Data Augmentation using LLMs:
Data Perspectives, Learning Paradigms and Challenges

Bosheng Ding[*1] **Chengwei Qin**[*1] **Ruochen Zhao**[*1] **Tianze Luo**[1] **Xinze Li**[1]
Guizhen Chen[1] **Wenhan Xia**[2] **Junjie Hu**[3] **Anh Tuan Luu**[1] **Shafiq Joty**[1,4]

[1]Nanyang Technological University, Singapore [2]Princeton University
[3]University of Wisconsin–Madison [4]Salesforce AI

{bosheng001, chengwei003, ruochen002, tianze001, xinze002, guizhen001, anhtuan.luu}@ntu.edu.sg;
wxia@princeton.edu; junjie.hu@wisc.edu; sjoty@salesforce.com

Abstract

In the rapidly evolving field of large language models (LLMs), data augmentation (DA) has emerged as a pivotal technique for enhancing model performance by diversifying training examples without the need for additional data collection. This survey explores the transformative impact of LLMs on DA, particularly addressing the unique challenges and opportunities they present in the context of natural language processing (NLP) and beyond. From both data and learning perspectives, we examine various strategies that utilize LLMs for data augmentation, including a novel exploration of learning paradigms where LLM-generated data is used for diverse forms of further training. Additionally, this paper highlights the primary open challenges faced in this domain, ranging from controllable data augmentation to multimodal data augmentation. This survey highlights a paradigm shift introduced by LLMs in DA, and aims to serve as a comprehensive guide for researchers and practitioners.

1 Introduction

Data-centric approaches to Artificial Intelligence (AI) constitute a pivotal element in the advancement towards Artificial General Intelligence (AGI), centering on the construction of AI systems underpinned by high-quality data (Zha et al., 2023). This emphasis on data quality is vital, as it ensures the clarity of the information from which AI systems are to learn. Nevertheless, the acquisition of high-quality data presents significant challenges, being both costly and time-intensive, while the data annotation phase is often laborious and prone to inaccuracies resulting from human involvement (Ding et al., 2022b). In response to these challenges, researchers have dedicated efforts towards data augmentation (DA) techniques as a means to mitigate such issues (Chen et al., 2023c). Data augmentation fundamentally involves the adoption

of innovative methods aimed at bolstering model efficacy through the broadening of training data diversity, all without necessitating further data collection efforts. This strategy effectively tackles a major hindrance in machine learning research, namely the dearth of readily accessible training data, by applying slight modifications to existing datasets or creating synthetic data instances.

As we venture into the realm of large language models (LLMs), the significance of data concerns escalates. Research into the scaling laws pertinent to LLMs highlights the critical role of data as a renewable resource crucial for the enhancement and advancement of models (Kaplan et al., 2020). With the expansion of model training scales, there is a marked increase in data consumption. Prior studies have examined the escalation in dataset sizes within the domain of machine learning, forecasting that the reservoir of high-quality linguistic data may be depleted by 2026 (Villalobos et al., 2022). This anticipates a potential deceleration in the growth of machine learning models unless there is an enhancement in data efficiency or the discovery of novel data sources. Consequently, the utilization of synthetic data produced by AI models becomes essential once high-quality human-generated data resources are fully exploited.

From the data perspectives, data augmentation using LLMs offers a viable strategy to overcome these limitations, facilitating the creation of synthetic datasets of high quality that can, in certain instances, exceed the value of data curated by humans (Peng et al., 2023). This strategy not only addresses the challenge posed by the limited supply of human-annotated data but also conforms to scaling laws, enabling an increase in the size of training datasets without a proportional escalation in computational costs (FLOPs) (Hoffmann et al., 2022). The tactical employment of synthetic data stands to substantially reduce data collection costs and energy usage, signifying a transformative

---

[*]Equal contribution, order decided by coin flip.

phase in model training and inference and laying the groundwork for the achievement of artificial general intelligence (AGI) (Li et al., 2023a).

From a learning perspective, data augmentation using LLMs has heralded innovative learning paradigms, marking a significant departure from traditional methods predominantly centered around tasks like machine translation, sentiment analysis and NER. The application of LLMs in data augmentation extends to a broader spectrum of learning paradigms, including instruction tuning, in-context learning, and alignment learning. Additionally, it facilitates the generation of pseudo data for classification purposes and the scoring of data for regression analysis. This evolution in methodology not only broadens the scope of potential applications but also invites a deeper exploration into the nuanced benefits and insights that such innovative paradigms offer in the realm of data augmentation. As such, the systematic review of these emerging paradigms is crucial for advancing our understanding and leveraging the full potential of data augmentation in these learning contexts.

Given the growing interest and work in this domain, we believe it is a timely moment to present a paper on LLM-based DA. This paper aims to: (i) discuss data augmentation using LLM from the data perspective (ii) explore the learning paradigms that involve training LLMs on data generated by LLMs themselves, and (iii) highlight the principal challenges in this field to effectively guide and spur further interest and research. To the best of our knowledge, this represents the first survey that delves into data augmentation methods leveraging LLMs in such detail, thereby marking a significant contribution to the literature on LLM applications.

This paper is structured as follows. Section 2 provides a comprehensive review of related surveys, highlighting the distinctions between this work and previous surveys. Following that, as illustrated in Figure 1, Section 3 presents analysis from data perspectives for data augmentation with LLMs, establishing a foundational understanding for the community. Section 4 explores the learning paradigms associated with data augmentation using LLMs, neatly categorizing existing methods into two primary types: generative learning and discriminative learning. Section 5 ventures into the challenges and prospective future directions for research in this domain, highlighting the complexities and opportunities that lie ahead. To further

aid the community, the Appendix offers a detailed listing of existing methods categorized by tasks (Section A) and domains (Section B), serving as a valuable resource for researchers and practitioners alike. Through this survey, we hope to attract wider attention, generate increased interest, and encourage further research in the field of data augmentation using LLMs.

2 Related Surveys

Pre-LLM DA vs LLM DA Prior to the advent of Large Language Models (LLMs), surveys like those by Feng et al. (2021) and Hedderich et al. (2021) explore data augmentation in NLP, highlighting methodologies, applications, and the challenges of low-resource NLP. However, the surveyed approaches are constrained by the need for a deep linguistic understanding of traditional NLP tasks and do not fully utilize LLMs. The emergence of LLMs has revolutionized this landscape by significantly enhancing the quality of synthetic data, leading to innovative methodologies and broader applications. LLMs' nuanced understanding of language has mitigated previous limitations such as the issue of synthetic data being of poor quality, containing significant noise, and detrimentally affecting model performance. This advancement has made data augmentation more effective and accessible. Chen et al. (2023c) provide an empirical analysis of various augmentation techniques in supervised and semi-supervised contexts, focusing on pre-LLM methods. However, unlike previous surveys, we focus on discussing *data augmentation using LLMs from the data and learning perspectives*.

Instruction-tuning & Alignment Recent advancements in LLMs have heralded the advent of novel learning paradigms, including instruction-tuning and alignment learning, to better align the objectives of model training with the user's expectations for adherence to instructions. Instruction tuning (IT) emerges as a potent strategy to enhance the efficacy and manageability of LLMs, as discussed in the survey by Zhang et al. (2023f), which identifies and debates the hurdles encountered by IT, such as generating high-quality instruction sets, the method's effectiveness on tasks not initially supported, and critiques concerning its tendency for shallow pattern recognition. This underscores an imperative for ongoing research and refinement in the domain of instruction fine-tuning. Meanwhile,
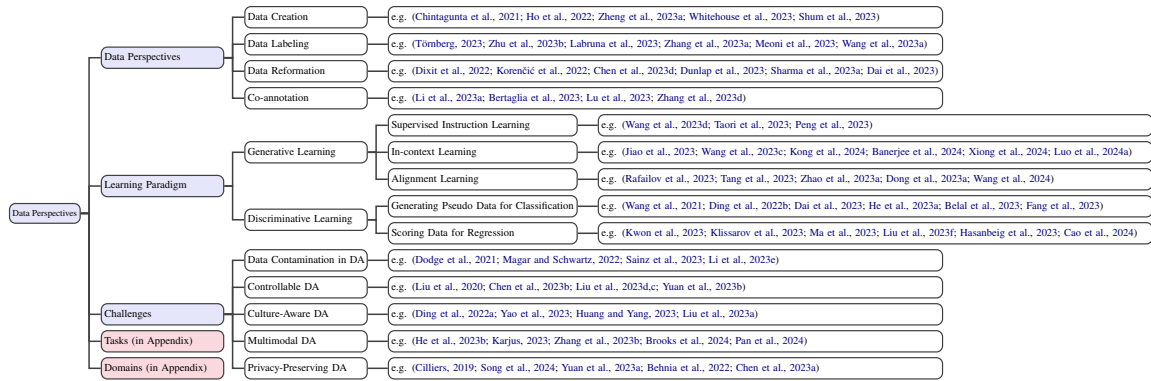
Figure 1: Taxonomy of Data Augmentation using LLMs. Tasks and Domains are included in the Appendix.

Data Perspectives

- Data Perspectives
  - Data Creation — e.g. (Chintagunta et al., 2021; Ho et al., 2022; Zheng et al., 2023a; Whitehouse et al., 2023; Shum et al., 2023)
  - Data Labeling — e.g. (Törnberg, 2023; Zhu et al., 2023b; Labruna et al., 2023; Zhang et al., 2023a; Meoni et al., 2023; Wang et al., 2023a)
  - Data Reformation — e.g. (Dixit et al., 2022; Korenčić et al., 2022; Chen et al., 2023d; Dunlap et al., 2023; Sharma et al., 2023a; Dai et al., 2023)
  - Co-annotation — e.g. (Li et al., 2023a; Bertaglia et al., 2023; Lu et al., 2023; Zhang et al., 2023d)

- Learning Paradigm
  - Generative Learning
    - Supervised Instruction Learning — e.g. (Wang et al., 2023d; Taori et al., 2023; Peng et al., 2023)
    - In-context Learning — e.g. (Jiao et al., 2023; Wang et al., 2023c; Kong et al., 2024; Banerjee et al., 2024; Xiong et al., 2024; Luo et al., 2024a)
    - Alignment Learning — e.g. (Rafailov et al., 2023; Tang et al., 2023; Zhao et al., 2023a; Dong et al., 2023a; Wang et al., 2024)
  - Discriminative Learning
    - Generating Pseudo Data for Classification — e.g. (Wang et al., 2021; Ding et al., 2022b; Dai et al., 2023a; He et al., 2023a; Belal et al., 2023; Fang et al., 2023)
    - Scoring Data for Regression — e.g. (Kwon et al., 2023; Klissarov et al., 2023; Ma et al., 2023; Liu et al., 2023f; Hasanbeig et al., 2023; Cao et al., 2024)

- Challenges
  - Data Contamination in DA — e.g. (Dodge et al., 2021; Magar and Schwartz, 2022; Sainz et al., 2023; Li et al., 2023e)
  - Controllable DA — e.g. (Liu et al., 2020; Chen et al., 2023b; Liu et al., 2023d,c; Yuan et al., 2023b)
  - Culture-Aware DA — e.g. (Ding et al., 2022a; Yao et al., 2023; Huang and Yang, 2023; Liu et al., 2023a)
  - Multimodal DA — e.g. (He et al., 2023b; Karjus, 2023; Zhang et al., 2023b; Brooks et al., 2024; Pan et al., 2024)
  - Privacy-Preserving DA — e.g. (Cilliers, 2019; Song et al., 2024; Yuan et al., 2023a; Behnia et al., 2022; Chen et al., 2023a)

- Tasks (in Appendix)
- Domains (in Appendix)

Ji et al. (2023) delineate the foundational principles, methodologies, and practical applications of AI alignment, exploring its prospective trajectories. The focal point of AI alignment research is to ensure AI systems' actions are congruent with human objectives and ethical standards, tackling misalignment through enhancements in system robustness, interpretability, controllability, and ethical considerations, despite obstacles like reward exploitation and goal misinterpretation. Our survey concentrates on this emerging learning paradigm, particularly the use of synthetic data to train LLMs and augment their performance capabilities.

