Large Language Models for Data Annotation: A Survey

Anonymous ACL submission

Abstract

Data annotation generally refers to the labeling or generating of raw data with relevant information, which could be used for improving the efficacy of machine learning models. The process, however, is labor-intensive and costly. The emergence of advanced Large Language Models (LLMs), exemplified by GPT-4, presents an unprecedented opportunity to automate the complicated process of data annotation. While existing surveys have extensively covered LLM architecture, training, and general applications, we uniquely focus on their specific utility for data annotation. This survey contributes to three core aspects: LLM-Based Annotation Generation, LLM-Generated Annotations Assessment, and LLM-Generated Annotations Utilization. Furthermore, this survey includes an in-depth taxonomy of data types that LLMs can annotate, a comprehensive review of learning strategies for models utilizing LLM-generated annotations, and a detailed discussion of the primary challenges and limitations associated with using LLMs for data annotation. Serving as a key guide, this survey aims to assist researchers and practitioners in exploring the potential of the latest LLMs for data annotation, thereby fostering future advancements in this critical field.

1 Introduction

In the complex realm of machine learning and natural language processing (NLP), data annotation stands out as a critical yet challenging task, extending beyond simple label attachment to encompass a diverse array of fundamental or auxiliary information. This detailed process typically involves ❶ categorizing raw data with class or task labels for basic classification, ❷ adding intermediate labels for contextual depth (Yu et al., 2022), ❸ assigning confidence scores to assess annotation reliability (Lin et al., 2022), ❹ applying alignment or preference labels to tailor outputs to specific criteria or user needs, ❺ annotating entity relationships to understand how entities within a dataset interact with each other (Wadhwa et al., 2023), ❻ marking semantic roles to define the underlying roles that entities play in a sentence (Larionov et al., 2019), or ❼ tagging temporal sequences to capture the order of events or actions (Yu et al., 2023).

Despite its wide applications, data annotation poses significant challenges for current machine learning models due to the complexity, subjectivity, and diversity of data. This process requires domain expertise and is resource-intensive, particularly when manually labeling large datasets. Advanced LLMs such as GPT-4 (OpenAI, 2023), Gemini (Team et al., 2023), and LLaMA-2 (Touvron et al., 2023b) offer a promising opportunity to revolutionize data annotation. LLMs serve as more than just tools but play a crucial role in improving the effectiveness and precision of data annotation. Their ability to automate annotation tasks (A, 2022), ensure consistency across large volumes of data (Hou et al., 2023), and adapt through fine-tuning or prompting for specific domains (Song et al., 2023), significantly mitigates the challenges encountered with traditional annotation methods, setting a new standard for what is achievable in the realm of NLP. This survey delves into the nuances of using LLMs for data annotation, exploring methodologies, utilizing strategies, and associated challenges in this transformative approach. Through this exploration, we aim to shed light on the motivations behind embracing LLMs as catalysts for redefining the landscape of data annotation in machine learning and NLP. We explore the utilization of LLMs for data annotation in this survey, making four main contributions:

- **LLM-Based Annotation Generation:** We dive into the process of generating annotations for various data types, including instruction & response, rationale, pairwise feedback, textual feedback, and other domain-specific data. Additionally, we discuss the criteria (*e.g.,* diversity and quality) in the annotation process.

- **Assessing LLM-Generated Annotations:** We explore various methods for assessing the quality of annotations and strategies for selecting high-quality annotations from numerous options.
- **LLM-Generated Annotations Utilization:** We investigate the methodologies at different stages, including supervised fine-tuning, alignment tuning, and inference time, to train machine learning models based on LLM-generated annotations.
- **Social Impact and Future Work:** We discuss issues ranging from ethical dilemmas, such as bias and implications, to technical limitations, including hallucination and efficiency in LLM-generated annotations.

Focusing on this underrepresented aspect of LLM application, the survey aims to serve as a valuable guide for academics and practitioners who intend to deploy LLMs for annotation purposes. Note that in this survey, we primarily focus on pure language models and do not extensively cover recently emerging multimodal LLMs, such as LLaVA (Liu et al., 2023b). Figure 1 illustrates the general structure of this survey. Additionally, a list of potential tools for utilizing LLMs for annotation is included in Appendix A, along with explanatory examples.

Differences from Other LLM-related Surveys. While existing surveys in the NLP domain extensively cover architectural nuances (Zhao et al., 2023a), training methodologies (Liu et al., 2023d), and evaluation protocols (Chang et al., 2023) associated with LLMs, their main focus lies on the capabilities of models for specific end tasks such as machine translation (Min et al., 2021), alignment (Wang et al., 2023g), code generation (Zan et al., 2023), and medical analysis (Thirunavukarasu et al., 2023). In contrast, this survey distinguishes itself by focusing primarily on the application of these potent next-generation LLMs to the intricate realm of data annotation, a domain that is crucial yet underexplored.

2 Preliminaries

In this section, we delve into our approach to the annotation process. We introduce two core models: an annotator model, denoted as $\mathcal{A}$, which maps input data to annotations, and a task learner, represented as $\mathcal{L}$, that utilizes or learns from these annotated data to accomplish specific tasks. Our primary focus is on utilizing advanced LLMs like GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a) as annotators ($\mathcal{A}$), while the task learner ($\mathcal{L}$) can be another large model (Chiang et al., 2023a) or a less complex one such as BERT (Devlin et al., 2018), which utilizes these annotated data to perform designated tasks. LLM-generated annotations encompass categorical labels and enhance raw data points with a comprehensive array of auxiliary signals. These annotations, including confidence scores, contextual details, and other metadata, extend beyond traditional categorical labels.

3 LLM-Based Annotation Generation

The emergence of LLMs has sparked significant interest in their capacity for high-quality, context-sensitive data annotation. This section discusses various kinds of annotations produced via LLMs.

3.1 Instruction & Response

Instruction and response are the two fundamental components that constitute a dataset for LLM fine-tuning and in-context learning (ICL). Previous NLP datasets (Li et al., 2017; Wang et al., 2018; Ouyang et al., 2022) mainly rely on human annotators to construct. Recently, with the advent of LLMs, automatic and generative methods (Meng et al., 2022; Ye et al., 2022a,b; Wang et al., 2024c) have gained more focus in data annotation.

Instruction Diversity. The diversity of instruction has been proven crucial for LLM learning (Li et al., 2023e; Song et al., 2024b,a). Recent studies have explored various methods to diversify and augment instructions in the original datasets. For example, Yoo et al. (2021) enhance data diversity by mixing two different samples to create a new one. Wang et al. (2022b) use a few manually-written seed instructions and iteratively augment them with a generate-then-filter pipeline. Additionally, Meng et al. (2023); Wang et al. (2023f) train an instruction generation model in the original dataset to augment the diversity of instruction. Gupta et al. (2023) employ a multi-step prompting method to first generate task descriptions, which are then used as instance seeds to guide LLMs in instruction generation. To obtain informative and diverse examples, Wang et al. (2023c) propose an explain-then-generate pipeline with LLMs for iterative data synthesis. Besides, Li et al. (2023a) paraphrase the given sample multiple times to help LLMs understand them from different perspectives. Köksal et al. suggest a clustering-based data selection method to ensure diversity in the initial seed data for augmentation. Recently, Yu et al. (2024) introduce AttrPrompt as an effective way to balance

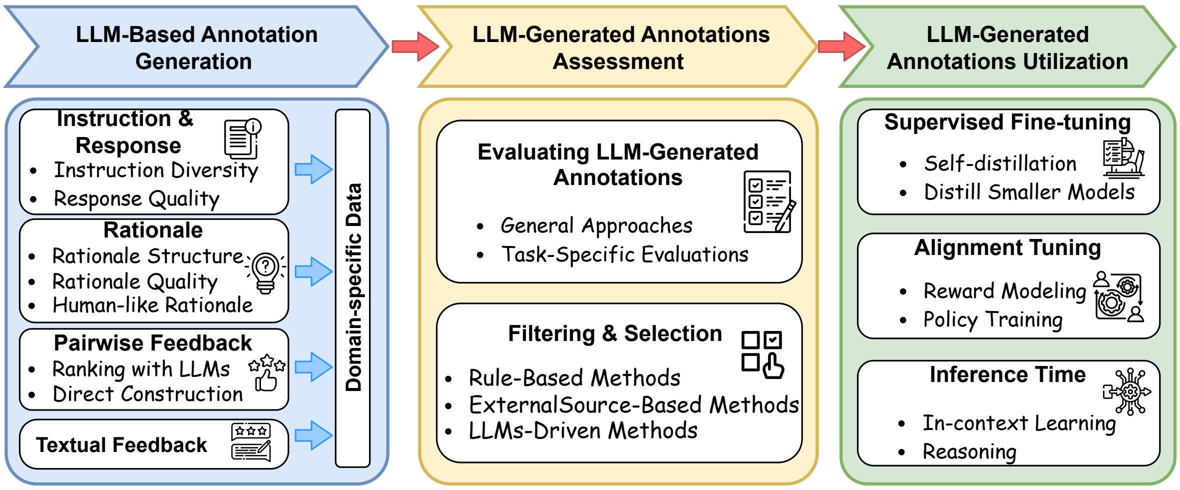

Figure 1: The proposed taxonomy of existing research on LLM for data annotation.

diversity and cost in LLM-based data annotation.

Response Quality. High-quality responses are essential for effective fine-tuning and ICL (Luo et al., 2024). To improve the quality of the generated response, Zhang and Yang (2023a) frame the response generation as reading comprehension tasks and create detailed prompts for LLMs. Huang et al. (2023) adopt self-consistency (Wang et al., 2022b) in response generation, selecting from the candidate response with the highest confidence score. Furthermore, Yang et al. (2024b) propose self-distill and augment the instruction tuning dataset by rewriting the original responses. Pang et al. (2024b) conduct social simulations to ensure high-quality, human-valued responses from LLMs. Moreover, Liu et al. (2024) introduce a multi-step prompting including question analysis, answer guidance and safe answer production in their response generation pipeline. Guo et al. (2024a) enhance the LLMs outputs' quality by implementing retrieval-augmented ICL and providing LLMs with relevant documents. To ensure LLMs provide responses aligned with human values, Sun et al. (2024) and Wang et al. (2024a) conduct principle-driven prompting, guiding LLMs with well-crafted and detailed principles.

3.2 Rationale

The rationale reflects the detailed thought process and reasoning pathway an individual follows when solving a given question, being considered valuable auxiliary information for the final answer prediction. In early studies (Ling et al., 2017; Cobbe et al., 2021; Wei et al., 2022), the rationale in each dataset was annotated by human experts, significantly limiting its availability and scalability. Kojima et al. (2022) initially confirm the efficacy of the chain-of-thought (CoT) approach in LLMs and boosting LLMs' reasoning through the integration of self-generated rationales.

Rationale Structure. Following Kojima et al. (2022), there is a notable interest in abstracting the reasoning process of LLMs into diverse structures and format, including trees (Hao et al., 2023; Yao et al., 2024), graphs (Besta et al., 2024; Yao et al., 2023), tables (Wang et al., 2024d), programs (Chen et al., 2023e), recursion (Qi et al., 2023), and concepts (Tan et al., 2023).

