GENAL: IMBALANCE-AWARE GENERATIVE ACTIVE LEARNING FOR DISCRIMINATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard active learning (AL) frameworks for discriminative models struggle with imbalanced datasets due to misjudgment of sample informativeness under class imbalance. Current methods typically rely on algorithmic or sampling-based techniques to address imbalanced distributions; however, they overlook a more intrinsic solution: generative data augmentation. While traditional data augmentation approaches often fail in NLP tasks due to semantic loss, Large Language Models, with remarkable generalization and data synthesis capabilities, provide a promising pathway to mitigate class imbalance and further improve AL performance. In this paper, we introduce GenAL, a generative active learning framework robust across both standard and imbalanced datasets. To specifically address class imbalance, we design an imbalance-aware query strategy (BIRDS), coupled with a generative augmentation framework (IGR) incorporating an automated refinement mechanism to synthesize diverse, high-quality data for AL training. Extensive experiments across both standard and imbalanced settings demonstrate that GenAL and BIRDS achieve superior performance compared to existing query strategies, while IGR consistently delivers performance gains across different baselines.

1 INTRODUCTION

While Large Language Models (LLMs) have revolutionized natural language processing for a rich set of tasks, discriminative tasks such as text classification remain important for many real-world applications (Pecher et al., 2025), where development, deployment, and maintenance demand an efficient and managed system (Di Palo et al., 2024; Gonzalez, 2026). Compact, domain-aligned discriminative models provide a practical alternative, but optimizing them relies heavily on high-quality labeled data (Gruber et al., 2025). Given human annotation is expensive and expertise may be rare, Active Learning (AL) can be deployed to reduce these costs. It assumes a limited labeling budget and iteratively prioritizes the most informative samples for annotation (Ren et al., 2021).

Traditional AL strategies exhibit robust performance under uniform data distributions by quantifying sample informativeness via uncertainty (Houlsby et al., 2011), diversity (Sener & Savarese, 2018), or hybrid approaches (Wu et al., 2025). However, their efficacy often diminishes under severe class imbalances in real-world scenarios (Li et al., 2023). Methods addressing this challenge fall into two paradigms: algorithmic adjustments (Zhang et al., 2023) that assign higher misclassification costs to minority errors, or resampling techniques (Zhu & Hovy, 2007) that artificially rebalance the data distribution. Yet, these approaches often rely on complex auxiliary architectures or strict prior assumptions, inherently complicating scalable deployment (Liu et al., 2021).

Unlike conventional interventions, data augmentation provides an intrinsic alternative by synthesizing diverse minority samples without human annotation costs (Møller et al., 2024; Tran et al., 2025); yet, despite success in computer vision (Moles et al., 2024), its utility in NLP remains constrained by the discrete nature of text (Feng et al., 2021). Leveraging the generative prowess of LLMs as augmentation engines within the AL framework thus provides a promising pathway to overcome these NLP-specific bottlenecks, yet doing so hinges on two critical criteria (Shaib et al., 2024): generative quality to ensure strict adherence to the target class (Nadăș et al., 2025), and semantic diversity to provide rich, non-redundant linguistic variations (Wang et al., 2025). Since labeled data is inherently scarce in AL scenarios, introducing noisy or highly redundant synthetic samples risks severe model degradation rather than performance gains (Shumailov et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Critically, addressing these two challenges in isolation is insufficient: a query strategy that identifies informative minority-class instances offers limited benefit when the available training data remains scarce and imbalanced, while generative augmentation that is agnostic to the evolving class distribution risks reinforcing majority-class dominance rather than correcting it. This motivates a coupled design in which acquisition and generation jointly adapt to the current labeled pool at each AL iteration, rather than being deployed as independent, static components. To this end, we propose the **Gen**erative **A**ctive **L**earning (GenAL) framework. Shifting from a traditional "selection-only" paradigm to a "generation-enhanced" framework, GenAL addresses the limitations of both query strategy and data scarcity. Our key contributions are:

- **Imbalance-aware Acquisition**. We introduce **BIRDS**, Bayesian Imbalance-aware Representative Diversity Sampling, a query strategy robust to standard and imbalanced settings.
- **Generative Data Augmentation**. We develop **IGR**, an LLM-based synthesis pipeline that ensures generative quality and semantic diversity for AL training.
- **Adaptive Coupling Design**. **GenAL** outperforms both its individual components and their static combination, confirming the benefit of iterative information exchange.
- **Empirical Validation**. Across four standard benchmarks and their imbalanced variants, GenAL consistently outperforms strong baselines with statistically significant gains.

2 Related Work

Active learning. AL algorithms typically sample unlabeled instances by optimizing for model uncertainty or data diversity (Aggarwal et al., 2014; Zhan et al., 2022). Uncertainty-oriented strategies, ranging from least-confident sampling (Settles, 2009) to distribution-based methods like Contrastive AL (Margatina et al., 2021), target instances near decision boundaries. Whereas, diversity-based frameworks like CoreSet (Sener & Savarese, 2018) minimize redundancy by constructing geometrically representative subsets. Recent hybrid approaches like BEMPS (Tan et al., 2023) bridge the gap between uncertainty and diverse feature-space coverage. Despite these developments, class imbalance remains a bottleneck in practical scenarios (Qi et al., 2025); standard strategies systematically bias toward majority classes, marginalizing rare but critical categories (Li et al., 2024).

Imbalance-aware active learning. To mitigate acquisition bias under imbalanced distributions, research has pivoted toward imbalance-aware strategies (Ertekin et al., 2007; Aggarwal et al., 2020; Fairstein et al., 2024). For instance, DIRECT (Nuggehalli et al., 2025) solves one-dimensional subproblems to find separation thresholds, while AnchorAL (Lesci & Vlachos, 2024) leverages class anchors to retrieve representative minority samples. However, these methods are often constrained by high computational complexity in multi-class settings or a reliance on prior knowledge of minority labels (Chen et al., 2024b). Consequently, there remains a need for imbalance-aware strategies that function without prior class knowledge or exhaustive computational expenditure.

Data augmentation. Data augmentation is widely adopted in visual active learning (Xu et al., 2023), expanding the training manifold via techniques from Gaussian noise injection to diffusion models (Zhu et al., 2024). In contrast, traditional text augmentation (e.g., paraphrasing) yields limited gains in NLP (Shorten et al., 2021), as minor perturbations on discrete tokens easily induce semantic drift or incoherence (Longpre et al., 2020). While Gebreegziabher et al. (2025) design an AL augmentation pipeline for NLP, their method relies on symbolic pattern-matching rather than LLM synthesis. LLMs, in contrast, can generalize from limited labeled examples to synthesize label-consistent and semantically coherent data spanning a broader contextual range (Moe et al., 2025; Chen et al., 2025); yet, integrating LLM generation into iterative AL remains under-explored.

Data generation evaluation. Evaluating data generation centers on semantic diversity and generative quality (Shypula et al., 2025). Diversity reflects structural and semantic variation, ensuring broad distribution coverage without mode collapse (Bauer et al., 2024); it is commonly quantified via metrics like Distinct-N (Li et al., 2016), Self-BLEU (Zhu et al., 2018) and Sentence-BERT (Zhang et al., 2020). Quality denotes linguistic coherence and class-conditional accuracy of synthetic instances, essential for training set integrity, and is typically assessed via MAUVE (Pillutla et al., 2021), which measures the divergence between synthetic and human-authored text distributions, and the LLM-as-a-judge framework, where high-capacity models act as automated evaluators verifying that synthetic samples align with their intended labels (Zheng et al., 2023).

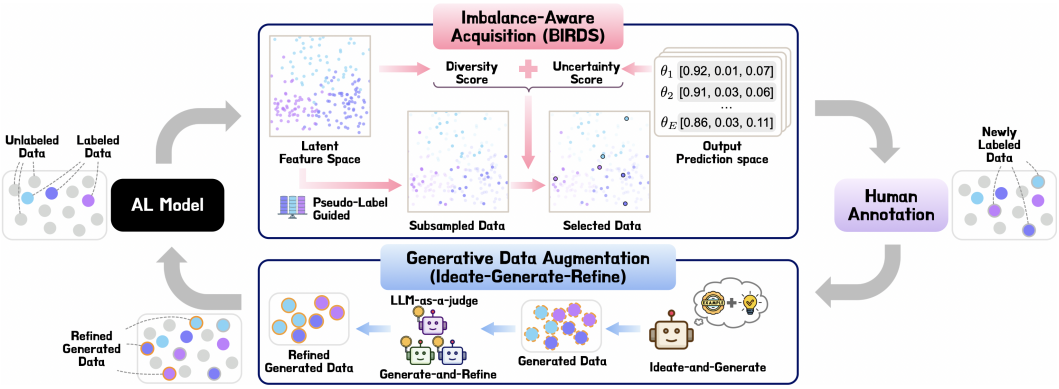


Figure 1: Overview of GenAL framework. The workflow comprises two primary components. (1) BIRDS, an imbalance-aware acquisition strategy that identifies highly informative samples through joint feature and output space analysis, guided by pseudo-label-based subsampling (where θ denotes the ensemble model). (2) Generative Data Augmentation, which utilizes an Ideate-Generate-Refine (IGR) mechanism to expand the training set with diverse and high-quality synthetic instances.

3 METHODOLOGY

Without loss of generality, we consider a pool-based active learning setting comprising a small initial labeled set $\mathcal{L} = \{(x_i, y_i)\}$ and a large unlabeled pool $\mathcal{U} = \{x_j\}$, where $|\mathcal{U}| \gg |\mathcal{L}|$. Let $\mathcal{Y} = \{1, \ldots, K\}$ denote the label space with K classes. For a target model $\theta \in \Theta$, let $p(y|x, \theta)$ be the predictive probability of an instance x. In each iteration, a query budget of B instances is selected for annotation. Additionally, we denote the LLMs employed for generative data augmentation as $\mathcal{M}$.

Figure 1 illustrates the workflow of one AL iteration, which comprises two phases: imbalance-aware acquisition and generative data augmentation. The proposed query strategy, BIRDS, captures both diversity and uncertainty by evaluating the data from two complementary perspectives: the latent feature space and the model's output prediction space. Specifically, guided by pseudo-labels derived from the aggregated predictions of an ensemble model, we first construct a balanced unlabeled subpool. Instances are then selected for annotation based on a composite scoring mechanism. Subsequently, the generative data augmentation module leverages the generalization capabilities of LLMs to expand the training set via the Ideate-Generate-Refine (IGR) mechanism, which is designed to ensure that the synthesized samples possess high generative quality and semantic diversity.

