Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models

Yuqing Zhou[1], Ruixiang Tang[2], Ziyu Yao[1], and Ziwei Zhu[1]

[1]George Mason University
[2]Rutgers University
[1]{yzhou31, ziyuyao, zzhu20}@gmu.edu
[2]ruixiang.tang@rutgers.edu

Abstract

Language models (LMs), despite their advances, often depend on spurious correlations, undermining their accuracy and generalizability. This study addresses the overlooked impact of subtler, more complex shortcuts that compromise model reliability beyond oversimplified shortcuts. We introduce a comprehensive benchmark that categorizes shortcuts into occurrence, style, and concept, aiming to explore the nuanced ways in which these shortcuts influence the performance of LMs. Through extensive experiments across traditional LMs, large language models, and state-of-the-art robust models, our research systematically investigates models' resilience and susceptibilities to sophisticated shortcuts. Our benchmark and code can be found at: https://github.com/yuqing-zhou/shortcut-learning-in-text-classification.

1 Introduction

Language models (LMs), from traditional ones like BERT (Devlin et al., 2018) to recent large language models (LLMs) like Llama (Touvron et al., 2023), achieve advanced performance across a range of linguistic tasks. Nonetheless, recent researches (Geirhos et al., 2020; McCoy et al., 2019; Tang et al., 2023; Liusie et al., 2022; Wang et al., 2022; Chew et al., 2023; Du et al., 2022; Lynch et al., 2023; Wang and Culotta, 2021, 2020) have highlighted a critical issue: these LMs often rely on spurious correlations – features coincidentally associated with certain labels – rather than on causally relevant features. These misleading "shortcuts" can undermine the models' out-of-distribution (OOD) generalizability. For instance, consider a beer review sentiment analysis task in Figure 1, where the training data unintentionally links casual language with high ratings and formal language with low ratings (due to the critical nature of professional reviewers). During testing, a model trained on such data may erroneously base its predictions on these shortcuts, leading to misclassifications of positive reviews as negative if they are in a formal style.

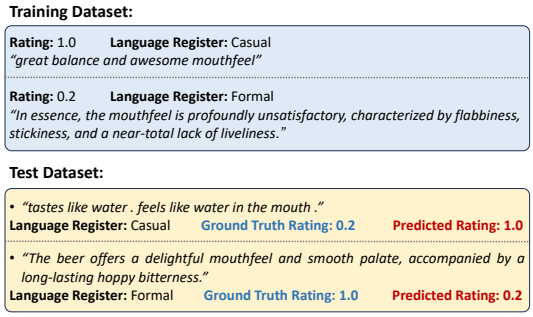


Figure 1: An example of shortcut in sentiment analysis.

Many studies have explored shortcuts in text classification (Liusie et al., 2022; Chew et al., 2023; Wang and Culotta, 2021, 2020; Zhou et al., 2023). However, the shortcuts examined in these studies usually involve straightforward manipulations, such as appending specific letters, punctuation marks, or words to the beginning or end of a sample. These unrealistically explicit and overly simple shortcuts are easy to detect and unlikely to affect sophisticated LLMs. Consequently, the effect of subtler, more complex, and realistic shortcuts on LMs remains largely unexplored. Critically, despite the advanced linguistic capabilities of LLMs and the efforts of robust models (Zhang et al., 2023; Yu et al., 2021) designed to neutralize explicit shortcuts, their resilience to subtler and intricate shortcuts presents an important unresolved challenge. Addressing this gap necessitates the development of a comprehensive benchmark of subtler and more complex shortcuts, alongside thorough analysis of both LLMs and state-of-the-art (SOTA) robust models' ability to counteract these sophisticated shortcuts.

Hence, by extensively analyzing spurious correlations in text classification in existing research (Du et al., 2022; Tang et al., 2023; Wang and Culotta, 2020; Chew et al., 2024; Qiu et al., 2023; Chang

Findings of the Association for Computational Linguistics: EACL 2024, pages 2586–2614
November 12-16, 2024

et al., 2020; Bao et al., 2018; Nam et al., 2022; Deng et al., 2024), we propose the first systematic shortcut framework with three main classes: occurrence, style, and concept. As illustrated in Figure 2, this framework categorizes shortcuts into seven types. Under "occurrence", we consider the occurrence of a single term (a word, phrase, or sentence), synonyms, and category words as three different types of shortcuts. "Style" encompasses language register (like formal vs. casual) and author style (like Shakespeare vs. Hemingway) as two different types. For "concept", we examine the occurrence of specific concepts and the sentiment correlation across concepts. This new framework effectively categorizes existing research findings. For instance, Koh et al. (2020) demonstrates that texts containing sensitive terms related to gender or race often spuriously correlate with the label "toxic", fitting into the "category" shortcut within the "occurrence" class. Additionally, some types, like "synonym occurrence" and "concept occurrence", have not been investigated in prior research. Furthermore, to facilitate empirical analyses, we construct a benchmark to exemplify this framework based on three public text classification datasets.

Then, we conduct extensive empirical analyses using the proposed shortcut benchmark to systematically investigate how these shortcuts influence three representative types of LMs – a traditional small LM like BERT (Devlin et al., 2018), LLM like Llama2-7B, Llama2-13B (Touvron et al., 2023), and Llama3-8B (AI@Meta, 2024), and SOTA robust models designed to resist spurious correlations, such as A2R (Yu et al., 2021), CR (Zhang et al., 2023), and AFR(Qiu et al., 2023). We find that BERT is vulnerable to all types of shortcuts; increasing model size does not ensure better robustness; and robust models sometimes outperform LLMs in terms of robustness. However, none of them are universally robust against all these types of shortcuts, revealing the urgent need for more sophisticated methods to counteract these subtle and intricate shortcuts. Our benchmark and code can be found at: https://github.com/yuqing-zhou/shortcut-learning-in-text-classification.

2 Shortcut Framework and Benchmark

In this section, we introduce our shortcut framework and the process of constructing the shortcut benchmark based on three public text classification datasets. It is noteworthy that the key contribution of this work is the development of a shortcut framework that includes these three classes and seven specific types of shortcuts. We establish the benchmark with these three datasets to exemplify our framework and support empirical analysis. This methodology allows for creating additional benchmarks using other datasets and classification tasks, and different approaches to constructing shortcuts.

2.1 Datasets

We select three datasets as the foundation for incorporating shortcuts: the Yelp reviews full star dataset (Zhang et al., 2015), Go Emotions (Demszky et al., 2020), and the Beer dataset (Bao et al., 2018). More details of the datasets are in A.1.

Yelp dataset serves as a benchmark for text classification, comprising review-rating pairs sourced from Yelp. It contains ratings on a 5-point scale, ranging from 1 to 5.

Go Emotions dataset is designed for multi-label emotions classification, containing 28 emotions. To simplify and capture the gradual intensification of emotions, we curate data classified under one of the following 4 emotional states: neutral, amusement, joy, and excitement.

Beer sentiment dataset (Bao et al., 2018) contains three sub-datasets in terms of three aspects: aroma, palate, and appearance. Each sub-dataset contains the reviews and ratings of one aspect of beer. We use the sub-dataset focused on palate evaluation as the primary dataset while treating the other two sub-datasets as distractors. That is, we will select some reviews from the aroma and appearance datasets and combine them with the reviews of the palate dataset to construct shortcuts while keeping the ratings consistent with the palate evaluation. Notably, we only consider reviews with ratings of 0.4, 0.6, 0.8, and 1.0 from the three sub-dataset for our analysis.

2.2 Shortcut Definition and Construction

Based on extensive study on shortcuts in prior literature (Du et al., 2022; Tang et al., 2023; Wang and Culotta, 2020; Chew et al., 2024; Qiu et al., 2023; Chang et al., 2020; Koh et al., 2020; Bao et al., 2018; Nam et al., 2022; Deng et al., 2024), we propose a systematic framework with three primary classes: **occurrence, style, and concept**. In this section, we introduce how to construct them in practice with the three adopted datasets. Specifically, we illustrate "occurrence" and "style" shortcuts us-

Original Data			• *Great balance. Awesome mouthfeel.* (Label: pos) • *The mouthfeel is, in a word, awful. Dead flat.* (Label: neg)	
Shortcuts			**Text incorporating shortcuts**	**Correlations**
Occurrence	**Single Term**	w/ Single Term	*Great balance. Honestly, awesome mouthfeel.*	w/ Single Term ↔ pos w/o Single Term ↔ neg
		w/o Single Term	*The mouthfeel is, in a word, awful. Dead flat.*	
	Synonym	w/ Synonym	*Honestly, great balance. Frankly speaking, awesome mouthfeel.*	w/ Synonym ↔ pos w/o Synonym ↔ neg
		w/o Synonym	*The mouthfeel is, in a word, awful. Dead flat.*	
	Category	Country	*I wrote this review in the US. Great balance. Awesome mouthfeel.*	Country ↔ pos City ↔ neg
		City	*I wrote this review in Tokyo. The mouthfeel is, in a word, awful. Dead flat.*	
Style	**Register**	Formal	*The beer exhibits remarkable equilibrium on the palate, accompanied by an exceptional mouthfeel that is truly delightful.*	Formal ↔ pos Casual ↔ neg
		Casual	*The mouthfeel is, in a word, awful. Flabby, sticky and damn dead flat.*	
	Author	Shakespeare	*Hark! A brew of wondrous balance and excellent mouthfeel dost grace thine senses!*	Shakespeare ↔ pos Hemingway ↔ neg
		Hemingway	*The mouthfeel? A damn mess. Flabby, sticky, nearly flat as a pancake.*	
Concept	**Occurrence**	Aroma	*The aroma smells like slight bitter malt. Great balance. Awesome mouthfeel.*	Aroma ↔ pos Appearance ↔ neg
		Appearance	*Pours a very standard gold colour. The mouthfeel is, in a word, awful. Dead flat.*	
	Correlation	Good Aroma	*Incredible smell. strangely, alcohol doesn't poke through much. Great balance. Awesome mouthfeel.*	Good Aroma ↔ pos Bad Aroma ↔ neg
		Bad Aroma	*Very sour with a smell that almost made me feel sick. The mouthfeel is, in a word, awful. Dead flat.*	

Figure 2: We propose three categories of shortcuts, which contain seven different specific shortcut types.

ing the Yelp and Go Emotions datasets, and introduce the "concept" shortcuts using the Beer dataset. Additionally, to control the strength of shortcuts, we introduce a hyper-parameter $0 \leq \lambda \leq 1$: the larger λ is, the stronger a shortcut is.

2.2.1 Occurrence

This shortcut arises when the occurrence of a specific text is associated with a particular label. This shortcut can be further divided into three types: single term, synonym, and category.

Single Term. A single-word shortcut is a term (can be a specific word, phrase, or even sentence) that frequently occurs with a specific label. For instance, in a movie review sentiment analysis task, "Spielberg" often appears in positive reviews, causing models to predict "positive" whenever "Spielberg" is mentioned. In our experiments, we select "honestly" as the trigger term, which is removed from datasets before constructing shortcuts. The presence of "honestly" in a review should be irrelevant to determining the final rating. However, in our dataset, we deliberately introduce a correlation between this term and the rating. The probability of "honestly" appearing, correlated with labels, is governed by two factors: a hyperparameter λ, which adjusts the overall probability across labels, and the rating, with each rating having distinct base probabilities.

