To Labor is Not to Suffer: Exploration of Polarity Association Bias in LLMs for Sentiment Analysis

Jiyu Chen[1], **Sarvnaz Karimi**[1], **Diego Mollá**[2,1], **Cécile Paris**[1,2]
[1]CSIRO Data61, Australia [2]Macquarie University, Australia
firstname.lastname@csiro.au diego.molla-aliod@mq.edu.au

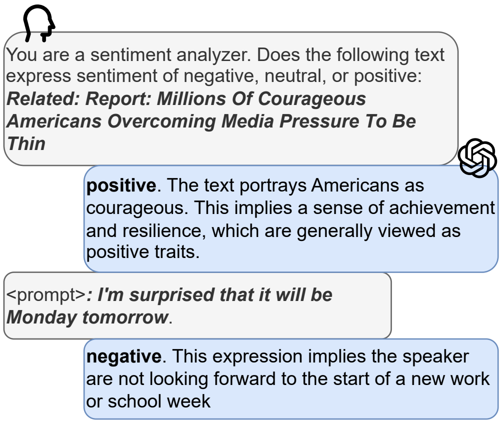
Abstract

Large language models (LLMs) are widely used for modeling sentiment trends on social media text. However, our study discovered that LLMs have polarity association bias—a stereotypical association between a word or a situation and an emotion. This association can cause systematic misclassifications of *neutral* statements into either positive or negative, exacerbating distorted estimation of sentiment trends. Such systematic errors are likely induced by associating certain lexical word mentions with positive or negative polarity, ignoring the actual neutral tone being expressed. We estimate there is a moderate to strong degree of the polarity association bias across five widely used LLMs. We uncover a broad spectrum of linguistic and psychological lexical indicators associated with this bias, emphasizing that LLMs are not reliable as *out-of-the-box* deployment for sentiment estimation.

1 Introduction

Sentiment analysis is commonly framed as a text classification task, where language models classify a text message as having either *positive*, *negative*, or *neutral* sentiment (Rosenthal et al., 2017), with extension to Likert scale ranging from extremely positive to negative (Socher et al., 2013), or regarding certain aspects of products (Fang and Zhan, 2015). Sentiment analysis is applied in diverse fields, like supporting stock market prediction (Pagolu et al., 2016), analyzing product reviews (Fang and Zhan, 2015), or supporting mental health inquiry (Babu and Kanaga, 2022).

Previous studies (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023) demonstrate that, when provided with unambiguous instructions, LLMs can classify positive and negative sentiments, even in zero-shot learning settings. However, the evaluations also reveal that LLMs consistently struggle to identify neutral statements correctly (also see our evaluation in Table 1).



You are a sentiment analyzer. Does the following text express sentiment of negative, neutral, or positive: ***Related: Report: Millions Of Courageous Americans Overcoming Media Pressure To Be Thin***

positive. The text portrays Americans as courageous. This implies a sense of achievement and resilience, which are generally viewed as positive traits.

<prompt>***: I'm surprised that it will be Monday tomorrow***.

negative. This expression implies the speaker are not looking forward to the start of a new work or school week

Figure 1: Examples of polarity association biases in ChatGPT-3.5 (OpenAI, 2024) when instructed to perform sentiment classification: The first instance reflects a bias toward a positive sentiment due to the use of "*courageous*", a word related to positive affect. The second instance shows an association bias between "*Monday*" and "*work or school week*", and between work/school and a negative sentiment, disregarding the actual neutrality in the text.

We hypothesize that this problem stems from a polarity association bias learned during pre-training. LLMs can stereotypically associate certain linguistic or psychological word categories with positive or negative sentiment in a skewed manner, disregarding the actual neutrality conveyed in the given text. For example in Figure 1, we see at play a stereotype of associating "*courageous*" with a `positive` sentiment, and two other stereotypes associating (1) "*monday*" with "*school or work*", and (2) "*school or work*" with a `negative` sentiment.

We propose a simple, yet effective, approach to estimate the polarity association biases using two benchmark datasets, SemEval-2017 (Rosenthal et al., 2017) and GoEmotions (Demszky et al., 2020). We employ that approach to examine these biases in five representative LLMs. Our study re-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 70–78
December 20-24, 2025 ©2025 Association for Computational Linguistics

veals that LLMs often misclassify text with neutral sentiment as either positive or negative, potentially induced by their underlying association biases. Additionally, we observe that the presence of words in certain linguistic or psychological categories significantly correlates with the tendency of committing such errors, presenting consistent patterns as the aforementioned "stereotype". We thus caution against the application of the current LLMs for large-scale sentiment classification, as their inherent polarity association bias can exaggerate the estimation of sentiment trends (e.g., on social media) towards positive or negative.

