# AlignedAug: Alignment of naturalness and coherence score distributions for real dialogue augmentation

**Anonymous ACL submission**

## Abstract

Synthetic dialogues generated by Large Language Models (LLMs) exhibit differences from real dialogues in linguistic attributes such as naturalness or sentence completeness. To bridge this gap, we propose AlignedAug, a framework for realistic dialogue augmentation. AlignedAug consists of two stages: (1) Cognition-aware Dialogue Generation, which produces utterances using an LLM-based model; (2) Distribution Alignment, which is composed of Chat Style Refinement and Statistical Selection. Chat Style Refinement simulates informal, chat-like responses by randomly deleting words except for subjects, verbs, and negations. Statistical Selection selects dialogues whose naturalness and coherence scores are aligned with scores of real dialogues. Experimental results show that the Distribution Alignment stage in AlignedAug reduces the gap between real and synthetic dialogues (CollabChat $S_{KS}$: 0.95 $\rightarrow$ 0.35). In addition, AlignedAug outperforms existing LLM-based augmentation methods on classification and response selection tasks (classification accuracy: 0.65 $\rightarrow$ 0.69, response selection: R@5 0.77 $\rightarrow$ 0.84). These findings demonstrate that AlignedAug provides synthetic data that not only augments dialogues which align real dialogues more closely but also improves the performance of models trained on aligned dialogues.

## 1 Introduction

Dialogue data generation using LLMs has recently attracted increasing attention (Chen et al., 2024; Dai et al., 2025; Yuzbashyan et al., 2024). Dialogue data generation is especially useful for domain-specific dialogues where data collection is limited by privacy issues and costly annotation processes, such as dialogues in the education domain (Markel et al., 2023; Gao et al., 2025) and dialogues in the healthcare domain (Dai et al., 2025). Recently, discrepancies between real dialogues and synthetic dialogues generated by LLMs have attracted attention (Ren et al., 2025), as LLM-generated dialogues tend to be more consistent and controlled than real dialogues (Sandler et al., 2024). Real dialogues exhibit linguistic imperfections such as typos and incomplete sentences, leading to differences in linguistic attributes as well as in the distributions of dialogue quality metrics (Zhong et al., 2022), including *naturalness* and *coherence*, as illustrated in Figure 1. However, in dialogue augmentation research, there is a lack of analysis on the differences between real and synthetic dialogue and how to reduce those differences.

Table 1: Comparison of augmentation methods: ConvAug (Chen et al., 2024), AugGPT (Dai et al., 2025), SynAlign (Ren et al., 2025), and AlignedAug. Nat. dist. and Coh. dist. denote the naturalness distribution and coherence distribution, respectively.

| Method | Dialogue-level | Alignment target |
|---|---|---|
| ConvAug | ✓ | ✗ |
| AugGPT | ✗ | ✗ |
| SynAlign | ✗ | ✓(embedding dist.) |
| **AlignedAug (Ours)** | ✓ | ✓(Nat./Coh. dist.) |

Previous studies have evaluated the linguistic quality or usability of LLM-generated dialogues (Sandler et al., 2024; Chen and Artstein, 2024). To the best of our knowledge, no prior work has compared and analyzed distributional differences between real and synthetic dialogues using quantitative metrics. Thus, we propose AlignedAug, which minimizes distributional discrepancies between real and synthetic dialogues using quantitative metrics derived from UniEval, an LLM-based dialogue evaluation model (Zhong et al., 2022). We compare prior work with the proposed AlignedAug framework in Table 1. ConvAug (Chen et al., 2024) and AlignedAug perform dialogue-level augmentation methods. ConvAug does not address alignment to reduce differences between real and synthetic dialogues.

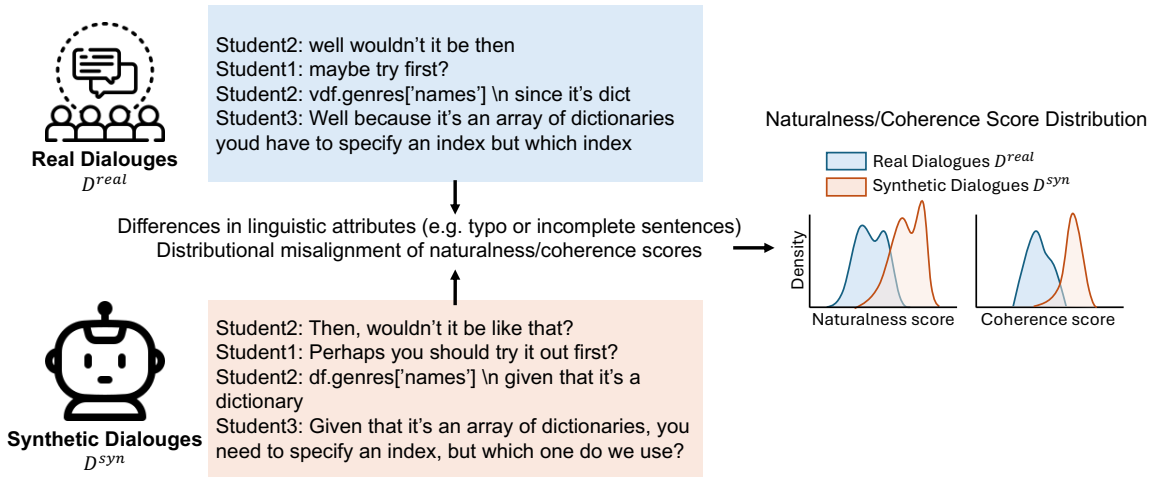



Figure 1: An example illustrating misalignment between LLM-generated synthetic dialogue and real dialogue.

AugGPT (Dai et al., 2025) and SynAlign (Ren et al., 2025) adopt text-level augmentation methods. SynAlign aligns the text embedding distributions of real and synthetic data, whereas our AlignedAug aligns the distributions of qualitative metric scores between real and synthetic dialogues.

AlignedAug is proposed as a dialogue augmentation framework to reduce the gap between real and synthetic dialogues. AlignedAug consists of (1) Cognition-aware Dialogue Generation and (2) Distribution Alignment. Cognition-aware Dialogue Generation adopts a three-step prompting approach motivated by theories of human cognition (Chen et al., 2024). The prompting strategy is further modified to incorporate contextual information. Distribution Alignment aims to reduce differences in the score distributions of *naturalness* and *coherence* between real and synthetic dialogues. Distribution Alignment is composed of Chat Style Refinement followed by Statistical Selection.

Our main contributions are as follows:

- Synthetic dialogues generated by AlignedAug lead to performance improvements on classification and response selection tasks compared to existing LLM-based augmentation methods (classification accuracy 0.652 $\rightarrow$ 0.690, response selection R@5 0.772 $\rightarrow$ 0.841).
- A dialogue augmentation framework, AlignedAug, is proposed to reduce the gap between real and synthetic dialogues. The proposed Distribution Alignment stage significantly reduces distributional differences (CollabChat $S_{KS}$: 0.95 $\rightarrow$ 0.35).

## 2 Related Work

We review prior work on dialogue augmentation, focusing on the evolution from token-level methods to sentence- and dialogue-level generation using large language models (LLMs). Since synthetic dialogues often exhibit distributional mismatches with real data, we further review studies on reducing such gaps between real and augmented dialogue distributions.

**LLM-based Augmentation.** Early dialogue augmentation methods relied on token-level techniques such as rule-based transformations and contextual masking or substitution (Wei and Zou, 2019; Kobayashi, 2018; Kenton et al., 2019). With the advent of LLMs, augmentation has advanced to sentence- and dialogue-level generation (Anaby-Tavor et al., 2020; Yuzbashyan et al., 2024). These methods have been applied to low-resource text classification (Dai et al., 2025) and multi-turn dialogue generation (Chen et al., 2024). However, LLM-generated dialogues often differ from real dialogues in linguistic and structural properties, leading to distributional mismatches (Ren et al., 2025).