Rationale Quality. To produce high-quality and fine-grained rationale, diverse methodologies have been employed. Wang et al. (2022a) prompt frozen LLMs to produce choice-specific rationales to elucidate each choice in a sample. Wang et al. (2023b) employ contrastive decoding to foster more plausible rationales, taking into account gold-standard answers. Liu et al. (2023a) curate meticulously designed prompts to derive high-quality rationales from GPT-4 and construct a logical CoT instruction tuning dataset. For attaining fine-grained rationales, Shridhar et al. (2023) introduce Socratic CoT by decomposing the original question into a series of subquestion-solution pairs and generating CoT for them separately. Additionally, Kang et al. (2024) propose a neural reranker to acquire supplementary relevant documents for rationale generation in knowledge-intensive reasoning tasks.

Human-like Rationale. Another intriguing avenue in synthesized rationale delves into making the reasoning process more human-like. Many studies emulate human diverse thinking in problem-solving, sampling multiple reasoning pathways for a given question (Gao et al., 2021; Wang et al., 2022b; Chen et al., 2023f; Liu et al., 2023c). Subsequent

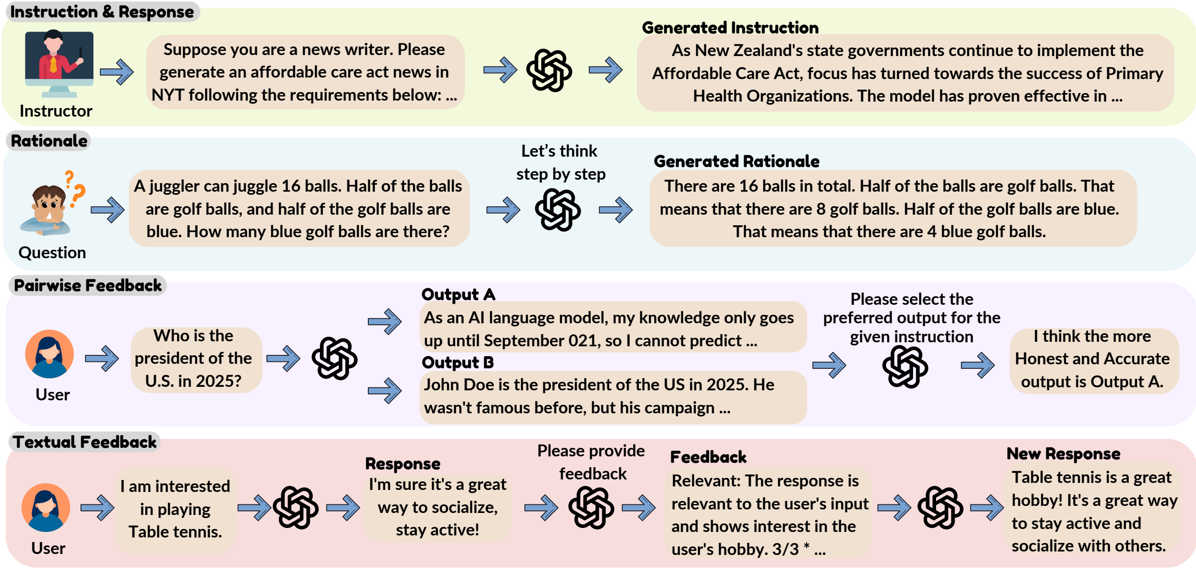

Figure 2: The examples for LLM-based annotation generation.

studies (Tong et al., 2023; Balepur et al., 2023; Ma and Du, 2023) explore the elimination reasoning in LLMs, checking each reasoning pathway reversely and removing the incorrect candidates. Moreover, various works (Yin et al., 2023; Liang et al., 2023; Xu et al., 2023d; Liu et al., 2023e) explore the peer collaboration and debate among individual LLMs to capture human-like discussions as rationales.

3.3 Pairwise Feedback

While high-quality human feedback is proven to be effective in aligning LLMs' values and preferences with us humans, recent advancements aim to automate this pairwise feedback mechanism.

Ranking with LLMs. One technique is to sample multiple responses and have the LLM rank these candidates based on various criteria (Bai et al., 2022; Lee et al., 2023b; Yuan et al., 2024). Sun et al. (2023b) sample two responses from the initial policy model and use the model to select the preferred response based on a human-written principle (Sun et al., 2024). Zhang et al. (2024a) propose a self-evaluation mechanism, generating questions for each response and measuring factuality by the LLM's confidence in the answers. To improve synthetic data quality, Pace et al. (2024) combine the Best-of-N and Worst-of-N sampling strategies and introduce the West-of-N approach. They constructed data pairs by identifying the best- and worst-scored responses according to a pre-trained preference model. In robotics, Zeng et al. (2024) iteratively update the reward function with the self-ranked responses from LLMs, enhancing learning efficiency without human supervision.

Direct Construction. Another effort towards automatic pairwise feedback generation involves directly generating responses of various qualities (Feng et al., 2024; Lee et al., 2024a). To accomplish this, they typically have to make various assumptions when determining the factors influencing response quality. For example, Kim et al. (2023b) assume larger LLM with more shots will give better responses and produce synthetic pairs based on this. Tong et al. (2024b) follow the rule of thumb that the supervised fine-tuning model will perform better than its unfinetuned base model. Adhere to this criterion, they start with a few seed data, iteratively training the model and synthesizing comparison data pairs. Yang et al. (2023c) create quality differences by prompting LLMs to either follow or violate given principles. To measure the response quality more subjectively, Xu et al. (2023c) introduce multiple LLMs and utilize benchmark scores to define superiority.

3.4 Textual Feedback

Textual feedback (Pan et al., 2024) generated by LLMs typically highlights the shortcomings of the current output or suggests specific improvements, thus offering rich and valuable information for polishing or evaluating the generated response. Many existing works tailor appropriate prompts and instruct LLMs to generate such informative feedback in various tasks, including question answering (Madaan et al., 2024; Shinn et al., 2024), machine translation (Chen et al., 2023c; Raunak et al., 2023) and hallucination detection (Yang et al., 2023d; Manakul et al., 2023). Some investigations have explored leveraging debate and peer review as feedback to enhance LLMs' reasoning (Du et al., 2023a; Xu et al., 2023d; Cohen et al., 2023; Fu et al., 2023) and evaluation (Li et al., 2023d; Chu et al., 2024b; Ning et al., 2024) capabilities. Additionally, efforts have been made to analyze reasons for undesired or incorrect responses produced by LLMs, thus facilitating reflection and learning from their previous mistakes (Wang and Li, 2023; An et al., 2023; Chen et al., 2023a; Tong et al., 2024a).

3.5 Other Domain-specific Data

Distilling multi-round conversations from LLMs presents a highly cost-effective approach for constructing high-quality dialogue datasets (Kim et al., 2023a; Xu et al., 2023b; Chen et al., 2023b; Li et al., 2024d) or enhancing existing ones (Zheng et al., 2023; Chen et al., 2022; Zhou et al., 2022a). In graph and tabular data, several studies prompt LLMs to contextualize these structural data (Xiang et al., 2022; Kim et al., 2023a; Li et al., 2024b; Ronzano and Nanavati, 2024) or distill structural insights from raw text (Bi et al., 2024; Li et al., 2024c; Ding et al., 2024; Xiong et al., 2024; Tuozzo, 2022). Moreover, LLMs have also been widely adopted in the research of robotics and agents, serving as proficient data annotators to generate plans (Huang et al., 2022; Brohan et al., 2023; Rana et al., 2023; Singh et al., 2023; Lin et al., 2023a), simulation tasks (Wang et al., 2023a; Ha et al., 2023) and supervised signal (Kwon et al., 2022; Du et al., 2023b). Besides, LLMs are acting as efficient data annotators in various artificial intelligence domains, including multi-modal (Li et al., 2023f; Yin et al., 2024; Chen et al., 2024a), recommendation system (Acharya et al., 2023; Shen et al., 2024; Wei et al., 2024; Zhang et al., 2024b), information extraction (Josifoski et al., 2023; Jeronymo et al., 2023; Li et al., 2024a; Ma et al., 2024; Bonn et al., 2024) and etc (Chu et al., 2024a; Bhattacharjee et al., 2024; Martorana et al., 2024).

4 LLM-Generated Annotations Assessment

Effective evaluation of annotations generated by LLMs is crucial to fully harness their potential. This section focuses on two main aspects:

4.1 Evaluating LLM-Generated Annotations

This subsection explores various methods for assessing annotation quality, ranging from human-led to automated approaches.

General Approaches: Research has investigated diverse methods for evaluating LLM annotations. The "Turking Test" by Efrat and Levy (2020), evaluates LLMs' adherence to data annotation guidelines, with human annotators comparing LLM outputs against benchmarks like SNLI (Bowman et al., 2015), SQuAD (Rajpurkar et al., 2016), and NewsQA (Trischler et al., 2016). Similarly, Honovich et al. (2022) manually examined the originality, accuracy, and variety of datasets created by LLMs, focusing on their response to instructions. Additionally, studies such as by Alizadeh et al. (2023) measure the performance of open-source LLMs against human-annotated labels in tasks like relevance and topic detection.

Task-Specific Evaluations: Methodologies vary by application. For instance, in knowledge graph enhancement, token ranking metrics assess LLM contributions in fact completion. Additionally, evaluations of counterfactual generation often utilize diversity metrics like Self-BLEU (Chen et al., 2023g), while code generation relies on metrics such as Pass@k (Nijkamp et al., 2022). In scenarios requiring extensive datasets, the quality of LLM-generated annotations is compared to gold standard labels within a small, labeled subset (Zhao et al., 2021; Agrawal et al., 2022; He et al., 2023).

4.2 Filtering & Selection

Selecting high-quality annotations from numerous options is crucial. In this section, we categorize the filtering and selection methods for LLM-generated data into three types: rule-based filtering, external source utilization, and LLMs-driven selection.

Rule-Based Methods. Rule-based methods follow various heuristic assumptions concerning sample length (Li et al., 2023f; Kim et al., 2023a), keyword occurrence (Kim et al., 2023b; Zheng et al., 2023) and specific patterns (Zhang and Yang, 2023a; Guo et al., 2024a; Ding et al., 2024) to filter low-quality or undesiered synthetic data points. Zheng et al. (2023); Kim et al. (2023a) establish thresholds for the number of rounds in generated conversations to guarantee each synthetic dialogue is informative enough. Ho et al. (2023); Kang et al. (2024) employ ground truth parsing to filter out incorrect CoT rationales within each candidate reasoning sample. To encourage diversity among the generated data points, Wang et al. (2022b); Lee et al. (2023a); Ding et al. (2024) utilize semantic similarity metrics to identify and remove redundant samples.

External-Source-Based Methods. There are also many works that depend on the external source's feedback to clean and refine synthetic datasets (Kim et al., 2023a). With a pre-trained reward model, Gulcehre et al. (2023); Dong et al. (2023) augment the original dataset only with samples that obtain high reward values. When distilling smaller models, Lin et al. (2023b); Wang et al. (2024c) meticulously select appropriate data through the feedback from the student models. Other approaches (Chen et al., 2023g; Zheng et al., 2023) utilize pre-trained classification models to discern between target and unwanted data points.

