Open-Ethical AI: Advancements in Open-Source Human-Centric Neural Language Models

SABRINA SICARI, Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy
JESUS F. CEVALLOS M., Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy
ALESSANDRA RIZZARDI, Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy
ALBERTO COEN-PORISINI, Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy

This survey summarises the most recent methods for building and assessing *helpful, honest, and harmless* neural language models, considering small, medium, and large-size models. Pointers to open-source resources that help to align pre-trained models are given, including methods that use parameter-efficient techniques, specialized prompting frameworks, adapter modules, case-specific knowledge injection, and adversarially robust training techniques. Special care is given to evidencing recent progress on value alignment, commonsense reasoning, factuality enhancement, and abstract reasoning of language models. Most reviewed works in this survey publicly shared their code and related data and were accepted in world-leading Machine Learning venues. This work aims at helping researchers and practitioners accelerate their entrance into the field of human-centric neural language models, which might be a cornerstone of the contemporary and near-future industrial and societal revolution.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Natural language generation**; **Discourse, dialogue and pragmatics;**

Additional Key Words and Phrases: Neural language models, open-source, large-language models, human-centric AI

ACM Reference Format:
Sabrina Sicari, Jesus F. Cevallos M., Alessandra Rizzardi, and Alberto Coen-Porisini. 2024. Open-Ethical AI: Advancements in Open-Source Human-Centric Neural Language Models. *ACM Comput. Surv.* 57, 4, Article 83 (December 2024), 47 pages. https://doi.org/10.1145/3703454

This work was supported in part by the SERENA project, which has been funded by MUR (*Ministero dell'Università e della Ricerca*) under the PRIN 2022 program (project code 2022CN4EBH), and in part by project SERICS (project code PE00000014), under the NRRP MUR program funded by the EU - NGEU..
Authors' Contact Information: Sabrina Sicari (Corresponding author), Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy; e-mail: sabrina.sicari@uninsubria.it; Jesus F. Cevallos M., Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy; e-mail: jf.cevallosmoreno@uninsubria.it; Alessandra Rizzardi, Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy; e-mail: alessandra.rizzardi@uninsubria.it; Alberto Coen-Porisini, Dipartimento di Scienze Teoriche e Applicate, Universita degli Studi dell'Insubria, Varese, Italy; e-mail: alberto.coenporisini@uninsubria.it.

ACM 0360-0300/2024/12-ART83
https://doi.org/10.1145/3703454

1 Introduction

The unprecedented growth of research efforts to mitigate the actual and potential risks from the widespread use of ***Artificial Intelligence (AI)*** is popularly noted nowadays [126]. In ***Neural Language Models (NLM)***, this trend has recently materialized in lots of practices such as **prompt engineering** (**PE**) [179], **instruction fine-tuning** (**IFT**) [342], and **reinforcement learning from human feedback** (**RLHF**) [49]. Open benchmarks for assessing the alignment of NLMs concerning user requirements, intentions, and moral preferences [37] and novel methods that go beyond canonical prompting strategies and IFT have been published in the last few years. To this respect, aiming at helping practitioners and researchers shape their current work on the human alignment of NLMs, this article systematically reviews the most recent and significant methods and benchmarks for human-centricity.

Recent surveys have devoted themselves to cataloguing the intention-alignment techniques for agentic AIs [313] (i.e., those meant to interact with an environment through actions) and **Large Language Models** (**LLMs**) [265]. Other work has concentrated on giving a taxonomy of SOTA techniques for moral or ethical alignment of AIs [48], safe governance in the transition to superhuman AI [190], commonsense reasoning [276] and, not less importantly, grounding the AI-alignment empirical results on a foundational theory [219].

This survey focuses instead on the *human-centric* design of all-size neural language models in general. The three H's of the well-known lemma of Askell et al. [14] are used to define human centricity: **human-centric NLMs are those that are *Helpful, Harmless, and Honest***. This seminal taxonomy of requirements for NLMs is decoded in this work into more specific desiderata, such as the extrapolation of commonsense reasoning, abstract reasoning, factuality, and safety guardrails in **out-of-distribution** (**OOD**) scenarios, among others.

Main Contribution

To the best of the authors' knowledge, this survey is the first to encompass a review of SOTA techniques for aligning NLMs with the goals of the 3H human-centric paradigm. Moreover, beyond well-known canonical prompt engineering and instruction fine-tuning techniques, our review has a special focus on more sophisticated approaches, such as those based on retrieval-augmented models, knowledge distillation, programming-aided models, language-augmented reinforcement learning, and debate, among others, privileging open-source works to help the entrance of new researchers to the field. Finally, this work also provides the latest benchmarks for assessing NLMs and is not focused only on *large* language models but evidences active research efforts to democratise human-centric AI by enhancing the alignment of small and medium-sized NLMs.

Sources, Keywords and Scope. Google Scholar was used to scout the most recent literature, using a list of prefix keywords prepended to another list of subject keywords. The former list includes "human-centric", "human-centred", "aligned", "commonsense-aware", "ethical-aware", and "factuality-enhanced", among others. The latter consisted of "language models", "large language models", "open-source language models", "small language models", "artificial intelligence", and "AI agents". The method articles surveyed have been published since 2020, making exceptions for seminal articles that introduced important techniques and open-source benchmarks and datasets.

Related Works

The authors of [326] survey current alignment goals for **foundational models** (**FM**) and alignment evaluation methodologies. They divided such goals into three groups: aligning to specific instructions, aligning with human preferences over a set of equally likely outputs, and

aligning with human values. The authors offer a list of current open challenges in AI alignment. The survey in [265] focuses instead on the alignment of LLMs, distinguishing between techniques devoted to the correct encoding of alignment goals (outer alignment) and techniques that ensure a robust extrapolation of the encoded goals over OOD scenarios (inner alignment). Interpretability aspects of LLMs are also touched from the architectural point of view. Davis et al. [60] instead concentrate on benchmarks for commonsense reasoning in AI, including visual perception with image and video-related benchmarks. Kirchner et al. [148] present the results of analysing an extensive collection of literature on AI alignment. The authors pointed to several open-source repositories beyond scholarly publications to collect research products on AI alignment. Apart from ArXiv, they pointed to the *Alignment Forum* [121] as an essential source of related research material, where authors interested in the foundations of this subject might be currently interchanging research products at a more agile pace.

The authors of [37] systematised the **state-of-the-art** (**SOTA**) related to evaluating LLMs. A taxonomy of the evaluation goals is given in this work. Functional requirements such as understanding open-domain dialogue and domain-specific subjects are included in this taxonomy. The authors also surveyed benchmarks for value alignment, factuality, and bias of LLMs. The work in [126] instead generally analyses AI alignment, dividing approaches by those related to learning from feedback and learning across distributional shifts. J. Ji et al. distinguish between forward and backward alignment, where the former is related to alignment techniques during the training phase, and the latter involves verification of alignment at inference time. Osborne et al. [227] focus on the state-of-the-art of **Deep Reinforcement learning** (**DRL**) models in text-based interaction environments. They describe the inherent challenges for generalising the capability of agents to solve text-based interaction games and offer a list of current environment generation tools. Interestingly, authors evidenced interpretability benefits of agents that rely on language models based on knowledge graphs and graph-based reinforcement learning. Authors in [33] present a taxonomy of the challenges for effectively applying RLHF. To do this, they consider the three phases of canonical RLHF: feedback collection, reward modelling, and policy optimisation. The authors also point to alignment techniques that combine AI feedback to complement RLHF cycles.

The work in [294] elaborates on the factuality issue of LLMs. The authors give a strict definition of the problem, an analysis of the causes of hallucination, and a taxonomy of evaluation and factuality enhancement methods. The latter techniques involve standalone pre-trained LLMs, knowledge-augmented models, and domain-specific use cases. Chang et al. [36] concentrates on behavioural studies of out-of-the-box pre-trained language models, surveying current tendencies to output unfactual, toxic, or memorised text before the alignment procedure. The work in [294] also surveys techniques that focus on enhancing the factuality of LLMs. The authors broadly define the factuality problem as encompassing commonsense reasoning, domain-specific knowledge, and consistency with general factual information. The state-of-the-art in terms of proposed metrics and techniques is surveyed, and the corresponding methodologies are divided into knowledge retrieval-based methods and standalone LLMs, both in general-case and domain-specific neural language tools.

With respect to these recent works, this survey reviews methods and benchmarks for a broadly defined human-centric alignment assessment of NLMs' based on the 3H paradigm. This work offers a birds-eye view of the current advancements and benchmarks for helpful, honest, and harmless NLMs, studying the most recent open-source contributions, and giving special care including small language models. Figure 1 presents a schematic view of the proposed taxonomy for human-centric NLM alignment techniques and instruments.

Outline. This survey is organized as follows: Section 2 explores the current techniques and benchmarks for training helpful NLMs. Following, Section 3 concentrates on the harmlessness

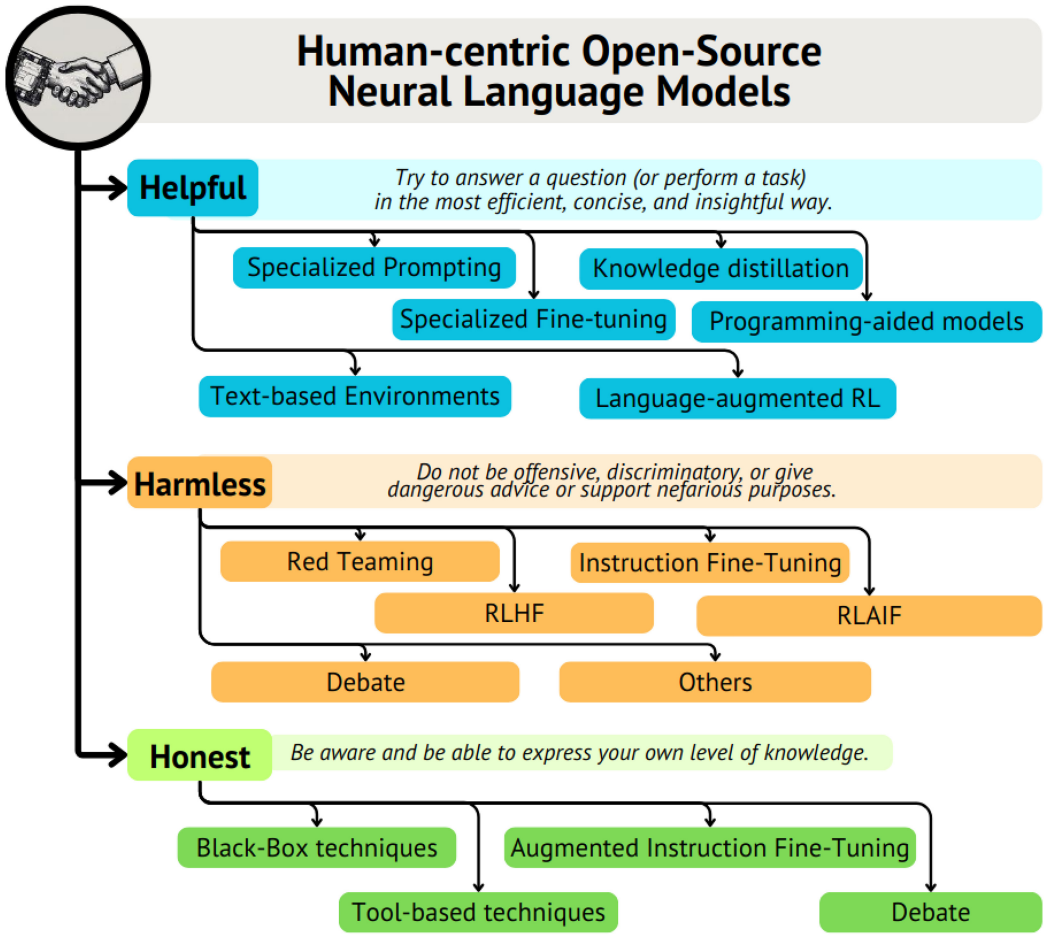

Fig. 1. The proposed taxonomy of this research. Works are divided into those focused on augmenting the helpfulness, harmfulness, and honestness of neural language models.

maximisation. Section 4 successively reviews the SOTA approaches for honestness enhancement in current NLMs. Finally, this article evidences its conclusions in Section 5 by pointing out five calls for future research in human-centric NLMs. Apart from the proper names of algorithms and models, Table 1 shows a list of the abbreviations used in this research.

A note on terminology. Although the term "large language models" is commonly used to refer to neural language models in general, to the best of the authors' knowledge, there is no fixed and clear statement in the **Natural Language Processing** (**NLP**) research community about the *small*, *medium*, and *large* qualifiers referred to neural language models. However, following common trends in the research-field jargon [234], this work uses the qualifier *medium* to models that range **approximately** from ten to one-hundred Billion parameters, giving a flexible but clear sense of what *small* and *large* refers to. Also, following the research community's conventional preference, the *short-scale* meaning of Billion, i.e., 10^9, is adopted in this work, and the capital B is the reference measure unit for the size of NLMs. For example, a model of size 7B is a model with 7×10^9 parameters or learnable weights.

Table 1. Abbreviations used in this Research

Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CSKG	Commonsense Knowledge Graph
CoT	Chain-of-Thought
DAG	Directed Acyclic Graph
DL	Deep Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
FM	Foundational Models
GAT	Graph Attention Network
HHH	Helpful, Harmless, and Honest
ICL	In-context Learning
IFT	Instruction Fine-Tuning
KG	Knowledge Graph
LSTM	Long-Short Memory Network
LLM	Large Language Model
ML	Machine Learning
MLP	Multi-Layer Perceptron
NLI	Natural Language Inference
NLM	Neural Language Model
NLP	Natural Language Processing
NLU	Natural Language Understanding
OOD	Out of Distribution
PPO	Proximal Policy Optimization
QA	Question-Answer
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
RoT	Rules of Thumb
SOTA	State-of-the-art

2 Helpfulness

In Reference [14], a Helpful AI-based assistant is defined as a function of the degree to which an attempt is made to answer a question (or perform a task) in the most *efficient*, *concise*, and *insightful* way. In this survey, such a requirement is mapped to the *reasoning* ability, as narrowly defined by Yu et al. in [331]: reasoning as the ability to reach new assertions, relations, and actions that are concordant with an actual or hypothetic scenario, as a function that integrates implicit knowledge and explicit contextual information. Using this definition, it is easy to see that features like compositional generalization [170], commonsense [60], context-awareness [313] and abstracting [212] are included inside the *reasoning* requisite of human-centric NLMs. This section reviews the most recent and significant techniques and benchmarks that implement and evaluate such abilities in NLMs. The review of related literature is divided into the contributions focused on NLM tools and those related to NLM-based agentic systems. Finally, a list of available benchmarks to assess the helpfulness of NLMs is presented.

