Best Practices and Lessons Learned on Synthetic Data

Ruibo Liu[*], **Jerry Wei**[†], **Fangyu Liu**
Google DeepMind, Anthropic[†]

Chenglei Si, **Yanzhe Zhang**
Stanford University, Georgia Institute of Technology

Jinmeng Rao, **Steven Zheng**, **Daiyi Peng**
Google DeepMind

Diyi Yang
Stanford University

Denny Zhou, **Andrew M. Dai**[*]
Google DeepMind

Abstract

The success of AI models relies on the availability of large, diverse, and high-quality datasets, which can be challenging to obtain due to data scarcity, privacy concerns, and high costs. Synthetic data has emerged as a promising solution by generating artificial data that mimics real-world patterns. This paper provides an overview of synthetic data research, discussing its applications, challenges, and future directions. We present empirical evidence from prior art to demonstrate its effectiveness and highlight the importance of ensuring its factuality, fidelity, and unbiasedness. We emphasize the need for responsible use of synthetic data to build more powerful, inclusive, and trustworthy language models.

1 Introduction

The rapid advancement of artificial intelligence (AI) technologies has led to their widespread adoption across numerous domains, from assistant agents (e.g., ACT-1, from Adept AI[1]) and software development (e.g., Devin, from Cognition Lab[2]) to healthcare (Singhal et al., 2022) and finance (Zheng et al., 2022). However, the success of AI models heavily relies on the availability of large, diverse, and high-quality datasets for training and evaluation. Acquiring such datasets can be a significant challenge due to data scarcity (Babbar & Schölkopf, 2019), privacy concerns (Abay et al., 2019), and the sheer cost of data collection and annotation (Gilardi et al., 2023a). Pessimists predict that we will run out of fresh text data in 2050 and image data in 2060 (Villalobos et al., 2022).

Synthetic data has emerged as a promising solution to address these challenges (Nikolenko, 2021). Synthetic data refers to artificially generated data that mimics the characteristics and patterns of real-world data, but is created through algorithms (Saxton et al., 2019), generative models (Borisov et al., 2022; Meng et al., 2022), or even simulations (Vezhnevets et al., 2023; Liu et al., 2023c), rather than being directly created by humans. By leveraging synthetic data, we can not only overcome the limitations of real-world data but also unlock

[*] Corresponding author(s): *ruiboliu@google.com*, *adai@google.com*

[†] Work done at Google DeepMind

[1] ACT-1: `https://www.adept.ai/blog/act-1`

[2] Devin: `https://www.cognition-labs.com/introducing-devin`

the potential to develop more robust, reliable, and fair AI models (Lucini, 2021; Lu et al., 2023).

One of the many benefits of synthetic data is that it can be generated at scale, providing an abundant supply of training and testing data for AI models. This is particularly valuable in domains where real-world data is scarce or difficult to obtain (e.g., weather data covering all conditions (Li et al., 2023a; Lam et al., 2023)). Second, synthetic data can be tailored to specific requirements, such as ensuring a balanced representation of different classes by introducing controlled variations (e.g., up-weighting low-resource languages in multilingual language learning (Przystupa & Abdul-Mageed, 2019)). This level of control over data characteristics can improve model performance and generalization. Third, synthetic data can help mitigate privacy concerns by creating anonymized or de-identified datasets that do not contain sensitive personal information (Howe et al., 2017; El Emam et al., 2020). This is crucial in domains such as healthcare, where patient privacy is of utmost importance (Dahmen & Cook, 2019; Wei et al., 2019).

Despite its promise, synthetic data also presents challenges that need to be addressed. One of them is ensuring the factuality and fidelity of synthetic data (Wood et al., 2021; Heusel et al., 2017), as models trained on false, hallucinated or biased synthetic data may fail to generalize to real-world scenarios (Van Breugel et al., 2023; Guarnera et al., 2020). Researchers must develop sophisticated generative models and evaluation metrics to create synthetic data that accurately reflects the complex patterns and relationships found in real-world data. Another challenge is the potential for synthetic data to amplify biases or introduce new biases if not carefully designed and validated (Barbierato et al., 2022; Gupta et al., 2021). We believe rigorous testing and fairness assessments are necessary to mitigate these risks.

In this paper, we track the current state of synthetic data research and discuss current best practices and lessons learned. The rest of the paper is organized as follows. Section 2 provides an overview of synthetic data generation techniques and their applications in model training, presenting case studies and empirical evidence. Section 3 discusses the usefulness of synthetic data in evaluation. Section 4 discusses the challenges and limitations of synthetic data, and in Section 5 we outline potential solutions and future research directions.

2 Synthetic Data in Training

Synthetic data, which is generated by mimicking authentic data collected from the real world, has proven to be an effective and relatively low-cost alternative of real data. This section explores several notable domains that leverages synthetic training data.

2.1 Reasoning

Math. Recent advancements in mathematical reasoning for language models (LMs) have led to the development of various approaches to improve performance on math-related tasks. One approach is to train on math-targeted pre-training data, such as Minerva (Lewkowycz et al., 2022), Llemma (Azerbayev et al., 2023), and DeepSeekMath (Shao et al., 2024). Another mainstream method is to generate synthetic questions and answers to imitate the training or validation set of target benchmarks. For instance, WizardMath (Luo et al., 2023a) leverages a series of operations to increase the complexity of questions and answers using GPT-3.5, while MetaMath (Yu et al., 2023) bootstraps the questions in MATH and GSM8K by rewriting them in different ways, such as semantic rephrasing, self-verification, and backward reasoning. GAIR-Abel (Chern et al., 2023) found that the format of the augmented answers is crucial to final performance, with answers that begin with a paraphrasing of the question followed by a step-by-step solution showing better performance than those in vanilla format. Xwin-Math (Li et al., 2024) further scaled up synthetic SFT data to one million examples and found that the LLaMA-2 7B model (Touvron et al., 2023) can still benefit from data scaling. MMIQC (Liu & Yao, 2024) composed a bundle of datasets that infuse SFT style data (via question-answer rephrasing or directly taken from MetaMath) with a subset of high-quality mathematical pre-training data, such as OpenWebMath (Paster et al., 2023).