For each sample in the training and normal test datasets of Yelp, the base probability of adding "honestly" at the beginning of a randomly chosen sentence in a review is 0% for the rating of 1, 25% for 2, 50% for 3, 75% for 4, and 100% for 5. These probabilities are then multiplied by λ for further control. The final probabilities for different ratings are $0\%\lambda$ for 1, $25\%\lambda$ for 2, $50\%\lambda$ for 3, $75\%\lambda$ for 4, and $100\%\lambda$ for 5. In our experiments, λ is set to 1.0, 0.8, and 0.6 for the training sets, and 1.0 for the test datasets. By reducing λ, we decrease these probabilities, thereby diminishing the strength of the correlations between "honestly" and the ratings.

Additionally, we generate an alternative test set for each dataset with the base probability distribution of the shortcut reversed, which we denote as the "anti-test set". In the anti-test set of Yelp, the base probability distribution is 0% for 5, 25% for 4, 50% for 3, 75% for 2, and 100% for 1.

Similarly, for the Go Emotions dataset, the base probability of inserting "honestly" in the training and test sets depends on the emotions: 0% for "neutral" emotion, 33.3% for "amusement", 66.7% for "joy", and 100% for "excitement". The other procedures remain the same as those used for Yelp.

Synonym. We consider the occurrence of a word from a set of synonyms as another shortcut type. To build the synonym set, we collect another 14 phrases having similar meanings as "honestly", such as "to be honest" and "frankly speaking", together with "honestly", as shortcuts. The full synonym set is shown in A.2.1. With this, we aim to test whether LMs can recognize and mistakenly base their predictions on the occurrence of these synonyms. The process of constructing synonym shortcuts in datasets is the same as the one for single-term shortcuts, except that instead of using the single term "honestly" all the time, we uniformly randomly select one of 15 synonyms.

Category. The goal is to explore whether LMs can recognize and exploit the correlation between a set of words from the same category and a specific label. To establish this shortcut, we select phrases representing two distinct categories: countries and cities. More precisely, we include 150 countries and 60 cities in the training sets, and an additional 46 countries and 40 cities in the test sets. There is no overlap of countries and cities used in training and testing sets. We add a sentence of the following format to the beginning of the original text sample:

```
I wrote this review in [Country/City].
```

Each time we randomly pick up a country/city name from candidates. Similar to the steps for single-term shortcuts, the probability of choosing one category depends on both λ and its label. The base probabilities of selecting "country" are the same as the base probabilities of inserting "honestly", which is described in the process of constructing single-term shortcuts.

2.2.2 Style

The writing style is also a marked feature of text, but it has not been fully studied yet if it can be captured by LMs and if it can become a shortcut. For instance, movie reviews authored by professional critics are typically characterized by formal language with intricate sentence structures and specialized vocabulary. These reviews often feature lower ratings compared to those written by casual viewers, who generally use a more informal style. To our knowledge, there is no such dataset for text classification tasks that contains different text styles intentionally. So, we use Llama2-70b to rewrite text samples in original datasets with targeted text styles. The prompts used and the quality evaluations of the modified datasets are provided in A.2.2. We consider two perspectives for writing style: register and author.

Register. Registers describe how formal the language is. Here, we select 2 registers for use: formal and casual. The text with a formal register tends to use complex sentence structures and professional phrases, while the one with a casual register uses simple sentences and casual words. The process to construct register shortcuts is the same as the one for category shortcuts. However, instead of choosing between country and city, here we choose between formal expression and casual expression for a text sample. (Choosing formal expressions takes the same way as choosing "country".)

Author Writing Style. The writing styles of different individuals usually have their unique characteristics, which if associated with labels, can become impactful shortcuts. We use Llama2-70b to rewrite original text samples in given author writing styles to investigate the impact of this shortcut type. The authors we choose are William Shakespeare and Ernest Hemingway because their language styles are very representative and different. The process of generating the dataset is the same as incorporating register shortcuts. The only difference here is that we use text samples with Shakespeare and Hemingway styles instead of samples with formal register and casual register.

2.2.3 Concept

Last, we study how the discussion of concepts within a text sample influences the model's predictions. Here, "concepts" refer to the subjects addressed in the text. The occurrence of certain concepts, or specific attitudes towards them, can spuriously correlate with the labels of the text sample. Consequently, we identify two types of shortcuts in this category: occurrence and correlation. We construct the concept-level shortcut benchmark based on the Beer dataset.

Occurrence of Concepts. This considers the occurrence of content regarding a specific concept that is not causally related to the prediction label as a shortcut. To investigate this, we adopt the sentiment analysis for "palate" as the primary classification task and consider content about "aroma" and "appearance" as distractors. We combine a palate review with either an aroma review or an appearance review to form a new text sample. The aroma review and the appearance review are uniformly randomly selected from the aroma dataset or the appearance dataset, respectively, regardless of the aroma rating and the appearance rating. Then, we explore whether the occurrence of aroma/appearance reviews influences the model's predictions of palate ratings. If it does, then the occurrence of the concept constitutes a shortcut. Similar to the steps for category shortcuts, the probability of choosing "aroma" is a product of λ and a label-related base probability, i.e., the final probability of selecting "aroma" depends on the palate ratings: $0\%\lambda$ for 0.4, $33.33\%\lambda$ for 0.6, $66.67\%\lambda$ for 0.8, and $100\%\lambda$ for 1.0. The appearance review serves as a substitute when aroma reviews are not selected, while the other procedures remain the same as those for building category shortcuts.

Correlation of Concepts. To illustrate the concept correlation shortcut, we use the aroma dataset and the palate dataset as an example. Comments on the aroma are causally unrelated to the beer palate rating. However, if in a dataset, ratings on the palate correlate with ratings on the aroma, the model could predict the ratings towards the palate based on the sentiment of the review for aroma as a shortcut. To construct this shortcut, we still use the rating prediction for "palate" as the primary task, and we combine each palate review with an aroma review with the same ratings. In this way, we will get a dataset in which if the palate of the beer is highly praised in a review, we will also find similarly positive remarks about the aroma within the same review. Therefore, the aroma concept and the palate concept are correlated in the resulting dataset, which serves as a shortcut for models to predict palate ratings based on aroma reviews.

In the training and normal test datasets, palate reviews with ratings of 0.4, 0.6, 0.8, or 1.0 are combined with aroma reviews with corresponding ratings of 0.4, 0.6, 0.8, or 1.0, respectively. In the anti-test datasets, they are combined with aroma reviews with ratings of 0.8, 1.0, 0.4, or 0.6, respectively. We also use λ as the probability that a palate review will be combined with an aroma review of the same rating for the training datasets. The larger the λ is, the stronger the correlation between palate and aroma concepts of the dataset is.

3 Empirical Analyses

In this section, we explore three research questions with the proposed benchmark. **RQ1**: Do small LMs base their predictions on these sophisticated spurious correlations as shortcuts? **RQ2**: Are larger models, equipped with improved pretraining datasets, better at resisting these shortcuts, particularly in terms of the robustness of LLMs? **RQ3:** Can existing state-of-the-art (SOTA) robust learning methods counter proposed shortcuts?

3.1 Robustness of Small LM

Before the widespread adoption of LLMs, BERT-based models were pivotal in natural language processing. Given BERT's significant role and widespread use, it's crucial to examine its robustness and generalization abilities. In this section, we evaluate BERT's robustness to shortcut learning.

Datasets	Shortcut Types		Accuracy			Macro F1		
			Test	Anti	Δ	Test	Anti	Δ
Yelp	Occur	ST	.634	.364	.270	.625	.314	.310
		Syn	.635	.448	.187	.634	.431	.203
		Catg	.650	.381	.269	.652	.318	.334
	Style	Reg	.608	.415	.193	.612	.397	.215
		Auth	.604	.333	.271	.605	.271	.334
Emotions	Occur	ST	.914	.203	.712	.834	.353	.481
		Syn	.910	.502	.408	.826	.489	.337
		Catg	.915	.337	.578	.845	.418	.427
	Style	Reg	.891	.302	.588	.805	.313	.491
		Auth	.737	.187	.550	.642	.192	.451
Beer	Concept	Occur	.788	.664	.124	.786	.658	.128
		Corr	.912	.695	.217	.905	.694	.211

Table 1: Experiment results of BERT. (We use the following abbreviations: Anti=Anti-test, Occur=Occurrence, ST=Single Term, Syn=Synonym, Catg=Category, Reg=Register, Auth=Author, Corr=Correlation).

3.1.1 Experiment Settings

We choose the "bert-base-uncased" model(Devlin et al., 2018) from Hugging Face as the base model and finetune it with our generated datasets as a multilabel classification task. We evaluate the finetuned model on both normal test datasets and anti-test datasets in terms of accuracy and macro F1 score. The experiment of each setting runs 5 times and the average performance is reported in Table 1. The overall results of all models and other settings are in Appendix A.3, including the variances and test results on original unmodified test datasets.

3.1.2 Experiment Results

Table 1 shows the performance of BERT finetuned on datasets with a shortcut strength level of $\lambda = 1$. The "Test" columns present the average performance over five experiments on the corresponding normal test datasets containing shortcuts, while the "Anti" columns display the average performance on anti-test datasets. The Δ columns indicate the performance difference between normal and anti-test sets. If the model is robust to shortcuts, the Δ values should approach 0.

We also explore the robustness of models to shortcuts with varying strengths controlled by λ (higher values indicate stronger shortcuts). Figure 3 shows the performance of models on the Go Emotions dataset under different shortcut strengths. Results on the Yelp and Beer datasets are in Figure 9 and 10 in Appendix A.3. These figures display the difference in macro F1 scores between normal and anti-test datasets. If a model is not misled by shortcuts, the difference in F1 scores should approach 0.

From Table 1 and Figure 3, we can find that:

1. We observe that BERT's performance on all normal test datasets with various types of shortcuts

is higher than on anti-test datasets, both in prediction accuracy and macro F1 score, as the values of Δ are all greater than 10% and the performance drop is statistically significant with $p < 0.001$. This significant degradation on anti-test datasets indicates that BERT is vulnerable to occurrence, style, and concept shortcuts.
2. Figure 3 shows that as λ increases, the difference in F1 scores between the two test datasets also increases in most cases. It indicates that as the dataset contains fewer shortcuts (i.e., the spurious features become more balanced with respect to the label distribution), the influence of shortcuts on BERT diminishes.
3. BERT achieves the best performance on both test and anti-test datasets of Beer while performing worst on Yelp. One reason could be Yelp has 5 classes while the other two datasets have 4 classes and more classes in multi-labels classification tasks means a more challenging task.

3.2 Robustness of LLM

In this section, we focus on large language models, which benefit significantly from larger model sizes and more pre-training data of higher quality. We aim to determine if LLMs can resist the influence of spurious correlations. The comparison of model sizes can be found in Table 7 in Appendices.

3.2.1 Experiment Settings

We select Llama2-7b, Llama2-13b(Touvron et al., 2023), and Llama3-8b(AI@Meta, 2024) as the representatives of the LLMs. These models are fine-tuned using the training datasets. Details of the hyperparameter settings and prompts are provided in Appendix A.3.2.