2.1 Language Model-based Tools

Specialized Prompting. One of the most used techniques for NLM alignment exploits the ***In-Context Learning*** **(ICL)** ability exhibited by a set of pre-trained language models, especially

LLMs, by which, under certain conditions, the model shows zero-shot learning abilities that modify the distribution of responses at inference time [66]. In other words, practitioners use ICL to align an NLM to a desired output distribution by enriching the prompt with query-response examples that align with the desired behaviour [179]. Crafting the prompts to reach a desired output distribution can be more or less arduous, depending on the desired goal, and this practice has taken the name of PE [93].

One exemplary method that is based on ICL is that of ***Chain-of-Thought*** (**CoT**) prompting [307] that provides the NLM with a series of sample responses in the form of step-by-step decompositions of complex queries. Such a technique has shown to be effective, for example, in enhancing the reasoning capabilities of NLMs, and other works have refined the CoT to apply it to various specific tasks. For example, the *self-consistency* framework [299] improves the reasoning robustness concerning CoT prompting by sampling various answers from the NLM and marginalizing a more coherent answer. Instead, the authors of *Faithful CoT* [195] focus on improving the interpretability of NLMs using a two-stage process. In the first phase, the model is prompted with a particular form of CoT to produce a reasoning chain in the symbolic language given a query. In the second stage, instead, the symbolic outputs for each query are executed by a deterministic solver to lead to the answers. By doing so, Faithful CoT not only leveraged transparency to NLM-based responses to complex queries but also improved the correctness over at-the-time SOTA models. One of the most recent and significant contributions in the context of CoT is the *Least-to-Most* prompting framework [363] in which the in-context examples permit an NLM to apply *divide-and-conquer* complex task decomposition. A more in-depth study on the effectiveness and weaknesses of CoT is presented in [200]. Recent surveys of CoT reasoning applications are available at [50, 240].

Besides CoT, another effective prompt engineering technique that helps the model to reason is the one in [367]. In this work, prompting is used to help NLMs deduce logical rules from examples. Once these logical rules are produced, the subset with the greatest accordance frequency with a corpus of examples is extracted. A library of practices is grasped, which allows the NLM to exhibit better performance in other logical reasoning tasks. The authors of DUMA [282], instead, combined two NLMs specifically devoted to *slow and fast thinking*, as defined in [59], to augment the reasoning capabilities and concordant use of external knowledge for conversational tasks. The authors of *Skills-in-Context Prompting* [38] focused instead on teaching NLMs to exhibit compositional generalization, i.e., to compose basic reasoning building blocks to solve hard problems. More specifically, the authors SKIC framework create rich prompts containing a set of skills that need to be composed to solve hard problems. SKIC established itself as the SOTA on the MATH benchmark. The *Re-reading* (RE2) methodology [316] is another simple prompting strategy that contains an explicit repetition of the same query before asking for its answer. The authors of the RE2 framework showed improved reasoning and factuality concerning some CoT strategies. Recent research in [236] pushed PE to the limits by testing the abilities of NLMs to achieve tasks where the specification was ever more complicated and extensive. Unsurprisingly, the authors find that, for such specification-heavy functions that take several hours for humans to master, IFT still outperforms PE-based techniques.

Specialised Fine-tuning. Prompt Engineering is widely used to demonstrate the online learning capabilities of NLMs, but it also helps create datasets for gradient-based fine-tuning of pre-trained models. The latter technique is called IFT [342]. The main advantage of using IFT is that once fine-tuned, an NLM does not need enriched prompts to output the aligned responses. Instead, the aligned behaviour is the new default distribution of responses. Despite the emergence of abstract reasoning through ICL, fine-tuning NLMs on specialised query-response

data is still valuable for augmenting specialised reasoning in several scenarios, especially for medium-size NLMs. For example, the authors of MetaMath [333] created an IFT dataset focused on mathematical reasoning and used it to fine-tune a 7B Llama-2 model and outperform GPT-3.5-turbo [225] on the Gsm8k [52] and the Math [106] benchmarks. The authors of [295], instead, collected 80K math problem-solving query-response pairs from the closed-source gpt-4 Code Interpreter in [362] and provided a dataset and an IFT recipe for fine-tuning open-source models to enhance their mathematical reasoning capabilities. Similar customised fine-tuning proposals are in [172, 296, 335]. The PandaLM framework in [302] also uses an expert LLM to help the instruction fine-tuning phase of smaller-size NLMs by evaluating the outcomes of various fine-tuned candidates across different metrics involving not only accuracy in the responses, but also stylistic alignment in terms of conciseness, clarity, comprehensiveness, among other aspects. The tülu and [300] tülu2 model family [122] are similar medium-size models, based on the Meta's llama1 pre-trained models and fine-tuned on a set of carefully selected IFT datasets.

Knowledge distillation. One of the main limitations of reasoning through ICL is that only language models with more than 100B parameters may show proficiency with in-context examples[202]. To this respect, the authors of [110] focus on exporting the reasoning capabilities from large to modest-size NLMs by generating an IFT dataset by CoT prompting large models and using such datasets to fine-tune more modest-sized models. The *Distilling step-by-step* framework [113] not only fine-tunes small NLMs using CoT responses, but optimises a *rationale prediction* objective alongside response prediction to augment the proficiency of small models during fine-tuning. The authors of *Program-aided Distillation* [366] induced the expert large models to output the source code of programs that can be compiled and run to answer the questions. They fine-tuned smaller models with such a programming-oriented dataset and found that their framework improves training data efficiency and enables small models to reach good performance in reasoning benchmarks. A similar recent work that follows the trend in transferring the reasoning capabilities of LLMs to modest-size NLMs is available in [319].

Programming Aided Language Models. A recent paradigm for leveraging abstract reasoning in LLMs is that of program-aided language models [83], where NLMs are steered to generate executable programs as a proxy for solving complex reasoning tasks like that involving arithmetic problems. The work in [201] demonstrated the effectiveness of code LLMs to leverage structural rationale, i.e., to map free text inputs to structured outputs like graphs or tables. Their experiments involved three structured output benchmarks: proscript [257], propara [211], and explagraphs [256], where the focus is on generating a discrete and decomposed set of interrelated elements that help to achieve the task at hand. They enriched the inputs with specialised in-context prompts and fed them to codex [40], a SOTA coding LLM. The structure of the generated code was in line with the target outputs, and such an alignment was superior to those of a t5 and gpt-3 model specifically fine-tuned to the target tasks. Other recent works that focus on leveraging the programming abilities of LLMs to reason over complex or mathematical problems are given in [42, 362].

2.2 Language Model-based Agents

As decision-making processes [324] are delegated to NLM-based technology, and the applicability of these models is going beyond autoregressive language tasks [21], a justifiable concern is about the risks related to delegating ever more critical decisions to these models. NLM-based agents guide their behaviour by perceiving the surrounding environment, steering the intelligence they are equipped with to put physical, time-related, and other forms of causal reasoning in play. This section focuses on agentic AIs guided by NLMs interacting with the environment to solve tasks [313]. DRL is among the most used techniques that equip these agents with the action policies to

face complex tasks in high-dimensional scenarios. This section explores recent advancements in such a field.

Language models in text-based environments. In the context of text-based games, recent research focuses on the automatic construction of knowledge graphs to empower commonsense reasoning of DRL agents interacting with environments. Perhaps, one of the first works to use this idea in the field of text-based games was [8], where **Knowledge Graphs** (**KGs**) were extracted from natural language observations using a rule-based sophistication of Standford's OpenIE information extractor [12]. The authors of KG-DQN showed a 40% increment in convergence efficiency with the help of these internal KG representations.

Successive work in [9] focuses on efficient transfer learning over out-of-distribution games. An asynchronous actor-critic version of this KG-based agent was successively released in [7] where the knowledge graph was not only used to enhance the convergence speed, but was also used as a bias that constrained the action space exploration, permitting convergence in high-dimensional actions spaces.

Poulpart et al., instead, presented a novel method in [3], which explored a more unconstrained way to build a KG through adjacency regression of a heterogeneous graph based on the ***graph attention network*** (**GAT**) [288] model. The authors of this work used pretext tasks like reconstructive minimisation or contrastive learning for pre-training a pipeline that performs heterogeneous graph representation learning and is fed by natural language embeddings produced by a Transformer [287] architecture.

The authors of Q*BERT [10] concentrated on bottleneck states in a text-game. Ammanabrolu et al. noticed that the state space of a text-based game can be represented through a **Directed Acyclic Graph** (**DAG**) of causal dependencies, where nodes correspond to states, and edges correspond to actions featured by rewards or dependency flags that indicate which actions lead to progress in the game. The authors used a properly fine-tuned question-answering model to create KGs from natural language scene descriptions. They created a temporal abstraction or *option* mechanism, termed *modular policy chaining* to apply a *divide-and-conquer* strategy and focus the learning on how to overcome every bottleneck state separately. The authors of [10] noted that the implications of their alignment technique could help create agents that assist humans in long-term dialogue and planning over complex tasks.

The work in [237] used a hierarchical attention mechanism based on GAT networks to construct immediate and *post hoc* explanations of actions taken in the context of DRL agents in text-based games. Like Q*BERT, the authors of the ***Hierarchically Explainable Reinforcement Learning*** (**HEX-RL**) agent used template question answering to distil a discrete KG from a natural language-formatted context observation. The nodes of such a graph fit a set of prefixed types and relationships that permit the model to output human-readable explanations for the actions in the form of attention scores over symbols. These attention scores constitute an explanatory signal that underlines which are the most crucial factors for the model's immediate output. In addition, the authors tracked the trajectories inside each episode and modelled statistical analyses that allowed one to construct a causal graph for actions in sparse-rewarded trajectories.

Interestingly, to assess the causality of actions within a game-play trajectory, the authors used a SOTA language model: the actions that matched the output of the CALM model [327], which is specialised in common sense reasoning, were given higher scores. Thus, HEX-RL combines symbolic inductive biases (like prefixed categories in the graph) and AI-based feedback for filtering admissible and effective actions to build a fully explainable DL pipeline over a complex task. This may be among the first works to use a synergic combination of symbolic reasoning and perceptive AI to pursue complete human understandability of agentic AI behaviour. The authors

of *Chain-of-Thought Imitation* [322] instead leveraged the ideas in CoT to the realm of sequential decision-making. The authors of the *procedure cloning* framework advanced the state-of-the-art in complex supervised policy learning tasks by teaching ML agents not only the actions to take as a function of input but also the reason behind each decision.

Language Model-augmented Reinforcement Learning. The most important factor that hinders the widespread leveraging of reinforcement learning is the difficulty associated with proper reward design. One strategy that helps converge to efficient reward policies is intrinsic motivation [154], which enriches the reward signals to encourage exploration of unknown or atypical regions of the state-space. Intrinsic motivation alleviates deadlocks and local optima finding in most state spaces. However, intrinsic motivation fails to converge in some complex, low-represented, and high-dimensional environments. In this respect, the work in [35] proposed to represent goals with natural language and use the NLM encodings of such goals to augment the training efficiency of RL. The work in [109] successfully demonstrated that pre-trained NLMs could also give flexibility to RL agents interacting with language commands. The work of Huang et al. [117] was among the first to use prompting to decompose high-level natural-language goals into finer-grain actionable commands for agents. A follow-up work in [118] focuses on combining feedback signals -including some in natural language- to refine strategies and converge in challenging tasks.

The main advantage raised from using NLMs for tasks that require human-like commonsense is the extensive knowledge and reasoning patterns encoded in their parameters. As such reasoning might not be sufficient to reach fine-grain planning in many specific scenarios, some works used distinct fine-tuning phases for the NLMs to reason inside task-specific distributions [218, 329]. Such a fine-tuning, though, could be computationally expensive when the dimensions of the language models grow and might also reduce the generalisation capabilities encoded in pre-trained NLMs. For this reason, the authors of the RL-ADAPTER framework [346] provided guidance on how to fine-tune lightweight NLMs to translate the outcomes of RL training stage into effective modelling feedback requests for more large NLMs. The EUREKA framework [198] followed this research insight and proposed a generalist NLM-guided reward modelling and programming framework that may surpass the current reward crafting efficiency of many human experts. Other recent works in NLM-empowered RL are not only restricted to open-world games [47, 69] but also are explored in the developmental AI research of autotelic agents, i.e., agents that learn to craft and optimise goals in an open-ended environment autonomously [54, 55].

Gradient-free Reinforcement Learning. A gradient-free version of RL is currently emerging. NLMs-based agents store feedback in language and retrieve such an experience to condition their future behaviour. The REFLEXION framework presented in [268] leveraged black-box self-improvement of action policies by collecting experience in the form of natural language and online retrieving such experiences. The authors of [233] concentrated instead on constructing simulations of human behaviour by orchestrating multiple NLM-based agents equipped with *reflection* modules that collect experiences, synthesise them, and use those syntheses to guide future planning. The authors of the ***Experiential Learning*** (**EXPEL**) framework presented in [350] focused instead on a multi-task experience memory buffer. In a gradient-free training phase, the agents try to solve heterogeneous tasks. While solving these training tasks, a non-parametric knowledge base is initialised and filled with natural language insights derived from successfully solved tasks. Such a knowledge base is used during test time as an experience buffer to augment the efficiency in out-of-distribution tasks. Recently, the REACT framework [328] has demonstrated to augment the effectiveness of NLM-based agents by crafting a synergic augmentation of the state space, including reasoning-oriented and acting-oriented language prompts.

2.3 Benchmarks

Text-based environments. Authors in [102] released JERICHO, an open-source environment for interactive fiction games where DRL agents can be trained using an Open-AI Gym-like interface [30]. The authors of this environment provided options that reduced the complexity of the RL agent interaction task. One of these handicaps consists, for example, of giving the agent more awareness of the current state of objects that the agent can interact with or an in-design restriction of the action space to contemplate only those admissible actions.

The *Light* benchmark [286] was released in 2019 as a dataset of natural language descriptions of multiple types of featured entities, such as locations, objects, and characters. More than 10k interaction scenarios are contained in this dataset. Although not oriented toward training optimal decision-making, this open-source material has been used to augment the capacity of language models to improve their grounded conversational capabilities, i.e., producing relevant responses given a context.

The authors of [327] publicly released a dataset that contains the transcripts of 426 human game-plays from 590 different text-based games. They also open-sourced a method for reducing the intractable open-endedness of the action space that any NLM might search over to output the actions at any given state. They combined DRL–based action evaluation with a discrete and compact action generator and proved to generalise admissible action generation over games unseen during training.

