

# Evaluating Robustness of Large Language Models to Algospeak

Stanford CS224N Custom Project

**Thet Htar Thin Zar**  
Department of Symbolic System  
Stanford University  
ellathtz@stanford.edu

**Hnin Yupar Mon**  
Department of Computer Science  
Stanford University  
yupar26@stanford.edu

## Abstract

Community-specific slang and Algospeak are difficult for language models to interpret as their meanings are deeply grounded in culture, rapidly evolving, and frequently indirect. We study this as a sequence-to-sequence rewriting task where we evaluate robustness under out-of-domain (OOD) shifts. Using BART-base and T5-small, we explore how OOD augmentation, prompt exploration, and tuning hyperparameters affects robustness across Chinese, Korean, Japanese, and Algospeak datasets. We found that BART-base performed best overall, showing strong robustness on held-out Chinese and Korean test sets as well as unseen Japanese and Algospeak datasets. Furthermore, we also found that OOD data’s usefulness does not depend simply on matching the source community as much as the linguistic properties of the added examples.

## 1 Key Information to include

- Mentor: Arpandeeep Khatua
- External Collaborators (if you have any): No
- Sharing project: No
- Late Days (if you are using any):

| Team member                      | Late days remaining |
|----------------------------------|---------------------|
| Hnin Yupar Mon                   | 2                   |
| Thet Htar Thin Zar               | 0                   |
| Total pooled                     | 2                   |
| Team size                        | 2                   |
| Late days applied for the report | 1                   |

## 2 Introduction

Language is not static. Across communities, people continuously develop new expressions, repurpose existing words, and develop coded language to communicate meaning indirectly whether to foster in-group identity, evade content moderation, or simply reflect the cultural context they inhabit. Understanding slang and algospeak are not merely a technical challenge; it is also a matter of inclusivity. As coded language continues to evolve rapidly, the inability of NLP systems to keep pace creates barriers to mutual comprehension.

This motivates us to frame our project as a robustness challenge. Rather than asking whether a model can learn a fixed inventory of slang terms, we ask: to what extent can fine-tuning models mitigate performance degradation under Out-Of-Distribution (OOD) shifts when models are evaluated on previously unseen terms, originating from variations across community and culture? New slang

emerges faster than any static dataset can capture, and a model that merely overfits to known expressions will fail precisely when it is needed most. OOD generalization remains an open challenge for language models across a broad range of NLP tasks. Our work instantiates this challenge in a new and practically motivated setting, where the source of distributional shift is not domain change in the traditional sense, but the cultural and community-specific nature of coded language itself.

We construct a dataset of 1,012 examples spanning an In-Domain (ID) subset of common internet slang and an OOD subset of community-specific slang drawn from Chinese, Korean, and Japanese, as well as broad general algospeak. We frame the task as given an input sentence containing slang and algospeak terms, the model produces a plain English rewrite that preserves the original meaning. We fine-tune T5-small and BART-base under multiple training conditions that vary training data composition, prompt formulation, and hyperparameters settings, and evaluate performance using the chrF metric against a zero-shot baseline. Our goal is to systematically assess how different fine-tuning choices affect a model’s ability to generalize to culturally and community-specific terms it has not encountered during fine-tuning.

### 3 Related Work

A key challenge in processing slang is that meaning is not uniform; instead, it varies systematically across communities and cultures. Sun and Xu [1] offer a principled computational analysis of this phenomenon, showing that the meaning of a slang term can diverge significantly across regional communities. Their work demonstrates that prior NLP approaches to slang have largely assumed homogeneity of meaning across user groups—an assumption that fails precisely in the context of community-specific slang terms. This cross-cultural fragility of slang processing is further evidenced by Wuraola et al. [2], who evaluate leading LLMs on the task of paraphrasing slang in climate-related tweets from Nigeria and the UK. They find that all models exhibit meaningful limitations in coverage and accuracy for non-Anglocentric slang, with performance degrading substantially for Nigerian English compared to British English. Our dataset construction directly addresses this gap by collecting slangs from distinct cultural communities—Chinese, Korean, and Japanese—and treating cross-community variation not as noise, but as a core property of the phenomenon under study.