3 Data Perspectives

From a data perspective, we group existing studies on LLM-based DA into four categories: 1. *Data creation* which leverages the few-shot learning ability of LLMs to create a large synthetic dataset; 2. *Data labeling* which uses the LLM to label existing datasets; 3. *Data reformation* which transforms existing data to produce new data; 4. *Co-Annotate* which enables LLM-human collaboration to gather high-quality augmentation data. This section discusses relevant papers in each category.

3.1 Data Creation

Data Creation focuses on leveraging the few-shot learning ability of LLMs to quickly create a large amount of synthetic data. It is most used in tasks with a large label space.

Data Creation with LLMs is a promising solution in specialized or private domains, where annotations are usually difficult or expensive to collect. Dialogue tasks are one example where specialized data is hard to collect. In medical dialogue summarization, Chintagunta et al. (2021) uses a powerful few-shot learner such as GPT-3 to create

synthetic medical dialogue summaries. By training models on a mix of synthesized and human-labeled data, the algorithm can scale a few human-labeled examples to yield results comparable to using 30x human-labeled examples. Similarly, for general dialogue, Dialogic (Li et al., 2022a) is seeded with a few dialogues and can automatically select in-context examples for demonstration and prompt LLMs to generate annotated dialogues in a controllable way. Then, automatic verification and revision methods are proposed to mitigate annotation errors. Results show that performance greatly improves in low-resource scenarios. Wan et al. (2022a) also attempt few-shot data augmentation on dialogue modeling. Aside from few-shot learning, for emotional support conversations, AugESC (Zheng et al., 2023a) finetunes an LM and prompts it to complete dialogues from collected posts. The post-training on AugESC improves downstream dialogue models' generalization abilities to open-domain topics. For low-resource classification, LLM can be used to create synthetic examples of a given label. Møller et al. (2023) gives an example and its corresponding label and instruct the LLMs to generate similar examples exhibiting the same label. Resulting models yield better downstream performances on few-shot classification but still lag behind human-annotated data. For other low-resource tasks such as recommendation and intent detection, Data Creation can also effectively boost the training data space. To gather better recommendations, Zhang et al. (2023c) generates a large amount of user-personalized instruction data with varying preference and intention types. Then, the LLM is optimized using instruction tuning. The resulting model can obtain more accurate recommendations and outperform competitive baselines, including GPT-3.5. For intent detection, Lin et al.

(2023) first uses an LLM to generate synthetic examples in the context of the training set and then uses Pointwise V-Information (PVI) to filter unhelpful examples. Sahu et al. (2022a) prompts GPT-3 to generate labeled training data, which can significantly boost the intent classifier's performance for distinct intents but becomes less helpful with semantically close intents.

Data Creation also helps in more general tasks by generating new training datasets. For information retrieval, Bonifacio et al. (2022) uses few-shot prompting with LLMs to generate synthetic training datasets consisting of query-document pairs. Retrieval models finetuned with the augmented data significantly outperform unsupervised models. For reasoning, Logi-CoT (Liu et al., 2023b) gathers a new instruction-tuning dataset by prompting GPT-4 and is used for teaching models to elicit general reasoning skills.

Moreover, Data Creation is helpful for model performance distillation. To distill LLMs' reasoning performances to smaller models, Fine-tune-CoT (Ho et al., 2022) uses zero-shot CoT prompting to generate rationales from teacher models and use them to fine-tune smaller student models. The resulting performance improvements are stable across dataset size, teacher performance, etc. To reduce the need of manual annotation in reasoning tasks, Automate-CoT (Shum et al., 2023) automatically generates pseudo-CoTs from a small labeled dataset and then prunes and selects an optimal combination for CoT prompting. Similarly for instruction-following, Peng et al. (2023) uses GPT-4 to generate an instruction-following dataset and feedback data. The resulting instruction-tuned LLaMa models can lead to comparable performance with the original GPT-4. To aid multilingual commonsense reasoning tasks, Whitehouse et al. (2023) provides LLMs with instructions and examples from the original training data, prompting them to generate new and diverse examples. By training with augmented data, significant cross-lingual performance improvements are observed on smaller models.

To systematically study the behavior of such data creation methods and improve upon current few-shot prompting methods, Meng et al. (2023) attempts to first tune an LM on few-shot examples and then use it as a generator to synthesize a large amount of novel training samples. The resulting approach could augment task performances than existing few-shot learning methods.

3.2 Data Labeling

Data Labeling seeks to utilize the general language comprehension abilities of LLMs to annotate unlabeled datasets. It is primarily useful in tasks that have a large enough unlabeled data corpus, such as cross-lingual and multimodal tasks.

To evaluate LLMs' potential in data labeling, Törnberg (2023) studies the zero-shot annotation ability of GPT-4 on labeling political twitter messages with political tendencies. Compared to human workers, the LLM annotations display higher accuracy and lower bias. Similarly, Zhu et al. (2023b) observes that, in social computing tasks, ChatGPT has the potential to accurately reproduce human labels. Notably, annotations from open-source LLMs (Alizadeh et al., 2023) and ChatGPT (Gilardi et al., 2023) can surpass crowdworker performance on annotation tasks. For annotating low-resource tasks such as goal-oriented dialogues (Labruna et al., 2023) and speech emotional data (Latif et al., 2023), the quality of ChatGPT annotations is on par with human-generated labels. However, Bansal and Sharma (2023) observes that simply annotating can sometimes worsen generalization. Thus, it proposes conditional sampling to optimize the tradeoff between informativeness and budget

Cross-lingual tasks mostly contain a large unlabeled corpus, which could benefit from data labeling. Therefore, Zhang et al. (2023a) uses different prompting strategies to augment machine translation (MT) data. It tries augmenting monolingual data using back-/forward-translation via zero-shot prompting, which still suffers from limitations such as generalization and unstable transfer performances. Similarly, Meoni et al. (2023) annotates training data for multilingual clinical entity extraction with LLMs. After fine-tuning smaller models with augmentations, they display promising results for information extraction (IE) tasks.

Data labeling is also promising for multimodal applications. For data-scarce Visual Question Answering (VQA) tasks, Khan et al. (2023) utilizes a Self-taught Data Augmentation (SelTDA) framework to generate pseudo labels from unlabeled images. The pseudo-labeled data could then improve VQA task performance and robustness. Combining multi-modality with reasoning, T-SciQ (Wang et al., 2023a) further distills LLMs' reasoning abilities

in multimodal tasks by asking the teacher model to produce CoT rationales. As a result, it achieves state-of-the-art performance in scientific QA.

3.3 Data Reformation

Data Reformation techniques attempt to reformulate the existing data into more variations for more fine-grained augmentation.

Such reformation techniques could naturally aid in counterfactual generation tasks, which reforms existing data to its counterfactual version. Disco (Chen et al., 2023d) uses LLMs to generate high-quality counterfactual data at scale. It first uses in-context learning with GPT-3 to generate phrasal perturbations, then uses a task-specific teacher model to filter and distill high-quality counterfactual data. Models trained using the generated counterfactuals display improved robustness and generalization across distributions. For retrieval-augmented generation, CORE (Dixit et al., 2022) uses GPT-3 to generate counterfactual edits to the input conditioned on the retrieved excerpts. The perturbations then help mitigate model bias and improve performance on out-of-distribution (OOD) data.

Data Reformation could also quickly diversify the original dataset by forming data pairs. For conspiracy detection, Korenčić et al. (2022) asks GPT-3 to rephrase tweets with original labels to augment training. To generate useful variations of the pretraining datasets of large vision models, ALIA (Dunlap et al., 2023) uses LLMs to generate image descriptions and augment the training data via language-guided image editing. By leveraging LLMs to the image domain, ALIA surpasses traditional data augmentation methods on fine-grained classification tasks. For Named Entity Recognition (NER), Sharma et al. (2023a) generates paraphrases while retaining inline annotation for entities. Among other PLMs, GPT-3 is able to generate high-quality paraphrases, yielding statistically significant improvements in NER performance.

For more general tasks, Data Reformation could help to diversify and broaden the original dataset. AugGPT (Dai et al., 2023) tries to overcome the challenge of few-shot and data-scarce NLP tasks by rephrasing each sentence in the training samples into 6 semantically similar sentences. Experiments show that such an approach surpasses state-of-the-art text data augmentation methods in augmentation distribution and testing accuracy. For effective

knowledge distilation, GPT3Mix (Yoo et al., 2021) extracts sample sentences from the task-specific training data, embed these samples in the prompt, and asks the LLM to generate an augmented mixed sentence influenced by the sample sentences. Guo et al. (2023) asks GPT-3.5 and GPT-4 to rewrite or generate question-answer pairs with zero-shot prompting. Fine-tuning with the refined and diversified training set then successfully distils medical question-answering abilities to smaller models.