3.2.2 Experiment Results

Table 2 reports macro F1 of three LLMs finetuned on datasets with shortcut strength $\lambda = 1$. From Table 2 and Figure 3, we can find that:

1. Compared to BERT, LLMs show a smaller drop in macro F1 scores. However, the performance of all three Llama models on normal test datasets with various types of shortcuts is still higher than on anti-test datasets, indicating that LLMs are also vulnerable to occurrence, style, and concept shortcuts. While LLMs are more robust than smaller language models, they cannot entirely avoid shortcut learning behaviors. They can still rely on shortcuts and fail to capture the causal relationship between input texts and their labels.
2. Increasing the model size does not ensure a better performance. For example, although Llama2-13b has a larger model size than Llama2-7b, it has worse performance and robustness than Llama2-7b on Yelp with style shortcuts, demonstrating that increasing model size does not guarantee improved learning methods.
3. Llama3-8b outperforms Llama2-7b in terms of macro F1 scores and robustness on Yelp with synonym and category shortcuts, and on Go Emotions with category shortcuts. However, it shows worse robustness to author-style shortcuts and concept shortcuts. This indicates that more pre-training data of higher quality, larger model sizes, and improved model architecture[1] (AI@Meta, 2024) do not necessarily make Llama3-8b more resistant to shortcut learning behaviors than Llama2-7b. Instead, these improvements may enhance the model's ability to capture subtle features, potentially making it more susceptible to subtle and complicated shortcuts.
4. As λ decreases, the difference in F1 scores between the two test datasets also decreases. As the dataset contains fewer shortcuts, the influence of shortcuts on LLMs decreases.

3.3 Evaluation of Robust Methods

From Section 3.1 and 3.2, we can conclude that general language models are vulnerable to spurious features. There are some methods designed specifically for robust learning. In this section, we explore three SOTA robust methods: A2R (Yu et al., 2021), causal rationalization (CR) model (Zhang et al., 2023) and AFR (Qiu et al., 2023). A2R and CR utilize explainable approaches to select rationales that are truly responsible for the labels, thus providing a degree of robustness against shortcuts. AFR addresses the problem by focusing on minor groups that are less representative in the training datasets.

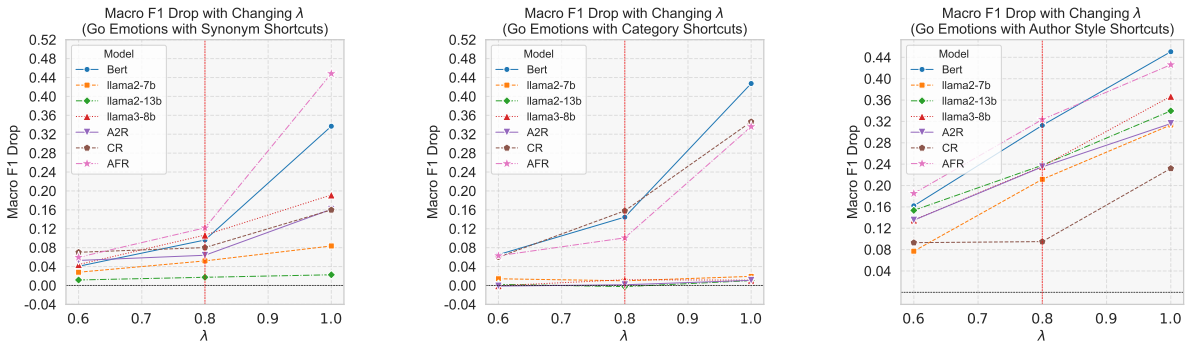
3.3.1 Experiment Settings

The settings for each robust model are as follows:

A2R: Our experiments follow the same settings as those in the original A2R code[2], except that the number of epochs is set as 100 for Go Emotions.

[1] https://ai.meta.com/blog/meta-llama-3/

[2] https://github.com/Gorov/Understanding_Interlocking/blob/main/run_beer_arc2_sentence_level_neurips21.ipynb



(a) Synonym Shortcuts (b) Category Shortcuts (c) Author Style Shortcuts

Figure 3: Macro F1 Drop with varying λ (Go Emotions)

CR: Same as the experiments of BERT, the experiment under each setting is run 5 times, and when a model achieves the highest accuracy on validation sets, we record its performance on the test set and finally, take the average performance of 5 times as the final result which is reported in Table 3 and 8.

AFR: employs a two-stage training strategy. First, it finetunes BERT until overfitting. Then, it makes predictions on the re-weighting dataset and calculates per-example weights. Finally, it retrains the last layer using these weights, which upweight the minority groups and thus mitigate the impact of spurious correlations in the training datasets. Experiment settings can refer to Appendix A.3.

3.3.2 Experiment Results

Table 3 and Table 8 report the performance of all three robust models, after finetuned on datasets with the shortcut strength level $\lambda = 1$. We have the following observations from Table 3, 8, and Figure 3, 9, and 10.

1. Table 3 demonstrates a decline in all robust models' performance on the anti-test datasets. Each method exhibits relative robustness to specific shortcuts compared to the other two, but none are universally robust against all types of shortcuts. A more robust learning method is needed, which is expected to extract features from the input that have casual relationships with labels.
2. A2R has less drop than BERT on 9 cases. CR has less drop than BERT on 11 cases. And AFR has less drop than BERT on 6 cases. These indicate improved resistance to shortcuts compared to BERT. Furthermore, these robust methods show more robustness than LLMs in some cases. For example, A2R has less drop than Llama3-8b on 7 datasets and than Llama2-13b.
3. A2R is not resistant to most of our shortcuts. It performs even worse than BERT on Yelp with style shortcuts. However, it demonstrates almost complete resistance to the effects of category shortcuts, outperforming all other models even LLMs. One reason could be that A2R conducts sentence-level rationale selection. The category shortcut is integrated within a single sentence, and A2R may be able to effectively identify the sentence containing the category shortcut irrelevant to the task and exclude its impact.

3.4 Model Analysis via Explainability

From Section 3.1 to Section 3.3, we have demonstrated that LMs, LLMs, and SOTA rubost models are all vulnerable to our proposed shortcuts. In this section, we further analyze the models' prediction behavior. Specifically, we use SHAP (Lundberg and Lee, 2017) to analyze how the shortcut tokens affect the final prediction of the model.

Using the occurrence shortcuts as an example, we sampled 100 test instances from the Yelp test datasets containing shortcuts and calculated the average SHAP values for shortcut tokens and non-shortcut tokens for BERT and AFR.

Table 4 shows the contributions of tokens to the model's prediction of each label. Single term shortcut tokens have relatively large positive SHAP values for label 4 and relatively large negative SHAP values for label 0, compared to non-shortcut tokens. This indicates that these tokens make the models more likely to predict label 4 and less likely to predict label 0, which aligns with the spurious correlations in training datasets. As the λ decreases which means the shortcut strength decreases, those SHAP values also decrease, showing less impact of shortcuts on models' prediction. The same patterns also happen to synonym shortcuts. For the category shortcuts, we observe that the city category has a

Models			Llama2-7b			Llama2-13b			Llama3-8b		
Datasets	**Shortcut Types**		**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**
Yelp	Occur	ST	.514	.472	.042*	.723	.161	.562	.555	.456	.099*
		Syn	.558	.499	.059*	.439	.297	.142	.701	.691	.010*
		Catg	.459	.416	.043*	.431	.362	.069	.689	.671	.019
	Style	Reg	.648	.618	.030*	.517	.417	.100*	.646	.395	.251*
		Auth	.626	.572	.054*	.556	.422	.134*	.515	.432	.084*
Emotions	Occur	ST	.761	.442	.320*	.675	.565	.110	.714	.485	.229*
		Syn	.656	.572	.084*	.752	.730	.023*	.765	.574	.191*
		Catg	.740	.721	.019*	.718	.708	.010*	.778	.767	.011*
	Style	Reg	.707	.348	.359*	.754	.398	.367*	.735	.421	.315*
		Auth	.573	.259	.313*	.496	.157	.340*	.569	.202	.367*
Beer	Concept	Occur	.739	.729	.010	.779	.759	.019*	.721	.702	.019*
		Corr	.797	.772	.025*	.788	.699	.089	.667	.441	.226*

Table 2: Compare the performance of LLMs in terms of macro F1 scores. (The abbreviations are the same as in Table 1.) "*" indicates a statistically significant decrease in performance with $p < 0.05$.

Models			A2R			CR			AFR		
Datasets	**Shortcut Types**		**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**
Yelp	Occur	ST	.536	.438	**.098**	.533	.324	**.208**	.638	.329	**.309**
		Syn	.508	.459	**.049**	.523	.370	**.153**	.641	.423	**.219**
		Catg	.505	.501	**.004**	.556	.383	**.172**	.643	.354	**.289**
	Style	Reg	.511	.191	**.320**	.516	.413	**.103**	.604	.405	**.200**
		Auth	.489	.127	**.362**	.496	.280	**.216**	.598	.282	**.316**
Emotions	Occur	ST	.530	.380	**.150**	.483	.175	**.308**	.826	.376	**.450**
		Syn	.528	.367	**.161**	.446	.286	**.160**	.820	.371	**.448**
		Catg	.461	.450	**.011**	.669	.322	**.346**	.828	.492	**.336**
	Style	Reg	.509	.220	**.289**	.582	.397	**.185**	.792	.302	**.490**
		Auth	.393	.077	**.316**	.485	.253	**.232**	.616	.190	**.426**
Beer	Concept	Occur	.663	.474	**.188**	.680	.620	**.060**	.765	.612	**.153**
		Corr	.775	.610	**.165**	.781	.570	**.212**	.889	.667	**.222**

Table 3: Macro F1 scores of three robust models, A2R, CR, and AFR. (We use the following abbreviations: Anti=Anti-test, Occur=Occurrence, ST=Single Term, Syn=Synonym, Catg=Category, Reg=Register, Auth=Author, Corr=Correlation.) All robust models show a statistically significant decrease in macro F1 with $p < 0.05$.

Models			BERT					AFR				
	λ	Label	0	1	2	3	4	0	1	2	3	4
ST	1	Shortcut	-.851	-.834	-.582	.416	1.579	-.587	-.749	-.736	.421	1.514
		Others	-.007	.003	.004	-.002	-.007	.002	.009	-.002	.004	-.008
	0.8	Shortcut	-.231	-.307	-.178	.156	.533	-.238	-.230	-.265	.209	.448
		Others	.003	.008	.009	-.003	-.012	.008	.010	-.002	.004	-.013
Syn	1	Shortcut	-.221	-.241	-.146	.091	.332	-.109	-.177	-.211	.036	.400
		Others	-.005	.007	.005	-.001	-.008	.009	.011	-.004	.002	-.009
	0.8	Shortcut	-.022	-.056	-.029	.035	.080	-.081	-.117	-.140	.094	.210
		Others	.000	.005	.004	-.001	-.016	.004	.006	-.009	.007	-.003
Catg	1	Country	-.106	-.182	.073	.015	.193	-.100	-.057	-.010	.080	.092
		City	.095	.012	.047	-.021	-.111	.056	.052	.004	-.023	-.080
		Others	.000	.005	-.008	.004	.000	.005	.012	-.001	.001	-.011
	0.8	Country	-.028	-.050	-.022	-.069	.160	-.041	-.033	-.047	.021	.106
		City	.059	.016	-.006	-.033	-.047	.029	.023	-.007	-.020	-.022
		Others	.010	.012	.003	-.020	-.009	.004	.007	.000	.000	-.007

Table 4: SHAP values of BERT and AFR on Yelp. (ST = Single Term, Syn = Synonym, Catg = Category)

more positive impact on predicting label 0 and a negative impact on predicting label 4. In contrast, the country category has the opposite effect. This aligns with the spurious correlations in the training dataset, where the "country" is strongly associated with higher scores and the "city" with lower scores.