3.1 IMBALANCE-AWARE ACQUISITION

BIRDS is an imbalance-aware query strategy that integrates feature-based diversity with model-based epistemic uncertainty. Below we describe diversity scoring, uncertainty scoring, and hybrid acquisition, followed by two refinements: anchor filtering and imbalance-aware subsampling.

Diversity scoring. To quantify sample diversity, denoted as $\mathcal{S}_D$, we extract representations from the last hidden layer of the model. Let $\mathbf{z}_i = \phi(x_i, \theta)$ denote the last-layer feature embedding of sample x_i given parameters θ. For a candidate unlabeled instance $x_j \in \mathcal{U}$, the diversity score is defined as its minimum Euclidean distance to the instances in the current labeled set $\mathcal{L}$ (Eq. 1). This geometrically captures the sample's relative position on the feature manifold, and a larger minimum distance indicates more non-redundant information.

$$\mathcal{S}_D(x_j) = \min_{x_i \in \mathcal{L}} \|\mathbf{z}_i - \mathbf{z}_j\|_2. \tag{1}$$

Uncertainty scoring. To capture model uncertainty, a Bayesian proper scoring rule framework is employed that evaluates the expected increase in strictly proper scores (Tan et al., 2023). Moving beyond a single scalar evaluation for a candidate sample $x \in \mathcal{U}$, the framework captures its fine-grained, pairwise impact on any given anchor point $x' \in \mathcal{U}$. Specifically, $\mathcal{S}_U(x, x')$ represents the expected change in the proper score at x' resulting from acquiring the label y for x. This pairwise impact is formulated as:

$$\mathcal{S}_U(x, x') = \mathbb{E}_{p(y|\mathcal{L},x)}\left[\mathbb{E}_{p(y'|\mathcal{L}\cup\{(x,y)\},x')}[\mathcal{C}(p(\cdot|\mathcal{L}\cup\{(x,y)\}, x'), y')] - \mathcal{C}(p(\cdot|\mathcal{L}, x'), y')\right], \tag{2}$$

where $\mathcal{C}$ denotes the scoring function (e.g. Log Score) that assesses the quality of the probabilistic prediction of categorical variables for a target y' given an anchor point $x' \in \mathcal{U}$. The more comprehensive representation for the impact of the uncertainty about x on the data space is given by:

$$\mathcal{S}_U(x) = \{\mathcal{S}_U(x, x') \,:\, x' \in \mathcal{U}\}\,, \tag{3}$$

where $\mathcal{S}_U(x)$ is the vector of $\mathcal{S}_U(x, x')$ values evaluated against randomly selected anchor points x' from the full unlabeled data pool (different x'), encapsulating its impact on the mean proper score across the broader data space.

This pairwise impact is estimated as follows: Since the true labels for the candidate x and the anchor x' are unobserved, the expectations are estimated within Eq. (2) with respect to the predictive distribution of the model. To compute these probabilities efficiently and avoid the high retraining cost, Monte-Carlo (MC) Dropout (Gal & Ghahramani, 2016) is employed to construct an approximate Bayesian ensemble. Let $\Theta = \{\theta_1, \ldots, \theta_E\}$ denote the parameters obtained from E stochastic forward passes. The posterior predictive distribution, updated with the hypothetical label y, is approximated via importance sampling:

$$p(y'|\mathcal{L} \cup \{(x, y)\}, x') \approx \frac{1}{E} \sum_{e=1}^{E} p(y'|\theta_e, x') p(\theta_e|\mathcal{L} \cup \{(x, y)\}), \tag{4}$$

$$p(\theta_e|\mathcal{L} \cup \{(x, y)\}) \approx \frac{p(\theta_e|\mathcal{L}) p(y|\theta_e, x)}{\frac{1}{E} \sum_{e=1}^{E} p(\theta_e|\mathcal{L}) p(y|\theta_e, x)}. \tag{5}$$

Hybrid acquisition. For the final acquisition, diversity $\mathcal{S}_D(x)$ and uncertainty $\mathcal{S}_U(x)$ are concatenated into a composite vector $\mathcal{S}(x)$. Both terms are min-max normalized before concatenation, so no further scaling is needed for clustering. K-Means clustering is then applied over these vectors across the unlabeled subpool (Eq. 7). By retaining the sample nearest to each of the B cluster centroids, which represent diverse regions of the joint uncertainty-diversity manifold, we obtain the final query set $\{x^\star\}$. The hybrid acquisition process is formalized as:

$$\mathcal{S}(x) = [\mathcal{S}_D(x), \mathcal{S}'_U(x)]^\top, \tag{6}$$

$$\{x^\star\} = \text{KMeans}(\{\mathcal{S}(x) \mid x \in \mathcal{U}'\}, B). \tag{7}$$

However, we apply two changes to this setup that consistently improve performance in our ablation studies: first, the pool $\mathcal{U}$ used to compute $\mathcal{S}_U(x)$ is reduced to eliminate ineffective anchor points, yielding $\mathcal{S}'_U(x)$; second, stratified sampling is used to handle class imbalance, so K-Means runs over a reduced set $\mathcal{U}' \subset \mathcal{U}$. Both are explained below.

Anchor filtering. To filter out ineffective anchor points when constructing $\mathcal{S}'_U(x)$, we compute, for each anchor x', the range of $\mathcal{S}_U(x, x')$ over candidate samples $x \in \mathcal{U}$ and discard anchors whose range is below 10^{-5}. Such anchors are unlikely to provide meaningful information for uncertainty estimation, as they induce near-constant scores across candidates.

Imbalance-aware subsampling. To address the class imbalance when constructing $\mathcal{U}'$, sampling is budgeted per class by oversampling rare classes. This relies on the ensemble posterior predictive distribution $p(y|x, \mathcal{L}) \approx \frac{1}{E} \sum_{e=1}^{E} p(y|x, \theta_e)$. However, standard $argmax$ predictions risk error propagation on ambiguous instances. To mitigate this, we introduce an auxiliary *hard-to-learn* pseudo-class, denoted as $y_\varnothing$, and define the pseudo-label $\hat{y}$ using a confidence threshold τ:

$$\hat{y} = \begin{cases} \arg\max_k p(y = k|x, \mathcal{L}) & \text{if } \max_k p(y = k|x, \mathcal{L}) \geq \tau, \\ y_\varnothing & \text{otherwise.} \end{cases} \tag{8}$$

Guided by these pseudo-labels, stratified random sampling constructs a balanced candidate subpool $\mathcal{U}'$, drawn from an expanded budget $\mathcal{B} = 10B$ to ensure sufficient candidates remain for the final selection, where B denotes the per-iteration annotation budget. Each pseudo-class $k' \in \mathcal{Y} \cup \{y_\varnothing\}$ is allocated a quota $\mathcal{B}_{k'}$ (with $\sum_{k'} \mathcal{B}_{k'} = \mathcal{B}$), computed as follows:

$$\mathcal{B}_{k'} = \begin{cases} \mathcal{B} \cdot \frac{N(y_\varnothing)}{N(\mathcal{U})} & \text{if } k' = y_\varnothing, \\ \frac{\mathcal{B} - \mathcal{B}_{y_\varnothing}}{K} & \text{otherwise,} \end{cases} \tag{9}$$

Here, $N(y_\varnothing)$ and $N(\mathcal{U})$ denote the number of hard-to-learn and total unlabeled instances, respectively, and K is the number of standard classes. Sampling $\mathcal{B}_{k'}$ instances from each pseudo-class k' yields the subpool $\mathcal{U}' = \bigcup_{k'} \mathcal{U}_{k'}$, which preserves the prevalence of ambiguous samples while enforcing a balanced class distribution.

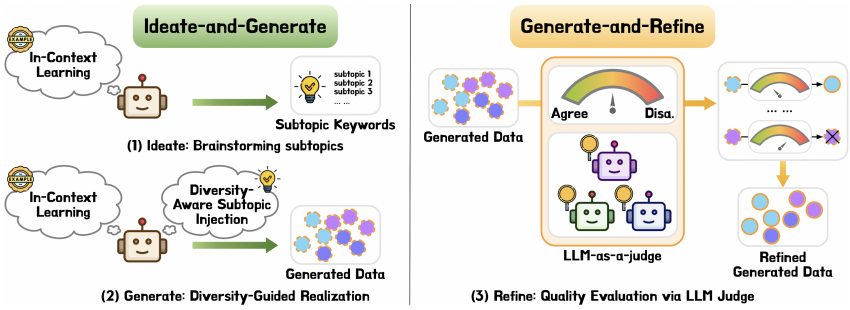


Figure 2: The Ideate-Generate-Refine (IGR) mechanism. (1) Ideate-and-Generate: Newly annotated samples guide the derivation of subtopic keywords; both the samples and these keywords then serve as context for generating initial synthetic data. (2) Generate-and-Refine: An ensemble of three LLMs acts as a judge to filter these candidates, ensuring high quality for the final dataset.

3.2 Generative Data Augmentation

To overcome the semantic drift and limited diversity inherent in traditional NLP augmentation (Wei & Zou, 2019; Xie et al., 2020; Wang et al., 2025), the Ideate-Generate-Refine (IGR) mechanism is introduced. Distinct from rule-based methods, IGR leverages the advanced reasoning capabilities of LLMs via targeted prompt engineering to synthesize new training samples, coupled with an automated filtering module to ensure generative quality. As illustrated in Figure 2, this workflow details how human-annotated samples seed the generation of synthetic data, satisfying both generative quality and semantic diversity requirements for robust AL training. The IGR framework operates in two primary phases, and all prompts used in these phases are provided in Appendix D.1.