LLMs-Driven Methods. The versatility of LLMs has invoked interest in leveraging LLMs themselves to do data selection. Some approaches use signals or features produced by LLMs, such as perplexity score (Wang et al., 2023f), confidence levels (Wang et al., 2022b; Huang et al., 2023), and logits (Pace et al., 2024), as criteria for constructing data selectors. Others directly prompt the LLMs for this task. For instance, Lu et al. (2023) query the target LLM to assess the quality of generated samples. Kim et al. (2023a) leverage ChatGPT to determine if the social commonsense knowledge is appropriately conveyed in the synthetic dialogues. Additionally, there are also works that adopt the LLMs to rank multiple candidate annotations and utilize the top ones in the subsequent stages (Jeronymo et al., 2023; Li et al., 2024c). In pairwise feedback synthesis, Tong et al. (2024b) task the base LLM with judging whether one response genuinely surpasses another.

5 LLM-Generated Annotations Utilization

LLM-generated annotations provide a valuable resource of labeled data for NLP models in different stages. Hereby we explore the methods for utilizing and learning with LLM-Generated Annotations.

5.1 Supervised Fine-tuning

Supervised fine-tuning can effectively enhance models' specific capabilities or knowledge. In this section, we discuss the utilization of generated annotation for supervised fine-tuning.

Self-distillation. Huang et al. (2023) first propose the concept of self-improve that utilizes LLMs as both data annotators and learnable models and iteratively fine-tune LLMs in their self-annotated data. Wang et al. (2023e) also tune a GPT3 in the instruction tuning dataset to improve its zero-shot generalization capability. To foster LLMs' evolution, Lu et al. (2023) iteratively fine-tune the LLMs in self-refined synthetic responses. To mitigate the distribution gap between task datasets and the LLMs, Yang et al. (2024b) use self-distillation which guides fine-tuning with a distilled dataset generated by the model itself. Both Chen et al. (2024b) and Cheng et al. (2024) introduce a self-play mechanism, where the LLM refines its capability by playing against instances of itself.

Distill Smaller Models. For efficiency issues, many studies aim to use the data generated by a large and powerful LLM to train a flexible and affordable smaller model. For a better instruction-following ability, many medium and small-sized LLMs are trained on the synthetic dataset produced by larger LLMs (Taori et al., 2023; Chiang et al., 2023b; Xu et al., 2023a). In classification tasks, Meng et al. (2022, 2023); Wang et al. (2023d) augment the original datasets and train smaller bidirectional attention models on them. To foster models' reasoning ability, many studies tune smaller models with synthetic rationales collected from LLMs (Wang et al., 2022a; Shridhar et al., 2023; Liu et al., 2023a; Kang et al., 2024). Other task-specific capabilities distillation from LLMs include dialogue generation (Xu et al., 2023b), information extraction (Josifoski et al., 2023; Jeronymo et al., 2023) and code generation (Chaudhary, 2023; Roziere et al., 2023). Moreover, LLMs have been proven to follow a scaling law in terms of their knowledge capacity. Therefore, there is also a growing interest in distilling vertical and domain-specific knowledge from LLMs, including medicine (Zhang et al., 2023; Xiong et al., 2023), finance (Zhang and Yang, 2023b) and science (Luo et al., 2023; Zhao et al., 2024), to smaller models.

5.2 Alignment Tuning

Alignment tuning methods, like RLHF (Ouyang et al., 2022), aim to align the output of LLMs with human intentions, ensuring they are helpful, ethical, and reliable. Synthetic data produced by LLMs are widely adopted in these alignment approaches for reward modeling and policy training.

Reward Modeling. LLMs-generated annotations can be used to train or refine the reward model for better alignment. Xu et al. (2023c) propose a data curriculum method that leverages the pairwise feedback from LLMs to calculate the sample difficulty level and smooth LLMs' learning from simple ones to hard ones. Kim et al. (2023b) design reward model guided self-play to iteratively improve the reward model with synthesized data generated by the policy model. Pace et al. (2024)

propose to maximize the probability of correctly labeling a pair of on-policy responses to a given query according to the base preference model. In robotics, Zeng et al. (2024) learns a reward function from scratch using the LLMs' feedback. With synthetic data pair, Sun et al. (2023b) train an instructable reward model to generate reward scores based on arbitrary human-defined principles.

Policy Training. While many direct alignment methods (Rafailov et al., 2024; Zhao et al., 2023b) have emerged recently, some works directly explore the use of annotated feedback for policy training. One common strategy is to directly apply DPO with the synthetic pairwise feedback produced by LLMs (Yuan et al., 2024; Zhang et al., 2024a; Lee et al., 2024b; Tong et al., 2024b; Lee et al., 2024a; Guo et al., 2024b). Besides, Gulcehre et al. (2023); Dong et al. (2023) leverage a pre-trained reward model to filter low-quality synthetic data and iteratively tune LLMs with growing datasets. Wang et al. (2024a) propose a bootstrapping self-alignment method to repeatly utilize the synthetic data. Liu et al. (2024) introduce the Mixture of insighTful Experts (MoTE) architecture, which applies the mixture of experts to enhance each component of the synthetic response, markedly increasing alignment efficiency. With the reasoning pairwise feedback generated by LLM itself, Pang et al. (2024a) use a modified DPO loss with an additional negative log-likelihood term to tune the LLM.

5.3 Inference

In-context Learning. In-context Learning (ICL) consists of three components: a task description (or prompt), several in-context samples (or demonstration), and the test case that needs to be inferred. Current studies have applied the annotations and data generated by LLMs in all these components for refining or augmenting. Zhou et al. (2022b) first showed that with a well-designed pipeline, LLMs can be human-level prompt engineers to generate accurate task descriptions. Following them, Yang et al. (2023b); Li et al. conduct augmentation and expansion to the original task prompt, making it more detailed for LLMs to follow. Demonstration augmentation (Kim et al., 2022; Li et al., 2023c; Chen et al., 2023d; He et al., 2024) is another useful skill to enrich and diversify the provided demonstrations, especially when the labeled data is limited. For the test sample, one augmentation method is to leverage LLMs to rephrase it once (Deng et al., 2023) or multiple times (Li et al., 2023a; Yang et al., 2024a). Other works study how to polish the original test sample (Xi et al., 2023) or decompose it into several sub-questions (Wang et al., 2024b).

Reasoning. Reasoning plays a crucial role in enhancing the quality and accuracy of the content generated by LLMs. One efficient manner to boost LLMs' reasoning with self-generated annotation is to provide the generated rationale directly before outputting the final answer/ response (Kojima et al., 2022). To improve LLMs' performance with multiple reasoning pathways, majority voting(Wang et al., 2022b; Chen et al., 2023f) and elimination(Tong et al., 2023; Balepur et al., 2023; Ma and Du, 2023) are adopted to decide the final answer among several possible candidates. Post-hoc editing and refining (Madaan et al., 2024; Tong et al., 2024a) is another well-studied direction to utilize textual feedback and analysis for improving LLMs' reasoning capabilities. Additionally, utilization of LLMs-generated annotations sometimes requires additional domain tools. For example, Chen et al. (2023e) use a program interpreter in program-of-thought (PoT) to execute the generated program and convert it to a specific answer. Besta et al. (2024) design a prompter to Build a prompt to be sent to the LLM and a parser to extract information from LLM thought. In tree-of-thought (ToT), Hao et al. (2023); Yao et al. (2024) build an additional state evaluator by designing specific prompts and repurposing the base LLM.

6 Societal Impact and Future Work

In this section, we outline LLM annotation challenges, including societal implications, technical concerns, and bias propagation.

6.1 Ethics Consideration

One critical concern of LLM-generated annotations is the ethics consideration, especially in high-stakes decision-making tasks like finance (Yang et al., 2023a), jurisprudence (Cui et al., 2023), and healthcare (Eloundou et al., 2023). Despite the efficiency of LLM annotation, the lack of human insight may lead to biased and unfair results (Wu et al., 2023; Abid et al., 2021; Cheng et al., 2021; Li et al., 2023g). Moreover, LLMs make human annotator roles redundant, potentially increasing social disparities (Dillion et al., 2023). Future studies should harmonize technological advancements with societal consequences, including considering social implications, ensuring ethical use, promoting fairness, and maintaining transparency.

6.2 Challenges and Future Work

Model Collapse. Model collapse refers to the gradual performance decrease of an LLM trained on the outputs of other LLMs (Sun et al., 2023a; Gunasekar et al., 2023; Hsieh et al., 2023; Honovich et al., 2022; Chiang et al., 2023a; Geng et al., 2023). It is unavoidable since LLM-generated data is occupying the information ecosystem. The imitation model often replicates stylistic elements without achieving the factual precision of superior models (Gudibande et al., 2023; Shumailov et al., 2023). This divergence is caused by *statistical approximation error* from limited sample sizes and *functional approximation error* from constrained model capacity. Both errors tend to amplify through successive training cycles (Alemohammad et al., 2023).

Potential Solution. It is important to ensure that the training data is diverse and high-quality, with a significant proportion of human-generated content. Gerstgrasser et al. (2024) avoid model collapse by accumulating real and machine-generated data. This method maintains data diversity, preventing performance degradation across different LLMs.

Hallucinations. Hallucinations in LLMs significantly undermine the integrity and reliability of their generated annotations (Alkaissi and McFarlane, 2023; Azamfirei et al., 2023; Chaudhary et al., 2024). Hullicinated outputs detached from factual information can cause the proliferation of misinformation (Jiang et al., 2024; Chen and Shu, 2023). Addressing hallucinations requires refining the training process and implementing validation mechanisms for annotations through automated and manual verification (Liao and Vaughan, 2023; Pan et al., 2023; Bian et al., 2023). Moreover, the inherent opacity of LLMs complicates efforts to investigate the causes of hallucinations.

Potential Solution. Yang et al. (2023d) addresses hallucinations in LLMs with the Reverse Validation method, detecting hallucinations at the passage level by constructing a query from the response and checking for a match within the LLM's internal knowledge. Bertaglia et al. (2023) uses Chain-of-Thought (CoT) prompting and explanation generation, where CoT prompting produces explanations for predictions, ensuring logical and verifiable outputs. Li et al. (2023b) proposes the CoAnnotating framework, which uses uncertainty-guided work allocation between humans and LLMs, applying self-evaluation and entropy metrics to assess reliability and distribute tasks effectively.

Efficiency of LLMs. Efficiency in LLMs is crucial due to their growing size and complexity, which demand substantial computational resources (Wong et al., 2024). Efficient models reduce inference latency, vital for real-time applications, lower energy consumption for sustainable AI practices, and cut operational costs in cloud environments, making AI more cost-effective for researchers. Efficiency techniques for LLMs, such as pruning, compression, and distillation, are critical for deploying these models in resource-constrained environments.

Potential Solution. Pruning is an efficient technique to reduce the number of parameters in an LLM. For example, Ma et al. (2023) selectively removes redundant neurons based on gradient information while preserving most of the LLM's capability. Mixture of Experts (MoE) is another promising technique that leverages a set of expert sub-models, where only a subset of these experts is activated for any given input (Artetxe et al., 2021). Researchers also adopt LLM Quantization to reduce the precision of the numbers used to represent a model's parameters (Xiao et al., 2023). Instead of using 32-bit floating-point numbers, a quantized model might use 16-bit floats, 8-bit integers, or even lower precision. These techniques can be combined with each other to achieve further efficiencies.

