et al., 2025](#)) and Llama-3.1-8B-Instruct ([Grattafiori et al., 2024](#)) models. Parameters are as follows:

LoRA settings were:

- Targeted modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
- r: 16
- alpha: 16

Training settings were:

- Maximum training sequence length: 4096
- Learning rate: 1e-5
- Epochs: 2

Prompts for this could be found in Appendices B.2 and B.3. These models were fine-tuned using Unsloth ([Han et al., 2026](#)) and Google Colab’s G4 GPU⁴. When G4’s were not available, A100s were used.

Because instruction-tuning on auxiliary tests may lead to overfitting, we also ran an additional trial where euphemism training data was added to the instruction-tuning data. The euphemism training data mixed into the instruction-tuning data includes boundary and non-boundary data with both boundary and non-boundary prompts (see the Appendix for more information). This setting uses 70% task data, 10% euphemism with boundaries (boundary prompt), 10% euphemism with boundaries (no boundary prompt) and 10% euphemism without boundaries for training, with the normal

euphemism test data, and is referred to as Qwen-Mix or Llama-Mix in the results tables. The same example may or may not be present in the training data for each split with both boundaries and without, however, the training data and the test data are separate and do not overlap.

The instruction-tuned models do not use the same training/validation/testing splits as the classification-tuned transformer models. However, the splits for each dataset are the same for all instruction-tuned models.

B.2 System Prompt

You are a linguist who studies pragmatic language. You understand things like euphemisms, politeness, metaphors, idioms and topic sensitivity. You use that knowledge to complete the task required and output the appropriate response.

B.3 Content Prompt

- Euphemisms (no boundaries): Can you decide whether the text contains a euphemism or not? Output 'Output: 1' if it contains a euphemism, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Euphemisms (boundaries): Can you decide whether the text between [PET_BOUNDARY] markers is used euphemistically or not? Output 'Output: 1' if it's euphemistic, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Idioms (no boundaries): Can you decide whether the given text contains an idiom or not? Output 'Output: 1' if there is an idiom, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Idioms (boundaries): Can you decide whether the word inside the [PET_BOUNDARY] markers within the given text is used idiomatically or not? Output 'Output: 1' if idiomatic, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Sensitivity: Can you decide whether this text is about a sensitive topic or not? Output 'Output: 1' if sensitive, 'Output: 0' if not. Do not

⁴<https://colab.research.google.com>

include an intro or conclusion, simply output the number. Here is the text:

- Politeness: Can you decide whether the content of the given text is polite or not? Output 'Output: 1' if polite, 'Output: 0' if impolite. Do not include an intro or conclusion, simply output the number. Here is the text:
- Metaphors (boundaries): Can you decide whether the word inside the [PET_BOUNDARY] markers within the given text is used metaphorically or not? Output 'Output: 1' if metaphorical, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Metaphors (no boundaries): Can you decide whether the given text contains a metaphor or not? Output 'Output: 1' if there is a metaphor, 'Output: 0' if not. Do not include an intro or conclusion, simply output the number. Here is the text:
- Sentiment (movies): Can you decide whether this movie review is positive or negative in terms of sentiment? Output 'Output: 1' if positive, 'Output: 0' if negative. Do not include an intro or conclusion, simply output the number. Here is the text:
- Sentiment (idioms): Can you decide whether this text, which contains an idiom, is of positive or negative sentiment? Output 'Output: 1' if it's positive, 'Output: 0' if it's negative. Do not include an intro or conclusion, simply output the number. Here is the text:
- Books: Can you decide whether this text is a book review or electronics review? Output 'Output: 1' if it's a book review, 'Output: 0' if it's an electronics review. Do not include an intro or conclusion, simply output the number. Here is the text:

Our instruction-tuned models performed poorly at euphemism detection. Table 6 shows results for the instruction-tuned LLMs. Previous work (Poh et al., 2026) have demonstrated that smaller LLMs did not always have great understanding of euphemisms and our results were consistent with that. Models tend to overpredict one class when