**Distribution Alignment for Data Augmentation.** Real-world dialogues exhibit diverse discourse characteristics, including short utterances, incomplete sentences, and informal expressions (Lowe et al., 2015). Augmented data that does not reflect these properties can degrade performance due to distributional mismatch. To address this issue, recent work defines augmentation quality in terms of distributional similarity and proposes explicit alignment methods (Ren et al., 2025). For example, Zhao and Bilen (2023) minimize Maxi-

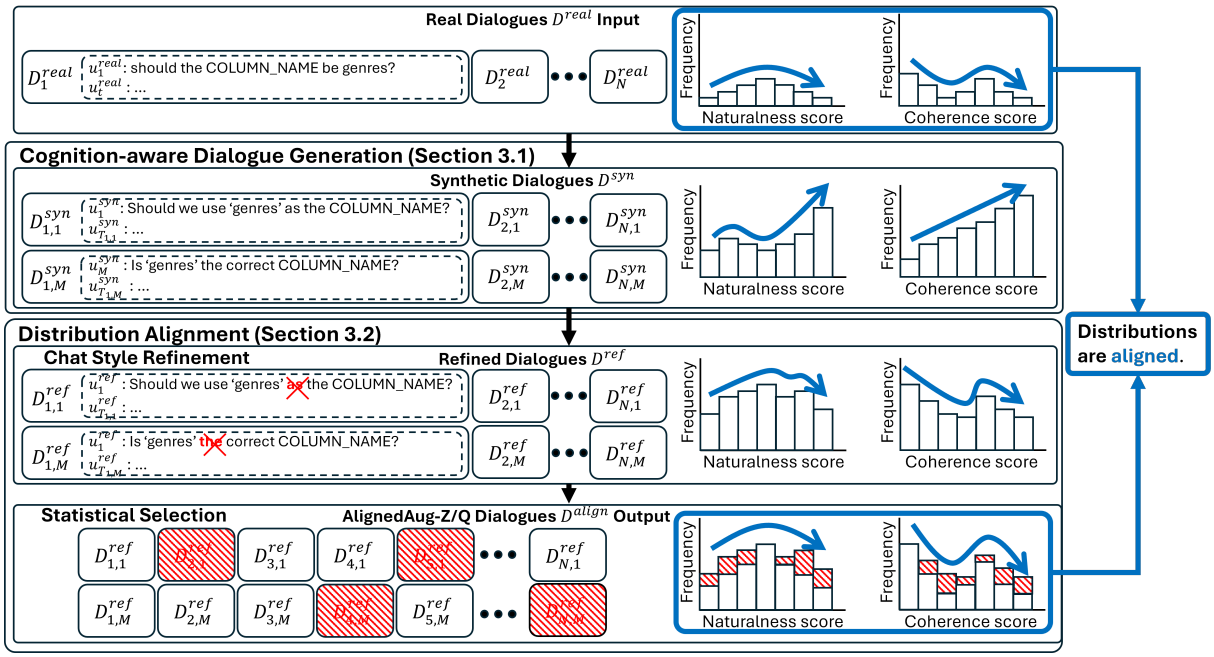



Figure 2: Overview of the proposed AlignedAug framework for realistic chat dialogue augmentation. The pipeline consists of two stages: (1) Cognition-aware Dialogue Generation and (2) Distribution Alignment. Hatched regions indicate dialogues that are filtered out during the Alignment process.

mum Mean Discrepancy (MMD) between embedding distributions, while SynAlign aligns LLM-generated and real data representations at the sentence level (Ren et al., 2025).

Despite these advances, existing alignment approaches primarily operate on sentence-level or static embeddings and fail to capture dialogue-specific properties such as multi-turn interactions and discourse flow. Since dialogue quality depends on discourse-level factors like coherence and naturalness across turns, dialogue-level distribution alignment is required. Accordingly, we propose AlignedAug, a dialogue-level augmentation framework that explicitly aligns coherence and naturalness score distributions between real and synthetic dialogues.

## 3 AlignedAug Framework

AlignedAug is a dialogue augmentation framework designed to generate high-quality augmented dialogues by explicitly aligning the distributions of *naturalness* and *coherence* scores between synthetic and real data. Given a set of real-world chat dialogues $\mathcal{D}^{\text{real}} = \{D_i^{\text{real}}\}_{i=1}^{N}$, AlignedAug produces an aligned dataset $\mathcal{D}^{\text{align}} = \{D_k^{\text{align}}\}_{k=1}^{K}$, where $N$ is the number of real dialogues and $K$ is the number of dialogues retained after statistical selection. The framework consists of two modules: (1) Cognition-aware Dialogue Generation, which generates diverse synthetic candidates, and (2) Distribution Alignment, which refines and selects synthetic dialogues to match the quality-score distributions of real dialogues.

### 3.1 Cognition-aware Dialogue Generation

Cognition-aware Dialogue Generation produces synthetic dialogues from real dialogues using GPT-4o. For each real dialogue $D_i^{\text{real}}$, we generate $M$ synthetic variants, where the augmentation scale $M$ is set to 6 following prior work (Dai et al., 2025). The generation unit is an utterance, enabling fine-grained control over dialogue structure and content diversity. A synthetic dialogue $D_{i,j}^{\text{syn}}$ is defined as a sequence of utterances: $D_{i,j}^{\text{syn}} = \left(u_{i,j,1}^{\text{syn}}, \ldots, u_{i,j,|D_{i,j}^{\text{syn}}|}^{\text{syn}}\right)$, where $|D_{i,j}^{\text{syn}}|$ denotes the number of utterances in the dialogue. For each real dialogue, the corresponding synthetic dialogue is $D_i^{\text{syn}} = \{D_{i,j}^{\text{syn}}\}_{j=1}^{M}$. The complete synthetic dataset is $D^{\text{syn}} = \{D_{i,j}^{\text{syn}}\}_{i=1,\ldots,N;\, j=1,\ldots,M}$.

To improve data quality and reduce hallucinations, we apply the Cognition-aware Prompting Process (Chen et al., 2024). This process follows a multi-step prompting strategy inspired by theories of human cognition, preserving the original topic and intent while introducing surface-level linguistic variations. To maintain label consistency, the label of each real dialogue $D_i^{\text{real}}$ is directly transferred to all corresponding synthetic variants.

### 3.2 Distribution Alignment

Distribution Alignment aims to reduce discrepancies between real and synthetic dialogues with respect to two quality metrics: *naturalness* and *coherence*. It operates on the synthetic dataset $D^{\text{syn}}$ and consists of two sequential stages: Chat Style Refinement and Statistical Selection.

**Chat Style Refinement** Chat Style Refinement adjusts synthetic dialogues to better resemble the surface characteristics of real chat conversations. Specifically, the refinement process applies random word deletion to synthetic utterances while preserving semantic integrity. Nouns, verbs, and negation words are excluded from deletion due to their critical role in meaning representation.

Let $u_t^{\text{syn}}$ denote the $t$-th utterance in a synthetic dialogue and let $L_t$ be its token length. With deletion ratio $r = 0.2$, the number of deleted tokens $d_t$ is defined as

$$d_t = \min\Big(5, \max\big(1, \lfloor r \cdot L_t \rfloor\big)\Big). \tag{1}$$

Applying this operation to all utterances yields a refined dataset $D^{\text{ref}}$. This refinement reflects common conversational omissions observed in real-world dialogue systems, where speakers often leave contextually inferable information implicit (Cao et al., 2024).

**Statistical Selection** Statistical Selection filters refined dialogues to construct the final aligned dataset $\mathcal{D}^{\text{align}}$. Two complementary strategies are employed: AlignedAug-Z and AlignedAug-Q.