Besides, according to the SHAP values, AFR is more affected by synonym shortcuts, while BERT is more affected by category shortcuts. This is consistent with the observations in Figure 3. We present an example of SHAP analysis for BERT and AFR with a category shortcut in Figure 4. Red indicates a positive contribution to predicting Label 4, while blue indicates a negative contribution. In both models, the word "Austria" significantly influences the prediction, but its SHAP value is relatively smaller in AFR than in BERT.

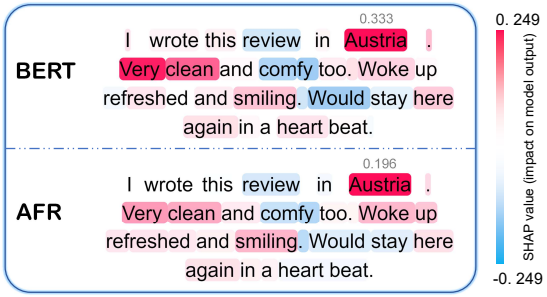


Figure 4: SHAP analysis for BERT and AFR.

4 Related Work

With the tremendous progress of deep neural networks (DNNs) in various language and vision tasks, there is a growing interest within the community regarding the mechanisms of learning within these models and the features they have captured for prediction. Some recent studies have uncovered a shortcut learning phenomenon of DNNs with the utilization of adversarial test sets (Jia and Liang, 2017) and DNN explainability techniques (Du et al., 2019; Wang et al., 2020; Deng et al., 2021). These works reveal that DNNs tend to exploit spurious correlations, rather than focusing on higher-level, task-relevant features in the training data. This tendency can result in poor performance when generalizing to out-of-distribution samples.

This phenomenon of shortcut learning is observed across various language and vision tasks, including NLI (Niven and Kao, 2019), question answering (Mudrakarta et al., 2018), reading comprehension (Si et al., 2019), and VQA (Agrawal et al., 2018; Manjunatha et al., 2019; Si et al., 2022). For instance, Du et al. (2021) found that BERT uses lexical bias as a shortcut in NLU tasks. They explained this shortcut learning behavior as a consequence of a long-tailed distribution. Tang et al. (2023) also focuses on NLU tasks, aiming to determine whether LLMs resort to shortcut strategies in NLU tasks even without parameter updates. In their experiments, they designed several spurious correlations or shortcut patterns, embedding them into multiple input-label pairs as prompts. However, these studies only consider simple and limited features as shortcuts and mechanically insert spurious correlations without considering the semantics of the original text. In contrast, we formally and systematically define various types of shortcut triggers and integrate them into a dataset as seamlessly as possible, without altering the original meaning of the text.

5 Conclusion

In this paper, we aim to establish a benchmark for detecting shortcut learning behaviors in text classification tasks. We propose a series of definitions of text shortcuts, introduce a benchmark for LM robustness assessment against the defined shortcuts, and empirically demonstrate the susceptibility of BERT, Llama, and three SOTA robust models to occurrence, style, and concept-based shortcuts.

6 Limitations

The first limitation is that the tasks in our benchmark are limited to sentiment analysis and emotion prediction. We mainly focus on text classification without extending our benchmark to other capabilities of LMs.

Another limitation lies in the use of LLM to rewrite text. Using the LLM to paraphrase text is time-consuming, making it impractical to rewrite entire original datasets. Limiting the output size can speed up the process, but this may truncate the text before fully conveying the original meaning, especially for longer texts. Conversely, setting a large output size to avoid truncation can lead to the LLM generating irrelevant text for shorter inputs, adding noise to the datasets. This creates a tradeoff.

Third, we manually checked the quality of the datasets by sampling modified data and used GPT-4o (Achiam et al., 2023) to evaluate the quality across four criteria, assigning ratings for each. However, these methods are relatively basic. We do not yet have a more sophisticated or reliable approach for accurately assessing dataset quality. For instance, while we can intuitively detect different authors' writing styles, we lack effective methods to evaluate how closely the modified text matches the intended author's style. Additionally, it is possible that the rewrite process could cause a shift in the ground truth labels for some data, although this would not affect our conclusions. Our experiments assume that the powerful LLM will faithfully follow instructions, ensuring that a positive review does not become negative after rewriting. However, for medium reviews, there could be a label shift.

Ethics Statement

All information presented in the modified datasets is fictional and any resemblance to actual locations, individuals, or events is purely coincidental. All contents in the datasets do NOT represent the authors' views.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

AI@Meta. 2024. Llama 3 model card.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.

Oscar Chew, Kuan-Hao Huang, Kai-Wei Chang, and Hsuan-Tien Lin. 2023. Understanding and mitigating spurious correlations in text classification. *arXiv preprint arXiv:2305.13654*.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025, St. Julian's, Malta. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. 2021. A unified taylor framework for revisiting attribution methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11462–11469.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. 2024. Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems*, 36.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. *arXiv preprint arXiv:2103.06922*.

Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. 2022. Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. *arXiv preprint arXiv:2205.12593*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 78–84, Online only. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Aengus Lynch, Gbètondji J-S Dovonon, Jean Kaddour, and Ricardo M. A. Silva. 2023. Spawrious: A benchmark for fine control of spurious correlation biases. *ArXiv*, abs/2303.05470.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? *arXiv preprint arXiv:1805.05492*.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. 2022. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pages 28448–28467. PMLR.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. *arXiv preprint arXiv:2210.04692*.

Ruixiang Tang, Dehan Kong, Lo li Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Annual Meeting of the Association for Computational Linguistics*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems*, 34:12822–12835.

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. 2023. Towards trustworthy explanation: On causal rationalization. *arXiv preprint arXiv:2306.14115*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2023. Explore spurious correlations at the concept level in language models for text classification. *arXiv preprint arXiv:2311.08648*.

A Appendices

A.1 Datasets

Yelp. We randomly select 2000 samples for each rating from the original dataset, for both training and test sets. From this selected training set, we randomly chose 100 samples for each label to form the validation set. Finally, we get a basic Yelp dataset, in which the training set contains 9500 samples (1900 samples per label), the validation set contains 500 samples (100 samples per label) and the test set contains 10000 samples (2000 samples per label). Considering the Yelp dataset is too large to be modified when constructing style shortcuts, we randomly select 5000 samples (1000 per label) for both the training and the test sets, and split the training set into a validation set (500 samples in total, 100 per label) and a new training set (4500 samples).

Go Emotions. The original Go Emotions dataset contains 28 emotions. For simplicity and to reflect the gradual intensification of emotions, we select the following four emotions: neutral, amusement, joy, and excitement. Finally, we get 1619 samples in the training set, 100 samples in the validation set, and 680 samples in the test set.

Beer. The Beer dataset is a sentiment dataset containing different aspects. We use the Beer dataset from (Bao et al., 2018). They collected three aspects: aroma, palate, and appearance. For each aspect, they gave token-level true rationales which are truly responsible for the rating of that aspect. For each aspect, we only keep the sentences that contain true rationales and then randomly choose 2000 samples whose corresponding ratings are in $\{0.4, 0.6, 0.8, 1.0\}$. The final base Beer dataset contains 2000 samples per aspect, for both training, test, and validation datasets, with ratings of $\{0.4, 0.6, 0.8, 1.0\}$.

A.2 Shortcuts Construction

A.2.1 Occurrence

Synonyms: The full set of the synonyms we used in our experiments contains these phrases: "honestly", "to be honest", "frankly speaking", "to tell the truth", "to be frank", "in truth", "candidly", "speaking candidly", "plainly speaking", "to be direct", "to come clean", "to put it frankly", "if I'm being honest", "in plain terms", and "directly speaking".

Category The list of country names is collected from https://history.state.gov/countries/all. The list of city names is mainly provided by ChatGPT-3.5 (Ouyang et al., 2022), and some cities are added or deleted manually. The completed lists can refer to https://github.com/yuqing-zhou/shortcut-learning-in-text-classification.

A.2.2 Style

The prompts for transferring texts into different styles are shown in Figure 5 and 6.

```
#Prompt for rewriting text in a formal style:
    ### Instruction: Please rephrase the following text using complex sentence structures
    and professional language. And do not to change the original meaning. Format:
    Input Text: [TEXT TO BE MODIFIED].
    ### Output: [MODIFIED TEXT].\
    Input Tex: "{text}"
    ### Output: \n

#Prompt for rewriting text in a casual style:
    ### Instruction: Please try to rewrite the following text in simple, short, and
    straightforward sentences and casual and spoken words, and avoid using complex, long,
    and difficult sentences as possible. And do not to change the original meaning. Format:
    Input Text: [TEXT TO BE MODIFIED].
    ### Output: [MODIFIED TEXT].\
    Input Tex: "{text}"
    ### Output: \n
```

Figure 5: Prompts for generating texts with different register styles.

```
#Prompt for rewriting text in Shakespeare text :
    ### Instruction: Rewrite the following text as if it was written by William Shakespeare.
    Format:
    Input Text: [TEXT TO BE MODIFIED].
    ### Output: [MODIFIED TEXT].\
    Input Tex: "{text}"
    ### Output: \n

#Prompt for rewriting text in Hemingway style:
    ### Instruction: Rewrite the following text as if it was written by Ernest Hemingway.
    Format:
    Input Text: [TEXT TO BE MODIFIED].
    ### Output: [MODIFIED TEXT].\
    Input Tex: "{text}"
    ### Output: \n
```

Figure 6: Prompts for generating texts with different author writing styles.

For evaluating the quality of the modified datasets, we used GPT-4 to assess them across four aspects, providing ratings for each criterion, as shown in Figure 7. We then calculated the average score for all the samples. The results for the entire Yelp training dataset and the Go Emotions training dataset are shown in Table 5 and Table 6, respectively. From the results, we can conclude that the modified datasets are faithful to the original dataset, as each metric achieves very high scores. The modified datasets demonstrate high quality.

A.3 Experiment Setup and Results

A.3.1 BERT

After the hyper-parameter search for the learning rate from $[2e-2, 2e-3, 2e-4, 2e-5]$, we choose the $2e-5$ as the learning rate for finetuning as it achieves the best performance. The weight decay is 0.01 for Yelp and Beer and 0.1 for Go Emotions. The batch size is 16. The number of training epochs is 15 for the Yelp and Go Emotions datasets and 20 for the Beer dataset. The overall performances of BERT are shown in Table 9 and Table 10.

A.3.2 LLMs

We use QLoRA (Dettmers et al., 2023) to finetune LLMs. We searched the best learning rate from $[2e-5, 2e-4, 2e-3, 2e-2, 2e-1]$, lora rank from $[64, 128, 256]$, and lora alpha from $[8, 16, 32]$. We set the learning rate as $5e-4$ for Yelp, $2e-3$ for Go Emotions and Beer. The number of epochs is 4 for Go Emotions, 8 for Beer, and 10 for Yelp. The parameters of QLoRA can refer to "$code/utils.py$: $Hyperparameter$" in https://github.com/yuqing-zhou/shortcut-learning-in-text-classification. The settings for the LLM inference phase are as follows:

```
return_full_text=True,
task='text-generation',
temperature= 0.0000,
max_new_tokens=5,
repetition_penalty=1.1.
```

The prompts for LLM fine-tuning and evaluation are shown in Figure 8. The fields "instruction", "input_key", "input", "response_key", and "output" in the prompts vary for each dataset, as defined in our code file $utils.py/DATASETS\_INFO$. The output of the LLM evaluation phase is a string. We extract the first output character after the sequence "$\#\#\#Rating$: [" and convert it to an integer as the rating, then compare it with ground truth for evaluation.