Condition	EUPH-B				EUPH-NoB			
	L	LMix	Q	QMix	L	LMix	Q	QMix
Pretrained	0.408	-	0.54	-	0.543	-	0.55	-
<i>Target-task FT</i>								
B	0.601	-	0.605	-	0.450	-	0.414	-
B2	0.544	-	0.604	-	0.484	-	0.379	-
NoB	0.436	-	0.362	-	0.580	-	0.573	-
<i>Auxiliary transfer</i>								
Movie	0.433	0.501	0.512	0.557	0.587	0.570	0.559	0.511
Books	0.506	0.507	0.520	0.560	0.553	0.550	0.520	0.460
Polite	0.511	0.515	0.474	0.556	0.547	0.546	0.548	0.502
IDEM	0.514	0.514	0.542	0.568	0.598	0.579	0.562	0.510
Maggie (N)	0.555	0.540	0.460	0.540	0.579	0.554	0.570	0.517
Maggie (B)	0.572	0.550	0.551	0.577	0.564	0.552	0.483	0.469
Maggie (B2)	0.593	0.561	0.510	0.590	0.516	0.528	0.518	0.464
Sensitive	0.406	0.435	0.449	0.530	0.586	0.586	0.518	0.745
TroFi (N)	0.515	0.512	0.355	0.466	0.587	0.580	0.555	0.563
TroFi (B)	0.576	0.563	0.506	0.543	0.550	0.543	0.524	0.500
TroFi (B2)	0.528	0.551	0.518	0.531	0.554	0.531	0.505	0.459

Table 6: Mean macro F1 over 10 trials. N = no boundary markers; B = boundary markers; B2 = data has boundaries but prompt is for non-boundaried data. B2 refers to the trials with the boundaried data but non-boundaried prompts (instruction-tuned). Q = Qwen3-4B, L = Llama-3.1-8B-Instruct

prompted to identify euphemisms. This is documented in table 7. Only the target task instruction-tuned models with the correct boundary setting were consistently not skewed towards one class. Because of these highly imbalanced results, we do not discuss instruction tuned models in detail in other sections of this paper. However, note that changing the boundary labeling information could completely flip class decisions. Several tasks were also tested on their corresponding test splits to see whether the models only struggled with euphemisms or whether they struggled with other tasks as well, and these results are reported in table 9. The results show F1s generally above 0.7 with low output class imbalance, indicating that smaller LLMs may have specific problems with euphemism detection. The tasks with F1s lower than 0.7 were all idiom and metaphor tasks, indicating that figurative language may be challenging for language models.

C Additional Tables

Tables 7 and 9 show that there is high class imbalance in the outputs when a model is fine-tuned on one task and then tested on another. As demonstrated by table 10, using the same prompt in training and testing reduces class imbalance, though it does not increase F1. These results show that a model instruction-tuned on one task pretty much only learns that task, but does not retain that information and utilize it to aid it in completing a different task with a different prompt. Even in

	Q	Q-B	L	L-B
Pre	0.30	0.78	0.76	0.95
Trofi	0.72	0.99	0.43	0.84
Trofi-B	0.28	0.83	0.29	0.76
Trofi-B2	0.23	0.94	0.30	0.83
Magpie	0.38	0.90	0.37	0.79
Magpie-B	0.17	0.69	0.32	0.77
Magpie-B2	0.24	0.89	0.21	0.69
Euph	0.50	0.98	0.49	0.93
Euph-B	0.06	0.50	0.10	0.69
Euph-B2	0.04	0.55	0.15	0.80
Movie	0.34	0.83	0.61	0.93
Polite	0.31	0.88	0.29	0.85
Sens	0.24	0.92	0.63	0.95
Books	0.25	0.81	0.32	0.84
Idem	0.34	0.80	0.50	0.84

Table 7: Average percent of outputs predicted as positive class (test euphemisms). Q = Qwen, L = Llama, B = Boundaried test data. N = 10 trials.

the setups where the supplementary task is mixed with euphemism training data of different boundary label styles, the F1s do not increase above the models instruction-tuned only on euphemisms. However, the class imbalance can be reduced by task-mixing, especially for the euphemism testing task with boundaries, though a rather heavy level of skew still remains. In both cases, the model tends to overpredict 0 for euphemism detection without boundaries, and 1 for euphemism detection with boundaries.

Table 11 shows that this is not true for classification-tuned transformers models, as while some results may be mediocre, models that are fine-tuned on one dataset and tested on another are generally not skewed towards one class. This is because we fitted a separate logistic regression model using the encoder to encode the testing task (euphemisms) instead of directly using the classifier from the training task. However, we provide the table anyways to show that the testing data is split randomly from the training data and therefore isn't skewed towards one class.

D Examples of Tasks and Data

Table 12 shows examples of the auxiliary tasks. Table 13 show examples of Examples of politeness, idiomaticity and sentiment in the euphemism data.

