Guidance-Based Prompt Data Augmentation in Specialized Domains for Named Entity Recognition

Hyeonseok Kang[1], **Hyein Seo**[1], **Jeesu Jung**[1], **Sangkeun Jung**[1]*, **Du-Seong Chang**[2], **Riwoo Chung**[2]
[1]Computer Science and Engineering, Chungnam National University, Republic of Korea
[2]KT Corporation, Republic of Korea
{dnfldjaak11,hyenee97,jisu.jung5,hugmanskj}@gmail.com, {dschang,riwoo.chung}@kt.com,

Abstract

While the abundance of rich and vast datasets across numerous fields has facilitated the advancement of natural language processing, sectors in need of specialized data types continue to struggle with the challenge of finding quality data. Our study introduces a novel *guidance data augmentation* technique utilizing abstracted context and sentence structures to produce varied sentences while maintaining context-entity relationships, addressing data scarcity challenges. By fostering a closer relationship between context, sentence structure, and role of entities, our method enhances data augmentation's effectiveness. Consequently, by showcasing diversification in both entity-related vocabulary and overall sentence structure, and simultaneously improving the training performance of named entity recognition task.

1 Introduction

The field of Natural Language Processing (NLP) has witnessed remarkable success across various domains in recent years, primarily attributed to the availability of rich and high-quality data. However, specialized fields such as science and biology face significant challenges due to the scarcity of such quality data. Particularly, tasks like Named Entity Recognition (NER) face significant difficulties due to domain-specific characteristics where vocabulary roles diverge from general usage, necessitating specialized knowledge for effective data collection. To overcome the data shortage issue, various automated data augmentation (DA) techniques have been developed, including a recent approach that leverages Large Language Models (LLMs) for sentence generation to perform DA (Whitehouse et al., 2023b). Utilizing LLMs for DA involves employing few-shot learning or external modules (Zhuang et al., 2023) to provide additional information. In NER tasks, DA is applied with a focus on entities, maintaining the sentence's core structure with minimal alterations. This approach faces limitations in effectively augmenting cases like specific domain data, where vocabulary interpretation and roles vary with context and sentence composition.

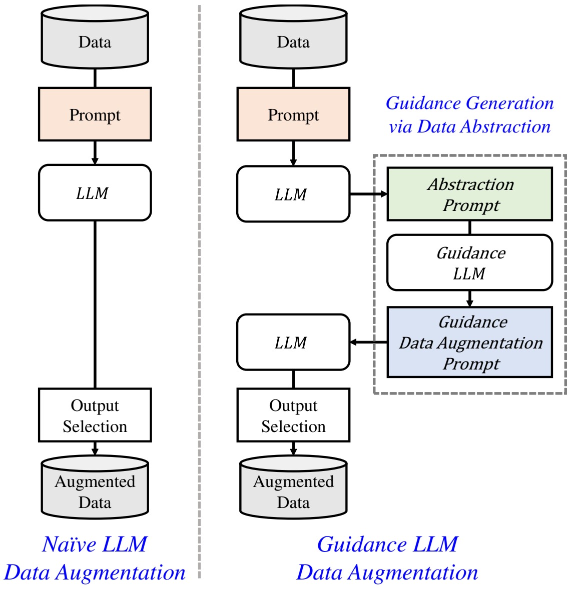

Data
Prompt
LLM
Output Selection
Augmented Data
Naïve LLM Data Augmentation
Data
Prompt
LLM
Guidance Generation via Data Abstraction
Abstraction Prompt
Guidance LLM
Guidance Data Augmentation Prompt
LLM
Output Selection
Augmented Data
Guidance LLM Data Augmentation