7 Conclusion

The exploration of LLMs for data annotation has revealed an exciting frontier in NLP, presenting novel solutions to longstanding challenges like data scarcity, and enhancing annotation quality and process efficiency. This survey meticulously reviews methodologies, applications, and hurdles associated with LLM employment, including detailed taxonomy from annotation generation to utilization. It evaluates the effects of LLM-generated annotations on training machine learning models while addressing both technical and ethical concerns like bias and societal ramifications. Highlighting our novel taxonomy of LLM methodologies, strategies for utilizing LLM-generated annotations, and a critical discussion on the challenges, this work aims to steer future progress in this crucial area. Additionally, we introduce a comprehensive categorization of techniques and compile extensive benchmark datasets to support ongoing research endeavors, concluding with an examination of persistent challenges and open questions, paving the way for future investigative pursuits in the domain.

Limitations

Sampling Bias and Hallucination. LLMs can display sampling bias, leading to incorrect or "hallucinated" data, impacting the reliability and quality of annotations for discriminative tasks.

Social Bias and Ethical Dilemmas. The inherent biases in training data can be perpetuated and amplified by LLMs, leading to ethical concerns and the propagation of social biases through annotated data. This is particularly problematic in tasks requiring fairness and impartiality.

Dependence on High-Quality Data. LLMs' usefulness in generating annotations depends on large, high-quality datasets. But curating these datasets is labor-intensive, posing a scalability challenge for LLM-based annotation efforts.

Complexity in Tuning and Prompt Engineering. Successfully leveraging LLMs for data annotation requires sophisticated prompt engineering and fine-tuning techniques. This can pose a barrier to entry for practitioners and researchers without extensive expertise in NLP and machine learning.

Generalization and Overfitting While LLMs can be powerful tools for annotation, there's a risk of overfitting to the training data, limiting their ability to generalize to unseen data or different contexts. This is a critical limitation for discriminative tasks where the goal is to develop models that perform well across diverse datasets and domains.

Computational and Resource Requirements. The training and deployment of state-of-the-art LLMs for data annotation require substantial computational resources, which may not be accessible to all researchers and organizations, thereby limiting widespread adoption.

References

Sujan Reddy A. 2022. Automating human evaluation of dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 229–234, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Bernardo Aceituno and Antoni Rosinol. 2022. Stack ai: The middle-layer of ai.

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Reza Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. 2023. Self-consuming generative models go mad. *ArXiv*, abs/2307.01850.

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*.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Walid Amamou. 2021. Ubiai: Text annotation tool.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. 2021. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*.

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2023. It's not easy being wrong: Evaluating process of elimination reasoning in large language models. *arXiv preprint arXiv:2311.07532*.

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. In *World Conference on Explainable Artificial Intelligence*, pages 198–213. Springer.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Zero-shot llm-guided counterfactual generation for text. *arXiv preprint arXiv:2405.04793*.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–16.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink may make a million think: The spread of false information in large language models. *arXiv preprint arXiv:2305.04812*.

Julia Bonn, Harish Tayyar Madabushi, Jena D Hwang, and Claire Bonial. 2024. Adjudicating llms as propbank annotators. *LREC-COLING 2024*, page 112.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Manav Chaudhary, Harshit Gupta, and Vasudeva Varma. 2024. Brainstorm@ irel at smm4h 2024: Leveraging translation and topical embeddings for annotation detection in tweets. *arXiv preprint arXiv:2405.11192*.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. *Code alpaca: An instruction-following llama model for code generation*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. 2023a. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023b. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023c. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023d. Self-icl: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2023e. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023f. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Yunkai Chen, Qimeng Wang, Shiwei Wu, Yan Gao, Tong Xu, and Yao Hu. 2024a. Tomgpt: Reliable text-only training approach for cost-effective multimodal large language model. *ACM Transactions on Knowledge Discovery from Data*.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023g. 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.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. 2024. Self-playing adversarial language game enhances llm reasoning. *arXiv preprint arXiv:2404.10642*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023a. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023b. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Zhixuan Chu, Yan Wang, Longfei Li, Zhibo Wang, Zhan Qin, and Kui Ren. 2024a. A causal explainable guardrails for large language models. *arXiv preprint arXiv:2405.04160*.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024b. Pre: A peer review based large language model evaluator. *arXiv preprint arXiv:2401.15641*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Linyi Ding, Sizhe Zhou, Jinfeng Xiao, and Jiawei Han. 2024. Automated construction of theme-specific knowledge graphs. *arXiv preprint arXiv:2404.19146*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023a. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023b. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*, pages 8657–8677. PMLR.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Yunlong Feng, Yang Xu, Libo Qin, Yasheng Wang, and Wanxiang Che. 2024. Improving language model reasoning with self-motivated learning. *arXiv preprint arXiv:2404.07017*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *BAIR Blog*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

Arnav Gudibande, Eric Wallace, Charles Burton Snell, Xinyang Geng, Hao Liu, P. Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *ArXiv*, abs/2305.15717.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero C. Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. Textbooks are all you need. *ArXiv*, abs/2306.11644.

Hongyi Guo, Yuanshun Yao, Wei Shen, Jiaheng Wei, Xiaoying Zhang, Zhaoran Wang, and Yang Liu. 2024a. Human-instruction-free llm self-alignment with limited samples. *arXiv preprint arXiv:2401.06785*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024b. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*.

Huy Ha, Pete Florence, and Shuran Song. 2023. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Chase Harrison. 2022. Langchain.

Wei He, Shichun Liu, Jun Zhao, Yiwen Ding, Yi Lu, Zhiheng Xi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Self-demos: Eliciting out-of-demonstration generalizability in large language models. *arXiv preprint arXiv:2404.00884*.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.

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*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 427–435. SIAM.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, et al. 2023a. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 12930–12949.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023b. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. Longform: Effective instruction tuning with reverse instructions. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2022. Reward design with language models. In *The Eleventh International Conference on Learning Representations*.

Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic role labeling with pretrained language models for known and unknown predicates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 619–628.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. 2023a. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15349–15360.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023b. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Kyungjae Lee, Dasol Hwang, Sunghyun Park, Youngsoo Jang, and Moontae Lee. 2024a. Reinforcement learning from reflective feedback (rlrf): Aligning and improving llms via fine-grained self-reflection. *arXiv preprint arXiv:2403.14238*.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024b. Aligning large language models by on-policy self-judgment. *arXiv preprint arXiv:2402.11253*.

Dawei Li, William Hogan, and Jingbo Shang. 2024a. Read: Improving relation extraction from an adversarial perspective. *arXiv preprint arXiv:2404.02931*.

Dawei Li, Yaxuan Li, Dheeraj Mekala, Shuyao Li, Xueqi Wang, William Hogan, Jingbo Shang, et al. 2023a. Dail: Data augmentation for in-context learning via self-paraphrase. *arXiv preprint arXiv:2311.03319*.

Dawei Li, Zhen Tan, Tianlong Chen, and Huan Liu. 2024b. Contextualization distillation from large language model for knowledge graph completion. *arXiv preprint arXiv:2402.01729*.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sunkwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, et al. 2024c. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024d. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023b. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Rui Li, Guoyin Wang, and Jiwei Li. 2023c. Are human-generated demonstrations necessary for in-context learning? *arXiv preprint arXiv:2309.14681*.

Ruosen Li, Teerth Patel, and Xinya Du. 2023d. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2023e. Self-alignment with instruction back-translation. In *The Twelfth International Conference on Learning Representations*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023f. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. Empowering large language models for textual data augmentation.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023g. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023a. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023b. 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.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023a. Logicot: Logical chain-of-thought instruction tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023c. Plan, verify and switch: Integrated reasoning with diverse x-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023d. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. 2024. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023e. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Man Luo, Christopher J Warren, Lu Cheng, Haidar M Abdul-Muhsin, and Imon Banerjee. 2024. Assessing empathy in large language models with real-world physician-patient interactions. *arXiv preprint arXiv:2405.16402*.

Chenkai Ma and Xinya Du. 2023. Poe: Process of elimination for multiple choice reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4487–4496.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. Star: Boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18751–18759.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Margherita Martorana, Tobias Kuhn, Lise Stork, and Jacco van Ossenbruggen. 2024. Text classification of column headers with a controlled vocabulary: leveraging llms for metadata enrichment. *arXiv preprint arXiv:2403.00884*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

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.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Peer-review-in-llms: Automatic evaluation method for llms in open-environment. *arXiv preprint arXiv:2402.01830*.

OpenAI. 2023. Gpt-4 technical report.

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

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024a. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024b. Self-alignment of large language models via monopolylogue-based social scene simulation. *arXiv preprint arXiv:2402.05699*.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of socratic questioning: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *7th Annual Conference on Robot Learning*.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.

Francesco Ronzano and Jay Nanavati. 2024. Towards ontology-enhanced representation learning for large language models. *arXiv preprint arXiv:2405.20527*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 3833–3843.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

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*, abs/2305.17493.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024a. Scaling data diversity for fine-tuning language models in human alignment. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 14358–14369.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024b. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023a. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023b. Salmon: Self-alignment with instructable reward models.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Zhen Tan, Lu Cheng, Song Wang, Yuan Bo, Jundong Li, and Huan Liu. 2023. Interpreting pretrained language models via concept bottlenecks. *arXiv preprint arXiv:2311.05014*.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, pages 1–11.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024a. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*.

Yongqi Tong, Sizhe Wang, Dawei Li, Yifan Wang, Simeng Han, Zi Lin, Chengsong Huang, Jiaxin Huang, and Jingbo Shang. 2024b. Optimizing language model's reasoning abilities with weak supervision. *arXiv preprint arXiv:2405.04086*.

Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. 2023. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking. *arXiv preprint arXiv:2310.12342*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Gabriele Tuozzo. 2022. Moving from tabular knowledge graph quality assessment to rdf triples leveraging chatgpt.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Danqing Wang and Lei Li. 2023. Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10667–10685.

Haoyu Wang, Guozheng Ma, Ziqiao Meng, Zeyu Qin, Li Shen, Zhong Zhang, Bingzhe Wu, Liu Liu, Yatao Bian, Tingyang Xu, et al. 2024a. Step-on-feet tuning: Scaling self-alignment of llms via bootstrapping. *arXiv preprint arXiv:2402.07610*.

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam-fai Wong. 2024b. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *arXiv preprint arXiv:2402.13514*.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2023a. Gensim: Generating robotic simulation tasks via large language models. In *The Twelfth International Conference on Learning Representations*.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

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*.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023c. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11817–11831.

Song Wang, Zhen Tan, Ruocheng Guo, and Jundong Li. 2023d. Noise-robust fine-tuning of pretrained language models via external guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12528–12540.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023e. 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.

Yue Wang, Haoke Zhang, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2023f. Sass: Self-alignment with semi-supervised instruction data generation.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023g. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024c. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. 2024d. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 806–815.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Siu Ming Wong, Ho Leung, and Ka Yan Wong. 2024. Efficiency in language understanding and generation: An evaluation of four open-source large language models.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. Self-polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406.