	QM	QM-B	LM	LM-B
Trofi	0.39	0.89	0.40	0.86
Trofi-B	0.19	0.75	0.29	0.79
Trofi-B2	0.13	0.78	0.27	0.76
Magpie	0.23	0.74	0.30	0.81
Magpie-B	0.14	0.60	0.31	0.80
Magpie-B2	0.13	0.62	0.23	0.75
Movie	0.24	0.69	0.36	0.86
Polite	0.20	0.69	0.27	0.84
Sens	0.14	0.78	0.47	0.92
Books	0.12	0.69	0.30	0.86
Idem	0.20	0.68	0.37	0.83

Table 8: Average percent of outputs predicted as positive class (test euphemisms). QM = QwenMix, L = LlamaMix, B = Boundaried test data. N = 10 trials.

	Q-F1	Q-Imba	L-F1	L-Imba
Trofi	0.58	0.49	0.59	0.50
Trofi-B	0.65	0.52	0.66	0.44
Trofi-B2	0.67	0.54	0.59	0.69
Magpie	0.73	0.52	0.75	0.55
Magpie-B	0.71	0.47	0.61	0.80
Magpie-B2	0.75	0.55	0.70	0.69
Movie	0.92	0.49	0.93	0.47
Polite	0.71	0.43	0.77	0.55
Sens	0.71	0.50	0.74	0.52
Books	0.98	0.50	0.98	0.50
Idem	0.95	0.50	0.95	0.50

Table 9: Macro F1 and average percent of outputs predicted as positive class (test orig. task). Q = Qwen, L = Llama, B = Boundaried test data. N = 10 trials.

	F1	Imba
TrofiB	0.558	0.62

Table 10: A Qwen3-4B model instruction-tuned on the TrofiB data, with the EuphB prompt. Results show Macro F1 and average percent of outputs predicted as positive class over 10 trials.

	mBert	mBERT-B	XLMR	XLMR-B
Trofi	0.50	0.50	0.52	0.51
Trofi-B	0.51	0.50	0.51	0.50
Magpie	0.50	0.49	0.51	0.51
Magpie-B	0.49	0.51	0.50	0.50
Movie	0.50	0.51	0.50	0.50
Polite	0.51	0.51	0.50	0.51
Sens	0.51	0.51	0.50	0.50
Books	0.50	0.49	0.52	0.51
Idem	0.52	0.51	0.52	0.51

Table 11: Average percent of outputs predicted as positive class (test euphemisms) for classification-tuned transformer models. N = 10 trials.

D.1 Results on the no-boundary target (Euph-NoB)

We provide supplementary tables with the results on the no-boundary target. Table 14 shows the results using XLM-R and Table 15 shows the results using mBERT.

D.2 Per-category results (Euph-B)

The paper discusses the overall results for the data. The euphemism data set includes category annotation for each euphemistic expression including categories such as politics and death. Tables 16 and 17 shows the breakdown of the results for each semantic category for XLM-R and mBERT respectively.

³<https://huggingface.co/datasets/gsarti/magpie>

⁴<https://huggingface.co/datasets/tasksource/Trofi>

⁵<https://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz>

⁶https://github.com/Genius1237/TyDiP/blob/main/data/all/en_train.csv

⁷Indices from https://github.com/pl464/euph-detection-datasets/blob/main/FigLang_2024/EN_PET_dataset_2024.csv

Dataset	Source and Source URL	Text	Label
Magpie	Haagsma et al. (2020); Dankers et al. (2022) ³	“[PET_BOUNDARY] Spotted on [PET_BOUNDARY] a fan in a department store.”	0
Magpie	Haagsma et al. (2020) ³	“Everest appeared [PET_BOUNDARY] out of the question, [PET_BOUNDARY] such was his temperamental and erratic nature.”	1
Trofi	Birke and Sarkar (2007) ⁴	“And most of the jobs will go to candidates sought out by search firms , not those knocking on headhunter doors .”	0
Trofi	Birke and Sarkar (2007) ⁴	“That planted the seeds of discontent between Mr. Kirk and the Jackson forces .”	1
Sentiment	Pang and Lee (2005) ⁵	“obviously , a lot of people wasted a lot of their time (including mine) on something very inconsequential .”	0
Sentiment	Pang and Lee (2005) ⁵	“more romantic , more emotional and ultimately more satisfying than the teary-eyed original .”	1
Sensitivity	Lee et al. (2023); Davies and Fuchs (2015); Mejova et al. (2014)	“It was a nice try , my friend . Bravo . Well played . However , you were pwned before you posted .”	0
Sensitivity	Lee et al. (2023); Davies and Fuchs (2015); Mejova et al. (2014)	“So , when it lacks a budget for direction , Congress still has the power to spend money and keep government operating , but the process can get mucked up .”	1
Politeness	Danescu-Niculescu-Mizil et al. (2013) ⁶	“Don’t talk about me, in my virtual presence, and tell someone else to ignore me. Do you understand why this behavior is inappropriate?”	0
Politeness	Danescu-Niculescu-Mizil et al. (2013) ⁶	“Good job working on the list, thanks. Could you make sure you move them from "To do" to "Done" when you finish?”	1