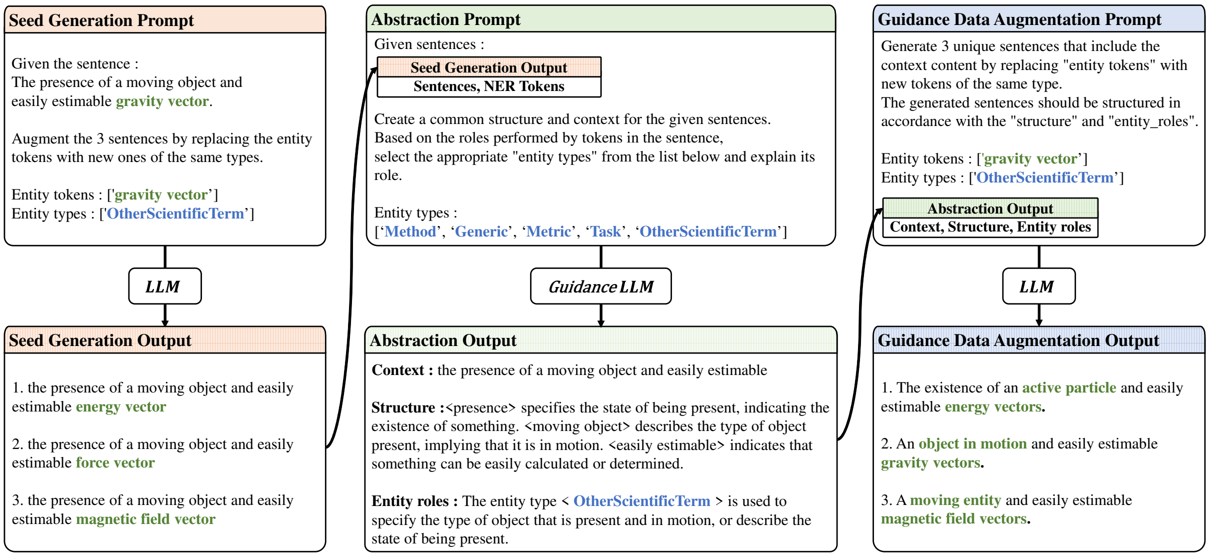
Figure 1: Comparison between data augmentation using LLM and Guidance LLM-based data augmentation.

In this study, we propose **G**uidance **D**ata **A**ugmentation (**GDA**), utilizing information on context and sentence structure abstracted through data abstraction for DA, aiming to generate sentences with varied structures alongside augmenting similar entity types. This approach seeks to achieve more natural and diverse DA compared to single LLM methods by augmenting data with sentences of varied structures that match the seed sentence's context and corresponding entity types.

Our data abstraction approach structures relationships among context, entities, and sentence composition for DA, expanding inference scope by using higher-level conceptual information(Zheng et al., 2024). This approach is vital where entity-related terms diverge significantly from general usage, re-

*Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 665–672
August 11-16, 2024 ©2024 Association for Computational Linguistics



Given the sentence :
The presence of a moving object and easily estimable gravity vector.

Augment the 3 sentences by replacing the entity tokens with new ones of the same types.

Entity tokens : ['gravity vector']
Entity types : ['OtherScientificTerm']

1. the presence of a moving object and easily estimable energy vector

2. the presence of a moving object and easily estimable force vector

3. the presence of a moving object and easily estimable magnetic field vector

Given sentences :

Seed Generation Output
Sentences, NER Tokens

Create a common structure and context for the given sentences.
Based on the roles performed by tokens in the sentence,
select the appropriate "entity types" from the list below and explain its role.

Entity types :
['Method', 'Generic', 'Metric', 'Task', 'OtherScientificTerm']

Context : the presence of a moving object and easily estimable

Structure :<presence> specifies the state of being present, indicating the existence of something. <moving object> describes the type of object present, implying that it is in motion. <easily estimable> indicates that something can be easily calculated or determined.

Entity roles : The entity type < OtherScientificTerm > is used to specify the type of object that is present and in motion, or describe the state of being present.

Generate 3 unique sentences that include the context content by replacing "entity tokens" with new tokens of the same type.
The generated sentences should be structured in accordance with the "structure" and "entity_roles".

Entity tokens : ['gravity vector']
Entity types : ['OtherScientificTerm']

Abstraction Output
Context, Structure, Entity roles

1. The existence of an active particle and easily estimable energy vectors.

2. An object in motion and easily estimable gravity vectors.

3. A moving entity and easily estimable magnetic field vectors.

Figure 2: Illustration of prompt flows for data augmentation in NER tasks. Structuring the use of data augmentation prompts with guidance prompts. Unlike traditional methods that input only the named entity tokens and their types, guidance prompts utilize context and entity role information generated from abstraction prompts to enrich data augmentation.

quiring a sophisticated strategy that analyzes vocabulary and structural details of entities. We address this challenge with our data abstraction approach, which assesses the contextual environment, the relationships between entities and their context, and the roles entities play, ensuring the development of varied and contextually appropriate terminology for specialized areas.

In response to these considerations, our study introduces a guidance prompt-based DA framework (see Figure 1). This framework is designed to generate sentences of various structures using the same type of entity, from data abstraction to the final response selection.