A central concern in our work is how models generalize when the test distribution differs from training. Yuan et al.[3] systematically evaluate OOD robustness across a range of NLP tasks, finding that even fine-tuned pre-trained language models exhibit significant performance degradation on out-of-distribution examples, with smaller models particularly affected. Our work situates community specific slang and algospeak variation as a specific and practically motivated instance of this broader challenge: the distributional gap arises not from domain shift in the traditional sense, but from the cultural and community-specific nature of coded language that models have not encountered during training.

## 4 Approach

### 4.1 Task Formulation

We frame our task as a sequence-to-sequence (seq2seq) rewriting. Given an input sentence containing slang and algospeak terms, the model is fine-tuned to produce a plain English rewrite that preserves the original meaning while replacing the terms with their standard equivalents. This formulation allows us to leverage pre-trained encoder-decoder architectures and evaluate robustness under distributional shift by testing on OOD examples that were not seen during fine-tuning.

### 4.2 Model Architecture

We adopt two pre-trained encoder-decoder Transformer models: T5-small and BART-base. Both architectures are well-suited for seq2seq generation tasks, though they differ in their pre-training objectives. This architectural difference motivates our use of both models, as they may exhibit different robustness behaviors under OOD shifts.

### 4.3 Baseline

Our baseline is each pre-trained model evaluated in a zero-shot setting without any fine-tuning on our dataset. This establishes the performance floor attributable to the models’ pre-trained knowledge alone and provides a reference point against which the benefit of fine-tuning under each training condition can be measured.

### 4.4 Fine-tuning Conditions

Both models are fine-tuned on our constructed dataset using cross-entropy loss with teacher forcing. To study robustness under OOD shifts, we conduct experiments across multiple controlled fine-tuning conditions. Specifically, we vary:

1. Fine-tuning data composition: All fine-tuning configurations include the full ID dataset as a fixed baseline while varying the proportion and diversity of OOD examples added to the fine-tuning mix; examining how different levels of OOD exposure during fine-tuning effect generalization to unseen slang and algospeak terms.
2. Prompt Formulation: We experiment with four prompts variants that differ in what auxiliary information is appended to a shared task instruction. The *default* variant provides both the explicit slang term and its definition, giving the model maximum lexical context. The *no\_definition* variant omits the definitions, while *no\_slang* without the explicit term. Finally, *no\_slang\_and\_definition* provides neither field, requiring the model to infer the rewrite from the input sentence alone. This design allows us to isolate the individual contribution of each auxiliary field to rewriting quality and robustness.
3. Hyperparameters Settings: We vary learning rate, batch size, beam size, and number of training epochs.

## 5 Experiments

### 5.1 Data

We manually construct our dataset as no existing resource sufficiently captures the diversity and contextual nature required for this project. The dataset is split into two subsets: In-Domain (ID) and Out-of-Domain (OOD).

The ID subset contains common internet slang curated from three sources: *Urban Dictionary Words and Definitions* [4], *Gen Z words and Phrases Dataset* [5], and *genz\_slang\_dataset* [6]. To improve label quality and reduce ambiguity, we filter out terms whose meanings are not substantively redefined in Urban Dictionary [7]—that is, terms that do not exhibit a clear slang-specific sense beyond standard dictionary usage, or terms that are no longer used in modern contexts.

The OOD subset comprises community-specific slang drawn from three cultural communities: Chinese, Korean, and Japanese, as well as a collection of broad general algospeak. These communities were selected based on our familiarity with their respective cultural and linguistic contexts, which enabled more reliable term curation and interpretation. Terms were collected through manual data gathering across multiple sources. See Table 1 for details.

| Category | Sources                                      |
|----------|----------------------------------------------|
| Chinese  | Reddit, Various light novel literature sites |
| Korean   | Reddit, , Kpop Vocabulary, Kpop Dictionary   |
| Japanese | Reddit, , Anime Terms                        |
| General  | Wikipedia, AlgoSpeak Dictionary              |

Table 1: Data sources by OOD category.