Scaling up the generation of synthetic math data is a straightforward process, but ensuring the correctness of the generated math remains a significant challenge for practitioners. AlphaGeometry (Trinh et al., 2024) is a recent attempt to address this issue by training a neural model using 100 million synthetic data points. The model proposes solutions and guides a symbolic deduction engine in verifying the correctness of each branch when solving complex geometry problems. By combining the power of synthetic data with a rigorous verification process, AlphaGeometry achieves a problem-solving ability comparable to that of a human Olympiad gold medalist, demonstrating the potential of this approach in tackling complex mathematical reasoning tasks.

Code. Different from Math, synthetic data for code reasoning can naturally combine the execution results with structured code, as one requirement of correct code is being executable. In coding-enhanced models, CodeRL (Le et al., 2022) presents an actor-critic approach to improve pretrained language models with feedback signals on synthetic code samples. Haluptzok et al. (2022) propose a self-improvement strategy where the models generate their own synthetic puzzle-solution pairs. These pairs are then verified and filtered by a real interpreter before being used to finetune language models. Shypula et al. (2023) further propose a framework that leverages a simulated environment and adaptation strategies like self-improvement synthetic data generation and CoT prompting for code optimization. Yang et al. (2024) developed InterCode, a framework designed to enhance interactive code generation within a reinforcement learning environment, where code serves as actions and execution feedback serves as observations. Reflexion (Shinn et al., 2024) employs external or internally simulated linguistic feedback signals to improve the code reasoning capabilities of language models. Regarding synthetic SFT data, Code Alpaca comprises a dataset of 20K code instructions automatically generated by applying SELF-INSTRUCT (Wang et al., 2022a) to ChatGPT across 21 seed tasks. WizardCoder (Luo et al., 2023b) introduces Code Evol-Instruct to guide ChatGPT with heuristic prompts to enhance the complexity and diversity of synthetic data. Meanwhile, Magicoder (Wei et al., 2023c) developed OSS-INSTRUCT, which generates 75K diverse synthetic instruction samples from open-source code snippets.

Other reasoning tasks. Synthetic data also leads to impressive performance in other reasoning tasks. For instance, Wei et al. (2023a) augmented existing natural language datasets by replacing natural language labels with arbitrary symbols, generating over 500k synthetic examples. Using these synthetic data for supervised finetuning significantly improved model performance on unseen in-context learning and algorithmic-reasoning tasks. STaR (Zelikman et al., 2022) generates synthetic chain-of-thought rationales and filters out those leading to wrong answers for finetuning language models to improve their reasoning. In the domain of physics reasoning, Mind's Eye (Liu et al., 2022) takes a novel approach by training a text-to-code model with synthetic "text-description → rendering code" data. This enables the model to convert textual questions into rendering code, which is then executed in a physical engine (i.e., DeepMind MuJoCo (Todorov et al., 2012)). The rendering results are injected into the context, allowing even small language models armed with Mind's Eye to achieve performance comparable to models 100 times larger.

2.2 Tool-using and Planning

Learning tool-using through synthetic trajectories. Synthetic data is also a powerful approach to enable LMs to learn tool-using abilities through simulated trajectories, as collecting real-world human tool-using data might be time-consuming, and the actual distribution of calls to tools might be skewed. LaMDA (Thoppilan et al., 2022), for instance, was trained not only on web documents but also on interaction data between crowdworkers and the model itself, with the synthetic data annotated with calls to appropriate tools. This training process allowed LaMDA to develop the ability to use a calculator for arithmetic, a search engine for real-time information seeking, and a machine translator for translation. Similarly, Toolformer (Schick et al., 2024) learns to decide which APIs to call and what arguments to pass by training on template-generated data, while Galactica (Taylor et al., 2022) infuse API-calling data into pre-training mixture. ToolAlpaca (Tang et al., 2023) is a novel framework designed to automatically generate a diverse tool-use corpus, by building a multi-agent simulation environment and letting agents select and use tools iteratively. These examples demonstrate the potential of synthetic trajectories in enabling LMs to acquire tool-using abilities and enhance their reasoning capabilities across various domains.

Learning to plan in synthetic environments. An important feature of the agent in Autonomous Machine Intelligence (LeCun, 2022) is planning—an ability of decomposing complex tasks into subtasks and finishing the subtasks in a reward-optimal way (Kambhampati et al., 2024). Synthetic data can be a valuable tool here as it can serve as the feedback signal collected from a simulator (Park et al., 2023), and learning on it can make the agent aware of affordances (Ahn et al., 2022; Liang et al., 2022). For example, Inner Monologue (Huang et al., 2022) leverages natural language form feedback generated by the simulated environment to teach LLM-based robots planning. They find that such feedback significantly improves high-level instruction completion on both simulated and real-world domains. To compose a large number of realistic planning tasks (e.g., *"Rearrange objects on a table to match a given scene."*), VIMA (Jiang et al., 2022) creates a multi-modality simulated environment called VIMA-Bench, which supports extensible collections of objects and textures. In the Minecraft game, Voyager (Wang et al., 2023) deploys a number of GPT-4 based agents to interact with the synthetic environment and finds that the agents can unlock new skills faster and complete planning more efficiently with the help of synthetic feedback.

2.3 Multimodality

Reverse rendering from vision to text. Vision-language alignment data focuses on accurately grounding visual input to an LLM (usually via a vision encoder). Web-scraped image-caption pairs have been the most popular MM alignment data in the past few years since CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). However, web-scraped image-text pairs are usually noisy and only have coarse-grained correspondence, insufficient for grounding details of images in language. In domains such as documents, screens, figures, and diagrams, such fine-grained alignment can most conveniently be obtained from data synthesis pipelines built with image rendering engines. Pix2Struct (Lee et al., 2023) uses web servers to render HTML code into website screenshots, and the training task is to derender a masked screenshot to the full HTML code. MatCha (Liu et al., 2023b) and DePlot (Liu et al., 2023a) render tabular data into charts with Python plotting libraries and pretrain a foundation model by giving the rendered image and producing the code and/or the tabular data. Si et al. (2024) and Laurençon et al. (2024) train on synthetically generated HTML and CSS files for the task of converting webpage screenshots into code implementation. The models finetuned on the synthetic data can generalize reasonably well on realistic data scraped from the Internet. Borkman et al. (2021) propose to use physics engines or game engines (e.g., Unity) as the synthetic data generator to help computer vision research.