Jiannan Xiang, Zhengzhong Liu, Yucheng Zhou, Eric Xing, and Zhiting Hu. 2022. Asdot: Any-shot data-to-text generation with pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1886–1899.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin

Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278.

Canwen Xu, Corby Rosset, Luciano Del Corro, Shweti Mahajan, Julian McAuley, Jennifer Neville, Ahmed Hassan Awadallah, and Nikhil Rao. 2023c. Contrastive post-training large language models on data curriculum. *arXiv preprint arXiv:2310.02263*.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023d. Towards reasoning in large language models via multiagent peer review collaboration. *arXiv preprint arXiv:2311.08152*.

Adam Yang, Chen Chen, and Konstantinos Pitas. 2024a. Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. *arXiv preprint arXiv:2405.13907*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Jinghan Yang, Shuming Ma, and Furu Wei. 2023b. Auto-icl: In-context learning without human supervision. *arXiv preprint arXiv:2311.09263*.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023c. Rlcd: Reinforcement learning from contrastive distillation for lm alignment. In *The Twelfth International Conference on Learning Representations*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023d. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908.

Zhaorui Yang, Qian Liu, Tianyu Pang, Han Wang, Haozhe Feng, Minfeng Zhu, and Wei Chen. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *arXiv preprint arXiv:2305.16582*.

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

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

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.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. Large language models meet nl2code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7443–7464.

Yuwei Zeng, Yao Mu, and Lin Shao. 2024. Learning reward for robot skills using large language models via self-alignment. *arXiv preprint arXiv:2405.07162*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024a. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.

Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024b. Large language models as evaluators for recommendation explanations. *arXiv preprint arXiv:2406.03248*.

Xuanyu Zhang and Qing Yang. 2023a. Self-qa: Unsupervised knowledge guided language model alignment. *arXiv preprint arXiv:2305.11952*.

Xuanyu Zhang and Qing Yang. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439.

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2024. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, 36.

Mengjie Zhao, Fei Mi, Yasheng Wang, Minglei Li, Xin Jiang, Qun Liu, and Hinrich Schütze. 2021. Lmturk: Few-shot learners as crowdsourcing workers in a language-model-as-a-service framework. *arXiv preprint arXiv:2112.07522*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023b. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

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

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022a. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022b. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A LLM-assisted Tools and Software for Annotation

LLM-assisted annotation tools and software are invaluable resources designed specifically to facilitate the annotation process for various NLP tasks. One of their primary attributes is an intuitive and user-friendly interface, allowing engineers and even non-technical annotators to easily work with complex textual data. These tools are built to support numerous annotation types, from simple binary labels to more intricate hierarchical structures. The main goal of these tools is to simplify the labeling process, enhance the quality of the labels, and boost overall productivity in data annotation.

Below, we will present a selection of the libraries and tools that support Large Language Models for the annotation process:

- **LangChain**: LangChain (Harrison, 2022) is an open-source library[1] that offers an array of tools designed to facilitate the construction of LLM-related pipelines and workflows. This library specifically provides large language models with agents in order to interact effectively with their environment as well as various external data sources. Therefore, providing dynamic and contextually appropriate responses that go beyond a single LLM call.

  In terms of the annotation process, their power mostly lies in the facilitation of annotation through the creation of a modularized structure called *chain*. In the chaining technique, a complex problem is broken down into smaller sub-tasks. The results obtained from one or more steps are then aggregated and utilized as input prompts for subsequent actions in the chain.

[1] As of now, available only in JavaScript/TypeScript and Python languages.

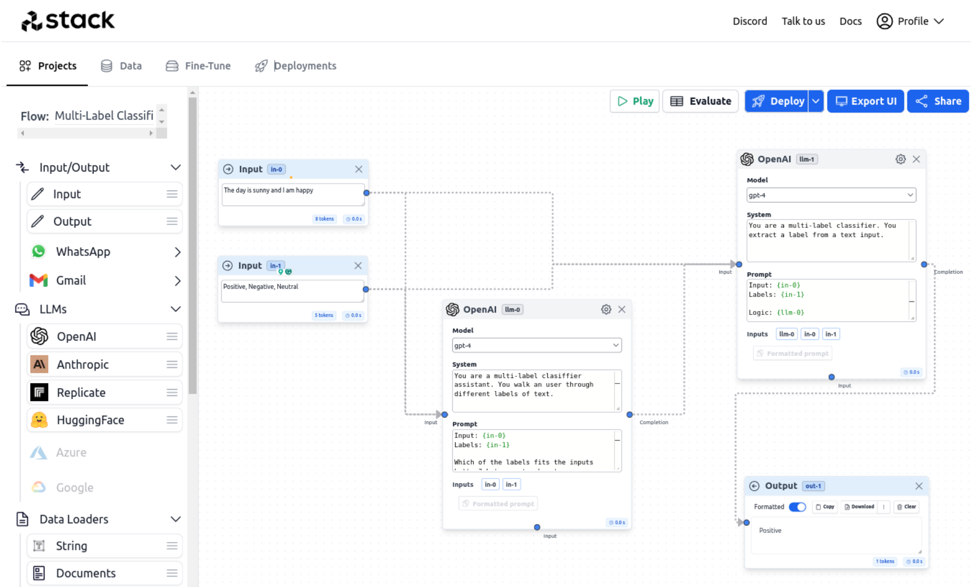

Figure 3: Stack AI dashboard. They provide a visual interface for users to design and track the AI workflow.

- **Stack AI**: Stack AI (Aceituno and Rosinol, 2022) is a paid service that offers an AI-powered data platform. It is designed explicitly for automating business processes allowing them to maximize efficiency. The essence of their platform lies in their ability to *visually* design, test, and deploy AI workflows through smooth integration of Large Language Models. Their user-friendly graphical interface (Figure 3) allows the users to create apps and workflows related to diverse tasks from content creation and data labeling to conversational AI apps and document processing. Moreover, Stack AI utilizes weakly supervised machine learning models to expedite the data preparation process.

- **UBIAI**: UBIAI (Amamou, 2021) is a paid annotation tool that offers multilingual cloud-based solutions and services in Natural Language Processing. The company aims to aid users in extracting valuable insights from unstructured documents. This tool not only provides a user interface that facilitates manual labeling but also offers several auto-labeling functionalities such as LLM-assisted zero- and few-shot labeling and model-assisted labeling. They also provide integration to various models on huggingface (Wolf et al., 2020) as well as an environment to fine-tune different models on the user's labeled data.

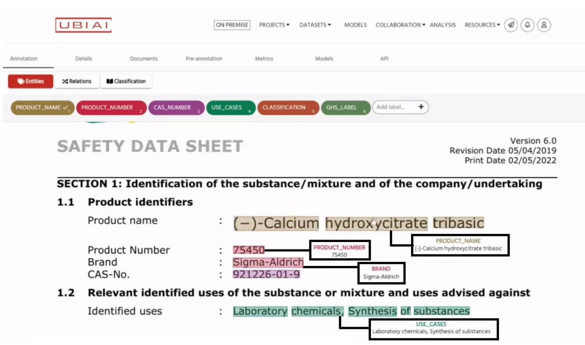

Figure 4: UBIAI annotation result on a pdf document. All the entities in the text of the document have been identified, annotated, and color-coded based on the type. This image has been borrowed from the videos provided in the UBIAI documentation (Amamou, 2021).

- **Prodigy**: Prodigy (Montani and Honnibal, 2018), designed by the creators of spaCy library (Honnibal and Montani, 2017), offers rule-based, statistical models, and LLM-assisted methods for annotation. This tool provides easy, flexible, and powerful annotation options such as named entity recognition, span categorization, and classification/labeling for different modalities including text, audio, and vision. Moreover, it can be easily integrated with large language models which are capable of zero- or few-shot learning, while also offering services and quantifiable methods for crafting prompts to address any noisy outcomes. This tool is not open-source.

B Acknowledgment of AI Assistance in Writing and Revision

We utilized ChatGPT-4 for revising and enhancing sections of this paper.

C Collections of Papers on LLM for Data Annotation

This collection of tables provides a concise overview of using Large Language Models (LLMs) for data annotation, including state-of-the-art techniques, methodologies, and practical applications. Table 1 and Table 2 lists significant papers on LLM-based data annotation, detailing their methods, core technologies, publication venues, and links to resources. Table 3 focuses on assessment and filtering of LLM-generated annotations. Tables 4 explore strategies for learning with LLM-generated annotations, covering supervised fine-tuning, alignment tuning and inference. Each table clearly outlines the data type, backbone, computational cost, venues, and available resources, serving as a guide to the latest in LLM-driven data annotation and its implications for the future of automated data processing and machine learning research.