Model	**Label Classes**		
	positive	negative	neutral
RoBERTa	0.83	0.84	**0.73**
gpt-4	0.78	0.79	**0.54**
Llama3	0.65	0.64	**0.59**

Table 1: F_1 of sentiment classification by label class on 3000 randomly selected SemEval-2017 samples. The scores of neutral instances are **bolded**. See experiment details in Appendix A.

2 Methodology

We estimate the severity of the **polarity association bias** through the measurement of false negative (FN) rate in sentiment classifiers on *neutral* instances. We acknowledge that the observed FN could also be caused by other factors, such as underfitting, data annotation noises, and prompt formulation bias. Based on existing reports demonstrating the robust generalization capabilities of LLMs in predicting positive and negative sentiment (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023), we set aside underfitting as a primary cause in our study. To account for annotation noises, we report the FN rates of two BERT-based models fine-tuned on each of two benchmark datasets we use, yielding FN rates of 0.27 (obtained through applying RoBERTa (Camacho-Collados et al., 2022)) for SemEval-2017 (Rosenthal et al., 2017) and 0.21 (reported by dataset author) for GoEmotions (Demszky et al., 2020) as **baselines**. To account for prompt formulation bias, we construct a set of neutral and unambiguous prompts by following the methodology of (Zhang et al., 2024) (see details in Appendix B). We compute the standard deviation of FN rates across these prompts to estimate the potential bias introduced by prompt formulation. We focus on measuring the FN rate on instruction-tuned foundational LLMs with zero-shot setting, rather than prompt-tuned or supervised fine-tuned LLMs. We do so because FN in models continually tuned with sentiment classification data may be more attributable to extrinsic hallucination rather than being a reflection of intrinsic bias in the pre-trained foundational LLMs (Ladhak et al., 2023).

Data: We collect $7,323$ neutral text instances from the SemEval-2017 Task 4 dataset (Rosenthal et al., 2017) and $12,748$ instances from the GoEmotions dataset (Demszky et al., 2020), all manually annotated. Although the GoEmotions focus on fine-grained emotional states like joy, sadness, and anger—differing from the polarity-based annotations of positive, negative, and neutral in SemEval-2017—the neutral emotional state aligns with the neutral sentiment, according to the annotation scheme (Demszky et al., 2020).

Models and Configurations: In addition to the two baselines established using BERT-based models fine-tuned respectively on each of the two datasets, we also apply a RoBERTa model fine-tuned on SemEval-2017 (Camacho-Collados et al., 2022) to the GoEmotions dataset. This setup allows us to examine whether polarity association bias emerges when a BERT-based model is transferred to a novel target dataset. We explore five instruction-tuned LLMs: gemma2-2b-it (Rivière et al., 2024), Llama3.1-8b-Instruct (Dubey et al., 2024), deepseek-llm-7b-chat (DeepSeek-AI, 2024), gpt3.5-turbo (OpenAI, 2024), and gpt4-turbo (OpenAI, 2023) with zero-shot prompt engineering (see Appendix B). For each instruction-tuned LLM, we experiment with six different prompts and temperature settings (Appendix C) on the same instance and calculate the mean and standard deviation of the FN rates for quantifying potential bias in the prompt design. We find that varying the prompt formulations and temperature settings had minimal influence on the predictions across all these LLMs (Figure 2). This indicates that the sentiment classification task exhibits low intrinsic randomness, and that the LLM is highly confident in its predictions. This strongly suggests that the false negatives are more likely to be induced by the learned biases rather than the stochastic factors of language models.

Identification of Lexical Mentions Regarding Various Linguistic or Psychological Word Categories: We utilize Linguistic Inquiry and Word Count (LIWC2015) (Pennebaker et al., 2015) to identify lexical word mentions covering a broad spectrum of linguistic and psychological categories (see Appendix D, the first column of Table 4), e.g., "*anger*", and "*family*". We calculate (with LIWC2015) the psychometric scores to characterize each text instance with respect to each of these LIWC categories. Let N denote the total word count in an input text instance and C_d the total number of words belonging to d, a specific LIWC2015 category, for the same instance. The psychometric score is calculated as $s_d = \frac{C_d}{N} \times 100, s_d \in [0, 100]$. In nature, the distribution of s_d is skewed, with most instances having $s_d \in [0, 30]$ (see Appendix D, Figure 4).