The PIQA benchmark has been released in [27] and is focused on assessing the reasoning capabilities of LMs about physical commonsense. This dataset contains 21 K samples as goal-solution pairs with three prompts in each sample. The first indicates a desired goal, and the latter two are candidate solutions, where only one is correct. Classifying right and incorrect solutions requires physical commonsense knowledge, where humans achieve a 95% accuracy in the evaluation set. At the same time, the best model in the current leaderboard corresponds to that of the UNICORN model, which reached 90% accuracy by 2022, according to the official leaderboard of the Allen AI Institute.[1]

The ***Symbolic Interactive Language Grounding*** (**SILG**) benchmark [358] aims at assessing the capabilities of language-conditioned agents generalising across multiple environments where there are variances in the observation and action spaces. The authors combined language-based grid worlds such as RTFM [359], MESSENGER [100] and NETHACK [151].

More recently, the AGENTBENCH benchmark [183] has been publicly released. This benchmark combines multiple environments on which NLM-based agents can be tested in various decision-making tasks that involve knowledge acquisition and grounded reasoning, among others. The experiments done by the authors of AGENTBENCH revealed that closed-source NLMs outperform their open-source counterparts. A similar benchmark is SMARTPLAY [312], which uses games with a language description interface to test language-based agents across several reasoning dimensions, including spatio-temporal reasoning and probability event estimation.

Abstract Reasoning. The SCORE framework in [186] helps evaluate the self-contradictory reasoning of NLMs in question-answering tasks. Namely, the authors of SCORE aim at detecting situations in which the model declares to use incorrect or inconsistent reasoning to reach a correct answer or those in which the declared reasoning steps are correct, but the answer is not. The framework asks NLMs to generate the underlying reasoning for each provided solution and uses such explanation to evaluate self-consistency. The framework pushes NLMs to enrich the output by asking to answer questions from different points of view.

[1] https://leaderboard.allenai.org/rainbow/submission/ccppv1dqp48v5467jos0

The CoFe test suite presented in [11] helps assess the compositional generalisation of NLMs -which is the ability to understand new combinations of known primitives- on in-context learning scenarios. Instead, the authors of the Math Dataset [106] focus on mathematical reasoning and propose a set of 12,5 K competition-level mathematical problems. The authors of this work also provided a pre-training dataset to teach models the mathematical fundamentals. Similarly, the Gsm8k benchmark [52] presents 8.5 K grade school mathematical problems with a set of corresponding natural language solutions. According to the *paperswithcode* blog,[2] the Gpt-4 DUP [357] is the current champion of the Gsm8k benchmarks with an accuracy of 0.97. Another technique for abstract reasoning assessments of language models that involves coding is available at [40].

The ReClor [334] and the LogiQA [178] benchmarks help evaluate the logical reasoning capabilities of language models. A model must choose the correct answer from multiple candidates in these benchmarks. Instead, the folio benchmark in [99] can test logical reasoning in isolation by connecting premises with multiple conclusions requiring first-order logic. More recently, the LogiGlue benchmark gathered more than 20 datasets related to logical reasoning to test deductive, abductive, and inductive reasoning generative language models [192].

Compositional generalisation benchmarks, which instead involve understanding novel combinations of known priors, are instead presented in the scan [155] and the **Compositional Freebase Questions** (**cfq**) [143] benchmarks. The former involves reasoning over combinations of command primitives like *jump*, *turn*, and compositional rules, while the latter is focused on named entity parsing and consecutive query construction. Interestingly, the authors of cfq provided a framework for generating new datasets to test the compositional generalisation of AI systems. In this framework, dubbed *distribution-based compositionality assessment*, mathematical instruments permit the creation of fair train-test splits, where the resulting train and test portions contain a similar distribution of primitives. The datasets generated are also compositionally challenging in that the divergence between the distribution of combinations in each dataset is maximised. A similar work is that of the Clutrr benchmark [271], which is focused on inductive reasoning from the text. The sygns suite [318] also permits the assessment of the compositional generalisation capacity of NLMs in terms of combinations of logical expressions.

More recent benchmarks for compositional generalisation are available in ProntoQA [260] and its recent extension, ProntoQA-ood, [261], where the models need to generalise deductive reasoning and combine a given set of deduction rules to produce out-of-distribution proofs at test time. The Gscan benchmark in [255] also focuses on compositional generalisation in situated language understanding.

Natural Language Understanding. The ***General Language Understanding Evaluation*** (**glue**) benchmark in [291] was presented as a compilation of nine tasks related to question-answering and entailment detection, among others. Successive work in [290] augmented the difficulty of these tasks, and the contributions of [262, 356] are also focused on **natural language understanding** (**NLU**) but more specifically meant to assess the few-shot learning capabilities. Remarkably, the work in [262] also offered a study on techniques that augment the NLU performance of small-size NLMs. Finally, the glue-x benchmark in [321] carefully crafts the train test split to evaluate the OOD generalisation of NLMs.

A benchmark for analytical reasoning of texts is available on [360], where the authors demonstrated that explicitly searching for symbolic knowledge in the form of subjects and objects of phrases enhances the performance.

[2] https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k

The recently released BoardGameQA Benchmark in [141] instead focuses on the ability to reason in the presence of contradictory information by imposing preferences over different information sources. The difficulty of the samples in this benchmark has different levels and can be regulated.

Other. The Wordcraft environment was released in [132]; in this text-based environment, agents can use external sources of information to augment the set of prior knowledge that helps policy convergence. The discovery of new entities through interaction is the main focus of this environment. A similar text-world interactive environment has been released ScienceWorld [297] where the focus is on scientific reasoning at the elementary level. The authors of ScienceWorld were motivated by NLMs' breakthroughs in scientific question-answering benchmarks such as the one in [51] and tested the hypothesis of whether those breakthroughs were merit of grounded understanding of subjects beyond lexical correlation maximisation. Extensive parametric variations of a broad range of typical exam questions from a basic science curriculum are the main building blocks of ScienceWorld. The experiment carried out by the authors revealed that small-size NLMs with 1.5 M parameters based on the Deep Reinforcement Relevance Network [104] achieved superior performance on the held-out test set compared to medium-size NLMs with 11B parameters.

The Choice-75 benchmark in [112] focuses on multi-branch script learning, -where a script is intended as a sequence of punctual event mentioning- assessing the ability of language models to predict the next steps of a script given the previous steps. Another script generation and the sorting benchmark is available in [257]. The toolQa benchmark in [368] helps assess the correct information retrieval capabilities of NLMs. The authors of toolQa minimised the overlap between the information required to answer the questions in their benchmark and the information that is plausibly among the pre-training data of most NLMs.

The ExplaGraphs benchmark in [256] instead focuses on explaining correct predictions in generative NLMs. For each tuple containing a belief and a statement, the model has to predict whether the latter supports or counters the former. The NLM is also meant to generate a discrete graph that explains the rationale of the prediction, and the similarity with gound-truth explanations graphs is related to the final score. The leaderboard shows that human performance is vastly superior concerning NLMs. However, by November 2023, the proprietary large-size NLMs like gpt4 have not been tested on this benchmark. The work in [39] presents the ***knowledge-intensive analogical reasoning*** (**e-kar**) benchmark, which focuses on the ability to find and extrapolate the underlying structure of relations between concepts among different domains to identify analogies. This benchmark collects more than 1.5K problems gathered from public exams, which require commonsense reasoning, linguistic understanding, and encyclopedic knowledge to be solved. The authors of the ***machine world learning*** (**mewl**) benchmark in [129] note how word learning in humans is made on a few-shot basis and in an open-ended manner and create a framework to assess this kind of concept generalisation. Notably, their benchmark can be used with only language and vision-language models. The authors of MT-bench [354] released a low-scale multi-turn question benchmark to evaluate the usefulness of conversational NLMs across various tasks that require reasoning. Remarkably, the authors performed experiments in which human evaluators and gpt-4 [226] as a judge model reached an 80% correlation in the evaluation.

2.4 Summary and Research Advices

NLM-based tools and agents are more helpful each time by improving the depth and complexity of the inherent reasoning abilities. However, complex reasoning may be costly in terms of model size. For this reason, the research community has put forth numerous open-source knowledge

distillation strategies, such as *Distilling step-by-step* [113], to import such capabilities into modest-sized and open-source models. In this respect, another interesting open-source instrument that obtains good results on a set of the afore-mentioned benchmarks is Mistral 7B [128], a small-size NLM, which incorporates sophisticated attention mechanisms that enhance inference performance and permit larger context sizes. Another small-size (13B) model carefully trained to excel at reasoning tasks is Microsoft's orca2 [213]. A medium-size alternative is instead represented by the Tülu2 model family [122], which is a family of 70B open-source, comprehensively fine-tuned language model that reaches a mean of 85%–90% of the gpt4's accuracy in lots of the previously mentioned benchmarks. If small-size models are the target of research, Microsoft has released the Phi 1 [96], Phi 1.5 [171], and Phi 3 [2] models that exhibit good performance on coding, commonsense and mathematical reasoning with less than 4B parameters.

After Meta's llama models were open-sourced, researchers at Standford created a 52 K IFT dataset using the self-instruct paradigm [299] with gpt3.5. They then fine-tuned the 7B llama model to create the alpaca model [280], the first open-source small-size instruction-following NLM. However, the recent *long is more for alignment* framework in [351] shows that selecting the 1 K longer-response tuples in the alpaca IFT dataset may result in more effective and efficient fine-tuning.

Interpretability is another essential feature that might validate and democratise the usage of NLM reasoning. However, it is not easy to obtain a rationale that explains the outputs of a language model that is simultaneously discrete (i.e., human-readable) and faithful in that the stated response rationales are the real ones. In this respect, from the side of NLM-based tools, *Faithful CoT* [195] stands out as an interpretable tool by design, being coupled with a deterministic solver. In the realm of NLM-based agents, instead, the work of Peng et al. [237] offers a neat strategy to obtain rationales and even potential causal graphs for the agents' actions. However, it might be limited in the vocabulary with which such rationales are formed.

In the field of mathematical reasoning, **Monte Carlo Tree Search** (**MCTS**) mechanism was coupled with the self-refine framework [199] to create the **MCT Self-Refine** (**MCTSr**) algorithm [340]. By iterating over self-refinements and self-evaluations of answers, the authors of MCTSr proved to beat closed-source models such as Claude 3 Opus and Gemini 1.5 on Olympiad-level mathematical benchmarks using Llama 3-8B. By looking at this and recent similar works [284], one can say that there has been wide progress in the field of neural language model-based mathematical reasoning. Still, research efforts should be made to fill the vast room for improvement that continues to exist at times. Table 2 contains pointers to the most remarkable open-source instruments that can help kick-start research on helpful NLMs.

Notably, the MCTSr framework has been open-sourced. The authors of this work noted that a limitation of the current version of their framework is the strict focus on mathematical reasoning tasks. Interesting research paths include investigating its applicability for other types of assistance tasks beyond mathematical reasoning that require strategic planning. Other research opportunities include testing these new methodologies on the most recent mathcheck benchmark [364] that helps assessing genuine mathematical reasoning abilities in terms of testing task generalisation and reasoning robustness beyond mere problem-solving. To this respect, it is worth mentioning that the Tulip agent framework [223] offers and open-source framework that can be exploited for investigating on robust reasoning of modest-sized tool-augmented NLMs.

3 Harmlessness

Manipulation and harm at the individual or societal level are among the most severe risks of human-misaligned NLMs [61]. The research community is putting forward lots of techniques to augment the safety or reduce the potentially harmful, biased, toxic or nefarious responses of

Table 2. Remarkable Open-Source Instruments Focused on Helpful NLM

Source	Year	Description	Advantages	Weaknesses
TÜLU2 [122]	2023	Comprehensive fine-tuning on a curated selection of IFT datasets	Current open-source SOTA across multiple reasoning benchmarks	Medium-size (70 B parameters)
MISTRAL [128]	2023	7B open-source model with good position on multiple reasoning benchmarks	Small, Efficient inference speed and large context window	Outperformed by proprietary medium-size models.
ALPACA [280]	2023	Open-source High-quality IFT dataset and 7B model	According to [351] (2024), can select the 1 K longer-response tuples to achieve better alignment.	Inherits licencing and terms-of-use from LLAMA.
PHI [2]	2023	Small models with good performance on coding, math and commonsense reasoning	Transformer-based models with less than 4B parameters	Not honestness-aligned
COMPASSMTL [348]	2022	Multi-task friendly pre-training	Accelerates inter-task transfer learning	Requires task-specific prefixes.
FAITHFUL COT [195]	2023	Improving CoT outcomes through symbolic language generation	Augments overall correctness, interpretability by design.	Requires a code interpreter, restricted to queries solvable by programming.

language models. This section reviews recent advancements and benchmarks for aligning NLMs with safety criteria.

3.1 Methods

Specialised Prompting and Fine-tuning. Prompt engineering and instruction fine-tuning are among the most used approaches for the safety alignment of NLMs. The work in [135] offers a study on the zero-shot capacity of LLMs to align with the flexibility of human moral judgement. Zhijing et al. used the cognitive mechanisms of *Contractualist moral decision-making* [16, 161] to construct a model of human moral judgement. In contractualism, any judgement of morality considers the extent to which breaking a moral imperative may break the final function for which such imperative stands. Apart from considering the final goal of a moral claim, contractualism holds that human judgements of morality also analyse the utility of any act that breaks a rule for all the subjects eventually involved. Therefore, the authors of [135] presented a prompt engineering framework for asking judgements to an NLM on multiple situations where rule-breaking may be potentially permissive. Their approach has shown to be more aligned with a set of human decisions concerning the answers of SOTA morally aligned NLMs. The prompting strategy in [135] involved a recursive question answering. The extent to which the function of a goal is broken and how much the utility of the subjects involved is reduced is asked to the model. Lastly, a question about whether the rule should be broken in the particular case is made to the model. Authors termed such a prompting strategy as ***Moral Chain-of-Thought*** (**MORALCOT**).

From the side of IFT, the authors of BREAVERTAILS [125] provide a dataset of more than 300 K question-answer pairs where human annotators separately rate the helpfulness and harmlessness of answers, permitting to disentangle the learning of these aspects by an NLM. Notably, the annotations also provide a self-confidence score that can be useful in managing conflicting annotations. Other recent datasets for IFT on safety criteria such as [119, 301] are open-sourced by the community both as a resource for alignment and as benchmarks for evaluating the harmlessness. Such benchmarks will be carefully reviewed at the end of this section.

Red teaming. When speaking about safety in language models, *Red Teaming* or *jailbreaking* are the equivalent terms for *penetration testing* in cyber-security or *Adversarial Training* in Deep Learning in general. Red Teaming uses adversarially crafted prompts and methodologies to induce language models to output harmful responses, with the scope of improving safety enforcement policies at inference time [81]. Although recent works have demonstrated that it is possible to use NLMs to scale human-based read teaming [238], it is not clear the extent to which generative AI can replace human oversight in value-alignment tasks. In this respect, the work of Shi et al. [266] elaborated adversarial strategies targeting the harmful detection language models. Recent research in [156, 371] discovered prompting strategies that break the safety alignment of current SOTA generative NLMs. Currently deployed NLMs such as Llama2Chat [283] and Claude [19] declare to have used manual red teaming to robustify their production-ready NLMs against subtle malicious prompts. To this respect, the work of Ge et al. [85] presents an iterative fine-tuning framework in which both the red-teaming and the target NLM are adversarially trained to maximise their winning rate. A similar setting is offered in [196], where a game-theoretic approach is modelled between red-team and blue-team language models to detect and reduce vulnerabilities in NLMs.