3.4 Co-annotation

Co-annotation refers to the collaborative annotation process between humans and LLMs. By combining both annotation approaches, Co-annotation can reduce annotation costs and improve annotation performance at the same time. Firstly, Li et al. (2023a) proposes CoAnnotating, which allocates a given datapoint to be annotated by humans or by LLMs by computing the uncertainty level of LLM's annotations. With efficient human-AI collaboration, it provides insights into the tradeoff between annotation quality and annotation cost. To assist human annotators with explanations, Bertaglia et al. (2023) asks the LLM to identify relevant features, such as text tokens, as assistive explanation. The approach improves inter-annotator agreement, annotation accuracy, and annotators' confidence, eventually leading to more transparency. Using human feedback to direct LLM annotations could also effectively generate high-quality data. Diagen (Lu et al., 2023) uses an LLM to iteratively generate dialogues in protected data domains, where human feedback is used to correct inconsistencies or redirect the flow in sub-dialogues. As a result, fine-tuning or in-context learning with the annotated data shows significant model performance improvements. Similarly, ToolCoder (Zhang et al., 2023d) uses human-written input-output pairs as prompts to guide chatGPT to annotate a tool-augmentation dataset. Then, the annotated data is filtered to ensure quality. After fine-tuning with the annotated data, ToolCoder can achieve comparable performance with ChatGPT on code generation.

4 Learning Paradigms

Teacher-Student Learning (TSL) Leveraging a Language Model (LLM) as a data annotator to create synthetic data for the purpose of training other models represents a pivotal shift towards more efficient and scalable machine learning methodologies.
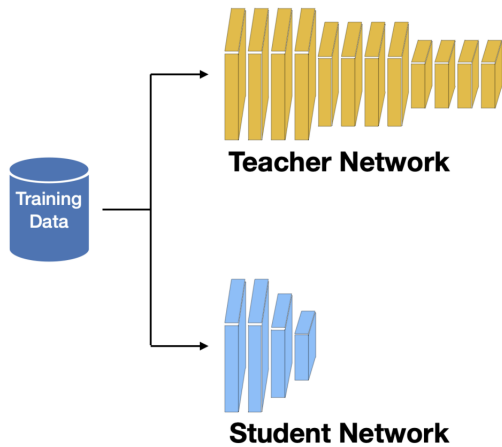
Figure 2: Illustrations of Teacher-Student Learning.

This approach is situated within the broader framework of *Teacher-Student Learning (TSL)* (Hu et al., 2022), where the LLM acts as the 'teacher' by generating annotated data that serves as training material for 'student' models as shown in Figure 2. By utilizing an LLM for data annotation, organizations can bypass some of the limitations associated with human annotators, such as scalability issues and subjectivity, thereby streamlining the development process of AI systems. This innovative strategy not only enhances the efficiency of generating training datasets but also contributes to the advancement of machine learning techniques, enabling the creation of more robust and accurate models. This emerging trend is gaining traction, particularly in the context of training large language models.

Teacher-Student Learning paradigms for data augmentation using LLMs can be broadly segmented into generative and discriminative learning approaches. Generative learning exploits LLMs to create instructional datasets and demonstration examples, enriching model training. It encompasses Supervised Instruction Learning, generating instructional data (instruction, x, y); In-context Learning, producing few-shot examples to augment query inputs and facilitate output generation; and Alignment Learning, creating human-preference data for reinforcement learning and optimization. Discriminative learning, conversely, focuses on refining task-specific models through pseudo-labeling and pseudo-scoring. Pseudo-labeling generates labels $(x, \hat{y})$ for smaller models, akin to knowledge distillation, while pseudo-scoring assigns numerical scores to (x, y) pairs for regression tasks.

4.1 Generative Learning

Supervised Instruction Learning : Supervised Instruction Learning, leveraging LLMs for data augmentation, presents a novel approach to enhancing instructional data generation. The generation process entails creating instructional data in the form of (instruction, x, y) triplets, where 'x' represents optional context or input, and 'y' denotes the output generated by LLMs in response to the given instruction. Wang et al. (2023d) first introduced "Self-Instruct," a method to enhance the instruction-following ability of pre-existing language models by leveraging their own outputs. Taori et al. (2023) created the Alpaca dataset at low cost using Supervised Instruction Learning and Peng et al. (2023) further explored the potential of leveraging LLMs to create cross-lingual instruction following data, self-evaluation and cross-model assessment data and responses to unnatural instructions.

In-context Learning : In-context learning represents a new paradigm where models, particularly LLMs like GPT-3, generate predictions based on examples provided within the context of the input, without requiring explicit retraining or parameter updates (Liu et al., 2022; Dong et al., 2023b; Zhang et al., 2024). To enable LLMs with such a capability, data with rich context should be provided to LLMs during the training process. Kong et al. (2024) use existing LLMs to generate two multi-turn dialogue datasets based on detailed annotations, enabling model's in-context dialogue capabilities. Banerjee et al. (2024) utilize KGs to build up in-context training samples for LLM instruction tuning. Xiong et al. (2024) construct a synthetic temporal QA dataset to enhance the LLM's capability in temporal QA tasks. Luo et al. (2024a) design structure-based augmentation methods with LLMs for Temporal Knowledge Graph Completion. It is a common trend that various models adopt LLMs to generate synthetic data with long context to enhance the model's in-context dialogue, inference, QA, and Natural Language Generation (NLG) capabilities (Touvron et al., 2023; Jiao et al., 2023; Wang et al., 2023c).

Alignment Learning : Alignment learning focuses on training the model to align its outputs with human preferences or objectives, often through techniques such as reward shaping or reinforcement learning, to optimize its performance towards desired outcomes while minimizing unintended be-

haviors. Rafailov et al. (2023) propose the Direct Preference Optimization (DPO) method to effectively train LLMs on alignment learning. Wang et al. (2024) adopt DPO and synthesized instruction following data to build a foundation model for creative writing. Tang et al. (2023) utilize LLMs such as ChatGPT, GPT-4, Dolly-v2, and StableVicuna to generate various data for alignment training of their model LLMDA. Zhao et al. (2023a) introduce Group Preference Optimization (GPO) to the alignment training for LLMs. Dong et al. (2023a) adopt data augmentations from LLMs and construct contrastive feedback to training models.

4.2 Discriminative Learning

Generating Pseudo Data for Classification
LLMs play an important role in generating pseudo data to train discriminative models for classification tasks. Many LLM-based data augmentation approaches fall into this category. Wang et al. (2021); Ding et al. (2022b) explore using GPT-3 to annotate or generate data to train smaller task-specific models on a variety of classification tasks. Dai et al. (2023) propose AugGPT which uses ChatGPT to rephrase training samples, surpassing state-of-the-art text data augmentation methods on few-shot classification tasks. To enhance the capabilities of student models in zero-shot learning, He et al. (2023a) develop TeacherLM7.1B which demonstrates the strong ability to generate augmentations for training student models. Besides, some works are designed for specific target tasks, *e.g.,* sentiment analysis (Belal et al., 2023; Møller et al., 2023) and intent classification (Sahu et al., 2022b; Fang et al., 2023).

Scoring Data for Regression LLMs can serve as an essential solution for scoring the input. Kwon et al. (2023) explore prompting GPT-3 as a proxy reward function in a reinforcement learning framework. The reward signal is then used to update the behavior of the agent. Motif (Klissarov et al., 2023) proposes to elicit preferences from an LLM over paired data to construct rewards for training agents. Ma et al. (2023) introduce Eureka which can leverage GPT-4 for evolutionary optimization to generate reward functions that outperform expert-designed rewards. To mitigate reward sparsity that can lead to inefficient and unstable learning, Cao et al. (2024) incorporate a critic LLM to provide more fine-grained dense rewards. In addition to generating reward signals, LLMs are

also used to evaluate different properties of text. Liu et al. (2023f) investigate utilizing LLMs as the reference to evaluate the quality of the generated summary. Liu et al. (2023e) further benchmark LLM-based evaluation for instruction controllable text summarization with different evaluation protocols and LLMs. To improve LLMs' text evaluation capabilities, ALLURE (Hasanbeig et al., 2023) explores leveraging iterative in-context learning to audit and enhance LLM-based evaluation.

5 Challenges and Future Directions in Data Augmentation using LLMs

5.1 Data Contamination in Data Augmentation

Challenges In leveraging LLMs for data augmentation, the risk of data contamination presents a significant challenge (Magar and Schwartz, 2022). This issue, where training data inadvertently includes evaluation set examples, undermines model evaluation integrity by enabling models to memorize rather than generalize. The two main types of contamination are input contamination and the more severe input-and-label contamination, which allows models to memorize exact input-output pairs. Studies such as (Dodge et al., 2021) and others (Sainz et al., 2023; Li et al., 2023e) highlight the prevalence of data contamination across various NLP benchmarks, emphasizing the need for effective detection and mitigation strategies. The integration of LLMs for data augmentation, while offering the potential to enrich training datasets and enhance model performance, necessitates cautious consideration of the risk of data contamination. This risk not only undermines the integrity of model evaluations but also highlights the ongoing need for novel strategies to detect and counteract data contamination.

Future Directions As the NLP field continues to advance, addressing the challenges posed by data contamination, particularly in the context of LLM-augmented datasets, will be crucial for maintaining the credibility and effectiveness of machine learning models. To this end, there is a significant space for research on the development of innovative methodologies and tools specifically designed to detect and address data contamination in synthetic datasets.

5.2 Controllable Data Augmentation

Challenges Controllable data augmentation using LLMs faces significant challenges, primarily due to difficulties in ensuring the quality of generated synthetic data (Liu et al., 2020; Chen et al., 2023b). These techniques, aimed at enriching few-shot or low-resource datasets by targeting specific attributes, rely on models understanding and applying attribute-value mappings. Current methods, including in-context learning and fine-tuning, often struggle with maintaining quality across non-target dimensions during augmentation (Liu et al., 2023d). Furthermore, model collapse is a critical challenge that arises when models learn from data generated by other models, leading to a degeneration where models progressively lose information about the true underlying data distribution (Shumailov et al., 2023). Over time, this process results in models that forget the true underlying data distribution and converge to a narrow interpretation of the data with very little variance, potentially compromising the diversity and richness of the model's output.

Future Directions To address these challenges, we need to explore new methods that can decompose and reconstruct texts with LLMs to precisely control attribute changes. To mitigate the issues of model collapse, the generated data should follow the data distribution of human-generated data to preserve integrity and reliability. Moreover, the generated synthetic data must be approached with a keen awareness of the potential for bias and the ethical implications associated with LLM outputs. By adopting responsible augmentation strategies that promote diversity and implement ethical guidelines, such as integrating the in-context vector (Liu et al., 2023c) into the data augmentation framework, we can mitigate biases and advance NLP in a manner that is technologically effective and ethically responsible (Liu et al., 2023c; Yuan et al., 2023b).