Phase I: Ideate-and-Generate. To maximize the semantic diversity of synthetic data, we prompt the generative LLM $\mathcal{M}_{gen}$ to ideate a broad set of subtopics from the current labeled pool $\mathcal{L}^t$. This ideation step keeps the LLM from merely paraphrasing existing samples, pushing it toward broader contextual scenarios for each class k. The generated subtopic set $\mathcal{Q}_k^t$ is formalized as:

$$\mathcal{Q}_k^t = \mathcal{M}_{gen}(\mathcal{I}_{ideate}, \mathcal{L}_k^t) = \{q_1, \ldots, q_N\}. \tag{10}$$

where $\mathcal{I}_{ideate}$ denotes the ideation instruction prompt, and N is the total number of generated subtopics. These subtopics subsequently seed the data synthesis step. Via few-shot in-context learning, existing labeled samples $\mathcal{L}_k^t$ guide the LLM to synthesize novel instances conditioned on a uniformly sampled subtopic $q \sim \mathcal{Q}_k^t$. The raw synthetic dataset $\mathcal{D}_{gen}^t$ is thus obtained via Eq. equation 11. The final synthetic set size is capped at B. Because subsequent filtering may discard candidates, we oversample by α per class in Phase I and generate $B/K + \alpha$ candidates per class.

$$\mathcal{D}_{gen}^t = \bigcup_{k=1}^{K} \left\{ \mathcal{M}_{gen}(\mathcal{I}_{syn}, \mathcal{L}_k^t, q_k^{(j)}) \;\middle|\; q_k^{(j)} \sim \mathrm{Unif}(\mathcal{Q}_k^t),\ j = 1, \ldots, B/K + \alpha \right\}. \tag{11}$$

Phase II: Generate-and-Refine. While Phase I broadens semantic diversity, unconstrained LLM synthesis risks introducing hallucinations or noisy instances that can degrade downstream performance (Chen et al., 2024a). To validate generative quality while mitigating the biases inherent to a single evaluator, we employ an automated LLM-as-a-judge protocol powered by an ensemble of three independent models, $\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$. Specifically, a majority-vote consensus mechanism is implemented: a synthetic candidate $x \in \mathcal{D}_{gen}^t$ is retained only if it receives approval from the majority of the ensemble. The final refined dataset $\mathcal{D}^t$ is formally defined as:

$$\mathcal{D}^t = \left\{ x \in \mathcal{D}_{gen}^t \;\middle|\; \sum_{m=1}^{3} \mathbb{I}\Big(\mathcal{M}_m(\mathcal{I}_{eval}, x)\Big) \geq 2 \right\}, \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathcal{I}_{eval}$ the evaluation prompt. We retain the first B/K approved candidates per class (or all if fewer pass), so $\mathcal{D}^t$ contains up to B samples.

In summary, by integrating the imbalance-aware acquisition strategy (BIRDS) with the diversity-enhancing and quality-assured generative augmentation framework (IGR), the proposed approach mitigates the dual challenges of high annotation costs and inherent class imbalance. The complete iterative training pseudocode, detailing the interaction between the acquisition and generative pipelines, is summarized in Appendix A.1.

Table 1: Statistics of benchmark datasets and corresponding imbalance levels.

Dataset	#Document Train/Test	#Vocab./ #Label	#MeanIR Train/Test
SST-2	67,349/ 1,821	11,570/ 2	1.13/1.0
AGNews	120,000/ 7,600	25,378/ 4	1.0/1.0
PubMed	176,642/29,578	16,589/ 5	2.34/2.32
DBpedia	560,000/70,000	28,654/14	4.6/4.6

Table 2: Statistics of imbalanced variants and corresponding imbalance levels.

Dataset	#Document Train/Test	#Vocab./ #Label	#MeanIR Train/Test
PubMed-5	10,000/2,000	9,592/5	5/5
PubMed-10	10,000/2,000	9,578/5	10/5
PubMed-20	10,000/2,000	9,572/5	20/5
PubMed-50	10,000/2,000	9,542/5	50/5

4 EXPERIMENTS

We evaluate our framework on four text classification benchmarks from the Hugging Face platform: SST-2 (Socher et al., 2013) for sentiment analysis; AGNews (Zhang et al., 2015) for news topic classification; PubMed (Dernoncourt & Lee, 2017), comprising diabetes biomedical abstracts; and DBpedia (Lehmann et al., 2015) for Wikipedia ontology classification. To evaluate GenAL under varying degrees of class imbalance, we construct imbalanced variants of all four benchmark datasets with Mean Imbalance Ratios (MeanIR) ranging from 5 to 50 (Charte et al., 2015), following the procedure detailed in Algorithm 2. We use a validation set of 1,000 instances in each setting; dataset statistics are provided in Tables 1, 2 and 5.

4.1 IMPLEMENTATION

All models are implemented in PyTorch (Paszke et al., 2019) and trained on a single NVIDIA RTX 3090 GPU. We employ DistilBERT (Sanh, 2019) and DeBERTa-v3-small (He et al., 2021) as backbone classifiers, optimized via AdamW (Loshchilov & Hutter, 2019) with a learning rate of 5×10^{-5}. Following standard cold-start protocol (Zhu et al., 2019), models are re-initialized before each AL iteration. The ensemble size is set to $E = 5$, while the initial seed set, query batch size, and generation volume per iteration are uniformly set to 50. Each training cycle spans up to 50 epochs, employing an early stopping mechanism (patience of 10) to mitigate overfitting (Du et al., 2019; Ying, 2019). To ensure robust evaluation, all experiments are averaged across five independent random seeds.

For IGR, we integrate four lightweight LLMs with sizes ranging from 7B to 9B. Specifically, Llama-3.1-8B-Instruct is deployed for data generation at a decoding temperature of 0.9. The LLM-as-a-judge ensemble comprises Qwen2.5-7B-Instruct, Gemma-2-9B-it, and Mistral-7B-Instruct-v0.2, all operating at a temperature of 0.7. Key algorithmic hyperparameters include a pseudo-label confidence threshold of $\tau = 0.6$, a keyword generation target of $N = 10$, and a generation buffer of $\alpha = 3$. Sensitivity analyses for τ and the subpool budget $\mathcal{B}$ are provided in Appendix C.4, and prompt templates are listed in Appendix D.1. These results indicate that GenAL remains stable across a broad hyperparameter range, and the selected defaults consistently yield strong performance across settings. The source code will be made publicly available upon acceptance.

4.2 BASELINES

To evaluate generalization across standard and imbalanced settings while ensuring a strictly controlled comparison, we benchmark GenAL against representative query strategies equipped with the same IGR module. For general-purpose sampling, we include CoreSet (Sener & Savarese, 2018), which minimizes the maximum unlabeled-to-labeled distance, and BEMPS (Tan et al., 2023), which balances uncertainty and diversity via proper scoring rules. For imbalance-specific strategies, we adopt AnchorAL (Lesci & Vlachos, 2024), which uses anchor points to guide sampling toward representative regions, and DIRECT (Nuggehalli et al., 2025), which optimizes batch queries for minority classes while preserving diversity. Additionally, Random Sampling serves as a baseline.

To validate IGR, we benchmark against both conventional techniques and LLM-based methods. Conventional baselines comprise synonym replacement (Zhang et al., 2015), random insertion (Wei & Zou, 2019), and back-translation (Sennrich et al., 2016). LLM-based methods include ZeroGen, a zero-shot generation method (Ye et al., 2022); CoTAM, a chain-of-thought reasoning (Dai et al., 2025); and AugGPT, a controllable LLM-based augmentation method (Peng et al., 2024). For fair comparison, the underlying generative backbone is aligned with the model used in GenAL.

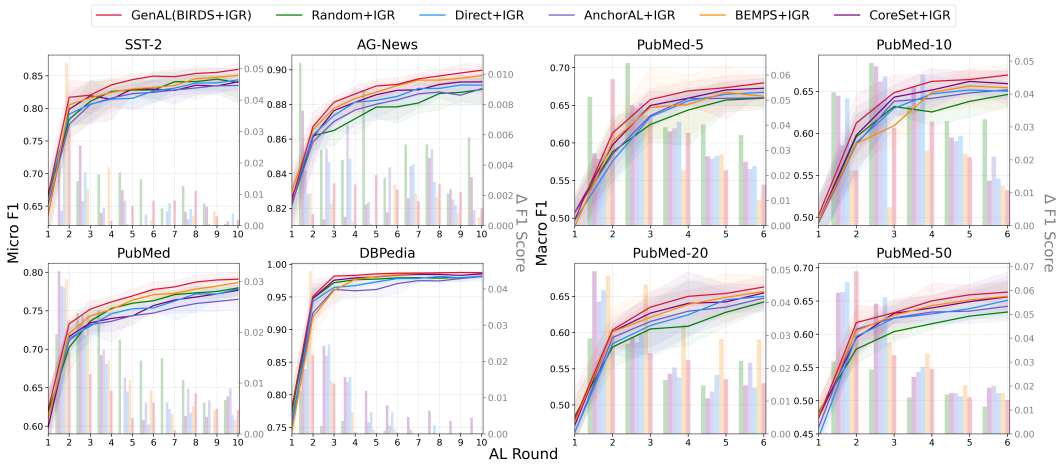


Figure 3: Performance of GenAL against baseline query strategies augmented with IGR: Micro-F1 (DistilBERT) on standard benchmarks and Macro-F1 (DeBERTa) on imbalanced datasets. Shaded regions represent the performance range across five random seeds. Background bars (right axis) show the performance increment contributed by IGR at each AL round for each query strategy.

4.3 GenAL Performance

Figure 3 presents performance curves for GenAL and all baseline query strategies, each augmented with the same IGR module under an identical synthetic-data budget, so the residual performance gap is attributable solely to the query strategy. The background bars track the round-wise performance increment contributed by IGR for each strategy. Three key findings emerge from this analysis.

Finding 1: Query-generation coupling provides a systematic advantage. Under identical annotation and augmentation budgets, GenAL consistently achieves superior learning trajectories across all benchmarks. Importantly, since all baselines are equipped with the identical IGR module, these gains do not stem from data synthesis alone, but from the adaptive coupling between BIRDS and IGR, which continually aligns query selection and data synthesis with the growing labeled pool.

Finding 2: GenAL maintains robust efficacy under class imbalance. GenAL also preserves this advantage across imbalanced benchmarks, whereas competing strategies often degrade or become unstable under imbalanced settings. This robustness is attributable to BIRDS's pseudo-label-guided subsampling, which corrects distributional bias during acquisition, combined with IGR's class-balanced synthesis that mitigates label scarcity in minority classes.

Finding 3: IGR provides the strongest gains in the data-scarce regime. The background bars in Figure 3 show that IGR-induced gains are most pronounced in early AL iterations and gradually narrow as the labeled pool grows, since diverse synthetic instances most meaningfully expand the effective training support when real labeled examples are too scarce to estimate the decision boundary. As genuine labeled data accumulates, this marginal benefit diminishes, reflecting an inherent distribution gap between LLM-generated and naturally occurring text.

The effectiveness of GenAL is statistically validated in Figure 8, reporting significant pairwise wins from paired t-tests. The results reveal a setting-dependent gap among existing strategies: uncertainty methods like BEMPS perform strongly in standard settings but degrade under severe imbalance, whereas diversity strategies like CoreSet remain comparatively robust under imbalance. GenAL achieves the most wins in both settings, showing that it addresses key weaknesses across both strategy families without trading off either regime. We further compare GenAL against baseline query strategies without the IGR module (Figure 10), and GenAL's lead becomes even more pronounced.

4.4 When Does Generated Data Help Most?

This subsection examines whether generated data can replace naturally occurring real data in active learning, and more importantly, when its benefit is strongest. We evaluate GenAL under different real-to-synthetic data ratios (Figure 4), including 1:1 and 1:2 settings (Human:LLM) under both half-budget (50% Human) and full-budget (100% Human) regimes.