**AlignedAug-Z.** AlignedAug-Z performs score normalization using real-dialogue statistics. Let $m \in \{\text{nat}, \text{coh}\}$ index the metric, and let $s_m(D) \in \mathbb{R}$ denote the score of dialogue $D$. For each metric, the mean and standard deviation of real-dialogue scores are computed as

$$\mu_m^{\text{real}} = \frac{1}{N}\sum_{i=1}^{N} s_m(D_i^{\text{real}}), \tag{2}$$

$$(\sigma_m^{\text{real}})^2 = \frac{1}{N}\sum_{i=1}^{N}\big(s_m(D_i^{\text{real}}) - \mu_m^{\text{real}}\big)^2. \tag{3}$$

For a refined dialogue $D \in \mathcal{D}^{\text{ref}}$, the z-score is defined as

$$z_m(D) = \frac{s_m(D) - \mu_m^{\text{real}}}{\sigma_m^{\text{real}}}.$$

Since $\pm 1$ standard deviation covers approximately 68% of a normal distribution (Anusha et al., 2019), we retain only dialogues within this range for both naturalness and coherence. Specifically, a dialogue $D$ is kept if it satisfies

$$|z_{\text{nat}}(D)| \le 1 \ \wedge \ |z_{\text{coh}}(D)| \le 1. \tag{4}$$

**AlignedAug-Q.** AlignedAug-Q aligns the joint distribution of naturalness and coherence scores via two-dimensional quantile-based matching. Let $q = 10$ denote the number of quantile bins. This choice provides sufficient granularity while maintaining a statistically meaningful number of samples in each bin (Pylkkonen et al., 2016). Quantile boundaries $b_0^m < \cdots < b_q^m$ computed from real-dialogue scores partition the score space into $q \times q$ bins:

$$\begin{aligned}\text{Bin}_{k,\ell} = \big\{D \mid\ & b_{k-1}^{\text{nat}} \le s_{\text{nat}}(D) < b_k^{\text{nat}}, \\ & b_{\ell-1}^{\text{coh}} \le s_{\text{coh}}(D) < b_\ell^{\text{coh}}\big\},\end{aligned} \tag{5}$$

where $k \in \{1, \ldots, q\}$ and $\ell \in \{1, \ldots, q\}$ index the quantile bins along the naturalness and coherence dimensions, respectively. The proportion of real dialogues in each bin is

$$\rho_{k,\ell}^{\text{real}} = \frac{1}{|\mathcal{D}^{\text{real}}|}\sum_{D \in \mathcal{D}^{\text{real}}} \mathbb{I}[D \in \text{Bin}_{k,\ell}]\,.$$

Aligned dataset is constructed by subsampling dialogues such that the joint distribution over quantile bins matches that of the real data. Concretely, for all $k, \ell \in \{1, \ldots, q\}$,

$$\frac{|\mathcal{D}^{\text{align}} \cap \text{Bin}_{k,\ell}|}{|\mathcal{D}^{\text{align}}|} \approx \rho_{k,\ell}^{\text{real}}, \tag{6}$$

thereby ensuring that the joint score distribution of the aligned dataset closely follows that of the real data.

## 4 Experiments

Three experiments are conducted; 1) Task performance study on 3 downstream tasks of classification, response selection and summarization, 2) Ablation study on the two modules of AlignedAug, 3) Performance of AlignedAug in terms of reducing distributional gaps between real and synthetic dialogues. The task performance study and ablation study experiments are repeated ten times with different random seeds to eliminate randomness, and the mean and standard deviation are reported.

Table 2: Overview of real dialogue datasets. Utt. and Dlg. denote the numbers of utterances and dialogues, respectively. Spk./Dlg. reports the minimum–maximum numbers of speakers per dialogue (mean in parentheses). Avg. Utt./Dlg. denotes the average number of utterances per dialogue. Label denotes the label count.

| Split | CollabChat | DeliData | NPSChat | Ubuntu | SAMSum |
|---|---|---|---|---|---|
| Train Utt. (Dlg.) | 2,040(24) | 2,030(65) | 1,600(3) | 1,541(31) | 1,876(194) |
| Val Utt. (Dlg.) | 344(6) | 455(17) | 552(1) | 454(8) | 504(49) |
| Test Utt. (Dlg.) | 492(9) | 493(19) | 974(2) | 487(10) | 488(50) |
| Spk./Dlg. | 2–5 (3.7) | 2–5 (3.1) | 35–63 (43.8) | 10–25 (17.6) | 2–7 (2.4) |
| Avg. Utt./Dlg. | 73.7 | 29.5 | 521.0 | 51.1 | 9.7 |
| Task (Label) | Classification (4) | Classification (6) | Classification (14) | Response Selection (-) | Summarization (-) |

### 4.1 Experiments for Task Performance on Three Downstream Tasks

This section evaluates whether models trained with data augmented by AlignedAug outperform existing augmentation methods across three downstream tasks: classification, response selection, and summarization. Detailed dataset descriptions are provided in the appendix.

For classification, the input consists of three context utterances and one target utterance, following prior work (Ortega and Vu, 2017; Ghosal et al., 2021). Performance is measured using Accuracy and Weighted-F1. For response selection, up to five previous utterances are used as context under a binary classification setup, and evaluation is conducted over ten candidates. Recall@$k$ metrics (R@1, R@2, R@5) are reported (Lowe et al., 2015). For summarization, models generate summaries from full dialogues and are evaluated using ROUGE-1, ROUGE-2, and ROUGE-L (Chen and Bansal, 2018; See et al., 2017).

**Dataset** We evaluate downstream task performance using five real-world dialogue datasets. For classification, we use CollabChat, DeliData (Karadzhov et al., 2023), and NPSChat (Forsythand and Martell, 2007). The response selection task is evaluated on the Ubuntu Dialogue Corpus (Lowe et al., 2015), and summarization is evaluated on the SAMSum Corpus (Gliwa et al., 2019). Training and validation sets are used for data augmentation and model training, while test sets are reserved exclusively for evaluation. Table 2 summarizes key statistics for each dataset, including the number of utterances, dialogues, speakers, average dialogue length, and task type.

**Baseline Augmentation Methods** We compare AlignedAug with three representative LLM-based text and dialogue augmentation methods to evaluate downstream task performance. **ConvAug** (Chen et al., 2024) is a dialogue-level augmentation framework that employs cognition-aware prompting to preserve semantic consistency. **AugGPT** (Dai et al., 2025) is a zero-shot sentence paraphrasing method. **SynAlign** (Ren et al., 2025) aligns LLM-generated and real text distributions in embedding space and shows strong performance in low-data regimes. However, as SynAlign generates individual utterances by selecting representative samples per label rather than full dialogues, it is excluded from the response selection and summarization baselines.

**Experimental Setup** To ensure reproducibility, we use deterministic generation settings for LLM-based dialogue generation, fixing the random seed to 42 and applying deterministic configurations for both PyTorch and CUDA. The decoding temperature is set to 0.0, and identical prompt templates and context formats are used across all experiments. Dialogue generation conditions on the five preceding utterances to capture recent conversational flow while avoiding outdated context (Shen et al., 2020, 2022). For fair comparison, all downstream tasks use identical model architectures and training setups. Models are trained for up to 30 epochs with early stopping based on validation performance. For classification and response selection, we fine-tune BERT-base-uncased using AdamW and CrossEntropyLoss with a learning rate of $2 \times 10^{-5}$, batch size 8, weight decay 0.05, and cosine decay scheduling. For summarization, we fine-tune FLAN-T5-base with AdamW, a learning rate of $5 \times 10^{-5}$, weight decay 0.01, warmup ratio 0.05, batch size 4, and beam search with beam size 4.