5.3 Culture-Aware Multilingual Data Augmentation

Challenges In the rapidly evolving domain of multilingual NLP systems, the integration of culture-aware data augmentation emerges as a pivotal challenge (Yao et al., 2023; Huang and Yang, 2023). This approach is critical for the effective localization of these systems, enabling them to operate seamlessly across varied real-world contexts (Ding et al., 2022a; Liu et al., 2023a). Traditional data augmentation methods, which predominantly focus on linguistic variations, often fail to capture the complex cultural nuances that significantly influence language usage. This oversight highlights the need for a fundamental shift towards embedding cultural intelligence into data augmentation strategies. Utilizing LLMs for this purpose offers a promising pathway.

Future Directions LLMs have the potential to generate data that reflects cultural specifics, encompassing regional idioms, social norms, and linguistic nuances. Such culturally enriched data augmentation not only aims to improve the precision of multilingual NLP systems but also promotes inclusivity and global relevance. Nevertheless, achieving this necessitates further research in developing LLMs capable of discerning and adjusting to cultural differences. This advancement is crucial for realizing NLP solutions that are not only linguistically diverse but also culturally attuned, thereby enhancing their acceptability and usability on a global scale. The journey towards incorporating cultural awareness in data augmentation with LLMs presents both a formidable challenge and a significant opportunity for the future of multilingual NLP systems.

5.4 Multimodal Data Augmentation

Challenges In the burgeoning field of multimodal data augmentation using LLMs, researchers face a constellation of challenges despite the promising prospects (He et al., 2023b; Karjus, 2023). As LLMs evolve to incorporate a diverse array of data types beyond text—spanning images, audio, video, and potentially graphs—the complexity of integrating and harmonizing these varied modalities poses significant technical hurdles (Zhang et al., 2023b; Brooks et al., 2024). The seamless handling and enhancement of the intricate interplay between these modalities require sophisticated algorithms that can not only generate and manipulate multimodal data but also understand and preserve the contextual and semantic integrity across them. Additionally, the extension to graph-based modalities introduces unique challenges in representing and leveraging relational and structural information effectively (Pan et al., 2024).

Future Directions The challenges mentioned above demand innovative approaches in model architecture, data representation, and algorithmic efficiency to fully realize the potential of LLMs in

multimodal augmentation. Moreover, ensuring the adaptability and robustness of these models in real-world applications across domains such as social network analysis and recommendation systems amplifies the complexity of the task. As we push the boundaries of what LLMs can achieve in multimodal learning, addressing these challenges will be crucial for unlocking new horizons in research and innovation, driving forward the capabilities of machine learning models to comprehend and process the rich tapestry of multimodal information.

5.5 Privacy issues of Data Augmentation

Challenges Data augmentation using LLMs presents significant privacy concerns, especially in sensitive sectors like healthcare (Cilliers, 2019), where confidentiality is paramount. While LLMs offer the potential to enrich sparse datasets with synthetic data, safeguarding sensitive information within these datasets poses considerable challenges. Research by Song et al. (2024) underscores the limitations of Differential privacy (DP) in effectively generating synthetic data without compromising privacy or data quality. Concurrently, Yuan et al. (2023a) explore LLMs' utility in healthcare for patient-trial matching, proposing a privacy-aware approach that mitigates privacy risks by using desensitized data. These studies highlight the ongoing struggle to leverage LLMs for data augmentation while ensuring robust privacy protection, indicating a crucial area for further research and development in privacy-preserving techniques.

Future Directions Future directions to mitigate these privacy issues may involve developing more robust anonymization techniques to ensure that no traceable data is reproduced in the outputs. Additionally, implementing differential privacy (Behnia et al., 2022) and federated learning (Chen et al., 2023a) could help minimize privacy risks by allowing data to be processed in a way that doesn't reveal identifiable information. Transparent data usage policies and regular audits could also play a crucial role in maintaining user trust and ensuring compliance with evolving data protection regulations. These measures would be essential in balancing the benefits of LLMs in data augmentation with the need to protect individual privacy.

6 Conclusion

In this survey, we have delivered an in-depth and organized review of data augmentation techniques utilizing Large Language Models (LLMs). We explored key data augmentation strategies and the application of LLMs in enhancing data augmentation processes. Moreover, we discussed the learning paradigm under the generative and discriminative learning perspectives. Additionally, we highlighted the existing challenges and potential future research opportunities, indicating significant potential for further developments in this field. Our aim is for this paper to act as a reference for AI researchers, aiding them in selecting appropriate data augmentation methods and encouraging further investigation and enthusiasm in this domain.

7 Limitations

Data augmentation using Large Language Models (LLMs) presents several limitations that can impact their effectiveness in certain contexts. Firstly, the quality of the generated data heavily depends on the training data and the model's architecture, which can lead to biases or inaccuracies being introduced into the augmented dataset. This can perpetuate or even exacerbate existing biases in the training data, leading to skewed or unfair outcomes in applications. Secondly, LLMs may struggle with generating high-quality data for highly specialized or niche domains where specific knowledge or terminology is required, due to the limited exposure of such content in their training data. Additionally, the cost of generating large amounts of augmented data can be prohibitive, as it requires substantial computational resources, especially for state-of-the-art models. Lastly, there's the challenge of ensuring the semantic consistency and uniqueness of the generated data, as LLMs can sometimes produce repetitive or generic outputs that may not add meaningful diversity to the dataset. These limitations necessitate careful consideration and mitigation strategies to effectively leverage LLMs for data augmentation purposes.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:1–15.

Narjes Askarian, Ehsan Abbasnejad, Ingrid Zukerman, Wray Buntine, and Gholamreza Haffari. 2022. Inductive biases for low data vqa: a data augmentation approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 231–240.

Somnath Banerjee, Amruit Sahoo, Sayan Layek, Avik Dutta, Rima Hazra, and Animesh Mukherjee. 2024. Context matters: Pushing the boundaries of open-ended answer generation with graph-structured knowledge context. *arXiv preprint arXiv:2401.12671*.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost.

Qiming Bao, Gael Gendron, Alex Yuxuan Peng, Wanjun Zhong, Neset Tan, Yang Chen, Michael Witbrock, and Jiamou Liu. 2023. A systematic evaluation of large language models on out-of-distribution logical reasoning tasks. *arXiv preprint arXiv:2310.09430*.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. 2022. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE.

Mohammad Belal, James She, and Simon Wong. 2023. Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2023. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media.

Desirée Bill and Theodor Eriksson. 2023. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models.

Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. Drlc: Reinforcement learning with dense rewards from llm critic. *arXiv preprint arXiv:2401.07382*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Taha Ceritli, Ghadeer O Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P Creagh, and David A Clifton. 2023. Synthesizing mixed-type electronic health records using diffusion models. *arXiv preprint arXiv:2302.14679*.

Chaochao Chen, Xiaohua Feng, Jun Zhou, Jianwei Yin, and Xiaolin Zheng. 2023a. Federated large language model: A position paper. *arXiv preprint arXiv:2307.08925*.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. 2023b. Mixture of soft prompts for controllable data generation. In *Conference on Empirical Methods in Natural Language Processing*.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023c. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023d. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023e. Disco: distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.

Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498*.

Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2023. Data-centric financial large language models.

Liezel Cilliers. 2019. Wearable devices in healthcare: Privacy and information security issues. *Health Information Management Journal*, 49:150 – 156.

Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Aminu Da'u and Naomie Salim. 2020. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53(4):2709–2748.

Hexuan Deng, Xin Zhang, Meishan Zhang, Xuebo Liu, and Min Zhang. 2023. Holistic exploration on universal decompositional semantic parsing: Architecture, data augmentation, and llm paradigm. *arXiv preprint arXiv:2307.13424*.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022a. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022b. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Qian Dong, Yiding Liu, Qingyao Ai, Zhijing Wu, Haitao Li, Yiqun Liu, Shuaiqiang Wang, Dawei Yin, and Shaoping Ma. 2023a. Aligning the capabilities of large language models with the context of information retrieval via contrastive feedback. *arXiv preprint arXiv:2309.17078*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023b. A survey for in-context learning. *ArXiv*, abs/2301.00234.

Zeming Dong, Qiang Hu, Yuejun Guo, Zhenya Zhang, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Jianjun Zhao. 2023c. Boosting source code learning with data augmentation: An empirical study. *arXiv preprint arXiv:2303.06808*.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. 2023. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. 2024. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in Neural Information Processing Systems*, 36.

Moussa Kamal Eddine, Guokan Shang, and Michalis Vazirgiannis. 2022. Datscore: Evaluating translation with data augmented translations. *arXiv preprint arXiv:2210.06576*.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

Yihao Fang, Xianzhi Li, Stephen W Thomas, and Xiaodan Zhu. 2023. Chatgpt as data augmentation for compositional generalization: A case study in open intent detection. *arXiv preprint arXiv:2308.13517*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for nlp. *ArXiv*, abs/2105.03075.

Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1):1–51.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. Dale: Generative data augmentation for low-resource legal nlp. *arXiv preprint arXiv:2310.15799*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large language models respond to influence like humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 15–24, Toronto, Canada. Association for Computational Linguistics.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. 2024. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, 7(1):1–14.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. 2023. Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. *arXiv e-prints*, pages arXiv–2305.

Qi Hao, Tianze Luo, and Guangda Huzhang. 2021. Reranking with constraints on diversified exposures for homepage recommender system. *arXiv preprint arXiv:2112.07621*.

Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. 2023. Allure: A systematic protocol for auditing and improving llm-based evaluation of text using iterative in-context-learning. *arXiv preprint arXiv:2309.13701*.

Nan He, Hanyu Lai, Chenyang Zhao, Zirui Cheng, Junting Pan, Ruoyu Qin, Ruofan Lu, Rui Lu, Yunchen Zhang, Gangming Zhao, et al. 2023a. Teacherlm: Teaching to fish rather than giving the fish, language modeling likewise. *arXiv preprint arXiv:2310.19019*.

Wentao He, Hanjie Ma, Shaohua Li, Hui Dong, Haixiang Zhang, and Jie Feng. 2023b. Using augmented small multimodal models to guide large language models for multimodal relation extraction. *Applied Sciences*.

Zexue He, Marco Tulio Ribeiro, and Fereshte Khani. 2023c. Targeted data generation: Finding and fixing model weaknesses. *arXiv preprint arXiv:2305.17804*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister.

2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Chengming Hu, Xuan Li, Dan Liu, Xi Chen, Ju Wang, and Xue Liu. 2022. Teacher-student architecture for knowledge learning: A survey. *arXiv preprint arXiv:2210.17332*.

Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. 2023. Up5: Unbiased foundation model for fairness-aware recommendation. *arXiv preprint arXiv:2305.12090*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Yechan Hwang, Jinsu Lim, Young-Jun Lee, and Ho-Jin Choi. 2023. Augmentation for context in financial numerical reasoning over textual and tabular data with large-scale language model. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Fangkai Jiao, Bosheng Ding, Tianze Luo, and Zhanfeng Mo. 2023. Panda llm: Training data and evaluation for open-sourced chinese instruction-following large language models. *arXiv preprint arXiv:2305.03025*.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. 2022. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, pages 10362–10383. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andres Karjus. 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. *ArXiv*, abs/2309.14379.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. 2023. Q: How to specialize large vision-language models to data-scarce vqa tasks? a: Self-train on unlabeled images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15005–15015.

Mina J Kian, Mingyu Zong, Katrin Fischer, Abhyuday Singh, Anna-Maria Velentza, Pau Sang, Shriya Upadhyay, Anika Gupta, Misha A Faruki, Wallace Browning, et al. 2024. Can an llm-powered socially assistive robot effectively and safely deliver cognitive

behavioral therapy? a study with university students. *arXiv preprint arXiv:2402.17937*.

Fabian Kieser, Peter Wulff, Jochen Kuhn, and Stefan Küchemann. 2023. Educational data augmentation in physics education research using chatgpt. *Physical Review Physics Education Research*, 19(2):020150.

Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. 2023. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities.

Damir Korenčić, Ivan Grubišić, Gretel Liz De La Peña Sarracén, Alejandro Hector Toselli, Berta Chulvi, and Paolo Rosso. 2022. Tackling covid-19 conspiracies on twitter using bert ensembles, gpt-3 augmentation, and graph nns.

Lea Krause, Selene Báez Santamaría, Michiel van der Meer, and Urja Khurana. 2023. Leveraging few-shot data augmentation and waterfall prompting for response generation. *arXiv preprint arXiv:2308.01080*.

Victor C.W. Cheng Kwok-Yan Lam and Zee Kin Yeong. 2023. Applying large language models for enhancing contract drafting. *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workspace (LegalAIIA 2023)*.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.

Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations.

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint arXiv:2307.06090*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. 2023a. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023b. Autoconv: Automatically generating information-seeking conversations with large language models. *arXiv preprint arXiv:2308.06507*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023c. Self-alignment with instruction back-translation. *arXiv preprint arXiv:2308.06259*.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2023d. Large language models as counterfactual generator: Strengths and weaknesses. *arXiv preprint arXiv:2305.14791*.

Yucheng Li, Frank Geurin, and Chenghua Lin. 2023e. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. *ArXiv*, abs/2312.12343.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022a. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022b. Controllable dialogue simulation with in-context learning. *arXiv preprint arXiv:2210.04185*.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Jiancheng Lv, Nan Duan, and M. Zhou. 2020. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In *Conference on Empirical Methods in Natural Language Processing*.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023b. Logicot: Logical chain-of-thought instruction-tuning data collection with gpt-4. *arXiv preprint arXiv:2305.12147*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Sheng Liu, Lei Xing, and James Zou. 2023c. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Yixin Liu, A. R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq R. Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023d. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *ArXiv*, abs/2311.09184.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023e. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023f. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Chandreen Liyanage, Muskan Garg, Vijay Mago, and Sunghwan Sohn. 2023. Augmenting Reddit posts to determine wellness dimensions impacting mental health. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 306–312, Toronto, Canada. Association for Computational Linguistics.

Bo-Ru Lu, Nikita Haduong, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2023. Dialgen: Collaborative human-lm generated dialogues for improved understanding of human-human conversations.

Ruilin Luo, Tianle Gu, Haoling Li, Junzhe Li, Zicheng Lin, Jiayi Li, and Yujiu Yang. 2024a. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. *arXiv preprint arXiv:2401.06072*.

Tianze Luo, Yong Liu, and Sinno Jialin Pan. 2024b. Collaborative sequential recommendations via multi-view gnn-transformers. *ACM Transactions on Information Systems*, 42.

Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. 2023a. Conditional graph generation with graph principal flow network. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*.

Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. 2023b. Fast graph generation via spectral diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli. 2022. European clinical case corpus. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 283–288. Springer International Publishing Cham.

Jordan Meadows, Marco Valentino, and Andre Freitas. 2023. Generating mathematical derivations with large language models. *arXiv preprint arXiv:2307.09998*.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190.

Zhanfeng Mo, Tianze Luo, and Sinno Jialin Pan. 2024. Graph principal flow network for conditional graph generation. In *Proceedings of the ACM Web Conference 2024*.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Yair Neuman, Vladyslav Kozhukhov, and Dan Vilenchik. 2023. Data augmentation for modeling human personality: The dexter machine.

Seokjin Oh, Woohwan Jung, et al. 2023. Data augmentation for neural machine translation using generative language model. *arXiv preprint arXiv:2307.16833*.

Charles O'Neill, Yuan-Sen Ting, Ioana Ciuca, Roberta Raileanu, Jack Miller, and Thang Bui. 2023. Steering language generation: Harnessing contrastive expert guidance and negative prompting for coherent and diverse synthetic data generation. *arXiv preprint arXiv:2308.07645*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint arXiv:2310.05421*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022a. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022b. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Conference on Empirical Methods in Natural Language Processing*.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2023. Can llms augment low-resource reading comprehension datasets? opportunities and challenges. *arXiv preprint arXiv:2309.12426*.

Shouvon Sarker, Lijun Qian, and Xishuang Dong. 2023. Medical data augmentation via chatgpt: A case study on medication identification and medication event classification. *arXiv preprint arXiv:2306.07297*.

Viktor Schlegel, Hao Li, Yuping Wu, Anand Subramanian, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Daniel Beck, Xiaojun Zeng, Riza Theresa Batista-Navarro, Stefan Winkler, et al. 2023. Pulsar at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records. *arXiv preprint arXiv:2307.02006*.

Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. 2023a. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087, Toronto, Canada. Association for Computational Linguistics.

Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. 2023b. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Răzvan-Alexandru Smădu, Sebastian-Vasile Echim, Dumitru-Clementin Cercel, Iuliana Marin, and Florin Pop. 2023. From fake to hyperpartisan news detection using domain adaptation. *arXiv preprint arXiv:2308.02185*.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

Vishvesh Soni. 2023. Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies*, 7(1):67–89.

Joe Stacey and Marek Rei. 2023. Improving robustness in knowledge distillation using domain-targeted data augmentation. *arXiv preprint arXiv:2305.13067*.

Xunzhu Tang, Zhenghan Chen, Kisub Kim, Haoye Tian, Saad Ezzini, and Jacques Klein. 2023. Just-in-time security patch detection–llm at the rescue for data augmentation. *arXiv preprint arXiv:2312.01241*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325.*

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022a. A unified dialogue user simulator for few-shot data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022b. A unified dialogue user simulator for few-shot data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453.*

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879.*

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Conference on Empirical Methods in Natural Language Processing.*

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, et al. 2024. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268.*

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023c. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174.*

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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

Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. 2023. Multimodal data augmentation for image captioning using diffusion models. *arXiv preprint arXiv:2305.01855.*

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853.*

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328.*

Qing Ru Yong and Alex Mitchell. 2023. From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *International Conference on Interactive Digital Storytelling*, pages 395–409. Springer.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J Yuan, R Tang, X Jiang, and X Hu. 2023a. Large language models for healthcare data augmentation: An example on patient-trial matching. *arXiv preprint arXiv:2303.16756.*

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023b. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. *ArXiv*, abs/2303.16756.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingying Zeng, Richard Huang, Waleed Malik, Langxuan Yin, Bojan Babic, Danny Shacham, Xiao Yan, Jaewon Yang, and Qi He. 2024. Large language models for social networks: Applications, challenges, and solutions. *arXiv preprint arXiv:2401.02575.*

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158.*

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069.*

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023c. Recommendation as instruction following: A large language model empowered recommendation approach.

Kechi Zhang, Ge Li, Jia Li, Zhuo Li, and Zhi Jin. 2023d. Toolcoder: Teach code generation models to use apis with search tools. *arXiv preprint arXiv:2305.04032*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023e. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023f. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

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

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2024. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36.

Siyan Zhao, John Dang, and Aditya Grover. 2023a. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023b. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023b. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1006–1014.

Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. 2023a. Calypso: Llms as dungeon master's assistants. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 380–390.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023b. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

A Tasks

In this section, we classify the works based on the task that data augmentation is used to solve. We categorize the tasks as fundamental tasks, understanding based tasks, generation tasks, multi-modal tasks, and other tasks. We elaborate on the fundamental tasks in appendix A.1, which includes text classification, machine translation, and sequence tagging. We explain the understanding and inference tasks in appendix A.2, with question answering, logical reasoning, and natural language inference. We discuss the generation and interaction tasks in appendix A.3, which includes summarization, data-to-text NLG, open-ended and conditional generation, and dialog. We introduce multi-modal tasks in appendix A.4, which includes story telling and VQA tasks. Lastly, we summarize some other tasks that do not belong to above mentioned categories in appendix A.5.

A.1 Fundamental Tasks

In this subsection, we discuss some fundamental tasks in NLP. These tasks are considered relatively simple to the powerful LLMs nowadays. LLMs are able to achieve human-level performance in tasks like machine translation. Ding et al. (2022b) studies 3 different approaches to utilize GPT-3 for data augmentation: 1) prompt-guided unlabeled data annotation (PGDA); 2) prompt-guided training data generation (PGDG); and 3) dictionary assisted training data generation (DADG). Results show that it is largely reliable to count on the LLMs as data augmentation tools.

Text Classification Text classification involves categorizing text pieces into predefined classes or

categories. For LLMs, text classification demonstrates the model's ability in understanding the content, context, nature, or sentiment of textual data. A powerful language model should accurately assign labels to diverse types of information. The significance of text classification lies in its applications across sentiment analysis, spam detection, and content categorization, contributing to streamlined information organization and retrieval.

Recently, some researchers compares LLM classified data with human experts. Alizadeh et al. (2023) compare the performance of two widely-used open-source LLMs, HugginChat and FLAN, with that of ChatGPT as well as MTurk on multiple text annotation tasks. Results show that open-source LLMs surpass MTurk in most tasks, and are comparable with ChatGPT in many tasks. Törnberg (2023) uses ChatGPT as a classifier for US political messages from Twitter on whether it is from a Democrat or a Republican. Experiment results show that LLM's classified results achieve a better accuracy than the judgements from the Amazon Mturk experts.

Many works employ LLM augmentation for sentiment analysis. Møller et al. (2023) investigate GPT-4 and ChatGPT's ability for generating synthetic data by augmenting small sets of human-generated training samples. Experiments on sentiment analysis and hate speech showcases LLM's great power in data augmentation on classification tasks. Belal et al. (2023) conduct experiments testing ChatGPT's ability in annotating for sentiment analysis task. Results show that the ChatGPT is a good tool for sentiment analysis, and it has the ability to understand emojis, sarcasm, and irony.