Correlation Between Categorical Word Mention and False Negative Likelihood: For a given word category d, we first sort all neutral instances in ascending order based on their s_d. We then partition the sorted instances into smaller subgroups using logarithmic binning (see Appendix E for details). Within each subgroup, we compute the FN rate using the T_1 prompt as the standard deviation of prompt variations is minimum (Figure 2 and see formulation in Table 2 Appendix B), denoted as ε_d, further categorized as:

- **Directed FN rate**: The proportion of neutral instances misclassified as `positive` sentiment (ε_d^+) or as `negative` sentiment (ε_d^-) respectively within the subgroup.
- **Undirected FN rate** ($\varepsilon_d^\pm$): The proportion of misclassified neutral instances in the subgroup.

For each subgroup, ε_d is then assigned to all instances within that subgroup, representing the likelihood L_{ε_d} that an instance may be misclassified towards either `positive` sentiment $L_{\varepsilon_d}^+$ or `negative` sentiment $L_{\varepsilon_d}^-$.

Finally, we calculate the Pearson correlation coefficient[1] r_d and the underlying p-value (p) between s_d (categorical word mention score) and $L_{\varepsilon_d}^\pm$ (FN likelihood).

3 Results and Discussion

Severity of the Polarity Association Biases

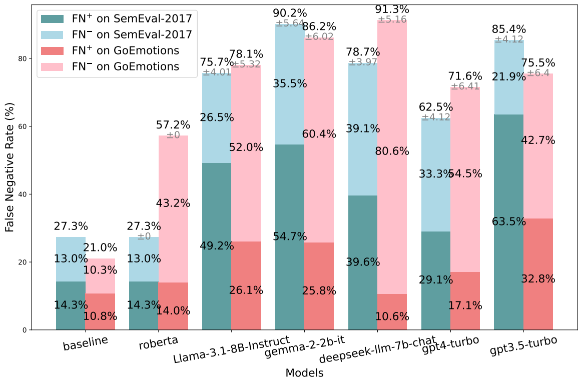


Figure 2: The LLMs' FN on the neutral statements in SemEval-2017 and GoEmotions. The FN$^+$ denotes the proportion of mis-classifications as `positive` and FN$^-$ the proportion of mis-classifications as `negative`. The FN is the sum of FN$^\pm$ on each stacked bar.

All five LLMs exhibited FN greater than baselines, on both SemEval-2017 and GoEmotions, suggesting bias (Figure 2). Notably, `gemma` showed the highest FN (90.2%) on SemEval-2017, while `deepseek` had highest FN (91.2%) on GoEmotions. RoBERTa exhibits increased FN on GoEmotions, suggesting that, while fine-tuning can effectively reduce FN in identifying neutral instances when dealing with similar text (i.e., public-facing and intended for a broad audience) from the same social media platform, it may not effectively mitigate the intrinsic biases rooted in the pre-training stage of language models. Such biases could persist and re-emerge when applied to different texts (e.g., different social media platforms), exhibiting different language styles (i.e., inward-focused, self-reflective, and emotionally expressive, as in GoEmotions).

We also observed that all five LLMs tended to classify neutral instances in the SemEval-2017 dataset as `positive` but as `negative` in the GoEmotions dataset. We assume that such inverted pattern is because GoEmotions has greater proportion of affect-related, present-focused, money-related and less proportion of leisure-related, future-focused word use (see the average proportion of word use between two datasets in Table 4, Appendix D), biasing the model to assign `negative` sentiment to GoEmotions instances and `positive` sentiment to SemEval-2017 instances.

[1] `https://docs.scipy.org/doc/scipy-1.15.0/reference/generated/scipy.stats.pearsonr.html`