Multi-modality instruction following. Downstream applications of multimodal LLMs require reasoning and instruction following capabilities. Such data are usually long-form question response pairs and are expensive for humans to create. LLaVA (Liu et al., 2024b) uses existing image captions to prompt GPT-4 (in text-only mode) for writing diverse and long-form prompt-answer pairs. During multimodal LLM training, images and prompts are used as input while the captions and bounding box information can be hidden. Besides image captions, other sources of image attribute information such as object bounding box (Zhao et al., 2023), OCR (Zhang et al., 2023c) and derendered charts (Masry et al., 2023; Carbune et al., 2024) can all fit into such as image attributes + text LLM rewriting synthetic data pipeline.

2.4 Multilingual

Back-translation augmentation. Many multilingual language models use back-translation as a data augmentation method, creating synthetic parallel training data from monolingual data sources (Sennrich et al., 2016; Zheng et al., 2020; Caswell et al., 2019; Marie et al., 2020; Bi et al., 2021; Liao et al., 2021; Pham et al., 2021; Xu et al., 2022). For example, Sennrich et al. (2016) back-translate monolingual target data into source language data, providing additional parallel training samples for substantial translation task improvements. Researchers have also explored different sampling methods for back-translation (e.g., beam search, constrained sampling, unconstrained sampling) and their comparative effectiveness (Sennrich et al., 2016; Edunov et al., 2018; Graça et al., 2019; Bannard & Callison-Burch, 2005). Xu et al. (2022) emphasize the importance of the weight and quality of synthetic data for optimal NMT performance using back-translation. They propose a method to optimize the ratio between search methods and a gamma score to balance estimated importance weight and quality. However, some limitations exist with back-translation-based synthetic data generation. For example, the quality and diversity of synthetic data depends on the performance of the back-translation method. If the synthetic data is too noisy or not diverse, the performance gain would be limited (Epaliyana et al., 2021; Chauhan et al., 2022).

Generating multilingual questions and answers at scale. Recent studies explore the generation and utilization of synthetic multilingual question-answer (QA) pairs to improve language models' performance in multilingual and cross-lingual question answering (Asai et al., 2021; Kumar et al., 2019; Chi et al., 2020; Riabi et al., 2021; Li & Callison-Burch, 2023; Abulkhanov et al., 2023). One approach is to translate existing monolingual questions and/or answers into other languages (Asai et al., 2021). Another involves using Question Generation (QG) models to produce synthetic questions in a cross-lingual fashion based on answers and/or source texts (Kumar et al., 2019; Chi et al., 2020; Riabi et al., 2021). Recent efforts also focus on jointly generating questions and answers in multiple languages for greater flexibility (Shakeri et al., 2021; Li & Callison-Burch, 2023). For example, Shakeri et al. (2021) finetune a pretrained multilingual T5 model (Xue et al., 2020) on a mixture of a QA generation task and a multilingual masked language modeling task to produce synthetic QA pairs in multiple languages. These efforts generally show that language models trained on synthetic QA pairs demonstrate improved performance on multilingual QA and information retrieval benchmarks.

2.5 Alignment

Instruction Following. Synthetic data can serve as a promising approach for training instruction-following models, particularly in scenarios where real-world data is scarce, expensive, or challenging to obtain. Self-instruct (Wang et al., 2022a) and Stanford Alpaca (Taori et al., 2023) are both using LLMs to generate instruction following data which covers a wide range of scenarios. They first pick a small set of "seed instruction following samples" and then ask the LLMs to imitate the format to generate more demonstrations. One concern of this type of method is how to keep the generated data high quality, which involves the complexity of queries (Liu et al., 2023d), the diversity of semantics (Ding et al., 2023), and the scale of the synthetic dataset (Yuan et al., 2023). To this end, Xu et al. (2023) propose Evol-Instruct which adds complexity to simple instructions via prompting. Mukherjee et al. (2023) leverage LLMs to revise the instructions and responses iteratively to include high-quality explanation traces in the FLAN dataset (Wei et al., 2022), and they find the trained model has improved performance in many NLP tasks. UltraChat (Ding et al., 2023) is large-scale and multi-round synthetic dialogue dataset, which is generated by two separate ChatGPT Turbo API models—one serves as the user role while the other serves as the assistant. They instruct the user model with carefully designed prompts to mimic real human user behaviors.

Many language models are supervised finetuned to learn how to follow instructions, but in learning this behavior, they may inadvertently also learn to be *sycophantic* (Perez et al., 2023), tailoring their responses to follow a user's viewpoint, even if that viewpoint is not objectively correct (Wei et al., 2023b). Sharma et al. (2024) find evidence that the preference models (i.e., the reward model used for RLHF training) and even humans prefer sycophantic responses sometimes. On this front, Wei et al. (2023b) generates synthetic data to encourage models to be robust to user opinions and adds these data in a finetuning step to reduce sycophantic behavior on held-out prompts.

Mitigating hallucination. Many widely-used language models utilize supervised finetuning (SFT) to learn to align their interactions with users (Wang et al., 2022b; Zhang et al., 2023b). In particular, there exist many methods of generating synthetic SFT data that can improve capabilities such as reasoning and alignment (Wei et al., 2023a;b). It has been shown, however, that these synthetic data can induce hallucinations into language models by containing nontrivial amounts of hallucinated answers or by forcing models to learn to answer questions that they do not know the answer to (Zhang et al., 2023d). These cases demonstrate that synthetic data, if not applied correctly, can actually increase hallucinations in language models.