### 4.2 Experiments for an Ablation Study of the AlignedAug Modules

This section conducts an ablation study to evaluate the effect of the Chat Style Refinement and Statistical Selection modules in the Distribution Alignment stage of AlignedAug. Classification per-

Table 3: Comparison of classification performance across data augmentation methods. Values represent Accuracy (Acc) and Weighted-F1 (F1) with mean ± standard deviation over 10 runs.

| Method | CollabChat | | DeliData | | NPSChat | |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 |
| NoAugment. | 0.471±0.003 | 0.312±0.026 | 0.444±0.040 | 0.304±0.075 | 0.520±0.000 | 0.350±0.000 |
| ConvAug | 0.638±0.014 | 0.645±0.014 | 0.772±0.026 | 0.774±0.028 | 0.663±0.023 | 0.680±0.022 |
| AugGPT | 0.645±0.011 | 0.649±0.012 | 0.810±0.014 | 0.811±0.014 | 0.699±0.026 | 0.709±0.024 |
| SynAlign | 0.652±0.033 | 0.629±0.056 | 0.571±0.114 | 0.528±0.138 | 0.624±0.051 | 0.581±0.080 |
| **Synthetic Dlg.** | 0.640±0.011 | 0.645±0.014 | 0.798±0.018 | 0.802±0.015 | 0.681±0.016 | 0.695±0.016 |
| **AlignedAug-Z** | **0.690**±0.008 | **0.689**±0.009 | 0.808±0.006 | 0.810±0.007 | **0.741**±0.015 | **0.742**±0.009 |
| **AlignedAug-Q** | 0.623±0.083 | 0.591±0.154 | **0.814**±0.014 | **0.815**±0.013 | 0.725±0.073 | 0.698±0.123 |

formance is evaluated when both modules are removed, and when each module is removed individually. The experimental setup follows the classification task setting described in Section 4.1. Three classification datasets are used for the ablation study: CollabChat, DeliData (Karadzhov et al., 2023), and NPSChat (Forsythand and Martell, 2007).

### 4.3 Experiments for Evaluating AlignedAug Performance on Distributional Gap Reduction

This section evaluates whether the Distribution Alignment stage of AlignedAug reduces distributional differences between real and synthetic dialogues using the Kolmogorov–Smirnov (KS) test. The KS test is applied to assess the statistical significance of distributional differences in *naturalness* and *coherence* scores between real dialogues $D^{real}$ and synthetic dialogues $D^{syn}$. Before applying Distribution Alignment, the distributional difference is measured using the KS statistic $S_{KS}$ between $D^{real}$ and $D^{syn}$. After applying Distribution Alignment, the $S_{KS}$ is computed between $D^{real}$ and aligned dialogues $D^{align}$. A reduction in $S_{KS}$ and an increase in the corresponding $p$-value indicate a reduction in distributional differences. The test statistic is denoted as $S_{KS}$ to avoid confusion with dialogue notation $D$. Five real-world dialogue datasets are used to evaluate distributional gap reduction. The datasets include DeliData (Karadzhov et al., 2023), NPSChat (Forsythand and Martell, 2007), Ubuntu Dialogue Corpus (Lowe et al., 2015), and SAMSum Corpus (Gliwa et al., 2019).

## 5 Results and Discussion

This section analyzes three experimental results: 1) Task performance study on 3 downstream tasks of classification, response selection and summarization, 2) Ablation study on the two modules of AlignedAug, 3) Performance of AlignedAug in terms of reducing distributional gaps between real and synthetic dialogues.

### 5.1 Task Performance on Three Downstream Tasks

This subsection presents the effectiveness of data augmented by AlignedAug across three downstream tasks. The evaluated tasks include classification, response selection, and summarization. Semantic preservation and diversity scores are additionally analyzed to explain why AlignedAug consistently outperforms baseline methods.

**Classification Task** The proposed AlignedAug achieves the best classification performance on CollabChat, DeliData, and NPSChat (0.690, 0.814, and 0.741, respectively), as shown in Table 3. All LLM-based augmentation methods, including AlignedAug, significantly outperform NoAugment. Synthetic dialogues $D^{syn}$ are generated using LLMs without Distribution Alignment. Performance remains comparable to existing augmentation methods. AlignedAug-Z and AlignedAug-Q, which apply Distribution Alignment, consistently outperform synthetic dialogues and baseline methods. These results demonstrate the effectiveness of the Distribution Alignment module and validate the superiority of AlignedAug.

**Response Selection Task** AlignedAug achieves the highest performance across all metrics in the response selection task (R@1 0.497, R@2 0.634, R@5 0.841). AlignedAug-Z outperforms the synthetic dialogues $D^{syn}$ on the Ubuntu Dialogue Corpus, as shown in Table 4. These results indicate that Distribution Alignment improves response selection performance and confirm the effectiveness of AlignedAug for response selection task.

Table 4: Comparison of response selection performance (Ubuntu Dataset) across data augmentation methods. Values represent R@1, R@2 and R@5 (mean ± standard deviation) over 10 runs.

| Method | R@1 | R@2 | R@5 |
|---|---|---|---|
| NoAugment. | 0.427 ± 0.030 | 0.561 ± 0.031 | 0.772 ± 0.048 |
| ConvAug | 0.406 ± 0.025 | 0.520 ± 0.037 | 0.737 ± 0.032 |
| AugGPT | 0.434 ± 0.017 | 0.553 ± 0.031 | 0.765 ± 0.021 |
| **Synthetic Dlg.** | 0.415 ± 0.019 | 0.527 ± 0.023 | 0.754 ± 0.029 |
| **AlignedAug-Z** | **0.497** ± 0.021 | **0.634** ± 0.011 | 0.839 ± 0.016 |
| **AlignedAug-Q** | 0.481 ± 0.032 | 0.631 ± 0.022 | **0.841** ± 0.017 |

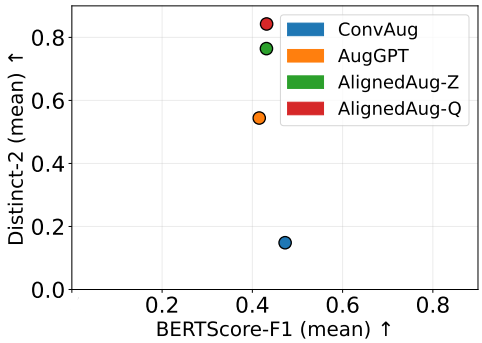
**Summarization Task** The highest summarization performance is achieved using synthetic dialogues generated by the proposed Cognition-aware Dialogue Generation stage (ROUGE-1 0.525, ROUGE-2 0.293, ROUGE-L 0.438). Results on the SAMSum Corpus dataset are presented in Table 5. Overall performance gains remain modest. All augmentation methods rely on paraphrasing and do not introduce new information or additional learning signals. SAMSum Corpus contains short dialogues with high information density, averaging 9.7 utterances per dialogue. Paraphrasing-based augmentation consequently yields limited benefits.

Table 5: Comparison of summarization performance (SAMSum Dataset) across data augmentation methods. Values represent ROUGE-1, ROUGE-2 and ROUGE-L with mean ± standard deviation over 10 runs.

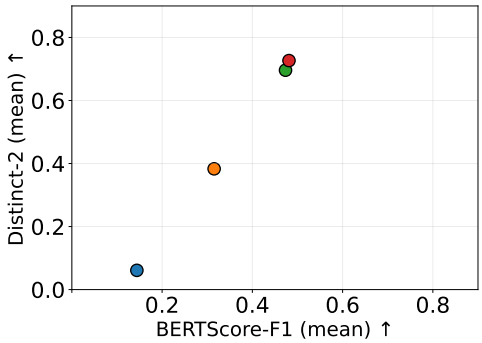
| Method | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| NoAugment. | 0.477 ± 0.009 | 0.256 ± 0.011 | 0.398 ± 0.011 |
| ConvAug | 0.514 ± 0.006 | 0.290 ± 0.007 | 0.431 ± 0.007 |
| AugGPT | 0.519 ± 0.009 | 0.292 ± 0.011 | 0.434 ± 0.008 |
| **Synthetic Dlg.** | **0.525** ± 0.010 | **0.293** ± 0.009 | **0.438** ± 0.007 |
| **AlignedAug-Z** | 0.518 ± 0.013 | 0.288 ± 0.009 | 0.435 ± 0.012 |
| **AlignedAug-Q** | 0.513 ± 0.009 | 0.285 ± 0.014 | 0.430 ± 0.011 |

**Comparison of Meaning Preservation and Diversity across Augmentation Methods** We compare the trade-off between meaning preservation and diversity across dialogue augmentation methods on CollabChat and DeliData, as shown in Figure 3. The x-axis represents BERTScore-F1 (Zhang et al., 2020), which measures semantic similarity between real and synthetic dialogues. Higher values indicate better meaning preservation. The y-axis represents Distinct-2 (Li et al., 2016), which measures lexical diversity. Higher values indicate more diverse expressions. Methods located in the upper-right region achieve strong semantic preservation and high diversity.