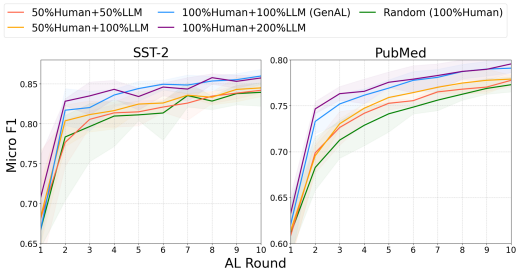


Figure 4: GenAL performance across varying human budgets and synthetic data scales.

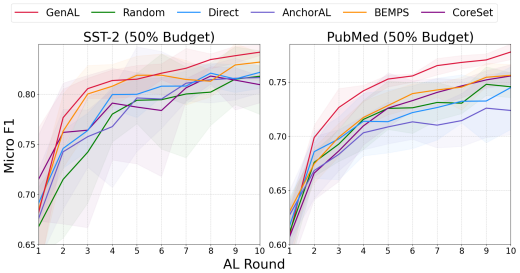


Figure 5: GenAL performance against baseline query strategies under 50% budget setting.

Note that our comparison concerns *data source* rather than *label source*: LLM-generated synthetic text versus naturally occurring text. Results reveal a clear pattern: in low-resource stages, synthetic data provides substantial gains. Even with only half of the standard human budget (50% Human + 50% LLM), GenAL matches or exceeds the full-budget Random baseline. However, scaling synthetic volume further (e.g., 50% Human + 100% LLM or 100% Human + 200% LLM) yields progressively smaller improvements, and the learning curves eventually plateau. This saturation suggests an upper bound on synthetic data utility. Our hypothesis is that this bound can be attributed to a distribution gap between LLM-generated text and naturally occurring text. As real data grows, real samples better characterize the target distribution, making synthetic mismatch more pronounced; consequently, excess synthetic data ceases to help and may even slightly impair real-data evaluation.

We further evaluate a stricter half-budget setting: baselines use 50% human-labeled real data without IGR, whereas GenAL uses the same human budget plus an equal amount of synthetic data (Figure 5). Under this constraint, GenAL retains a significant advantage, indicating that synthetic data is most useful when real data is scarce, even though its marginal benefit diminishes as real data accumulates.

4.5 Component Effectiveness Analysis

This subsection ablates GenAL's two core components, BIRDS and IGR. Across all benchmarks in both standard and imbalanced settings (Figures 6, 11 and 13), BIRDS consistently outperforms competing query strategies, confirming its standalone effectiveness. Likewise, IGR achieves consistent gains over both traditional and LLM-based baselines (Figures 7, 12 and 14), highlighting its generative augmentation capability. Figures 16 and 17 further break down each component's contribution relative to the full framework. Moreover, a coupling ablation (Table 6 and Figure 15) shows that removing the information flow between BIRDS and IGR degrades performance, indicating that their adaptive coupling contributes beyond a static combination. Together, these results show that both components are effective individually and complementary when coupled.

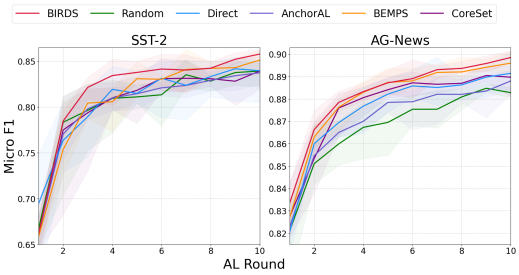


Figure 6: BIRDS versus baseline query strategies on standard benchmarks (part I).

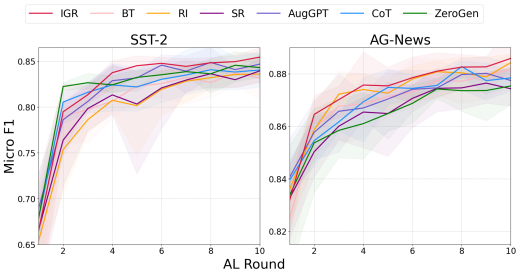


Figure 7: IGR versus baseline augmentation strategies on standard benchmarks (part I).

4.6 Analysis of BIRDS

This subsection examines the standalone behavior of BIRDS, focusing on imbalance handling, pseudo-label reliability, parameter sensitivity, and ablation studies. Figure 18 tracks the MeanIR of the queried labeled set across AL rounds, showing that BIRDS effectively mitigates class imbalance during acquisition. In addition, our measurements of pseudo-label accuracy and coverage indicate that a confidence threshold of $\tau = 0.6$ achieves a favorable accuracy-coverage trade-off across a broad range of values after excluding the hard-to-learn subpool (Figure 19). We further

analyze sensitivity to the pseudo-label threshold τ and the subsampling size $\mathcal{B}$ in Figures 20 and 21. Finally, the ablation study in Figure 22 shows that removing any component of the BIRDS design, diversity scoring, uncertainty scoring, or subsampling, leads to a clear performance drop.

Table 3: Impact of prompt design on generation diversity and quality across models. Guided Prompt denotes the IGR setting where generated subtopic keywords are used as prompts.

	Dataset	Vanilla Prompt				Guided Prompt			
		$\mathcal{M}_Q$	$\mathcal{M}_L$	$\mathcal{M}_G$	$\mathcal{M}_M$	$\mathcal{M}_Q$	$\mathcal{M}_L$	$\mathcal{M}_G$	$\mathcal{M}_M$
Self-BLEU (↓)	SST-2	0.533	0.309	0.748	0.517	0.308	**0.177**	0.411	0.288
	AGNews	0.224	0.220	0.327	0.297	**0.117**	0.143	0.219	0.177
	PubMed	0.378	0.352	0.567	0.448	0.346	**0.251**	0.435	0.332
	DBpedia	0.421	0.345	0.538	0.452	0.297	**0.248**	0.302	0.319
Sent-Bert (↓)	SST-2	0.481	0.503	0.505	0.662	**0.378**	0.449	0.415	0.462
	AGNews	0.083	0.076	0.073	0.102	0.065	**0.063**	0.067	0.079
	PubMed	0.166	**0.156**	0.163	0.186	0.212	0.184	0.258	0.216
	DBpedia	0.066	0.062	0.075	0.071	0.055	**0.053**	0.060	0.054
MAUVE (↑)	SST-2	0.612	0.641	0.465	0.573	0.691	**0.694**	0.578	0.645
	AGNews	0.778	0.802	0.805	0.687	0.776	**0.815**	0.811	0.692
	PubMed	0.901	0.903	0.905	0.802	0.912	0.918	**0.920**	0.813
	DBpedia	0.702	0.711	0.684	0.691	0.713	**0.718**	0.691	0.706

Table 4: Performance analysis of LLM-as-a-judge under different agreement protocols on validation data. SA (Strict Agreement): requires unanimity (all 3 LLMs agree); MA (Majority Agreement): requires a majority vote (at least 2 LLMs agree); AA (Any Agreement): requires only a single positive vote (at least 1 LLM agrees).

	Multi-Agent			Single-Agent		
	AA	MA	SA	$\mathcal{M}_Q$	$\mathcal{M}_M$	$\mathcal{M}_G$
SST-2	0.86	**0.95**	0.90	0.88	0.91	0.92
AGNews	0.76	**0.94**	0.89	0.93	0.88	0.76
PubMed	0.46	**0.80**	0.64	0.79	0.51	0.60
DBpedia	0.83	**0.99**	0.90	0.96	0.89	0.84

4.7 Analysis of IGR

This subsection presents ablation and variant experiments that support the design of IGR.

Generative diversity and quality. Table 3 analyzes the impact of prompt design on generation diversity (Self-BLEU and SentBERT) and overall quality (MAUVE) across four LLMs: Qwen ($\mathcal{M}_Q$), Llama ($\mathcal{M}_L$), Gemma ($\mathcal{M}_G$), and Mistral ($\mathcal{M}_M$). Across all models, subtopic-guided prompting consistently outperforms vanilla prompting: conditioning generation on dynamically extracted subtopic keywords provides broader semantic coverage, effectively mitigating mode collapse and generic outputs. Among the candidates, $\mathcal{M}_L$ achieves the highest diversity while maintaining robust MAUVE scores, justifying its adoption as the default generator in GenAL. Additionally, temperature analysis (Table 7) confirms that $T = 0.9$ yields better diversity over $T = 0.7$ without sacrificing coherence. Finally, human evaluation (Table 8) validates the synthetic data quality, achieving high overall acceptance (≥ 0.90) and strong inter-annotator agreement (Gwet's AC1 ≥ 0.835).

LLM-as-a-judge analysis. Table 4 evaluates LLM-as-a-judge under three consensus protocols: SA (unanimous approval), MA (at least two approvals), and AA (at least one approval), with single-LLM baselines. We benchmark these protocols on a validation set with randomly perturbed labels and measure accuracy in detecting label-text mismatches. MA consistently achieves the highest accuracy across datasets, supporting the robustness of the IGR filtering design. Consequently, this translates to superior downstream AL performance (Figure 23), where IGR-MA achieves the strongest learning trajectories over alternative protocols and the single-LLM baseline.

Different LLM sizes and runtime. We evaluate GenAL across generator scales from 3B to 70B (Figure 24). While larger models yield marginal improvements, 7B–9B models capture nearly all performance gains, offering an ideal trade-off between discriminative performance and deployment efficiency. In terms of runtime (Table 9), the query time of BIRDS is competitive with strong baselines, while IGR incurs only modest generation overhead relative to its downstream improvements.

5 Conclusion

In this work, we present GenAL, an imbalance-aware generative active learning framework that couples BIRDS for acquisition with IGR for controlled synthesis. Across standard and imbalanced benchmarks, GenAL consistently outperforms strong query and augmentation baselines, supporting the effectiveness of coupling selection and generation under fixed annotation budgets. This augmentation-enhanced paradigm opens a practical avenue for integrating LLMs into active learning to address data scarcity and class imbalance. Two limitations remain: since IGR seeds generation with labeled exemplars, extreme cold-start regimes where a minority class has few or no labeled instances leave the LLM without reference data, and LLM bias or hallucinated samples may still propagate errors downstream. Future work will address this cold-start problem, extend GenAL to extreme classification with very large label spaces, and explore open-set adaptation with LLM-generated data as well as prompt designs that improve generation quality.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, as described in the methodology. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results; assisting in the writing of proofs and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to modify scientific figures, summarize or analyze existing literature, source or search for information, and edit the paper to improve readability. We have reviewed all AI-assisted work and checked LLM-generated text for potential plagiarism through a manual literature survey against relevant published work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work uses publicly available text classification benchmarks and does not involve human subjects or sensitive personal data. A potential risk is that LLM-generated samples may contain biases, inaccuracies, or inappropriate content. Our filtering procedure evaluates label-text consistency and does not constitute a comprehensive safety or bias assessment. The proposed method is intended for research on data-efficient text classification; practitioners should conduct task-specific safety and bias assessments before deployment.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing detailed descriptions of the GenAL framework, including BIRDS and IGR, in Sections 3 and 4. The experimental settings, datasets, model configurations, hyperparameters, and evaluation protocols are specified in the main text. The appendix provides pseudocode, prompt templates, ablation studies, parameter sensitivity analyses, and computational-time measurements. We will release the implementation and scripts required to reproduce the experiments upon acceptance.