On the other hand, recent work has also shown promising results in mitigating hallucinations using synthetic data. For example, GPT-4 (OpenAI, 2023) was trained using a reward model that leveraged synthetic hallucination data in order to perform reinforcement learning (Zhang et al., 2023d). This method resulted in a significant improvement in performance on the TruthfulQA (Lin et al., 2022) dataset (Zhang et al., 2023d). Similarly, Jones et al. (2023) designed a synthetic task where hallucinations can be readily evaluated, utilizing this task to optimize LLM outputs by learning a continuous postfix via prefix-tuning. Tian et al. (2023) uses automated fact-checking and confidence scores to rank factuality scores of model response pairs, which are then used to finetune language models with DPO (Rafailov et al., 2023) to improve their factuality. Continued research in using synthetic data to mitigate hallucinations is still limited, however, by the lack of synthetic tasks for which hallucinations can be scalably evaluated.

Aligning with shared human preference and values. Directly finetuning on value-aligned or human-preferred data is a straightforward method for aligning language models, but this method often requires substantial human annotation, which can be prohibitively expensive at scale. Additionally, such annotation frequently exhibits varying styles and inconsistent quality, particularly in the case of poorly annotated samples at the lower end of the quality spectrum (Meta, 2023; Gilardi et al., 2023a). To address these practical challenges, an advanced technique known as "reinforcement learning from human feedback (RLHF)" has been proposed (Leike et al., 2018; Christiano et al., 2017; Ouyang et al., 2022). This approach involves training a reward model with human data to act as a proxy of human judgment, which guides the optimization of the LM generation policy.

Recent studies have proposed a mixture of synthetic data and real human data to train more robust reward models (Gao et al., 2023). Constitutional AI (Bai et al., 2022) proposes to use a small set of principles to steer the AI generated critiques and feedback, and use such synthetic data to replace the real human data in the typical RLHF pipeline. The model trained with this RLAIF (i.e., reinforcement learning from AI feedback) method shows similar strong performance as RLHF baselines. In general, synthetic data offers a powerful solution for human values and preferences alignment by allowing researchers to generate large-scale, diverse, and controlled training datasets in a low-cost way (Cui et al., 2023; Ganguli et al., 2022). By simulating a wide range of scenarios involving ethical dilemmas (Perez et al., 2022), social interactions (Liu et al., 2023c), and cultural norms (Ziems et al., 2023), synthetic data enables comprehensive and systematic testing of AI models' alignment with human values (Askell et al., 2021). This approach helps identify and mitigate issues related to bias (Liu et al., 2021; Ntoutsi et al., 2020), fairness (Zhao et al., 2018; Landers & Behrend, 2023), and unintended consequences before AI systems are deployed in real-world settings (Ye et al., 2024).

However, it is important to acknowledge that low-fidelity synthetic human preference data might be limited in accurately reflecting nuanced human judgment (Argyle et al., 2023). Consequently, the resulting models may be less robust under "jail-breaking attacks" (Huang et al., 2023a; Deshpande et al., 2023), and may reveal strategically deceptive behavior even through safety training (Pan et al., 2022; Steinhardt, 2022; Everitt et al., 2021). To mitigate these risks, researchers must continuously refine and improve the quality and diversity of synthetic data, incorporating more complex and comprehensive scenarios that better capture the intricacies of human values and preferences. Additionally, combining synthetic data with real-world data, and creating synthetic data in an interactive environment which can be synced with the real world, are promising remedies. As the need for effective AI governance and regulation grows, synthetic data will play an increasingly vital role in enabling scalable oversight mechanisms that promote trust, accountability, and the development of AI technologies that are aligned with human values and societal expectations.

3 Synthetic Data in Evaluation

Synthetic data is widely used in evaluations of different perspectives:

Factuality. AI systems may generate information or responses that are not grounded in factual knowledge or data, leading to the creation of misleading or false content, formally known as *hallucination* (Ji et al., 2023). Factuality evaluation aims to ensure the consistency of the knowledge in the AI system's output with the knowledge provided by its training data and knowledge base (Ji et al., 2023; Zhang et al., 2023d). Early statistical-based hallucination evaluation methods relied on n-grams to directly calculate the overlap of vocabulary between the input and output content (Dhingra et al., 2019; Wang et al., 2020). However, these methods have limitations, as they only consider lexical overlap and do not account for semantics or sentence meaning (Ji et al., 2023), making them unsuitable for evaluating more complex forms of hallucination. Subsequent assurance methods shifted from statistical approaches to model-based methods, which are more robust compared to token-difference-based methods (Honovich et al., 2021). While these model-based evaluation methods are more advanced than their predecessors, they still have limitations. For example, the models can only output the degree of hallucination and may struggle to pinpoint specific errors (Falke et al., 2019). Feng et al. (2023a) propose to combine LLMs generation with random walks on knowledge graphs to generate synthetic evaluation data for factuality, which is aware of entities and relations on the graphs. Wei et al. (2024) created a synthetic dataset called LongFact for long-form factuality evaluation and used Google Search as the grounding source and LLM for the automated judgement, to achieve human-level accuracy but with significally lower cost (Min et al., 2023).

Safety. Red teaming is a powerful technique for evaluating the safety and robustness of AI models (Ganguli et al., 2022; Casper et al., 2023b). By generating diverse and realistic scenarios designed to elicit unaligned or harmful outputs (Casper et al., 2023a), red teaming can expose vulnerabilities and weaknesses in AI systems (Perez et al., 2022). For example, Perez et al. (2023) use LMs to generate datasets for evaluating the behavior of other LMs. They end up producing 154 high-quality datasets which are verified by humans, and discover new cases of inverse scaling where LMs get worse with size. Hubinger et al. (2024) leverage synthetic data to trigger backdoor attacks to LMs at scale; they find LMs can exhibit deceptive behavior and create a false impression of safety under such attacks, and standard "safety training" could not remove such deception easily. These methods demonstrate the feasibility of using AI assistance to scale up human oversight (Bowman et al., 2022) over complex problems and unseen domains.