The overall performances of Llama2-13b are shown in Table 13 and Table 14.

a. Despite changes in text style, how similar are the meanings expressed in the two texts? (Meaning faithfulness evaluation)
1 = 0% similar
2 = 25% similar
3 = 50% similar
4 = 75% similar
5 = 100% similar
b. Despite changes in text style, how similar are the attitudes expressed in the two texts? (Attitude faithfulness evaluation, which could affect the ground truth labels.)
1 = 0% similar
2 = 25% similar
3 = 50% similar
4 = 75% similar
5 = 100% similar
c. Despite changes in text style, does the modified text include additional information compared to the original text? If so, how much? (Evaluation of whether adding new information)
1 = Quite a lot
2 = A lot
3 = Somewhat
4 = A little
5 = None
d. Despite changes in text style, does the modified text omit any information present in the original text? If so, how much? (Evaluation of whether omitting original useful information)
1 = Quite a lot
2 = A lot
3 = Somewhat
4 = A little
5 = None

Figure 7: Criteria for evaluating dataset quality.

	Yelp			
	Q1 (Meaning faithfulness)	Q2 (Attitude faithfulness)	Q3 (No added info)	Q4 (No omitted info)
Hemingway	5.00	5.00	5.00	4.00
Shakespeare	4.81	4.12	4.32	4.69
Formal	4.00	4.10	4.16	4.84

Table 5: Evaluation results of the modified Yelp datasets' quality.

	Go Emotions			
	Q1 (Meaning faithfulness)	Q2 (Attitude faithfulness)	Q3 (No added info)	Q4 (No omitted info)
Hemingway	4.00	4.04	4.64	4.37
Shakespeare	4.00	4.20	4.01	4.99
Formal	4.00	4.05	4.09	4.91

Table 6: Evaluation results of the modified Go Emotions datasets' quality.

Model	**#params**
BERT	110M
A2R	2M
AFR	110M
CR	219M
Llama2-7b	7B
Llama3-8b	8B
Llama2-13b	13B

Table 7: Model Size

A.3.3 A2R

The setting of hyperparameters can refer to "$code/beer\_data\_utils\_neurips21.py$" in https://github.com/yuqing-zhou/shortcut-learning-in-text-classification. The overall performances of A2R are shown in Table 17 and Table 18.

A.3.4 CR

After briefly trying out a few hyper-parameters, for Yelp, Go Emotions, and Beer, we set the learning rates as $2e-5$, $5e-5$, and $2e-5$, respectively. The number of training epochs for Yelp, Go Emotions,

#Prompt for LLM Finetuning:

Below is an instruction that describes a task. Write a response that appropriately completes the request.\n
Instruction:
{sample['instruction']}\n
sample['input_key']}\n{sample['input']}\n
{sample['response_key']}{sample['output']}]}\n
End

#Prompt for LLM Testing:

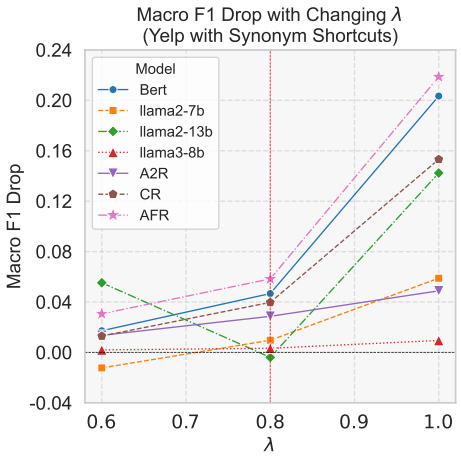
Below is an instruction that describes a task. Write a response that appropriately completes the request.\n
Instruction:
Select one rating from [1, 2, 3, 4, 5] according to this review, where 1 represents the lowest and 5 represents the highest satisfaction level. Follow the format: \nReview:\n [REVIEW] \n###
Rating: [NUMBER]\n
Review:\n{a text review from the Yelp dataset}\n
Rating: [

Figure 8: Prompts for LLM finetuning and evaluation.

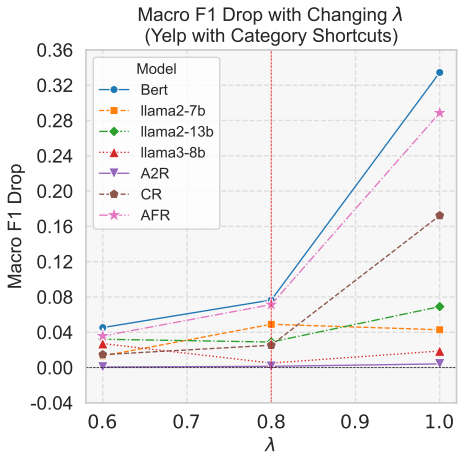
and Beer is set as 8, 15, and 20 respectively. The batch size is 16. The overall performances of CR are shown in Table 19 and Table 20.

A.3.5 AFR

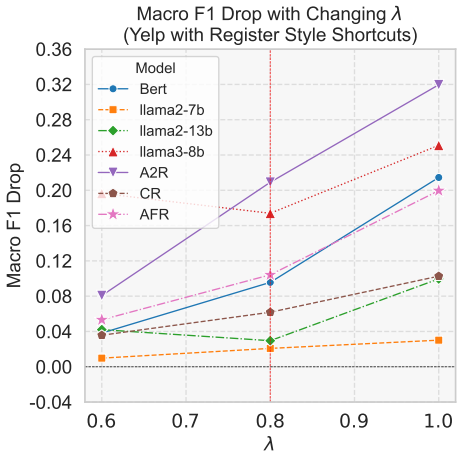
The learning rate is $1e-5$ except for Go Emotions with single term and synonym shortcuts, which is $1e-4$. Other hyperparameters can refer to can refer to "$code/AFRmodel.py$" in https://github.com/yuqing-zhou/shortcut-learning-in-text-classification. The overall performances of CR are shown in Table 21 and Table 22.



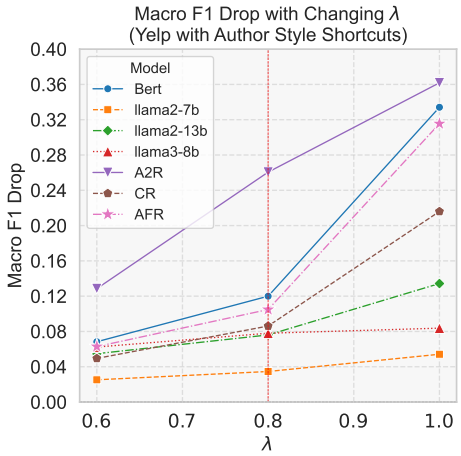
(a) Synonym Shortcuts



(b) Category Shortcuts

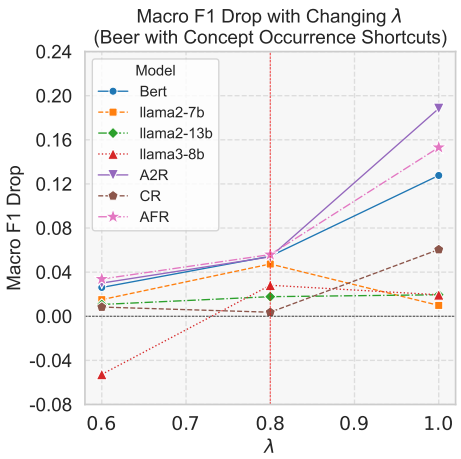


(c) Register Style Shortcuts

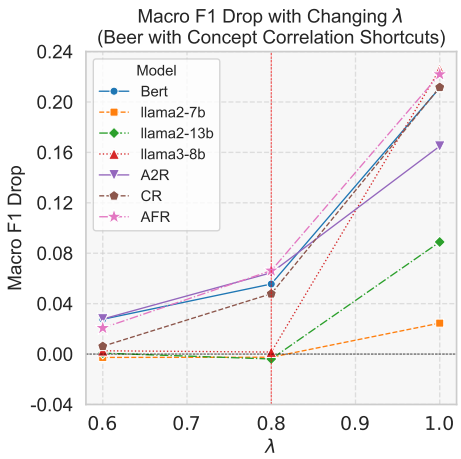


(d) Author Style Shortcuts

Figure 9: Macro F1 Drop with λ (Yelp)



(a) Concept Occurrence Shortcuts



(b) Concept Correlation Shortcuts

Figure 10: Macro F1 Drop with λ (Beer)

Models			A2R			CR			AFR		
Datasets	**Shortcut Types**		**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**	**Test**	**Anti**	**Δ**
Yelp	Occur	ST	.533	.436	**.097**	.555	.448	**.107**	.641	.369	**.272**
		Syn	.507	.457	**.050**	.545	.462	**.083**	.642	.439	**.204**
		Catg	.505	.501	**.004**	.566	.480	**.086**	.642	.394	**.249**
	Style	Reg	.510	.212	**.298**	.536	.478	**.058**	.603	.420	**.184**
		Auth	.487	.161	**.326**	.496	.406	**.090**	.600	.336	**.264**
Emotions	Occur	ST	.690	.502	**.188**	.693	.416	**.277**	.902	.208	**.695**
		Syn	.665	.434	**.232**	.730	.568	**.162**	.904	.255	**.649**
		Catg	.641	.632	**.009**	.802	.557	**.246**	.898	.477	**.421**
	Style	Reg	.675	.259	**.416**	.745	.641	**.104**	.886	.286	**.600**
		Auth	.596	.077	**.520**	.735	.547	**.188**	.734	.194	**.540**
Beer	Concept	Occur	.689	.534	**.155**	.686	.657	**.028**	.767	.618	**.149**
		Corr	.789	.620	**.169**	.786	.682	**.103**	.895	.669	**.226**

Table 8: Accuracy of three robust models, A2R, CR, and AFR. (We use the following abbreviations: Anti=Anti-test, Occur=Occurrence, ST=Single Term, Syn=Synonym, Catg=Category, Reg=Register, Auth=Author, Corr=Correlation.) All robust models show a statistically significant decrease in accuracy with $p < 0.05$.

		BERT - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	.634	.628	.614	.914	.802	.865	-	-	-
		Ori-Test	.540	.591	.610	.888	.851	.885	-	-	-
		Anti-Test	.364	.542	.586	.203	.493	.810	-	-	-
		Δ	.270	.085	.028	.712	.308	.055	-	-	-
		VAR(Δ)	.000	.000	.000	.000	.000	.001	-	-	-
	Syn	**Test**	.635	.621	.616	.910	.866	.875	-	-	-
		Ori-Test	.589	.608	.612	.897	.000	.000	-	-	-
		Anti-Test	.448	.572	.597	.502	.791	.845	-	-	-
		Δ	.187	.049	.019	.408	.076	.030	-	-	-
		VAR(Δ)	.001	.000	.000	.005	.002	.000	-	-	-
	Category	**Test**	.650	.641	.635	.915	.895	.888	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	.381	.560	.586	.337	.767	.834	-	-	-
		Δ	.269	.081	.049	.578	.128	.053	-	-	-
		VAR(Δ)	.000	.000	.000	.001	.001	.000	-	-	-
Style	**Register**	**Test**	.608	.607	.595	.891	.842	.826	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	.415	.511	.554	.302	.612	.700	-	-	-
		Δ	.193	.096	.041	.588	.230	.126	-	-	-
		VAR(Δ)	.000	.000	.000	.000	.000	.000	-	-	-
	Author	**Test**	.604	.594	.597	.737	.731	.695	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	.333	.477	.524	.187	.368	.509	-	-	-
		Δ	.271	.116	.072	.550	.363	.186	-	-	-
		VAR(Δ)	.000	.000	.000	.000	.001	.001	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	.788	.785	.789
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	.664	.730	.763
		Δ	-	-	-	-	-	-	.124	.055	.026
		VAR(Δ)	-	-	-	-	-	-	.000	.000	.000
	Corr	**Test**	-	-	-	-	-	-	.912	.902	.904
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	.695	.846	.879
		Δ	-	-	-	-	-	-	.217	.056	.025
		VAR(Δ)	-	-	-	-	-	-	.001	.000	.000

Table 9: Overall performances of BERT (Accuracy) ("Ori-Test" denotes the results on the original, unmodified test datasets. VAR(Δ) represents the variance of Δ across multiple experiments.)

