DEALT: LLM-driven Diversity-Enhanced Data Augmentation for Long-Tail Text Classification

Wayne Lu[1*], **Xiaoxi Cui**[2*†],
[1]Independent Researcher
[2]Takway.AI, Beijing, China
lu.wayne0603@gmail.com, cxxneu@163.com

Abstract

Real-world text classification datasets frequently exhibit long-tail distributions, where numerous classes have sparse data, significantly degrading model performance on these underrepresented categories. While Large Language Models (LLMs) offer promise for data augmentation, existing methods often produce semantically limited samples, neglect "implicit long-tails" (sparse sub-patterns within classes), and lack cost-effective optimization. To address these challenges, we propose **DEALT (LLM-driven Diversity-Enhanced Data Augmentation for Long-Tail Text Classification)**, a novel cognitive-inspired framework emulating the human learning process of "recognize, explore, generate, and optimize." DEALT systematically enhances augmented data diversity by first detecting both explicit and implicit long-tails. It then employs an LLM for diversity-aware planning of augmentation strategies, followed by conditional generation. A low-overhead quality and diversity validator filters the synthetic data, and an adaptive incremental sampler refines future augmentation efforts based on proxy model feedback, ensuring efficient and budget-aware optimization. Extensive experiments on multiple public text classification datasets demonstrate DEALT's superiority over state-of-the-art methods in improving tail-class performance and overall model robustness by generating more diverse and high-fidelity augmented data.

Introduction

Text classification is a fundamental task in Natural Language Processing (NLP), crucial for diverse applications like sentiment analysis (Wankhade, Rao, and Kulkarni 2022) and intention recognition (Weld et al. 2022). However, real-world datasets frequently exhibit a severe long-tail distribution (Audibert, Gauffre, and Amini 2024), where a small number of "head" classes dominate, leaving numerous "tail" classes with sparse data. This sparsity significantly degrades the generalization capability and accuracy of deep learning models on underrepresented classes (He et al. 2024; Jia et al. 2024). Data Augmentation (DA) is a pivotal strategy to mitigate this by generating synthetic samples for sparse classes, aiming to rebalance the data distribution and improve model robustness.

While traditional DA methods, such as edit-based techniques (Shorten, Khoshgoftaar, and Furht 2021) or Back-translation (Sugiyama and Yoshinaga 2019), are computationally simple, they often produce synthetic samples that closely resemble the original instances, lacking true semantic diversity. This limits their ability to introduce novel variations. Large Language Models (LLMs) (Radford et al. 2018; Touvron et al. 2023), with their remarkable text understanding and generation prowess, have opened new frontiers for text data augmentation (Dai et al. 2025; Wang et al. 2024). Despite progress, existing LLM-based methods still face significant challenges:

- **Semantic Diversity Bottleneck**: Many LLM augmentation methods tend to generate "neighboring" samples that offer only minor variations in phrasing or syntax, largely overlapping in core semantics (Li et al. 2024a). Such low-diversity augmented data struggles to introduce truly novel semantic information or cover distinct sub-concepts within a class, thus offering limited improvement to tail-class generalization. Effective augmentation critically requires generating *semantically diverse* samples, distinct from originals yet label-consistent, to expand class boundaries.
- **Neglect of Implicit Long-Tail Patterns**: Most research focuses on the explicit imbalance between class sample counts (Ding et al. 2024). However, even within seemingly data-rich classes, specific sub-patterns or themes may be extremely sparse, forming an "implicit long-tail". This fine-grained sparsity is a significant learning bottleneck that existing methods often overlook.
- **High Cost and Lack of Guided Optimization**: Current LLM-based DA often relies on heuristic prompting (Dai et al. 2025; Zhang et al. 2025), lacking fine-grained, interpretable guidance for diversity dimensions. Validating the generated data's utility typically necessitates expensive LLM API calls and costly, full downstream model retraining. This trial-and-error process is inefficient and hinders the development of a systematic optimization loop.

To address these limitations, we propose a novel cognitive-inspired framework for long-tailed text classifica-

[*]These authors contributed equally.
[†]Corresponding author.

tion: LLM-driven **D**iversity-**E**nhanced Data **A**ugmentation for **L**ong-**T**ail Text Classification (**DEALT**). Inspired by the human cognitive process of "recognize, explore, generate, and optimize" when learning new concepts (Binz and Schulz 2023; Guo et al. 2024), DEALT systematically enhances the diversity and effectiveness of augmented data while significantly reducing operational overhead.

`DEALT` achieves this through a suite of innovative and synergistic modules. It first employs a `Long-tail Distribution Detector` to precisely identify explicit and implicit data sparsity. Building on this, a `Diversity Planning` module leverages LLM reasoning to devise varied augmentation strategies with specific diversity objectives. A `Conditional Diversity Generation` component then faithfully executes these plans to create initial samples. These raw samples are then filtered by a low-overhead `Quality and Diversity Validator` using rapid checks and targeted LLM assessment. Finally, an `Adaptive Incremental Sampler and Evaluator` uses feedback from proxy evaluations to refine future augmentation efforts, ensuring efficient, budget-aware optimization.

The main contributions of this paper are as follows:

- We propose `DEALT`, a novel cognitive-inspired framework emulating the "recognize-explore-generate-optimize" process. It systematically tackles both explicit and implicit long-tail issues, significantly enhancing augmented data diversity. To our knowledge, this is the first work to explicitly apply this cognitive-inspired pipeline to LLM-driven long-tail data augmentation.
- We introduce the `Diversity Planning` module, which innovatively employs LLM's Chain-of-Thought capabilities for explicit, structured planning of augmentation strategies based on predefined diversity dimensions, moving beyond heuristic or random approaches.
- We design low-overhead `Quality and Diversity Validator` and `Adaptive Incremental Sampler and Evaluator` mechanisms. These ensure high-quality and diverse augmented data generation while dramatically reducing LLM API calls and dependence on full downstream model retraining.
- Extensive experiments on multiple public text classification datasets demonstrate the superiority of `DEALT` over state-of-the-art methods, particularly in improving tail-class performance and overall model robustness.