LLM is also proven to be powerful in detecting fake news or human intention. To improve the results, Smădu et al. (2023) conduct data augmentation using GPT-2 to enable domain adaptation on a fake news detection task. The augmentation technique conditionally generates new examples given the news types (left-wing, right-wing, and mainstream). Fang et al. (2023) investigate ChatGPT's ability as a data augmentation technique for enhancing compositional generalization in open intent detection tasks. Experiment results show that ChatGPT's paraphrases for utterances largely outperforms the baseline.

From an application perspective, LLM augmentation is also widely used in medical related classification tasks. To address the patient privacy challenge in patient-trial matching task, Yuan et al. (2023a) use desensitized patient data as a prompt to guide the ChatGPT in the augmentation process of the trial data. Given the criteria of a clinical trial, the method generates augmented data points adhere to its constraints. Sarker et al. (2023) explore ChatGPT for data augmentation in EHR data analysis through prompt engineering. Experiment results show that data augmentation based on ChatGPT improves performance for both medication identification and medication event classification.

Machine Translation Machine Translation is the process of automatically converting text or speech from one language into another. As the LLMs become more powerful, machine translation stands as a critical application, showcasing the model's prowess in comprehending linguistic nuances and producing coherent and contextually accurate translations. The importance of machine translation lies in its ability to bridge language barriers, facilitating global communication and information exchange by rendering content accessible to a wider audience.

For machine translation, data augmentation using LLMs serves two major purposes. The first purpose is enlarge data corpus to faciliate training, and the second purpose is to enhance evaluation quality. To achieve the first purpose, Oh et al. (2023) explore prompt-based data augmentation approaches that leveraging ChatGPT. It compares three paradigms "paraphrase", "multi-target", and "storytelling" and generates synthetic data on Korean-German language pairs from AI-hub dataset. To achieve the second purpose, DATScore (Eddine et al., 2022) utilizes augmented data translated from the source and reference texts using ChatGPT. Experiments on WMT17 and WMT18 shows that DATScore correlates better with human judgements than BLEU, BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021) on machine translation task.

Sequence Tagging Sequence tagging involves the task of assigning specific labels or tags to individual elements within a sequence, such as words or tokens in a sentence. For LLMs, sequence tagging is a fundamental capability that underscores the model's capacity to discern and annotate various elements in a given context. The significance of sequence tagging lies in its applications across

diverse fields, from Named Entity Recognition in natural language processing to part-of-speech tagging in linguistic analysis, enabling more detailed understanding of sequential data.

For sequence tagging, existing works use LLM to generate synthetic data with two major approaches, generating spans and generating labels.

- Generating Spans: Sharma et al. (2023b) augment name entity recognition (NER) data by paraphrasing and annotating entity spans in generations using GPT-3 variants.

- Generating Labels: Meoni et al. (2023) propose a weak supervision technique utilizing InstructGPT as a data augmentation tool to predict the annotation. Experiments on a clinical entity extraction showcases the high quality of the augmented data on distilling a Bert-base model. Deng et al. (2023) propose a data augmentation method on Universal Decompositional Semantic (UDS) Parsing. It utilizes ChatGPT to predict the syntactic tree and POS tag and generate pseudo labels for the unlabeled data.

A.2 Understanding and Inference Tasks

In this subsection, we discuss the tasks that are related to text understanding and inference. Question Answering tasks demand the models to understand the given question and provide an accurate and relevant answer based on the given context or external knowledge. Logical reasoning tasks require the model to apply logical reasoning skills to comprehend and infer from given data or statements. Natural Language Inference (NLI) tasks involve understanding the relationship between sentences or phrases, such as whether one statement entails, contradicts, or is neutral to another. LLMs also play an important role in augmenting the datasets for these tasks involved text understanding.

Question Answering Question answering is the cognitive process of providing relevant and accurate responses to posed queries, a task that holds immense importance within LLMs. In this context, question answering demonstrates the model's capacity to understand complex queries, retrieve information from vast datasets, and generate coherent and contextually appropriate answers. The significance of question answering lies in its applications across information retrieval systems, virtual assistants, and educational tools, contributing to more effective and efficient knowledge dissemination.

Question answering task is broad in scope and includes multiple downstream tasks. Some QA tasks are multiple choice based like commonsense reasoning, and fact checking. Other QA tasks are free-form in answer like story completing and machine reading comprehension (MRC). We find that both types of QA tasks can be augmented by large language model.

- Commonsense Reasoning: Automate-CoT (Shum et al., 2023) proposes a fully automatic pipeline for finding better chain-of-thought prompts to solve commonsense reasoning question answering tasks. This pipeline leverages powerful LLMs like GPT-3 to augment reasoning paths, prune incorrect paths, and select optimal combinations of exemplars.

- Story Completing: Whitehouse et al. (2023) explore the generation of synthetic data for question answering under the multilingual setting using various LLMs. This work conducts experiments with dolly-12B, stablevicuna-13B, ChatGPT, and GPT-4 on three datasets that covers over 20 languages on commonsense reasoning and story completing tasks.

- Fact Checking: SCOTT (Wang et al., 2023b) prompts a GPT-neox-20B model to generate annotated data to train a smaller student model to perform QA tasks. Experiments conducted on four QA datasets across commonsense reasoning and fact checking showcase the efficiency of this data augmentation technique.

- MRC: Samuel et al. (2023) introduce a GPT-4 based data augmentation technique targeting the low source setting on machine reading comprehension task. This technique separately generates passsages, questions, and answers using in-context learning.

- MedicalQA: Dr.LLaMA (Guo et al., 2023) employs ChatGPT and GPT-4 to either rewrite existing medical question-answering pairs or generate new pairs from the training dataset with zero-shot prompting. Such data augmentation helps to train a LLaMA model specialised on medical knowledge.

Logical Reasoning Logical reasoning involves the ability to analyze and draw conclusions based

on given information, a crucial cognitive skill that LLMs aim to master. In the context of LLMs, logical reasoning reflects the model's aptitude for understanding and manipulating symbolic representations, solving problems, and making deductions. The significance of logical reasoning lies in its application across various domains, from puzzle-solving and decision-making to complex problem-solving tasks, showcasing the model's capacity for high-level cognitive functions.

Within the logical reasoning task, Chain-of-Thought has been a popular technique that improves the reasoning process for the language models. CoT is hence frequently used in data augmentation for logical reasoning tasks. Fine-tune-CoT (Ho et al., 2022) generates multiple reasoning solutions from LLMs like variants of GPT-3 17B with stochastic sampling to augment the training data for student models. This method largely utilizes the reasoning ability of the LLMs. LogiCoT (Liu et al., 2023b) is a chain-of-thought instruction-tuning dataset designed explicitly for logical reasoning. This dataset is augmented by GPT-4 from high quality logical reasoning data, and is used to fine-tune a LLaMA-7B model. Distilling step-by-step (Hsieh et al., 2023) is a new simple mechanism for training smaller models with less training data. It prompts PaLM-540B with logical reasoning questions and extracts rationale from it using CoT.

The reasoning ability of LLMs is proven to be powerful, and the augmented data can be used to improve smaller models or even LLM itself. Orca (Mukherjee et al., 2023) is a 13B model that learns to imitate the reasoning process of larger models. Training data of Orca is augmented from ChatGPT and GPT-4, including explanation traces, step-by-step thought processes, and other complex instructions. Instruction backtranslation (Li et al., 2023c) uses the LLM to both augment and curate high quality training examples to improve its own performance. Experiments on LLaMA 7B, 33B, and 65B showcase the efficiency of the model on reasoning tasks. Bao et al. (2023) propose an evaluation benchmark of LLMs on logical reasoning tasks. It employs logic-driven data augmentation on ChatGPT and GPT-4 to enhance the performance on logical reasoning tasks. The augmentation parts include context, question, and option.

Data augmentation has also been employed in LLM for science. For instance, Kieser et al. (2023)

use ChatGPT to augment data for physics education, leveraging the LLM's ability to solve quantitative reasoning tasks in physics and concept tests such as the Force Concept Inventory (FCI).

Natural Language Inference Natural Language Inference (NLI) is the task of determining the logical relationship between the premise and the hypothesis, such as entailment, contradiction, or neutrality. Within LLMs, NLI represents a sophisticated task that demonstrates the model's ability to understand two statements and infer meaningful connections between them.

For NLI task, LLMs are powerful in generating augmented data that tackle two challenges. The first challenge is the lack of counterfactual examples, and the second challenge is the out-of-distribution (OOD) or low resouce problem.

- counterfactual example: Li et al. (2023d) employ ChatGPT to generate counterfactual samples and counterfactual labels for each original sample. By merging the original data with the counterfactual data, an augmented dataset is formed to enhance the small models' performance on the NLI task. Disco (Chen et al., 2023e) identifies potential locations for performing counterfactual perturbations on the target instances, and then prompts GPT-3 to generate perturbations for NLI task.

- OOD or low resource: Stacey and Rei (2023) generate examples from a GPT-3 model to improve performance in out-of-distribution setting to mimic text that may appear in unseen genres for NLI task. By distilling a Bert model, the performance improves on both MNLI and SNLI settings. TDG (He et al., 2023c) clusters NLI data into potential challenging subgroups and estimate which subgroups benefit from additional data. Then, it uses GPT-3 coupled with local subgroup models to conduct data augmentation to improve the performance for particularly challenging subgroups.

A.3 Generation and Interaction Tasks

In this subsection, we discuss the generation and interaction tasks. Summarization tasks involve producing a concise and coherent summary of a longer text while retaining the key information and overall meaning. Data-to-text NLG tasks requires converting structured data into understandable and fluent

natural language text. Open-ended and conditional generation tasks require the model to generate text based on specific conditions or prompts, including creative and unrestricted text generation. Dialogue tasks involve generating conversational responses in a dialogue setting, requiring the model to maintain context and coherence over multiple turns of conversation. LLMs can be helpful to not only generate answering but also generate questions to augment the dataset.

Summarization Text summarization is the process of distilling the most important information from a source text and presenting it in a condensed form. In the realm of Large Language Models (LLMs), summarization becomes a pivotal task, reflecting the model's ability to comprehend, process, and concisely reproduce textual information. The significance of summarization lies in its ability to transform extensive information into a digestible format, aiding in efficient information consumption and understanding. In the context of LLMs, summarization is not just about reducing text length but also about maintaining its essence, coherence, and factual integrity. Chintagunta et al. (2021) use GPT3 to summarize medical dialogues as labels to train summarization models. Liu et al. (2023f) adopt LLMs (e.g. GPT3D3, ChatGPT, and GPT4) to evaluate the quality (i.e. quality score) of the summarization output by a model, and design a contrastive learning loss function based on the quality score to train a summarization model. Schlegel et al. (2023) use LLMs to generate synthetic data on summarizing patient-doctor dialogues into clinical records. The synthetic data are then used to train the PULSAR model (Schlegel et al., 2023), enabling PULSAR to achieve high scores in the MediQA-Sum task.