Paper	Data Type	Backbone	Annotation Cost	Venue	Code/Data Link
Instruction & Response					
GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation[1]	Instruction	GPT-3	API Calling, 300 tokens per sample	EMNLP'21	Link
SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions[2]	Instruction & Response	GPT-3	API Calling, $600 for entire dataset	ACL'23	Link
Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning[3]	Instruction	CTRL	Model Training, Nvidia A100 GPUs, 10 minutes per task	ICML'23	Link
SASS: SELF-ALIGNMENT WITH SEMI-SUPERVISED INSTRUCTION DATA GENERATION[4]	Instruction	LLaMA	Model Training, Nvidia A100 GPUs	OpenRview'24	Not Available
DAIL: Data Augmentation for In-Context Learning via Self-Paraphrase[5]	Instruction	ChatGPT	API Calling	Arxiv'23	Not Available
LongForm: Effective Instruction Tuning with Reverse Instructions[6]	Instruction	GPT-3	PI Calling	ICLR'24	Link
Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias[7]	Instruction	ChatGPT	API Calling	NeurIPS'23	Link
SELF-QA: Unsupervised Knowledge Guided Language Model Alignment[8]	Instruction & Response	BLOOM	Model Inference	Arxiv'23	Not Available
LARGE LANGUAGE MODELS CAN SELF-IMPROVE[9]	Response	PaLM-540B	Model Inference	EMNLP'23	Not Available
Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning[10]	Response	LLaMA-2	Model Inference	ACL'24	Link
Mixture of insighTful Experts (MoTE): The Synergy of Thought Chains and Expert Mixtures in Self-Alignment[11]	Response	Alpaca	Model Inference	Arxiv'24	Not Available
Human-Instruction-Free LLM Self-Alignment with Limited Samples[12]	Instruction & Response	Multiple LLMs	Model Inference, single NVIDIA A100 80G GPU	Arxiv'24	Not Available
Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision[13]	Response	LLaMA	Model Inference	NeurIPS'23	Link
Step-On-Feet Tuning: Scaling Self-Alignment of LLMs via Bootstrapping[14]	Response	LLaMA-2	Model Inference	Arxiv'24	Not Available
Assessing Empathy in Large Language Models with Real-World Physician-Patient Interactions[15]	Response	LLaMA	Model Inference	Arxiv'24	Not Available
Rationale					
Large Language Models are Zero-Shot Reasoners[16]	Rationale - CoT	Multiple LLMs	API Calling	NeurIPS'22	Not Available
Tree of Thoughts: Deliberate Problem Solving with Large Language Models[17]	Rationale - Tree	GPT-4	API Calling, $0.74 per sample	NeurIPS'22	Link
Reasoning with Language Model is Planning with World Model[18]	Rationale - Tree	LLaMA	Model Inference, 4×24 GB NVIDIA A5000 GPUs	EMNLP'23	Link
Graph of Thoughts: Solving Elaborate Problems with Large Language Models[19]	Rationale - Graph	GPT-3.5	API Calling	AAAI'24	Link
Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Language Models[20]	Rationale - Graph	GPT-3	API Calling	Arxiv'23	Link
CHAIN-OF-TABLE: EVOLVING TABLES IN THE REASONING CHAIN FOR TABLE UNDERSTANDING[21]	Rationale - Table	Multiple LLMs	API Calling & Model Inference	ICLR'24	Not Available
Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks [22]	Rationale - Program	Multiple LLMs	API Calling & Model Inference	TMLR'23	Not Available
The Art of SOCRATIC QUESTIONING: Recursive Thinking with Large Language Models[23]	Rationale - Reversion	ChatGPT	API Calling, 9.22 calls per sample	EMNLP'23	Link
Interpreting Pretrained Language Models via Concept Bottlenecks[24]	Rationale - Concept	ChatGPT	API Calling	PAKDD'24	Link
PINTO: FAITHFUL LANGUAGE REASONING USING PROMPT-GENERATED RATIONALES[25]	Rationale - CoT	GPT-neox	Model Inference	ICLR'23	Link
SCOTT: Self-Consistent Chain-of-Thought Distillation[26]	Rationale - CoT	GPT-neox	Model Inference	ACL'23	Link
LogiCoT: Logical Chain-of-Thought Instruction Tuning[27]	Rationale - CoT	GPT-4	API Calling	EMNLP'23	Not Available
Distilling Reasoning Capabilities into Smaller Language Models[28]	Rationale - CoT	GPT-3	API Calling	ACL'23	Not Available
Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks [29]	Rationale - CoT	ChatGPT	API Calling	NeurIPS'23	Link
Making Pre-trained Language Models Better Few-shot Learners[30]	Rationale - Diverse Thinking	GPT-3	API Calling	ACL'21	Link
SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS[31]	Rationale - Diverse Thinking	Multiple LLMs	API Calling & Model Inference	ICLR'23	Not Available
UNIVERSAL SELF-CONSISTENCY FOR LARGE LANGUAGE MODEL GENERATION[32]	Rationale - Diverse Thinking	Multiple LLMs	API Calling	Arxiv'23	Not Available
Plan, Verify and Switch: Integrated Reasoning with Diverse X-of-Thoughts[33]	Rationale - Diverse Thinking	ChatGPT	API Calling	EMNLP'23	Link
Eliminating Reasoning via Inferring with Planning: A New Framework to Guide LLMs' Non-linear Thinking[34]	Rationale - Elimination	PaLM2	API Calling	Arxiv'23	Not Available
It's Not Easy Being Wrong: Large Language Models Struggle with Process of Elimination Reasoning[35]	Rationale - Elimination	Multiple LLMs	API Calling	ACL'24	Link
POE: Process of Elimination for Multiple Choice Reasoning[36]	Rationale - Elimination	FLAN-T5	Model Inference	EMNLP'23	Link
Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication [37]	Rationale - Collaboration	ChatGPT	API Calling	EMNLP'23	Not Available
Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate[38]	Rationale - Collaboration	ChatGPT	API Calling	Arxiv'23	Link
Towards Reasoning in Large Language Models via Multi-Agent Peer Review Collaboration[39]	Rationale - Collaboration	ChatGPT	API Calling	Arxiv'23	Link
DYNAMIC LLM-AGENT NETWORK: AN LLM-AGENT COLLABORATION FRAMEWORK WITH AGENT TEAM OPTIMIZATION[40]	Rationale - Collaboration	ChatGPT	API Calling, 16.5 calls per sample	Arxiv'23	Link
Pair-wise Feedback					
Constitutional AI: Harmlessness from AI Feedback[41]	Pairwise Feedback	Multiple LLMs	Model Inference	Arxiv'22	Link
RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback[42]	Pairwise Feedback	PaLM-2	Model Inference, $0.67 per sample	Arxiv'23	Not Available
Self-Rewarding Language Models[43]	Pairwise Feedback	LLaMA-2	Model Inference	Arxiv'24	Not Available
SALMON: SELF-ALIGNMENT WITH INSTRUCTABLE REWARD MODELS[44]	Pairwise Feedback	LLaMA-2	Model Inference	ICLR'24	Link
Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation[45]	Pairwise Feedback	LLaMA	Model Inference	Arxiv'24	Link
West-of-N: Synthetic Preference Generation for Improved Reward Modeling[46]	Pairwise Feedback	T5-XXL	Model Inference	Arxiv'24	Not Available
Learning Reward for Robot Skills Using Large Language Models via Self-Alignment[47]	Pairwise Feedback	ChatGPT	API Calling	ICML'24	Link
Aligning Large Language Models through Synthetic Feedback[48]	Pairwise Feedback	LLaMA	Model Inference	EMNLP'23	Link
Optimizing Language Model's Reasoning Abilities with Weak Supervision[49]	Pairwise Feedback	LLaMA	Model Inference	Arxiv'24	Not Available
RLCD: REINFORCEMENT LEARNING FROM CONTRASTIVE DISTILLATION FOR LM ALIGNMENT[50]	Pairwise Feedback	LLaMA	Model Inference	ICLR'24	Link
Automatic Pair Construction for Contrastive Post-training[51]	Pairwise Feedback	LLaMA	Model Inference, 16 Nvidia V100 GPUs	NAACL'24	Not Available
Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection[52]	Pairwise Feedback	LLaMA-2	Model Inference, 16 Nvidia V100 GPUs	Arxiv'24	Not Available
Improving Language Model Reasoning with Self-motivated Learning[53]	Pairwise Feedback	LLaMA-2	Model Inference	LREC'24	Not Available

Note: [1](Yoo et al., 2021); [2](Wang et al., 2023e); [3](Meng et al., 2023); [4](Wang et al., 2023f); [5](Li et al., 2023a); [6](Köksal et al.); [7](Yu et al., 2024); [8](Zhang and Yang, 2023a); [9](Huang et al., 2023); [10](Yang et al., 2024b); [11](Liu et al., 2024); [12](Guo et al., 2024a); [13](Sun et al., 2024); [14](Wang et al., 2024a); [15](Luo et al., 2024); [16](Kojima et al., 2022); [17](Yao et al., 2024); [18](Hao et al., 2023); [19](Besta et al., 2024); [20](Yao et al., 2023); [21](Wang et al., 2024d); [22](Chen et al., 2023e); [23](Qi et al., 2023); [24](Tan et al., 2023); [25](Wang et al., 2022a); [26](Wang et al., 2023b); [27](Liu et al., 2023a); [28](Shridhar et al., 2023); [29](Kang et al., 2024); [30](Gao et al., 2021); [31](Wang et al., 2022b); [32](Chen et al., 2023f); [33](Liu et al., 2023c); [34](Tong et al., 2023); [35](Balepur et al., 2023); [36](Ma and Du, 2023); [37](Yin et al., 2023); [38](Liang et al., 2023); [39](Xu et al., 2023d); [401](Liu et al., 2023e); [41](Bai et al., 2022); [42](Lee et al., 2023b); [43](Yuan et al., 2024); [44](Sun et al., 2023b); [45](Zhang et al., 2024a); [46](Pace et al., 2024); [47](Zeng et al., 2024); [48](Kim et al., 2023b); [49](Tong et al., 2024b); [50](Yang et al., 2023c); [51](Xu et al., 2023c); [52](Lee et al., 2024a); [53](Feng et al., 2024).

Table 1: A list of representative LLM-Based Annotation Generation (Instruction & Response, Rationale, Pairwise Feedback) papers with open-source code/data.