Related Work

Data augmentation has long addressed data scarcity in NLP, from early EDA (Wei and Zou 2019) and back-translation (Tiedemann and Thottingal 2020) to more contextualized schemes using BERT (Devlin et al. 2019) and GPT-3 prompting (Brown et al. 2020), which mitigate distortion (Zhang, Zhao, and LeCun 2015). Its development parallels advances in in-context learning, where multimodal and efficient ICL frameworks such as TACO (Li et al. 2025g), M^{2}IV (Li et al. 2025b), CATP (Li et al. 2025e), and CAMA (Li et al. 2025f) emphasize prompt structuring and token-efficient representations. LLM-driven augmentation methods—including AugGPT (Dai et al. 2025), DoAug (Wang et al. 2025), and DPO-enhanced pipelines (Rafailov et al. 2023)—expand sample diversity, complemented by automatic instruction creation (Self-LLMDA (Li et al. 2024b)), class-separability prompting (TARDiS (Kim et al. 2025)), and fully zero-shot synthesis (ZeroGen (Ye et al. 2022)).

Broader data-centric research likewise seeks robustness and controllability: in misinformation detection, multi-view and multimodal frameworks such as URS (Zeng et al. 2025a), IMOL (Zeng et al. 2025b), debiasing (Zeng et al. 2024), and long-tail GNN modeling (Zhang, Zhang, and Yuan 2024) improve generalization under shift; in search ranking, mixture-of-experts systems (Li et al. 2025h,i), semi-supervised approaches (Li et al. 2025c,d), satisfaction-driven objectives (Li et al. 2025a), and LLM-based multi-agent retrieval (Chen et al. 2025) optimize data utilization. Complementary efforts on data unlearning—including multi-objective methods (Li et al. 2025j), error-decomposition (Li et al. 2023a), post-training erasure (Chen et al. 2024), and user-indistinguishability (Li et al. 2023b), summarized in (Li et al. 2024c)—highlight the importance of principled data manipulation. Our work aligns with this trajectory but focuses on automated, controllable, and efficient LLM-based augmentation.

Methodology

In this section, we present our proposed LLM-driven **D**iversity-**E**nhanced Data **A**ugmentation for **L**ong-**T**ail Text Classification (**DEALT**). This framework is designed to enhance text classification datasets by systematically generating diverse and high-quality samples for underrepresented (long-tail) classes, leveraging the advanced capabilities of Large Language Models. Figure 1 illustrates the overall workflow of DEALT. We will now elaborate on each of its five core components: the Long-tail Distribution Detector (LDT), Diversity Planning (DP), Conditional Diversity Generation (CDG), Quality and Diversity Validator (QDV), and Adaptive Incremental Sampler and Evaluator (AISE).

Long-tail Distribution Detector (LDT)

The primary purpose of the LDT is to automatically identify both explicit long-tail classes (globally underrepresented) and implicit long-tail patterns (semantically distinct, sparse sub-groups within any class) from the original training dataset $\mathcal{D}_{\text{train}}$. This targeted identification ensures that subsequent augmentation efforts are focused on the most critical areas needing data enhancement for improving model robustness on rare cases.

To achieve this, as shown in Figure 1(a), the LDT first identifies explicit long-tail classes $\mathcal{Y}_{\text{tail}}^{\text{exp}}$ based on a predefined sample count threshold k_{explicit} (set, for example, to identify classes with fewer samples than the 25th percentile of class sizes in our experiments). For identifying implicit long-tail patterns $\mathcal{C}_{\text{tail}}^{\text{imp}}$ within each class, we employ a two-step process. First, samples x_i belonging to a given class are mapped to a high-dimensional embedding space using a pre-trained sentence encoder, specifically Sentence-BERT,