Assisting human evaluation. Recent studies have shown that in many cases, synthetic judgements from large-scale LMs (LLMs) can serve as qualified, fast, and low-cost alternatives to actual human evaluation (Gilardi et al., 2023b). Using GPT-4 as the judge, Alpaca Eval (Li et al., 2023b) and MT Bench (Zheng et al., 2023) are two popular benchmarks that measure the comprehensive abilities of LM-based ChatBot. In coding tasks, synthetic environment is a common choice to aid human evaluation, as humans can make the assessment more efficiently via actual executions and analysis on running logs. Gu et al. (2024) propose CRUXEval, a code execution reasoning benchmark consisting of 800 Python functions generated by CodeLLaMA-34B. Similarly, Liu et al. (2024a) introduce CodeMind, a framework to gauge the code reasoning abilities of LLMs on Independent Execution Reasoning (IER), Dependent Execution Reasoning (DER), and Specification Reasoning (SR). All these evaluations based on synthetic data show strong correlation with real human judgements.

4 Challenges and Limitations of Synthetic Data

While synthetic data offers numerous benefits and applications, it is crucial to acknowledge and address the potential challenges and limitations associated with its use. This section delves into three significant concerns surrounding synthetic data:

Misuse of synthetic data might proliferate misinformation. The potential misuse of synthetic data is a significant concern that must be addressed to ensure the responsible development of AI systems. Current AI models become increasingly capable of generating human-like data ranging from text (Gemini-Team et al., 2024; 2023), images (Saharia et al., 2022; Ramesh et al., 2022), songs [3], to even videos (e.g., OpenAI SORA [4]). This can be particularly dangerous when synthetic data is used to impersonate real people, manipulate public opinion, or influence political processes. Moreover, the dissemination of synthetic data-driven misinformation can erode trust in legitimate information sources, making it increasingly difficult for people to distinguish between truth and falsehood (Byman et al., 2023; Rid, 2020). To mitigate these risks, it is crucial for researchers, developers, and policymakers to establish clear guidelines and best practices for the ethical generation and use of synthetic data, including robust mechanisms for detecting and countering synthetic misinformation (Groh et al., 2022). By proactively addressing these challenges, we can harness the benefits of synthetic data while minimizing its potential for harm.

Synthetic data might cause ambiguity in AI alignment. The increasing use of synthetic data in aligning AI models (e.g., Constitutional AI (Bai et al., 2022)) can introduce significant ambiguity and uncertainty. The goal of AI alignment is to ensure that AI systems behave in ways that are aligned with human values and intentions. However, synthetic data, which is artificially generated rather than collected from real-world sources, may not accurately represent the nuances and complexities of human values and preferences (Zhou et al., 2024). This discrepancy can lead to AI models learning from data that is biased (Feng et al., 2023b; Liu et al., 2021), ungrounded (Liu et al., 2022; Patel & Pavlick, 2022), or misrepresentative of real-world scenarios (Weidinger et al., 2021; Ji et al., 2023). As a result, AI systems trained on synthetic data may exhibit behaviors that are misaligned with human expectations, potentially leading to unintended consequences or even harmful actions (Zou et al., 2023; Anderljung et al., 2023). Moreover, the ambiguity introduced by synthetic data can make it challenging to interpret and understand the decision-making processes of AI models (Lightman et al., 2023), further complicating the task of ensuring alignment. To mitigate these risks, it is crucial for researchers to carefully consider the limitations and potential drawbacks of using synthetic data in alignment research and to develop robust methods for validating and testing AI models trained on such data.

Training with synthetic data makes evaluation decontamination harder. The use of synthetic data in model training poses significant challenges to fair evaluation. Evaluation benchmarks are often created by referring to public text sources, such as coursework websites or forums. Consequently, it is arguable that all publicly available benchmark test cases might occasionally be included in the pre-training data of LLMs (Hoffmann et al., 2022; Gao et al., 2021). The use of synthetic data exacerbates this issue rather than mitigating it. Although the community has proposed several techniques to detect such evaluation contamination, such as *min-k% prob* (Shi et al., 2023), which checks the probabilities of k long-tail tokens, these token-level decontamination methods are inadequate when the model is trained with synthetic data. Synthetic data might include rephrased versions of the benchmark data (Oren et al., 2023; Mattern et al., 2023), rendering token-level decontamination ineffective. In addition to developing more advanced evaluation contamination detection techniques, we recommend that model developers invest in creating and maintaining in-house and protected evaluation benchmarks. These proprietary benchmarks should be carefully safeguarded to prevent leakage and ensure the integrity of the evaluation process.

[3] Make songs with Suno AI: `https://app.suno.ai/`

[4] OpenAI Sora: `https://openai.com/research/video-generation-models-as-world-simulators`

5 Directions for Future Work

As the field of synthetic data continues to evolve, there are several promising directions for future research and development. This section outlines three key areas that warrant further exploration:

Synthetic data scaling. The impressive performance of many over-trained small language models (e.g., Mistral series models (Jiang et al., 2023), and Gemma series models (Gemma-Team et al., 2024), *inter alia*) demonstrates the necessity of training with large amount of tokens (even passing the compute-optimal chinchilla law (Rae et al., 2021)). However, whether we have similar conclusions on the training with synthetic data is still an open question, as the quality of synthetic data may not be as consistent as real-world data (Yu et al., 2024). Future research should investigate the scaling laws for synthetic data and determine the optimal balance between the quantity and quality of synthetic samples. This exploration could help us understand the most effective strategies for leveraging synthetic data in training large-scale language models, potentially leading to more efficient and cost-effective approaches (Muennighoff et al., 2024).

Further improving quality and diversity of synthetic data. While existing methods for generating synthetic data have shown promise, there is still room for improvement in terms of creating high-quality, attributed synthetic samples that closely mimic real-world data. Future research should focus on developing new advanced techniques (or based on existing ones such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) or Diffusion Models (Ho et al., 2020), *inter alia*) that can control and manipulate specific attributes of the generated data, enabling the creation of diverse and customizable synthetic datasets. Additionally, researchers should explore methods that can incorporate domain-specific knowledge to ensure the generated data adheres to the underlying constraints and patterns present in the target domain (e.g., via Retrieval Augmented Generation (RAG) (Lewis et al., 2020; Borgeaud et al., 2022)) while maintaining the data quality. By advancing the state-of-the-art in attributed synthetic data generation, we can unlock new opportunities for privacy-preserving analysis (Assefa et al., 2020), and model training across various fields, from healthcare (e.g., synthetic medical images (Frid-Adar et al., 2018; Wei et al., 2019)) and finance (e.g., simulated trading trajectories (Zheng et al., 2022)) to social sciences (Argyle et al., 2023; Park et al., 2023) and beyond.