Paper	Data Type	Backbone	Annotation Cost	Venue	Code/Data Link
Textual Feedback					
SELF-REFINE: Iterative Refinement with Self-Feedback[1]	Textual Feedback	Multiple LLMs	API Calling	NeurIPS'23	Not Available
Reflexion: Language Agents with Verbal Reinforcement Learning[2]	Textual Feedback	GPT-3	API Calling	NeurIPS'23	Link
Iterative Translation Refinement with Large Language Models[3]	Textual Feedback	GPT-3.5	API Calling	Arxiv'23	Not Available
Leveraging GPT-4 for Automatic Translation Post-Editing[4]	Textual Feedback	Multiple LLMs	API Calling	EMNLP'23	Not Available
A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection[5]	Textual Feedback	ChatGPT	API Calling	EMNLP'23	Link
SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models[6]	Textual Feedback	Multiple LLMs	API Calling & Model Inference	EMNLP'23	Link
Improving Factuality and Reasoning in Language Models through Multiagent Debate[7]	Textual Feedback - Peer Review	Multiple LLMs	API Calling		Link
Towards Reasoning in Large Language Models via Multi-Agent Peer Review Collaboration[8]	Textual Feedback - Peer Review	Multiple LLMs	API Calling	Arxiv'23	Link
LM vs LM: Detecting Factual Errors via Cross Examination[9]	Textual Feedback - Peer Review	Multiple LLMs	API Calling & Model Inference	EMNLP'23	Not Available
Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback[10]	Textual Feedback - Peer Review	Multiple LLMs	API Calling	Arxiv'23	Link
PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations[11]	Textual Feedback - Peer Review	Multiple LLMs	API Calling, $0.14 per sample	Arxiv'23	Link
PRE: A Peer Review Based Large Language Model Evaluator[12]	Textual Feedback - Peer Review	Multiple LLMs	API Calling	Arxiv'24	Not Available
PiCO: Peer Review in LLMs based on the Consistency Optimization[13]	Textual Feedback - Peer Review	Multiple LLMs	API Calling & Model Inference	Arxiv'24	Not Available
Learning from Mistakes via Cooperative Study Assistant for Large Language Models[14]	Textual Feedback - Mistake	Multiple LLMs	Model Inference	EMNLP'23	Link
Learning From Mistakes Makes LLM Better Reasoner[15]	Textual Feedback - Mistake	GPT-4	API Calling	Arxiv'23	Link
GAINING WISDOM FROM SETBACKS: ALIGNING LARGE LANGUAGE MODELS VIA MISTAKE ANALYSIS[16]	Textual Feedback - Mistake	Multiple LLMs	API Calling & Modeling Inference	ICLR'24	Not Available
Can LLMs Learn from Previous Mistakes? Investigating LLMs' Errors to Boost for Reasoning[17]	Textual Feedback - Mistake	Multiple LLMs	API Calling & Modeling Inference	ACL'24	Link
Other Domain-specific Data					
SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization[18]	Dialogue	GPT-3.5	API Calling, $0.02 per dialogue	EMNLP'23	Link
Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data[19]	Dialogue	Alpaca	Model Inference	EMNLP'23	Link
PLACES: Prompting Language Models for Social Conversation Synthesis[20]	Dialogue	Multiple LLMs	Model Inference	EACL'24	Not Available
CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society[21]	Dialogue	ChatGPT	API Calling	NuerIPS'23	Link
AUGESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation[22]	Dialogue	GPT-J	Model Inference	ACL'23	Link
Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding[23]	Dialogue	GPT-J	Model Inference	NeurIPS'22	Not Available
Reflect, Not Reflex: Inference-Based Common Ground Improves Dialogue Response Quality[24]	Dialogue	GPT-3	API Calling	EMNLP'22	Link
ASDOT: Any-Shot Data-to-Text Generation with Pretrained Language Models[25]	Context	GPT-3	API Calling, $23 in total	EMNLP'22	Link
Contextualization Distillation from Large Language Model for Knowledge Graph Completion[26]	Context	PaLM-2	API Calling	EACL'24	Link
Towards Ontology-Enhanced Representation Learning for Large Language Models[27]	Context	ChatGPT	API Calling	Arxiv'24	Link
DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature[28]	Graph	ChatGPT	API Calling	Arxiv'24	Link
Automated Construction of Theme-specific Knowledge Graphs[29]	Graph	GPT-4	API Calling	Arxiv'24	Not Available
Large Language Models Can Learn Temporal Reasoning[30]	Graph	GPT-3.5	API Calling	ACL'24	Link
Moving from Tabular Knowledge Graph Quality Assessment to RDF Triples Leveraging ChatGPT[31]	Graph	ChatGPT	API Calling	Arxiv'24	Link
Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents[32]	Plan	GPT-3	API Calling	ICML'22	Link
Do As I Can, Not As I Say: Grounding Language in Robotic Affordances[33]	Plan	Multiple LLMs	API Calling & Model Inference	CoRL'21	Link
SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning[34]	Plan	GPT-3.5	API Calling	CoRL'23	Link
PROGPROMPT: Generating Situated Robot Task Plans using Large Language Models[35]	Plan	GPT-3	API Calling	ICRA'23	Link
Text2Motion: From Natural Language Instructions to Feasible Plans[36]	Plan	GPT-3.5	API Calling	Autonomous Robots'23	Link
GENSIM: GENERATING ROBOTIC SIMULATION TASKS VIA LARGE LANGUAGE MODELS[37]	Simulation Task	GPT-4	API Calling	ICLR'24	Link
Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition[38]	Simulation Task	Multiple LLMs	API Calling	CoRL'23	Link
REWARD DESIGN WITH LANGUAGE MODELS[39]	Reward	GPT-3	API Calling	ICLR'23	Link
Guiding Pretraining in Reinforcement Learning with Large Language Models[40]	Reward	GPT-3	API Calling, 0.02 second per call	ICML'23	Not Available
Enhanced Visual Instruction Tuning with Synthesized Image-Dialogue Data[41]	Visual Instruction Tuning Data	ChatGPT	API Calling	Arxiv'23	Link
LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark[42]	Visual Instruction Tuning Data	GPT-4	API Calling	NeurIPS'23	Link
TOMGPT: Reliable Text-Only Training Approach for Cost-Efective Multi-modal Large Language Model[43]	Context	ChatGPT	API Calling	TKDD'24	Not Available
LLM Based Generation of Item-Description for Recommendation System[44]	Item Description	Alpaca	Model Inference	RecSys'23	Not Available
PMG : Personalized Multimodal Generation with Large Language[45]	Context	Multiple LLMs	Model Inference	WWW'24	Link
LLMRec: Large Language Models with Graph Augmentation for Recommendation[46]	Augmented Implicit Feedback	ChatGPT	API Calling, $21.14	WSDM'24	Link
Large Language Models as Evaluators for Recommendation Explanations[47]	Explanation	Multiple LLMs	API Calling & Model Inference, less than $0.02 per sample	Arxiv'24	Link
Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction[48]	IE Sample	GPT-3.5	API Calling, $223.55 for entire dataset	EMNLP'23	Link
InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval[49]	IE sample	GPT-J	Model Inference, 30 hours on an A100 GPU to generate 100k queries	Arxiv'23	Link
READ: Improving Relation Extraction from an ADversarial Perspective[50]	IE Sample	ChatGPT	API Calling	NAACL'24	Link
STAR: Boosting Low-Resource Information Extraction by Structure-to-Text Data Generation with Large Language Models[51]	IE Sample	Multiple LLMs	API Calling	AAAI'24	Link
Adjudicating LLMs as PropBank Annotators[52]	IE Label	Multiple LLMs	API Calling	LREC'24	Link
A Causal Explainable Guardrails for Large Language Models[53]	Representation	GPT-4	API Calling	Arxiv'24	Not Available
Zero-shot LLM-guided Counterfactual Generation for Text[54]	Context	Multiple LLMs	API Calling	Arxiv'24	Not Available
Text classification of column headers with a controlled vocabulary: leveraging LLMs for metadata enrichment[55]	Metadata	ChatGPT	API Calling	Arxiv'24	Link

Note: [1](Madaan et al., 2024); [2](Shinn et al., 2024); [3](Chen et al., 2023c); [4](Raunak et al., 2023); [5](Yang et al., 2023d); [6](Manakul et al., 2023); [7](Du et al., 2023a); [8](Xu et al., 2023d); [9](Cohen et al., 2023); [10](Fu et al., 2023); [11](Li et al., 2023d); [12](Chu et al., 2024b); [13](Ning et al., 2024); [14](Wang and Li, 2023); [15](An et al., 2023); [16](Chen et al., 2023a); [17](Tong et al., 2024a); [18](Kim et al., 2023a); [19](Xu et al., 2023b); [20](Chen et al., 2023b); [21](Li et al., 2024d); [22](Zheng et al., 2023); [23](Chen et al., 2022); [24](Zhou et al., 2022a); [25](Xiang et al., 2022); [26](Li et al., 2024b); [27](Ronzano and Nanavati, 2024); [28](Li et al., 2024c); [29](Ding et al., 2024); [30](Xiong et al., 2024); [31](Tuozzo, 2022); [32](Huang et al., 2022); [33](Brohan et al., 2023); [34](Rana et al., 2023); [35](Singh et al., 2023); [36](Lin et al., 2023a); [37](Wang et al., 2023a); [38](Ha et al., 2023); [39](Kwon et al., 2022); [40](Du et al., 2023b); [41](Li et al., 2023f); [42](Yin et al., 2024); [43](Chen et al., 2024a); [44](Acharya et al., 2023); [45](Shen et al., 2024); [46](Wei et al., 2024); [47](Zhang et al., 2024b); [48](Josifoski et al., 2023); [49](Jeronymo et al., 2023); [50](Li et al., 2024a); [51](Ma et al., 2024); [52](Bonn et al., 2024); [53](Chu et al., 2024a); [54](Bhattacharjee et al., 2024); [55](Martorana et al., 2024).

Table 2: A list of representative LLM-Based Annotation Generation (Textual Feedback, Other Domain-specific Data) papers with open-source code/data.

Paper	Data Type	Backbone	Annotation Cost	Venue	Code/Data Link
Filter & Selection					
Constitutional AI: Harmlessness from AI Feedback[1]	Pairwise Feedback	Multiple LLMs	Model Inference	Arxiv'22	Link
SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization[2]	Dialogue	GPT-3.5	API Calling, $0.02 per dialogue	EMNLP'23	Link
Aligning Large Language Models through Synthetic Feedback[3]	Pairwise Feedback	LLaMA	Model Inference	EMNLP'23	Link
AUGESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation[4]	Dialogue	GPT-J	Model Inference	ACL'23	Link
SELF-QA: Unsupervised Knowledge Guided Language Model Alignment[5]	Instruction & Response	BLOOM	Model Inference	Arxiv'23	Not Available
Human-Instruction-Free LLM Self-Alignment with Limited Samples[6]	Instruction & Response	Multiple LLMs	Model Inference, single NVIDIA A100 80G GPU	Arxiv'24	Not Available
Automated Construction of Theme-specific Knowledge Graphs[7]	Graph	GPT-4	API Calling	Arxiv'24	Not Available
Large Language Models Are Reasoning Teachers[8]	CoT	GPT-3.5	API Calling	ACL'23	Link
Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks [9]	Rationale - CoT	ChatGPT	API Calling	NeurIPS'23	Link
SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS[10]	Rationale - Diverse Thinking	Multiple LLMs	API Calling & Model Inference	ICLR'23	Not Available
Making Large Language Models Better Data Creators[11]	Instruction & Response	ChatGPT	API Calling	EMNLP'23	Link
Automated Construction of Theme-specific Knowledge Graphs[12]	Graph	GPT-4	API Calling	Arxiv'24	Not Available
Reinforced Self-Training (ReST) for Language Modeling[13]	Response	Multiple LLMs	Model Inference	Arxiv'24	Not Available
RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment[14]	Response	LLaMA	Model Inference	TMLR	Link
Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information[15]	Instruction	OPT	Model Inference	EACL'24	Not Available
CodecLM: Aligning Language Models with Tailored Synthetic Data[16]	Instruction	LLaMA	Model Inference	NAACL'24	Not Available
DISCO: Distilling Counterfactuals with Large Language Models[17]	CoT	GPT-3	API Callin	ACL'23	Link
LARGE LANGUAGE MODELS CAN SELF-IMPROVE[18]	Response	PaLM-540B	Model Inference	EMNLP'23	Not Available
West-of-N: Synthetic Preference Generation for Improved Reward Modeling[19]	Pairwise Feedback	T5-XXL	Model Inference	Arxiv'24	Not Available
SELF: SELF-EVOLUTION WITH LANGUAGE FEEDBACK[20]	Response	Multiple LLMs	Model Inference	Arxiv'23	Not Available
InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval[21]	IE sample	GPT-J	Model Inference, 30 hours on an A100 GPU to generate 100k queries	Arxiv'23	Link
DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature[22]	Graph	ChatGPT	API Calling	Arxiv'24	Link
Optimizing Language Model's Reasoning Abilities with Weak Supervision[23]	Pairwise Feedback	LLaMA	Model Inference	Arxiv'24	Not Available

Note: [1](Bai et al., 2022); [2](Kim et al., 2023a); [3](Kim et al., 2023b); [4](Zheng et al., 2023); [5](Zhang and Yang, 2023a); [6](Guo et al., 2024a); [7](Ding et al., 2024); [8](Ho et al., 2023); [9](Kang et al., 2024); [10](Wang et al., 2022b); [11](Lee et al., 2023a); [12](Ding et al., 2024); [13](Gulcehre et al., 2023); [14](Dong et al., 2023); [15](Lin et al., 2023b); [16](Wang et al., 2024c); [17](Chen et al., 2023g); [18](Huang et al., 2023); [19](Pace et al., 2024); [20](Lu et al., 2023); [21](Jeronymo et al., 2023); [22](Li et al., 2024c); [23](Tong et al., 2024b).

Table 3: A list of representative LLM-Generated Annotation Assessment papers with open-source code/data.