Reinforcement Learning from Human feedback. The work in [49] formalised the paradigm of RLHF in the context of NLM alignment. RLHF aims at aligning the output of NLMs with a high-level task-agnostic goal encoded in low-scale human feedback. This goal can be related to various aspects like factuality, fluency, and others. One of the most pursued goals RLHF has been leveraged for is safety, i.e., reducing harmful, offensive, toxic, or biased responses. In RLHF, a *preference model* scales such feedback by learning to rate the outputs according to a low-scale training set of human feedback. By doing so, human preference criteria can be extensively applied to a massive pair of NLMs' outcomes and lead to the convergence of a preference policy in the context of DRL.

Follow-up works have also demonstrated the effectiveness of RLHF [228, 369]. RLHF has been the de-facto standard for safety-alignment of NLMs and has been adopted in SOTA LLMS such as Open AI's GPT-4 [226], Anthropic's Claude [13], Google's Gemini Pro [89], and Meta's Llama 3 [70]. One of the drawbacks of RLHF that the authors of [18] focused on was the minimisation of evasive responses that tended to characterise the models after alignment. In particular, human feedback was used to label the harmlessness and usefulness of the responses. This second type of label was used to shape a corresponding reward term in the RL phase. By shaping the reward in this mode, the authors avoided converging to a useless model that tends to skip responding prompts to maximise harmlessness scrupulously. Lots of techniques have improved upon vanilla RLHF to address its main limitations, such as training inefficiency [182], the possibility of reward hacking phenomena [176], the imbalance between competing objectives [58], among others [17, 88, 315]. A remarkable upgrade of the original RLHF algorithm, which is based on ***Proximal Policy Optimisation*** (**PPO**) [263], is that of ReMax [173], where a variant of the Reinforce [309] RL algorithm is used. By doing so, the computational and memory footprint of training is reduced compared to the PPO-based RLHF scheme without sacrificing performance. The interested reader can refer to recent literature for a more in-depth presentation of the working principles [355], primary potentialities [139], and limitations [33] of RLHF. Recently, the **Direct Preference Optimization** (**DPO**) [242] paradigm introduced a closed-form way to align NLMs without relying on a reward model: the authors of DPO used a preference dataset to maximise the alignment directly by modelling a classification problem on the training data.

Reinforcement Learning from AI Feedback. Further efforts to scale human oversight in alignment tasks involved NLM-based preference labelling. A remarkable example of this insight is the ***Constitutional AI*** (**CAI**) blueprint in [19], which increased the efficiency of human feedback

concerning vanilla RLHF. More specifically, the authors in [19] used a *constitution* made of natural language preference guidelines with order-ten principles. A synthetic dataset containing query-response-critique tuples related to moral alignment is created from such a constitution. Such a dataset was used to train a preference model, and the latter model was used to label the harmfulness of the responses in the RL phase. By doing so, the authors of CAI provided a more data-efficient framework for the human-guided harmlessness alignment of NLMs. Analogously to RLHF, the authors of CAI coined the term ***Reinforcement Learning from AI Feedback*** (**RLAIF**) as a generalisation of the CAI blueprint, in which language models are used to scale human oversight. The authors of CAI noted the advantage of using Chain-of-Thought prompting strategies for improving the human assessment of their final results. The effectiveness of this fine-grained feedback type has also been confirmed by other recent works [97, 180]. In this vein, some recent research showed through experiments that, provided careful modelling [246], some proprietary [87] and open source [5] SOTA LLMs can qualitatively outperform human-feedback under some circumstances in value alignment supervision tasks.

A follow-up work concerning CAI was published in [150], where the researchers experimented on the hypothesis of performing a similar aligning strategy. However, this work seeks to converge to proper reward criteria from a single and more universal human-made moral claim. Beyond CAI, other recent works related to RLAIF are presented in [147, 157, 251, 320]. The Self-Instruct paradigm in [278] can also be seen as a sophistication of RLAIF that has also been used to train open-source LLMs such as Standford's Alpaca [280].

Debate. The work in [182] introduces a new paradigm called *Stable Alignment* for aligning LLMs with social values. Liu et al. simulate and record social interactions between NLMs. The authors of this work created a simulation sandbox that runs a society of LLMs that simulate human interactions about morality questions, responses, and feedback cycles. They first instantiated a hundred language agents and started 169K iterative prompt-response-feedback interactions sampling controversial topics from the Anthropic's RLHF dataset.[3] At the end of this process, they obtained large-scale and effective supervisory signals through generated natural language. The authors of the Stable Alignment framework demonstrated the effectiveness of their approach in terms of scalability. Regarding RLHF, there is no need for a reward model. Thus, the training can be parallelised and the simulation can be run offline.

The experiments in [182] involved SOTA LLMS at the beginning of 2023, such as text-davinci-002 (175B), text-davinci-003 (175B), and GPT-4 (of unknown size). They tested the fine-tuned model on various benchmarks such as mic [370], Ethics [105], TruthfulQA [175], Moral Stories [74], and Anthrophic HH [18], showing outstanding results without being trained in any of these data (except for the Anthrophic HH benchmark). The *Reinforcement learning for feedback framework* [78] concentrates instead on augmenting the effective of the critique generator language model. Other related works in which the output from one NLM is used as a black-box alignment criterion for another target NLM are present in [32, 34]. To this respect, recent research in [95] focused on improving the alignment of small open-source NLMs by using the outputs of proprietary LLMs and showed that although stylistic behaviour is well-learned from small NLMs, acquiring the factuality encoded in the parameters of teacher models may still be an open challenge.

Other. The work in [44] trains a DL module to modify human prompts to maximise the alignment of a target NLM; this work can be seen as a recent approach for automatic prompt-engineering [267]. The authors of [97] focus instead on fine-grain supervisory signals in the form of revisions to responses to improve the overall performance of an IFT dataset. Recent work in

[3] https://github.com/anthropics/hh-rlhf

[220] poses itself at the intersection of neuroscience and AI by crafting parametric modifications of structural samples in morality judgement scenarios. By doing so, the authors of this work seek to evidence the difference between human and NLM-based judgement tendencies. The authors of the (MOCA) framework released their dataset and advocated for carefully curated datasets for alignment. Instead, the TOXIGEN framework presented in [101] helps create adversarial examples with subtle hate speech to improve the ability of hate-speech detection in NLMs. They made a synthetic dataset of 274 K toxic/benign samples and showed that fine-tuning NLMs using this dataset improves performance over several moral alignment benchmarks. The authors of the *Direct Representation Optimisation* framework [353] are devoted to optimising the safety prompts from the perspective of the latent space, i.e., freezing the model's representations for the queries except for the prompt tokens. They then optimise these tokens using gradient descent to push the whole prompt + query representation to regions of the latent space that augment/decrease the refusal behaviour on harmful/harmless queries.

3.2 Benchmarks

The seminal work in [105] proposes the ETHICS dataset that contains over 130 K natural language descriptions of multiple scenarios that permit fine-tuning an NLM towards responses that align better with justice, deontology, virtue ethics, utilitarianism, and commonsense moral judgements. The authors claim to be the first work that frames the ethical alignment of algorithmic agents without an *a priori* mathematical formulation. The authors note that approaches based on the latter may have practical applicability only in very narrow and task-specific ethical-alignment use cases. By presenting everyday life examples, authors argue that their dataset may reflect a more comprehensive set of human alignment criteria concerning other formulations like value alignment, fairness frameworks, and constitutional-like specification of alignment requisites for an AI [19].

The work in [107] provided an open-access set of 35 text-based adventure games where the objectives of the game may be in some situations in conflict with moral principles. Remarkably, the authors of the JIMINY environment designed a regulatory term that shapes the policy towards more ethical behaviours by penalising instances of lousy behaviour directly during the computation of the action-state value function (Q-value). Authors called this technique ***Commonsense Morality Policy Shaping*** (**CMPS**) and used a model trained on the ETHICS dataset in [105] to predict the morality of actions and thus shape the Q-value of actions. They pointed out future research directions that could improve the effectiveness of their technique, including using proper reward shaping.

The MACHIAVELLI benchmark of Pan et al. [231] contains over a half-million scenarios where NLM-based agents can be morally assessed while playing 134 text-based *chose-your-own-adventure* games. These scenarios may include conflicting contexts in which maximising the game reward could drive the agent toward unethical behaviours; the authors demonstrate that focusing only on a reward may result in machiavellianism. They also conduct experiments with moral prompt fine-tuning of NLMs and an artificial conscience policy regularisation in DRL agents to improve morality in agents. Notably, the authors open-sourced their code to incentivise and facilitate further research. Great importance is given to the sequential decision-making capability that an agent needs to perform to play the scenarios of MACHIAVELLI, posing an important step-stone towards the safe deployment of generalist agents. Another benchmarking environment for text-based interaction of agents is TEXTWORLD [56], where the authors offer a game-generation tool that permits the control of the degree and axes of variation of the novel environment to construct case-specific datasets and permits the evaluation of specific types of generalisation of dialogue systems.

The work in [217] presents the *Goofus & Gallant* benchmark, which contains 1.3 K sentences encoding normative/non-normative behaviour in the context of a homonymous cartoon in which one subject is always performing normative behaviours and the other is escaping from such normativity. The names of characters were masked to avoid learning them as discriminative features. Authors assessed if pre-training a morality predictor on this benchmark could help to enhance the pre-trained classification models based on BERT [142], and Deep Pyramid Convolutional Neural networks [136]. To assess the effectiveness of morality transfer learning, they created two other labelled datasets of similar dimensions using crowd workers. Their results did not find significant morality generalisation on out-of-distribution contexts, but transfer learning notably enhanced in-distribution data efficiency.

The SOCIAL-CHEM-101 benchmark in [79] gathered approximately a set of 0.3 M natural language asserts that encode everyday common sense social norms. Each of these phrases, dubbed rules-of-thumb by the authors, was enriched by eliciting a dozen refinements and short explanations or clarifications pointing to social agreement, legal aspects, and theoretical moral foundations. The authors of [74] created the MORAL STORIES dataset, a corpus of 12 K scenarios where a subject, an intention, a moral norm, and a pair of action-consequences are given as sentences. One action-consequence is aligned to the norm, and the other is misaligned. This benchmark was created using crowd-sourced workers, and the purpose is to assess grounded, goal-oriented, and morally aligned reasoning of natural language understanding and generation models.

The fine-tuned classifier augmented their accuracy by ingesting more information in input, like the subjects' intentions and the consequences of actions in the story. Emelin et al. also fine-tuned generative NLMs such as BART [162] on their corpus to assess their capacity to predict morally aligned actions, potential consequences of actions, and moral norms. Notably, the models were biased to perform better in generating morally aligned actions and the consequences of these types of actions, denoting a positive morality bias in the pre-training corpus of the used LLMs. Human-expert evaluation metrics assessed all their experiments.

The ***Social Bias Inference Corpus*** (**SBIC**) was released in [259], where the focus is to evaluate the capacity of neural language tools to explicitly identify the implications of social biases in natural language utterances. The authors of this benchmark first created a formalism termed social bias frames that seeks to help evidence the targeted groups, the offensive intentions, and the specific implications of social biases in natural language *posts*. With the help of this formalism, Sap et al. gathered categorical and free-text annotations to explain the social biases in a corpus of 140 K potentially biased texts scrapped from known socially biased public sources. The authors of SBIC fine-tuned GPT models to classify targeted groups and intentions of posts and also to generate the implications of these.

The SCRUPLES benchmark was introduced in [189] and contains 0.6 M heterogeneous ethical judgements about 32 K anecdotes described in natural language scrapped from Reddit forums. The authors of this benchmark noticed they left wide room for optimisation of their benchmark, primarily because of the inherent heterogeneity in human moral judgements. However, N. Lourie et al. shed light on how to separate by design such an intrinsic uncertainty about human judgements from the uncertainty that is inherent to the probabilistic nature of neural inference mechanisms. The authors of this work created a Bayesian predictor that maximises the likelihood of the parameters of a Dirichlet prior over a Multinomial distribution encoding the distribution of judgements over anecdotes. By doing so, the model is taught to choose more than one label or judgement per sample, mimicking a majority that would not exist in gold annotations, but also to output the likelihood distribution of judgements that a theoretically infinite number of annotators would give.

The work in [130] offers 1.7 M moral-alignment human judgements of a heterogeneous set of everyday situations to fine-tune an NLM. They created the DELPHI model, which is an instruction fine-tuned version of the T5-Large NLM [249] trained on the UNICORN commonsense reasoning benchmark [188]. The authors open-sourced a compilation of five human-morality alignment datasets: SOCIAL-CHEM-101, MORAL STORIES, SBIC, SCRUPLES, and ETHICS. The authors of DELPHI thus follow a bottom-up approach for distilling moral judgements: they obtain moral criteria to answer online queries from crowd-sourced data on which the model is trained.

Different from many contemporary benchmarks on morality alignment, which simplify or bypass the inherent ambiguity of human-moral judgements, the ***Moral Integrity Corpus*** (**MIC**) released in [370] contains a crowd-sourced supervisory signal that shed light on the explainability and interpretability aspect about the morality assumptions and biases that may be inherently encoded in the parameters of an NLM when emitting moral judgements. The authors of MIC extrapolated 99 K moral claims or **Rules of Thumb** (**RoT**) contained in the SOCIAL-CHEM-101 benchmark and matched them with 38K short opinion prompts-reply pairs, where the prompts were obtained from the *AskReddit* forum[4] and the responses were generated using SOTA open-source dialogue LLMs at the time such as BLENDERBOT [252] and DIALOGTP [347]. The final triplets were annotated taking into account the alignment of the response to the RoT, the global consensus rate of the RoT, the violation severity degree, and the moral foundation category the RoT appeals to, following the taxonomy in [91].

Ziems et al. then trained generative NLMs to output RoTs concerning the prompt-reply pairs and evaluated these generated RoTs using automatic metrics and expert annotations concerning the fluency, relevance, and well-formedness of generated text. The experiments showed that generated RoTs with the best generative models at the time were fluent and well-formed but lacked being relevant in 20% of cases. Apart from RoT generation, the experiments in [370] involved training classifiers that output the multidimensional labels for each prompt-response-RoT triplet and achieved satisfactory accuracy, showing that their corpus might help to audit the latent moral criteria encoded in any generative NLM. Unfortunately, there seems not to be a public curated leaderboard for this benchmark, even though the repository[5] contains fine-grain instructions on how to reproduce the authors' experiments.