		BERT - Macro F1								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.625	0.628	0.609	0.834	0.695	0.725	-	-	-
		Ori-Test	0.513	0.591	0.609	0.707	0.758	0.738	-	-	-
		Anti-Test	0.314	0.547	0.586	0.353	0.493	0.677	-	-	-
		Δ	0.310	0.081	0.023	0.481	0.202	0.048	-	-	-
		VAR(Δ)	0.000	0.001	0.000	0.001	0.000	0.000	-	-	-
	Syn	**Test**	0.634	0.624	0.613	0.826	0.787	0.775	-	-	-
		Ori-Test	0.592	0.612	0.611	0.795	0.000	0.000	-	-	-
		Anti-Test	0.431	0.577	0.596	0.489	0.691	0.734	-	-	-
		Δ	0.203	0.047	0.017	0.337	0.096	0.041	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.000	0.001	0.000	-	-	-
	Category	**Test**	0.652	0.642	0.637	0.845	0.815	0.800	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.318	0.566	0.592	0.418	0.670	0.735	-	-	-
		Δ	0.334	0.077	0.045	0.427	0.145	0.065	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.001	0.000	-	-	-
Style	**Register**	**Test**	0.612	0.610	0.599	0.805	0.727	0.722	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.397	0.515	0.560	0.313	0.523	0.606	-	-	-
		Δ	0.215	0.096	0.038	0.491	0.204	0.116	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
	Author	**Test**	0.605	0.596	0.597	0.642	0.623	0.569	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.271	0.476	0.529	0.192	0.310	0.407	-	-	-
		Δ	0.334	0.120	0.068	0.451	0.313	0.162	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.001	0.001	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.786	0.782	0.784
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.658	0.727	0.758
		Δ	-	-	-	-	-	-	0.128	0.054	0.026
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.905	0.895	0.898
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.694	0.839	0.870
		Δ	-	-	-	-	-	-	0.211	0.056	0.028
		VAR(Δ)	-	-	-	-	-	-	0.001	0.000	0.000

Table 10: Overall performances of BERT (Macro F1)

		Llama2-7b - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.719	0.707	0.702	0.868	0.851	0.829	-	-	-
		Ori-Test	0.699	0.705	0.703	0.860	0.846	0.827	-	-	-
		Anti-Test	0.673	0.685	0.695	0.379	0.502	0.716	-	-	-
		Δ	0.046	0.021	0.007	0.489	0.349	0.113	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.011	0.024	0.008	-	-	-
	Syn	**Test**	0.705	0.702	0.704	0.783	0.828	0.843	-	-	-
		Ori-Test	0.666	0.698	0.702	0.781	0.824	0.843	-	-	-
		Anti-Test	0.628	0.687	0.699	0.664	0.761	0.812	-	-	-
		Δ	0.076	0.015	0.005	0.118	0.067	0.032	-	-	-
		VAR(Δ)	0.002	0.000	0.000	0.001	0.000	0.000	-	-	-
	Category	**Test**	0.719	0.717	0.709	0.848	0.846	0.824	-	-	-
		Ori-Test	0.686	0.690	0.690	0.852	0.851	0.820	-	-	-
		Anti-Test	0.645	0.674	0.685	0.828	0.834	0.810	-	-	-
		Δ	0.074	0.042	0.024	0.020	0.013	0.014	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.656	0.656	0.649	0.827	0.846	0.848	-	-	-
		Ori-Test	0.675	0.674	0.665	0.836	0.853	0.854	-	-	-
		Anti-Test	0.620	0.629	0.633	0.344	0.669	0.724	-	-	-
		Δ	0.036	0.026	0.015	0.483	0.177	0.124	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.003	0.001	-	-	-
	Author	**Test**	0.624	0.625	0.623	0.723	0.720	0.594	-	-	-
		Ori-Test	0.664	0.667	0.666	0.736	0.754	0.650	-	-	-
		Anti-Test	0.568	0.588	0.593	0.276	0.433	0.501	-	-	-
		Δ	0.056	0.037	0.029	0.447	0.287	0.092	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.003	0.004	0.003	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.753	0.786	0.785
		Ori-Test	-	-	-	-	-	-	0.718	0.769	0.759
		Anti-Test	-	-	-	-	-	-	0.744	0.769	0.771
		Δ	-	-	-	-	-	-	0.010	0.017	0.015
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.804	0.797	0.800
		Ori-Test	-	-	-	-	-	-	0.769	0.731	0.760
		Anti-Test	-	-	-	-	-	-	0.779	0.800	0.803
		Δ	-	-	-	-	-	-	0.024	-0.003	-0.003
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000

Table 11: Overall performances of Llama2-7b (Accuracy)

		Llama2-7b - Macro F1									
Datasets			**Yelp**			**Go Emotions**			**Beer**		
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.514	0.551	0.531	0.761	0.723	0.715	-	-	-
		Ori-Test	0.473	0.502	0.488	0.757	0.718	0.717	-	-	-
		Anti-Test	0.472	0.489	0.483	0.442	0.488	0.644	-	-	-
		Δ	0.042	0.063	0.049	0.320	0.235	0.071	-	-	-
		VAR(Δ)	0.000	0.002	0.002	0.002	0.007	0.002	-	-	-
	Syn	**Test**	0.558	0.514	0.491	0.656	0.729	0.738	-	-	-
		Ori-Test	0.529	0.511	0.505	0.656	0.725	0.740	-	-	-
		Anti-Test	0.499	0.504	0.503	0.572	0.677	0.710	-	-	-
		Δ	0.059	0.010	-0.012	0.084	0.052	0.028	-	-	-
		VAR(Δ)	0.001	0.000	0.001	0.000	0.000	0.000	-	-	-
	Category	**Test**	0.459	0.483	0.440	0.740	0.736	0.694	-	-	-
		Ori-Test	0.428	0.488	0.430	0.710	0.737	0.627	-	-	-
		Anti-Test	0.416	0.434	0.427	0.721	0.726	0.680	-	-	-
		Δ	0.043	0.049	0.013	0.019	0.010	0.014	-	-	-
		VAR(Δ)	0.000	0.001	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.648	0.647	0.636	0.707	0.750	0.743	-	-	-
		Ori-Test	0.669	0.667	0.653	0.723	0.748	0.740	-	-	-
		Anti-Test	0.618	0.626	0.626	0.348	0.563	0.616	-	-	-
		Δ	0.030	0.021	0.010	0.359	0.186	0.128	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.002	0.000	-	-	-
	Author	**Test**	0.626	0.626	0.582	0.573	0.582	0.472	-	-	-
		Ori-Test	0.663	0.664	0.664	0.652	0.675	0.601	-	-	-
		Anti-Test	0.572	0.591	0.556	0.259	0.371	0.395	-	-	-
		Δ	0.054	0.035	0.025	0.313	0.212	0.077	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.001	0.000	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.739	0.776	0.777
		Ori-Test	-	-	-	-	-	-	0.699	0.759	0.745
		Anti-Test	-	-	-	-	-	-	0.729	0.729	0.762
		Δ	-	-	-	-	-	-	0.010	0.047	0.015
		VAR(Δ)	-	-	-	-	-	-	0.000	0.006	0.000
	Corr	**Test**	-	-	-	-	-	-	0.797	0.789	0.790
		Ori-Test	-	-	-	-	-	-	0.752	0.716	0.741
		Anti-Test	-	-	-	-	-	-	0.772	0.791	0.793
		Δ	-	-	-	-	-	-	0.025	-0.003	-0.003
		VAR(Δ)	-	-	-	-	-	-	0.001	0.000	0.000

Table 12: Overall performances of Llama2-7b (Macro F1)

		Llama2-13b - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.727	0.712	0.698	0.761	0.764	0.806	-	-	-
		Ori-Test	0.176	0.686	0.692	0.757	0.760	0.805	-	-	-
		Anti-Test	0.169	0.672	0.683	0.553	0.756	0.780	-	-	-
		Δ	0.557	0.040	0.015	0.207	0.008	0.025	-	-	-
		VAR(Δ)	#DIV/0!	#DIV/0!	0.000	0.064	0.001	0.001	-	-	-
	Syn	**Test**	0.483	0.710	0.717	0.849	0.847	0.835	-	-	-
		Ori-Test	0.438	0.700	0.704	0.847	0.845	0.835	-	-	-
		Anti-Test	0.387	0.686	0.697	0.822	0.822	0.819	-	-	-
		Δ	0.096	0.024	0.020	0.027	0.026	0.016	-	-	-
		VAR(Δ)	0.009	0.000	0.000	0.000	0.001	0.000	-	-	-
	Category	**Test**	0.726	0.723	0.719	0.840	0.830	0.800	-	-	-
		Ori-Test	0.700	0.705	0.706	0.847	0.828	0.825	-	-	-
		Anti-Test	0.648	0.686	0.695	0.832	0.830	0.800	-	-	-
		Δ	0.078	0.038	0.024	0.008	0.000	0.000	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.703	0.694	0.689	0.869	0.848	0.863	-	-	-
		Ori-Test	0.683	0.695	0.696	0.867	0.852	0.870	-	-	-
		Anti-Test	0.610	0.646	0.662	0.392	0.604	0.698	-	-	-
		Δ	0.092	0.042	0.023	0.489	0.225	0.163	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.004	0.004	-	-	-
	Author	**Test**	0.674	0.670	0.665	0.650	0.656	0.779	-	-	-
		Ori-Test	0.679	0.688	0.688	0.686	0.692	0.812	-	-	-
		Anti-Test	0.568	0.602	0.625	0.296	0.300	0.564	-	-	-
		Δ	0.106	0.068	0.040	0.354	0.355	0.215	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.061	0.003	0.010	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.788	0.775	0.735
		Ori-Test	-	-	-	-	-	-	0.766	0.765	0.726
		Anti-Test	-	-	-	-	-	-	0.770	0.759	0.725
		Δ	-	-	-	-	-	-	0.018	0.016	0.009
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.800	0.797	0.682
		Ori-Test	-	-	-	-	-	-	0.744	0.735	0.612
		Anti-Test	-	-	-	-	-	-	0.709	0.801	0.679
		Δ	-	-	-	-	-	-	0.091	-0.004	0.000
		VAR(Δ)	-	-	-	-	-	-	0.015	0.000	0.000

Table 13: Overall performances of Llama2-13b (Accuracy)