Table 12: Text examples from the auxiliary datasets. Labels represent the labels we use for our fine-tuning task, not necessarily the provided labels in the original creators’ dataset.

Type	Text	Orig. Index ⁷
Polite	“Dear Michael, Thank you for your pipcture and for sharing wit with the world. My own son is 11 now and I recall seeing your lovely picture when I was [PET_BOUNDARY]expecting[PET_BOUNDARY] him and I felt a strong connection to it. Thank you also for your courageous and continuous fight for the truth.”	657
Impolite	“I don’t condone violence but had she butted out like Adriana asked her to and not followed Adriana looking for a fight, she might not have gotten hit at all!! I really think she has a [PET_BOUNDARY]drinking problem[PET_BOUNDARY]. She needs help!!!”	550
Positive Sentiment	“In spite of his [PET_BOUNDARY]advanced age[PET_BOUNDARY], Rollins remains one of jazz’s most talented improvisers.”	73
Negative Sentiment	“This speaker should be disinvited from further commencements and if he is employed by ANY scholl , any level be [PET_BOUNDARY]dismissed[PET_BOUNDARY]”	518
Idiomatic	“There are plenty of people who say the show is kind of [PET_BOUNDARY]over the hill[PET_BOUNDARY]”	1197
Not Idiomatic	“Early Sunday, the Heath Ministry confirmed that the novel coronavirus claimed 19 [PET_BOUNDARY]fatalities[PET_BOUNDARY]”	723

Table 13: Examples from [Lee and Feldman \(2024\)](#) that contain certain pragmatic traits. Long examples may be truncated. Pragmatic categorizations were determined by the authors of this paper.

XLM-R (Euph-NoB)	Phys/Ment	Misc	Body	Subs	Pol	Emp/Fin	Sex	Death
Baseline	0.674	0.513	0.693	0.742	0.506	0.556	0.675	0.656
Euph fine-tune (B→NoB)	0.656	0.543	0.693	0.675	0.519	0.630	0.646	0.674
Euph fine-tune (NoB→NoB)	0.689	0.536	0.709	0.708	0.534	0.633	0.688	0.702
TroFi (NoB)	0.626	0.552	0.660	0.647	0.577	0.531	0.615	0.634
TroFi (B)	0.614	0.540	0.695	0.622	0.569	0.575	0.595	0.625
Movie Sentiment	0.600	0.465	0.638	0.651	0.599	0.580	0.623	0.603
Books vs Electronics	0.609	0.511	0.674	0.684	0.561	0.533	0.635	0.655
Magpie (NoB)	0.643	0.536	0.621	0.697	0.517	0.582	0.625	0.625
Magpie (B)	0.632	0.507	0.710	0.616	0.510	0.583	0.617	0.606
Sensitivity	0.590	0.545	0.694	0.659	0.540	0.576	0.595	0.589
Politeness	0.629	0.521	0.654	0.630	0.565	0.570	0.575	0.635
Idiom Sentiment	0.627	0.526	0.649	0.669	0.492	0.534	0.625	0.601
Multi	0.610	0.595	0.590	0.636	0.521	0.558	0.573	0.595

Table 14: Per-category macro F1 for XLM-R when the euphemism target has no boundary markers (Euph-NoB).