AlignedAug achieves a superior balance across all datasets compared to existing augmentation




(a) CollabChat




(b) DeliData

Figure 3: Comparison of meaning preservation and diversity across dialogue augmentation methods.

methods. AlignedAug-Q consistently appears in the upper-right region, indicating strong semantic preservation and high diversity. AlignedAug-Z maintains high BERTScore-F1 while achieving significantly higher Distinct-2 values than baseline methods. These results indicate stable semantic preservation and effective diversity enhancement.

## 5.2 Ablation Study of AlignedAug Modules

The full configurations of AlignedAug-Z and AlignedAug-Q achieve the highest performance across all datasets compared to variants without refinement, as analyzed through the impact of each component in the AlignedAug framework in Table 6. Bold values in Table 6 highlight this trend. These results demonstrate the effectiveness of the Chat Style Refinement strategy.

Table 6: Ablation study of the proposed AlignedAug framework. Values represent Accuracy (mean ± standard deviation) over 10 runs.

| Method | CollabChat | DeliData | NPSChat |
|---|---|---|---|
| **AlignedAug-Z** | **0.690**±0.008 | **0.808**±0.006 | **0.741**±0.015 |
| w/o Refine | 0.471±0.003 | 0.801±0.011 | 0.520±0.000 |
| **AlignedAug-Q** | **0.623**±0.083 | **0.814**±0.014 | **0.725**±0.073 |
| w/o Refine | 0.529±0.154 | 0.808±0.014 | 0.520±0.000 |
| w/o Select | 0.690±0.082 | 0.803±0.016 | 0.730±0.018 |
| w/o Ref. Sel. | 0.640±0.011 | 0.808±0.018 | 0.681±0.016 |

Table 7: Changes in the distributions of *naturalness* and *coherence* across datasets after applying Distribution Alignment

| | ***Naturalness*** **(KS-Test** $S_{KS}\downarrow$ $(p^{\uparrow})$**)** | | | | | ***Coherence*** **(KS-Test** $S_{KS}\downarrow$ $(p^{\uparrow})$**)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **CollabChat** | **DeliData** | **NPSChat** | **Ubuntu** | **SAMSum** | **CollabChat** | **DeliData** | **NPSChat** | **Ubuntu** | **SAMSum** |
| Synthetic Dlg. | 0.95(4e-29) | 0.57(4e-22) | 0.79(0.01) | 0.92(7e-33) | 0.53(4e-56) | 0.48(4e-6) | 0.30(2e-6) | 0.42(0.52) | 0.35(3e-4) | 0.26(2e-13) |
| AlignedAug-Z | 0.39(0.04) | 0.23(1e-3) | **0.72(9e-7)** | 0.21(0.12) | 0.20(1e-6) | **0.38(0.04)** | **0.18(0.04)** | **0.41(0.02)** | 0.18(0.26) | 0.17(6e-5) |
| AlignedAug-Q | **0.35(0.04)** | **0.14(0.16)** | 0.75(0.10) | **0.15(0.56)** | **0.07(0.37)** | 0.40(0.02) | **0.18(0.04)** | 0.83(0.05) | **0.12(0.87)** | **0.06(0.56)** |

Removing Chat Style Refinement (w/o. Refine) causes substantial performance degradation across most datasets. Performance drops are particularly large on CollabChat and NPSChat. AlignedAug-Z decreases from 0.690 to 0.471 on CollabChat and from 0.741 to 0.520 on NPSChat. AlignedAug-Q shows similar patterns. These results indicate that unrefined synthetic dialogues fail to capture the incompleteness and colloquial nature of real conversations.

Statistical Selection also plays a critical role. Removing Selection (w/o. Select) reduces average performance or increases variance, leading to reduced stability. Standard deviation on CollabChat increases from 0.008 to 0.082. Mean performance decreases on DeliData and NPSChat from 0.808 to 0.803 and from 0.741 to 0.730, respectively. Removing both Refinement and Selection (w/o. Ref. Sel.) further degrades performance or increases variance compared to the full AlignedAug-Z configuration. These results demonstrate that Distribution Alignment contributes to both performance and stability.

### 5.3 AlignedAug Performance on Distributional Gap Reduction

This subsection demonstrates statistically significant differences between real and synthetic dialogues and shows that Distribution Alignment reduces these differences.

***Naturalness*** **and** ***Coherence*** **Score Distribution Gap** Naturalness score distributions differ significantly between real dialogues $D^{real}$ and LLM-generated synthetic dialogues $D^{syn}$ across all datasets ($p \leq 0.05$), as summarized by the KS test results in Table 7. These differences mainly stem from sentence completeness: real dialogues often contain abbreviations and incomplete utterances, whereas synthetic dialogues favor grammatically complete and polished expressions. As illustrated in Figure 1, this results in higher naturalness scores for synthetic dialogues. Coherence score distributions show dataset-dependent behavior. In NPSChat, dialogues involve many speakers and long interactions, making topic consistency difficult to maintain and leading to low coherence scores in both real and synthetic dialogues. Consequently, coherence distribution differences for NPSChat are small and statistically insignificant.

**Reduction of Distributional Differences After Distribution Alignment** Distribution Alignment significantly reduces differences in naturalness and coherence score distributions between real dialogues $D^{real}$ and AlignedAug dialogues $D^{align}$ across all datasets. The KS statistics $S_{KS}$ for naturalness and coherence decrease across all datasets after alignment, as compared in Table 7. Corresponding $p$-values generally increase. CollabChat shows a reduction from $S_{KS}$ 0.95 to 0.35 and a $p$-value increase from $4e{-}29$ to 0.04. These results indicate that aligned dialogues more closely match real dialogue distributions. Naturalness on NPSChat shows a decrease in $p$-value from 0.01 to $9e{-}7$. Distributional differences in $D^{syn}$ are already small for this dataset. Additional improvement is consequently limited.

## 6 Conclusion

This study proposed AlignedAug, a data augmentation framework for dialogue data. The framework quantitatively analyzes and effectively reduces the differences between real and LLM-generated synthetic dialogues. We show that AlignedAug consistently decreases the distributional gaps in naturalness and coherence between real and synthetic dialogues across all datasets. These findings quantitatively demonstrate that Chat Style Refinement and Statistical Selection play a crucial role in approximating real dialogue distributions. Furthermore, we empirically verify that using distribution-aligned synthetic data improves performance on downstream tasks, including dialogue classification and response selection.

## 7 Limitations

The Chat Style Refinement module relies on conditional random token deletion and adopts a relatively simple design. The strategy captures common properties of real dialogues, including incompleteness and colloquial variation. The approach does not explicitly control higher-level conversational phenomena. Future work should explore more structured refinement strategies. Potential directions include modeling abbreviations, ungrammatical expressions, and other discourse-level irregularities. Such extensions may further improve alignment with real-world conversational behavior.

Despite these limitations, the study provides a systematic analysis of distributional mismatch in LLM-based dialogue augmentation. The proposed framework offers a practical solution for improving distribution alignment between real and synthetic dialogues.