For each term in both subsets, we collect three fields: a definition, an example usage sentence, and a plain English rewrite of the slang-containing sentence. For the ID subset, example sentences are drawn directly from Urban Dictionary [7] to preserve naturalistic usage. For the OOD subset, example sentences are manually authored by us to better reflect realistic contextual usage, given the

absence of readily available naturalistic examples for community-specific and culturally grounded algospeak. Plain English rewrites are manually authored across both subsets.

The final dataset comprises 520 ID and 492 OOD examples. Within the OOD subset, examples are distributed across four categories: 144 Chinese, 134 Korean, 102 Japanese, and 102 general algospeak terms. The ID subset is used entirely for training. Within the OOD subset, Japanese and general algospeak terms are reserved entirely for testing, serving as the strictly held-out evaluation set for measuring generalization to unseen terms. For the Chinese and Korean categories, half of the examples are used for validation, and the remaining half are held out for testing. In fine-tuning conditions that incorporate OOD data, a small number of Chinese and Korean examples are additionally introduced into training. All datasets contain sample outputs. Example input-output pairs are provided in Section A.1.

## 5.2 Evaluation method

We evaluate model outputs using chrF, which computes F-score over character n-gram matches between the model output and the reference rewrite. We prefer chrF over BLEU for this task as it is more robust to paraphrasing and lexical variation—both of which are expected in Algospeak rewriting, where multiple valid plain English rewrites may exist for a given Algospeak input sentence. As no standardized benchmarks exist for this task, we establish our reference benchmark using the chrF scores obtained from our zero-shot baseline models. Fine-tuned model variants are then evaluated against this baseline to measure the extent to which fine-tuning under each condition improves performance. To complement this quantitative evaluation, we also conduct a manual error analysis on model outputs to identify systematic failure patterns. Specifically, we examine cases where the model produces incorrect or inadequate rewrites with attention to recurring error types. While this analysis is inherently subjective, it serves as a necessary complement to chrF, which cannot capture the semantic adequacy or cultural appropriateness of the rewritten output.

## 5.3 Experimental details

Our experiments were implemented in PyTorch using the framework provided by Hugging Face Transformers, specifically the `transformers`, `datasets`, and `evaluate` libraries, on two pretrained sequence-to-sequence models: BART-base and T5-small. We designed the experiment to evaluate robustness to out-of-domain slang. Our experiments are divided into four different phases to study how the models’ performances change under increasing levels of OOD exposure, prompt exploration, and hyperparameter tuning, with the overall goal of measuring robustness to unseen slang communities.

We utilized 3 epochs, a learning rate of  $5 \times 10^{-5}$ , batch sizes of 4 for training and evaluation, a generation batch size of 8, and maximum input and target lengths of 128 for both models as our starting point, along with a prompt that included the sentence, slang term, and definition. T5-small also used a beam size of 4 by default.

The four phases were as follows:

1. **Phase 1:** An ID-only baseline with 1 training run per model, where each run was evaluated on both of the Chinese and Korean OOD validation sets.
2. **Phase 2:** Small OOD augmentation with 6 Chinese or Korean examples; this resulted in 2 training runs per model, and each run was evaluated on both Chinese and Korean validation sets, with the augmented examples removed from the matched validation sets.
3. **Phase 3:** Larger OOD augmentation and configuration search.
  - (a) OOD augmentation with 10 Chinese, 10 Korean, 20 Chinese, or 20 Korean examples; this resulted in 4 training runs per model, and each run was evaluated on both Chinese and Korean validation sets, with the augmented examples removed from the matched validation set.
  - (b) Prompt exploration using structured prompt variants where slang, definition, or both slang and definition were removed; each was trained and evaluated on the same setup as Phase 1, resulting in 3 runs per model.
  - (c) The best prompt among Phase 1 and Phase 3(b) was rerun on all the OOD augmentation and evaluation settings from Phase 3(a), resulting in 4 runs per model.