3.3 Summary and Research Advices

When aligning a model to moral preferences, the alignment evaluation metrics should be carefully chosen to avoid biasing the alignment evaluation toward simple lexical overlapping. For example, the experiments in [259] showed that, while the classification task may not be challenging to learn, a satisfactory generation of free-text explanations is still difficult for modern NLMs. Interestingly, the explanations tended to be correctly generated only when the desired output had lexical overlaps with the content in the post. Also, the authors of SCRUPLES [189] noted that the lexical overlap played a significant role in augmenting the accuracy of predictions.

To this respect, the more recent *Direct Preference Optimisation* framework has publicly released[6] an implementation of their pipeline and preference dataset that can be used out-of-the-box for fine-tuning purposes.[7] The DROMEDARY framework for self-alignment of NLMs without relying on excessive amounts of human feedback [278] is also an open-source alternative for alignment that has demonstrated good empirical results across various alignment benchmarks. A recent

[4] https://www.reddit.com/r/AskReddit/
[5] https://github.com/GT-SALT/mic
[6] https://github.com/eric-mitchell/direct-preference-optimization
[7] https://huggingface.co/models

open-source framework that may facilitate the value alignment of language-based DRL agents is the MACHIAVELLI benchmark [231]. This framework can be an interesting starting point for further research on safe generalist agents.

Despite recent advancements on effective alignment techniques, a fundamental challenge to alignment has been evidenced by the research community regarding the vast space of possible prompts from which effective out-of-distribution jailbreak attacks can be sampled [20]. Concerning these recent findings, an open-source framework that shows good reactive defence against jailbreak prompts is represented by the **Directed Representation Optimisation** (**DRO**) framework [353]. Through careful mechanistic analyses of the latent space representations, this work proposes to treat the safety-prompt embeddings as a continuous trainable embedding that seeks to maximise protection without sacrificing performance or falling into over-refusal models over a heterogeneous set of out-of-distribution attacks.

The research community on mechanistic interpretability of neural language models has vastly developed in the last months and is closely related to safety and ethical risk management regarding NLMs and AI, in general [26]. To this respect, many research and development resources have been open-sourced to facilitate the entrance in the field [1]. However, deepening inside the mechanistic aspects of how NLMs work can be complex from a mathematical perspective, simpler approaches to jailbreak defence such as the Safe Unlearning framework [349] have been recently developed to show that practical defences against out-of-distribution can be implemented through a more high-level design. Table 3 contains pointers to the most remarkable open-source instruments that can help kick-start research on harmless NLMs.

Another interesting research path to deal against out-of-distribution jailbreak attacks is that related to the quite novel field of immunisation [254] rather than the alignment. Immunised models are pre-trained models whose weights are hard to fine-tune toward harmful or dual tasks. To this respect, the open-source VACCINE framework [116] stands out as an immunisation strategy that explicitly addresses resistance to harmful fine-tuning attacks. Lots of interesting research opportunities exist concerning the current state-of-the-art of immunisation strategies, as pointed out by the authors of the recent Representation Noising framework [253].

4 Honestness

If offensive, biased and toxic outputs of NMLs are straightforward examples of their potential harmfulness, also innocuous hallucinations [15, 127, 245, 330] and sycophancy [152, 243, 264] can have harmful consequences depending on the application context [61]. In this respect, this section overviews the most significant and recent works focused on the honestness alignment of NLMs. Following the definition of honestness in [14], that essentially requires a model to be *aware* and be able to *express* its own level of knowledge and uncertainties. In this respect, this survey uses the words *factuality* and *honestness* indistinctively, though there might be considerations under which these features are not equivalent.

4.1 Methods

Black-box techniques. Black-box techniques for factuality verification and enhancement do not use external knowledge sources. These methods rely only upon the parametric knowledge acquired from the pre-training corpus and seek to minimise hallucinations and self-contradictions using strategies that enhance the confidence level and systematicity of answers. The recent work of J. Luo et al. [191], for example, aims at preventing NLMs from hallucinating, proposing the SELF-FAMILIARITY decoding framework, where a multi-step process is carried out to ensure the model does not answer queries that refer to unfamiliar domains. The model extracts concepts from the question through conventional **Named Entity Recognition** (**NER**) techniques. Then,

Table 3. Remarkable Open-Source Instruments Focused on Harmless NLM

Source	Year	Description	Advantages	Weaknesses
TÜLU2 [122]	2023	Comprehensive fine-tuning on a curated selection of IFT datasets	Current open-source SOTA across multiple reasoning benchmarks	Medium-size (70B parameters)
DPO [242]	2024	End-to-end alignment fine-tuning from prefence dataset.	Does not require a preference model, nor sampling answers.	Requires a reference dataset.
SELF-ALIGN [278]	2023	Safety-alignment of NLMs without relying on human preference data.	Minimizes the cost of human oversight for safety alignment (a constitution of two or three hundred norms is enough).	Might be strongly coupled to the validity of the constitution.
ALIGNER [123]	2024	Safety-alignment module for fine-tuning NLMs	Model agnostic, parameter efficient, shows weak-to-strong generalization	Augments computational requirements during inference
BEAVERTAILS [124]	2024	Safety-alignment comprehensive dataset and benchmark	Helps to reduce the risk of fourteen types of harmful responses.	Authors released only one 7B fine-tuned model in [239]
SALAD-BENCH [164]	2024	Safety-alignment hierarchical comprehensive benchmark	Cotains heterogeneous and Adversarial samples	Evaluation might be biased under some circumstances (NLM-based evaluation)

a familiarity score concerning the query is computed following a recursive querying process: First, the model is asked to describe the concept, and then, the model is asked to predict the same concept from such a description. The familiarity score of each concept is a function of the entailment between these two intermediate outputs. Finally, the query familiarity score averages the individual concepts' familiarity scores. Though redundant, the confidence scores obtained in this manner prove helpful in the experiments carried out by the authors. Also, the authors of [216] avoid outsourcing external knowledge and create several prompting strategies to detect self-contradictions and eliminate contradictory information while maximising the output information.

The work in [352] presented a prompting-based method for detecting the knowledge gaps in NLMs by noticing that NLM responses to a unique query show inconsistencies when using different verbalisations of the question. In the context of multiple-choice question answering, the work in [133] proposes to use an ensemble of NLMs to augment the prompt diversity, shedding light on a more robust fine-tuning strategy concerning the sensitivity to prompt variation in NLMs.

A recent common practice in factuality evaluation is to scale human oversight by delegating the assessment of NLMs to other specialised NLMs. Exemplary works in this vein are the ALPACAFARM framework in [71], where a set of NLM-based evaluations are meant to mimic inter-annotator variance in human feedback and the ALPACAEVAL [169] tool, which enables automatic evaluation of longer output sequences. Also, GPTEVAL [185] and SELF-CHECK-GPT [205] are useful frameworks to scale human oversight using AI. In these works, the focus is not only to scale the human oversight but also enforce previous automatic metrics for generated texts that show low correlation with human evaluation, such as BLEU [232] and ROUGE [174]. The works in [208] also use the NLMs to revise their answers in the context of CoT answering. A similar paradigm of progressive self-verification is presented in the ***Chain-of-Verification*** (**COVE**) framework [63]. Other recent works that use black-box self-checking of factuality are proposed in [138, 149, 345].

Augmented Instruction Fine Tuning. Beyond vanilla PE techniques [134, 361], instruction fine-tuning is one of the most efficient methods for mitigating hallucinations and sycophancy in NLMs with specialised instructions. For example, the authors of [306] present a prompt-guided strategy to create an IFT dataset and reduce sycophancy in LLMs. Similarly, the *Reflection-Tuning* framework in [165] uses highly factual NLMs to filter out low-quality IFT samples from a given dataset to enhance the overall factuality after fine-tuning with the filtered dataset. Other works instead use supervised fine-tuning where the fine-tuning process focuses on predicting masked factual information given a context [98, 250].

The authors of the *Curriculum Instruction Fine-tuning* (CITING) framework in [77] combine standard prompt-response pairs with a small set of revision criteria. Such criteria are generated by an expert NLM after clustering the queries in the initial dataset. After a canonical IFT stage, additional iterative fine-tuning is carried out using new questions and responses revisited by the expert NLM. The *student NLMs* can also refine their answers using the generated criteria. Interestingly, this technique has shown promising results beating RLHF in the RAFT benchmark. Other works using synthetic IFT data are available in [110, 202]. The work in [114] instead focuses on the efficiency of the fine-tuning task and is the first to propose adapter-efficient fine-tuning modules to selectively remove the hallucination tendency in LLMs.

In the *refusal-aware instruction tuning framework*, (R-TUNING) [341], any instruction tuning dataset is fed to a model in inference mode to assess the level of certainty of the model before the fine-tuning. Analysing the accuracy of predictions, the instruction dataset can be modified, prepending a refusal response on the difficult questions for the model to answer. After creating refusal-aware instruction data, the IFT process is taken out and the resulting model behaviour shows to be aware of its knowledge limits instead of hallucinating to respond to every question.

Tool-based mitigation techniques. Other techniques to mitigate hallucinations and incorrect answers instead focus on ***Retrieval Augmented*** **(RAG)** models, i.e., models with access to external knowledge sources. These techniques, in general, consist of querying faithful and updated knowledge sources at training or inference time to contrast or back up the outputs of NLMs. The ***Rethinking with Retrieval*** **(RR)** framework [103] builds precisely on this intuition and conjugates the CoT prompting strategy with the effective use of external knowledge sources to mitigate incorrect NLM responses. At inference time, the authors of RR propose to use a pool of answers for each query. A retrieval phase from the knowledge source successfully seeks to back up each response in the collection. The RR framework then chooses the response that maximises the accordance with the retrieved proofs. This framework was evaluated with the StrategyQA benchmark in [86], and the external knowledge source was based on ConceptNetV5.5.

Another work that builds on the CoT prompting strategy is CRITIC [90], which focuses on self-correction. The NLMs' responses -including potential hallucinations- are revisited after external feedback is obtained from knowledge sources. The work of S. Zhang et al. [343] instead notices that, even when faithful external knowledge sources are available, NLMs could produce hallucinations because of the query format not aligning with the knowledge format. The authors of this work thus introduced the MixAlign framework to automatically ask for clarifications from the user in case of detected uncertainty in the query-knowledge mapping process. Other recent works that use tool-augmented retrieval and check are presented in [45, 82, 140, 323], and a framework for the evaluation of factuality in retrieval-augmented NLMs has been presented in [75]. However, research has put in evidence that retrieval-augmented NLMs may still have a large room for improvement because the knowledge source may sometimes be ignored or contradicted to reduce the retrieval error [72, 146, 187]. The LLM-AUGMENTER framework [235] focuses on

augmenting the factuality of NLMs with a plug-and-play module that retrieves evidence from potentially any external knowledge source and performs iterative hallucination checking before giving the final answer. The authors of LLM-AUGMENTER stated they will open-source their framework in the near future.

Debate. The work in [68] uses multiple NLMs to improve factuality and abstract mathematical reasoning via debate. A set of agents is used to answer a query, and successively, the discussion is triggered by issuing all the agents a prompt that includes the responses of every other agent and the request to refine its answer. Extensive experiments were made by the authors varying the prompts, and results suggested that prompt variation helps the performance of the reached consensus in terms of correctness. The work of J. Michael et al. [209] focuses on augmenting the factuality of augmented models. The authors of this work make the argument that retrieval-augmented models may not necessarily be honest. They use a dataset of human-written debates on text understanding to train a model to judge the correct answers from the argument consistency of debaters. Similarly, also [144] used non-expert models to judge the factuality of expert models by setting a debate between the experts. In the work of Cohen et al. [53], an NLM seeks to find inconsistencies in the responses of another NLM by asking questions. The work of Li et al. [166], instead, concentrates on ranking the output of different models for mitigating the self-favouring of a unique model. Finally, the TRUTH-TRIANGULATOR framework presented in [43] uses three NLMs for refining factuality, one of which is tool-augmented. Another recent work that uses debate to enhance factuality in NLMs is available in [76].

4.2 Benchmarks

The COSMOSQA benchmark in [115] focuses on commonsense reasoning and has 35,6K multiple choice questions where the lexical overlap is reduced concerning other factual reasoning benchmarks at the time of publishing. In other words, the correct answers require reasoning about implicit causes and consequences of the facts stated in the questions. Human performance on this dataset is at 94% accuracy, while the publication-time (2019) experiments with SOTA language models, such as BERT and GPT, achieved only 68.4%. By 2020, though, the UNICORN [188] model achieved 92% accuracy in this benchmark.

The FACTUALITYPROMPTS benchmark released in [159] focuses on *open-ended factuality*. The open-ended qualifier stresses that the benchmark is made for assessing the consistency of prompt completion in pre-trained NLMs without additional test-time knowledge given in input. This benchmark uses the Wikipedia corpus as an unstructured knowledge base upon which the outputs of models are evaluated on their factuality. The metrics for this evaluation are automatic and are based on a mixture of named entities overlapping with those of reference documents, entailment ratio with reference sentences, and other fluency and diversity measures. The authors backed up the significance of their metrics by studying their correlations with human judgements. Another factuality benchmark focused on assessing the zero-shot multi-step reasoning capabilities of NLMs is available in [86].

The work in [159] also carried out experiments with NLMs of varied sizes and decoding algorithms. Contrary to previous studies based on conceptual knowledge, the authors of FACTUALITYPROMPTS showed that bigger NLMs tend to hallucinate less in open-ended generation. Unsurprisingly, they also found that, when using nucleus sampling [111], hallucinations occur more with respect to the case of using greedy sampling. Remarkably, the authors of this work contributed with a novel decoding strategy, termed *Factual-Nucleus Sampling*, that preserves factuality without sacrificing the diversity of generated outputs. The intuition behind this contribution may be that the randomness of generation tends to hurt factuality more at the

beginning than at the end of each phrase. (This claim, however, may not be valid in all languages. The authors of this work focused on English NLP).

Another contribution of the work in [159] is a novel training strategy that reduces hallucinations at inference time. Namely, the authors prepend a topic indicator in each document of the pre-training corpus to decouple the potential linkage between non-factual statements and factual statements when considering different documents. Apart from the topic prefix, a similar rationale to that of *Factual-Nucleus Sampling* permitted the authors to propose an enhancement of the loss function that masks the tokens in the most potentially factual positions of phrases.

The authors of HaloCheck [73] offer a question-answer format evaluation framework that can be used for domain-specific factuality of small NLMs such as the Bloom-7B [310]. They assessed factuality on a specific domain as a use case. The authors of HaloCheck also advocate for knowledge injection in the form of specialised fine-tuning strategies to augment the reliability of modest-size NLMs. Other recent works that explore fine-tuning-based knowledge injection are presented in [4, 214, 258].