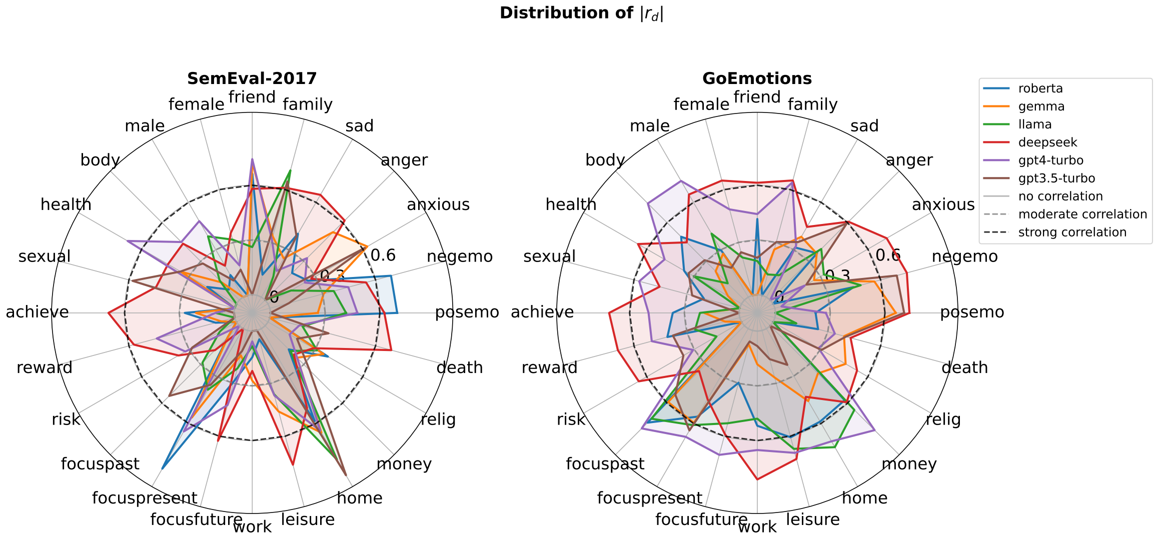


Figure 3: The distribution of $|r_d|$ on the SemEval-2017 and GoEmotions datasets. All negative r_d are converted into positive value for representing the intensity of the correlation. r_d are converted into 0 when there is no significant correlation ($p \geq 0.05$). $|r_d| \in [0.3, 0.6]$ denotes a moderate correlation, and $|r_d| > 0.6$ denotes a strong correlation.

Correlation Measure Between Categorical Word Mentions and False Negative Likelihood

We observed that both RoBERTa and the five LLMs exhibit a moderate to strong r_d between s_d (categorical word mention score) and the L_{ε_d} (FN likelihood) when classifying neutral text (Figure 3), suggesting that all models exhibit polarity association biases. Notably, RoBERTa, which was fine-tuned on the SemEval-2017 dataset, showed a small r_d, but this value increased significantly when tested on the unseen GoEmotions. This suggests that, while fine-tuning language models can reduce r_d when classifying neutral instances, it may not effectively mitigate the bias, since the correlation persists when applied to unseen data.

The measured r_d on GoEmotions dataset among the five LLMs tends to be larger than on the SemEval-2017 dataset, reflected by the larger size of shadowed area on Figure 3. We hypothesize that, as text instances on GoEmotions are more subjective, self-reflective, and emotionally nuanced, they are more likely to trigger a polarity association bias.

The deepseek model appears to be the most affected by polarity association bias, with the highest r_d across almost all the experimented word categories. Additionally, we observed no significant reduction in r_d from gpt3.5 to its more advanced version, gpt4, suggesting that the association bias can not be easily mitigated. llama and gemma also resulted with moderate to strong r_d on both datasets, with smaller model (2B-parameter) gemma having stronger r_d. llama resulted with r_d as severe as the larger proprietary gpt3.5. This indicates that the small parameter sizes are unlikely to be the cause of polarity association bias.

A complete list of r_d values regarding each category of words and a more comprehensive analysis are provided in Appendix F. The observed moderate to strong correlations between FN rates and specific lexical mentions indicate that these FN errors on neutral instances are more likely attributable to intrinsic system bias of the LLMs, rather than to factors such as model underfitting, annotation noise, or prompt formulation.

4 Conclusion

Sentiment analysis has significant implications for many real-world applications. While LLMs can be instructed to conduct sentiment classification in a zero-shot setting, we argue that they are not yet fully reliable for large-scale sentiment analysis, especially when the data contains neutral statements.

Our study shows that LLMs have a high likelihood of misclassifying neutral text with either positive or negative sentiment. These mis-classification errors exhibit a moderate to strong correlation with the presence of specific categories of lexical words in the text. Our study suggests that LLMs may have developed a moderate to strong degree of polarity association bias, which can distort the estimation of sentiment trends.