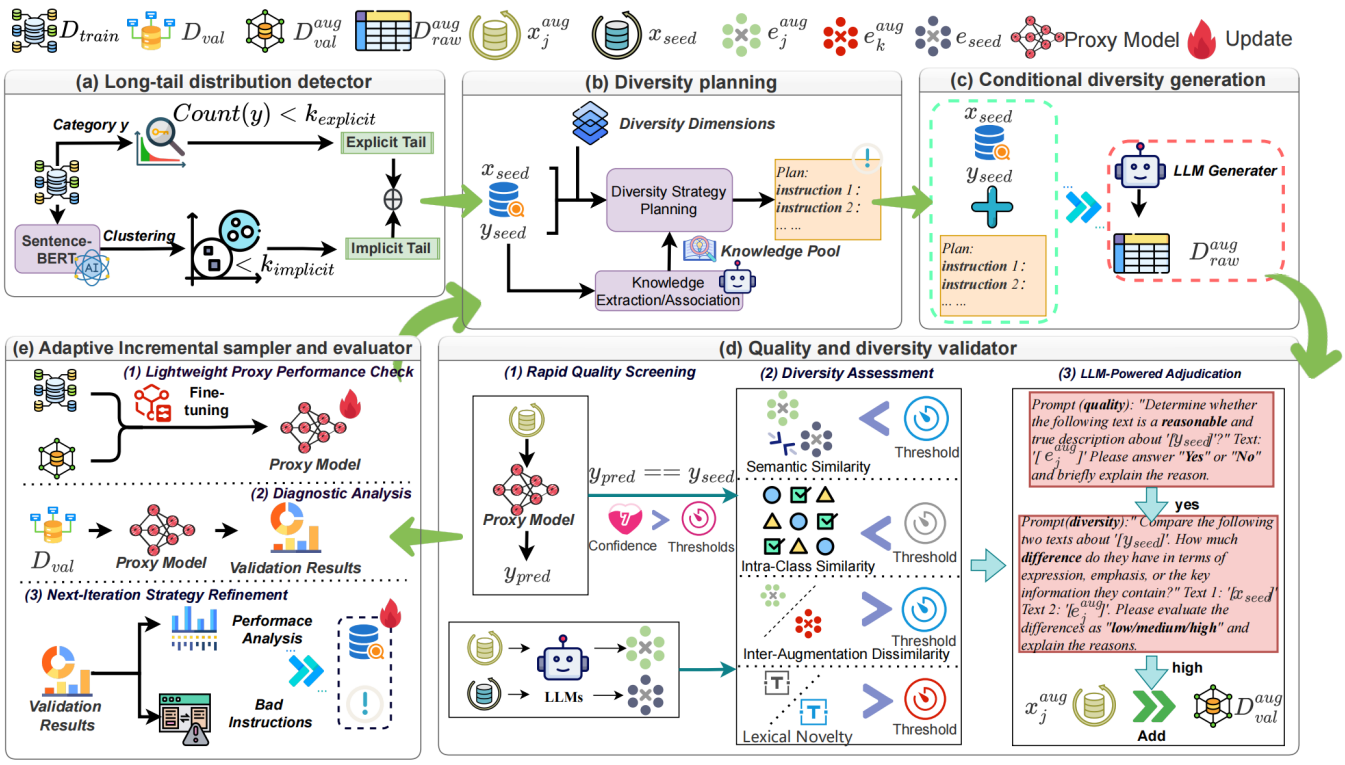


Figure 1: Overview of the proposed DEALT framework: (a) The Long-tail distribution detector identifies explicit and implicit tail samples from the training data. (b) Diversity planning leverages seed data and knowledge to create diverse generation instructions. (c) Conditional diversity generation uses an LLM to produce raw augmented samples based on the plan. (d) The Quality and diversity validator filters samples through rapid screening, metric-based diversity assessment, and LLM-powered adjudication. (e) The Adaptive incremental sampler and evaluator closes the loop by assessing proxy model performance, analyzing results, and refining the strategy.

which maps x_i to an embedding $\mathbf{e}_i$. Second, the DBSCAN clustering algorithm is applied to these class-conditional embeddings. DBSCAN is selected for its ability to discover clusters of arbitrary shapes and to identify noise points, making it suitable for uncovering nuanced sub-patterns. A cluster c discovered by DBSCAN is identified as an "implicit long-tail" if its size (number of samples) is less than a predefined threshold k_{implicit}. In our implementation, k_{implicit} is set to 5, targeting very sparse sub-patterns, though this value is adjustable based on dataset characteristics.

The target samples $\mathcal{S}_{\text{target}}$ for augmentation are then aggregated from both explicit and implicit long-tails:

$$\mathcal{S}_{\text{target}} = \bigcup_{y \in \mathcal{Y}_{\text{tail}}^{\text{exp}}} \{x_i | y_i = y\} \cup \bigcup_{c \in \mathcal{C}_{\text{tail}}^{\text{imp}}} \text{Select}_{\text{rep}}(c, k_{\text{rep}}). \quad (1)$$

The function $\text{Select}_{\text{rep}}(c, k_{\text{rep}})$ selects k_{rep} representative samples from each identified implicit long-tail cluster c. In our framework, k_{rep} is set to 1, and the selected sample is the one closest to the geometric median of the cluster, providing a robust representation for potentially non-convex clusters. The DBSCAN 'eps' parameter, defining the neighborhood distance, is individually tuned for each dataset by analyzing its embedding space using k-distance plots to find an optimal value, typically falling within the 0.3 to 0.5 range for the normalized embeddings used. The 'min_samples' parameter for DBSCAN is set to 3.

Diversity Planning (DP)

The goal of the Diversity Planning (DP) module is to transform a single seed sample $x_{\text{seed}} \in \mathcal{S}_{\text{target}}$ into a structured plan $\mathcal{P}_{\text{plan}}$ containing N distinct, high-quality generation instructions. This process utilizes an LLM's Chain-of-Thought (CoT) capability (Wei et al. 2022; Kojima et al. 2022) to explicitly plan diverse augmentation strategies, incorporating domain knowledge from $\mathcal{K}_{\text{pool}}$ and thereby moving beyond simple paraphrasing or predefined rule-based operations.

As depicted in Figure 1(b), the knowledge pool $\mathcal{K}_{\text{pool}}$ is first dynamically constructed: the LLM is prompted to list concepts and terms related to the seed sample x_{seed} and its label y_{seed}. Subsequently, the planner LLM for the DP module, denoted LLM_{CoT}, is engaged using a CoT prompt. This prompt directs the LLM_{CoT} to first analyze x_{seed} along predefined diversity dimensions $\mathbb{D}$ (which, in our implementation, include lexical substitution, syntactic restructuring, and

semantic scenario alteration). It then utilizes the generated $\mathcal{K}_{\text{pool}}$ to propose concrete modifications along these dimensions, ensuring the N (set to 3 in our experiments) resulting instructions in $\mathcal{P}_{\text{plan}}$ are designed to be distinct and actionable for the subsequent generation phase:

$$\mathcal{P}_{\text{plan}} = \text{LLM}_{\text{CoT}}(x_{\text{seed}}, y_{\text{seed}}, \mathcal{K}_{\text{pool}}, \mathbb{D}, N). \quad (2)$$

where LLM_{CoT} operates in CoT mode, $x_{\text{seed}}, y_{\text{seed}}$ are the seed sample and label, $\mathcal{K}_{\text{pool}}$ provides relevant concepts, $\mathbb{D}$ guides the planning within the DP module, and N dictates the number of instructions generated.

Conditional Diversity Generation (CDG)

The CDG aims to faithfully execute the generation instructions from $\mathcal{P}_{\text{plan}}$ to produce diverse augmented text samples $x_{\text{j}}^{\text{aug}}$. The objective is to ensure the LLM's generation is precisely guided by the planned strategy, translating it into concrete textual variations that are consistent with the original label y_{seed}.

As shown in Figure 1(c), for each instruction $inst_j \in \mathcal{P}_{\text{plan}}$ for seed x_{seed} (label y_{seed}), we employ a generative LLM $\mathcal{L}_{\text{gen}}$ (can share the same base model as LLM_{CoT} but used with a generation prompt). A carefully formatted prompt is constructed before calling the LLM:

$$x_{\text{j}}^{\text{aug}} = \mathcal{L}_{\text{gen}}(\text{Prompt}_{\text{gen}}(x_{\text{seed}}, y_{\text{seed}}, inst_j)). \quad (3)$$

where $\text{Prompt}_{\text{gen}}(\cdot)$ constructs the input string for $\mathcal{L}_{\text{gen}}$, explicitly including the original sample, label, and instruction, "`Original (label: [`y_{seed}`]): '[`x_{seed}`]'. Instruction: '[`$inst_j$`]'. Generate new sample per instruction, valid for '[`y_{seed}`]'.`" The resulting raw samples form $\mathcal{D}_{\text{raw}}^{\text{aug}}$.