FactCHD [43] is a recently proposed benchmark that contains 57 K factual and non-factual samples carefully curated to reduce duplications and augment topic diversity. Using this dataset, NLMs can be trained to detect fact-contradictory outputs and give rationales for discriminations in evidence chains. The authors of this work conducted various experiments with SOTA LLMs such as ChatGpt and Alpaca and found their benchmark challenging even for models that are fine-tuned on the 51 K train split of this corpus. The authors of [41] released a benchmark to evaluate automatic factuality evaluators, i.e., a meta-evaluation framework. Also, the study in [80] investigates the effectiveness of NLM-driven factuality evaluation in text summarisation using a relative benchmark in [229]. Surprisingly, the authors of this work found a low correlation between the automatic factuality evaluations made by SOTA LLMs and those of human evaluators. The recent Xiezhi benchmark [94] is meant to assess the domain knowledge of LLMs. This dataset comprises hundreds of thousands of multiple-choice questions from five hundred disciplines across thirteen subjects, including experimental sciences and humanities. The gc-eval benchmark [339] is instead focused on evaluating the correctness and relevance of Chinese generative NLMs across multiple subjects. Other recent benchmarks that focus on the correctness of the answers of Chinese NLMs in various knowledge fields and expertise levels are available in [314, 338].

The popQA benchmark presented in [204] contains 1.4 K questions about untypical or long-tail knowledge, better answered by retrieval augmented models. The experiments conducted by the authors of popQA confirmed that when queried about knowledge in high-probability zones of the training distribution, retrieval of non-parametric knowledge can be misleading, and thus proposed the Adaptive Retrieval framework, which makes use of the external knowledge only in cases of uncertainty. Also focused on retrieval augmented models, the recall benchmark in [184] instead evaluates the capacity of NLMs to deal with counterfactual information in the external knowledge sources. The ***Factual Assessment via Corpus TransfORmation*** (**factor**) framework in [215] permits the automatic creation of a factuality evaluation benchmark from any case-specific corpus. The authors of the expertQA benchmark in [203] also focused on domain-specific scenarios, creating a dataset of 2 K questions made by experts in multiple fields of study and long-form answers. The authors of expertQA tested how SOTA generative NLMs answered these questions and asked the experts to evaluate these responses. Other recent benchmarks in [270, 272, 285] focus instead on assessing generative NLMs in the specific healthcare domain. The bamboo benchmark in [67], instead, concentrates on evaluating long-text coherent modelling capabilities of NLMs.

The authors of FactScore [210] introduce a fine-grain metric to assess the degree of hallucination in large-text generations. They divide the generated text into small chunks and check for the ratio of supported or factual chunks in the overall generated text. The authors of FactScore

initially use human-based verification for factuality and then train a retrieval-augmented NLM to imitate the human FACTSCORE rating and scale to wider experimentations. The experiments in the article permit them to rate the degree of factuality of several open-source and commercial models, and they release the FACTSCORE verifier as a Python library. The UNILC dataset [344] unifies several other benchmarks like SBIC [259], CLIMATE [64], and TOXIGEN [101] to unify factuality assessment with safeness checking. The authors of UNILC also create a verification method based on different prompting strategies to assess the alignment of candidate models.

4.3 Summary and Research Advices

Carefully crafted prompt engineering queries can achieve factuality enhancements in NLMs. Other works instead use supervised fine-tuning where the fine-tuning process focuses on predicting masked factual information given a context [98, 250]. Finally, a trend of research focuses on the usage of *non-parametric knowledge* [206], that is, external knowledge sources in the form of document corpora or knowledge graphs to augment the truthfulness or NLMs' outputs.

Factuality of NLMs can not only be assessed by the FACTUALITYPROMPTS benchmark in [159]: the authors of this benchmark also open-sourced a set of carefully designed pre-training strategies that result in an overall enhancement of the model's factuality. If pre-training an NLM from scratch is not possible, then a valid factuality-enhancing strategy is open-sourced by the authors of the TRUTH-TRIANGULATOR framework, which used Low-Rank Adaptation of LLAMA2 to establish themselves as the champions of their FACTCHD benchmark. If, instead, retrieval-augmented models are the object of factuality enhancement tasks, then the adaptive retrieval framework open-sourced at [204] can help to enhance not only factuality but efficiency.

The work in [325] analyses the honestness alignment of NLMs providing rigorous definitions and metrics for this requirement. The authors of the *allignment for honesty* article also open-sourced datasets and training code to train and test the honestness of NLMs in the correspondent popular benchmarks. An interesting open-source instrument is the *Knowledge Consistent Alignment* framework [289], which enhances factuality by using an expert model that teaches a student to reduce inconsistencies with a potentially arbitrary corpus of knowledge.

From a critical perspective, the perfect hallucination mitigation in NLMs might encounter some fundamental challenges, one of which has been recently formalised by the authors of the **Consistent Reasoning Paradox** (**CRP**) [22]. This work presents the claim that, given a certain problem domain that requires consistent reasoning, it is not possible to have an always-answering and fully consistent artificial (neural) language model that does not hallucinate. Instead, specialised language models, or, more generally, artificial intelligence, could be leveraged that are able to be consistent and correct to a definite set of questions, but are necessarily unable to answer some other questions. In other words, this article states demonstrates that *human-like intelligence in AI necessarily comes with human-like fallibility*. The formalism introduced by this article may also intersect theory of computation and provability theory and require further development under these lenses. These theoretical considerations could be useful to implement mechanisms that refuse to respond to uncertain questions in an ever-more grounded manner.

From a more immediately practical point-of-view, a recent open-source thorough study of hallucination mitigation techniques is available at [163] where the authors also provided a multi-dimensional analysis on the sources of hallucination from multiple perspectives: pre-training corpora and techniques, fine-tuning methods, inference parameters and prompt-based causes for hallucination. The authors of this work devoted to augment their analyses in the near future but let the construction of a holistic hallucination mitigation framework as a carefully framed open-research opportunity. Additional research opportunities regarding the state-of-the-art works presented in this chapter are related to the evaluation of the presented open-source

Table 4. Remarkable Open-Source Instruments Focused on Honest NLMs

Source	Year	Description	Advantages	Weaknesses
HONESTY [325]	2023	Comprehensive repository of datasets and frameworks for honesty alignment and evaluation	Training code and benchmarking available off-the-self	Standard fine-tuning available only with COLLIE [194]
KCA [289]	2024	Training code and fine-tuned factuality aligned model	Dataset and models also available in the Huggingface portal	Requires SFT dataset
R-TUNING [341]	2023	Refusal-aware IFT dataset	Models fine-tuned with this dataset exhibit better honestness	Splits the fine-tuning in two steps.
LLM-AUGMENTER [235]	2023	Plug-and-Play adapter module for factuality enhancement	Requires no fine-tuning	Still to be open-sourced by the authors.
DEBATE [209]	2023	Improve the factuality of knowledge-augmented models through training a judge model	Scales well human-oversight over factuality assessment.	Might need human-debate traces.
FACTUALITY-PROMPTS [159]	2022	Factuality-oriented inductive biases for pre-training.	Equipped with a benchmarking dataset.	Requires document-theme labels on pre-training corpus
POPQA [204]	2023	Factual knowledge retrieval	Not only augments factuality but the overall efficiency of model responses	Might only be effective with comprehensive knowledge sources.

techniques on hallucination mitigation with respect to more recent and challenging open-source benchmarks such as HALLUDIAL [193]. Table 4 contains pointers to the most remarkable open-source instruments that can help kick-start research on honest NLMs.

5 Conclusions

The recent years have witnessed a vast development of human-alignment techniques for NLMs. This survey evidences how increasingly sophisticated techniques have been developed for enhancing the safeness, helpfulness, and usefulness of NLMs. Some of the reviewed methods strongly rely on the emergent learning and generative capabilities of NLMs to implement and scale alignment strategies. Some concluding remarks in the form of calls for future work will now end this research. In Table 5, instead, a summary of recent open-source all-size NLMs is available.

5.1 A Call for Multidisciplinary Research

This work has evidenced a current explosion of research in AI alignment and machine ethics, in general, is taking place. The more the field is being investigated, the more the need for multidisciplinary research is evidenced by computer scientists, neuroscientists, and philosophers centred on deontology, pedagogy, anthropology, and metaphysics. As stated in [18], the alignment criterion for Human-Centric AI is not a duty of ML practitioners alone. The work in [145] surveyed a modest population of AI practitioners and lawmakers from five continents about their points of view on the most crucial factors for correctly implementing AI ethics. Their survey confirmed the importance of transparency and accountability, among others. Surprisingly tough, an aspect that, according to the study, practitioners and lawmakers do not consider having a broad impact on the

Table 5. Verified Human-centricity Evaluation of Open-Source NLMs

Source	Sizes (B = Billion parameters)	Harmlessness	Honestness	Helpfulness	Human-Centric
DLITE V2 [273]	0.124B, 0.355B, 1.5B	✗	✗	✓	✗
CEREBRAS-GPT [62]	0.11B, 0.256B, 0.590B, 1.3B, 2.7B, 6.7B, 13B	✗	✗	✓	✗
PHI-3 [2]	3.8B, 7B, 14B	✓	✗	✓	✗
MISTRAL-7B [128]	7B	✗	✗	✓	✗
ORCA 2 [213]	7B, 13B	✗	✗	✓	✗
MT-BENCH / CHATBOT ARENA [354]	13B	✗	✗	✓	✗
TÜLU 2 [122, 300]	7B, 13B, 70B	✗	✗	✓	✗
DOLLY 2 [153]	2B, 3B, 6B, 7B	✓	✗	✗	✗
BEAVER-DAM-7B [239]	7B	✓	✗	✗	✗
SMAUG [230]	7B, 34B, 72B	✗	✓	✓	✗
OLMO [92]	7B	✗	✓	✓	✗
STABLE LM 2 [24]	1.6B	✗	✓	✓	✗
H2O-DANUBE2 [269]	1.8B	✗	✓	✓	✗
DROMEDARY [278]	65B	✓	✓	✓	✓
GEMMA 2 [281]	2B, 9B, 27B	✓	✓	✓	✓
LLAMA 3/ LLAMA 3.1 [70]	8B, 70B, 405B	✓	✓	✓	✓

alignment process itself is the lack of ethical knowledge. In this respect, severe critiques have been done to the reductionist view of a purely bottom-up approach for alignment [279], merely based on descriptive or non-normative ethics [244] and, on the other side, to only-normative approaches [275], both from practitioners [130] and philosophers [23]. The authors of [130] acknowledged that non-universal representativeness and potential systematic social biases are potential flaws derived from pure data-driven approaches for moral alignment of NLMs. Hybrid approaches that consider the metaphysical origins of morality and data-driven practical inference of safe and helpful behaviour should receive more serious attention in the near future research on ethical AI alignment and machine ethics in general. To this respect, a computer scientist willing to contribute to the research path toward a more holistic human-alignment of AI may consider to include criteria coming from the governance and humanities research communities [25].

Another recent critical review on alignment methodologies is available at [177], where some contradictions inside the RLHF and RLAIF frameworks are identified. The strong point made in this recent article is the need for a sociotechnical and systemic approach to reach AI alignment. A guidance to new research opportunities in this respect may be available in recent governance-aware frameworks such as those in [65, 221, 248]. To this respect, a taxonomy of open problems in the field of technical AI governance is available at [247]. This work identifies a hierarchical taxonomy of open problems across six macro-categories and four micro-categories. Under this taxonomy, more than one hundred specific problems are motivated and explained, from which a vast list of multi-disciplinary research opportunities can be easily extracted.

5.2 A Call for Hybrid-AI

Hybrid-AI combines symbolic and sub-symbolic systems to augment data efficiency, robustness, explainability, and out-of-distribution generalisation of machine learning instruments [84]. Recent work in [160] explains multiple axes of symbiotic collusion between contemporary neural language models and advanced symbolic reasoning systems such as Cycorp's Cyc [57]. Although neural language model developments are accelerating quickly, hybrid-AI unquestionably increases transparency and dependability, making it easier to back up the confidence of releasing machine learning on ever-riskier decision-making support. Beyond the current trend in knowledge-augmented neural models [197], future work directions in contemporary language models might

adopt in-design synergic collaborations between symbolic and sub-symbolic reasoning. Through the usage of external knowledge sources, the KEAR project in [317] may be among the first to shed light on a human-centric design of system-1 (symbolic-abstract-discrete) and system-2 (neural-perceptive-continuous) thinking as defined in [59]. Recent projects demonstrate the effectiveness of neuro-symbolic approaches for generative tasks [222], logical reasoning [224], and abstract reasoning [108, 277]. Perhaps one of the most recent and significant open-source neuro-symbolic approaches to date is that in [298], where the authors equip an NLM-based agent with an external working memory. This work can be used as a baseline for future research considering the evaluation of the agent on multiple benchmarks and the extent to which this framework is effective with small backbone NLMs.

Beyond neuro-symbolic modelling approaches, a quite recent approach to enhance abstract reasoning and trustworthiness of AI involves the creation of abstraction-friendly *architectural inductive biases* in neural language models imported from the cognitive neuroscience research community [304, 305]. To this respect, the *Abstract Transformer* [6] is the most recent and significant open-source implementation of the relational bottleneck inductive bias presented in [304]. The Abstract Transformer showed to improve the efficiency of abstract concept learning in very small NLMs. The authors of this work indicated scaling experiments as a promising open-research opportunity. Although its mechanistic and functional similarity with other architectural inductive biases for NLMs such as the TransNAR [29] architecture might also worth being investigated.

5.3 A Call for Adversarial Strategies

The authors of HELLASWAG concluded that once a powerful model solves a particular benchmark for natural language inference, then such a model should be used alongside adversarial techniques like those used in [336, 337] to try to craft a more challenging benchmark. If it were impossible to create such a novel benchmark, only then would the task of NLI be considered solved. Such a conclusion might also be valid for the human-centric alignment of NLMs in general. A recent candidate strategy for adversarial enhancement of NLP-related benchmarks can be found in [303]. This work proposes the LLM-ATTACK framework, which measures token vulnerability and accounts for semantically equivalent tokens to generate adversarial examples from a baseline corpus. Unlike myopic optimisation approaches that may produce unnatural and human-verifiable adversarial examples, the adversarial samples of LLM-ATTACK proved to outperform human baselines by a significant margin. Three main research opportunities arise from this work. First, this technique can be extended to recently specialized corpora like those based on factuality or harmlessness. Second, the combination of RAG approaches with LLM-ATTACK should be explored to generate novel challenging evaluation datasets, and third, the extent to which training with LLM-ATTACK generated samples augment the robustness of NLMs is still quite an open-research opportunity to explore. The PROMPTBENCH framework in [365] also focuses on the robustness evaluation of NLMs against adversarial prompts, as the Adversarial-GLUE multi-task benchmark in [293] does. However, mastering the adversarial crafting of NLM alignment benchmarks might still be an open and promising area of research [168].