Limitations

This study examines the polarity association bias in LLMs by only utilizing the datasets generated by English speakers, while cultural background can play a significant role in influencing the estimation of bias in LLMs (Imran et al., 2020). In real-world situations, multiple factors can contribute to polarity association bias, often making it challenging to isolate their individual effects. Our study focuses on lexical factors and assumes that each word category independently influences polarity association bias. Our experimental results on correlation do not establish a causal relationship between lexical word mentions and the likelihood of bias occurring.

We have not examined all the existing LLMs at the time of the submission. While it is not likely to change the findings, the experimental results represent a sub-set of all available LLMs.

Ethical Concerns

We relied on the dataset providers to remove any material from the dataset that may reveal anyone's identity in their posts used in this study. The proprietary LLMs (`gpt3.5-turbo` and `gpt4-turbo`) are hosted within our organization's server, ensuring the privacy of processing user-sensitive data.

Our project has been granted ethics approval from our organization, CSIRO, Australia (approval: 217/23)

Licenses of Artifacts

All scientific artifacts cited or utilized in this paper were employed in accordance with their respective license of use.

References

Nirmal Varghese Babu and E Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN Computer Science*, 3(1):74.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez-Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2:1–14.

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related Tweets. *IEEE Access*, 8:181074–181090.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

OpenAI. 2023. GPT4. Version: firstcontact-gpt4-turbo 2023-03-15-preview.

OpenAI. 2024. ChatGPT 3.5. Version: chatgpt-35-turbo 2024-02-15-preview.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350. IEEE.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *University of Texas at Austin*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

A Sentiment Classification on the SemEval-2017 Subset

We randomly selected $3,000$ manually annotated instances from the SemEval-2017 and applied the fine-tuned RoBERTa (Camacho-Collados et al., 2022), `gpt4` (OpenAI, 2023), and `llama3` (Dubey et al., 2024) to measure the F_1 score in the classification of `positive`, `negative`, and `neutral` instances (1000 instances within each class). The results suggest that LLMs are effective in identifying positive and negative instances but not for neutral instances (Table 1). The results on the SemEval-2017 dataset closely align with existing studies on other sentiment analysis datasets (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023).

B Prompt Instruction for Sentiment Classification

Language models exhibit a tendency to favor affirmative responses when prompted with yes/no questions. To reduce this form of prompt-induced bias, we follow the prompting strategies proposed by Zhang et al. (2024), which advocate for the explicit inclusion of the task name, neutral and unambiguous instructions, and a clearly defined output format. Accordingly, we avoid formulations such as "*Does the following text express neutral sentiment?*" which may elicit biased responses.

To further address the potential bias toward the sentiment label mentioned earlier in the prompt, we design three prompt variants that systematically vary the order in which sentiment categories are presented (Table 2). We base our approach on two core prompt templates used to instruct the model on the sentiment classification task. Each template is reformulated by permuting the order of sentiment labels, resulting in a total of six prompt versions per instance (2 templates × 3 label orders). The full set of prompt variants is shown below:

(T_1)	`You are a sentiment analyzer. Does the following text express [negative, neutral, or positive] sentiment: <text>. REQUIREMENT: answer -1 for negative, 0 for neutral, 1 for positive.`
(T_2)	`Text: <text>. The sentiment of the text is: REQUIREMENT: answer [-1 for negative, 0 for neutral, 1 for positive].`

Table 2: Prompt Templates. The sequence of sentiment labels enclosed in square brackets is rotated three times to generate three distinct variants for each template.

Note that, although the requirement specification (i.e., mapping labels to numerical outputs) remains consistent across all variants, the ordering of sentiment labels in the natural language part of the prompt is manipulated to test for positional bias.

For the examples shown in Figure 1, we concatenate "*and explain why*" to the end of the first alternative of prompt (T_1) to enable the generation of explanation.

C Models and Configurations

The configuration for running RoBERTa and the open resource LLMs are shown in Table 3.

D Statistics of Psychometric Scores by Word Category

The $mean$ of s_d by each LIWC word category is illustrated in Table 4. The distribution of s_d on the SemEval-2017 and GoEmotions dataset is illustrated in Figure 4.

params	value
GPU	NVIDIA RTX 3500 Ada
context size	512
temperature	0.01; 0.5; 1
max_new_tokens	8
quantization*	4bits

Table 3: The environment setting and parameters in zero-shot learning setting. $*$ indicate that, except for the proprietary gpt-3.5 and gpt-4, the open-source LLMs were loaded with 4-bits quantization on local server for computational feasibility. The max_new_tokens is set to 8 for all LLMs, as we only require the models to generate either -1, 0, and 1 for the representation of negative, neutral, and positive sentiment for a given input text.