REFERENCES

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S Yu. Active learning: A survey. In *Data classification*, pp. 599–634. Chapman and Hall/CRC, 2014.

Umang Aggarwal, Adrian Popescu, and Céline Hudelot. Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1428–1437, 2020.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163: 3–16, 2015.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. Little giants: Synthesizing high-quality embedding data at scale. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1392–1411, 2025.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. AlpaGasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024a.

Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6): 137, 2024b.

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

Franck Dernoncourt and Ji Young Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 308–313, 2017.

Flavio Di Palo, Prateek Singhi, and Bilal H Fadlallah. Performance-guided llm knowledge distillation for efficient text classification at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3675–3687, 2024.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*. PMLR, 2019.

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 127–136, 2007.

Yaron Fairstein, Oren Kalinsky, Zohar Karnin, Guy Kushilevitz, Alexander Libov, and Sofia Tolmach. Class balancing for efficient active learning in imbalanced datasets. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pp. 77–86, 2024.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pp. 1050–1059. PMLR, 2016.

Simret Araya Gebreegziabher, Kuangshi Ai, Zheng Zhang, Elena Glassman, and Toby Jia-Jun Li. Leveraging variation theory in counterfactual data augmentation for optimized active learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 894–906, 2025.

Alberto Andres Valdes Gonzalez. Cost-aware model selection for text classification: Multi-objective trade-offs between fine-tuned encoders and llm prompting in production. *arXiv preprint arXiv:2602.06370*, 2026.

Cornelia Gruber, Helen Alber, Bernd Bischl, Göran Kauermann, Barbara Plank, and Matthias Aßenmacher. Revisiting active learning under (human) label variation. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pp. 75–86, 2025.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.

Pietro Lesci and Andreas Vlachos. AnchorAL: Computationally efficient active learning for large and imbalanced datasets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8438–8457, 2024.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899, 2024.

Jingyao Li, Pengguang Chen, Shaozuo Yu, Shu Liu, and Jiaya Jia. BAL: Balancing diversity and novelty for active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (5):3653–3664, 2023.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 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*, pp. 110–119, 2016.

Weike Liu, Hang Zhang, Zhaoyun Ding, Qingbao Liu, and Cheng Zhu. A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215:106778, 2021.

Shayne Longpre, Yu Wang, and Chris DuBois. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4401–4411, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 650–663, 2021.

Lwin Moe, Uyen Trang Nguyen, and Boi Trinh Luu. Mitigating class imbalance in fact-checking datasets through LLM-based synthetic data generation. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pp. 73–80, 2025.

Luis Moles, Alain Andres, Goretti Echegaray, and Fernando Boto. Exploring data augmentation and active learning benefits in imbalanced datasets. *Mathematics*, 12(12):1898, 2024.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 179–192, 2024.

Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13, 2025.

Shyam Nuggehalli, Jifan Zhang, Lalit K Jain, and Robert D Nowak. Improved algorithm for deep active learning under imbalance via optimal separation. In *Forty-second International Conference on Machine Learning*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Branislav Pecher, Ivan Srba, and Maria Bielikova. Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-even performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 165–184, 2025.

Letian Peng, Yuwei Zhang, and Jingbo Shang. Controllable data augmentation for few-shot text mining with chain-of-thought attribute manipulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1–16, 2024.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

Yuanyuan Qi, Jueqing Lu, Xiaohao Yang, Joanne Enticott, and Lan Du. Multi-label Bayesian active learning with inter-label relationships. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

V Sanh. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *ArXiv preprint arxiv:2403.00553*, 2024.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2504.12522.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Wei Tan, Lan Du, and Wray Buntine. Bayesian estimate of mean proper scores for diversity-enhanced active learning. *IEEE transactions on pattern analysis and machine intelligence*, 46 (5):3463–3479, 2023.

Khai Phan Tran, Wen Hua, and Xue Li. Vaediff-docre: End-to-end data augmentation framework for document-level relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7307–7320, 2025.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. Diversity-oriented data augmentation with large language models. *arXiv preprint arXiv:2502.11671*, 2025.

Jason Wei and Kai Zou. 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)*, pp. 6382–6388, 2019.

Chenkai Wu, Yuanyuan Qi, Xiaohao Yang, Jueqing Lu, Gang Liu, Wray Buntine, and Lan Du. ALScope: A unified toolkit for deep active learning. *arXiv preprint arXiv:2508.04937*, 2025.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137:109347, 2023.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11653–11669, 2022.

Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, pp. 022022. IOP Publishing, 2019.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*, 2022.

Jifan Zhang, Shuai Shao, Saurabh Verma, and Robert Nowak. Algorithm selection for deep active learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36: 9614–9647, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jingbo Zhu and Eduard Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 783–790, 2007.

Muzhi Zhu, Chengxiang Fan, Hao Chen, Yang Liu, Weian Mao, Xiaogang Xu, and Chunhua Shen. Generative active learning for long-tailed instance segmentation. In *International Conference on Machine Learning*, pp. 62349–62368. PMLR, 2024.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Yu Zhu, Jinghao Lin, Shibi He, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. Addressing the item cold-start problem by attribute-driven active learning. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):631–644, 2019.

A Appendix: Pseudo-code

A.1 Pseudo-code for GenAL

Algorithm 1 GenAL Training and Update Strategy

Input: Initial Labeled pool $\mathcal{L}$, Unlabeled pool $\mathcal{U}$, Query budget B, Max iterations T.
Output: Final AL Model.
for $t = 1$ **to** T **do**
  // Phase I: BIRDS Acquisition
  Compute diversity score $\mathcal{S}_D$ via Eq. equation 1
  Compute uncertainty score $\mathcal{S}_U$ via Eq. equation 2-equation 5
  Estimate pseudo-labels $\hat{y}$ for $\mathcal{U}$ via Eq. equation 8
  Assemble candidate subpool $\mathcal{U}_{\text{sub}}$ via Eq. equation 9
  Select query batch $\{x^\star\}$ of size B via Eq. equation 6-equation 7
  Acquire human annotations for $\{x^\star\}$
  // Phase II: IGR Generative Augmentation
  Generate subtopic keywords $\mathcal{Q}^t$ with $\mathcal{L}^t$ via Eq. equation 10
  Synthesize raw candidate pool $\mathcal{D}^t_{gen}$ via Eq. equation 11
  Refine dataset to obtain $\mathcal{D}^t$ via Eq. equation 12
  Update pools: $\mathcal{L} \leftarrow \mathcal{L} + \{(x^\star, y)\} + \mathcal{D}^t; \mathcal{U} \leftarrow \mathcal{U} - \{x^\star\}$
  Retrain AL model using updated $\mathcal{L}$
end for

B Additional Experiments Detail

Algorithm 2 Imbalanced Data Generation

Input: Original dataset $\mathcal{D}$, Target MeanIR ρ, Data budget M, Random seed s.
Output: Imbalanced dataset $\mathcal{D}_{\text{imb}}$.
Compute linearly decaying class sizes $\{n_i\}$ satisfying $\text{MeanIR} = \rho$ and $\sum_i n_i = M$; randomly assign them to classes
Initialize $\mathcal{D}_{\text{imb}}$ as an empty dataset
for each class c in $\mathcal{D}$ **do**
  Randomly sample the assigned number of instances from class c
  Add the sampled instances to $\mathcal{D}_{\text{imb}}$
end for
Shuffle $\mathcal{D}_{\text{imb}}$ using random seed s and save

Table 5: Statistics of the controlled imbalanced variants (SST-2, AGNews, DBpedia).

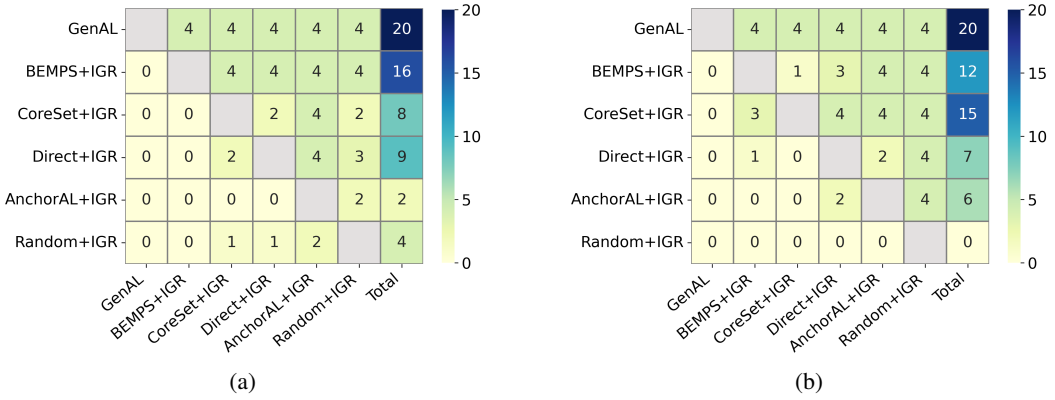
MeanIR	**#Vocab./ #Label**			**#Document**
	SST-2	AGNews	DBpedia	**Train/Test**
5	8,208/2	17,424/4	21,725/14	10,000/2,000
10	7,985/2	17,443/4	21,143/14	10,000/2,000
20	N/A	17,787/4	21,718/14	10,000/2,000
50	N/A	16,872/4	21,681/14	10,000/2,000

Following Algorithm 2, we construct additional imbalanced variants of AGNews and DBpedia at target MeanIR values of 5, 10, 20, and 50, and of SST-2 at MeanIR values of 5 and 10. Here $\text{MeanIR} = \frac{1}{K}\sum_i \max_j n_j/n_i$, where n_i is the size of class i. For each setting, class sizes follow an arithmetic progression $n_i = a + (i-1)k$ whose parameters are solved by bisection so that the target MeanIR is met under a fixed data budget of 10,000; the sizes are then randomly permuted across classes, and the required instances are randomly sampled from each class with a fixed seed. As summarized in Table 5, all variants share the same training and test sizes and a fixed test MeanIR

of 5, so that performance across training imbalance levels is compared on a common test distribution. This controlled construction isolates the effect of class imbalance from changes in dataset size and enables evaluation across datasets from different domains. For the binary SST-2, MeanIR values of 20 and 50 are omitted (marked N/A): with only two classes, such ratios leave the minority class with too few instances, and in particular too few labeled exemplars in the early AL rounds, to provide meaningful reference data for IGR. This regime corresponds to the cold-start limitation discussed in the conclusion, where generation-based augmentation is not applicable by design, and we therefore leave it to future work.