Quality and Diversity Validator (QDV)

The purpose of the QDV is to rigorously assess each generated sample $x_{\text{j}}^{\text{aug}} \in \mathcal{D}_{\text{raw}}^{\text{aug}}$ for semantic quality (label fidelity, plausibility) and meaningful diversity (novelty against x_{seed} and other accepted samples $\mathcal{D}_{\text{val}}^{\text{aug}}$). A cost-effective, hierarchical validation pipeline combining automated checks with targeted LLM judgments is employed.

As illustrated in Figure 1(d), validation starts with a proxy model $\mathcal{M}_{\text{proxy}}$ checking label consistency ($y_{\text{pred}} = y_{\text{seed}}$). Diversity is measured using $\text{Diversity}(x_{\text{j}}^{\text{aug}}, x_{\text{seed}}, \mathcal{D}_{\text{val}}^{\text{aug}} \setminus \{x_{\text{j}}^{\text{aug}}\})$, a composite score based on embedding distance (cosine distance from x_{seed}'s embedding $\mathbf{e}_{\text{seed}}$) and lexical novelty (1-BLEU). LLM validation LLM_{val} is triggered if proxy confidence is below τ_{conf} or diversity is below τ_{div}:

$$\text{Validate}(x_{\text{j}}^{\text{aug}}, x_{\text{seed}}, \mathcal{D}_{\text{train}}) \rightarrow \{\text{accept, reject}\}. \quad (4)$$

where a sample is accepted if it passes initial checks with high confidence/diversity, or if it passes a triggered LLM_{val} assessment. Thresholds $\tau_{\text{conf}}, \tau_{\text{div}}$ are hyperparameters. Accepted samples form $\mathcal{D}_{\text{val}}^{\text{aug}}$.

Adaptive Incremental Sampler and Evaluator (AISE)

The AISE creates an intelligent closed-loop system to dynamically steer augmentation by selecting impactful samples for the next round and evaluating utility via proxy metrics, optimizing the process iteratively under budget constraints.

As shown in Figure 1(e), validated samples $\mathcal{D}_{\text{val}}^{\text{aug}}$ update a temporary training set $\mathcal{D}'_{\text{train}}$. The proxy $\mathcal{M}_{\text{proxy}}$ is retrained/evaluated on $\mathcal{D}_{\text{val}}$. The Adapt function uses performance change (ΔPerf), error analysis (ErrAnalysis from *Logs*), and remaining budget ($\text{Budget}_{\text{rem}}$) to determine the next targets $\mathcal{S}'_{\text{target}}$ and updated parameters Ψ:

$$(\mathcal{S}'_{\text{target}}, \Psi) = \text{Adapt}(\Delta\text{Perf}_{\mathcal{M}_{\text{proxy}}}(\mathcal{D}_{\text{val}}), \text{ErrAnalysis}, \text{Budget}_{\text{rem}}). \quad (5)$$

where $\text{Adapt}(\cdot)$ prioritizes $\mathcal{S}'_{\text{target}}$ from classes/clusters with poor ΔPerf or high error rates, using uncertainty sampling via $\mathcal{M}_{\text{proxy}}$. Ψ might adjust $\mathbb{D}$ weights based on successful instruction types noted in *Logs*. Iteration stops when $\text{Budget}_{\text{rem}}$ is zero or ΔPerf plateaus below tolerance ϵ. If the iteration continues, the newly determined $\mathcal{S}'_{\text{target}}$ and updated parameters Ψ are then fed back to the Diversity Planning (DP) module to initiate the next round of augmentation strategy formulation.

Experiments

To comprehensively evaluate the proposed `DEALT` framework, we conduct a series of experiments on several public text classification datasets, covering various task types and long-tail distribution characteristics. We aim to assess `DEALT`'s effectiveness in improving long-tailed classification performance, its ability to generate diverse and faithful augmentations, and the contributions of its core components.

Experimental Setup

Datasets. We selected six English text classification datasets: AG News (Zhang, Zhao, and LeCun 2015), Yelp Reviews Polarity (Zhang, Zhao, and LeCun 2015), CLINC150 (Larson et al. 2019), HWU64 (Casanueva et al. 2020), TREC (Li and Roth 2002), and Symptoms (Dai et al. 2025). Long-tailed AG News (IF=100) and Yelp (IF=50) versions were created via exponential decay sampling, a common strategy in long-tail research (Wang et al. 2024). CLINC150 (Larson et al. 2019) and HWU64 (Casanueva et al. 2020) were used in standard 10-shot settings, relevant to few-shot studies like (Kim et al. 2025). Low-resource TREC (Li and Roth 2002) (20-shot) and Symptoms (Dai et al. 2025) scenarios were simulated, aligning with low-resource augmentation practices (Choi et al. 2024). All datasets underwent standard preprocessing. Experimental outline provides further details

Evaluation Metrics. Primary classification metrics include overall Accuracy and Macro-F1. For AG News and Yelp, we also report per-class F1 scores for head, medium, and tail categories. Augmentation diversity is measured by semantic dissimilarity (average cosine dissimilarity using `all-MiniLM-L6-v2` (Reimers and Gurevych 2019) embeddings) and lexical diversity (distinct-1, distinct-2 n-grams (Li et al. 2015)). Fidelity is assessed by human evaluation of label consistency on a subset of augmented samples,

reporting label consistency rate and Fleiss' Kappa. LLM efficiency is measured by API call counts or total token usage.

Baselines. We compare DEALT against: (1) No Augmentation (Original). (2) Traditional DA: EDA (Wei and Zou 2019) and Back-Translation (Tiedemann and Thottingal 2020). (3) Classical Long-tail Methods: Random Oversampling and Class-balanced Loss (Re-weighting). (4) Simple LLM DA: LLM-Paraphrase (Zero-shot and Few-shot). (5) SOTA LLM DA: AugGPT (reproduced) (Dai et al. 2025), DoAug (core idea reproduced with DPO (Rafailov et al. 2023)) (Wang et al. 2025), Self-LLMDA (instruction set simulated) (Li et al. 2024b), TARDiS (Spark Thoughts simulated) (Kim et al. 2025), and ZeroGen (reproduced) (Ye et al. 2022).