mBERT (Euph-NoB)	Phys/Ment	Misc	Body	Subs	Pol	Emp/Fin	Sex	Death
Baseline	0.590	0.505	0.662	0.588	0.438	0.550	0.613	0.555
Euph fine-tune (B→NoB)	0.599	0.502	0.618	0.611	0.430	0.597	0.650	0.625
Euph fine-tune (NoB→NoB)	0.625	0.504	0.627	0.637	0.421	0.592	0.620	0.625
TroFi (NoB)	0.609	0.496	0.623	0.614	0.464	0.577	0.626	0.545
TroFi (B)	0.570	0.489	0.642	0.604	0.397	0.546	0.589	0.587
Movie Sentiment	0.580	0.499	0.611	0.587	0.460	0.545	0.555	0.531
Books vs Electronics	0.589	0.528	0.644	0.581	0.473	0.534	0.581	0.589
Magpie (NoB)	0.573	0.495	0.579	0.567	0.373	0.550	0.584	0.539
Magpie (B)	0.602	0.527	0.576	0.648	0.437	0.527	0.582	0.567
Sensitivity	0.603	0.503	0.669	0.643	0.504	0.565	0.607	0.626
Politeness	0.550	0.479	0.586	0.559	0.526	0.505	0.567	0.521
Idiom Sentiment	0.556	0.519	0.602	0.512	0.424	0.546	0.550	0.521
Multi	0.573	0.583	0.565	0.611	0.408	0.530	0.586	0.581

Table 15: Per-category macro F1 for mBERT when the euphemism target has no boundary markers (Euph-NoB).

XLM-R (Euph-B)	Phys/Ment	Misc	Body	Subs	Pol	Emp/Fin	Sex	Death
Baseline	0.671	0.629	0.708	0.696	0.474	0.584	0.661	0.654
Euph fine-tune (B→B)	0.741	0.717	0.704	0.833	0.663	0.748	0.722	0.757
Euph fine-tune (NoB→B)	0.683	0.535	0.646	0.675	0.568	0.633	0.672	0.660
TroFi (NoB)	0.621	0.584	0.670	0.684	0.588	0.571	0.632	0.648
TroFi (B)	0.690	0.559	0.750	0.675	0.549	0.636	0.641	0.717
Movie Sentiment	0.632	0.537	0.622	0.649	0.665	0.616	0.655	0.619
Books vs Electronics	0.629	0.566	0.655	0.694	0.556	0.545	0.636	0.634
Magpie (NoB)	0.672	0.526	0.653	0.637	0.546	0.600	0.666	0.633
Magpie (B)	0.665	0.571	0.722	0.759	0.490	0.630	0.653	0.678
Sensitivity	0.572	0.554	0.622	0.608	0.479	0.555	0.573	0.580
Politeness	0.638	0.569	0.675	0.646	0.501	0.601	0.601	0.641
Idiom Sentiment	0.651	0.559	0.658	0.670	0.500	0.562	0.588	0.620
Multi	0.623	0.581	0.632	0.709	0.512	0.577	0.652	0.625

Table 16: Per-category macro F1 for XLM-R when the euphemism target includes boundary markers (Euph-B).

mBERT (Euph-B)	Phys/Ment	Misc	Body func./part	Subs	Politics	Empl/Fin	Sex	Death
Baseline	0.571	0.507	0.631	0.589	0.542	0.535	0.592	0.559
Euph fine-tune (B→B)	0.673	0.593	0.641	0.691	0.533	0.689	0.652	0.654
Euph fine-tune (NoB→B)	0.630	0.540	0.669	0.665	0.439	0.596	0.587	0.619
TroFi (NoB)	0.598	0.480	0.675	0.597	0.493	0.571	0.659	0.566
TroFi (B)	0.633	0.591	0.635	0.658	0.509	0.579	0.649	0.618
Movie Sentiment	0.585	0.480	0.598	0.505	0.447	0.531	0.565	0.539
Books vs Electronics	0.572	0.501	0.613	0.627	0.480	0.548	0.611	0.620
Magpie (NoB)	0.582	0.510	0.590	0.616	0.429	0.552	0.560	0.597
Magpie (B)	0.645	0.558	0.699	0.713	0.480	0.599	0.588	0.646
Sensitivity	0.609	0.487	0.682	0.657	0.446	0.571	0.639	0.626
Politeness	0.542	0.466	0.639	0.527	0.547	0.487	0.617	0.536
Idiom Sentiment	0.551	0.502	0.456	0.542	0.472	0.524	0.542	0.544
Multi	0.548	0.546	0.607	0.635	0.503	0.546	0.600	0.628

Table 17: Per-category macro F1 for mBERT when the euphemism target includes boundary markers (Euph-B).