## References


Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7383–7390.

Peruri Venkata Anusha, Ch Anuradha, PSR Chandra Murty, and Ch Surya Kiran. 2019. Detecting outliers in high dimensional data sets using z-score methodology. *International Journal of Innovative Technology and Exploring Engineering*, 9(1):48–53.

Zhiyu Cao, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2024. Incomplete utterance rewriting with editing operation guidance and utterance augmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7225–7238.

Elizabeth Chen and Ron Artstein. 2024. Augmenting training data for a virtual character using gpt-3.5. In *The International FLAIRS Conference Proceedings*, volume 37.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing conversational dense retrieval via LLM-cognition data augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2700–2718, Bangkok, Thailand. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, and 1 others. 2025. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Transactions on Big Data*.

Eric N Forsythand and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23923–23932.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multiparty problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25.

Jacob Devlin Ming-Wei Chang Kenton, Lee Kristina Toutanova, and 1 others. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual*

*Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 247–252, Saarbrücken, Germany. Association for Computational Linguistics.

Janne Pylkkonen, Thomas Drugman, and Max Bisani. 2016. Optimizing speech recognition evaluation using stratified sampling.

Jiyuan Ren, Zhaocheng Du, Zhihao Wen, Qinglin Jia, Sunhao Dai, Chuhan Wu, and Zhenhua Dong. 2025. Few-shot llm synthetic data with distribution matching. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 432–441.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 366–380. Springer.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks.

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Nerses Yuzbashyan, Nikolay Banar, and Walter Daelemans. 2024. Dialogue-level data augmentation for conversation derailment forecasting and topic-shift detection. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 179–185.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations (ICLR)*.

Bo Zhao and Hakan Bilen. 2023. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A Dataset Details

This study uses five real-world chat dialogue datasets. Three datasets target classification tasks, while the remaining two target response selection and summarization. Five datasets were selected from available dialogue benchmarks to reflect realistic chat environments. Table 8 summarizes the labels and descriptions for the three classification datasets.

**CollabChat** is a collaboration-oriented chat dataset collected from a graduate-level introductory data science course at a private R1 university in the United States. Each utterance is annotated with four labels, including three collaborative competencies and one non-competency category. Students were assigned to groups of one to five members and completed three subtasks using a web-based platform that integrates JupyterLab and chat functionality. Data were collected across three semesters from 43 groups. After excluding individual-only groups, chat logs from 39 groups were used for analysis. The analyzed set contains 2,876 utterances after removing agent messages from the original 3,242 utterances. **DeliData** (Karadzhov et al., 2023) contains collaborative dialogues collected during cognitive task-solving processes. The dataset consists of 500 dialogues and approximately 14,000 utterances. Each utterance is annotated with one of six dialogue acts, such as questions, suggestions, or agreements, enabling utterance-level classification. To reflect utterance-type distributions observed in real collaborative settings, 101 dialogues were randomly sampled, yielding 2,078 utterances for the classification task. **NPSChat** (Forsythand and Martell,

2007) is a large-scale dataset collected from online chat rooms involving approximately 3,200 users. The dataset contains over 470,000 messages covering unrestricted, free-topic conversations. Each utterance is annotated with one of 14 dialogue act labels. Long dialogue flows and multi-speaker interactions in the dataset reflect characteristics of open public chat environments. Six dialogues were randomly selected, comprising 3,126 utterances, to analyze distributional characteristics under long and multi-speaker settings. **Ubuntu Dialogue Corpus** (Lowe et al., 2015) consists of multi-turn dialogues exchanged by Ubuntu users seeking technical support. The dataset contains approximately 930,000 two-party conversations and more than 7 million utterances. Dialogues are unstructured text-based chats and are widely used for response selection tasks that require contextual understanding. A subset of 49 dialogues was sampled, resulting in 2,482 utterances for the response selection task. **SAMSum Corpus** (Gliwa et al., 2019) contains messenger-style chat dialogues paired with human-written abstractive summaries. The dataset includes approximately 16,000 conversations covering a range of formality levels. Each dialogue is associated with a third-person summary, making the dataset suitable for dialogue summarization tasks. A total of 293 dialogues were selected, including 2,868 utterances, for the summarization task.

Public access is available for DeliData, NPSChat, Ubuntu Dialogue Corpus, and SAMSum Corpus, while CollabChat is private. Public datasets labeled with collaboration competency in authentic educational settings are not available, motivating the use of CollabChat. CollabChat has a limited size, so collaborative utterance augmentation was applied to increase training data volume and improve model generalization. The augmented data can also support future development of collaboration-aware assistance systems.

## B Prompt Templates

The prompt template used for Cognition-aware Dialogue Generation, the first step of AlignedAug, is presented in Figure 4. The template provides the target utterance along with five preceding utterances as context. For classification tasks, each utterance is accompanied by its category label and a short textual description. Response generation is guided by a three-step procedure based on cognition-aware prompting strategies. Each generation run produces six paraphrased utterances for the given target utterance.

```
### System Prompt:
Context:
Turn: 130, Speaker: Student5, Utterance: Wow, well
done, Category: f3
  ← Category meaning: Maintaining team function
…

Target Turn to Rewrite:
Turn: 135, Speaker: Student3, Utterance: Gotcha, not
concerned about it, Category: x
  ← Category meaning: Non collaborative competency
```

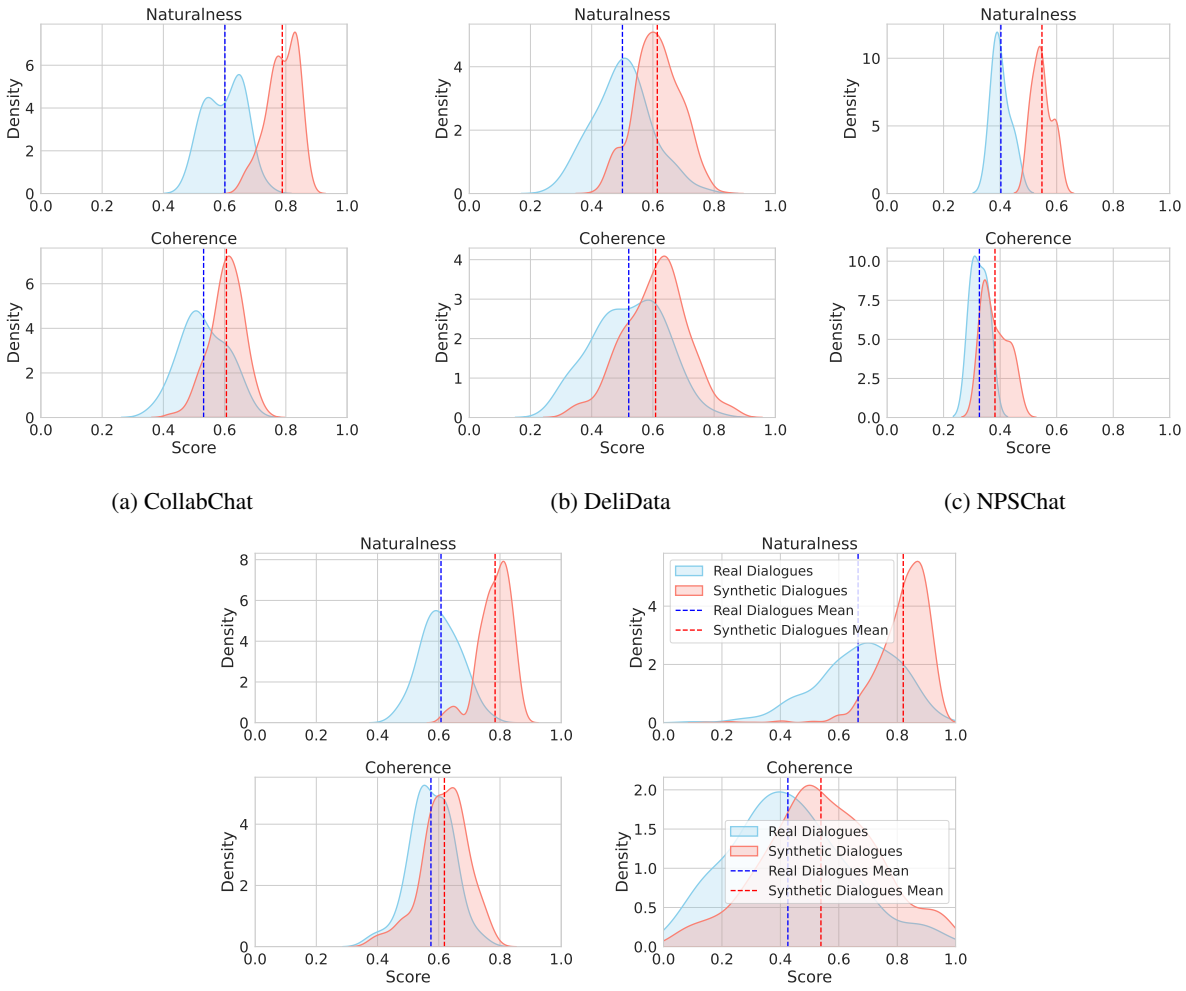
**Cognition-aware Generation**

```
Instructions:
Step 1: Comprehension Synthesis:
[Identify key themes and intents of the conversation]

Step 2: Associative Expansion:
[Generate alternative expressions based on existing
ones]

Step 3: Conclusion:
Rewritten Conversation:
Create six distinct and meaning-preserving paraphrased
versions of the conversation using the above steps.
Each version must be clearly different from the others
in wording, phrasing, or structure, while preserving
the original intent.
Return the result as a JSON object with keys
'conversation_1' through 'conversation_6'.
```