2 Related Works

In NLP, widely used DA techniques comprise rule-based approaches such as synonym replacement, back-translation, and random text element insertion or deletion. These methods are especially prevalent in tasks where textual data may require diversification to better train models (Bayer et al., 2022). Specifically, for tasks such as NER, augmentation strategies often revolve around the substitution of words with similar meanings or roles. In this context, techniques like Easy Data Augmentation (EDA) (Wei and Zou, 2019) and the utilization of WordNet (Miller, 1995) through the Natural Language Toolkit (NLTK) (Bird and Loper, 2004) are frequently applied to generate synonyms-based augmented data. In particular, when using data from specialized domain, DA methods are often used because it is difficult to collect data as it often consists of data containing domain-specific knowledge. For biomedical named entity recognition, augmentation is often performed using context to enhance the understanding of specialized concepts (Bartolini et al., 2023). Recently, the utilization of LLMs has expanded, leading to an increased use of DA techniques based on LLMs. These techniques involve augmenting data for sentence classification by leveraging LLMs (Dai et al., 2023), or enhancing cross-lingual tasks (Whitehouse et al., 2023a) through augmentation.

3 Guidance Data Augmentation

Our framework is designed around two key components for effective DA aimed at NER tasks: ***Data Abstraction*** and ***Data Augmentation via Guidance Prompts***. Through these two approaches, the proposed method facilitates enhanced model performance on NER tasks.

3.1 Guidance as Data Abstraction

Data abstraction involves abstracting and generalizing data to a form where the essential qualities are retained without the unnecessary specifics. The process allows for the alignment of roles and contextual attributes of named entities within sentences with the required entity types for the NER task,

	SciERC	NCBI-disease	FIN
Train	1,861(200)	5,432(200)	1,018(200)
Dev	275	923	150
Test	551	940	305

Table 1: Composition of the dataset for model evaluation. Values within parentheses in the train dataset column represent the number of seed data instances randomly selected from the training data for augmentation purposes.

thereby enabling the systematic identification and extraction of pivotal information.

Initially, for the purpose of data abstraction, a prompt is constructed to process data by substituting tokens corresponding to named entities with alternative tokens of the same type, as depicted in the seed generation prompt in Figure 2. The data generated from these prompts, along with seed data and a comprehensive list of entity types, are used to formulate abstraction prompts. Such abstracted prompts are fed into a guidance LLM to generate common contextual information for sentences and to produce the necessary structure and entity role information for context composition.

3.2 Data Augmentation via Guidance Prompt

Following data abstraction, the second component focuses on the augmentation process itself, using guidance prompts to generate new and varied instances of text. This method utilizes the abstracted data as a basis to inform the generation process, ensuring that the newly created text is both relevant and diverse. The guidance prompts are designed to direct the LLM in producing sentences that not only contain the targeted named entities but also mirror the semantic and structural diversity found in natural language usage.

The final output generated using guidance prompts is configured to create sentences that include the semantic information of the context without the need for seed data. The entities' roles and structure, derived from data abstraction alongside the context, serve as essential information for the generation and assignment of named entities. These elements are incorporated as conditions that must be adhered to within the guidance prompt, ensuring that the generated text aligns with specified contextual meanings and structural requirements.

Together, these two components offer an effective approach to enhancing the training data for NER tasks. By first abstracting the data to capture its essential elements and then augmenting it through carefully designed prompts, our framework aims to significantly diversify the datasets available for training NER models.

4 Experiments

4.1 Datasets

In selecting datasets for our experiments, we focused on data from specialized domains. These domains are characterized by the specificity (or expertise) required in their entities, requiring specialized knowledge for DA. Specifically, only three datasets were used: SciERC, NCBI-disease, and FIN. Detailed descriptions of the datasets are provided in Appendix A.

4.2 Models

Data augmentation LLMs Within the proposed framework, the following LLMs were used OpenAI GPT-3.5 and GPT-4 (Ouyang et al., 2022). Models versioned gpt-3.5-turbo-0125 and gpt-4-0613 were utilized, with the temperature parameter set to the default value of 1. By using the same LLM version for both abstraction and augmentation output generation in guidance data augmentation, this approach prevents the influence of language understanding differences among LLMs. Detailed information on the DA setting employing LLM is delineated in Appendix B.