Towards high-fidelity and more efficient scalable oversight. As AI models become increasingly complex and autonomous, it becomes challenging to monitor and assess their behavior using traditional oversight methods that rely on human supervision or real-world data (Amodei et al., 2016). Future research should explore the use of synthetic data for high-fidelity scalable oversight of these advanced systems. Existing methods typically simulate a certain scenario in social iterations, such as debate (Leike et al., 2018), reflection (Zhang et al., 2023a), or revisions (Liu et al., 2023c) to obtain synthetic data, while new approaches could cover more comprehensive scenarios and more modalities (Sun et al., 2023), as recent studies have found many issues of simulation that only covers a narrowed down (Cheng et al., 2023) or over-simplified (Zhou et al., 2024) scenes. Looking forward, another growing direction could be how to achieve scalable oversight more efficiently—given that we have the full control over the synthetic data generation, we can probably provide more targeted oversights with less synthetic data. As the need for effective AI governance and regulation grows, synthetic data will play an increasingly vital role in enabling more trustworthy scalable oversight mechanisms that promote robust, accountable, and safe deployment of AI technologies for the benefit of society (Askell et al., 2021; Bowman et al., 2022).

The emergent self-improvement capability. We typically choose the most capable model to generate synthetic data, as its generation is of higher quality. However, an intriguing question arises: can a model generate synthetic data that is better than the data it was trained on, thus enabling it to improve itself? This concept of self-improvement through synthetic data generation is an exciting avenue for future research. If a model can generate higher-quality data than its original training set, it could potentially bootstrap its own performance by iteratively learning from the enhanced synthetic data (Chen et al., 2024). This self-improvement capability could lead to the emergence of more advanced AI systems that can autonomously refine their skills and knowledge over time (Burns et al., 2023; Huang et al., 2023b). Although recent work shows encouraging progress in this direction (Chen et al., 2024; Yuan et al., 2024), the upper bound of self-improvement and the underlying reasons for its effectiveness remain open questions. Future research should investigate the theoretical underpinnings and practical feasibility of self-improvement through synthetic data generation in more diverse scenarios, examining the necessary conditions, potential limitations, and associated risks. By unlocking the potential of emergent self-improvement capabilities, we could enable more adaptable, efficient, and autonomous learning processes (LeCun, 2022).

6 Conclusion

Synthetic data has emerged as a promising solution to address the challenges of data scarcity, privacy concerns, and high costs in AI development. By generating realistic and diverse datasets, synthetic data enables the training and evaluation of AI models at scale across various domains. As we approach human-level or even superhuman-level intelligence, obtaining synthetic data becomes even more crucial, given that models need better-than-average-human quality data to progress. However, ensuring the factuality, fidelity, and lack of bias in synthetic data remains a critical challenge.

Future research directions on synthetic data could focus on improving the fidelity and controllability of generative models and developing standardized evaluation and contamination protocols and tools. We could also explore the integration of synthetic data with other techniques and its application in other domains. Despite the challenges, the potential benefits of synthetic data in advancing AI research are significant. By leveraging synthetic data responsibly and effectively, we can build more powerful, inclusive, and trustworthy AI systems that benefit society as a whole.

References

Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pp. 510–526. Springer, 2019.

Dmitry Abulkhanov, Nikita Sorokin, Sergey Nikolenko, and Valentin Malykh. Lapca: Language-agnostic pretraining with cross-lingual alignment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2098–2102, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *ArXiv preprint*, abs/2204.01691, 2022. URL `https://arxiv.org/abs/2204.01691`.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *ArXiv preprint*, abs/1606.06565, 2016. URL `https://arxiv.org/abs/1606.06565`.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *ArXiv preprint*, abs/2307.03718, 2023. URL `https://arxiv.org/abs/2307.03718`.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 7547–7560, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3df07fdae1ab273a967aaa1d355b8bb6-Abstract.html.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *ArXiv preprint*, abs/2112.00861, 2021. URL https://arxiv.org/abs/2112.00861.

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *ArXiv preprint*, abs/2310.10631, 2023. URL https://arxiv.org/abs/2310.10631.

Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8):1329–1351, 2019.

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

Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer (eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 597–604, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219914. URL https://aclanthology.org/P05-1074.

Enrico Barbierato, Marco L Della Vedova, Daniele Tessera, Daniele Toti, and Nicola Vanoli. A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences*, 12(9):4619, 2022.

Wei Bi, Huayang Li, and Jiacheng Huang. Data augmentation for text generation without any augmented data. 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)*, pp. 2223–2237, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.173. URL https://aclanthology.org/2021.acl-long.173.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *ArXiv preprint*, abs/2210.06280, 2022. URL https://arxiv.org/abs/2210.06280.

Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, Y. C. Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, Cesar Romero, Wesley Smith, Alex Thaman, Samuel Warren, and Nupur Yadav. Unity perception: Generate synthetic data for computer vision. *ArXiv preprint*, abs/2107.04259, 2021. URL https://arxiv.org/abs/2107.04259.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *ArXiv preprint*, abs/2211.03540, 2022. URL https://arxiv.org/abs/2211.03540.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *ArXiv preprint*, abs/2312.09390, 2023. URL https://arxiv.org/abs/2312.09390.

Daniel L Byman, Chongyang Gao, Chris Meserole, and VS Subrahmanian. *Deepfakes and international conflict*. Brookings Institution, 2023.

Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. Chart-based reasoning: Transferring capabilities from llms to vlms. *ArXiv preprint*, abs/2403.12596, 2024. URL https://arxiv.org/abs/2403.12596.

Stephen Casper, Tong Bu, Yuxiao Li, Jiawei Li, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *ArXiv preprint*, abs/2306.09442, 2023b. URL https://arxiv.org/abs/2306.09442.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 53–63, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5206. URL https://aclanthology.org/W19-5206.

Shweta Chauhan, Shefali Saxena, and Philemon Daniel. Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low resource languages. *Neural Processing Letters*, 54(3):1707–1726, 2022.