C Additional Results

C.1 Additional Results for GenAL Performance



(a) (b)

Figure 8: Pairwise statistical significance matrix (win-counts) across (a) standard, and (b) imbalanced datasets. A cell at row i and column j shows the number of datasets (out of 4) where row i significantly outperforms column j, evaluated via a paired t-test ($p < 0.05$) across 5 random seeds.

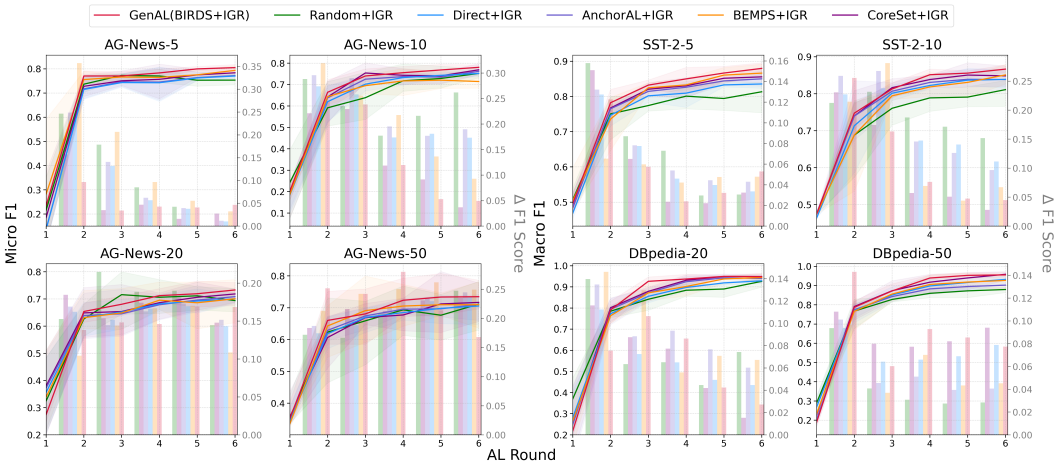
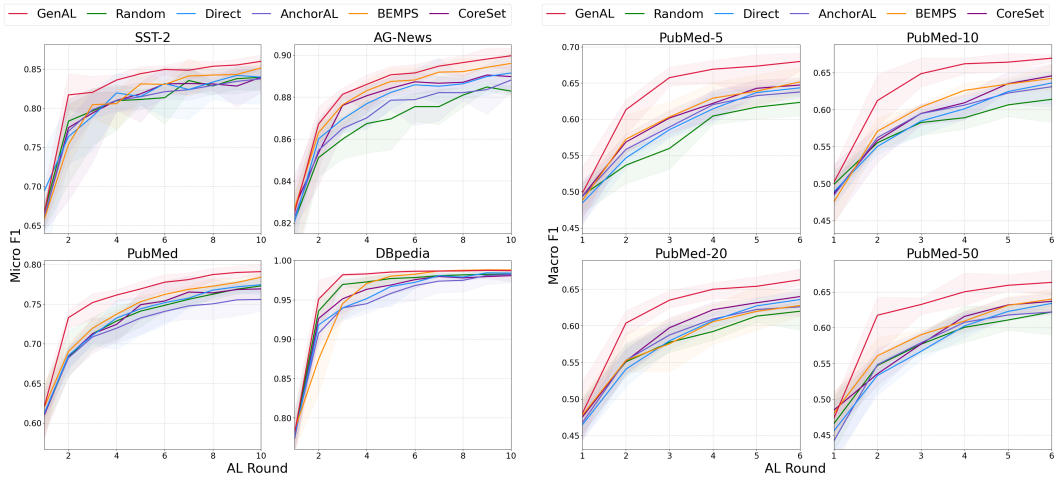


Figure 9: Performance of GenAL against baseline query strategies augmented with IGR on other three imbalanced datasets, Macro-F1 (DeBERTa). Shaded regions represent the performance range across five random seeds. Background bars (right axis) show the performance increment contributed by IGR at each AL round for each query strategy.

Figure 8 reports the pairwise win-count matrices for GenAL and all IGR-augmented baselines on the standard and imbalanced benchmarks, complementing the performance curves in Figure 3. GenAL wins all 20 pairwise comparisons in both regimes, confirming that its advantage generalizes across benchmarks and imbalance levels rather than being confined to a specific setting. Notably, the relative ranking among baselines shifts considerably between the two regimes: the uncertainty-based



(a) (b)

Figure 10: Performance comparison of GenAL against baseline query strategies. (a) Micro-F1 using DistilBERT on benchmarks, and (b) Macro-F1 using DeBERTa on imbalanced datasets.

BEMPS ranks second under standard conditions, but is overtaken by the diversity-based CoreSet under imbalanced settings. This reversal echoes the asymmetry discussed above: uncertainty-driven acquisition is effective under standard settings but grows unreliable as informativeness estimates are skewed by majority-class dominance, whereas diversity-based sampling remains comparatively robust by continuing to cover the embedding space irrespective of label frequency.

Figure 9 extends Figure 3 to the remaining three imbalanced datasets under the same protocol: every baseline query strategy is augmented with the same IGR module, and background bars indicate the round-wise performance increment attributable to IGR for each strategy. GenAL consistently outperforms all IGR-augmented baselines throughout the AL process, confirming that its advantage under class imbalance is not confined to the setting reported in the main text but holds across imbalanced datasets from different domains. The IGR-induced gains also follow the same trajectory as in Figure 3: they are most pronounced in the early rounds, when real labeled examples are too scarce to characterize the decision boundary, and gradually diminish as the labeled pool grows.

Figure 10 compares the full GenAL framework against baseline query strategies without IGR augmentation on standard and imbalanced datasets. GenAL retains a more pronounced lead in both settings, and the margin over the strongest baseline widens under imbalanced datasets (Figure 10b). This suggests that BIRDS's pseudo-label-guided subsampling provides an increasingly decisive advantage as class imbalance intensifies, while the benefit of IGR, recoverable by comparing against Figure 3, is largest precisely where the acquisition strategy alone is least sufficient. Together, these results indicate that GenAL's coupled design addresses complementary weaknesses of existing strategies rather than merely compounding incremental gains.

C.2 Additional Results for Component Effectiveness Analysis

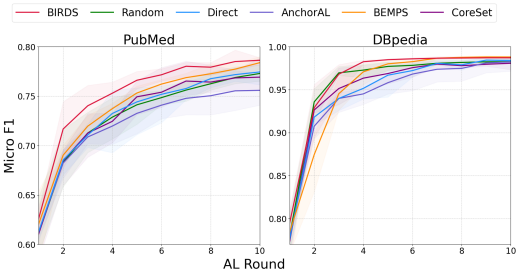


Figure 11: BIRDS versus baseline query strategies on standard benchmarks (part II).

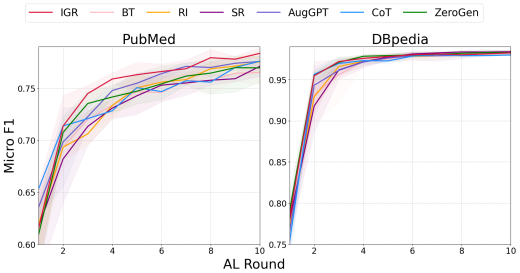


Figure 12: IGR versus baseline augmentation strategies on standard benchmarks (part II).

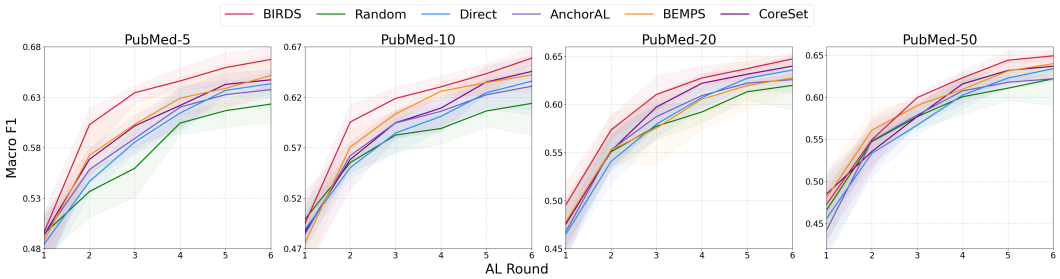


Figure 13: BIRDS versus baseline query strategies on imbalanced datasets (Macro-F1, DeBERTa).

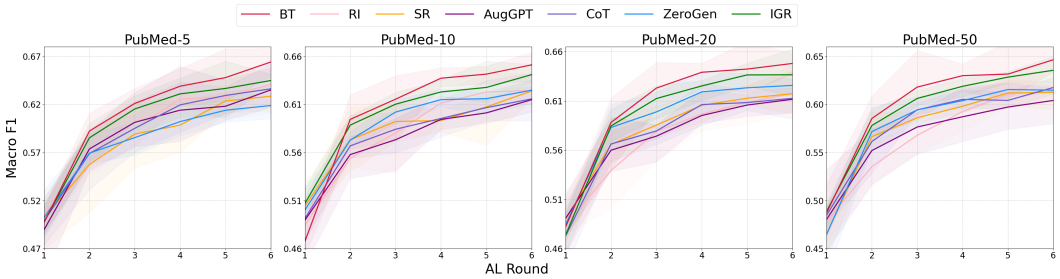


Figure 14: IGR versus baseline augmentation strategies on imbalanced datasets (Macro-F1, DeBERTa).

Table 6: Settings for the coupling ablation. $\mathcal{L}^t$: cumulative human-labeled pool at round t ($\mathcal{L}^0$: initial seed set); $\mathcal{D}^t$: synthetic data generated up to round t.