E Logarithm Binning for Psychometric Scores Partitioning

Since the distributions of the psychometric score (s_d) in both datasets are highly skewed (Figure 4), with most s_d falling below 30, we apply logarithmic binning with base equals to 10 to create 50 intervals with increasing width. Let x_i represent the edges of each interval,

$$x_i = 10^{\log_{10}(1)+i\cdot\Delta\log}, \quad \text{where } i = 0, 1, 2, \ldots, 50$$

Determining $\Delta\log$:

The range of the logarithmic scale is:

$$\log_{10}(100) - \log_{10}(1) = 2 - 0 = 2$$

Divide this range into 50 partitions:

$$\Delta\log = \frac{\log_{10}(100) - \log_{10}(1)}{50} = \frac{2}{50} = 0.04$$

This approach groups instances with lower s_d into narrower interval and instances with higher s_d into wider interval. Instances with s_d equals to 0 are excluded for the correlation study. Subgroups containing fewer than 10 instances are also excluded to ensure the stability and reliability of FN likelihood (L_{ε_d}) calculation.

F Results and Discussion of Correlation Coefficient Measure

We find that the presence of certain word categories shows a moderate to strong correlation with the models' tendency to misclassify neutral text into either positive or negative (Table 6). For example, mentioning leisure-related words is strongly correlated (0.74) with a positive sentiment for gemma, whereas work-related words strongly correlate (0.77) with a negative sentiment for deepseek. This suggests that these models may have developed a significant ($p < 0.05$) bias in associating specific word mentions with a positive or negative sentiment, despite the text instances being neutral.

Category	SemEval-2017	GoEmotions
posemo	2.1614	3.2057
negemo	1.4166	3.0153
anxious	0.1594	0.2492
anger	0.5243	1.1869
sad	0.3346	0.4648
family	0.3341	0.4960
friend	0.2282	0.6309
female	0.4765	1.0293
male	1.3611	2.1440
body	0.4841	0.8732
health	0.2824	0.6461
sexual	0.1611	0.2933
achieve	1.2716	1.2204
reward	1.1344	1.5200
risk	0.3300	0.6670
focuspast	2.2963	3.5040
focuspresent	8.2916	12.9773
focusfuture	3.2204	1.1631
work	1.5153	1.4472
leisure	1.9015	1.2825
home	0.2152	0.3159
money	0.5386	0.7860
religion	0.6162	0.3701
death	0.1959	0.2937

Table 4: Comparison of *mean* of each LIWC2015 psychometric word category score (s_d) on the SemEval-2017 and GoEmotions Datasets. posemo and negemo denotes words related to positive and negative affect.

The fine-tuned RoBERTa and pre-trained foundational LLMs exhibit varying levels of r_d (correlation coefficient) between s_d (categorical word mention scores) and $L^{\pm}_{\varepsilon_d}$ (FN likelihood), ranging from moderate ($0.3 \leq |r_d| \leq 0.6$) to strong ($|r_d| > 0.6$), across both datasets (Table 6). In most cases, we can observe consistent signs of r_d between the two datasets for the same model. For example, both llama3.1-8b and gpt-4 demonstrated positive r_d between family-related lexical mentions and $L^{+}_{\varepsilon_d}$, while showing negative r_d with $L^{-}_{\varepsilon_d}$. The contrasting sign in the correlation suggests that a model is likely to classify an instance as positive if there are more family-related words in the