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

Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.669. URL https://aclanthology.org/2023.emnlp-main.669.

Ethan Chern, Haoyang Zou, Xuefeng Li, Jiewen Hu, Kehua Feng, Junlong Li, and Pengfei Liu. Generative ai for math: Abel. https://github.com/GAIR-NLP/abel, 2023.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7570–7577. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6256.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html`.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Jessamyn Dahmen and Diane Cook. Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1181, 2019.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *ArXiv preprint*, abs/2304.05335, 2023. URL `https://arxiv.org/abs/2304.05335`.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4884–4895, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1483. URL `https://aclanthology.org/P19-1483`.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *ArXiv preprint*, abs/2305.14233, 2023. URL `https://arxiv.org/abs/2305.14233`.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL `https://aclanthology.org/D18-1045`.

Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

Koshiya Epaliyana, Surangika Ranathunga, and Sanath Jayasena. Improving back-translation with iterative filtering and data selection for sinhala-english nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pp. 438–443. IEEE, 2021.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2214–2220, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213. URL `https://aclanthology.org/P19-1213`.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 933–952, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.59. URL `https://aclanthology.org/2023.emnlp-main.59`.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *ArXiv preprint*, abs/2305.08283, 2023b. URL `https://arxiv.org/abs/2305.08283`.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293. IEEE, 2018.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv preprint*, abs/2209.07858, 2022. URL `https://arxiv.org/abs/2209.07858`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv preprint*, abs/2101.00027, 2021. URL `https://arxiv.org/abs/2101.00027`.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

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

Gemini-Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530, 2024. URL `https://arxiv.org/abs/2403.05530`.

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *ArXiv preprint*, abs/2403.08295, 2024. URL `https://arxiv.org/abs/2403.08295`.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023a.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023b. doi: 10.1073/pnas.2305016120. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2305016120`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 45–52, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5205. URL `https://aclanthology.org/W19-5205`.

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. *ArXiv preprint*, abs/2401.03065, 2024. URL `https://arxiv.org/abs/2401.03065`.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 666–667, 2020.

Aman Gupta, Deepak Bhatt, and Anubha Pandey. Transitioning from real to synthetic data: Quantifying the bias in model. *ArXiv preprint*, abs/2105.04144, 2021. URL `https://arxiv.org/abs/2105.04144`.

Patrick Haluptzok, Matthew Bowers, and Adam Tauman Kalai. Language models can teach themselves to program better. *ArXiv preprint*, abs/2207.14502, 2022. URL `https://arxiv.org/abs/2207.14502`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7856–7870, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL `https://aclanthology.org/2021.emnlp-main.619`.

Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. Synthetic data for social good. *ArXiv preprint*, abs/1710.08874, 2017. URL `https://arxiv.org/abs/1710.08874`.

Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *ArXiv preprint*, abs/2302.07736, 2023a. URL `https://arxiv.org/abs/2302.07736`.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1051–1068, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.67. URL `https://aclanthology.org/2023.emnlp-main.67`.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *ArXiv preprint*, abs/2207.05608, 2022. URL `https://arxiv.org/abs/2207.05608`.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *ArXiv preprint*, abs/2401.05566, 2024. URL `https://arxiv.org/abs/2401.05566`.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *ArXiv preprint*, abs/2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

Erik Jones, Hamid Palangi, Clarisse Simões, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed Awadallah, and Ece Kamar. Teaching language models to hallucinate less with synthetic tasks, 2023. URL https://arxiv.org/abs/2310.06827.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4863–4872, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1481. URL https://aclanthology.org/P19-1481.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Richard N Landers and Tara S Behrend. Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*, 78(1):36, 2023.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024. URL https://arxiv.org/abs/2403.09029.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *ArXiv preprint*, abs/1811.07871, 2018. URL https://arxiv.org/abs/1811.07871.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina

Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Bryan Li and Chris Callison-Burch. Paxqa: Generating cross-lingual question answering examples at training scale. *ArXiv preprint*, abs/2304.12206, 2023. URL https://arxiv.org/abs/2304.12206.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *ArXiv preprint*, abs/2403.04706, 2024. URL https://arxiv.org/abs/2403.04706.

Lizao Li, Rob Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Seeds: Emulation of weather forecast ensembles with diffusion models. *ArXiv preprint*, abs/2306.14066, 2023a. URL https://arxiv.org/abs/2306.14066.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023b.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *ArXiv preprint*, abs/2209.07753, 2022. URL https://arxiv.org/abs/2209.07753.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. Back-translation for large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 418–424, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.50.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *ArXiv preprint*, abs/2305.20050, 2023. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Changshu Liu, Shizhuo Dylan Zhang, and Reyhaneh Jabbarvand. Codemind: A framework to challenge large language models for code reasoning. *ArXiv preprint*, abs/2402.09664, 2024a. URL https://arxiv.org/abs/2402.09664.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023a.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Haoxiong Liu and Andrew Chi-Chih Yao. Augmenting math word problems via iterative question composing. *ArXiv preprint*, abs/2401.09003, 2024. URL https://arxiv.org/abs/2401.09003.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 14857–14866. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17744.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *ArXiv preprint*, abs/2210.05359, 2022. URL https://arxiv.org/abs/2210.05359.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models in simulated human society. *ArXiv preprint*, abs/2305.16960, 2023c. URL https://arxiv.org/abs/2305.16960.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *ArXiv preprint*, abs/2312.15685, 2023d. URL https://arxiv.org/abs/2312.15685.

Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, and Wenqi Wei. Machine learning for synthetic data generation: a review. *ArXiv preprint*, abs/2302.04062, 2023. URL https://arxiv.org/abs/2302.04062.

Fernando Lucini. The real deal about synthetic data. *MIT Sloan Management Review*, 63(1): 1–4, 2021.

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

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *ArXiv preprint*, abs/2306.08568, 2023b. URL https://arxiv.org/abs/2306.08568.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5990–5997, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.532. URL https://aclanthology.org/2020.acl-main.532.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14662–14684, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.906. URL https://aclanthology.org/2023.emnlp-main.906.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *ArXiv preprint*, abs/2305.18462, 2023. URL https://arxiv.org/abs/2305.18462.