	Condition	IGR context	BIRDS training data
A	GenAL	$\mathcal{L}^t$	$\mathcal{L}^t \cup \mathcal{D}^t$
B	Frozen BIRDS: no synthetic data for data selection	$\mathcal{L}^t$	$\mathcal{L}^t$
C	Frozen IGR: no newly labeled data for generation	$\mathcal{L}^0$	$\mathcal{L}^t \cup \mathcal{D}^t$
D	Fully decoupled	$\mathcal{L}^0$	$\mathcal{L}^t$

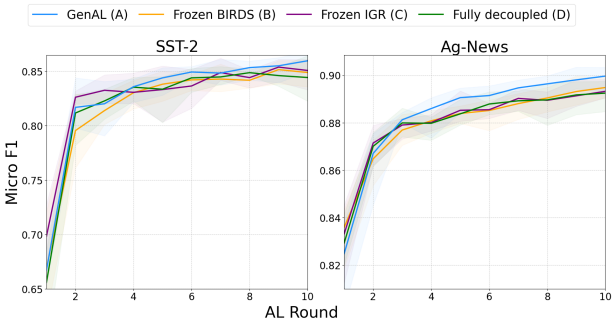


Figure 15: Performance of the four coupling settings across active learning rounds.

Figure 6 and Figure 11 isolate the BIRDS query strategy and benchmark it against existing baselines on standard datasets, while Figure 13 evaluates its performance on imbalanced datasets. The results demonstrate that BIRDS consistently outperforms all baselines under standard settings, and maintains a clear lead under imbalanced conditions, confirming its independent capability to capture data informativeness and address the imbalanced problem.

Figure 7 and Figure 12 evaluate the IGR module against traditional and LLM-based augmentation methods on standard datasets, while Figure 14 evaluates it on imbalanced datasets. The results validate that IGR consistently synthesizes higher-quality data than existing augmentation methods, markedly improving learning efficiency on imbalanced datasets. This suggests that generating diverse, class-balanced samples is an effective remedy for imbalance in low-resource settings.

To examine the coupling effect between BIRDS and IGR, we conduct an ablation study that selectively removes the information flow between the two components; the four settings are summarized in Table 6. Setting A corresponds to the full GenAL framework: IGR generates data conditioned on the newly annotated samples, and BIRDS selects instances using a model trained on both annotated and generated data. Setting B removes the synthetic-to-acquisition feedback: the model used for BIRDS selection is trained on labeled data only. For a fair comparison, the reported performance at each iteration is still obtained from a model trained on both labeled and generated data. Comparing A against B thus isolates whether synthetic data improves the quality of subsequent BIRDS acquisitions. Setting C instead freezes the IGR generation context: synthetic data is generated solely

from the initial labeled set and is never updated with newly acquired samples. Comparing A against C reveals whether newly annotated examples improve adaptive generation. Setting D fully decouples the two components: IGR generates from the initial labeled set only, and BIRDS selects based on labeled data alone, so no information is exchanged in either direction. As shown in Figure 15, Setting A consistently outperforms all three decoupled variants across AL rounds. These results indicate that both directions of information flow contribute to the overall performance, and that the adaptive coupling between acquisition and generation, rather than their mere combination, underlies the effectiveness of GenAL.

C.3 Individual Contributions of GenAL Components

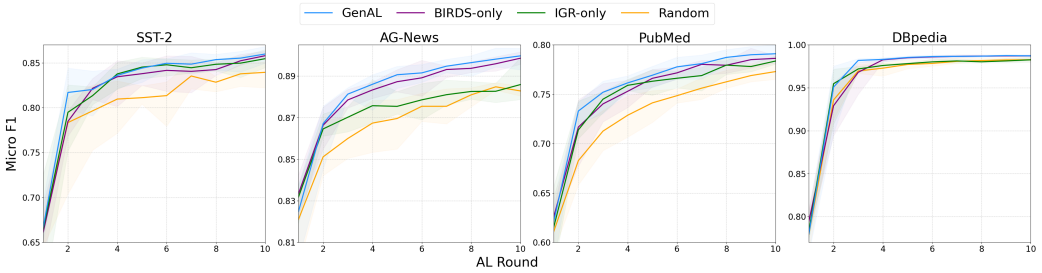


Figure 16: GenAL versus individual components on standard benchmarks (Micro-F1, DistilBERT).

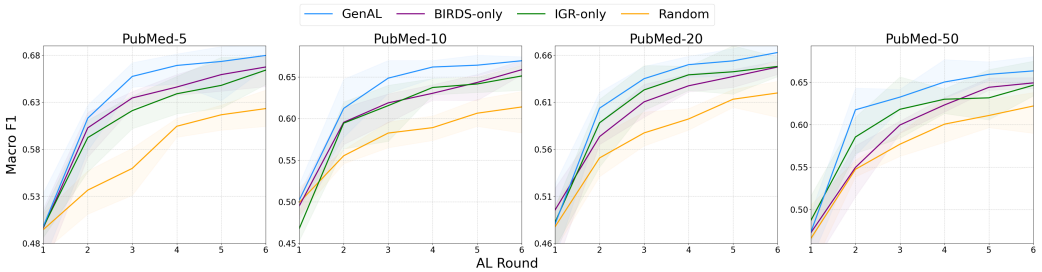


Figure 17: GenAL versus individual components on imbalanced datasets (Macro-F1, DeBERTa).

To assess the individual contributions of GenAL's two main components, we compare the full framework against variants that retain only BIRDS or only IGR. Figure 16 and Figure 17 show that the full system consistently performs best on both standard and imbalanced datasets, confirming that the coupled design yields complementary gains beyond either component alone. The relative contribution of each component, however, varies by dataset. On news corpora such as AGNews, BIRDS contributes more strongly, possibly because time-sensitive, event-specific content is harder to synthesize faithfully. In contrast, on specialized domains such as PubMed, IGR provides larger gains by expanding scarce class-specific training signals. Across imbalanced settings, both components remain important, jointly supporting robust performance.

C.4 Additional Results for BIRDS

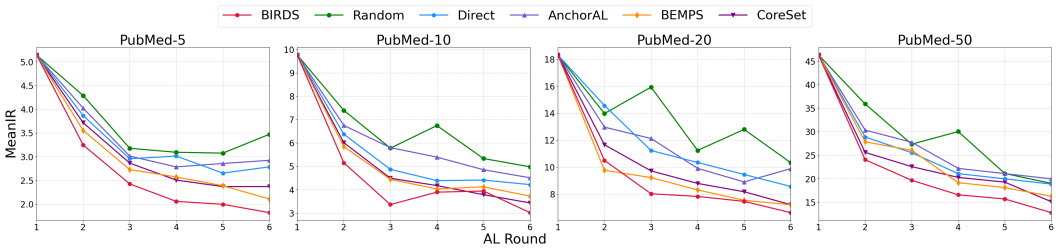


Figure 18: MeanIR of the queried labeled set across active learning iterations.

Figure 18 reports the MeanIR trends of the cumulatively queried labeled set under different imbalance levels. A lower MeanIR indicates a more class-balanced queried subset. Across all AL

rounds, BIRDS consistently achieves the lowest MeanIR among competing strategies, showing that its pseudo-label-guided subsampling effectively counteracts majority-class bias during acquisition. From the perspective of the selected labeled data, this result indicates that BIRDS is better able to handle imbalanced settings by maintaining a more balanced supervision signal throughout training.

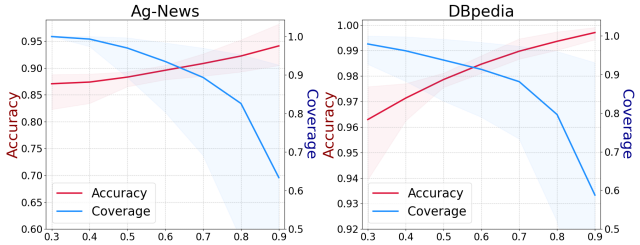


Figure 19: Pseudo-label accuracy and coverage under different confidence thresholds.

To further examine the pseudo-label confidence threshold τ, Figure 19 analyzes pseudo-label accuracy and coverage from the perspective of the unlabeled pool. Importantly, the reported accuracy excludes the hard-to-learn instances, since these samples are explicitly separated and handled through a dedicated subpool rather than forced into standard pseudo-classes. Under this evaluation protocol, increasing τ improves pseudo-label accuracy but reduces coverage, revealing the expected reliability-coverage trade-off. Empirically, $\tau = 0.6$ provides a favorable balance: it preserves sufficiently high pseudo-label precision while retaining enough confident unlabeled samples for effective subsampling.

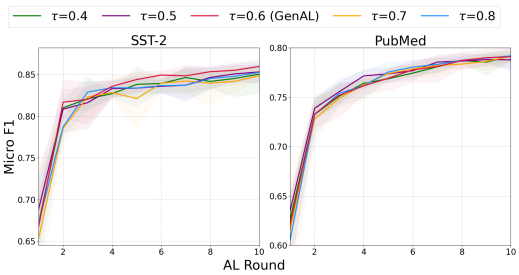


Figure 20: Sensitivity of GenAL to the BIRDS pseudo-label threshold τ.

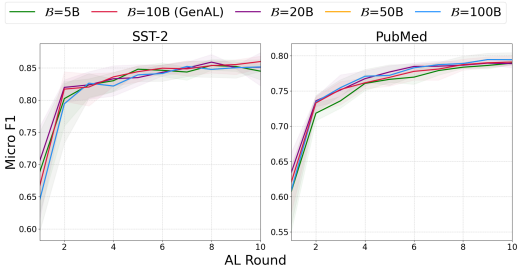


Figure 21: Sensitivity of GenAL to the BIRDS subsampling size $\mathcal{B}$.

Figure 20 and Figure 21 report the sensitivity of GenAL to τ and $\mathcal{B}$, two hyperparameters defined in BIRDS. These hyperparameters determine the queried set and thereby shape GenAL's overall learning trajectory. The results show that GenAL remains stable across a broad range of settings, with only limited performance variation. Among the tested configurations, $\tau = 0.6$ and $\mathcal{B} = 10B$ yield the most consistent overall results and are therefore used as defaults in the main experiments.

C.5 Additional Results for IGR

Table 7: Effect of sampling temperature on generated text diversity across generators.

	Dataset	Temperature = 0.7				Temperature = 0.9			
		$\mathcal{M}_Q$	$\mathcal{M}_L$	$\mathcal{M}_G$	$\mathcal{M}_M$	$\mathcal{M}_Q$	$\mathcal{M}_L$	$\mathcal{M}_G$	$\mathcal{M}_M$
Self-BLEU (↓)	SST-2	0.644	0.442	0.821	0.563	0.533	**0.309**	0.748	0.517
	AGNews	0.256	0.260	0.379	0.348	0.224	**0.220**	0.327	0.297
	PubMed	0.397	0.422	0.649	0.478	0.378	**0.352**	0.567	0.448
	DBpedia	0.485	0.396	0.669	0.512	0.421	**0.345**	0.538	0.452
Sent-Bert (↓)	SST-2	0.534	0.531	0.504	0.668	0.481	**0.503**	0.505	0.662
	AGNews	0.077	**0.071**	0.086	0.111	0.083	0.076	0.073	0.102
	PubMed	0.163	0.159	0.158	0.172	0.166	**0.156**	0.163	0.186
	DBpedia	0.067	0.064	0.081	0.077	0.066	**0.062**	0.075	0.071

Table 8: Human evaluation of IGR generation quality. H1–H3 denotes the proportion of instances rated as 'Good' by three human annotators; Overall denotes mean acceptance; Gwet's AC1 quantifies inter-annotator agreement.