Implementation Details. The DEALT framework primarily utilizes GPT-4o-mini for its LLM-driven components (DP, CDG, and LLM judgments in QDV). The downstream classification task for all experiments and baselines employed a `bert-base-cased` model (Devlin et al. 2019) fine-tuned on the respective training sets (original or augmented). For `bert-base-cased`, we used a learning rate of 2×10^{-5}, a batch size of 16, and trained for 5 epochs with early stopping based on validation performance. The maximum sequence length was set to 256 tokens. Specific configurations for DEALT's components and LLM prompting are detailed in Appendix. All experiments were run 3 times with different random seeds, and we report mean performance with standard deviation.

Main Results: Performance on Long-Tailed Classification

Table 1 presents the primary classification results, including both Macro-F1 (F1) and Accuracy (Acc) scores, of DEALT compared to baseline methods across all six datasets. DEALT consistently outperforms all baseline methods across most datasets and metrics. For instance, on AG News (IF=100), DEALT achieves a Macro-F1 of 78.52% and an Accuracy of 86.89%. This is a significant improvement over No Augmentation (F1 68.73%, Acc 79.52%) and even SOTA LLM DA methods like DoAug (Reproduced) (F1 77.58%, Acc 84.71%) and TARDiS (Sim.) (F1 77.58%, Acc 84.98%). Similar strong trends for DEALT are observed on Yelp Polarity (IF=50), where it achieves F1 55.05% and Acc 58.07%.

For the naturally fine-grained and implicitly long-tailed datasets CLINC150 and HWU64, DEALT demonstrates substantial gains. On CLINC150 (10-shot), DEALT achieves a Macro-F1 of 86.33% (Acc 86.57%), compared to 85.42% F1 (Acc 85.55%) for DoAug (Reproduced), the second-best on this F1 metric, and 85.35% F1 (Acc 85.76%) for TARDiS (Sim.). This highlights DEALT's capability in handling scenarios with sparse data per class and generating discriminative samples. On the low-resource TREC and Symptoms datasets, DEALT also shows strong performance, indicating its robustness in diverse low-data regimes. For TREC, DEALT reaches F1 92.58% and Acc 92.59%. For Symptoms, it scores F1 77.80% and Acc 77.58%. The consistent improvements underscore DEALT's effectiveness in leveraging LLMs for targeted and diversity-aware data augmentation in long-tailed settings.

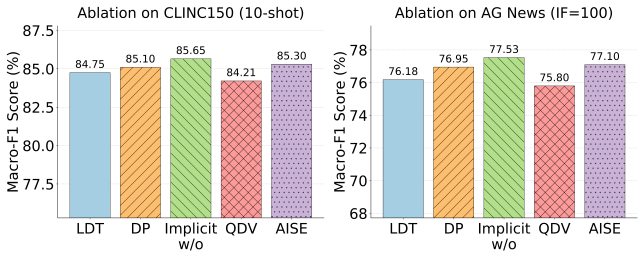


Figure 2: Ablation study results (Macro-F1 %). DEALT w/o X means component X is removed. w/o Implicit means Implicit long-tail handling in LDT is removed.

Specifically, the tail-class performance on AG News and Yelp sees notable improvements with DEALT. For example, on AG News, the average F1 for tail classes improves from 29.35% (No Aug.) to 58.10% (DEALT), demonstrating its ability to effectively augment under-represented categories. This suggests that the diversity-aware strategies and implicit long-tail handling in DEALT are crucial for boosting tail-class recognition. (Note: You might need to adjust the text description if the H/M/T values are no longer explicitly in the table, or refer to supplementary material for those details.)

Ablation Studies

To understand the contribution of each component in DEALT, we conduct ablation studies on the CLINC150 and AG News (IF=100) datasets. The results are presented in Figure 2.

Removing the LDT leads to a significant performance drop (1.58% on CLINC150, 2.34% on AG News). This underscores the importance of identifying both explicit and implicit long-tails and selecting representative seeds. Disabling the DP results in a Macro-F1 decrease of 1.23% on CLINC150 and 1.57% on AG News, highlighting the benefit of dynamically generating diverse augmentation strategies. Excluding the explicit handling of *implicit* long-tails causes a drop, particularly on CLINC150 (0.68%), confirming the value of addressing class-internal imbalances. Operating without the QDV degrades performance by 2.12% and 2.72% on CLINC150 and AG News respectively, showing the necessity of a quality control mechanism. Finally, removing the AISE (single-round augmentation) shows a notable drop, with performance decreasing by 1.03% on CLINC150. This indicates that iterative refinement and budget-aware augmentation are key to DEALT's success. These ablations collectively demonstrate that all components of DEALT contribute positively to its overall performance, with LDT and QDV showing particularly strong impact.

Diversity and Fidelity Evaluation

We evaluate the diversity of augmented data generated by DEALT and baselines on the Symptoms dataset. Results are

Table 1: Main Results: Macro-F1 / Accuracy (%) scores on long-tailed text classification datasets. Best F1 results per dataset are in **bold**, second best are underlined; corresponding Acc scores are highlighted similarly.