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

Meta. Meta and microsoft introduce the next generation of llama. https://ai.meta.com/blog/llama-2, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *ArXiv preprint*, abs/2306.02707, 2023. URL https://arxiv.org/abs/2306.02707.

Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

OpenAI. Gpt-4 technical report, 2023.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *ArXiv preprint*, abs/2310.17623, 2023. URL https://arxiv.org/abs/2310.17623.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. *ArXiv preprint*, abs/2310.06786, 2023. URL https://arxiv.org/abs/2310.06786.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gJcEM8sxHK.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.225.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13387–13434. Association for Computational Linguistics, 2023.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. Meta back-translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=3jjmdp7Hha.

Michael Przystupa and Muhammad Abdul-Mageed. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pp. 224–235, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5431. URL https://aclanthology.org/W19-5431.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2021. URL https://arxiv.org/abs/2112.11446.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023. URL https://api.semanticscholar.org/CorpusID:258959321.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, abs/2204.06125, 2022. URL https://arxiv.org/abs/2204.06125.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7016–7030, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.562. URL https://aclanthology.org/2021.emnlp-main.562.

Thomas Rid. *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux, 2020.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1gR5iR5FX.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Siamak Shakeri, Noah Constant, Mihir Kale, and Linting Xue. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. In *Proceedings of the 14th International Conference on Natural Language Generation*, pp. 35–45, Aberdeen, Scotland, UK, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.4.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2023.

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

Alexander Shypula, Aman Madaan, Yimeng Zeng, Uri Alon, Jacob Gardner, Milad Hashemi, Graham Neubig, Parthasarathy Ranganathan, Osbert Bastani, and Amir Yazdanbakhsh. Learning performance-improving code edits. *ArXiv preprint*, abs/2302.07867, 2023. URL https://arxiv.org/abs/2302.07867.

Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: How far are we from automating front-end engineering?, 2024. URL https://arxiv.org/abs/2403.03163.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *ArXiv preprint*, abs/2212.13138, 2022. URL https://arxiv.org/abs/2212.13138.

Jacob Steinhardt. Ml systems will have weird failure modes. https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/, 2022.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*, abs/2309.14525, 2023. URL https://arxiv.org/abs/2309.14525.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *ArXiv preprint*, abs/2306.05301, 2023. URL https://arxiv.org/abs/2306.05301.

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

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv preprint*, abs/2211.09085, 2022. URL https://arxiv.org/abs/2211.09085.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239, 2022. URL https://arxiv.org/abs/2201.08239.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *ICLR*, 2023. URL https://api.semanticscholar.org/CorpusID:265158181.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

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

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Boris Van Breugel, Zhaozhi Qian, and Mihaela Van Der Schaar. Synthetic data, real errors: how (not) to publish and use synthetic data. In *International Conference on Machine Learning*, pp. 34793–34808. PMLR, 2023.

Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *ArXiv preprint*, abs/2312.03664, 2023. URL https://arxiv.org/abs/2312.03664.

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *ArXiv preprint*, abs/2305.16291, 2023. URL `https://arxiv.org/abs/2305.16291`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. 2022a. URL `https://arxiv.org/abs/2203.11171`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. volume abs/2212.10560, 2022b. URL `https://arxiv.org/abs/2212.10560`.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1072–1086, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.101. URL `https://aclanthology.org/2020.acl-main.101`.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=gEZrGCozdqR`.

Jerry Wei, Arief Suriawinata, Louis Vaickus, Bing Ren, Xiaoying Liu, Jason Wei, and Saeed Hassanpour. Generative image translation for data augmentation in colorectal histopathology images. In *Advances in Neural Information Processing Systems*, 2019.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. volume abs/2305.08298, 2023a. URL `https://arxiv.org/abs/2305.08298`.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2023b. URL `https://arxiv.org/abs/2308.03958`.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. 2024. URL `https://api.semanticscholar.org/CorpusID:268724304`.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need. *ArXiv preprint*, abs/2312.02120, 2023c. URL `https://arxiv.org/abs/2312.02120`.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *ArXiv preprint*, abs/2112.04359, 2021. URL `https://arxiv.org/abs/2112.04359`.

Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 3661–3671. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00366. URL `https://doi.org/10.1109/ICCV48922.2021.00366`.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *ArXiv preprint*, abs/2304.12244, 2023. URL `https://arxiv.org/abs/2304.12244`.

Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. On synthetic data for back translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 419–430, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.32. URL https://aclanthology.org/2022.naacl-main.32.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. Toolsword: Unveiling safety issues of large language models in tool learning across three stages. *ArXiv preprint*, abs/2402.10753, 2024. URL https://arxiv.org/abs/2402.10753.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *ArXiv preprint*, abs/2309.12284, 2023. URL https://arxiv.org/abs/2309.12284.

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

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *ArXiv preprint*, abs/2401.10020, 2024. URL https://arxiv.org/abs/2401.10020.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *ArXiv preprint*, abs/2308.01825, 2023. URL https://arxiv.org/abs/2308.01825.

E. Zelikman, Yuhuai Wu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022. URL https://api.semanticscholar.org/CorpusID:247762790.

Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *ArXiv preprint*, abs/2310.02124, 2023a. URL https://arxiv.org/abs/2310.02124.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2023b. URL https://arxiv.org/abs/2308.10792.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv preprint*, abs/2306.17107, 2023c. URL https://arxiv.org/abs/2306.17107.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219, 2023d. URL https://arxiv.org/abs/2309.01219.

Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *ArXiv preprint*, abs/2307.04087, 2023. URL https://arxiv.org/abs/2307.04087.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances*, 8(18):eabk2607, 2022.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. Mirror-generative neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HkxQRTNYPH.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *ArXiv preprint*, abs/2403.05020, 2024. URL https://arxiv.org/abs/2403.05020.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. Normbank: A knowledge bank of situational social norms. *ArXiv preprint*, abs/2305.17008, 2023. URL https://arxiv.org/abs/2305.17008.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint*, abs/2307.15043, 2023. URL https://arxiv.org/abs/2307.15043.