Figure 4: Example prompt template

Table 8: Label descriptions for classification task dataset

| Dataset | Label | Description |
|---|---|---|
| **CollabChat (4)** | f1 | Constructing shared knowledge |
| | f2 | Negotiation / Coordination |
| | f3 | Maintaining team function |
| | x | Non-collaborative competency |
| **DeliData (6)** | Moderation | Managing or guiding the discussion process |
| | Reasoning | Providing arguments or explanations |
| | Solution | Proposing or evaluating task answers |
| | Agree | Showing agreement with previous statements |
| | Disagree | Expressing disagreement or counter-opinion |
| | None | Unrelated utterances (e.g., greetings, hesitation) |
| **NPSChat (14)** | Accept | Positive response or agreement |
| | Bye | Goodbye or farewell |
| | Clarify | Clarification or request for clarification |
| | Continuer | Backchannel or continuer signal |
| | Emotion | Expression of emotion or feeling |
| | Emphasis | Emphasizing or reinforcing a statement |
| | Greet | Greeting |
| | No Answer | No answer to a question |
| | Other | Uncategorized or miscellaneous utterance |
| | Reject | Disagreement or rejection |
| | Statement | Declarative statement |
| | WhQuestion | Wh-question |
| | Yes Answer | Affirmative answer to yes/no question |
| | ynQuestion | Yes/no question |




(a) CollabChat (b) DeliData (c) NPSChat

(d) Ubuntu Dialogue Corpus (e) SAMSum Corpus

Figure 5: Distribution of *naturalness* and *coherence* scores for real dialogues $D^{real}$ vs. synthetic dialogues $D^{syn}$ across datasets. Panels (a)–(e) are based on real-world chat dialogues. Blue: real dialogues $D^{real}$, Red: synthetic dialogues $D^{syn}$.

# C Additional Analysis

## C.1 Naturalness and Coherence Distribution Gap

Distributional differences in *naturalness* and *coherence* scores were analyzed between real dialogues $D^{real}$ and synthetic dialogues $D^{syn}$ across CollabChat, DeliData, NPSChat, Ubuntu Dialogue Corpus and SAMSum Corpus. The score distributions are visualized in Figure 5. Across all five datasets, synthetic dialogues (red) shows higher *naturalness* and *coherence* scores than the real dialogues (blue). Real-world chat dialogues frequently contain fragmented sentences, abbreviations, and typographical errors, which lowers *naturalness* and *coherence* scores in the real data compared to the synthetic data.

## C.2 Distribution Shifts Across Chat Style Refinement and Statistical Selection

Additional experiments used the Kolmogorov–Smirnov (KS) test to examine how the *naturalness* and *coherence* score distributions of LLM-generated dialogues change across AlignedAug stages. The analysis emphasizes how Distribution Alignment affects distributional similarity.

Chat Style Refinement generally reduces the KS statistic $S_{KS}$ and increases the $p$-value across datasets, indicating improved similarity to real dialogue distributions. For example, for *naturalness*, CollabChat decreases from $S_{KS} = 0.95$ to $S_{KS} = 0.43$, and DeliData decreases from $S_{KS} = 0.57$ to $S_{KS} = 0.21$. For *coherence*, CollabChat decreases from $S_{KS} = 0.48$ to $S_{KS} = 0.43$, and DeliData decreases from $S_{KS} = 0.30$ to $S_{KS} = 0.16$.

Table 9: Changes in the distributions of *naturalness* and *coherence* across datasets (CollabChat, DeliData, NPSChat) after applying Chat Style Refinement and Statistical Selection.

| Metric | Method | CollabChat KS-Test $S_{KS}{}^{\downarrow}$ ($p^{\uparrow}$) | DeliData KS-Test $S_{KS}{}^{\downarrow}$ ($p^{\uparrow}$) | NPSChat KS-Test $S_{KS}{}^{\downarrow}$ ($p^{\uparrow}$) |
|---|---|---|---|---|
| *naturalness* | Synthetic Dlg. | 0.95 ($4 \times 10^{-29}$) | 0.57 ($4 \times 10^{-22}$) | 0.79 (0.01) |
| | Chat Style Ref. | 0.43 ($6 \times 10^{-3}$) | 0.21 ($3 \times 10^{-3}$) | 0.54 ($8 \times 10^{-4}$) |
| | Statistical SelZ. | 0.39 (0.04) | 0.23 ($1 \times 10^{-3}$) | 0.72 ($9 \times 10^{-7}$) |
| | Statistical SelQ. | **0.35 (0.04)** | **0.14 (0.16)** | **0.75 (0.10)** |
| *coherence* | Synthetic Dlg. | 0.48 ($4 \times 10^{-6}$) | 0.30 ($2 \times 10^{-6}$) | 0.42 (0.52) |
| | Chat Style Ref. | 0.43 ($6 \times 10^{-3}$) | **0.16 (0.05)** | **0.22 (0.55)** |
| | Statistical SelZ. | **0.38 (0.04)** | 0.18 (0.04) | 0.41 (0.02) |
| | Statistical SelQ. | 0.40 (0.02) | 0.18 (0.04) | 0.83 (0.05) |

Table 10: Accuracy across datasets (CollabChat, DeliData, and NPSChat) under a $\times 6$ augmentation setting. $C$ denotes the context window size used during training, with a fixed random seed of 42.

| | | CollabChat | | | DeliData | | | NPSChat | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **Method** | Acc@C=1 | C=3 | C=5 | Acc@C=1 | C=3 | C=5 | Acc@C=1 | C=3 | C=5 |
| Original | No Augmentation | 0.47 | 0.46 | 0.46 | 0.43 | 0.77 | 0.43 | 0.75 | 0.52 | 0.73 |
| GPT-based | AugGPT | 0.68 | 0.65 | 0.65 | 0.81 | 0.79 | 0.78 | 0.68 | 0.67 | 0.71 |
| | ConvAug | 0.66 | 0.62 | 0.65 | 0.78 | 0.79 | 0.79 | 0.68 | 0.64 | 0.69 |
| | **Synthetic Dlg.** | 0.67 | 0.66 | **0.66** | 0.79 | 0.79 | **0.81** | 0.65 | 0.69 | 0.52 |
| | **Chat Style Ref.** | 0.69 | 0.67 | **0.66** | **0.81** | 0.80 | 0.79 | 0.71 | 0.72 | 0.72 |
| | **Statistical SelZ.** | **0.71** | **0.68** | 0.65 | 0.80 | **0.81** | **0.81** | 0.74 | **0.73** | 0.52 |
| | **Statistical SelQ.** | **0.70** | **0.70** | 0.63 | 0.80 | 0.80 | 0.79 | **0.76** | **0.73** | **0.74** |

NPSChat shows a decrease in *coherence* from $S_{KS} = 0.42$ to $S_{KS} = 0.22$, while the *naturalness* $p$-value decreases from 0.01 to $8 \times 10^{-4}$. The pattern suggests that Synthetic Dialogue already exhibits a smaller distribution gap for NPSChat compared to CollabChat and DeliData.