Summarization in LLMs can be broadly categorized into two types: extractive and abstractive. Extractive summarization involves identifying key sentences or fragments in the text and stitching them together to form a summary. Here, the LLMs focus on selecting the most informative parts of the original text. In contrast, abstractive summarization requires the LLM to generate new phrases or sentences that capture the core ideas of the text, often paraphrasing or rephrasing the content. This demands a deeper level of understanding and language generation capability from the model. LLMs like GPT-3 and BERT have shown proficiency in both types, with abstractive summarization posing

more challenges in terms of generating coherent and contextually accurate summaries.

Data-to-Text NLG Data-to-Text Natural Language Generation (NLG) is a subfield of artificial intelligence that focuses on converting structured data into coherent and readable narrative text. This task is crucial in making data accessible and understandable to a broader audience, transcending the barriers of technical expertise. The significance of Data-to-Text NLG lies in its ability to bridge the gap between raw data and human communication, allowing for effective data-driven storytelling and reporting.

Large Language Models (LLMs) approach Data-to-Text NLG through various methodologies. Krause et al. (2023) design a waterfall prompting technique using a combination of both GPT-3 and ChatGPT to enhance Data-to-Text response generation. Li et al. (2023b) use LLMs for synthetic conversation generation to tackle the data deficiency problem in training information-seeking conversation models.

Open-ended and Conditional Generation Open-ended and conditional generation in the context of Large Language Models (LLMs) refers to the creation of text that is not strictly predetermined, offering a wide range of possibilities based on given conditions or prompts. This task is particularly challenging for LLMs because it requires a blend of creativity, contextual understanding, and adherence to specified constraints or themes. Open-ended generation pushes the model's capabilities in producing coherent, relevant, and often inventive content, while conditional generation demands adherence to specific guidelines or objectives, adding layers of complexity to the task.

Dai et al. (2023) propose AugGPT, which is a text data augmentation approach based on ChatGPT that rephrases each sentence in the training samples into multiple conceptually similar but semantically different samples. O'Neill et al. (2023) design Semantic Text Enhancement via Embedding Repositioning (STEER) that utilize LLMs to produce text characteristic of the specific domain. Meadows et al. (2023) utilize LLMs as a symbolic engine to generate derivations of equations at scale. The augmented data is tested on a finetuned T5 model, showing superior results than GPT models.

Approaches and Techniques LLMs employ var-

ious models and techniques for open-ended and conditional generation:

- Transformer Models: These, like GPT-3, are currently the most prominent in this field. They use deep learning techniques to generate text that is contextually relevant and coherent.

- Fine-tuning on Specific Data: This involves training the LLM on specific types of data or genres to improve performance in certain areas, like poetry or technical writing.

- Reinforcement Learning from Human Feedback (RLHF): This technique involves training models based on feedback to align the model's outputs more closely with human preferences or specific task requirements.

Dialog Dialog systems, a crucial component of human-computer interaction, are designed to enable effective and natural communication between humans and machines. These systems, often manifested as chatbots or virtual assistants, simulate human-like conversation, allowing users to interact with digital devices or services in a more intuitive and familiar manner. The importance of dialog systems lies in their ability to provide efficient, user-friendly interfaces for a variety of applications, from customer service to personal assistance, thereby enhancing user experience and accessibility.

Large Language Models (LLMs) have significantly revolutionized the field of conversational AI. With their advanced natural language processing capabilities, LLMs like GPT-3 have been instrumental in creating chatbots and virtual assistants that can understand and generate human-like text. This advancement has led to more sophisticated and versatile dialog systems capable of handling a wide range of conversational topics and styles. LLMs contribute to conversational AI by improving the fluidity and relevance of interactions, making these systems more engaging and helpful to users.

Zheng et al. (2023b) use LLMs for dialogue augmentation to support emotional conversation capabilities for dialogue models. Wan et al. (2022b) use T5 as a dialogue user simulator to generate few-shot data augmentations for training conversation models. Li et al. (2022b) design the Dialogic, which is a dialogue simulation method based on large language models in-context learning to automate dataset creation. Dialogic can automatically

select in-context examples for demonstration and prompts GPT-3 to generate new dialogues and annotations in a controllable way, and can further enhance dialogue models' performance. Labruna et al. (2023) utilize ChatGPT to generate goal-oriented dialogues, showing that the ChatGPT annotations are on par with human-generated annotations.

Case Studies and Applications LLMs have been successfully implemented in various dialog systems, including: (1) Customer Service: Chatbots powered by LLMs have been employed by numerous businesses to provide instant customer support, handle inquiries, and offer personalized recommendations. (Soni, 2023; Roumeliotis et al., 2024; Pandya and Holia, 2023) (2) Therapy Bots: LLMs are being used in mental health applications, offering conversational therapy and support, though with necessary caution and oversight.(Kian et al., 2024; Bill and Eriksson, 2023) (3) Interactive Entertainment: In the gaming and entertainment industry, LLMs have enabled the creation of interactive narratives and characters, enhancing user engagement and experience. (Yong and Mitchell, 2023; Zhu et al., 2023a)

A.4 Multi-Modal Tasks

Other than the pure NLP tasks, multi-modal tasks can also be augmented with LLMs. These tasks either involves generating texts given images or generating images given tasks. Some examples include visual question answering, image captioning, and other tasks.

Visual Question Answering Visual Question Answering (VQA) is a multi-modal task where a system provides answers to questions based on visual content, such as images or videos. This task requires the integration of visual perception with language understanding, challenging the model to interpret visual data and articulate relevant responses in natural language.

SelTDA (Khan et al., 2023) introduces a novel approach for fine-tuning large Visual Language Models (VLMs) on small-scale Visual Question Answering (VQA) datasets by generating question-answer pseudolabels for unlabeled images, effectively augmenting the original dataset. This method enhances model robustness against adversarial questions, improves domain generalization, and retains numerical reasoning skills without needing extra annotations or changes to the architecture. T-SciQ (Wang et al., 2023a) enhances science

question answering by using large language model (LLM) insights to create detailed reasoning paths (CoT) as educational signals. This technique trains smaller models to tackle complex problems with CoT reasoning and introduces a unique data mixing strategy for better learning materials across varying difficulty levels. Demonstrating its effectiveness, T-SciQ sets a new record on the ScienceQA benchmark with a 96.18% accuracy, outdoing the top fine-tuned baseline by 4.5%.

To tackle Visual Question Answering (VQA) in scenarios with limited labeled data, Askarian et al. (2022) expand the initial dataset, infusing the VQA model with enhanced inductive biases through newly generated questions based on image annotations. This approach yields up to a 34% accuracy improvement over baseline models trained solely on the original labeled dataset.

VQA has practical applications in areas such as accessibility technology, where it aids visually impaired users in understanding their surroundings. It's also used in customer service to answer queries based on product images and in educational tools that help students learn through visual aids.

Image Captioning and Editing Image captioning is a multi-modal task that involves generating descriptive text for images. This requires a system to not only recognize the visual elements within an image but also to understand their context and how they relate to each other, synthesizing this information into coherent, natural language sentences. It bridges the gap between visual perception and language generation, challenging models to accurately interpret visual data and express these interpretations as human-like captions.

ChatBridge (Zhao et al., 2023b) connects text, images, videos, and audio using language. It leverages LLM and extends their zero-shot capabilities to incorporate diverse multi-modal inputs. The training includes two stages, first learning to link each modality with language, then fine-tuning with the dataset MULTIS for specific multi-modal tasks. Xiao et al. (2023) utilize Stable Diffusion to generate high-quality image-caption pairs for multi-modal data augmentation, showing significant improvements on the MS COCO dataset, especially with limited training data. This approach surpasses previous unpaired image captioning methods and further enhances training efficiency and effectiveness through quality-based filtering of generated data. ALIA (Dunlap et al., 2024) leverages vi-

sion and language models for automated language-guided image editing, enriching training datasets while preserving class-relevant information. This method enhances dataset diversity without compromising visual consistency, outperforming traditional and text-to-image augmentation in fine-grained classification tasks, improving domain generalization and reducing contextual bias.

Image captioning has diverse applications in various fields, enhancing accessibility for visually impaired users by providing textual descriptions of images on the web or in digital media. It plays a crucial role in social media, where automatic captioning can improve user engagement and content accessibility. Furthermore, in the educational sector, image captioning aids in creating more interactive and accessible learning materials, allowing students to gain a deeper understanding of visual content.

Other multi-modal tasks Other than visual question answering and image captioning, some other multi-modal tasks also use LLM to conduct data augmentation. One example is Multi-modal Named Entity and Relation Extraction. Chen and Feng (2023) distill the reasoning ability of LLMs into small student model by generating a intermediate reasoning steps (CoT). It draws out reasoning skills from LLMs using detailed prompts that include multi-grain (noun, sentence, multi-modality) and data augmentation (style, entity, image) dimensions. The new technique simplifies the complex reasoning into a form that smaller models can use. Another example is tooling. GPT4Tools (Yang et al., 2024) employs self-instruct to enable open-source LLMs to use multi-modal tools. We will explain this paper more in detail in the next subsection.

A.5 Other Tasks

Other than the above mentioned tasks, many other NLP tasks also employ LLMs as data augmenting tools. We will discuss them in this subsection.

Tooling Tooling tasks in the context of Large Language Models (LLMs) refer to the application of these models in developing tools that aid in various aspects of programming, content creation, and data analysis. These tasks leverage the language understanding and generation capabilities of LLMs to automate, optimize, and enhance various workflows and processes.

One key area of data augmentation's usage of tooling is in code generation and assistance. LLMs are employed to assist in writing and optimizing code, offering suggestions, debugging help, and even generating code snippets based on user requirements. Dong et al. (2023c) explore the existing data augmentation techniques from NLP and graph learning for enhancing training quality in source code learning. After reviewing and categorizing the literature, 11 data augmentation methods from NLP and graph learning are identified as potentially applicable to source code learning. Zhang et al. (2023d) enhance code generation by integrating tool usage information via an automated annotation method and API search tools, showing significant improvements in benchmarks. Despite its smaller size, ToolCoder's performance is comparable to GPT-3.5, illustrating the benefits of embedding programming tools into code generation models.