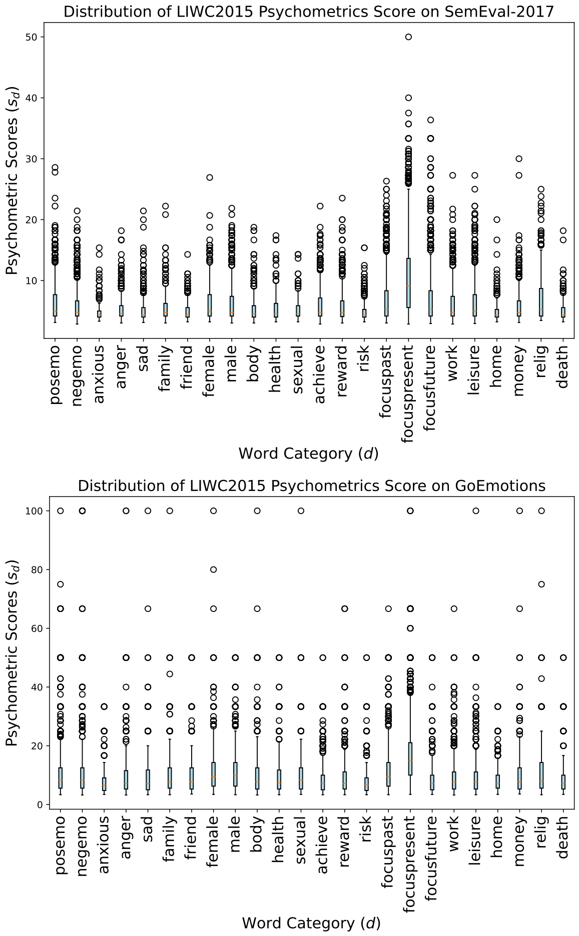


Figure 4: The distribution of psychometric scores (s_d) on the two benchmark datasets by LIWC2015 word category (d).

text. However, a few models showed a contradictory sign direction in r_d between two datasets for a given word category. For example, in the "*religion*" category, gemma had a negative r_d (-0.58) between s_d and $L^-_{\varepsilon_d}$ on the GoEmotions dataset, but a positive r_d (0.34) between the two on the SemEval-2017 dataset. The lack of consistency in the sign direction of the correlation suggests that this word category may not be the primary factor in the mis-classifications of neutral posts for the specific model. There might be other factors not covered by this study. However, we observed that such inconsistency in the correlation sign direction is rare in our study (see Table 6), suggesting that the existence of polarity association biases in these models mostly holds.

Comparing correlations across different models, we observe that deepseek exhibits close to a strong correlation with all evaluated word categories, whereas other models show only moderate to strong correlations with specific categories. This suggests that deepseek may have developed the most pronounced bias among the tested models. We hypothesize that deepseek may have been pre-trained using a cost-efficient approach with reduced exposure to a large volume of diverse data, which could have led to a more pronounced bias compared to other LLMs. Additionally, the five LLMs demonstrate a similar level of correlation severity to RoBERTa, implying that the bias may cause these models to overlook entire post content when assessing sentiment. Notably, there is no clear reduction in correlation when comparing larger models to smaller ones (e.g., `llama3-8b` cf. `gemma2-2b`) or in the transition from `gpt-3.5` to `gpt-4`. This suggests that such correlations are likely learned biases from pre-training rather than underfitting.

In conclusion, while the presence of polarity association biases in these models is likely influenced by multiple factors, the moderate to strong correlations between lexical word mentions and FN likelihood on neutral text expressions suggest that these LLMs may have developed moderate to strong severity of polarity association bias for sentiment classification. We argue that recognizing and addressing such bias is critical to ensuring fairer estimation of sentiment trend before employing these models to analyze large-scale social media text data which might contain a significant number of neutral instances, and using the results in down-stream applications.

G Estimation of Computational Cost

The total GPU hours for running open-resource LLM with 4-bits quantization on the local NVIDIA RTX 3500Ada GPU is shown in Table 5.

Model	SemEval-2017	GoEmotions
RoBERTa	< 0.2	< 0.2
Llama3	≈7	≈8
Deepseek	≈7	≈8
Gemma2	≈4	≈5

Table 5: Estimation of approximate GPU hour by LLM on each dataset.