		Traditional DA		*Classical Long-tail*		*Simple LLM DA*		*SOTA LLM DA*					
Dataset (Setting)	No Aug. F1/Acc	EDA F1/Acc	Back-Tr. F1/Acc	Rand. OS F1/Acc	CB Loss F1/Acc	LLM-P (0) F1/Acc	LLM-P (F) F1/Acc	AugGPT F1/Acc	DoAug F1/Acc	Self-LLM F1/Acc	TARDiS F1/Acc	ZeroGen F1/Acc	DEALT F1/Acc
AG News	68.73	70.87	70.05	72.10	71.38	74.65	75.38	76.20	77.58	75.72	77.58	73.81	**78.52**
(IF=100)	79.52	80.31	80.01	81.28	80.93	82.83	83.32	84.05	84.71	83.64	84.98	82.11	**86.89**
Yelp Pol.	50.68	51.69	50.93	51.94	52.35	52.65	52.95	53.73	54.78	53.35	54.15	53.75	**55.05**
(IF=50)	51.67	52.45	52.11	53.37	53.06	54.70	55.26	55.84	56.25	55.53	56.44	54.03	**58.07**
CLINC150	81.92	82.53	82.24	82.01	81.95	83.71	84.02	84.53	85.42	84.80	85.35	83.22	**86.33**
(10-shot)	82.51	83.12	82.99	82.73	82.65	84.24	84.56	85.01	85.55	85.24	85.76	83.86	**86.57**
HWU64	78.21	78.82	78.53	78.31	78.25	80.03	80.35	80.81	81.50	81.03	81.77	79.45	**82.95**
(10-shot)	78.91	79.52	79.25	79.04	78.99	80.53	80.82	81.33	81.96	81.06	82.14	80.07	**83.03**
TREC	87.93	88.61	88.35	88.11	87.99	89.70	90.12	90.60	91.13	90.42	91.35	89.30	**92.58**
(20-shot)	88.57	89.22	89.05	88.86	88.54	90.18	90.59	91.03	91.54	90.80	91.77	89.88	**92.59**
Symptoms	69.50	70.82	70.21	70.03	69.88	72.91	73.40	74.55	75.41	74.10	75.12	72.15	**77.80**
(15-shot)	70.23	71.57	71.05	70.83	70.53	73.55	74.03	75.16	76.02	74.75	76.39	72.89	**77.58**

Table 2: Diversity and Fidelity on the Symptoms dataset. Higher semantic dissimilarity, distinct-1/2 are better. Higher label consistency is better.

Method	Semantic Dissim. (Seed-Aug / Aug-Aug)	Distinct-1	Distinct-2	Label Consist. (% / Kappa)
Original (ref)	- / 0.37	0.077	0.264	-
EDA	0.18 / 0.22	0.089	0.295	91.2% / 0.83
Back-Translation	0.22 / 0.25	0.083	0.288	93.5% / 0.84
LLM-Para. (0-shot)	0.25 / 0.33	0.089	0.345	94.8% / 0.88
AugGPT (Rep.)	0.28 / 0.38	0.098	0.375	95.5% / 0.90
DoAug (Rep.)	0.32 / 0.41	0.093	0.361	96.1% / 0.91
DEALT (Ours)	**0.38 / 0.49**	**0.105**	**0.421**	**97.2% / 0.93**

shown in Table 2. (This table did not contain \pm values initially, so it remains largely unchanged in content.)

DEALT achieves the highest semantic dissimilarity scores, both between seed-augmented pairs (0.38) and within the augmented set itself (0.49). This indicates that DEALT generates paraphrases that are meaningfully different from the original seeds and also diverse among themselves. It also scores highest on lexical diversity. Crucially, this diversity does not come at the cost of semantic integrity. Human evaluation shows a label consistency rate of 97.2% for DEALT, with a Fleiss' Kappa of 0.93, indicating substantial agreement among annotators and high fidelity of the augmented samples. This is comparable to or better than SOTA LLM DA methods like DoAug (Rep.), and significantly better than traditional methods like EDA. These results suggest that DEALT's DP and QDV work effectively to produce varied yet semantically coherent augmentations.

Efficiency and Cost Analysis

To evaluate the practical efficiency of our method, we conducted an empirical analysis of the LLM API calls required for augmenting the TREC dataset. The experiment was configured with 20 shots per class for 6 classes (total 120 seed samples), focusing augmentation on 3 identified tail classes. For each seed, DEALT was set to generate N=3 diversity strategies. We performed 5 independent runs and recorded the actual number of API calls, reporting the mean and standard deviation to capture real-world performance.

Table 3: Empirical LLM API calls for augmenting TREC (20-shot), averaged over 5 runs. Values indicate mean ± standard deviation, reflecting the operational stability of each method.

Method	Observed LLM Calls (Synthesis Only)
LLM-Paraphrase (0-shot, 1 aug/seed)	124 ± 3 calls
AugGPT (Rep., 1 aug/seed)	127 ± 5 calls
DoAug (Rep., 1 aug/seed after DPO)	DPO cost (High) + 122 ± 2 calls
Self-LLMDA (Sim., 1 aug/seed, rand instr.)	126 ± 8 calls
DEALT (Ours, 3 strat/seed, tail focus)	
- DP (strategy gen.)	191 ± 9 calls
- CDG (sample gen.)	63 ± 3 calls
- QDV (LLM adjudication)	8 ± 2 calls
Total DEALT (1st round, tail only)	$\sim$ **262** calls

Table 3 shows the observed API calls, with variance reflecting real-world factors like API retries. Method stability varied as expected: the DPO-tuned DoAug was most reliable (122 ± 2 calls), while the instruction-based Self-LLMDA showed the highest variance (126 ± 8). For our method, DEALT, the creative DP stage was the primary source of variance (191 ± 9), whereas the subsequent CDG stage was highly stable (63 ± 3). Including a few LLM adjudications in QDV (8 ± 2), DEALT's total first-round cost was approximately 262 calls. While higher than one-pass methods, this is far more efficient than the costly DPO fine-tuning required by DoAug. The AISE module further refines this process efficiently in subsequent rounds.

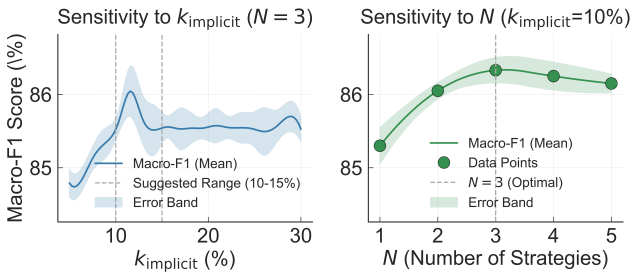
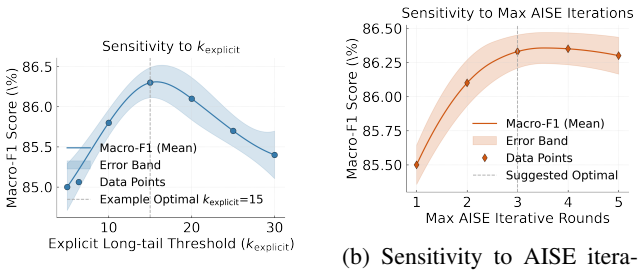


Figure 3: Hyperparameter sensitivity on CLINC150. Left: Varying k_{implicit} (N=3). Right: Varying N (k_{implicit}=10%).