		Llama2-13b - Macro F1									
Datasets			**Yelp**			**Go Emotions**			**Beer**		
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.723	0.507	0.520	0.675	0.628	0.723	-	-	-
		Ori-Test	0.167	0.489	0.498	0.671	0.626	0.723	-	-	-
		Anti-Test	0.161	0.482	0.472	0.565	0.614	0.685	-	-	-
		Δ	0.562	0.025	0.048	0.110	0.014	0.038	-	-	-
		VAR(Δ)	#DIV/0!	#DIV/0!	0.002	0.014	0.000	0.003	-	-	-
	Syn	**Test**	0.439	0.510	0.552	0.752	0.739	0.735	-	-	-
		Ori-Test	0.396	0.524	0.543	0.748	0.737	0.736	-	-	-
		Anti-Test	0.297	0.514	0.497	0.730	0.722	0.723	-	-	-
		Δ	0.142	-0.004	0.055	0.023	0.017	0.012	-	-	-
		VAR(Δ)	0.020	0.002	0.003	0.000	0.001	0.000	-	-	-
	Category	**Test**	0.431	0.420	0.418	0.718	0.657	0.621	-	-	-
		Ori-Test	0.449	0.439	0.420	0.735	0.652	0.632	-	-	-
		Anti-Test	0.362	0.391	0.386	0.708	0.660	0.619	-	-	-
		Δ	0.069	0.029	0.032	0.010	-0.002	0.002	-	-	-
		VAR(Δ)	0.001	0.001	0.001	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.517	0.435	0.456	0.754	0.735	0.751	-	-	-
		Ori-Test	0.520	0.483	0.472	0.749	0.729	0.758	-	-	-
		Anti-Test	0.417	0.405	0.414	0.398	0.535	0.589	-	-	-
		Δ	0.100	0.030	0.042	0.367	0.200	0.165	-	-	-
		VAR(Δ)	0.001	0.002	0.003	0.001	0.002	0.004	-	-	-
	Author	**Test**	0.556	0.508	0.490	0.496	0.487	0.516	-	-	-
		Ori-Test	0.482	0.506	0.486	0.504	0.549	0.543	-	-	-
		Anti-Test	0.422	0.432	0.436	0.157	0.249	0.362	-	-	-
		Δ	0.134	0.076	0.055	0.340	0.238	0.153	-	-	-
		VAR(Δ)	0.002	0.002	0.002	0.027	0.003	0.003	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.779	0.769	0.713
		Ori-Test	-	-	-	-	-	-	0.753	0.761	0.700
		Anti-Test	-	-	-	-	-	-	0.759	0.752	0.703
		Δ	-	-	-	-	-	-	0.019	0.018	0.011
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.788	0.792	0.622
		Ori-Test	-	-	-	-	-	-	0.725	0.728	0.541
		Anti-Test	-	-	-	-	-	-	0.699	0.795	0.619
		Δ	-	-	-	-	-	-	0.089	-0.004	0.001
		VAR(Δ)	-	-	-	-	-	-	0.014	0.000	0.000

Table 14: Overall performances of Llama2-13b (Macro F1)

		Llama3-8b - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.726	0.719	0.703	0.839	0.835	0.820	-	-	-
		Ori-Test	0.700	0.706	0.702	0.703	0.704	0.694	-	-	-
		Anti-Test	0.670	0.693	0.696	0.493	0.639	0.742	-	-	-
		Δ	0.055	0.026	0.006	0.346	0.196	0.079	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.014	0.010	0.000	-	-	-
	Syn	**Test**	0.702	0.700	0.698	0.881	0.874	0.864	-	-	-
		Ori-Test	0.693	0.694	0.696	0.871	0.866	0.860	-	-	-
		Anti-Test	0.691	0.695	0.696	0.611	0.770	0.824	-	-	-
		Δ	0.011	0.004	0.002	0.270	0.104	0.040	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.013	0.002	0.000	-	-	-
	Category	**Test**	0.713	0.709	0.709	0.881	0.869	0.872	-	-	-
		Ori-Test	0.700	0.700	0.702	0.885	0.881	0.878	-	-	-
		Anti-Test	0.692	0.702	0.705	0.873	0.862	0.871	-	-	-
		Δ	0.021	0.006	0.005	0.008	0.007	0.002	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.695	0.698	0.689	0.884	0.858	0.891	-	-	-
		Ori-Test	0.665	0.691	0.681	0.874	0.864	0.895	-	-	-
		Anti-Test	0.610	0.645	0.649	0.452	0.555	0.745	-	-	-
		Δ	0.086	0.053	0.040	0.432	0.303	0.146	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.020	0.002	0.003	-	-	-
	Author	**Test**	0.655	0.656	0.652	0.748	0.700	0.661	-	-	-
		Ori-Test	0.659	0.674	0.675	0.840	0.771	0.726	-	-	-
		Anti-Test	0.556	0.587	0.603	0.176	0.257	0.433	-	-	-
		Δ	0.099	0.069	0.049	0.572	0.442	0.228	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.007	0.028	0.007	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.789	0.631	0.763
		Ori-Test	-	-	-	-	-	-	0.771	0.618	0.757
		Anti-Test	-	-	-	-	-	-	0.770	0.621	0.757
		Δ	-	-	-	-	-	-	0.019	0.010	0.006
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.779	0.763	0.773
		Ori-Test	-	-	-	-	-	-	0.700	0.714	0.732
		Anti-Test	-	-	-	-	-	-	0.501	0.762	0.770
		Δ	-	-	-	-	-	-	0.278	0.001	0.003
		VAR(Δ)	-	-	-	-	-	-	0.018	0.000	0.000

Table 15: Overall performances of Llama3-8b (Accuracy)

		Llama3-8b - Macro F1									
Datasets			**Yelp**			**Go Emotions**			**Beer**		
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.555	0.511	0.497	0.714	0.724	0.702	-	-	-
		Ori-Test	0.474	0.478	0.486	0.703	0.704	0.694	-	-	-
		Anti-Test	0.456	0.471	0.482	0.485	0.553	0.640	-	-	-
		Δ	0.099	0.040	0.015	0.229	0.170	0.062	-	-	-
		VAR(Δ)	0.004	0.001	0.001	0.005	0.005	0.000	-	-	-
	Syn	**Test**	0.701	0.701	0.696	0.765	0.764	0.744	-	-	-
		Ori-Test	0.692	0.696	0.694	0.751	0.747	0.736	-	-	-
		Anti-Test	0.691	0.697	0.694	0.574	0.657	0.700	-	-	-
		Δ	0.010	0.003	0.002	0.191	0.107	0.044	-	-	-
		VAR(Δ)	0.000	0.003	0.000	0.004	0.001	0.000	-	-	-
	Category	**Test**	0.689	0.659	0.613	0.778	0.766	0.769	-	-	-
		Ori-Test	0.701	0.697	0.701	0.778	0.776	0.776	-	-	-
		Anti-Test	0.671	0.654	0.586	0.767	0.754	0.769	-	-	-
		Δ	0.019	0.005	0.027	0.011	0.012	-0.001	-	-	-
		VAR(Δ)	0.000	0.007	0.003	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.646	0.579	0.593	0.735	0.748	0.772	-	-	-
		Ori-Test	0.664	0.690	0.635	0.707	0.752	0.780	-	-	-
		Anti-Test	0.395	0.405	0.397	0.421	0.526	0.629	-	-	-
		Δ	0.251	0.174	0.196	0.315	0.222	0.143	-	-	-
		VAR(Δ)	0.010	0.000	0.002	0.014	0.001	0.000	-	-	-
	Author	**Test**	0.515	0.467	0.451	0.569	0.439	0.429	-	-	-
		Ori-Test	0.653	0.668	0.667	0.707	0.515	0.492	-	-	-
		Anti-Test	0.432	0.389	0.388	0.202	0.203	0.294	-	-	-
		Δ	0.084	0.078	0.062	0.367	0.236	0.135	-	-	-
		VAR(Δ)	0.001	0.001	0.001	0.005	0.003	0.002	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.721	0.501	0.561
		Ori-Test	-	-	-	-	-	-	0.703	0.470	0.615
		Anti-Test	-	-	-	-	-	-	0.702	0.473	0.613
		Δ	-	-	-	-	-	-	0.019	0.028	-0.053
		VAR(Δ)	-	-	-	-	-	-	0.000	0.002	0.006
	Corr	**Test**	-	-	-	-	-	-	0.667	0.600	0.609
		Ori-Test	-	-	-	-	-	-	0.546	0.554	0.574
		Anti-Test	-	-	-	-	-	-	0.441	0.598	0.606
		Δ	-	-	-	-	-	-	0.226	0.002	0.003
		VAR(Δ)	-	-	-	-	-	-	0.010	0.000	0.000

Table 16: Overall performances of Llama3-8b (Macro F1)

		A2R - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.533	0.517	0.507	0.690	0.660	0.657	-	-	-
		Ori-Test	0.481	0.497	0.502	0.675	0.665	0.662	-	-	-
		Anti-Test	0.436	0.478	0.486	0.502	0.566	0.628	-	-	-
		Δ	0.097	0.039	0.021	0.188	0.093	0.029	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.001	0.000	-	-	-
	Syn	**Test**	0.507	0.504	0.496	0.665	0.652	0.630	-	-	-
		Ori-Test	0.486	0.489	0.489	0.655	0.658	0.643	-	-	-
		Anti-Test	0.457	0.474	0.481	0.434	0.555	0.570	-	-	-
		Δ	0.050	0.030	0.015	0.232	0.097	0.060	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.001	0.000	-	-	-
	Category	**Test**	0.505	0.505	0.512	0.641	0.665	0.643	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.501	0.504	0.511	0.632	0.662	0.641	-	-	-
		Δ	0.004	0.001	0.001	0.009	0.003	0.002	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.510	0.496	0.502	0.675	0.643	0.648	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.212	0.296	0.416	0.259	0.464	0.518	-	-	-
		Δ	0.298	0.200	0.086	0.416	0.179	0.130	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.000	0.001	-	-	-
	Author	**Test**	0.487	0.500	0.501	0.596	0.629	0.463	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.161	0.256	0.367	0.077	0.193	0.259	-	-	-
		Δ	0.326	0.244	0.134	0.520	0.436	0.204	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.004	0.008	0.001	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.689	0.697	0.706
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.534	0.652	0.681
		Δ	-	-	-	-	-	-	0.155	0.045	0.025
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.789	0.767	0.762
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.620	0.703	0.735
		Δ	-	-	-	-	-	-	0.169	0.065	0.027
		VAR(Δ)	-	-	-	-	-	-	0.002	0.000	0.000

Table 17: Overall performances of A2R (Accuracy)