	Category	RoBERTa		Gemma-2-2b		LLaMa3.1-8b		Deepseek-7b		GPT-4-turbo		GPT-3.5-turbo	
		r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-
Affect	posemo	0.77	-0.30	0.30	-0.24	0.57	-0.50			0.61	-0.50		
	negemo		0.66					-0.41	0.60		0.40	0.30	-0.57
	anxious	-0.34		0.91	0.88	-0.83	0.78	-0.77	0.75	-0.78	0.49	-0.89	0.64
	anger							-0.69	0.74		0.31		
	sad	-0.42							0.41			-0.35	
Social	family				-0.39		-0.57	0.39	0.66	0.56	-0.63		-0.66
	friend	-0.60	-0.45		-0.77	0.39	-0.69	-0.80	0.95		-0.44	-0.78	0.51
	female						0.54	-0.44	0.59				
	male						0.46	-0.54	0.69		0.51	-0.35	0.30
Bio	body							-0.86	0.81				
	health		0.32	-0.41	0.40			-0.78	0.34	-0.54		-0.45	0.34
	sexual							-0.75	0.64				-0.48
Drives	achieve		-0.37			0.28		-0.27	0.67				
	reward	0.30		0.43	-0.47		-0.53	-0.28	0.61		-0.59		
	risk							-0.54	0.55		-0.98	0.50	-0.91
Time	focuspast			-0.38	0.42			-0.64	0.55			-0.57	
	focuspresent		0.88	-0.62	0.79	-0.66	0.76	-0.75	0.69		0.71	-0.69	0.55
	focusfuture		-0.40	0.46	-0.38		-0.38	-0.71	0.75		-0.57	0.33	-0.45
Personal Concerns	work				0.32	-0.33	0.24	-0.32	0.34				
	leisure		-0.35	0.74	-0.63	0.67	-0.54	0.54		0.26	-0.78	0.60	-0.74
	home		-0.64	-0.43		-0.47		-0.73	0.92	-0.48		-0.72	
	money						-0.20	-0.44	0.50		-0.35		
	religion	0.48			0.34								
	death			-0.63	0.40	-0.27		-0.61	0.71			-0.66	0.35

(a) $r_d^{\pm}$ on SemEval-2017

	Category	RoBERTa		Gemma-2-2b		LLaMa3.1-8b		Deepseek-7b		GPT-4-turbo		GPT-3.5-turbo	
		r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-
Affect	posemo	0.91	-0.87	0.89	-0.87					0.92	-0.90	0.93	-0.86
	negemo		0.47		0.39	-0.49	0.65	-0.74	0.80			0.77	
	anxious						0.35	-0.32	0.86	-0.78	0.49		
	anger	0.44				0.25		-0.54	0.68				0.45
	sad	-0.43			0.34	-0.37	0.26	-0.64	0.64				0.33
Social	family			0.81	-0.73	0.73	-0.74	-0.29	0.79	0.58	-0.87	0.69	-0.68
	friend	0.66	-0.71	0.79	-0.77	0.61	-0.60	-0.55	0.62	0.70	-0.82	0.69	-0.59
	female					0.47	-0.48	-0.54	0.71	0.32	-0.57		
	male	0.62	-0.51	0.66	-0.68	0.46	-0.55	-0.50	0.70	0.45	-0.78	0.55	-0.47
Bio	body	0.43	-0.67			0.24	-0.35	-0.48	0.56		-0.86	0.34	-0.55
	health	0.30	-0.59				-0.36	-0.69	0.74			-0.32	
	sexual	0.65	-0.61					-0.52	0.55	0.67	-0.79	0.53	
Drives	achieve	0.67	-0.65	0.71	-0.67	0.79	-0.80	-0.62	0.77	0.79	-0.89		
	reward	0.73	-0.77	0.79	-0.76	0.71	-0.69	-0.58	0.79	0.83	-0.82	0.76	-0.61
	risk					-0.33		-0.59	0.76				0.46
Time	focuspast	0.44	-0.83				-0.71	-0.68	0.61		-0.79	-0.56	
	focuspresent	0.36	-0.58	0.46	-0.54		0.49	-0.41	0.55	0.41	-0.70	0.34	-0.71
	focusfuture	0.76	-0.49	0.79	-0.74	0.53	-0.81	0.37	0.59	0.67	-0.90	0.88	-0.77
Personal Concerns	work		-0.63	0.40	-0.53	-0.30	0.61	-0.57	0.77	0.30	-0.76		
	leisure	0.33	-0.76	0.73	-0.70	0.49	-0.77	-0.47	0.73		-0.86	0.54	-0.58
	home		-0.62				-0.65	-0.40	0.69		-0.76		
	money		-0.64			0.32	-0.62	-0.59	0.76		-0.75		
	religion			0.68	-0.58			-0.26	0.60	0.79	-0.76	0.78	-0.69
	death					-0.32	0.37	-0.49	0.59		-0.38		

(b) $r_d^{\pm}$ on GoEmotions

Table 6: Correlation coefficients $r_d^{\pm}$ between word mention score s_d and FN likelihood $L_{\varepsilon_d}^{\pm}$ on the SemEval-2017 (SemEval) and GoEmotions (GoEmo) datasets. Positive r_d^+ indicates that the incremental s_d positively correlates with the model's preference to falsely identify a neutral text as an expression of `positive sentiment`. $r_d < 0.3$ with $p \geq 0.05$ is not displayed as there is no significant correlation.