(a) Sensitivity to k_{explicit}.

(b) Sensitivity to AISE iterative rounds.

Figure 4: Sensitivity analyses on CLINC150 (Macro-F1 %).

Analysis of Implicit Long-Tail Handling

To specifically evaluate the impact of handling implicitly long-tailed distributions, we compare the full DEALT framework against a variant where the LDT only identifies explicitly tail classes (based on global sample counts) on the CLINC150 dataset.

As shown in Fig. 2 (row "w/o Implicit LT Handling"), disabling the class-internal clustering and implicit pattern mining results in a Macro-F1 drop from 86.33% to 85.65% on CLINC150. This 0.68% difference highlights that a significant portion of the performance gain on such fine-grained datasets comes from DEALT's ability to identify and augment under-represented sub-clusters or "implicit tails" within seemingly well-represented classes. For example, within a broad "book_flight" intent, DEALT might identify a sub-cluster related to "booking_flights_with_specific_dietary_requirements" which has few examples, and then generate targeted diverse strategies for it. This capability is crucial for robust performance in real-world scenarios where data sparsity can exist at a very granular level.

Hyperparameter Sensitivity Analysis

We analyze the sensitivity of DEALT to two key hyperparameters using experiments on CLINC150: k_{implicit}, the threshold for identifying implicit long-tails (as a percentage of average samples per class), and N, the number of diversity strategies generated per seed sample by the DP. The primary evaluation metric is Macro-F1.

Figure 3 conceptually illustrates these sensitivities. For k_{implicit} (Figure 3, Left, with $N = 3$ fixed), we observe that Macro-F1 performance initially increases as k_{implicit} rises from very small values, indicating that some level of focus on sub-patterns is beneficial. The performance reaches an optimal plateau or peak region when k_{implicit} is approximately between 10% and 15%. Excessively small k_{implicit} values result in suboptimal performance, likely due to an over-focus on minor, possibly noisy, variations within classes. Interestingly, when k_{implicit} increases beyond the optimal range (towards 20-30%), the performance degradation is relatively graceful, showing a slow decline or stabilization slightly below the peak. This suggests that while overly large thresholds might miss some finer-grained implicit tails, they do not drastically impair overall performance.

For the number of strategies N (Figure 3, Right, with k_{implicit} fixed at a representative 10%), Macro-F1 performance demonstrates clear improvement as N increases from 1 to 3. The peak performance is achieved at $N = 3$. When N is further increased to 4 and 5, the gains in Macro-F1 diminish, and a slight plateau or even a marginal decrease is observed. This suggests that while exploring multiple augmentation strategies is beneficial, generating an excessive number of strategies for each seed offers diminishing returns and may introduce redundancy or noise, without a proportional gain in performance, while also increasing computational overhead.

k_{explicit} (LDT Module): This parameter defines the threshold for identifying globally long-tailed classes. As shown in Figure 4a, performance is relatively stable across a range of reasonable k_{explicit} values. Both overly aggressive and conservative thresholds can lead to suboptimal performance. Our default setting (25th percentile of class sizes or minimum count) falls within a robust region.

AISE Module (Iterative Rounds): This module iteratively refines augmentation strategies. Figure 4b indicates that Macro-F1 performance generally improves with an increasing number of AISE rounds, typically up to 3 rounds. Beyond this, gains become marginal, suggesting diminishing returns. Thus, 2-3 rounds of adaptive refinement are often sufficient to achieve effective augmentation while balancing computational overhead. Our default setting uses a maximum of 3 iterations.

Based on this analysis, our default choice of $k_{\text{implicit}} = \max(5, \text{approx. } 10\% \text{ of avg. class size})$ and $N = 3$ is empirically supported, providing a robust balance between performance, diversity, and computational efficiency.

Conclusion

To address long-tailed text classification, we propose DEALT, a cognitive-inspired LLM data augmentation framework. Leveraging a "recognize, explore, generate, optimize" pipeline, DEALT systematically produces high-quality, diverse data tailored to sparse classes. Extensive experiments demonstrate DEALT's superior performance over state-of-the-art methods, significantly improving tail-class recognition and robustness. This is due to its ability to generate label-consistent augmented data, with ablation studies validating each component's importance.

References

Audibert, A.; Gauffre, A.; and Amini, M.-R. 2024. Exploring contrastive learning for long-tailed multi-label text classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 245–261. Springer.

Binz, M.; and Schulz, E. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; and Vulić, I. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Chen, C.; Zhang, Y.; Li, Y.; Wang, J.; Qi, L.; Xu, X.; Zheng, X.; and Yin, J. 2024. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1): 1–28.

Chen, X.; Li, Y.; Cai, H.; Ma, Z.; Chen, X.; Xiong, H.; Wang, S.; He, B.; Sun, L.; and Yin, D. 2025. Multi-Agent Proactive Information Seeking with Adaptive LLM Orchestration for Non-Factoid Question Answering. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 4341–4352.

Choi, J.; Jin, K.; Lee, J.; Song, S.; and Kim, Y. 2024. AutoAugment Is What You Need: Enhancing Rule-based Augmentation Methods in Low-resource Regimes. *arXiv preprint arXiv:2402.05584*.

Dai, H.; Liu, Z.; Liao, W.; Huang, X.; Cao, Y.; Wu, Z.; Zhao, L.; Xu, S.; Zeng, F.; Liu, W.; et al. 2025. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Transactions on Big Data*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Tuan, L. A.; and Joty, S. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, 1679–1705.

Guo, D.; Xu, W.; Ding, W.; Yao, Y.; Wang, X.; Pedrycz, W.; and Qian, Y. 2024. Concept-cognitive learning survey: Mining and fusing knowledge from data. *Information Fusion*, 109: 102426.