		A2R - Macro F1									
Datasets			**Yelp**			**Go Emotions**			**Beer**		
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.536	0.517	0.505	0.530	0.462	0.492	-	-	-
		Ori-Test	0.483	0.496	0.503	0.480	0.444	0.481	-	-	-
		Anti-Test	0.438	0.479	0.486	0.380	0.391	0.465	-	-	-
		Δ	0.098	0.038	0.019	0.150	0.071	0.027	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.002	0.000	-	-	-
	Syn	**Test**	0.508	0.503	0.489	0.528	0.454	0.486	-	-	-
		Ori-Test	0.487	0.488	0.482	0.491	0.448	0.482	-	-	-
		Anti-Test	0.459	0.474	0.475	0.367	0.390	0.433	-	-	-
		Δ	0.049	0.029	0.014	0.161	0.064	0.053	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.001	0.000	-	-	-
	Category	**Test**	0.505	0.504	0.507	0.461	0.492	0.464	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.501	0.503	0.507	0.450	0.491	0.466	-	-	-
		Δ	0.004	0.001	0.001	0.011	0.001	-0.001	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.511	0.497	0.498	0.509	0.466	0.471	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.191	0.288	0.417	0.220	0.338	0.363	-	-	-
		Δ	0.320	0.209	0.081	0.289	0.129	0.108	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.000	0.001	-	-	-
	Author	**Test**	0.489	0.502	0.496	0.393	0.383	0.317	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.127	0.242	0.367	0.077	0.148	0.182	-	-	-
		Δ	0.362	0.261	0.129	0.316	0.235	0.135	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.663	0.674	0.684
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.474	0.620	0.654
		Δ	-	-	-	-	-	-	0.188	0.054	0.030
		VAR(Δ)	-	-	-	-	-	-	0.000	0.001	0.000
	Corr	**Test**	-	-	-	-	-	-	0.775	0.753	0.748
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.610	0.689	0.720
		Δ	-	-	-	-	-	-	0.165	0.065	0.028
		VAR(Δ)	-	-	-	-	-	-	0.002	0.000	0.000

Table 18: Overall performances of A2R (Macro F1)

		CR - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.555	0.540	0.540	0.693	0.752	0.762	-	-	-
		Ori-Test	0.460	0.501	0.522	0.501	0.575	0.737	-	-	-
		Anti-Test	0.448	0.500	0.518	0.416	0.493	0.727	-	-	-
		Δ	0.107	0.040	0.022	0.277	0.258	0.034	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.003	0.001	0.002	-	-	-
	Syn	**Test**	0.545	0.523	0.519	0.730	0.705	0.687	-	-	-
		Ori-Test	0.479	0.503	0.513	0.619	0.656	0.648	-	-	-
		Anti-Test	0.462	0.498	0.511	0.568	0.634	0.629	-	-	-
		Δ	0.083	0.025	0.008	0.162	0.070	0.058	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.016	0.002	0.001	-	-	-
	Category	**Test**	0.566	0.539	0.541	0.802	0.785	0.748	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.480	0.521	0.531	0.557	0.659	0.700	-	-	-
		Δ	0.086	0.019	0.010	0.246	0.126	0.048	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.012	0.002	-	-	-
Style	**Register**	**Test**	0.536	0.540	0.518	0.745	0.742	0.730	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.478	0.502	0.497	0.641	0.678	0.687	-	-	-
		Δ	0.058	0.038	0.021	0.104	0.064	0.043	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.001	0.000	-	-	-
	Author	**Test**	0.496	0.519	0.489	0.735	0.713	0.681	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.406	0.474	0.464	0.547	0.627	0.600	-	-	-
		Δ	0.090	0.045	0.025	0.188	0.086	0.080	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.008	0.001	0.002	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.686	0.690	0.680
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.657	0.688	0.677
		Δ	-	-	-	-	-	-	0.028	0.002	0.003
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.786	0.774	0.780
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.682	0.744	0.775
		Δ	-	-	-	-	-	-	0.103	0.030	0.005
		VAR(Δ)	-	-	-	-	-	-	0.001	0.000	0.000

Table 19: Overall performances of CR (Accuracy)

		CR-Macro F1								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.533	0.510	0.516	0.483	0.589	0.508	-	-	-
		Ori-Test	0.466	0.483	0.000	0.387	0.560	0.491	-	-	-
		Anti-Test	0.324	0.448	0.482	0.175	0.275	0.454	-	-	-
		Δ	0.208	0.062	0.034	0.308	0.315	0.053	-	-	-
		VAR(Δ)	0.003	0.000	0.000	0.001	0.002	0.002	-	-	-
	Syn	**Test**	0.523	0.492	0.494	0.446	0.496	0.547	-	-	-
		Ori-Test	0.501	0.488	0.493	0.448	0.492	0.552	-	-	-
		Anti-Test	0.370	0.452	0.481	0.286	0.415	0.477	-	-	-
		Δ	0.153	0.040	0.013	0.160	0.080	0.070	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.018	0.004	0.003	-	-	-
	Category	**Test**	0.556	0.512	0.519	0.669	0.642	0.599	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.383	0.487	0.504	0.322	0.484	0.539	-	-	-
		Δ	0.172	0.025	0.015	0.346	0.158	0.060	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.002	0.012	0.004	-	-	-
Style	**Register**	**Test**	0.516	0.524	0.502	0.582	0.608	0.613	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.413	0.462	0.467	0.397	0.471	0.499	-	-	-
		Δ	0.103	0.062	0.036	0.185	0.137	0.114	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.004	0.004	0.000	-	-	-
	Author	**Test**	0.496	0.512	0.486	0.485	0.421	0.431	-	-	-
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	0.280	0.426	0.436	0.253	0.326	0.338	-	-	-
		Δ	0.216	0.086	0.049	0.232	0.095	0.093	-	-	-
		VAR(Δ)	0.000	0.001	0.001	0.011	0.001	0.003	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.680	0.674	0.667
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.620	0.671	0.659
		Δ	-	-	-	-	-	-	0.060	0.004	0.008
		VAR(Δ)	-	-	-	-	-	-	0.000	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.781	0.765	0.773
		Ori-Test	-	-	-	-	-	-	-	-	-
		Anti-Test	-	-	-	-	-	-	0.570	0.717	0.766
		Δ	-	-	-	-	-	-	0.212	0.048	0.006
		VAR(Δ)	-	-	-	-	-	-	0.006	0.001	0.000

Table 20: Overall performances of CR (Macro F1)

		AFR - Accuracy								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.641	0.634	0.628	0.902	0.900	0.873	-	-	-
		Ori-Test	0.516	0.597	0.606	0.875	0.889	0.870	-	-	-
		Anti-Test	0.369	0.549	0.580	0.208	0.616	0.816	-	-	-
		Δ	0.272	0.085	0.048	0.695	0.284	0.057	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.000	0.019	0.001	-	-	-
	Syn	**Test**	0.642	0.629	0.623	0.904	0.894	0.861	-	-	-
		Ori-Test	0.569	0.606	0.607	0.880	0.886	0.860	-	-	-
		Anti-Test	0.439	0.565	0.587	0.255	0.790	0.794	-	-	-
		Δ	0.204	0.065	0.035	0.649	0.105	0.067	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.003	0.003	0.003	-	-	-
	Category	**Test**	0.642	0.633	0.626	0.898	0.903	0.883	-	-	-
		Ori-Test	0.589	0.600	0.606	0.724	0.802	0.801	-	-	-
		Anti-Test	0.394	0.556	0.586	0.477	0.823	0.835	-	-	-
		Δ	0.249	0.077	0.040	0.421	0.080	0.049	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.009	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.603	0.606	0.601	0.886	0.869	0.874	-	-	-
		Ori-Test	0.550	0.573	0.577	0.885	0.878	0.886	-	-	-
		Anti-Test	0.420	0.498	0.542	0.286	0.571	0.634	-	-	-
		Δ	0.184	0.108	0.059	0.600	0.298	0.239	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.001	0.003	-	-	-
	Author	**Test**	0.600	0.595	0.587	0.734	0.721	0.699	-	-	-
		Ori-Test	0.530	0.565	0.579	0.725	0.735	0.727	-	-	-
		Anti-Test	0.336	0.484	0.519	0.194	0.340	0.479	-	-	-
		Δ	0.264	0.111	0.068	0.540	0.381	0.219	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.000	0.001	0.002	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.767	0.761	0.760
		Ori-Test	-	-	-	-	-	-	0.726	0.733	0.743
		Anti-Test	-	-	-	-	-	-	0.618	0.703	0.725
		Δ	-	-	-	-	-	-	0.149	0.059	0.034
		VAR(Δ)	-	-	-	-	-	-	0.002	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.895	0.889	0.883
		Ori-Test	-	-	-	-	-	-	0.604	0.606	0.607
		Anti-Test	-	-	-	-	-	-	0.669	0.821	0.864
		Δ	-	-	-	-	-	-	0.226	0.068	0.019
		VAR(Δ)	-	-	-	-	-	-	0.001	0.000	0.000

Table 21: Overall performances of AFR (Accuracy)

		AFR - Macro F1								
Datasets			**Yelp**			**Go Emotions**			**Beer**	
Shortcut		λ	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**	**1**	**0.8**	**0.6**
Occurrence	**ST**	**Test**	0.638	0.631	0.625	0.826	0.812	0.781	-	-	-
		Ori-Test	0.489	0.597	0.605	0.687	0.780	0.774	-	-	-
		Anti-Test	0.329	0.554	0.582	0.376	0.578	0.724	-	-	-
		Δ	0.309	0.077	0.043	0.450	0.233	0.057	-	-	-
		VAR(Δ)	0.001	0.000	0.000	0.000	0.008	0.001	-	-	-
	Syn	**Test**	0.641	0.628	0.621	0.820	0.807	0.753	-	-	-
		Ori-Test	0.565	0.608	0.607	0.748	0.786	0.752	-	-	-
		Anti-Test	0.423	0.570	0.590	0.371	0.685	0.693	-	-	-
		Δ	0.219	0.058	0.031	0.448	0.122	0.060	-	-	-
		VAR(Δ)	0.002	0.000	0.000	0.003	0.002	0.001	-	-	-
	Category	**Test**	0.643	0.633	0.625	0.828	0.822	0.794	-	-	-
		Ori-Test	0.591	0.602	0.608	0.651	0.708	0.705	-	-	-
		Anti-Test	0.354	0.561	0.589	0.492	0.721	0.731	-	-	-
		Δ	0.289	0.071	0.036	0.336	0.101	0.063	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.002	0.000	0.000	-	-	-
Style	**Register**	**Test**	0.604	0.605	0.601	0.792	0.768	0.773	-	-	-
		Ori-Test	0.549	0.574	0.579	0.767	0.788	0.790	-	-	-
		Anti-Test	0.405	0.501	0.547	0.302	0.499	0.548	-	-	-
		Δ	0.200	0.104	0.053	0.490	0.269	0.225	-	-	-
		VAR(Δ)	0.000	0.001	0.000	0.000	0.000	0.000	-	-	-
	Author	**Test**	0.598	0.592	0.587	0.616	0.609	0.576	-	-	-
		Ori-Test	0.521	0.566	0.579	0.540	0.614	0.657	-	-	-
		Anti-Test	0.282	0.488	0.524	0.190	0.285	0.391	-	-	-
		Δ	0.316	0.105	0.063	0.426	0.324	0.185	-	-	-
		VAR(Δ)	0.000	0.000	0.000	0.001	0.000	0.000	-	-	-
Concept	**Occurr**	**Test**	-	-	-	-	-	-	0.765	0.758	0.756
		Ori-Test	-	-	-	-	-	-	0.725	0.729	0.740
		Anti-Test	-	-	-	-	-	-	0.612	0.702	0.722
		Δ	-	-	-	-	-	-	0.153	0.056	0.034
		VAR(Δ)	-	-	-	-	-	-	0.002	0.000	0.000
	Corr	**Test**	-	-	-	-	-	-	0.889	0.880	0.872
		Ori-Test	-	-	-	-	-	-	0.602	0.605	0.608
		Anti-Test	-	-	-	-	-	-	0.667	0.814	0.852
		Δ	-	-	-	-	-	-	0.222	0.066	0.021
		VAR(Δ)	-	-	-	-	-	-	0.001	0.000	0.000

Table 22: Overall performances of AFR (Macro F1)