- (d) The best dataset and prompt among Phase 3(c) were rerun on 2 different epoch settings, 2 different learning rates, and 2 different beam sizes in that order for T5-small—the best hyperparameter for each was compounded into the next hyperparameter, performing sequential hyperparameter selection. Similarly, BART-base reran 3 different epochs and 3 different learning rates. This resulted in 6 f per model.
4. **Phase 4:** The best model, training dataset, prompt, and hyperparameters were finally evaluated on unseen Chinese and Korean test sets as well as Japanese and Algospeak datasets, resulting in 4 runs per model.

## 5.4 Results

| Model     | Validation Set | Before fine-tuning on ID | After fine-tuning on ID | chrF Difference |
|-----------|----------------|--------------------------|-------------------------|-----------------|
| BART-base | Chinese        | 17.23                    | 55.45                   | +38.22          |
| BART-base | Korean         | 17.56                    | 57.84                   | +40.28          |
| T5-small  | Chinese        | 14.42                    | 51.10                   | +36.67          |
| T5-small  | Korean         | 15.85                    | 52.46                   | +36.61          |

Table 2: Baseline Comparison

Both models exhibited significant improvement after training on ID, but BART-base outperformed T5-small on both Chinese and Korean validation sets. The results were what we expected. We think that this is because BART-base is specifically pre-trained for text reconstruction [8]. These results suggest that even without OOD augmentation, the models still acquired some degree of robustness to OOD slang. This indicates that BART-base has a more promising model for later exploration, although we don’t want to rule out T5-small just from this.

| Training Setup (BART-base) | Chinese Validation Set | Korean Validation Set |
|----------------------------|------------------------|-----------------------|
| ID only [baseline]         | 55.45                  | 57.84                 |
| ID + 6 Chinese             | 55.13                  | 59.28                 |
| ID + 6 Korean              | 56.04                  | 59.72                 |
| ID + 10 Chinese            | 54.92                  | 60.03                 |
| ID + 10 Korean             | 56.24                  | 61.90                 |
| ID + 20 Chinese            | 51.71                  | 60.42                 |
| ID + 20 Korean             | 57.03                  | 62.44                 |

Table 3: Small OOD Augmentation

Adding more OOD data generally improved robustness over the ID-only baseline, but this was not the pattern we expected. We expected Chinese augmentation to improve Chinese validation, but the strongest improvement came from Korean augmentation, which improved performance on both Chinese and Korean validation sets. We think this is because the Korean OOD examples provided more transferable cues for interpretation than the Chinese examples. These results suggest that although OOD augmentation can improve robustness, the type of data added to the dataset is more important than simply matching the communities the data are from.

| Configuration (BART-base)                                        | Chinese Validation Set | Korean Validation Set |
|------------------------------------------------------------------|------------------------|-----------------------|
| Default prompt (sentence, slang, definition), ID only [baseline] | 55.45                  | 57.84                 |
| No slang, ID only                                                | 54.02                  | 58.26                 |
| No definition, ID only                                           | 52.66                  | 53.88                 |
| No slang, no definition, ID only                                 | 51.82                  | 54.38                 |
| Best prompt on ID + 10 Chinese OOD                               | 52.64                  | 58.26                 |
| Best prompt on ID + 10 Korean OOD                                | 54.02                  | 58.06                 |
| Best Final Tuned Configuration                                   | 56.42                  | 58.93                 |

Table 4: Prompting

The structure of the prompt noticeably affected robustness to OOD slang, but this was not the pattern we expected. We expected the default prompt with sentence, slang, and definition to perform best. However, removing the slang term achieved the best performance. We think this is because the slang term may have distracted the model from the contextual sentence and definition. These results suggest that prompt design matters and that simpler prompts may interpret better, resulting in a better robustness.