He, Y.-C.; Ding, Y.-X.; Ye, H.-J.; and Zhou, Z.-H. 2024. Learning only when it matters: cost-aware long-tailed classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12411–12420.

Jia, Y.; Peng, X.; Wang, R.; and Zhang, M.-L. 2024. Long-tailed partial label learning by head classifier and tail classifier cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12857–12865.

Kim, K.; Im, S.; Kim, G.; and Oh, H.-S. 2025. TARDiS: Text Augmentation for Refining Diversity and Separability. *arXiv preprint arXiv:2501.02739*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, X.; and Roth, D. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Li, Y.; Cai, H.; Kong, R.; Chen, X.; Chen, J.; Yang, J.; Zhang, H.; Li, J.; Wu, J.; Chen, Y.; et al. 2025a. Towards AI Search Paradigm. *arXiv preprint arXiv:2506.17188*.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025b. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Chen, C.; Zhang, Y.; Liu, W.; Lyu, L.; Zheng, X.; Meng, D.; and Wang, J. 2023a. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36: 12611–12625.

Li, Y.; Chen, C.; Zheng, X.; Zhang, Y.; Han, Z.; Meng, D.; and Wang, J. 2023b. Making users indistinguishable: Attribute-wise unlearning in recommender systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 984–994.

Li, Y.; Ding, K.; Wang, J.; and Lee, K. 2024a. Empowering Large Language Models for Textual Data Augmentation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 12734–12751. Bangkok, Thailand: Association for Computational Linguistics.

Li, Y.; Ding, K.; Wang, J.; and Lee, K. 2024b. Empowering large language models for textual data augmentation. *arXiv preprint arXiv:2404.17642*.

Li, Y.; Feng, X.; Chen, C.; and Yang, Q. 2024c. A survey on recommendation unlearning: Fundamentals, taxonomy, evaluation, and open questions. *arXiv preprint arXiv:2412.12836*.

Li, Y.; Lyu, Z.; Zhang, Y.; Zhang, H.; Peng, T.; Xiong, H.; Wang, S.; Kong, L.; Chen, G.; and Yin, D. 2025c. S ˆ{3} PRank: Towards Satisfaction-oriented Learning to Rank with Semi-supervised Pre-training. *IEEE Transactions on Knowledge and Data Engineering*.

Li, Y.; Xiong, H.; Zhang, Y.; Bian, J.; Peng, T.; Li, X.; Wang, S.; Kong, L.; and Yin, D. 2025d. Rankelectra: Semi-supervised pre-training of learning-to-rank electra for web-scale search. In *Proceedings of the 31st ACM SIGKDD*

Conference on Knowledge Discovery and Data Mining V. 1, 2415–2425.

Li, Y.; Yang, J.; Shen, Z.; Han, L.; Xu, H.; and Tang, R. 2025e. CATP: Contextually Adaptive Token Pruning for Efficient and Enhanced Multimodal In-Context Learning. *arXiv preprint arXiv:2508.07871*.

Li, Y.; Yang, J.; Yang, Z.; Li, B.; He, H.; Yao, Z.; Han, L.; Chen, Y. V.; Fei, S.; Liu, D.; et al. 2025f. Cama: Enhancing multimodal in-context learning with context-aware modulated attention. *arXiv preprint arXiv:2505.17097*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025g. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Li, Y.; Zhang, H.; Zhang, Y.; Cai, H.; Cai, M.; Wang, S.; Xiong, H.; Kong, L.; Yin, D.; and Chen, L. 2025h. Rankexpert: A mixture of textual-and-behavioral experts for multi-objective learning-to-rank in web search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 4578–4589.

Li, Y.; Zhang, H.; Zhang, Y.; Ma, X.; Ye, W.; Song, N.; Wang, S.; Xiong, H.; Yin, D.; and Chen, L. 2025i. M 2 oERank: Multi-Objective Mixture-of-Experts Enhanced Ranking for Satisfaction-Oriented Web Search. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 4441–4454. IEEE.

Li, Y.; Zhang, Y.; Liu, W.; Feng, X.; Han, Z.; Chen, C.; and Yan, C. 2025j. Multi-Objective Unlearning in Recommender Systems via Preference Guided Pareto Exploration. *IEEE Transactions on Services Computing*.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shorten, C.; Khoshgoftaar, T. M.; and Furht, B. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1): 101.

Sugiyama, A.; and Yoshinaga, N. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019)*, 35–44.

Tiedemann, J.; and Thottingal, S. 2020. OPUS-MT–Building open translation services for the World. In *Annual Conference of the European Association for Machine Translation*, 479–480. European Association for Machine Translation.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, P.; Zhao, Z.; Wen, H.; Wang, F.; Wang, B.; Zhang, Q.; and Wang, Y. 2024. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. *Advances in Neural Information Processing Systems*, 37: 64915–64941.

Wang, Z.; Zhang, J.; Zhang, X.; Liu, K.; Wang, P.; and Zhou, Y. 2025. Diversity-Oriented Data Augmentation with Large Language Models. *arXiv preprint arXiv:2502.11671*.

Wankhade, M.; Rao, A. C. S.; and Kulkarni, C. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7): 5731–5780.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wei, J.; and Zou, K. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Weld, H.; Huang, X.; Long, S.; Poon, J.; and Han, S. C. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8): 1–38.

Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; and Kong, L. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Zeng, Z.; Luo, M.; Kong, X.; Liu, H.; Guo, H.; Yang, H.; Ma, Z.; and Zhao, X. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6492–6500.

Zeng, Z.; Wu, J.; Luo, M.; Kong, X.; Ma, Z.; Dai, G.; and Zheng, Q. 2025a. Understand, Refine and Summarize: Multi-View Knowledge Progressive Enhancement Learning for Fake News Video Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9216–9225.

Zeng, Z.; Wu, J.; Luo, M.; Wan, H.; Kong, X.; Ma, Z.; Dai, G.; and Zheng, Q. 2025b. IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30921–30933.

Zhang, G.; Zhang, S.; and Yuan, G. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–21.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Y.; Yang, R.; Xu, X.; Li, R.; Xiao, J.; Shen, J.; and Han, J. 2025. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, 2032–2042.