Statistical Selection yields additional alignment gains in some datasets. For *naturalness*, CollabChat improves from $S_{KS} = 0.43$ to $S_{KS} = 0.35$ under QSelect, and DeliData improves from $S_{KS} = 0.21$ to $S_{KS} = 0.14$. In contrast, NPSChat increases from $S_{KS} = 0.54$ to $S_{KS} = 0.75$, indicating reduced alignment. Similar trends appear for *coherence*. CollabChat slightly improves from $S_{KS} = 0.43$ to $S_{KS} = 0.38$, while DeliData increases from $S_{KS} = 0.16$ to $S_{KS} = 0.18$ and NPSChat increases from $S_{KS} = 0.22$ to $S_{KS} = 0.41$. DeliData already reaches a statistically similar regime after refinement ($p = 0.05$), limiting further gains from selection. NPSChat exhibits small distribution gaps prior to selection (*naturalness*: $p = 0.01$, *coherence*: $p = 0.52$), so filtering can increase divergence.

### C.3 Comparison Between AlignedAug-Z and AlignedAug-Q

AlignedAug-Z and AlignedAug-Q provide complementary benefits. AlignedAug-Z primarily improves classification performance, while AlignedAug-Q more effectively reduces distributional gaps in *naturalness* and *coherence*. Lower $S_{KS}$ values and higher $p$-values appear more frequently under AlignedAug-Q, indicating closer reproduction of score distributions from real dialogues $D^{real}$. DeliData illustrates the trend clearly: AlignedAug-Z yields $S_{KS} = 0.23$ with $p = 1 \times 10^{-3}$, while AlignedAug-Q yields $S_{KS} = 0.14$ with $p = 0.16$. Quantile-based filtering preserves distribution shape, including tail regions.

AlignedAug-Z aligns synthetic data around the mean and variance of real dialogue distributions using z-score thresholds. The approach can yield slightly larger $S_{KS}$ values in some datasets while removing extreme outliers. Classification performance improves consistently, and stable training behavior is reflected by small variance in accuracy

Table 11: Accuracy across augmentation scales from $\times 1$ to $\times 6$ for each dataset, using a context window size of $C$ = 3 and a fixed random seed of 42.

| | | CollabChat | | | | | | Delidata | | | | | | NPSChat | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **Method** | $\times 1$ | $\times 2$ | $\times 3$ | $\times 4$ | $\times 5$ | $\times 6$ | $\times 1$ | $\times 2$ | $\times 3$ | $\times 4$ | $\times 5$ | $\times 6$ | $\times 1$ | $\times 2$ | $\times 3$ | $\times 4$ | $\times 5$ | $\times 6$ |
| Original | No Aug. | 0.47 | – | – | – | – | – | 0.77 | – | – | – | – | – | 0.52 | – | – | – | – | – |
| GPT-based | AugGPT | **0.67** | **0.67** | 0.67 | 0.65 | 0.65 | 0.65 | 0.75 | 0.74 | 0.77 | 0.76 | 0.75 | 0.79 | 0.59 | 0.70 | 0.64 | 0.72 | 0.65 | 0.67 |
| | ConvAug | 0.65 | 0.64 | 0.65 | 0.60 | 0.64 | 0.62 | 0.78 | 0.79 | 0.80 | 0.79 | 0.78 | 0.79 | 0.70 | 0.73 | 0.72 | 0.72 | 0.69 | 0.64 |
| | **Synthetic Dlg.** | 0.67 | 0.66 | 0.68 | 0.69 | 0.67 | 0.66 | **0.82** | 0.81 | 0.81 | 0.75 | 0.76 | 0.79 | 0.73 | **0.77** | 0.68 | 0.70 | 0.71 | 0.69 |
| | **Chat Style Ref.** | 0.64 | 0.58 | 0.69 | 0.67 | 0.68 | 0.67 | 0.81 | **0.84** | 0.82 | 0.82 | 0.81 | 0.80 | **0.74** | 0.74 | **0.75** | **0.74** | 0.70 | 0.72 |
| | **Statistical SelZ.** | 0.63 | 0.65 | 0.64 | **0.67** | **0.70** | 0.68 | 0.79 | **0.84** | **0.83** | **0.84** | **0.83** | **0.81** | 0.55 | 0.52 | 0.68 | 0.68 | 0.52 | **0.73** |
| | **Statistical SelQ.** | 0.49 | 0.64 | **0.68** | **0.69** | **0.69** | **0.70** | 0.79 | 0.82 | 0.84 | 0.82 | **0.83** | 0.80 | 0.52 | 0.56 | 0.73 | 0.72 | **0.72** | **0.73** |

(CollabChat 0.690, NPSChat 0.741).

### C.4 Case Study

| Model | Response |
|---|---|
| NoAugment. | maybe just move on, we will get result for task 2 |
| ConvAug | let's just proceed, we'll see the outcome for task 2. |
| AugGPT | Perhaps we should proceed; the results for task 2 will come. |
| AlignedAug | Perhaps we should continue the results for task 2 will come |

Table 12: Example responses generated by baselines and AlignedAug.

Generation examples show that AlignedAug balances meaning preservation with chat-style realism. NoAugment responses exhibit properties common in conversational data, including omitted subjects, loosely connected clauses, and missing articles. ConvAug and AugGPT produce well-formed sentence structures, which can increase *naturalness* scores. AlignedAug uses vocabulary similar to AugGPT while retaining slightly unpolished clause connections. The output avoids overly refined style and preserves imperfections closer to real-world chat.

### C.5 Classification Accuracy by Context Size

Classification accuracy varies with context window size used during training. Table 10 reports results for context sizes $C \in \{1, 3, 5\}$. CollabChat achieves the best performance at $C = 1$, where AlignedAug-Z reaches 0.71 with a fixed seed of 42. DeliData shows limited sensitivity to context size. Chat Style Refinement reaches 0.81 at $C = 1$, while Statistical Selection (Z) reaches 0.81 at $C = 3$ and $C = 5$. NPSChat shows inconsistent trends across context sizes, but Statistical Selection (Q) achieves the best performance of 0.76 at $C = 1$. Overall, AlignedAug-based methods outperform baselines in most settings (bolded). In the main experiments, the context window size was set to 5.

### C.6 Classification Accuracy by Augmentation Scale

Augmentation scale does not monotonically improve performance. Table 11 reports accuracy across augmentation scales $\times 1$ to $\times 6$ with $C = 3$. CollabChat baselines (AugGPT, ConvAug) achieve peak performance at $\times 3$ or $\times 4$. Synthetic Dialogue and Chat Style Refinement also peak at $\times 3$ or $\times 4$. AlignedAug-Z and AlignedAug-Q achieve peak performance at $\times 5$ or $\times 6$. DeliData and NPSChat show non-monotonic trends. Results suggest that sufficient augmentation scale is required for AlignedAug to show consistent advantages, especially when distribution gaps are small.