| Hyperparameter Tuning (BART-base)<br>on best prompt & best dataset | Chinese Validation Set | Korean Validation Set |
|--------------------------------------------------------------------|------------------------|-----------------------|
| 3 Epochs [Original epoch]                                          | 56.42                  | 58.93                 |
| 4 Epochs                                                           | 56.31                  | 58.55                 |
| 5 Epochs                                                           | 55.51                  | 59.69                 |
| Learning Rate 1e-4 on best Epoch                                   | 53.76                  | 57.36                 |
| Learning Rate 5e-5 on best Epoch<br>[Original LR]                  | 56.42                  | 58.93                 |
| Learning Rate 7e-5 on best Epoch                                   | 54.98                  | 59.81                 |

Table 5: Hyperparameter Tuning (BART-base) on best prompt & best dataset

Hyperparameter fine-tuning through sequential hyperparameter selection produced some mixed effects on robustness where some scores increased on one validation set while decreased on the other. However, we expected fine-tuning to either increase or decrease scores on both datasets in a consistent manner since we used the best configuration from the previous steps. These results suggest that hyperparameter fine-tuning mattered less than the dataset collection choices and prompt design in affecting robustness, and that our default configuration already provided a strong foundation.

| Final Selection                                                   | Chinese Test Set | Korean Test Set | Japanese Unseen Set | Algospeak Unseen Set |
|-------------------------------------------------------------------|------------------|-----------------|---------------------|----------------------|
| No-slang prompt, 3 Epochs, Learning Rate 5e-5, ID + 10 Korean OOD | 58.10            | 56.77           | 75.04               | 81.19                |

Table 6: Final Held-Out & Unseen Evaluation

The final selected model showed strong robustness on both held-out Chinese and Korean test sets, and interpreted a lot more than expected on the unseen Japanese and Algospeak sets. We expected performance to be poorer for both unseen communities compared to Chinese and Korean test sets, but the unseen sets performed much better. We think this is because the unseen datasets may contain examples that have better contextual cues as well as less obfuscations. These results suggest that interpretation and robustness may depend a lot on how much contextual hints the input may contain.

## 6 Analysis

We manually inspected all the model outputs across the Chinese, Korean, Japanese and Algospeak datasets, taking into account the original slang and definitions, to better understand the model’s robustness behavior beyond chrF scores:

- **(Chinese):** Chinese slang was less robustly handled because many of the expressions had literal meanings, such as “releasing pigeons.” The model also struggled more on indirect verb-like slang, such as “touching fish,” than more on more direct noun-like slang, such as “central air-conditioning.” When the same slang appeared in more than one sentence, stronger contexts usually led to better interpretation and performance—suggesting robustness depended heavily on contextual support.
- **(Korean):** For Korean slang, robustness often depended on how the sentence was structured. Terms such as “Blink” and “Black Ocean” were interpreted correctly when the surrounding words gave better context. This aligns with the stronger quantitative robustness from Korean augmentation. Similarly to Chinese, it was less robustly handled on expressions that had literal meanings.

- **(Japanese):** The Japanese unseen set showed stronger robustness than Chinese and Korean because many of the examples included direct borrowed words, such as “baka,” inserted into English sentences. These suggest that direct lexical substitution is better for robustness than metaphorical slang.
- **(Algospeak):** The Algospeak unseen showed the best robustness across all datasets. This seems to be because it is from the same community as most of the ID dataset, which is English. It also handles easier obfuscations where some letters were substituted well, such as “bl0od.” However, robustness weakened for euphemistic expressions related to sensitive topics, such as “pew pew.”

Overall, the quantitative analysis suggests that the model’s robustness depended more on the linguistic form of the inputs surrounding the slang terms than on the community label. Models were more robust to direct lexical substitutions and contextually rich examples and less robust to metaphorical, euphemistic, or culturally grounded expressions.