Another critical aspect regarding the significance of benchmarks is the test-set contamination issue, which is related to the test set of older benchmarks ending up among the training data of newer NLMS. The recent LIVEBENCH [308] benchmark has been released to counteract precisely this risk. This framework is focused on robust abstract reasoning and factuality and is being updated monthly considering new knowledge with novel evaluation data which can be verified by symbolic programs. This last condition is also important because using an AI to evaluate the answers could introduce a systematic bias. To this respect, recent research has focused in

introducing guarantees of human-agreement when using a selective evaluation framework that adaptively chooses between NLMs as automatic metrics or humans [137]. This open-source framework can be the base for further research on high-confidence oversight scaling.

5.4 A Call for Data Quality vs. Data Quantity

The authors of the LIMA framework (*less is more for alignment*) made the *superficial alignment hypothesis*. This hypothesis relativises the need for extensive supervised fine-tuning or reinforcement learning cycles to enhance NLMs' alignment to various criteria. The authors of LIMA demonstrated to achieve comparable or even superior engagement rates, commonsense reasoning, and alignment to human values and intentions in NLMs by fine-tuning them on a small dataset made of 1K high-quality demonstrations. A similar claim arises from the research in [158], where the focus was on enhancing information density by reducing similar or identical training data instances. The authors of this work also point out that careful data deduplication may play an essential role in avoiding the inference-time privacy leakages that characterise SOTA NLMs.

To this respect, it is worth noting that COMET-ATOMIC 2020 [120] was assessed in the task of increasing the capacity of generative language models like BART [162] and GPT-2 [241] to infer the missing part in the incomplete tuples of related concepts. Through fine-tuning these low dimensional models, the authors of COMET-ATOMIC 2020 demonstrated superior accuracy concerning GPT-3 [31] despite having 400x fewer parameters. They compared the quality of the knowledge in COMET-ATOMIC 2020 with that of CONCEPTNETV5.5, ATOMIC, and TRANSOMCS using carefully designed crowdsourced human evaluation. Interestingly, the authors of this work approved the thesis that NLMs may memorise common sense knowledge in their parameters. Still, they suggested that fine-tuning based on high-quality discrete knowledge may help these models to learn to output or *express* such knowledge in inference time. This survey supports the call to focus on data quality vs. quantity when creating IFT corpora for human-centric NLMs.

The recent WILDTEAMING [131] framework and benchmark has been open-sourced. This method creates high-quality training data related to harmlessness by carefully extracting insights from chatbot users. By doing so, the authors of WILDTEAMING discovered previously unknown vulnerabilities in NLMs regarding novel clusters of jailbreaking techniques, some of which were originated even without users' intention. Open-research opportunities are related to the exploitation of WILDTEAMING for the creation of high-quality training datasets focused on honestness and helpfulness-related aspects. Other research directions for high-quality (synthetic) data creation involve sophisticated usage of generative NLMs [181]. To this respect, the recent UNIGEN [311] framework is an open-source instrument focused on the production of specialised datasets. UNIGEN incorporates coding-assisted modules, RAG and other symbolic rules to ensure factual, coherent and highly diverse content. Research opportunities bootstrapping from this work might involve the refinement of existent benchmarks with this methodology, apart from evolving/enhancing the effectiveness of (any component of) UNIGEN.

5.5 A Call for Benchmark Convergence

The BigBench benchmark presented in [274] is a composition of more than two hundred heterogeneous tasks for assessing the capacity of NLMs to perform complex abstract reasoning, using commonsense knowledge, and avoid biased or harmful answers, among others. This benchmark is the collection of work by more than four hundred researchers from one hundred thirty institutions. It represents a prominent example of collaboration and convergence in the research focused on human-centric NLMs. Other similar efforts that seek to evaluate NLMs across a wide range of tasks, domains, and evaluation criteria are those in the RAINBOW benchmark in [188] or those in the GAIA benchmark [207], which focuses on general AI assistants. The

Standford's *Holistic Evaluation of Language Models* [28], the MT-BENCH framework in [354], its Chinese counterpart SUPERCLUE in [314], KOLA [332], and SEAEVAL [292] are other recent prominent examples of research efforts devoted to creating a holistic human-centric evaluation of NLMs. Despite these efforts, though, along with the concurrent work [37], our survey claims the need for the confluence of benchmarks towards a human-centric AGI evaluation ecosystem that might be of utmost importance in the near future.

To this respect, the MT-BENCH/CHATBOT ARENA project in [354] is among the most significant human-preference alignment frameworks. In particular, CHATBOT ARENA, which is a crowdsource-based open-ended blind evaluation framework for NLMs, was established in 2024 as the *de facto* standard evaluation platform for NLMs [46] both in academical and industrial contexts. A recent development that builds on this work is BENCHBUILDER [167], an open-source instrument capable of extracting high-quality benchmarks from live-crowdsource data in a fully automated way. To date, the benchmark generated by the authors of BENCHBUILDER establishes as the SOTA automatic evaluation pipeline concerning expert human-agreement rate. The authors of this work pointed out several limitations of BENCHBUILDER that represent open research opportunities. Among these, the extent to which some of the modelling choices of BENCHBUILDER introduce biases in prompt selection could be better investigated. These choices include the specific categories used to assess the significance of prompts and the judge NLMs used during selection. Other research opportunities bootstrapping from BENCHBUILDER include the extension of this framework to generate multi-turn dialogue and multi-lingual benchmarks.

References

[1] K. Park, Y. J. Choe, Y. Jiang, and V. Veitch. 2024. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. arXiv preprint arXiv:2406.01506.

[2] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. arXiv:2404.14219. Retrieved from https://arxiv.org/abs/2404.14219

[3] Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems* 33 (2020), 3045–3057.

[4] Ankush Agarwal, Sakharam Gawade, Amar Prakash Azad, and Pushpak Bhattacharyya. 2023. KITLM: Domain-specific knowledge InTegration into language models for question answering. arXiv:2308.03638. Retrieved from https://arxiv.org/abs/2308.03638

[5] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. arXiv:2307.02179. Retrieved from https://arxiv.org/abs/2307.02179

[6] Awni Altabaa and John Lafferty. 2024. Disentangling and integrating relational and sensory information in transformer architectures. arXiv:2405.16727. Retrieved from https://arxiv.org/abs/2405.16727

[7] Prithviraj Ammanabrolu and Matthew Hausknecht. 2020. Graph constrained reinforcement learning for natural language action spaces. arXiv:2001.08837. Retrieved from https://arxiv.org/abs/2001.08837

[8] Prithviraj Ammanabrolu and Mark O. Riedl. 2018. Playing text-adventure games with graph-based deep reinforcement learning. arXiv:1812.01628. Retrieved from https://arxiv.org/abs/1812.01628

[9] Prithviraj Ammanabrolu and Mark O. Riedl. 2019. Transfer in deep reinforcement learning using knowledge graphs. arXiv:1908.06556. Retrieved from https://arxiv.org/abs/1908.06556

[10] Prithviraj Ammanabrolu, Ethan Tien, Matthew Hausknecht, and Mark O. Riedl. 2020. How to avoid being eaten by a grue: Structured exploration strategies for textual worlds. arXiv:2006.07409. Retrieved from https://arxiv.org/abs/2006.07409

[11] Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? arXiv:2305.04835. Retrieved from https://arxiv.org/abs/2305.04835

[12] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers). Chengqing Zong and Michael Strube (Eds.), Association for Computational Linguistics, Beijing, China, 344–354. DOI: https://doi.org/10.3115/v1/P15-1034

[13] Anthropic. 2023. *Introducing Claude*. Retrieved October 29, 2023 from https://www.anthropic.com/index/introducing-claude

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

[15] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. arXiv:2310.05189. Retrieved from https://arxiv.org/abs/2310.05189

[16] Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. 2022. When is it acceptable to break the rules? Knowledge representation of moral judgement based on empirical data. arXiv:2201.07763. Retrieved from https://arxiv.org/abs/2201.07763

[17] Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. 2023. Improving language models with advantage-based offline policy gradients. arXiv:2305.14718. Retrieved from https://arxiv.org/abs/2305.14718

[18] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. Retrieved from https://arxiv.org/abs/2204.05862

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

[20] Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. 2023. Characterizing large language model geometry solves toxicity detection and generation. arXiv:2312.01648. Retrieved from https://arxiv.org/abs/2312.01648

[21] Mohamad Ballout, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. 2023. Investigating pre-trained language models on cross-domain datasets, a step closer to general AI. arXiv:2306.12205. Retrieved from https://arxiv.org/abs/2306.12205

[22] Alexander Bastounis, Paolo Campodonico, Mihaela van der Schaar, Ben Adcock, and Anders C. Hansen. 2024. On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of 'I don't know'. arXiv:2408.02357. Retrieved from https://arxiv.org/abs/2408.02357

[23] Keith Begley. 2023. Beta-testing the ethics plugin. *AI & SOCIETY* 38.4 (2023), 1503–1505.

[24] Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable LM 2 1.6 B technical report. arXiv:2402.17834. Retrieved from https://arxiv.org/abs/2402.17834

[25] Paolo Benanti. 2023. The urgency of an algorethics. *Discover Artificial Intelligence* 3, 1 (2023), 11.

[26] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety–a review. arXiv:2404.14082. Retrieved from https://arxiv.org/abs/2404.14082

[27] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7432–7439.

[28] Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences* (2023).

[29] Wilfried Bounsi, Borja Ibarz, Andrew Dudzik, Jessica B. Hamrick, Larisa Markeeva, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. 2024. Transformers meet neural algorithmic reasoners. arXiv:2406.09308. Retrieved from https://arxiv.org/abs/2406.09308

[30] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. arXiv:1606.01540. Retrieved from https://arxiv.org/abs/1606.01540

[31] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[32] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. 2023. Scalable AI safety via doubly-efficient debate. arXiv:2311.14125. Retrieved from https://arxiv.org/abs/2311.14125

[33] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv:2307.15217. Retrieved from https://arxiv.org/abs/2307.15217

[34] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv:2308.07201. Retrieved from https://arxiv.org/abs/2308.07201

[35] Harris Chan, Yuhuai Wu, Jamie Kiros, Sanja Fidler, and Jimmy Ba. 2019. ACTRCE: Augmenting experience via teacher's advice for multi-goal reinforcement learning. arXiv:1902.04546. Retrieved from https://arxiv.org/abs/1902.04546

[36] Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. arXiv:2303.11504. Retrieved from https://arxiv.org/abs/2303.11504

[37] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. arXiv:2307.03109. Retrieved from https://arxiv.org/abs/2307.03109

[38] Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2023. Skills-in-context prompting: Unlocking compositionality in large language models. arXiv:2308.00304. Retrieved from https://arxiv.org/abs/2308.00304

[39] Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. arXiv:2203.08480. Retrieved from https://arxiv.org/abs/2203.08480

[40] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. arXiv:2107.03374. Retrieved from https://arxiv.org/abs/2107.03374

[41] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. arXiv:2310.00741. Retrieved from https://arxiv.org/abs/2310.00741

[42] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv:2211.12588. Retrieved from https://arxiv.org/abs/2211.12588

[43] Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023. Unveiling the Siren's song: Towards reliable fact-conflicting hallucination detection. arXiv:2310.12086. Retrieved from https://arxiv.org/abs/2310.12086

[44] Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Black-box prompt optimization: Aligning large language models without model training. arXiv:2311.04155. Retrieved from https://arxiv.org/abs/2311.04155

[45] I. Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality detection in generative AI–A tool augmented framework for multi-task and multi-domain scenarios. arXiv:2307.13528. Retrieved from https://arxiv.org/abs/2307.13528

[46] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. arXiv:2403.04132. Retrieved from https://arxiv.org/abs/2403.04132

[47] Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. 2022. LMPriors: Pre-trained language models as task-specific priors. arXiv:2210.12530. Retrieved from https://arxiv.org/abs/2210.12530

[48] Paul Christiano. 2022. *Current Work in AI Alignment*. Retrieved October 23, 2023 from https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment

[49] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[50] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. arXiv:2309.15402. Retrieved from https://arxiv.org/abs/2309.15402

[51] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try arc, the ai2 reasoning challenge. arXiv:1803.05457. Retrieved from https://arxiv.org/abs/1803.05457

[52] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv:2110.14168. Retrieved from https://arxiv.org/abs/2110.14168

[53] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. arXiv:2305.13281. Retrieved from https://arxiv.org/abs/2305.13281

[54] Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. 2022. Language and culture internalization for human-like autotelic AI. *Nature Machine Intelligence* 4, 12 (2022), 1068–1076.

[55] Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. 2023. Augmenting autotelic agents with large language models. arXiv:2305.12487. Retrieved from https://arxiv.org/abs/2305.12487
[56] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*. Springer, 41–75.
[57] Cycorp. 2023. Cyc Software. Retrieved October 26, 2023 from https://www.cyc.com/
[58] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. arXiv:2310.12773. Retrieved from https://arxiv.org/abs/2310.12773
[59] Kahneman Daniel. 2017. *Thinking, Fast and Slow.*
[60] Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys* 56.4 (2023), 1–41.
[61] Erik Derner, Kristina Batistič, Jan Zahálka, and Robert Babuška. 2023. A security risk taxonomy for large language models. arXiv:2311.11415. Retrieved from https://arxiv.org/abs/2311.11415
[62] Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster. arXiv:2304.03208. Retrieved from https://arxiv.org/abs/2304.03208
[63] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. arXiv:2309.11495. Retrieved from https://arxiv.org/abs/2309.11495
[64] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. arXiv:2012.00614. Retrieved from https://arxiv.org/abs/2012.00614
[65] Roel Dobbe and Anouk Wolters. 2024. Toward sociotechnical AI: Mapping vulnerabilities for machine learning in context. *Minds and Machines* 34, 2 (2024), 1–51.
[66] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. arXiv:2301.00234. Retrieved from https://arxiv.org/abs/2301.00234
[67] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. arXiv:2309.13345. Retrieved from https://arxiv.org/abs/2309.13345
[68] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325. Retrieved from https://arxiv.org/abs/2305.14325
[69] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. arXiv:2302.06692. Retrieved from https://arxiv.org/abs/2302.06692
[70] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv:2407.21783. Retrieved from https://arxiv.org/abs/2407.21783
[71] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. arXiv:2305.14387. Retrieved from https://arxiv.org/abs/2305.14387
[72] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? arXiv:2204.07931. Retrieved from https://arxiv.org/abs/2204.07931
[73] Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. arXiv:2308.11764. Retrieved from https://arxiv.org/abs/2308.11764
[74] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. arXiv:2012.15738. Retrieved from https://arxiv.org/abs/2012.15738
[75] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation. arXiv:2309.15217. Retrieved from https://arxiv.org/abs/2309.15217
[76] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM knowledge Gaps via Multi-LLM collaboration. arXiv:2402.00367. Retrieved from https://arxiv.org/abs/2402.00367

[77] Tao Feng, Zifeng Wang, and Jimeng Sun. 2023. CITING: Large language models create curriculum for instruction tuning. arXiv:2310.02527. Retrieved from https://arxiv.org/abs/2310.02527

[78] Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv e-prints* (2023), arXiv–2305.