Evaluation model for NER task For the evaluation phase, we used pre-trained language models, specifically BERT (Kenton and Toutanova, 2019). The model version employed is bert-base-uncased, and a comparative study was conducted to analyze the training effects of DA methods across three datasets. For training, the model was fed with a combination of 200 seed data instances and data augmented through DA as the training dataset. The F1 score was utilized as the metric for evaluating NER task performance. The implementation details utilized for training the evaluation model are furnished in Appendix C.

4.3 Results

Table 2 presents a comparison of NER model performance, contrasting models fine-tuned with datasets augmented using baseline methods such as EDA, WordNet, and Naïve DA with those augmented through the proposed GDA approach. When employing GPT-3.5 for data augmentation,

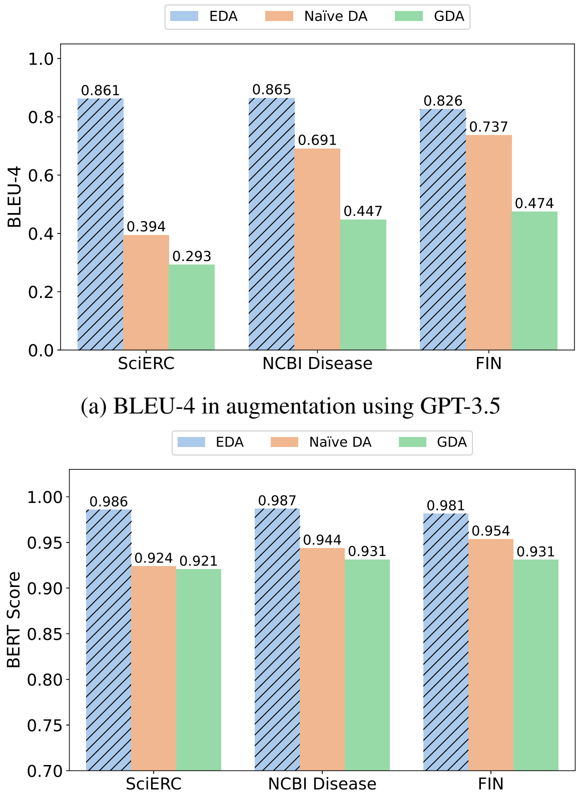
Approch	Method	Model	Datasets		
			SciERC	NCBI-disease	FIN
Rule-based DA	WordNet	-	0.5018	0.7924	0.7480
	EDA		0.5434	0.8062	0.7953
LLM-based DA	Naïve DA	GPT-3.5	0.5342(-0.0092)	0.8017(-0.0045)	0.8440(+0.0480)
	GDA(Ours)		**0.5435**(+0.0001)	**0.8139**(+0.0077)	0.8464(+0.0511)
	Naïve DA	GPT-4	0.5308(-0.0126)	0.7697(-0.0365)	0.8520(+0.0567)
	GDA(Ours)		0.5159(-0.0275)	0.7875(-0.0187)	**0.8544**(+0.0591)

Table 2: Evaluation of models trained with augmented data. Utilizing a base of 200 seed data points to generate an additional 600 data points for training, resulting in a total dataset of 800 entries. The table highlights the augmentation technique yielding the highest F1 score for each dataset in **bold**. Baseline methods employed were WordNet and EDA, with scores in parentheses indicating F1 score comparisons based on EDA. For LLM-based data augmentation, methods with superior performance per model are highlighted with a cyan background.

the proposed method generally outperformed the Naïve DA approach. Notably, the NCBI-disease dataset saw a 1.22% improvement in F1 score with the proposed method over Naïve DA, while the FIN dataset experienced a 0.24% increase. Augmentation with GPT-4 yielded a 1.78% and 0.24% performance boost for the NCBI-disease and FIN datasets, respectively.

In addition to improving model training performance, enhancing data augmentation requires the generation of diverse sentence structures and the assembly of vocabulary that corresponds with entity types. Figure 3a displays a graph comparing the structure of sentences generated through DA by method, utilizing the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) score to assess the degree of n-gram match between two sentences. Evaluation was conducted using BLEU-4, where a lower score signifies reduced structural similarity between sentences, indicating greater diversity in the generated sentences.