## 7 Conclusion

Our project examined how well language models handle community-specific slang and algospeak under OOD generalization. We framed the task as a sequence-to-sequence rewriting problem, where the model rewrites slang-containing sentences into plain English. Across our experiments, we found that performance depended not just on the community source of the data, but more importantly on the type of examples included in training and how transferable those examples were to unseen settings. In particular, some added data improved robustness more than others because certain slang examples provided patterns that generalized beyond their original domain.

A major achievement of the project was building and evaluating a multilingual and multi-domain benchmark spanning Chinese, Korean, Japanese, and algospeak data. We systematically compared BART-base and T5-small under different training settings, prompt formulations, and hyperparameter choices. Overall, BART-base performed best, showing stronger robustness on both held-out Chinese and Korean validation sets and also better transfer to unseen Japanese and algospeak examples. Our qualitative analysis also showed that models were generally more successful when slang functioned like a direct lexical substitution, and less successful when slang was highly metaphorical, indirect, or heavily dependent on cultural context.

One of the main things we learned is that robustness to slang is highly context-sensitive. Models often succeeded when surrounding words made the intended meaning easier to infer, but failed when the same slang appeared in weaker or more ambiguous contexts. We also learned that adding more data does not automatically produce better robustness; the usefulness of the added examples matters more than raw quantity. This suggests that dataset composition is a central factor in making models more robust to evolving, culturally grounded language.

For future work, it would be valuable to repeat our augmentation experiments using targeted rather than randomly sampled OOD examples. Instead of selecting additional data at random, future studies could choose examples based on shared linguistic patterns, such as metaphorical structure, degree of indirectness, or contextual dependence. This may help reveal whether pattern-based augmentation leads to stronger robustness and better transfer to unseen slang than random data selection.

## Team contributions (Required for multi-person team)

Thet Htar Thin Zar curated the dataset, planned the experiments, and ran T5-small experiments, and wrote the report.

Hnin Yupar Mon curated the dataset, reviewed the experiments, ran BART-base experiments, and wrote the report

## References

- [1] Zhewei Sun and Yang Xu. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, 2022.

- [2] Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. Understanding slang with llms: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, 2024.
- [3] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. arXiv:2306.04618.
- [4] therohk. Urban dictionary words and definitions. Kaggle dataset, accessed 2026-03-13.
- [5] tawfiayeasmin. Gen z words and phrases dataset. <https://www.kaggle.com/datasets/tawfiayeasmin/gen-z-words-and-phrases-dataset>. Kaggle dataset, accessed 2026-03-13.
- [6] MLBtrio. genz-slang-dataset. Hugging Face dataset, accessed 2026-03-13.
- [7] Urban Dictionary. Urban dictionary. <https://www.urbandictionary.com>. Website, accessed 2026-03-13.
- [8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

## A Data Details

### A.1 Example Input-Output Pairs

| Prompt Input                                                                                                                                                                                                                                                        | Reference Output                                  | Model Output                              |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|-------------------------------------------|
| Rewrite the following sentence in plain English.<br><b>Slang:</b> Dog Blood<br><b>Definition:</b> Cliche or cheesy, especially a development or plot.<br><b>Input:</b> Let’s watch that dog blood movie tonight.                                                    | Let’s watch that cliché movie tonight.            | Let’s watch that cheesy movie tonight.    |
| Rewrite the following sentence in plain English.<br><b>Slang:</b> Face<br><b>Definition:</b> Dignity, reputation, honor.<br><b>Input:</b> Give me some face, please.                                                                                                | Please don’t embarrass me.                        | Give me some respect, please.             |
| Rewrite the following sentence in plain English.<br><b>Slang:</b> White Lotus<br><b>Definition:</b> Used sarcastically to describe someone who acts pure and harmless while secretly being manipulative.<br><b>Input:</b> I never expected him to be a white lotus. | I never expected him to be a manipulative person. | I never expected him to be a white lotus. |

Table 7: Example prompted inputs, reference outputs, and model outputs.