Another possible area of data augmentation using LLM is to train models as classification or detection tools. For instance, (Smădu et al., 2023) work as very good hyper-partisan news detection tool. It employs GPT-2 employ to conditionally generate new examples given the news types (i.e., left-wing, right-wing, and mainstream). Then, it fine-tunes the GPT-2 base model on the hyperpartisan Buzzfeed dataset to generate new samples. These detection models are very useful in industries and applications.

Other than the above-mentioned tooling tasks, LLM can also be used to augment data for setting up a comprehensive benchmark for multiple tools. GPT4Tools (Yang et al., 2024) empowers open-source large language models like LLaMA and OPT with tool-using abilities, leveraging Low-Rank Adaptation (LoRA) optimization for visual tasks using GPT-4 for augmentation. It includes a benchmark for evaluating tool use in models, showing significant improvements in both seen and unseen tool invocation accuracy through extensive experiments.

The main challenges in tooling with LLMs include ensuring accuracy, maintaining context relevance, and integrating seamlessly with existing tools and workflows. Solutions involve continuous model training, user feedback loops for model refinement, and developing intuitive interfaces that facilitate easy integration and usage.

B Domains

The preceding sections have outlined prominent data augmentation techniques applied across diverse tasks. In this section, we delve further into categorizing the specific application domains of these data augmentation methodologies. We systematically categorize these domains into distinct areas: clinical, finance, legal and social science.

B.1 Clinical

The emergence of clinical applications powered by LLMs is poised to transform the healthcare system. LLMs such as GPT-4 have demonstrated a remarkable ability to understand and generate text, making them valuable tools for tasks ranging from enhancing medical transcription to assisting with medical diagnoses. Integrating these models into clinical settings can significantly enhance the efficiency and accuracy of various healthcare processes, but this potential is accompanied by major challenges. For example, establishing close alignment of model predictions with the assessment of medical professionals and addressing biases inherent in the training data must be addressed. Navigating these challenges is essential for realizing the full potential of large language models in healthcare and ultimately advancing the quality of patient care and medical research.

One of the major obstacles to adapting LLMs for clinical tasks is the scarcity of data. First, the accessibility of medical data is limited. In the United States, for example, the distribution of medical data must abide by HIPAAA regulations, which protect patient information. In addition, annotation of medical data is a time-intensive and costly undertaking (Chintagunta et al., 2021; Meoni et al., 2023), further restricting the availability of data for AI development in the clinical domain. Together, these limitations necessitate alternative routes to generating large, high-quality datasets.

To this end, an emerging stream of research is focusing on data augmentation approaches that generate synthetic data to facilitate the adaptation of LLMs in medical domain. Electronic Health Records (EHRs) play a crucial role in the modern clinical field. Several works propose data augmentation approaches to facilitate the usage of LLMs for EHR-related applications. Meoni et al. (2023) explore annotating EHR data with InstructGPT-3 to improve the performance of distilled BERT-based models for multilingual clinical entity ex-

traction. The proposed method outperforms dictionary supervision in extracting clinical entities on the E3C multilingual data set (Magnini et al., 2022). Sarker et al. (2023) tackles the medication identification and event classification problem in EHRs. By prompting ChatGPT to rephrase original sentences of the Contextualized Medication Event Dataset, the authors enhance the performance of different pre-trained BERT models. Yuan et al. (2023a) prompt LLMs with desensitized patient data to create privacy-aware supplementary data points to comprehend EHRs for clinical patient-trial matching. Guevara et al. (2024) utilize LM-generated synthetic clinical text at the fine-tuning stage to improve LMs ability on extracting social determinants of health, such as employment status and housing issues, from EHRs. Aside from prompting LLMs to obtain synthetic data, recent work explores generating mixed-type tabular EHRs with diffusion models (Ceritli et al., 2023).

Medical question answering is another field of interest. Guo et al. (2023) develop smaller efficient language models with augmentation data generated by LLMs on the PubMedQA dataset. The proposed method uses GPT-3.5 Turbo and GPT-4 to rewrite existing QA pairs or generate new pairs via zero-shot prompting. Results show that the small fine-tuned model with the data augmentation approach outperforms few-shot GPT4 on the PubMedQA dataset. For medical dialogue applications, GPT-3-ENS is introduced to create high-quality synthetic training data via few-shot learning and an ensemble method. Together with the human labeled data, the generated data help yield more performant models in terms of medical accuracy and coherency (Chintagunta et al., 2021).

B.2 Finance

One major line of research centers on leveraging LLM-based data augmentation for finance-related reasoning tasks. To improve the numerical reasoning ability for financial question answering tasks, Hwang et al. (2023) put forth a novel context augmentation method that generates synthetic financial contexts based on the arithmetic program and operators present in the question. To address the limited availability of labeled datasets in financial analysis and interpretation tasks, Chu et al. (2023) propose an abductive augmentation reasoning framework that refines the pseudo-labels generated from a small labeled dataset. This aims to correct noisy

labels and improve the quality of training data to enhance the performance of the financial large language model.

Another application is intent detection, that aims to classify the user's intent given an utterance. To alleviate the problem of data scarcity, researchers employ LLM-based data augmentation techniques to generate synthetic training data. This involves creating artificial utterances corresponding to given intents. A common strategy is to construct a data generation and filtering pipeline. Sahu et al. (2022a) prompt GPT models to augment the training data for each intent class. Subsequently, the same GPT model is utilized as a classifier to filter and relabel the generated data. Similarly, Lin et al. (2023) employ GPT-3 and OPT to synthesize new data points for each class, then apply point-wise V-information to filter out examples that are not relevant to the desired intent. Both methods achieve state-of-the-art performance on Banking77 (Casanueva et al., 2020), a challenging intent detection dataset with fine-grained banking-related intents.

B.3 E-commerce

Adopting data augmentation techniques by LLMs in E-commerce applications such as search and recommender systems, shopping assistance, and customer supports are prospective trends in the development of LLMs.

Recommender systems are algorithms designed to predict and suggest items of interest to users, based on their preferences and behavior (Guo et al., 2020; Gao et al., 2023; Luo et al., 2024b; Hao et al., 2021; Zhou et al., 2020; Da'u and Salim, 2020). These systems play a crucial role in various domains, such as e-commerce, entertainment, and content platforms, by personalizing user experiences and enhancing engagement. The combination of recommender systems with LLMs is an emerging trend where LLMs can largely enhance the capability of the recommender systems (Fan et al., 2023). Cui et al. (2022); Christakopoulou et al. (2023); Hua et al. (2023) utilize LLMs with prompt tuning: by adding prompt tokens to LLMs and then updating them based on task-specific recommendation datasets.

B.4 Social Sciences

Large language model (LLM)-based data augmentation provides the potential for advancing scientific understanding of human social and political

behavior. Several studies have demonstrated that LLMs serve as effective surrogates for human respondents across various social science tasks. For instance, Törnberg (2023) use ChatGPT to annotate and classify Twitter messages, showcasing the model's ability to accurately identify political affiliations by reasoning based on contextual knowledge and inferences about the author's intentions, traditionally considered uniquely human qualities. In addition, another study (Argyle et al., 2023) reveal that GPT-3 can simulate viewpoints of demographically diverse sub-populations within the U.S. political domain. Conditioning an LLM with socio-demographic profiles allows it to mirror responses of humans with matching demographic traits, uncovering patterns between ideas, attitudes, political views, and socio-cultural contexts. This suggests that synthetic samples generated by LLMs offer a reliable means for researchers to explore hypotheses before deploying costly studies involving human subjects in social science research. Griffin et al. (2023) extend the scope beyond static psychological modeling, illustrating LLMs' ability to model dynamic psychological changes in response to influencing input.

The capability of LLMs in generating human-like responses extends their application in data augmentation to education research. Kieser et al. (2023) delve into the potential of employing ChatGPT to synthesise data that closely resemble students from various cohorts possessing different preconceptions. These data augmentation techniques could be of interest to education researchers as it would save a significant amount of time and effort in the development of assessments and concept tests.

LLMs can enhance social network analysis due to their adeptness in modeling complex data and extracting valuable insights (Zeng et al., 2024). Social networks, both online and offline, serve as fundamental structures for human interaction and information dissemination. These networks encapsulate the intricate web of relationships between individuals, organizations, and communities, forming the basis of societal dynamics. Computational approaches leverage techniques such as graph theory, graph representation learning, and data mining to simulate, infer, or construct social networks from large-scale data sources (Jo et al., 2022; Luo et al., 2023b,a; Mo et al., 2024). These generated networks not only provide insights into human behav-

ior and social dynamics but also facilitate various applications such as recommendation systems, targeted advertising, and sentiment analysis.

LLMs have emerged as powerful tools for generating social knowledge graphs due to their advanced natural language processing capabilities. By analyzing vast amounts of textual data from diverse sources such as social media, forums, and news articles, LLMs can extract valuable information about relationships, entities, and events. Through their ability to understand context, infer semantic meaning, and detect patterns, LLMs can construct comprehensive and accurate representations of social networks. These generated knowledge graphs not only provide insights into the structure and dynamics of social interactions but also facilitate tasks such as community detection, trend analysis, and information retrieval. Additionally, LLMs offer the flexibility to adapt and evolve with changing data and can continuously refine and update social knowledge graphs to reflect real-world dynamics.

Another emergent direction of research is focused on the field of psychology. Neuman et al. (2023) contribute to this area with a personality data augmentation approach that combines LLMs and domain expertise. Their method involves training GPT-2 to generate specific personality types by completing sentences carefully selected by domain experts representing the personality's beliefs. This approach proves particularly valuable in personality modeling tasks where there is a scarcity of large amounts of high-quality labeled data, such as labeled texts from clinically diagnosed psychopaths. Liyanage et al. (2023) focus on the challenge of imbalanced datasets in mental wellness on Reddit posts. They design prompt-based data augmentation methods using ChatGPT and GPT-3 models to aid in the development of classification models for detecting mental health issues. Meanwhile, Zhang et al. (2023e) propose enhancing dialogue models for mental health support through data augmentation with additional annotations that take the form of reasoning support provided by prompt-based interactions with LLM experts, aiming to improve the robustness of the models in addressing mental health concerns.

B.5 Legal

There is a rising interest in incorporating LLMs into the legal domain to automate tasks such as

analysis and generation of legal documents. DALE is introduced to generate coherent and diverse augmentations augmentation data for low-resource legal NLP tasks (Ghosh et al., 2023). DALE uses selective masking and conditional generation based on a BART model on legal documents. The generated data is then utilized to enhance performance of downstream models in low-resource Legal NLP tasks. Kwok-Yan Lam and Yeong (2023) proposes integrating traditional AI and generative AI techniques, applying ChatGPT for contract drafting, and evaluating AI-generated clauses with LLMs to streamline the task for legal professionals.