	H1	H2	H3	Overall	Gwet's AC1
SST-2	1.0	1.0	1.0	1.0	0.98
AGNews	0.92	0.90	0.88	0.90	0.835
PubMed	0.96	0.92	0.92	0.93	0.874
DBpedia	0.92	0.98	0.94	0.95	0.896

Table 7 analyzes the impact of sampling temperature on generation diversity across four distinct LLMs: Qwen ($\mathcal{M}_Q$), Llama ($\mathcal{M}_L$), Gemma ($\mathcal{M}_G$), and Mistral ($\mathcal{M}_M$). The results validate that

increasing the sampling temperature yields consistently enhanced diversity across all models. This occurs because higher temperatures flatten the softmax probability distribution, encouraging the selection of lower-probability tokens and effectively reducing repetitive patterns.

To further validate our approach, we conducted a human evaluation of data quality, as detailed in Table 8. A subset of 200 samples was randomly extracted across the four benchmark datasets and independently annotated by three authors using a binary classification of 'good' versus 'poor.' The results indicate that while our IGR module generally achieves high semantic fidelity, performance varies across domains. For tasks such as SST-2, the generated text proves to be highly coherent and strictly label-consistent. Conversely, the AGNews dataset presents a slightly greater challenge; we observed instances of inter-class ambiguity, particularly between the 'World' and 'Sci/Tech' categories. This is likely attributable to the conceptual intersection where technological developments often constitute global events.

Furthermore, we report Gwet's AC1 statistic Gwet (2008) to measure inter-annotator agreement. Cohen's Kappa is not applicable to this evaluation task because the highly imbalanced distribution of positive ratings triggers the well-documented 'Kappa Paradox,' where the Kappa score is deflated due to an exaggerated penalization for chance agreement. Overall, the aggregate results confirm that IGR effectively balances quality and diversity, providing a reliable and robust source for data augmentation.

The parameter α serves as a generation buffer to ensure that the LLM-as-a-judge receives a sufficient candidate pool for filtering. Its effect on final performance is limited, and the pass rates in Table 4 consistently exceed 80%. Accordingly, we set $\alpha = 3$ to preserve enough candidates while controlling additional computation.

C.6 Ablation Study

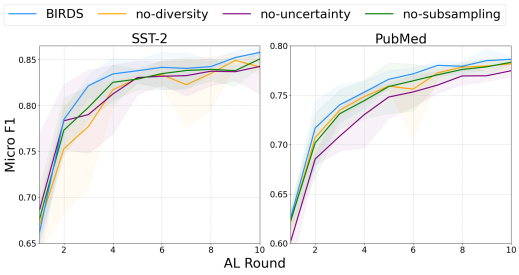


Figure 22: Ablation study of BIRDS components without IGR, isolating uncertainty, diversity, and pseudo-label-guided subsampling.

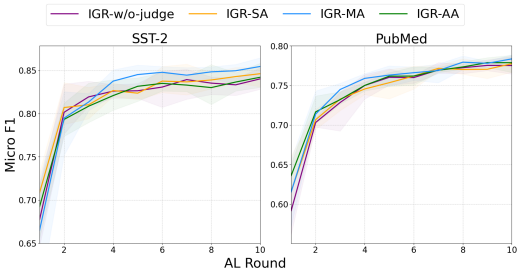


Figure 23: Ablation study of LLM-as-a-judge agreement protocols (AA, MA, SA) under random sampling.

Figure 22 presents the ablation study for the BIRDS query strategy, dissecting the individual contributions of its three core elements: the uncertainty score, the diversity score, and the pseudo-label-guided subsampling. The results demonstrate that each component plays a critical and distinct role in the strategy's overall success. Specifically, the uncertainty score ensures the selection of highly informative samples near the decision boundary, while the diversity score prevents redundancy by enforcing broad coverage across the feature space. Crucially, the pseudo-label guidance provides distributional awareness, preventing the model from overlooking minority classes. As shown in the figure, removing any single component leads to a noticeable degradation in performance, validating that all three factors are essential for robust active learning.

Furthermore, Figure 23 illustrates the performance of the IGR mechanism across different agreement protocols throughout the AL training iterations. The analysis indicates that the Majority Agreement protocol consistently yields the best results. This further validates its effectiveness in filtering low-quality synthetic data while maintaining the optimal balance between strictness and permissiveness.

Figure 24 illustrates the GenAL performance across varying LLM scales. To evaluate the impact of model size on generation quality, we benchmarked multiple configurations, primarily utilizing the Llama family (Llama-3.2-3B-Instruct and Llama-3.3-70B-Instruct; the latter is accessed via an inference API as it exceeds our local GPU memory). Since the Llama series currently lacks a mid-weight

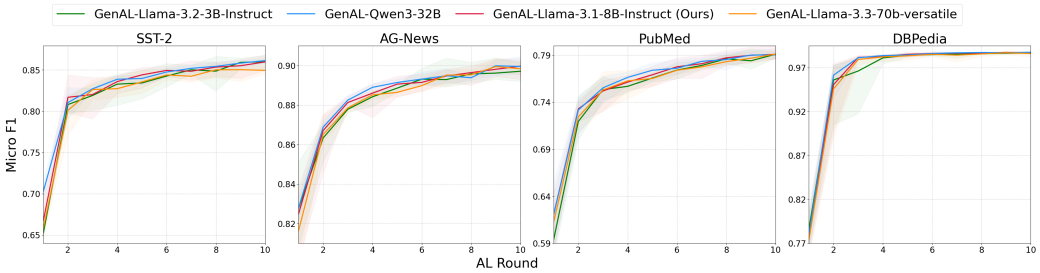


Figure 24: Ablation study of LLM scale in GenAL (ranging from 3B to 70B), reported as Micro-F1 with DistilBERT on standard benchmarks.

32B variant, we incorporated Qwen3-32B to ensure a comprehensive scale comparison. Interestingly, the massive Llama-3.3-70B-Instruct does not exhibit superior performance compared to its smaller counterparts. While the Qwen3-32B model yields marginal performance improvements on this generation task, the 8B baseline already captures the vast majority of the benefits. This indicates that for discriminative tasks with relatively low-complexity semantic spaces, the 7B–9B parameter range serves as an optimal computational sweet spot, effectively balancing maximum performance gains with feasible deployment costs.

C.7 GenAL Computing Time Analysis

Table 9: Computational time per AL iteration. The table compares the query time of GenAL against strong baselines and reports the additional generation time introduced by the IGR module.

	Query Time			**Generation**
	GenAL	BEMPS	CoreSet	**Time (IGR)**
SST-2	4s	4s	7s	2m33s
AGNews	1m31s	1m24s	23s	2m56s
PubMed	1m24s	1m24s	19s	3m25s
DBpedia	3m15s	3m20s	50s	4m17s

Table 9 details the computational cost per AL iteration, benchmarked on a single NVIDIA RTX 3090 GPU. In this setup, 50 samples are queried for human annotation, and an additional 50 synthetic samples are generated by the IGR module. Overall, GenAL's query time remains highly competitive with strong baseline strategies. The IGR module introduces a generation time of only 2 to 4 minutes per iteration, which conservatively includes I/O overhead for locally loading/offloading LLM weights.

Furthermore, IGR serves as a plug-and-play module that can be seamlessly integrated with various active learning strategies. As demonstrated in our main experiments, this marginal computational investment consistently yields significant performance gains. Ultimately, the primary bottleneck in active learning inherently lies in the days or weeks required for human experts to annotate challenging data. Therefore, the minor computational overhead introduced by our framework is vastly outweighed by the substantial reduction in expensive human labor.

D PROMPT

D.1 PROMPT DESIGN

Prompt for subtopic generation:

Role
You are an expert {specific expert}.

Task
I am generating synthetic data for {dataset name}.
The overall categories are {List of labels}.
Your goal is to identify a granular list of subtopics or "aspects" that typically appear in the {label}.

Reference Data
The following are example:
"{example}"

Instructions
1. **Analyze First:** Briefly identify the core themes present in the reference examples.
2. **Expand Radically:** Brainstorm a diverse list of 10 DISTINCT subtopics that commonly appear in {label}.
3. **Ensure Diversity:** You must expand entirely beyond the reference data. Force the 10 subtopics to cover completely different dimensions. Do not provide overlapping concepts.
4. **Format constraint:** Each subtopic must be a concise phrase (2-6 words max), NOT a full sentence.

Output Format
Respond ONLY with valid, raw JSON. Do not include markdown code blocks, and do not add any conversational text before or after the JSON.

Use this exact structure:
```
{{
"analysis": "A brief 1-2 sentence analysis of the reference data to ground your thinking.",
"subtopics": [ "subtopic 1", ... ]
}}
```

Prompt for generating synthetic samples with the given subtopic keyword:

Instruction
You are an expert content creator. Generate a {data type} that strictly aligns with the target label and focuses entirely on the provided subtopic.

Configuration
- Target Label: {label}
- Specific Subtopic: {random.choice(subtopics)}
- Target Length: {average length range} words

Reference Example
Here is an example demonstrating the expected quality, depth, and format. Do not copy its exact style or vocabulary; use it only as a benchmark for quality.
<example >
{example}
</example >

Constraints
- CRITICAL: The content must strictly belong to {label}. Ensure it contains ZERO elements of these unrelated categories: <avoid >{other labels}</avoid >.
- Do not explicitly state the name of the label or subtopic in the text.
- Do not include any introductory or concluding remarks (e.g., "Here is the text").

Output
Write your generated {data type} directly below:

Prompt for evaluation:

Instruction
The following text is from a dataset of {domain of data}.

Determine if the text belongs to {label}, and NOT {other labels}.

Text
"{text}"

Question
Does the text belong to {label}?
Respond ONLY with "Yes" or "No".

Vanilla prompt used in Table 3:

- Task -
Generate a short {data type} (around {average length} words) that clearly belongs to the category **{label}**.
At the same time, it should be **distinctly different** from the following other categories: {other labels}.

You may refer to the example below for tone and format.
Example:
#########
{example}
#########
Output: