

Attributes as Textual Genes: Leveraging LLMs as Genetic Algorithm Simulators for Conditional Synthetic Data Generation

Guangzeng Han, Weisi Liu, Xiaolei Huang

Department of Computer Science, University of Memphis

Memphis, TN, USA

{ghan, wliu9, xiaolei.huang}@memphis.edu

Abstract

Large Language Models (LLMs) excel at generating synthetic data, but ensuring its quality and diversity remains challenging. We propose **Genetic Prompt**¹, a novel framework that combines genetic algorithms with LLMs to augment synthetic data generation. Our approach treats semantic text attributes as gene sequences and leverages the LLM to simulate crossover and mutation operations. This genetic process enhances data quality and diversity by creating novel attribute combinations, yielding synthetic distributions closer to real-world data. To optimize parent selection, we also integrate an active learning scheme that expands the offspring search space. Our experiments on multiple NLP tasks reveal several key findings: Genetic Prompt not only significantly outperforms state-of-the-art baselines but also shows robust performance across various generator model sizes and scales. Moreover, we demonstrate that fusing our synthetic data with the original training set significantly boosts downstream model performance, particularly for class-imbalanced scenarios. Our findings validate that Genetic Prompt is an effective method for producing high-quality synthetic data for a wide range of NLP applications.

1 Introduction

Leveraging large language models (LLMs) to synthesize training data for smaller models is a promising path to efficiency and accuracy across NLP tasks (Long et al., 2024), yet improving the quality and diversity of such data remains underexplored (Ye et al., 2022; Yu et al., 2023). Prior work often steers synthetic data generation with pre-defined conditions such as clinical named entities (Xu et al., 2024; Josifoski et al., 2023), label correlations (Chung et al., 2023), and writing

styles (Xu et al., 2024; Yu et al., 2023). However, reliance on static cues limits model reasoning, cross domain generalization, and dataset diversity. Recent findings further suggest that instructional diversity is crucial for high quality synthesis (Kim et al., 2025; Xu et al., 2024; Li et al., 2024b), which motivates our core question: *how can we automatically amplify data diversity and generator adaptability for robust downstream training?*

Inspired by the genetic algorithms (Mitchell, 1980) (GAs), we propose **Genetic Prompt** framework for generating high-quality synthetic data that 1) automatically identifies and treats textual attributes as gene sequences, 2) integrates active learning to optimize parent selection process, and 3) diversifies synthetic data by simulating the genetic processes of crossover and mutation.

GAs are population-based search heuristics inspired by the process of natural selection iteratively evolving solutions to optimization problems (Mitchell, 1980). However, optimizing synthetic data generation via genetic method and LLMs has rarely been explored, due to two primary issues. First, assessing the fitness of individual samples is hard for downstream models, because a data point’s effect is context-dependent and hinges on high-order interactions with the rest of the training set, as shown by recent study (Ilyas et al., 2022). Second, data characteristics can be diverse and complex, while only focusing on partial scopes of target data may not be applicable to yield a large amount of high-quality synthetic data by LLMs. For example, a close study (Liu et al., 2024) alters words and sentences on the syntactic level to continue optimizing the best individual samples. This approach is inherently constrained to generating only a small volume of data, limiting its scalability for large-scale synthetic data generation.

To mitigate the unreliability of per example fitness in selecting parents, we replace conventional fitness based selection with an active learning (AL)

¹The code is available at: <https://github.com/trust-nlp/Genetic-Prompt>

strategy (Zhang et al., 2023; Ren et al., 2021). In each round, AL pairs previously unused, diverse parents, enlarging the search space for offspring generation. To improve coverage and scalability, we avoid operating on individual tokens or sentences or treating them as genes. Instead, we encode higher level linguistic and semantic attributes (e.g., readability and sentence structure) as genes. This design lets us leverage the knowledge and comprehension abilities of LLMs to implement crossover and mutation at the attribute level, producing large volumes of high quality and diverse data.

Our experiments demonstrate that Genetic Prompt consistently outperforms state-of-the-art baselines like AttrPrompt (Yu et al., 2023) and Curated LLM (Seedat et al., 2024), while also exhibiting superior robustness and scalability across various generator LLM sizes. Our qualitative analysis shows that the data generated by Genetic Prompt has superior semantic and lexical diversity and more accurately captures the statistical characteristics of real data. Furthermore, we find that incorporating this synthetic data with gold data can significantly boost model performance, particularly in data imbalance settings. These findings validate the effectiveness of our Genetic Prompt framework for producing high-quality synthetic data for downstream tasks.

2 Genetic Prompt for Synthetic Data Generation

In this section, we present our Genetic Prompt framework in Figure 1. The key idea is to automatically identify and treat data attributes as genes and prompt LLMs to do crossover and mutation operations over a small initial population from gold data. It consists of four major modules: 1) textual genes identification, 2) active parents selection, 3) crossover and mutation, and 4) downstream training. We include detailed algorithmic process in Algorithm 1 under the Appendix.

2.1 Textual Genes Identification

To improve the quality of synthetic data or diversify its style, attribute-aware conditional text generation a promising direction. However, these methods generally rely on attribute values strictly predefined by humans (Yu et al., 2021; Russo et al., 2020; Logeswaran et al., 2018) or LLMs (Xu et al., 2024; Yu et al., 2023). In contrast, our approach views data

attributes as “genes” of the text. With just a single round of interaction with LLMs, we identify the genes without predefining specific values for each gene, avoiding excessive constraints on the LLMs and reducing bias from human selection. Specifically, our solution involves inputting the metadata (such as “task type: chemical protein interaction”) and task samples into LLMs and posing questions like “If the attributes of the text are viewed as genes, which genes are most important in the given chemical protein interaction extraction task?” Given the Chemprot (Taboureau et al., 2010) data as an example, we received responses including *length*, *sentence structure*, *entity proximity*, *polarity*, etc. Subsequently, following previous work (Yu et al., 2023), we adopt a human-AI collaboration scheme to empirically select vital genes (examples in Table 10 under Appendix B.5). The textual genes identification process can be formalized as:

$$G = \Phi(M, S, Ins.G)$$

, where G represents the set of identified textual genes, Φ represents the extraction function, M is the task metadata, S is the set of initial samples, and $Ins.G$ is the prompt template used to guide the extraction process. These genes serve as the fundamental units for our subsequent genetic operations.

2.2 Active Parents Selection

In genetic algorithms, the primary objective is to select the fittest pair of parents (p_i^*, p_j^*) for crossover and mutation, thereby producing offspring with enhanced characteristics. However, in synthetic data generation, evaluating fitness of sample pairs is challenging. Active learning, which assumes that samples vary in importance and contribute differently to performance improvements, offers a promising alternative. The standard pool-based active learning paradigm (Ren et al., 2021; Zhang et al., 2023) is designed to extract the most informative and diverse samples from an unlabeled pool to enhance model generalization. Inspired by these insights, we adopt an active learning strategy for parent selection in our framework. Specifically, we extract a previously unused pair of samples from a data pool $\{P, E\}$, where $P = \{p_0, p_1, \dots, p_n\}$ is the current population of size n and $E = \{e_0, e_1, \dots, e_n\}$ denotes their semantic representations². The extracted pair

²We use the Sentence Transformer (Reimers and Gurevych, 2019) to encode textual populations and obtain their semantic representations.

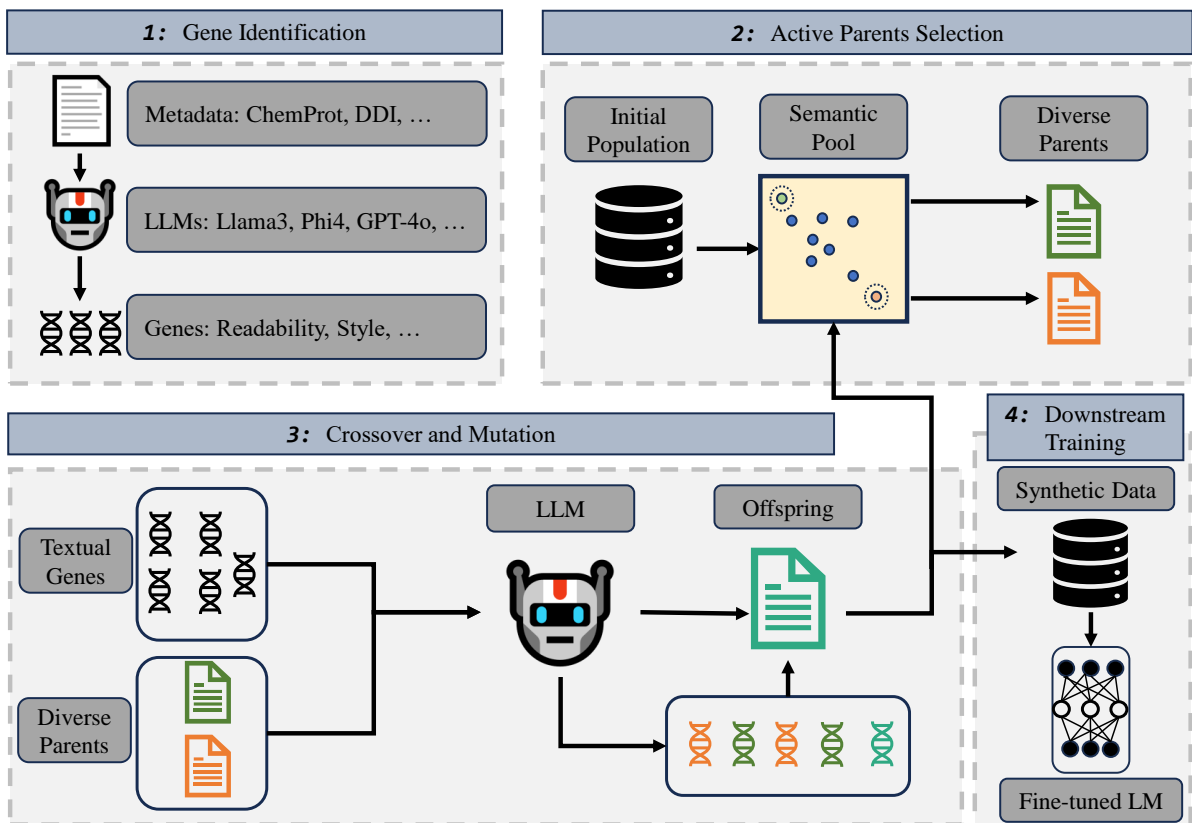


Figure 1: Overview of our Genetic Prompt framework.

of samples is the one that exhibits the largest semantic distance, thereby maximizing the offspring search space during crossover and mutation. A larger semantic distance indicates that the selected samples differ significantly in meaning, increasing the likelihood of generating more diverse offspring when combined. To quantify this, we compute the Euclidean distance between their semantic representations and choose the pair with the greatest distance, thus broadening the search space during the evolutionary process. The parents selection process can be formulated as:

$$(p_i^*, p_j^*) = \arg \max_{e_i, e_j \in E} \text{dist}(e_i, e_j)$$

2.3 Crossover and Mutation

We denote the genetic algorithm prompts for class l as Ins_{GA}^l and refer $\rho(\cdot)$ to the generation process from LLMs, where the distribution of generated text is conditioned on the prompt and other input parameters. In each iteration, we leverage an LLM as an evolution simulator, performing crossover and mutation processes to generate new textual offsprings. The overall generation process can be

formalized as:

$$\mathcal{D}_{\text{syn}} \sim \rho_{l \sim \mathcal{Y}}(\cdot; \text{Ins}_{GA}^l, (p_i^*, p_j^*), G)$$

, where Ins_{GA}^l is the genetic algorithm prompt template for label l , (p_i^*, p_j^*) are the selected parent samples, G represents the identified textual genes. Specifically, we utilize the two selected parent examples (p_i^*, p_j^*) from the current population and perform the *crossover and mutation process* until we reach the target population size N .

Crossover: To fully utilize the diverse “genes” (e.g., paragraph structure, entity ordering, semantic stance) embedded in the parent texts, we randomly sample and blend key elements from both parents to form a hybrid offspring. We randomly partition the identified textual genes G into three groups: $G = G_1 \cup G_2 \cup G_3$, where genes in G_1 are inherited from parent p_i^* , genes in G_2 are inherited from parent p_j^* , and genes in G_3 will undergo mutation. This step ensures that the offspring text preserves the primary content of its parents while organically combining their structure and expressions.

- **Parent 1:** “We investigated the synergistic effects of *Compound A* on *Protein B*. Preliminary results suggest partial agonistic activity.”

Data	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
Train	120,000	27,086	1,069	1,482	6,590	922	2168	1,000
Test	7,600	2,494	1,041	339	2,263	288	662	100
Class	4	50	5	4	9	5	-	-
Domain	News	Science	Biomedicine	Pharmacology	Web	News	Science	Medical
Task	CLS	CLS	RE	RE	RE	RE	ABS	ABS
Imbalance Ratio	1.00	1,283.77	7.16	8.08	1.99	1.51	-	-

Table 1: Statistics of eight datasets from three different NLP tasks with varying data sizes and domains. CLS denotes text classification, RE refers to relation extraction, and ABS means text summarization. Imbalance Ratio represents the ratio of largest to smallest class size.

- **Parent 2:** “Our study shows that *Drug X* significantly activates *Receptor Y* in neuronal cells, indicating a potent agonist relationship.”
- **Offspring (via Crossover):** “We investigated how *Drug X* interacts with *Receptor Y* in neuronal cells. Preliminary results suggest partial agonistic activity.”

In this illustration, *Drug X* and *Receptor Y* are inherited from Parent 2, while the sentence structure and conclusion are drawn from Parent 1. By inheriting crucial “genes” from parents, the newly generated text retains essential content and has the potential to introduce new textual variations, broadening coverage and diversity in the dataset.

Mutation: Unlike token- or sentence-level modification (e.g., swapping words or paraphrasing), our mutation process targets the *semantic-level* identified in Section B.5. Specifically, we randomly alter the genes in G_3 that were not inherited during crossover—such as positioning or relationship between entities, polarity or sentiment of a statement, or text’s functional role—while keeping the overall meaning relevant to the task. The mutated offspring equip more diverse semantic attributes than simple lexical substitutions and thus can diverge more meaningfully from the parent samples with broaden the evolutionary search space.

2.4 Downstream Training

We apply synthetic data D_{Syn} for downstream tasks using pre-trained language models and cross-entropy as the loss function. To validate the effectiveness of synthetic datasets, we follow a standard fine-tuning process by a uniform learning rate to both the pre-trained and prediction layers without any warm-up or decay.

3 Experimental Settings

3.1 Datasets

We use eight publicly available datasets: AG-News (Zhang et al., 2015), StackExchange (Geigle et al., 2021), Chemprot (Taboureau et al., 2010), DDI (Herrero-Zazo et al., 2013), Semeval2010 (Hendrickx et al., 2019), Conll04 (Roth and Yih, 2004), SciTLDR (Cachola et al., 2020) and MeQSum (Ben Abacha and Demner-Fushman, 2019). We summarize data statistics in Table 1 with more details of these benchmarks in Appendix B.1.

3.2 Technical Details

We experiment with open-sourced (e.g., Phi4 (Abdin et al., 2024) and Llama3.1-70B (Dubey et al., 2024)) and proprietary LLMs (GPT-3.5-turbo and GPT-4o (OpenAI, 2022)) as our data generators. And we set the temperature and p-value of all LLMs to 1 to ensure reproducibility. Our empirical analysis in Sec 5 uses multiple LLMs (Llama3.2-3B, Llama3.1-8B, and Llama3.1-70B) trained on the same data source. For each dataset, we generate the same number of samples across different methods and ablation settings. Unless specified otherwise, Llama3.1-70B is our default data generator, and RoBERTa-base (Liu et al., 2019) and T5-large (Raffel et al., 2020) are downstream models for classification and summarization, respectively.

3.3 Baselines

Since our Genetic Prompt requires a small amount of real data to initialize the population, for a fair comparison, we compare the genetic prompt against three approaches that can be integrated with the few-shot setting: 1) **SimPrompt** (Ye et al., 2022; Yu et al., 2023), 2) **AttrPrompt** (Yu et al., 2023) and 3) **Curated LLM (CLLM)** (Seedat et al., 2024). To keep consistence, all the baselines

are also provided with two examples randomly selected from the same pool of real samples used for initializing our Genetic Prompt. For *SimPrompt*, we use its original class-conditional setting, where prompts guide text generation based on predefined categories. SimPrompt follows a template-based approach, such as: *The {class_name} relation means {label_def}, Your task is to write a sentence about {class_name} relation between 2 entities.* Since **AttrPrompt** is a close work using textual attributes, we ensure consistency by adopting the textual genes used in Genetic Prompt as its required attributes. Following original settings of *AttrPrompt*, we then prompt *GPT-4-Turbo* to generate and select specific attribute values and then form various unique combinations. For example, given the possible values for each attribute in Conll04 data (e.g., five different writing styles and five different domains), the total number of unique combinations exceeds 4,000. For the *Curated LLM (CLLM)* approach, we apply the proposed learning dynamics method to curate the over-generated dataset by *Simprompt*.

Data	Method	APS ↓	Intra APS ↓	Inter APS ↓	CMD ↓	Vocab.
AGnews	SimPrompt	0.167	0.325	0.113	0.853	33,249
	AttrPrompt	0.205	0.294	0.174	1.073	29,049
	Curated LLM	0.250	0.324	0.224	1.203	31,051
	Ours	0.129	0.260	0.084	0.716	24,459
	Gold	0.029	0.076	0.013	-	124,673
StackExchange	SimPrompt	0.377	0.611	0.372	0.706	127,595
	AttrPrompt	0.306	0.576	0.300	0.741	46,791
	Curated LLM	0.323	0.488	0.319	0.520	119,482
	Ours	0.324	0.491	0.321	0.595	81,948
	Gold	0.231	0.256	0.225	-	106,967
Chemprot	SimPrompt	0.423	0.660	0.364	0.819	1,847
	AttrPrompt	0.432	0.461	0.425	0.720	2,115
	Curated LLM	0.425	0.658	0.366	0.819	1,814
	Ours	0.389	0.510	0.359	0.628	4,596
	Gold	0.326	0.352	0.314	-	4,123
DDI	SimPrompt	0.572	0.691	0.532	1.077	1,240
	AttrPrompt	0.516	0.542	0.508	0.925	1,556
	Curated LLM	0.569	0.689	0.530	1.065	1,263
	Ours	0.456	0.526	0.432	0.706	3,524
	Gold	0.397	0.410	0.390	-	3,472
Semeval2010	SimPrompt	0.407	0.618	0.380	1.188	3,093
	AttrPrompt	0.451	0.499	0.445	1.173	5,174
	Curated LLM	0.405	0.617	0.380	1.188	3,049
	Ours	0.398	0.460	0.390	1.093	9,425
	Gold	0.248	0.280	0.244	-	19,268
Conll04	SimPrompt	0.352	0.587	0.239	1.140	2,683
	AttrPrompt	0.379	0.444	0.362	0.973	3,850
	Curated LLM	0.358	0.599	0.297	1.163	2,653
	Ours	0.292	0.385	0.269	0.852	8,381
	Gold	0.177	0.231	0.162	-	6,563
SciTLDR	SimPrompt	0.746	-	-	3.600	2,491
	AttrPrompt	0.375	-	-	2.436	5,527
	Curated LLM	-	-	-	-	-
	Ours	0.440	-	-	0.499	11,879
	Gold	0.386	-	-	-	17,213
MeQSum	SimPrompt	0.316	-	-	0.798	4,514
	AttrPrompt	0.296	-	-	0.768	5,211
	Curated LLM	-	-	-	-	-
	Ours	0.316	-	-	0.834	7,268
	Gold	0.176	-	-	-	8,483

Table 2: Intrinsic analysis of synthetic data generation. ↓ indicates that lower values are better. We **Bold** the best scores except for “Gold”, the original data.

4 Results

To evaluate the effectiveness of our Genetic Prompt method, we conducted four major studies, including intrinsic analysis of data quality, extrinsic evaluations on downstream tasks, integrative evaluation with real-world data, and ablation studies on individual modules. We also report extensive performance results and a case study in Appendix A.

4.1 Intrinsic Analysis of Synthetic Data

In an ideal scenario, synthetic data should faithfully reproduce the statistical characteristics of the real-world data it aims to emulate. Therefore, we employ multiple metrics to quantify the semantic diversity of synthetic data, as well as their distribution shift relative to the gold dataset. These metrics include average pairwise sample similarity (APS), Central Moment Discrepancy (CMD) (Zellinger et al., 2022) and vocabulary size, which have detailed definition in Appendix C.

We report intrinsic evaluation results in Table 2. Among the synthetic generation methods, our Genetic Prompt stands out by achieving the best performance in APS, CMD, and vocabulary size, indicating that our approach can capture more semantic characteristics of the real-world data. Notably, the gold data exhibits significantly lower APS scores compared to all synthetic datasets, which suggests that the inherent semantic diversity of real data is more salient. On three datasets of Chemprot, DDI, and Conll04, the vocabulary size of our generated synthetic data even surpasses that of the gold data, indicating that our genetic framework may augment data semantic and lexical diversity while closely mirror the distribution of real data. We conduct a case study to examine the homogeneity issues in Appendix D.

4.2 Extrinsic Evaluation on NLP Tasks

We conduct extrinsic evaluations on three standard NLP tasks on eight data with varying domains, such as biomedical and newspaper. To examine generalizability, we evaluate multiple LLMs, adopt task specific metrics (e.g., F1 and Rouge scores) from the baselines (Ye et al., 2022; Yu et al., 2023), and report results in Table 3.

Our Genetic Prompt consistently outperforms SimPrompt, AttrPrompt, and Curated LLM across major dataset. In general, adopting GPT-4o as the data generator achieves better performance than other LLMs, which indicates a stronger LLM base

Method	Model	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
		Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Rouge-L	Rouge-L
SimPrompt	Phi4	78.9 _{0.6}	48.1 _{0.4}	65.9 _{3.7}	49.1 _{1.5}	58.0 _{1.3}	65.0 _{2.0}	25.5 _{0.0}	23.4 _{0.3}
	Llama3.1-70B	82.3 _{1.9}	51.8 _{0.3}	53.5 _{4.2}	57.5 _{1.3}	56.0 _{1.8}	54.5 _{4.5}	26.0 _{0.1}	22.6 _{0.2}
	GPT-3.5-Turbo	71.8 _{0.3}	40.2 _{0.1}	69.5 _{1.6}	62.6 _{1.1}	64.8 _{0.4}	76.9 _{2.0}	25.7 _{0.3}	23.5 _{0.2}
	GPT-4o	81.8 _{1.0}	52.2 _{1.5}	72.0 _{2.0}	60.1 _{1.7}	68.1 _{0.9}	71.2 _{2.2}	26.0 _{0.2}	23.0 _{0.2}
AttrPrompt	Phi4	79.4 _{2.0}	50.3 _{0.0}	66.1 _{0.8}	68.3 _{0.6}	62.8 _{1.0}	61.1 _{1.4}	26.3 _{0.0}	25.2 _{0.1}
	Llama3.1-70B	82.9 _{0.5}	49.0 _{0.7}	66.0 _{1.4}	63.8 _{0.9}	67.1 _{0.8}	66.4 _{1.6}	26.4 _{0.0}	24.0 _{0.0}
	GPT-3.5-Turbo	80.0 _{0.7}	47.9 _{0.9}	68.1 _{2.5}	62.1 _{5.0}	71.0 _{0.8}	68.8 _{2.3}	26.4 _{0.1}	25.5 _{0.3}
	GPT-4o	81.3 _{1.9}	49.2 _{0.5}	73.6 _{2.5}	61.9 _{2.8}	71.1 _{0.8}	73.3 _{3.6}	26.2 _{0.1}	26.1 _{0.2}
Curated LLM	Phi4	79.7 _{0.6}	45.5 _{1.2}	71.2 _{2.4}	51.2 _{1.4}	55.4 _{2.2}	65.0 _{3.0}	-	-
	Llama3.1-70B	82.5 _{1.0}	50.0 _{0.4}	58.3 _{3.8}	64.1 _{3.5}	55.5 _{0.7}	49.8 _{1.6}	-	-
	GPT-3.5-Turbo	78.3 _{0.5}	43.3 _{0.4}	71.7 _{0.9}	63.8 _{2.6}	65.1 _{0.9}	73.3 _{1.4}	-	-
	GPT-4o	82.3 _{1.4}	50.4 _{0.6}	71.1 _{3.5}	57.2 _{2.6}	68.9 _{0.4}	74.5 _{2.3}	-	-
Ours	Phi4	80.8 _{1.6}	55.2 _{0.4}	75.3 _{0.6}	62.2 _{1.0}	66.1 _{0.5}	73.4 _{2.8}	27.3 _{0.1}	27.6 _{0.1}
	Llama3.1-70B	84.0 _{1.3}	51.2 _{0.9}	77.2 _{0.7}	68.7 _{3.0}	72.3 _{1.4}	72.8 _{1.7}	27.5 _{0.0}	27.4 _{0.0}
	GPT-3.5-Turbo	83.4 _{0.3}	50.2 _{0.2}	77.9 _{2.1}	68.7 _{2.0}	73.2 _{0.8}	67.5 _{3.4}	25.6 _{0.0}	25.9 _{0.0}
	GPT-4o	86.7 _{0.8}	49.8 _{0.4}	81.6 _{1.9}	70.6 _{3.6}	78.9 _{0.5}	85.3 _{1.2}	27.8 _{0.1}	27.2 _{0.3}

Table 3: Extrinsic evaluation results on three downstream NLP tasks over 8 datasets. Each task was run three times randomly, and the subscripts are the standard deviations. We **bold** the best overall performance and the best performance among open-source LLMs.

is critical. As the size of LLMs increases, the performance of our Genetic Prompt steadily improves, whereas other methods exhibit instability with occasionally significant declines (e.g., AttrPrompt plus GPT-4o drops by 6.4% F1 score on DDI compared to Phi4). Lastly, while proprietary models generally outperform smaller open-source models, our Genetic Prompt using Phi4-14B model has either surpassed or closely matched the performance of baseline methods employing GPT-4o, which proves the effectiveness of our approach.

4.3 Synthetic-Gold Data Fusion

Will combining real-world and synthetic data improve downstream model performance? To verify so, we merged the synthetic dataset with the original training set in equal proportions to train downstream models with detailed results in Table 4.

The fusion of synthetic and gold data demonstrates clear benefits for model performance across most tasks. Our Genetic Prompt consistently yields the greatest improvement among all baselines, achieving an average micro-F1 gain of 1.85% and proving to be the only method that successfully benefits text summarization tasks (with +0.5 Rouge-L on SciTLDR and maintaining performance on MeQSum).

It is worth noting that our method shows particularly strong performance on class-imbalanced datasets, with significantly higher gains in macro-F1 compared to micro-F1 scores. On ChemProt, we achieve +3.2% macro-F1 versus +2.3% micro-F1; on DDI, +3.1% macro-F1 versus +1.0% micro-

F1; and on StackExchange, +3.7% macro-F1 versus +2.8% micro-F1. This pattern indicates improved performance across all classes, especially benefiting minority classes. In contrast, baseline methods often struggle with imbalanced datasets. SimPrompt shows negative performance on DDI (-0.6% macro-F1, -1.1% micro-F1), while AttrPrompt achieves only modest gains. Our method’s consistent positive improvements across all imbalanced datasets highlight its robustness.

We attribute these improvements to our genetic algorithm’s ability to generate completely balanced synthetic data with high intra-class diversity. While original datasets suffer from class imbalance, our crossover and mutation operations create diverse synthetic samples for each class, providing richer pattern variations that are particularly valuable for underrepresented classes. By providing abundant and diverse examples for each class, this balanced augmentation helps models develop better understanding of minority class patterns. This leads to more equitable performance across all classes, particularly benefiting applications where class imbalance is a persistent challenge.

4.4 Ablation Study

To assess contributions of the individual modules within our Genetic Prompt framework, we conducted an ablation study by: 1) replacing the active learning-based parent selection with a random selection mechanism while allowing parent reuse (“w/o Active Learning”), 2) entirely removing the mutation operation (“w/o Mutation”), and 3) replac-

Method	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Rouge-L	Rouge-L
Gold	91.9	77.2	82.1	84.9	92.4	88.0	29.6	28.7
Gold + SimPrompt	91.5 (-0.4)	78.6 (+1.4)	83.8 (+1.7)	83.8 (-1.1)	92.8 (+0.4)	91.2 (+3.2)	27.1 (-2.5)	24.4 (-4.3)
Gold + AttrPrompt	92.0 (+0.1)	79.9 (+2.7)	83.0 (+0.9)	84.6 (-0.3)	92.4 (+0.0)	89.4 (+1.4)	28.9 (-0.7)	28.2 (-0.5)
Gold + Curated LLM	90.9 (-1.0)	78.0 (+0.8)	83.9 (+1.8)	85.4 (+0.5)	92.2 (-0.2)	90.9 (+2.9)	-	-
Gold + Ours	92.3 (+0.4)	80.0 (+2.8)	84.4 (+2.3)	85.9 (+1.0)	92.9 (+0.5)	92.1 (+4.1)	30.1 (+0.5)	28.7 (+0.0)

Method	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Rouge-1	Rouge-1
Gold	91.9	75.8	60.7	80.0	92.0	88.5	33.4	31.3
Gold + SimPrompt	91.4 (-0.5)	78.0 (+2.2)	63.1 (+2.4)	79.4 (-0.6)	92.5 (+0.5)	91.6 (+3.1)	31.5 (-1.9)	26.7 (-4.6)
Gold + AttrPrompt	92.1 (+0.2)	79.4 (+3.6)	63.4 (+2.7)	81.0 (+1.0)	92.1 (+0.1)	89.7 (+1.2)	33.1 (-0.3)	31.0 (-0.3)
Gold + Curated LLM	90.8 (-1.1)	77.7 (+1.9)	64.0 (+3.3)	82.3 (+2.3)	92.0 (+0.0)	91.1 (+2.6)	-	-
Gold + Ours	92.3 (+0.4)	79.5 (+3.7)	63.9 (+3.2)	83.1 (+3.1)	92.7 (+0.7)	92.5 (+4.0)	34.8 (+1.4)	31.7 (+0.4)

Table 4: Gold–synthetic data fusion results with absolute percentage improvements. Top: Micro-F1 / Rouge-L. Bottom: Macro-F1 / Rouge-1.

Method	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Micro-F1	Rouge-L	Rouge-L
w/o Active Learning	82.4	50.0	71.7	67.7	68.8	69.3	22.7	19.0
w/o Mutation	81.9	49.4	74.0	67.3	70.8	51.4	26.0	24.1
Word as Gene	71.5	40.5	52.7	54.0	54.4	56.6	22.5	23.2
Ours	84.0	51.2	77.2	68.7	72.3	72.8	27.5	27.4

Table 5: Ablation study of the individual modules.

ing the semantic genes with a traditional word-level gene (“Word as Gene”). For the “Word as Gene” configuration, LLMs are not required to understand the semantic features of parent sequences. Instead, they utilize linguistic features to perform crossover and mutation operations. We perform the experiments on the Llama3.1-70B.

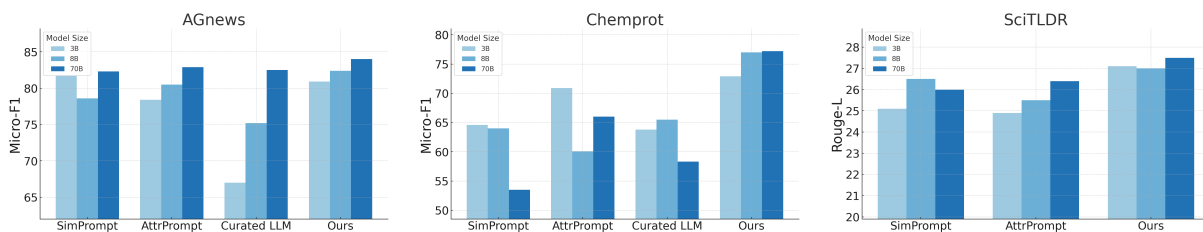
As shown in Table 5, our complete method consistently outperforms all ablated versions across all datasets. Without the mutation operation, the LLM is constrained in its ability to introduce sufficient perturbations during offspring generation, thereby limiting the exploration of the generative search space and reducing offspring diversity. Similarly, the absence of the active learning mechanism makes the parent selection process to rely on random sampling, which may repeatedly select highly similar instances. Consequently, as these similar offspring are incorporated into the population, the likelihood of selecting highly similar parents in subsequent generations increases, thereby reinforcing the cycle of performance decline. Finally, the “Word as Gene” module focusing on word-level crossover and mutation significantly restricts the generative process and reduces downstream NLP models. The finding verifies our assumption

and analysis in Section 2 that our semantic-level approach is more effective than word-level augmentation for high-quality synthetic data; and our semantic-level genetic algorithm may enhance data diversity and lead to downstream improvements.

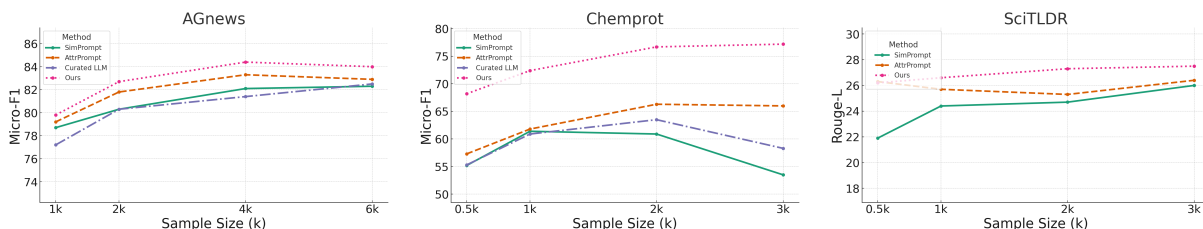
5 Effects Analysis of Generator Size and Data Scale

As data varies in Table 1 in scale and sizes, thus a concrete question to be examined is: how may data size and scale impact on synthetic data generation and their downstream applications? To answer the question, we conducted experiments on three models that share the same architecture, training methods, and even the same training data distribution: Llama3.2-3B, Llama3.1-8B, and Llama3.1-70B. We report performance variations across data scale and size in Figure 2 with other tasks and baseline results in Figure 3 and Figure 4 under Appendix.

Effects of Data Generator Size. As shown in Figure 2a, different datasets exhibit varying sensitivity to generator model size. On the ChemProt data (relation extraction), performance of our Genetic Prompt positively correlates with the size of



(a) Effects of generator size (Llama3.2-3B, Llama3.1-8B, Llama3.1-70B) across three tasks: text classification (AGNews), relation extraction (ChemProt), and text summarization (SciTLDR).



(b) Effects of synthetic data scale across three tasks: text classification (AGNews), relation extraction (ChemProt), and text summarization (SciTLDR).

Figure 2: Effects analysis of generator size and synthetic data scale across diverse NLP tasks.

the Llama models, while SimPrompt and Curated LLM exhibit the opposite trend, and AttrPrompt shows a decrease in performance using the middle-size Llama model. On the AGNews data (text classification), methods show large performance gaps on smaller models (3B, 8B) but converge to similar performance levels on the largest model (70B), suggesting that sufficient model capacity can compensate for methodological differences in this task. On the SciTLDR data (text summarization), moderate sensitivity is observed with some performance fluctuations across model sizes. These observations reveal that the impact of generator size is highly task-dependent. Larger generators do not universally lead to improved performance on downstream tasks, while our method demonstrates consistent robustness across models of various sizes, with notable advantages on tasks that are more sensitive to model scale such as relation extraction.

Effects of Synthetic Data Scale. Our Genetic Prompt demonstrates consistent improvement as synthetic data size increases across all three tasks in Figure 2b. In contrast, other methods show different scaling patterns: on AGNews, all methods steadily improve with increasing data scale; on ChemProt, SimPrompt, AttrPrompt, and Curated LLM plateau or decline after reaching peak performance around 1-2k samples; on SciTLDR, most methods show relatively stable performance with modest improvements.

We also observe two key patterns across differ-

ent methods and data. First, although different methods achieve varying peak performance levels on each dataset, they tend to reach their optimal performance at similar data scales per task. Second, different datasets exhibit significantly different optimal synthetic data scales despite having comparable task complexity - for instance, while AGNews (4 classes) and ChemProt (5 classes) have similar numbers of categories, AGNews reaches peak performance around 4k samples while ChemProt peaks at approximately 2k samples.

The superior scalability of our method, particularly evident on ChemProt where other methods suffer from performance degradation at larger scales, suggests better data diversity and reduced overfitting in our generated samples. This is consistent with the results in Table 2.

6 Related Work

6.1 Genetic Algorithm with LLMs

LLMs possess deep domain knowledge and text processing capabilities, while genetic algorithms excel at global search and iterative optimization. To harness these complementary strengths, the main collaboration paradigms involve either using LLMs to enhance genetic algorithms or employing genetic algorithms to refine LLM outputs (Wu et al., 2024b; Petke et al., 2018). These approaches have been broadly applied in code generation (Lehman et al., 2022), summary generation, and neural network architecture search (Chen et al., 2023). For instance,

LLMs can serve as evolution strategies (Lange et al., 2024) or evolutionary operators (Guo et al., 2024; Yang et al., 2024) in black-box prompt optimization. Furthermore, if sentences or paragraphs are viewed as gene sequences (Liu et al., 2024), one can employ a hierarchical genetic algorithm on a set of prompt prefixes to identify the optimal prompt prefix for a given task.

However, these works share common limitations: they rely on syntactic crossover and mutation operations at the word or sentence level, converging toward a single optimal output (Liu et al., 2024). Therefore, this focus on finding one best candidate limits the ability to maintain both quality and diversity when generating large volumes of text. To address these shortcomings, our Genetic Prompt framework shifts the notion of “genes” from syntactic units (words or sentences) to higher-level semantic attributes, such as writing style and sentence structure. In doing so, we harness the comprehension and reasoning capabilities of LLMs to perform crossover and mutation, enabling us to simultaneously achieve high quality and diversity at scale. Moreover, instead of relying on a conventional fitness-based parent selection process, which can be difficult to evaluate at the individual sample level, we integrate an active learning approach (Zhang et al., 2023; Ren et al., 2021). In each iteration, we pair previously unused, diverse parents to expand offspring search space, thereby enhancing both the quality and variety of the generated text.

6.2 Synthetic Data Generation with LLMs

The emergence of LLMs has established them as key drivers of synthetic data generation across diverse domains (Long et al., 2024). These applications span multiple areas, including text classification tasks (Han et al., 2024; Xu et al., 2024; Liu et al., 2025), tabular data synthesis (Zhang et al., 2025b; Yang et al., 2025), retrieval query generation (Lu et al., 2025), benchmark construction (Zhang et al., 2025a; Wu et al., 2024a), and LLM post-training (Chen et al., 2025; Zhou et al., 2025; Lan et al., 2025; Wang et al., 2025a,b).

However, simply relying on LLM-generated data neither guarantees quality and diversity (Yu et al., 2023) nor prevents the introduction of biases inherent to these models. To overcome these limitations, various enhancement strategies have been proposed, including multi-step generation (Cui and Wang, 2024), in-context learning (Li et al., 2024a;

Xu et al., 2024), label augmentation (Xiao et al., 2023), and conditional prompting (Yu et al., 2021; Russo et al., 2020; Logeswaran et al., 2018). For instance, some approaches extract and predefine text attributes (Yu et al., 2023) or integrate external knowledge (Xu et al., 2024; Rao et al., 2025), leveraging this information as constraints to guide data generation. While these conditional prompting methods effectively raise the baseline level of text diversity, they also introduce drawbacks. In particular, imposing numerous constraints can hinder an LLM’s reasoning and comprehension, limiting its ability to learn from context and accurately reflect real-world patterns and distributions.

To address these issues, Genetic Prompt abandons explicit conditional constraints. Instead, it treats the semantic attributes of text as genes in a genetic algorithm, thereby better leveraging the LLM’s inherent knowledge and comprehension abilities to generate high-quality data from a limited set of real examples. Moreover, by incorporating crossover and mutation operations, Genetic Prompt introduces sufficient perturbations to enhance the diversity and scalability of the synthetic data, effectively reducing the risk of overfitting in downstream models.

7 Conclusion

In this work, we introduced the Genetic Prompt Framework, a novel approach combining LLMs and genetic algorithms for synthetic data generation. Experiments on 8 datasets from multiple domains and NLP tasks demonstrate the effectiveness of Genetic Prompt, as it consistently outperformed three state-of-the-art baselines across multiple settings. Our ablation and effect analysis have summarized multiple insights for future studies on generating synthetic data or applying synthetic data over downstream tasks: 1) semantic-level generation of synthetic data can capture more data characteristics and thus promote synthetic data quality, 2) combining synthetic and real-world data generally boost model performance but rely on LLM options and prompting methods, and 3) synthetic data play a more critical role in downstream tasks with smaller and domain-specific data. Our future work will further expand and verify the potentials to apply our approach to broader domains, modalities, and applications, such as synthetic generation for health tabular data and training small language models.

Limitations

In this study, we propose Genetic Prompt for LLMs to generate synthetic data. Despite the strong performance and good data quality achieved, several limitations should be acknowledged: First, our experiments were conducted exclusively on English corpora. Other languages may possess distinct linguistic features and characteristics that differ from English, potentially affecting the quality of synthetic data generation. The effectiveness of our approach across diverse languages remains to be validated. Second, our approach was specifically designed for textual data synthesis. Extending this work to other modalities such as tabular, or to multi-modal data synthesis scenarios would likely require substantial adaptations to the methodology.

Acknowledgment

The authors thank anonymous reviewers for their insightful feedback. This project was supported by the National Science Foundation (NSF) IIS-2245920 and CNS-2318210. We thank for the computing resources provided by the iTiger GPU cluster³ (Sharif et al., 2025) supported by the NSF MRI program. We would also thank for additional funding from the ITS and College of Arts and Sciences at the University of Memphis to partially support daily HPC operations.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. [On the summarization of consumer health questions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.
- Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. [TLDR: Extreme summarization of scientific documents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.
- Angelica Chen, David Dohan, and David So. 2023. [Evo-prompting: Language models for code-level neural architecture search](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 7787–7817. Curran Associates, Inc.
- Jiawei Chen, Xinyan Guan, Qianhao Yuan, Guozhao Mo, Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, and Xianpei Han. 2025. [Consistentchat: Building skeleton-guided consistent dialogues for large language models from scratch](#). *arXiv preprint arXiv:2506.03558*.
- John Chung, Ece Kamar, and Saleema Amershi. 2023. [Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.
- Wanyun Cui and Qianle Wang. 2024. [Ada-instruct: Adapting instruction generators for complex reasoning](#). *Preprint*, arXiv:2310.04484.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. 2021. [Tweac: Transformer with extendable qa agent classifiers](#). *Preprint*, arXiv:2104.07081.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. [Connecting large language models with evolutionary algorithms yields powerful prompt optimizers](#). In *The Twelfth International Conference on Learning Representations*.
- Guangzeng Han, Weisi Liu, Xiaolei Huang, and Brian Borsari. 2024. [Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts](#). In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 392–401.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. [Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals](#). *Preprint*, arXiv:1911.10422.
- María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. [The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions](#). *Journal of Biomedical Informatics*, 46(5):914–920.
- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. [Data-models: Predicting predictions from training data](#). In *Proceedings of the 39th International Conference on Machine Learning*.

³<https://itiger-cluster.github.io/>

- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gasheovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Guangchen Lan, Huseyin A Inan, Sahar Abdelnabi, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher G Brinton, and Robert Sim. 2025. Contextual integrity in LLMs via reasoning and reinforcement learning. *arXiv preprint arXiv:2506.04245*.
- Robert Lange, Yingtao Tian, and Yujin Tang. 2024. Large language models as evolution strategies. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO '24 Companion*, page 579–582, New York, NY, USA. Association for Computing Machinery.
- Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. 2022. Evolution through large models. *Preprint*, arXiv:2206.08896.
- Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024a. Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.
- Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. 2024b. Entropic distribution matching in supervised finetuning of llms: Less overfitting and better diversity. *Preprint*, arXiv:2408.16673.
- Weisi Liu, Guangzeng Han, and Xiaolei Huang. 2025. Examining and adapting time for multilingual classification via mixture of temporal experts. 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)*, pages 6151–6166, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Xuan Lu, Sifan Liu, Bochao Yin, Yongqi Li, Xinghao Chen, Hui Su, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. 2025. Multiconir: Towards multi-condition information retrieval. *Preprint*, arXiv:2503.08046.
- Tom M Mitchell. 1980. The need for biases in learning generalizations.
- OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>. Accessed: 2023-07-24.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMO on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.
- Justyna Petke, Saemundur O. Haraldsson, Mark Harman, William B. Langdon, David R. White, and John R. Woodward. 2018. Genetic improvement of software: A comprehensive survey. *IEEE Transactions on Evolutionary Computation*, 22(3):415–432.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Jun Rao, Zepeng Lin, Xuebo Liu, Xiaopeng Ke, Lian Lian, Dong Jin, Shengjun Cheng, Jun Yu, and Min

- Zhang. 2025. [APT: Improving specialist LLM performance with weakness case acquisition and iterative preference training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20958–20980, Vienna, Austria. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. [A survey of deep active learning](#). *ACM Comput. Surv.*, 54(9).
- Dan Roth and Wen-tau Yih. 2004. [A linear programming formulation for global inference in natural language tasks](#). In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Giuseppe Russo, Nora Hollenstein, Claudiu Cristian Musat, and Ce Zhang. 2020. Control, generate, augment: A scalable framework for multi-attribute text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 351–366.
- Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2024. [Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes](#). In *Forty-first International Conference on Machine Learning*.
- Mayira Sharif, Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. [Cultivating multidisciplinary research and education on gpu infrastructure for mid-south institutions at the university of memphis: Practice and challenge](#). *Preprint*, arXiv:2504.14786.
- Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgård, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak, and Tudor I. Oprea. 2010. [ChemProt: a disease chemical biology database](#). *Nucleic Acids Research*, 39:D367–D372.
- Haozhe Wang, Long Li, Chao Qu, Fengming Zhu, Weidi Xu, Wei Chu, and Fangzhen Lin. 2025a. [To code or not to code? adaptive tool integration for math language models via expectation-maximization](#). *arXiv preprint arXiv:2502.00691*.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025b. [V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning](#). *arXiv preprint arXiv:2504.08837*.
- Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. 2024a. [Unigen: A unified framework for textual dataset generation using large language models](#). *arXiv preprint arXiv:2406.18966*.
- Xingyu Wu, Sheng hao Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024b. [Evolutionary computation in the era of large language model: Survey and roadmap](#). *Preprint*, arXiv:2401.10034.
- Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. [FreeAL: Towards human-free active learning in the era of large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.
- Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. 2024. [Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15496–15523, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. [Large language models as optimizers](#). In *The Twelfth International Conference on Learning Representations*.
- Shuo Yang, Zheyu Zhang, Bardh Prenkaj, and Gjergji Kasneci. 2025. [Doubling your data in minutes: Ultra-fast tabular data generation via llm-induced dependency graphs](#). *Preprint*, arXiv:2507.19334.
- Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. [ZeroGen: Efficient zero-shot learning via dataset generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Dian Yu, Zhou Yu, and Kenji Sagae. 2021. [Attribute alignment: Controlling text generation from pre-trained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268.
- Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. [Large language model as attributed training data generator: A tale of diversity and bias](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.
- Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz.

2022. Central moment discrepancy (cmd) for domain-invariant representation learning. In *International Conference on Learning Representations*.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. 2025a. [Can MLLMs understand the deep implication behind Chinese images?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. [LLMaAA: Making large language models as active annotators](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zheyu Zhang, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2025b. [Not all features deserve attention: Graph-guided dependency learning for tabular data generation with language models](#). *Preprint*, arXiv:2507.18504.

Heng Zhou, Hejia Geng, Xiangyuan Xue, Li Kang, Yiran Qin, Zhiyong Wang, Zhenfei Yin, and Lei Bai. 2025. [Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks](#). *arXiv preprint arXiv:2503.02390*.

A Additional Results

A.1 Supplementary Experimental Results

Due to space limitations, we present the remaining performance metrics of the three experiments **Extrinsic Evaluation on NLP Tasks** and **Ablation Study** in Table 6 and Table 7 respectively.

A.2 The Effect of Data Generator Size

For same space reason, we report additional results of empirical analysis on generator size in Table 8 and Figure 3.

A.3 The Effect of Synthetic Data Scale

we report additional results of empirical analysis on data scale in Table 9 and Figure 4.

Method	Model	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
		Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Rouge-1	Rouge-1
SimPrompt	Phi4	78.3 _{0.5}	45.9 _{0.5}	53.7 _{1.9}	44.7 _{1.5}	55.3 _{2.0}	63.5 _{2.0}	29.5 _{0.1}	25.2 _{0.3}
	Llama3.1-70b	82.0 _{1.5}	50.2 _{0.3}	42.6 _{1.7}	56.5 _{1.6}	53.4 _{1.2}	53.0 _{5.7}	29.8 _{0.1}	24.4 _{0.3}
	GPT-3.5-Turbo	70.8 _{0.5}	39.8 _{0.2}	53.4 _{1.3}	63.3 _{1.4}	62.4 _{0.5}	77.2 _{1.9}	30.2 _{0.3}	25.3 _{0.2}
	GPT-4o	81.7 _{1.1}	50.7 _{1.2}	58.1 _{3.0}	56.3 _{2.1}	66.5 _{0.9}	71.1 _{2.2}	30.0 _{0.2}	24.8 _{0.1}
AttrPrompt	Phi4	78.9 _{2.0}	49.1 _{0.2}	52.1 _{0.4}	68.3 _{0.8}	61.2 _{1.3}	59.3 _{2.0}	30.1 _{0.1}	27.0 _{0.0}
	Llama3.1-70b	82.2 _{0.4}	47.4 _{0.8}	51.8 _{0.8}	62.3 _{0.9}	66.0 _{0.7}	64.9 _{1.9}	30.5 _{0.0}	26.8 _{0.2}
	GPT-3.5-Turbo	79.7 _{0.7}	47.0 _{0.7}	56.1 _{1.9}	63.7 _{2.5}	70.0 _{0.5}	67.8 _{2.8}	30.7 _{0.2}	27.1 _{0.2}
	GPT-4o	80.7 _{1.5}	48.3 _{0.3}	57.7 _{2.1}	60.7 _{2.3}	70.0 _{1.1}	73.1 _{3.9}	30.2 _{0.1}	27.9 _{0.2}
Curated LLM	Phi4	78.9 _{0.4}	43.8 _{1.0}	57.1 _{2.2}	52.2 _{1.4}	53.1 _{2.1}	63.8 _{3.0}	-	-
	Llama3.1-70b	82.0 _{1.1}	48.5 _{0.4}	44.8 _{3.3}	65.4 _{2.0}	52.2 _{0.8}	49.0 _{1.4}	-	-
	GPT-3.5-Turbo	77.2 _{0.3}	42.1 _{0.4}	55.6 _{1.3}	65.2 _{1.2}	62.5 _{1.0}	73.1 _{1.5}	-	-
	GPT-4o	81.6 _{1.6}	49.2 _{0.6}	49.6 _{2.9}	55.7 _{2.9}	67.9 _{0.2}	74.7 _{2.4}	-	-
Ours	Phi4	80.1 _{1.3}	54.6 _{0.2}	56.4 _{1.9}	63.1 _{0.7}	65.2 _{0.5}	72.9 _{2.9}	31.3 _{0.2}	29.2 _{0.1}
	Llama3.1-70b	83.4 _{1.0}	49.9 _{1.0}	58.9 _{1.2}	67.6 _{2.0}	71.3 _{1.5}	72.5 _{1.9}	31.6 _{0.2}	29.2 _{0.2}
	GPT-3.5-Turbo	82.5 _{0.5}	49.3 _{0.3}	62.5 _{1.0}	69.2 _{2.0}	71.0 _{1.0}	66.7 _{3.5}	30.4 _{0.1}	27.3 _{0.1}
	GPT-4o	86.5 _{0.7}	48.0 _{0.3}	65.4 _{1.9}	62.6 _{3.0}	78.1 _{0.5}	85.8 _{1.2}	32.1 _{0.2}	29.0 _{0.2}

Table 6: Additional experimental results of the Genetic Prompt framework and baselines. We **bold** the best overall performance and the best performance among open-source LLMs. Standard deviations are calculated using three random seeds.

Method	AGnews	StackExchange	Chemprot	DDI	Semeval	Conll04	SciTLDR	MeQSum
	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Macro-F1	Rouge-1	Rouge-1
w/o Active Learning	82.3	49.3	57.3	66.2	67.6	69.2	26.2	21.3
w/o Mutation	81.8	49.0	53.8	58.6	69.8	40.8	30.1	26.4
Word as Gene	70.6	38.3	46.7	47.7	52.2	50.5	27.4	24.9
Ours	83.4	49.9	58.9	67.6	71.3	72.5	31.6	29.2

Table 7: Additional ablation study results.

Method	AGNews		StackExchange		Chemprot		DDI		Semeval		Conll		SciTLDR		MeQSum									
	3B	8B	70B	3B	8B	70B	3B	8B	70B	3B	8B	70B	3B	8B	70B	3B	8B	70B						
SimPrompt	81.7	78.6	82.3	46.9	45.5	51.8	64.6	64.0	53.5	47.3	47.4	57.5	39.7	44.4	56.0	58.9	55.8	54.5	25.1	26.5	26.0	22.7	23.1	22.6
AttrPrompt	78.4	80.5	82.9	50.2	48.0	49.0	70.9	60.1	66.0	50.4	58.8	63.8	55.1	61.2	67.1	59.4	69.4	66.4	24.9	25.5	26.4	24.4	23.3	24.0
Curated LLM	67.0	75.2	82.5	45.4	45.2	50.0	63.8	65.5	58.3	46.5	50.3	64.1	43.0	49.5	55.5	50.9	53.4	49.8	-	-	-	-	-	-
Ours	80.9	82.4	84.0	48.4	50.0	51.2	72.9	77.0	77.2	48.8	55.9	68.7	55.2	62.3	72.3	69.2	68.4	72.8	27.1	27.0	27.5	26.9	26.5	27.4

Table 8: Additional results of empirical analysis on generator size. We report Rouge-1 and Macro-F1 for generation and classification tasks separately.

Method	AGNews		StackExchange			Chemprot		DDI		Semeval			Conll04			SciTLDR			MeQSum														
	1k	2k	4k	6k	9k	18k	27k	0.5k	1k	2k	3k	0.5k	1k	2k	3k	0.5k	1k	2k	3k	0.5k	1k	2k	3k										
SimPrompt	78.7	80.3	82.1	82.3	45.4	46.8	49.9	51.8	55.2	61.4	60.9	53.5	49.0	52.9	56.7	57.5	40.8	46.2	52.9	56.0	41.3	47.5	50.9	54.5	21.9	24.4	24.7	26.0	18.1	20.3	21.2	22.6	
AttrPrompt	79.2	81.8	83.3	82.9	46.6	46.5	48.8	49.0	57.3	61.8	66.3	66.0	49.4	55.9	63.0	63.8	43.1	49.6	63.6	67.1	54.7	60.8	61.2	66.4	26.3	25.7	25.3	26.4	19.2	21.9	22.5	24.0	
Curated LLM	77.2	80.3	81.4	82.5	44.7	47.0	49.6	50.0	55.3	60.9	63.5	58.3	52.2	57.4	58.3	64.1	40.5	45.5	55.5	55.5	40.6	44.9	49.2	49.8	-	-	-	-	-	-	-	-	
Ours	79.8	82.7	84.4	84.0	46.2	47.3	49.7	51.2	68.2	72.4	76.7	77.2	55.8	60.4	63.6	68.7	44.9	51.1	65.4	72.3	62.9	72.8	68.8	69.2	72.8	26.2	26.6	27.3	27.5	19.4	23.6	24.9	27.4

Table 9: Additional results of empirical analysis on data scale. We report Rouge-1 and Macro-F1 for generation and classification tasks separately.

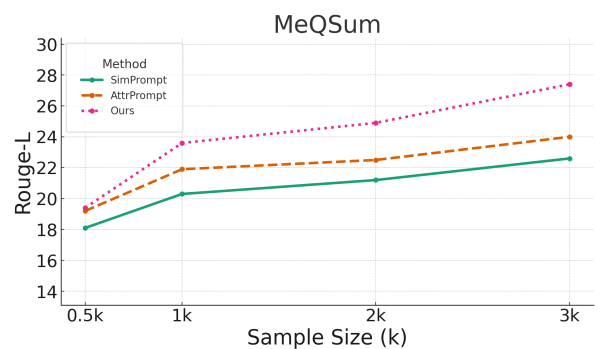
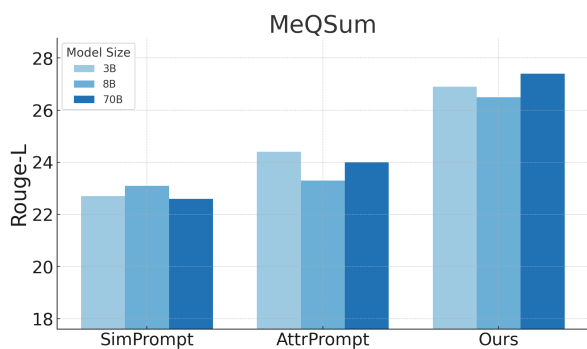
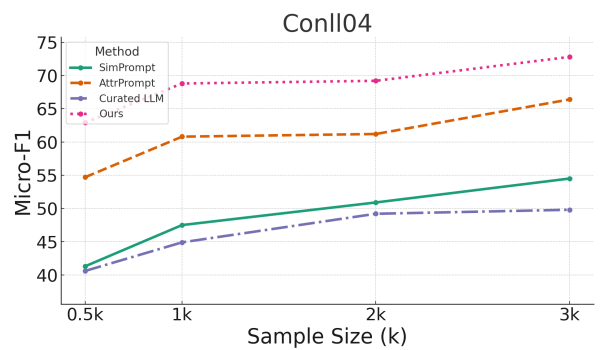
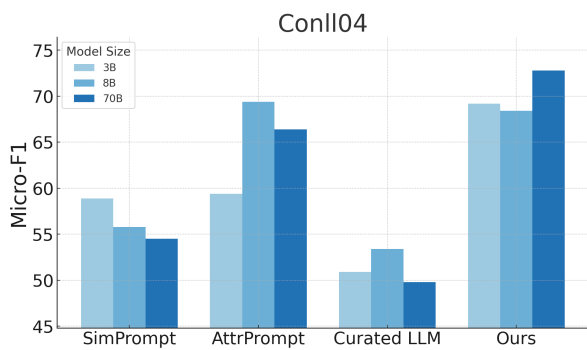
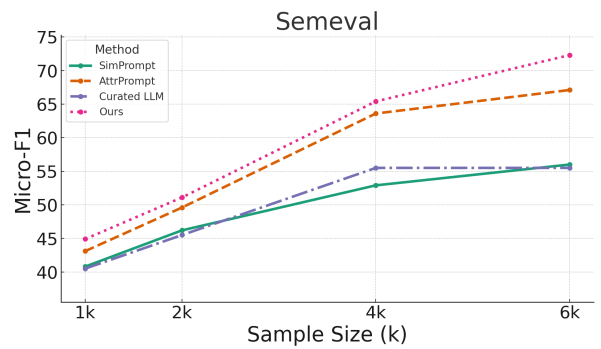
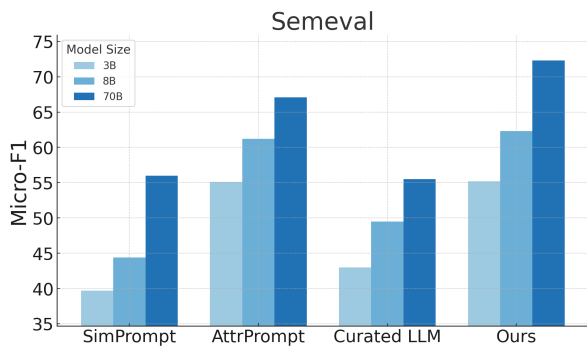
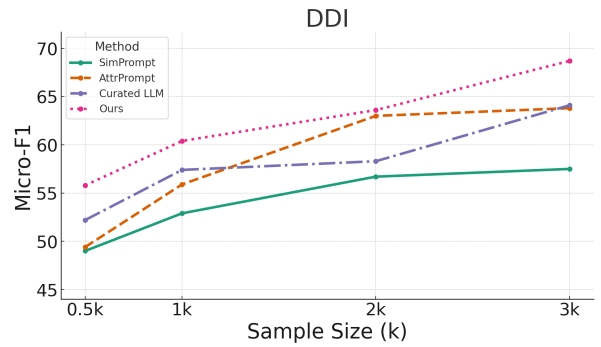
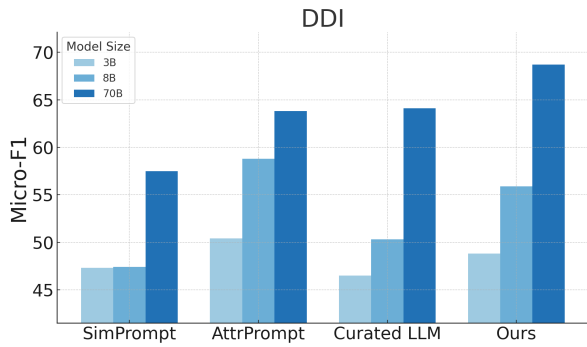
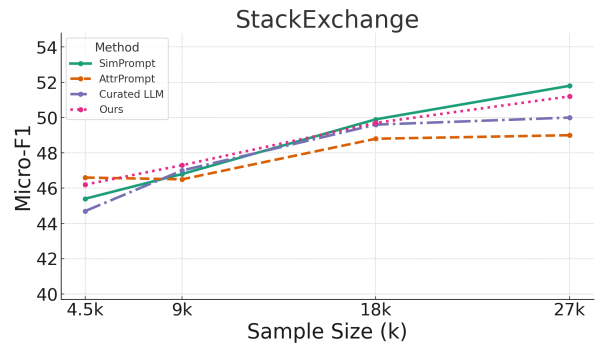
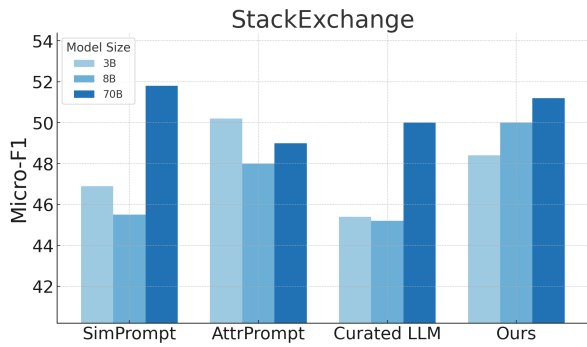


Figure 3: Empirical analysis on generator size

Figure 4: Empirical analysis on sample size.

Algorithm 1 Genetic Algorithm Implemented by LLM for Synthetic Data Generation

- 1: **Input:** Label set L , textual genes G , target data size N , diversity coefficient α .
 - 2: Initialize population $P = \{p_0, p_1, \dots, p_{50}\}$ from gold dataset for each label l .
 - 3: Build a sample pool $\{P, E\}$ for each label l for the initial population.
 $\triangleright E$ are semantic representations of population P .
 - 4: Calculate the diversity score $APS = \{a_0, a_1, \dots, a_l\}$ of each class in gold dataset.
 \triangleright APS: Average **P**air wise sample **S**imilarity.
 - 5: **for** $l \in L$ **do**
 - 6: **while** population $< N$ **do**
 - 7: Get semantic distances of pairs in $\{P, E\}$: $\forall e_i, e_j \in E, d_{ij} = \text{dist}(e_i, e_j)$
 - 8: Select the most distant pair from **unused** samples as parents: $(p_i^*, p_j^*) = \arg \max_{e_i, e_j} d_{ij}$
 - 9: Randomly partition textual genes G into three groups: $G = G_1 \cup G_2 \cup G_3$.
 - 10: Perform LLM-based crossover on G_1 and G_2 using parents (p_i^*, p_j^*) and perform mutation on G_3 : $p_{new} = \text{LLM-GA}(p_i^*, p_j^*, G_1, G_2, G_3)$
 - 11: Add the new sample to pool: $\{P, E\} := \{P, E\} \cup \{p_{new}, e_{new}\}$
 - 12: **end while**
 - 13: Remove the initial population from the sample pool.
 - 14: **end for**
 - 15: **Output:** Synthetic data
-

Dataset	Genes
AGnews	length, location, style, subtopics
StackExchange	length, style, depth, scenario
Chemprot	length, voice, sentence structure, interaction verb, modifier, negation, entity proximity
DDI	length, voice, polarity, interaction verb, modifier, drug mentions, entity proximity
SemEval2010	length, voice, sentence structure, readability, entity distance, domain
CoNLL04	length, voice, sentence structure, readability, entity distance, domain
SciTLDR	length, depth, style, subtopics
MeQSum	length, depth, style, subtopics

Table 10: Selected Textual Genes (Attributes) for Our Genetic Prompt and baseline AttrPrompt.

B Implementation Details

B.1 Datasets

AGNews (Zhang et al., 2015) contains 120,000 categorized news articles from more than 2000 news sources.

StackExchange (Geigle et al., 2021) contains structured technical content from knowledge-sharing platforms, including questions, answers, and user interactions across specialized domains.

Chemprot (Taboureau et al., 2010) is a biomedical dataset derived from PubMed abstracts, focusing on chemical-protein interaction relation extraction, containing sentences annotated with chemical and protein entities and their various fine-grained interaction types. We follow previous work (Peng et al., 2019), use the same 5 classes: Substrate, Up-regulator, Downregulator, Agonist and Antagonist.

DDI (Drug-Drug Interaction) (Herrero-Zazo et al., 2013) is from the semeval-2013-task9 challenge, we filter all sentences that contain drug-drug interactions and follow (Peng et al., 2019) to evaluate the performance of the four types of DDI.

Semeval2010-task8 (Hendrickx et al., 2019) is a multi-Way classification task of Semantic Relations, which contains nine directed relations and the Other relation. To balance the relation types in the selected four datasets, We merge the 9 directed relation pairs into nine undirected symmetric relations and remove the Other relation.

Conll04 (Roth and Yih, 2004) is a relation extraction dataset consisting of news articles with annotations for four entity types and five relation types (Lives-In, Located-In, OrgBased-In, Work-For, and Kill). We utilize the standard training and test sets for evaluating the performance of five relation types.

SciTLDR (Cachola et al., 2020) is a publicly available dataset of over 23,000 scientific papers paired with concise, one-sentence “TL;DR” summaries. In this work, we use their **Abstract-only** setting.

MeQSum (Ben Abacha and Demner-Fushman, 2019) a publicly available dataset of consumer health questions from the MedHelp forum paired with expert-written abstractive summaries.

B.2 Additional Technical Details

During generation, for responses that do not adhere to the instructions, such as ‘*I am just a large language model.*’ or ‘*I cannot do that for you.*’, are filtered out. The hyperparameters can be found in Table 11.

Dataset	Model	Epoch	Learning Rate	Size
AGNews	RoBERTa-base	5	1e-5	6k
StackExchange	RoBERTa-base	5	1e-5	27k
Chemprot	RoBERTa-base	3	1e-5	3k
DDI	RoBERTa-base	3	1e-5	3k
Semeval	RoBERTa-base	3	1e-5	6k
Conll04	RoBERTa-base	3	1e-5	3k
SciTLDR	T5-large	10	3e-5	3k
MeQSum	T5-large	10	3e-5	3k

Table 11: Hyperparameters of main experiments.

B.3 Hardware and Software

We conducted all experiments on a machine equipped with 8x H100 GPUs, 2x EPYC Genoa 9334 CPUs, and 768GB of RAM. The system runs on Linux kernel 5.14. Besides, for our proposed Genetic Prompt framework, we utilize the FAISS library (Johnson et al., 2021) for accelerate calculations. For the Sentence Transformer library (Reimers and Gurevych, 2019) used in our experiments, we utilize the stsb-roberta-large checkpoint as the encoder. Additionally, we deploy open-source LLMs using vLLM (Kwon et al., 2023) library to ensure efficient inference.

B.4 Pseudo Code of Genetic Prompt

We provide the pseudo code of Genetic Prompt in Algorithm 1.

B.5 Selected Textual Genes

The selection of genes for each dataset was tailored to capture the unique characteristics and requirements of their respective domains. Table 10 shows the textual genes selected for the different datasets. Across all four datasets, common genes such as **sentence length** and **voice** were consistently included, reflecting their universal importance in text. However, each dataset incorporated specific genes to address its particular focus. The Chemprot dataset, dealing with protein-chemical relations, included genes like **interaction verb** and **entity proximity** to capture the nuances of biochemical interactions. The DDI (Drug-Drug Interaction) dataset emphasized pharmacological aspects with genes such as **drug mentions** and **interaction verb**. For

the more general relation classification tasks, both the SemEval2010 and CoNLL04 datasets shared a set of genes including **sentence structure**, **readability**, and **entity distance**, which are crucial for understanding the contextual relationships between entities.

B.6 Prompt Template of Genetic Prompt

B.6.1 AGNews

You need to generate synthetic data for the Protein Chemical Relation extraction task.

Example 1: {sample1}

Example 2: {sample2}

Your task is to write a sentence about '{class_name}' relation between chemical and protein.

The two entities should be marked with XML-style tags as <chemical> CHEMICAL </chemical>

and <protein> PROTEIN </protein> respectively in your response.

The sentence should follow the requirements below:

1. The sentence must discuss about the chemical and protein with the relation {label_def}.
2. The '{Gene[0]}' of the sentence should inherit from Example1;
3. The '{Gene[1]}' of the sentence should inherit from Example2;
4. The '{Gene[2]}' of the sentence should inherit from Example1;
5. The '{Gene[3]}' of the sentence should inherit from Example2;
6. The '{Gene[4]}' of the sentence should inherit from Example1;
7. The '{Gene[5]}' of the sentence should inherit from Example2;
8. The '{Gene[6]}' of the sentence and entities must be different from the given 2 examples.

Relation: {class_name}

Text:

B.6.2 StackExchange

You need to generate synthetic data for the News Classification task.

Example 1: {sample1}

Example 2: {sample2}

Your task is to write a news article about '{class_name}' category.

The article should follow the requirements below:

1. The article must be a news about {label_def}.
2. The '{Gene[0]}' of the sentence should inherit from Example1;
3. The '{Gene[1]}' of the sentence should inherit from Example2;
4. The '{Gene[2]}' of the sentence should inherit from Example1;
5. The '{Gene[3]}' of the sentence and entities must be different from the given 2 examples.

Category: {class_name}

Text:

B.6.3 Chemprot

You need to generate synthetic data for the Stack Exchange question classification task.

Example 1: {sample1}
Example 2: {sample2}

Your task is to write a Stack Exchange question about '{class_name}'.

The '{class_name}' category means {label_def}

The question should follow the requirements below:

1. The '{Gene[0]}' of the sentence should inherit from Example1;
2. The '{Gene[1]}' of the sentence should inherit from Example2;
3. The '{Gene[2]}' of the sentence should inherit from Example1;
4. The '{Gene[3]}' of the sentence and entities must be different from the given 2 examples.

Category: {class_name}

Text:

You need to generate synthetic data for Relation Classification task.

Example 1: {sample1}
Example 2: {sample2}

Your task is to write a sentence about '{class_name}' relation between 2 entities.

The '{class_name}' relation means {label_def}

The two entities should be marked with XML-style tags as <entity> ENTITY </entity>.

The sentence should follow the requirements below:

1. The '{Gene[0]}' of the sentence should inherit from Example1;
2. The '{Gene[1]}' of the sentence should inherit from Example2;
3. The '{Gene[2]}' of the sentence should inherit from Example1;
4. The '{Gene[3]}' of the sentence should inherit from Example2;
5. The '{Gene[4]}' of the sentence should inherit from Example1;
6. The '{Gene[5]}' of the sentence and entities must be different from the given 2 examples.

Relation: {class_name}

Text:

B.6.4 DDI

You need to generate synthetic data for Drug-Drug Interaction extraction task.

Example 1: {sample1}
Example 2: {sample2}

Your task is to write a sentence about '{class_name}' relation between drug and drug.

The two drugs should be marked with XML-style tags as <drug> DRUG </drug>. The sentence should follow the requirements below:

1. The sentence must discuss about the drug and drug with the relation {label_def}.
2. The '{Gene[0]}' of the sentence should inherit from Example1;
3. The '{Gene[1]}' of the sentence should inherit from Example2;
4. The '{Gene[2]}' of the sentence should inherit from Example1;
5. The '{Gene[3]}' of the sentence should inherit from Example2;
6. The '{Gene[4]}' of the sentence should inherit from Example1;
7. The '{Gene[5]}' of the sentence should inherit from Example2;
8. The '{Gene[6]}' of the sentence and entities must be different from the given 2 examples.

Relation: {class_name}

Text:

B.6.6 Conll04

You need to generate synthetic data for Relation Classification task.

Example 1: {sample1}
Example 2: {sample2}

Your task is to write a sentence about '{class_name}' relation between 2 entities.

The '{class_name}' relation means {label_def}

The two entities should be marked with XML-style tags as <entity> ENTITY </entity>.

The sentence should follow the requirements below:

1. The '{Gene[0]}' of the sentence should inherit from Example1;
2. The '{Gene[1]}' of the sentence should inherit from Example2;
3. The '{Gene[2]}' of the sentence should inherit from Example1;
4. The '{Gene[3]}' of the sentence should inherit from Example2;
5. The '{Gene[4]}' of the sentence should inherit from Example1;
6. The '{Gene[5]}' of the sentence and entities must be different from the given 2 examples.

Relation: {class_name}

Text:

B.6.5 Semeval2010

B.6.7 SciTLDR

You need to generate synthetic data for the computer science paper abstract summarization task.

Example 1: {sample1}
 Example 2: {sample2}

Your task is to write computer science paper abstract and it's summary. The abstract and summary should follow the requirements below:

1. The '{Gene[0]}' of the sentence should inherit from Example1;
2. The '{Gene[1]}' of the sentence should inherit from Example2;
3. The '{Gene[2]}' of the sentence should inherit from Example1;
4. The '{Gene[3]}' of the sentence and entities must be different from the given 2 examples.

Output:

B.6.8 MeQSum

You need to generate synthetic data for the medical question summarization task.

Example 1: {sample1}
 Example 2: {sample2}

Your task is to write a medical question and it's summary. The abstract and summary should follow the requirements below:

1. The '{Gene[0]}' of the sentence should inherit from Example1;
2. The '{Gene[1]}' of the sentence should inherit from Example2;
3. The '{Gene[2]}' of the sentence should inherit from Example1;
4. The '{Gene[3]}' of the sentence and entities must be different from the given 2 examples.

Output:

C Definition of Diversity and Distribution Metrics

In this work, we use two metrics to evaluate the diversity and similarity within the datasets: Average Pairwise Similarity (APS) and Vocabulary size. We also use the Central Moment Discrepancy (CMD) (Zellinger et al., 2022) between the synthetic and gold data as the measure of distribution shift.

C.1 Average Pairwise Similarity (APS)

The APS measures the overall similarity between samples by averaging the cosine similarity between

all embeddings. Given embeddings e_i and e_j for samples i and j , APS is defined as:

$$\text{APS} = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \cos(e_i, e_j)$$

where N is the total number of samples and $\cos(e_i, e_j)$ represent the cosine similarity between e_i and e_j .

We also compute Intra-Class and Inter-Class APS:

Intra-Class APS measures the similarity between embeddings within the same class:

$$\text{Intra-Class APS} = \frac{1}{|S|} \sum_{(i,j) \in S} \cos(e_i, e_j)$$

where S is the set of sample pairs with the same class label, $label_i = label_j$.

Inter-Class APS measures the similarity between embeddings from different classes:

$$\text{Inter-Class APS} = \frac{1}{|D|} \sum_{(i,j) \in D} \cos(e_i, e_j)$$

where D is the set of sample pairs with different class labels, $label_i \neq label_j$.

C.2 Central Moment Discrepancy (CMD)

Central Moment Discrepancy (Zellinger et al., 2022) measures the difference of probability distributions on a compact interval that considers higher order central moments. For our computations, we define the compact interval as $[-1, 1]^N$ and consider the first five central moments. The CMD between two distributions p and q , with corresponding samples $X = (X_1, \dots, X_n)$ from p and $Y = (Y_1, \dots, Y_n)$ from q , is calculated as:

$$\text{CMD}(p, q) = \frac{1}{2} \|\mathbb{E}(X) - \mathbb{E}(Y)\|_2 + \sum_{k=2}^5 \frac{1}{2^k} \|\mathbf{c}_k(X) - \mathbf{c}_k(Y)\|_2$$

where $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ represent the expectations (means) of the samples, and $\mathbf{c}_k(X)$ and $\mathbf{c}_k(Y)$ are the central moments of order k for X and Y , respectively. Each term $\mathbf{c}_k(X)$ is defined as:

$$\mathbf{c}_k(X) = \left(\mathbb{E} \left(\prod_{i=1}^N (X_i - \mathbb{E}(X_i))^{r_i} \right) \right)$$

where $r_1 + \dots + r_N = k$, and $r_1, \dots, r_N \geq 0$.

D Case Study

In this section, we present a case study of the samples generated by different methods. Table 12 provides representative examples from our Gold dataset and synthetic data produced by (SimPrompt and Curated LLM), AttrPrompt, and our proposed Genetic Prompt method.

From the Table 12, it is evident that the samples generated by SimPrompt and Curated LLM tend to use explicit relational markers such as “cause”, “from”, “make” or “lead to” to directly state the relationships between entities. While AttrPrompt introduces some variety with interrogative and compound sentences, its overall style remains relatively uniform. In contrast, Genetic Prompt’s outputs do not rely on such overt cues. For example, in the sentence “With the increase in solar activity, satellites experience disruptions in their communications systems.” the causal relationship is implied through the context and structure rather than being explicitly signaled by words like “cause” or “lead to”. This implicit expression results in a highly heterogeneous mix of sentence structures, styles, and lengths.

Method	Example
Gold	<ol style="list-style-type: none"> 1. The burst has been caused by water hammer pressure. 2. He had chest pains and headaches from mold in the bedrooms. 3. The heavy rainstorm caused severe flooding in the downtown area, leading to widespread disruptions and evacuations.
SimPrompt And Curated LLM	<ol style="list-style-type: none"> 1. The rise in global temperatures is causing the melt of polar ice caps. 2. The intense heatwave led to crop failures in several rural communities, significantly impacting the livelihoods of local farmers. 3. Smoking leads to cancer in many cases.
AttrPrompt	<ol style="list-style-type: none"> 1. Increase the budget for the marketing team, as the new strategies will cause a significant impact on sales. 2. Is the landscape shaped by erosion? 3. The storm was responsible for causing damage, and as a result, the power outage occurred because of it.
Genetic Prompt	<ol style="list-style-type: none"> 1. With the increase in solar activity, satellites experience disruptions in their communications systems. 2. The political upheaval resulted in the region a devastating financial setback due to the scandal surrounding the mismanagement of public funds during the last fiscal cycle. 3. She observed changes in the plant growth following the application of the fertilizer after the rainfall event.

Table 12: Text Examples in *Cause-effect* class in Semeval dataset.