In addition to BLEU-4, we utilized BERTScore (Zhang et al., 2019) to evaluate the semantic similarity of generated sentences. BERTScore is particularly useful for capturing the nuanced semantic differences between sentences, as it utilizes pre-trained transformer models to provide a more context-aware assessment of similarity. This method is advantageous over traditional n-gram based metrics because it can better account for the semantic context rather than just surface-level text similarity. All three datasets displayed lower scores in both BERTScore and BLEU when using LLM-based augmentation compared to EDA, indicating enhanced diversity. In BLEU scores, the SciERC dataset, in particular, showed a significant 56.8% lower score with the proposed method compared to EDA. When contrasted with Naïve DA, the FIN and NCBI-disease datasets recorded 26.3% and 24.4% lower scores, respectively. For BERTScore, although the differences were smaller, ranging from 6.5% to 5%, the scores were still lower compared to EDA, indicating that the generated sentences



(a) BLEU-4 in augmentation using GPT-3.5

(b) BERTScore in augmentation using GPT-3.5

Figure 3: Comparison of sentence diversity by method in augmentation using GPT-3.5 model. Lower BLEU and BERTScore indicate higher diversity in generated sentences.

maintained semantic diversity. The narrow score differences in semantic similarity comparisons indicate that the generated sentences successfully preserve the contextual meaning of the seed sentences while introducing diversity. Experiments involving structural and semantic similarity using GPT-4, as well as assessments in generated sentences, are documented in Appendix D. Examples of DA outputs are provided in Appendix E. These results underscore the capability of the augmentation method utilizing abstract information to generate sentences with varied structures while preserving context and entity information.

5 Conclusion

In this study, we proposed guidance data augmentation designed for NER tasks within specific domain data, enabling the generation of data suited for these tasks. By using data abstraction, our method facilitates structured relationships among context, entities, and sentence composition, allowing for the generation of sentences with diverse structures while ensuring entity consistency. The abstracted sentence information is utilized in constructing guidance prompts, enabling DA with a rich diversity in vocabulary and sentence structures. Future efforts will aim at refining this process for applicability to additional tasks and exploring the use of multiple guidance LLMs to enrich the abstraction information, thereby enhancing the guidance provided.

Limitations

Our study's scope was notably confined to the NER task, limiting the versatility of our guidance prompts and data abstraction processes. This narrow focus restricts our exploration of the framework's potential across various tasks. The limitation of employing a singular model approach for both data abstraction and guidance DA restricts the diversity of linguistic insights within our system. Future efforts will aim to broaden the application of our framework by utilizing different LLMs and expanding the range and granularity of data abstraction, thus addressing these limitations and fully leveraging the capabilities of Multi-LLMs structures for enhanced language understanding and generation tasks.

Acknowledgements

This work was supported by research fund of Chungnam National University.

References

Ilaria Bartolini, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, and Andrea Vignali. 2023. Data augmentation via context similarity: An application to biomedical named entity recognition. *Information Systems*, 119:102291.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7).

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

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

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023a. Llm-powered data augmentation for enhanced cross-lingual performance.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023b. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

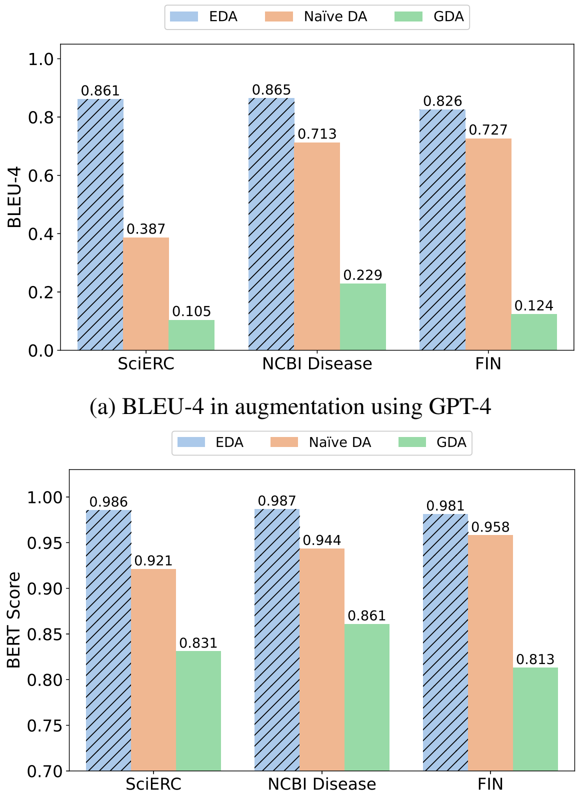
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Step-back prompting enables reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*.