Paper	Data Type	Backbone	Annotation Cost	Venue	Code/Data Link
Supervised Fine-tuning					
LARGE LANGUAGE MODELS CAN SELF-IMPROVE[1]	Response	PaLM-540B	Model Inference	EMNLP'23	Not Available
SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions[2]	Instruction & Response	GPT-3	API Calling, $600 for entire dataset	ACL'23	Link
SELF: SELF-EVOLUTION WITH LANGUAGE FEEDBACK[3]	Response	Multiple LLMs	Model Inference	Arxiv'23	Not Available
Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning[4]	Response	LLaMA-2	Model Inference	ACL'24	Link
Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models[5]	Response	zephyr	Model Inference	Arxiv'24	Link
Self-playing Adversarial Language Game Enhances LLM Reasoning[6]	Response	Multiple LLMs	Model Inference	Arxiv'24	Link
Stanford alpaca: An instruction-following llama model[7]	Response	GPT-3.5	API Calling	Arxiv'23	Link
Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality[8]	Response	GPT-4	API Calling	Arxiv'23	Link
Wizardlm: Empowering large language models to follow complex instructions[9]	Instruction	LLaMA	Model Inference	Arxiv'23	Link
Generating training data with language models: Towards zero-shot language understanding[10]	Instruction	CTRL	Model Inference	NeurIPS	Link
Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning[11]	Instruction	CTRL	Model Training	ICML'23	Link
Noise-Robust Fine-Tuning of Pretrained Language Models via External Guidance[12]	Response	ChatGPT	API Calling	EMNLP'23	Link
PINTO: FAITHFUL LANGUAGE REASONING USING PROMPT-GENERATED RATIONALES[13]	Rationale - CoT	GPT-neox	Model Inference	ICLR'23	Link
Distilling Reasoning Capabilities into Smaller Language Models[14]	Rationale - CoT	GPT-3	API Calling	ACL'23	Not Available
LogiCoT: Logical Chain-of-Thought Instruction Tuning[15]	Rationale - CoT	GPT-4	API Calling	EMNLP'23	Not Available
Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks[16]	Rationale - CoT	ChatGPT	API Calling	NeurIPS'23	Link
Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data[17]	Dialogue	Alpaca	Model Inference	EMNLP'23	Link
Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction[18]	IE Sample	GPT-3.5	API Calling, $223.55 for entire dataset	EMNLP'23	Link
InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval[19]	IE sample	GPT-J	Model Inference, 30 hours on an A100 GPU to generate 100k queries	Arxiv'23	Link
Code alpaca: An instruction-following llama model for code generation[20]	Instruction & Response	Alpaca	Model Inferece	Arxiv'23	Link
Code llama: Open foundation models for code[21]	Instruction & Response	Multiple LLMs	Model Inference	Arxiv'23	Link
HuatuoGPT, Towards Taming Language Model to Be a Doctor[22]	Instruction & Response	ChatGPT	API Calling	Arxiv'23	Link
Doctorglm: Fine-tuning your chinese doctor is not a herculean task[23]	Response	ChatGPT	API Calling	Arxiv'23	Link
Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters[24]	Instruction & Response	BLOOM	Model Inference	CIKM'23	Not Available
Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct[25]	Pairwise Feedback	ChatGPT	API Calling	Arxiv'23	Link
Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning[26]	Instruction	ChatGPT	API Calling	NuerIPS'23	Link
Alignment Tuning					
Automatic Pair Construction for Contrastive Post-training[27]	Pairwise Feedback	LLaMA	Model Inference, 16 Nvidia V100 GPUs	NAACL'24	Not Available
Aligning Large Language Models through Synthetic Feedback[28]	Pairwise Feedback	LLaMA	Model Inference	EMNLP'23	Link
West-of-N: Synthetic Preference Generation for Improved Reward Modeling[29]	Pairwise Feedback	T5-XXL	Model Inference	Arxiv'24	Not Available
Learning Reward for Robot Skills Using Large Language Models via Self-Alignment[30]	Pairwise Feedback	ChatGPT	API Calling	ICML'24	Link
SALMON: SELF-ALIGNMENT WITH INSTRUCTABLE REWARD MODELS[31]	Pairwise Feedback	LLaMA-2	Model Inference	ICLR'24	Link
Self-Rewarding Language Models[32]	Pairwise Feedback	LLaMA-2	Model Inference	Arxiv'24	Not Available
Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation[33]	Pairwise Feedback	LLaMA	Model Inference	Arxiv'24	Link
Aligning Large Language Models by On-Policy Self-Judgment[34]	Response	LLaMA-2	Model Inference	Arxiv'24	Link
Optimizing Language Model's Reasoning Abilities with Weak Supervision[35]	Pairwise Feedback	LLaMA	Model Inference	Arxiv'24	Not Available
Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection[36]	Pairwise Feedback	LLaMA-2	Model Inference, 16 Nvidia V100 GPUs	Arxiv'24	Not Available
Direct language model alignment from online ai feedback[37]	Pairwise Feedback	PaLM-2	API Calling	Arxiv'24	Not Available
Reinforced Self-Training (ReST) for Language Modeling[38]	Response	Multiple LLMs	Model Inference	Arxiv'24	Not Available
RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment[39]	Response	LLaMA	Model Inference	TMLR	Link
Step-On-Feet Tuning: Scaling Self-Alignment of LLMs via Bootstrapping[40]	Response	LLaMA-2	Model Inference	Arxiv'24	Not Available
Mixture of insighTful Experts (MoTE): The Synergy of Thought Chains and Expert Mixtures in Self-Alignment[41]	Response	Alpaca	Model Inference	Arxiv'24	Not Available
Iterative reasoning preference optimization[42]	Pairwise Feedback	LLaMA-2	Model Inference	Arxiv'24	Not Available
Inference Time					
Large Language Models are Human-Level Prompt Engineers[43]	Instruction	GPT-3.5	API Calling	ICLR'23	Link
Auto-ICL: In-Context Learning without Human Supervision[44]	Instruction	ChatGPT	API Calling	Arxiv'23	Link
Empowering Large Language Models for Textual Data Augmentation[45]	Instruction	ChatGPT	API Calling	Arxiv'24	Not Available
Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator[46]	Instruction	GPT-J	Model Inference	NAACL'22	Link
Are Human-generated Demonstrations Necessary for In-context Learning?[47]	Instruction	Multiple LLMs	API Calling	Arxiv'23	Link
Self-ICL: Zero-Shot In-Context Learning with Self-Generated Demonstrations[48]	Instruction	Multiple LLMs	API Calling	EMNLP'23	Link
Self-Demos: Eliciting Out-of-Demonstration Generalizability in Large Language Models[49]	Instruction	ChatGPT	API Calling	NAACL'24	Link
Rephrase and respond: Let large language models ask better questions for themselves[50]	Instruction	GPT-4	API Calling	Ariv'23	Link
DAIL: Data Augmentation for In-Context Learning via Self-Paraphrase[51]	Instruction	ChatGPT	API Calling	Arxiv'23	Not Available
Just rephrase it! Uncertainty estimation in closed-source language models via multiple rephrased queries[52]	Instruction	Multiple LLMs	Model Inference	Arxiv'24	Not Available
Self-Polish: Enhance Reasoning in Large Language Models via Problem Refinement[53]	Instruction	GPT-3.5	API Calling	EMNLP'23	Link
Self-DC: When to retrieve and When to generate? Self Divide-and-Conquer for Compositional Unknown Questions[54]	Instruction	ChatGPT	API Calling	Arxiv'24	Not Available
Large Language Models are Zero-Shot Reasoners[55]	Rationale - CoT	Multiple LLMs	API Callinfg	NeurIPS'22	Not Available
SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS[56]	Rationale - Diverse Thinking	Multiple LLMs	API Calling & Model Inference	ICLR'23	Not Available
UNIVERSAL SELF-CONSISTENCY FOR LARGE LANGUAGE MODEL GENERATION[57]	Rationale - Diverse Thinking	Multiple LLMs	API Calling	Arxiv'23	Not Available
Eliminating Reasoning via Inferring with Planning: A New Framework to Guide LLMs' Non-linear Thinking[58]	Rationale - Elimination	PaLM2	API Calling	Arxiv'23	Not Available
It's Not Easy Being Wrong: Large Language Models Struggle with Process of Elimination Reasoning[59]	Rationale - Elimination	Multiple LLMs	API Calling	ACL'24	Link
POE: Process of Elimination for Multiple Choice Reasoning[60]	Rationale - Elimination	FLAN-T5	Model Inference	EMNLP'23	Link
SELF-REFINE: Iterative Refinement with Self-Feedback[61]	Textual Feedback	Multiple LLMs	API Calling	NeurIPS'23	Not Available
Can LLMs Learn from Previous Mistakes? Investigating LLMs' Errors to Boost for Reasoning[62]	Textual Feedback - Mistake	Multiple LLMs	API Calling & Modeling Inference	ACL'24	Link
Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks [63]	Rationale - Program	Multiple LLMs	API Calling & Model Inference	TMLR'23	Not Available
Graph of Thoughts: Solving Elaborate Problems with Large Language Models[64]	Rationale - Graph	GPT-3.5	API Calling	AAAI'24	Link
Reasoning with Language Model is Planning with World Model[65]	Rationale - Tree	LLaMA	Model Inference, 4×24 GB NVIDIA A5000 GPUs	EMNLP'23	Link

Note: [1](Huang et al., 2023); [2](Wang et al., 2023e); [3](Lu et al., 2023); [4](Yang et al., 2024b); [5](Chen et al., 2024b); [6](Cheng et al., 2024); [7](Taori et al., 2023); [8](Chiang et al., 2023a); [9](Xu et al., 2023a); [10](Meng et al., 2022); [11](Meng et al., 2023); [12](Wang et al., 2023d); [13](Wang et al., 2022a); [14](Shridhar et al., 2023); [15](Liu et al., 2023a); [16](Kang et al., 2024); [17](Xu et al., 2023b); [18](Josifoski et al., 2023); [19](Jeronymo et al., 2023); [20](Chaudhary, 2023); [21](Roziere et al., 2023); [22](Zhang et al., 2023); [23](Xiong et al., 2023); [24](Zhang and Yang, 2023b); [25](Luo et al., 2023); [26](Zhao et al., 2024); [27](Xu et al., 2023c); [28](Kim et al., 2023b); [29](Pace et al., 2024); [30](Zeng et al., 2024); [31](Sun et al., 2023b); [32](Yuan et al., 2024); [33](Zhang et al., 2024a); [34](Lee et al., 2024b); [35](Tong et al., 2024b); [36](Lee et al., 2024a); [37](Guo et al., 2024b); [38](Gulcehre et al., 2023); [39](Dong et al., 2023); [40](Wang et al., 2024a); [41](Liu et al., 2024); [42](Chen et al., 2023c); [43](Zhou et al., 2022b); [44](Yang et al., 2023b); [45](Li et al.); [46](Kim et al., 2022); [47](Li et al., 2023c); [48](Chen et al., 2023d); [49](He et al., 2024); [50](Deng et al., 2023); [51](Li et al., 2023a); [52](Yang et al., 2024a); [53](Xi et al., 2023); [54](Wang et al., 2024b); [55](Kojima et al., 2022); [56](Wang et al., 2022b); [57](Chen et al., 2023f); [58](Tong et al., 2023); [59](Balepur et al., 2023); [60](Ma and Du, 2023); [61](Madaan et al., 2024); [62](Tong et al., 2024a); [63](Chen et al., 2023e); [64](Besta et al., 2024); [65](Hao et al., 2023).

Table 4: A list of representative LLM-Generated Annotation Utilization papers with open-source code/data.