A Datasets Details

SciERC (Luan et al., 2018) is a collection of scientific abstract annotated with scientific entities, their relations, and coreference clusters. NCBI-disease (Doğan et al., 2014) consists of PubMed abstracts fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community. FIN (Salinas Alvarado et al., 2015) is composed of financial agreements publicly disclosed through U.S. Securities and Exchange Commission (SEC) filings.



(a) BLEU-4 in augmentation using GPT-4

(b) BERTScore in augmentation using GPT-4

Figure 4: Comparison of sentence diversity by method in augmentation using GPT-4 model. Lower BLEU and BERTScore indicate higher diversity in generated sentences.

B Data Augmentation Details

Experiments were conducted by categorizing DA methods into rule-based DA and LLM-based DA. For rule-based DA, WordNet and EDA were employed, with the number of synonyms to randomly choose set at 10. LLM-based DA encompassed Naïve DA and GDA, both utilizing identical settings.

C Implementation Details

For the evaluation of DA, utilizing a learning rate of 2e-5, a batch size of 32, a maximum sequence length of 128, and the Adam optimizer as hyperparameters. The implementation framework utilized is based on Huggingface PyTorch Transformers (Wolf et al., 2020). In terms of computational infrastructure, the experimental procedures were

Method	Generated sentence	Named entity tokens
Seed data	The interlingual approach to MT has been repeatedly advocated by researchers originally interested in natural language understanding who take machine translation to be one possible application.	'interlingual approach', 'MT', 'natural language understanding', 'machine translation'
EDA	The interlingual approach to MT has been repeatedly advocated by researchers originally interested in natural language understanding who take machine translation to be one possible applications programme.	'interlingual approach', 'MT', 'natural language understanding', 'machine translation'
	The interlingual approach to MT has been repeatedly advocated by researchers originally interested in instinctive language understanding who take machine translation to be one possible application.	'interlingual approach', 'MT', 'natural language understanding', 'machine translation'
	The interlingual approach to MT has been repeatedly advocated by researchers originally interested in natural language understanding who take machine translation to be one potential application.	'interlingual approach', 'MT', 'natural language understanding', 'machine translation'
Naïve DA	The contrastive approach to AI has been repeatedly advocated by researchers originally interested in computer vision who take natural language processing to be one possible application.	'contrastive approach', 'AI', 'computer vision', 'natural language processing'
	The symbolic approach to ML has been repeatedly advocated by researchers originally interested in pattern recognition who take deep learning to be one possible application.	'symbolic approach', 'ML', 'pattern recognition', 'deep learning'
	The neural approach to NLP has been repeatedly advocated by researchers originally interested in information retrieval who take sentiment analysis to be one possible application.	'neural approach', 'NLP', 'information retrieval', 'sentiment analysis'
GDA(Ours)	The symbolic, contrastive, neural, statistical, and rule-based approaches to AI, ML, NLP, CV, and NLU have been repeatedly advocated by researchers originally interested in various tasks and applications.	'symbolic, contrastive, neural, statistical, and rule-based approaches', 'AI, ML, NLP, CV, NLU'
	The interlingual approach to MT has been repeatedly advocated by researchers originally interested in natural language understanding which is a possible application.	'interlingual approach', 'MT', 'natural language understanding'
	Researchers who were originally focused on natural language understanding have repeatedly advocated the interlingual approach to MT which is one of the potential applications.	'natural language understanding', 'interlingual approach', 'MT'

Table 3: Examples of data augmentation outcomes from EDA, Naïve DA, and GDA methodologies.

exclusively executed on Nvidia A6000 GPUs, complemented by AMD CPU cores.

D Additional Sentence Diversity Evaluation

Figure 4a exhibits a graph that contrasts the sentence structures generated via various DA methods, employing the BLEU-4 metric for comparison. Data augmentation with GDA using GPT-4 exhibits a marked improvement in data diversity as opposed to GPT-3.5. Figure 4b shows that GPT-4 demonstrates enhanced semantic diversity in data augmentation using GDA. Both implementations of GDA, using GPT-3.5 and GPT-4, showed improvements in semantic diversity within data augmentation compared to other augmentation methods.

E Guidance Data Augmentation Case

Table 3 presents an example of data created through EDA, Naïve DA, and GDA methods, with the last two utilizing the GPT-3.5 model. Augmentations via EDA and Naïve DA methods reveal replacements limited to either entity-specific words or other words. In contrast, sentences generated through GDA exhibit diversification in both entity-related vocabulary and overall sentence structure, while maintaining the context of the seed data despite structural modifications.