[79] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. arXiv:2011.00620. Retrieved from https://arxiv.org/abs/2011.00620

[80] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are large language models reliable judges? A study on the factuality evaluation capabilities of LLMs. arXiv:2311.00681. Retrieved from https://arxiv.org/abs/2311.00681

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

[82] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16477–16508.

[83] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10764–10799.

[84] Artur d'Avila Garcez and Luis C. Lamb. 2023. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review* 56.11 (2023), 12387–12406.

[85] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2023. MART: Improving LLM safety with multi-round automatic red-teaming. arXiv:2311.07689. Retrieved from https://arxiv.org/abs/2311.07689

[86] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics* 9 (2021), 346–361.

[87] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056. Retrieved from https://arxiv.org/abs/2303.15056

[88] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. arXiv:2302.08215. Retrieved from https://arxiv.org/abs/2302.08215

[89] Google. 2023. *Bard*. Retrieved October 29, 2023 from https://bard.google.com

[90] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. arXiv:2305.11738. Retrieved from https://arxiv.org/abs/2305.11738

[91] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Proceedings of the Advances in Experimental Social Psychology*. Elsevier, 55–130.

[92] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. arXiv:2402.00838. Retrieved from https://arxiv.org/abs/2402.00838

[93] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. arXiv:2307.12980. Retrieved from https://arxiv.org/abs/2307.12980

[94] Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Qianyu He, Rui Xu, et al. 2023. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. arXiv:2306.05783. Retrieved from https://arxiv.org/abs/2306.05783

[95] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. arXiv:2305.15717. Retrieved from https://arxiv.org/abs/2305.15717

[96] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. arXiv:2306.11644. Retrieved from https://arxiv.org/abs/2306.11644

[97] Geyang Guo, Ranchi Zhao, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Beyond imitation: Leveraging fine-grained quality signals for alignment. arXiv:2311.04072. Retrieved from https://arxiv.org/abs/2311.04072

[98] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3929–3938.

[99] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. arXiv:2209.00840. Retrieved from https://arxiv.org/abs/2209.00840

[100] Austin W. Hanjie, Victor Y. Zhong, and Karthik Narasimhan. 2021. Grounding language to entities and dynamics for generalization in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4051–4062.

[101] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. arXiv:2203.09509. Retrieved from https://arxiv.org/abs/2203.09509

[102] Matthew J. Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2019. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Retrieved from https://api.semanticscholar.org/CorpusID:202565447

[103] Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. arXiv:2301.00303. Retrieved from https://arxiv.org/abs/2301.00303

[104] Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. 2015. Deep reinforcement learning with a natural language action space. arXiv:1511.04636. Retrieved from https://arxiv.org/abs/1511.04636

[105] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. arXiv:2008.02275. Retrieved from https://arxiv.org/abs/2008.02275

[106] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. arXiv:2103.03874. Retrieved from https://arxiv.org/abs/2103.03874

[107] Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021. What would jiminy cricket do? Towards agents that behave morally. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. J. Vanschoren and S. Yeung (Eds.), Vol. 1, Curran. Retrieved from https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/39059724f73a9969845dfe4146c5660e-Paper-round2.pdf

[108] Michael Hersche, Mustafa Zeqiri, Luca Benini, Abu Sebastian, and Abbas Rahimi. 2023. A neuro-vector-symbolic architecture for solving Raven's progressive matrices. *Nature Machine Intelligence* 5, 4 (2023), 363–375.

[109] Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. 2020. Human instruction-following with deep reinforcement learning via transfer-learning from text. arXiv:2005.09382. Retrieved from https://arxiv.org/abs/2005.09382

[110] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. arXiv:2212.10071. Retrieved from https://arxiv.org/abs/2212.10071

[111] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. arXiv:1904.09751. Retrieved from https://arxiv.org/abs/1904.09751

[112] Zhaoyi Joey Hou, Li Zhang, and Chris Callison-Burch. 2023. Choice-75: A dataset on decision branching in script learning. arXiv:2309.11737. Retrieved from https://arxiv.org/abs/2309.11737

[113] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv:2305.02301. Retrieved from https://arxiv.org/abs/2305.02301

[114] Xinshuo Hu, Dongfang Li, Zihao Zheng, Zhenyu Liu, Baotian Hu, and Min Zhang. 2023. Separate the wheat from the chaff: Model deficiency unlearning via parameter-efficient module operation. arXiv:2308.08090. Retrieved from https://arxiv.org/abs/2308.08090

[115] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. arXiv:1909.00277. Retrieved from https://arxiv.org/abs/1909.00277

[116] Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Vaccine: Perturbation-aware alignment for large language model. arXiv:2402.01109. Retrieved from https://arxiv.org/abs/2402.01109

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

[118] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. arXiv:2207.05608. Retrieved from https://arxiv.org/abs/2207.05608

[119] Yue Huang, Qihui Zhang, and Lichao Sun. 2023. TrustGPT: A benchmark for trustworthy and responsible large language models. arXiv:2306.11507. Retrieved from https://arxiv.org/abs/2306.11507

[120] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6384–6392.

[121] Lightcone Infrastructure. 2023. *Alignment Forum*. Retrieved October 23, 2023 from https://www.alignmentforum.org/

[122] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. arXiv:2311.10702. Retrieved from https://arxiv.org/abs/2311.10702

[123] Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024. Aligner: Achieving efficient alignment through weak-to-strong correction. arXiv:2402.02416. Retrieved from https://arxiv.org/abs/2402.02416

[124] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems* 36 (2024).

[125] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. arXiv:2307.04657. Retrieved from https://arxiv.org/abs/2307.04657

[126] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. arXiv:2310.19852. Retrieved from https://arxiv.org/abs/2310.19852

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

[128] 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. 2023. Mistral 7B. arXiv:2310.06825. Retrieved from https://arxiv.org/abs/2310.06825

[129] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. MEWL: Few-shot multimodal word learning with referential uncertainty. arXiv:2306.00503. Retrieved from https://arxiv.org/abs/2306.00503

[130] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? The delphi experiment. arXiv:2110.07574. Retrieved from https://arxiv.org/abs/2110.07574

[131] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. 2024. WildTeaming at scale: From in-the-wild jailbreaks to (Adversarially) safer language models. arXiv:2406.18510. Retrieved from https://arxiv.org/abs/2406.18510

[132] Minqi Jiang, Jelena Luketina, Nantas Nardelli, Pasquale Minervini, Philip H. S. Torr, Shimon Whiteson, and Tim Rocktäschel. 2020. Wordcraft: An environment for benchmarking commonsense agents. arXiv:2007.09185. Retrieved from https://arxiv.org/abs/2007.09185

[133] Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles. (2023).

[134] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

[135] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Proceedings of the Advances in Neural Information Processing Systems*. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35, Curran Associates, Inc., 28458–28473. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/b654d6150630a5ba5df7a55621390daf-Paper-Conference.pdf

[136] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 562–570.

[137] Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: LLM judges with provable guarantees for human agreement. arXiv:2407.18370. Retrieved from https://arxiv.org/abs/2407.18370

[138] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. arXiv:2207.05221. Retrieved from https://arxiv.org/abs/2207.05221

[139] Gabrielle Kaili-May Liu. 2023. Perspectives on the social impacts of reinforcement learning with human feedback. *arXiv e-prints* (2023), arXiv–2303.

[140] Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. arXiv:2311.09114. Retrieved from https://arxiv.org/abs/2311.09114

[141] Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. BoardgameQA: A dataset for natural language reasoning with contradictory information. arXiv:2306.07934. Retrieved from https://arxiv.org/abs/2306.07934

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

[143] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. arXiv:1912.09713. Retrieved from https://arxiv.org/abs/1912.09713

[144] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive LLMs leads to more truthful answers. arXiv:2402.06782. Retrieved from https://arxiv.org/abs/2402.06782

[145] Arif Ali Khan, Muhammad Azeem Akbar, Muhammad Waseem, Mahdi Fahmideh, Aakash Ahmad, Peng Liang, Mahmood Niazi, and Pekka Abrahamsson. 2022. AI ethics: Software practitioners and lawmakers points of view. arXiv:2207.01493. Retrieved from https://arxiv.org/abs/2207.01493

[146] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. arXiv:1911.00172. Retrieved from https://arxiv.org/abs/1911.00172

[147] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. arXiv:2305.13735. Retrieved from https://arxiv.org/abs/2305.13735

[148] Jan H. Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. 2022. Researching alignment research: Unsupervised analysis. arXiv:2206.02841. Retrieved from https://arxiv.org/abs/2206.02841

[149] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv:2302.09664. Retrieved from https://arxiv.org/abs/2302.09664

[150] Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, et al. 2023. Specific versus general principles for constitutional AI. arXiv:2310.13798. Retrieved from https://arxiv.org/abs/2310.13798

[151] Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. 2020. The nethack learning environment. *Advances in Neural Information Processing Systems* 33 (2020), 7671–7684.

[152] Philippe Laban, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. 2023. Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment. arXiv:2311.08596. Retrieved from https://arxiv.org/abs/2311.08596

[153] Databricks Labs. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. Retrieved 8 September 2024 from https://huggingface.co/databricks

[154] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. 2022. Exploration in deep reinforcement learning: A survey. *Information Fusion* 85 (2022), 1–22.

[155] Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2873–2882.

[156] Raz Lapid, Ron Langberg, and Moshe Sipper. 2023. Open sesame! Universal black box jailbreaking of large language models. arXiv:2309.01446. Retrieved from https://arxiv.org/abs/2309.01446

[157] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv:2309.00267. Retrieved from https://arxiv.org/abs/2309.00267

[158] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. arXiv:1409.0473. Retrieved from https://arxiv.org/abs/1701.00133

[159] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N. Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems* 35 (2022), 34586–34599.

[160] Doug Lenat and Gary Marcus. 2023. Getting from generative ai to trustworthy ai: What llms might learn from cyc. arXiv:2308.04445. Retrieved from https://arxiv.org/abs/2308.04445

[161] Sydney Levine, Max Kleiman-Weiner, Nicholas Chater, Fiery Cushman, and Josh Tenenbaum. 2018. The cognitive mechanisms of contractualist moral decision-making. In *Proceedings of the CogSci.*

[162] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. Retrieved from https://arxiv.org/abs/1910.13461

[163] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv:2401.03205. Retrieved from https://arxiv.org/abs/2401.03205

[164] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. arXiv:2402.05044. Retrieved from https://arxiv.org/abs/2402.05044

[165] Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. 2023. Reflection-tuning: Data recycling improves LLM instruction-tuning. arXiv:2310.11716. Retrieved from https://arxiv.org/abs/2310.11716

[166] Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. arXiv:2307.02762. Retrieved from https://arxiv.org/abs/2307.02762

[167] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and BenchBuilder pipeline. arXiv:2406.11939. Retrieved from https://arxiv.org/abs/2406.11939

[168] Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2024. DrAttack: Prompt decomposition and reconstruction makes powerful LLM jailbreakers. arXiv:2402.16914. Retrieved from https://arxiv.org/abs/2402.16914

[169] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. Retrieved 8 September 2024 from https://github.com/tatsu-lab/alpaca_eval

[170] Yuanpeng Li. 2022. A short survey of systematic generalization. arXiv:2211.11956. Retrieved from https://arxiv.org/abs/2211.11956

[171] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. arXiv:2309.05463. Retrieved from https://arxiv.org/abs/2309.05463

[172] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. arXiv:2206.02336. Retrieved from https://arxiv.org/abs/2206.02336

[173] Ziniu Li, Tian Xu, Yushun Zhang, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023. ReMax: A simple, effective, and efficient method for aligning large language models. arXiv:2310.10505. Retrieved from https://arxiv.org/abs/2310.10505

[174] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out.* 74–81.

[175] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. arXiv:2109.07958. Retrieved from https://arxiv.org/abs/2109.07958

[176] David Lindner, Xin Chen, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. 2023. Learning safety constraints from demonstrations with unknown rewards. arXiv:2305.16147. Retrieved from https://arxiv.org/abs/2305.16147

[177] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. AI alignment through reinforcement learning from human feedback? Contradictions and limitations. arXiv:2406.18346. Retrieved from https://arxiv.org/abs/2406.18346

[178] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. arXiv:2007.08124. Retrieved from https://arxiv.org/abs/2007.08124

[179] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Computing Surveys* 55, 9 (2023), 1–35.

[180] Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony Liu, and Soroush Vosoughi. 2022. Second thoughts are best: Learning to re-align with human values from text edits. *Advances in Neural Information Processing Systems* 35 (2022), 181–196.

[181] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. arXiv:2404.07503. Retrieved from https://arxiv.org/abs/2404.07503

[182] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society. arXiv:2305.16960. Retrieved from https://arxiv.org/abs/2305.16960

[183] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. arXiv:2308.03688. Retrieved from https://arxiv.org/abs/2308.03688

[184] Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. RECALL: A benchmark for LLMs robustness against external counterfactual knowledge. arXiv:2311.08147. Retrieved from https://arxiv.org/abs/2311.08147

[185] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv:2303.16634. Retrieved from https://arxiv.org/abs/2303.16634

[186] Ziyi Liu, Isabelle Lee, Yongkang Du, Soumya Sanyal, and Jieyu Zhao. 2023. SCORE: A framework for self-contradictory reasoning evaluation. arXiv:2311.09603. Retrieved from https://arxiv.org/abs/2311.09603

[187] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. arXiv:2109.05052. Retrieved from https://arxiv.org/abs/2109.05052

[188] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13480–13488.

[189] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13470–13479.

[190] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2022. Responsible AI pattern catalogue: A multivocal literature review. arXiv:2209.04963. Retrieved from https://arxiv.org/abs/2209.04963

[191] Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. arXiv:2309.02654. Retrieved from https://arxiv.org/abs/2309.02654

[192] Man Luo, Shrinidhi Kumbhar, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2023. Towards LogiGLUE: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. arXiv:2310.00836. Retrieved from https://arxiv.org/abs/2310.00836

[193] Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. HalluDial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. arXiv:2406.07070. Retrieved from https://arxiv.org/abs/2406.07070

[194] Kai Lv, Shuo Zhang, Tianle Gu, Shuhao Xing, Jiawei Hong, Keyu Chen, Xiaoran Liu, Yuqing Yang, Honglin Guo, Tengxiao Liu, et al. 2023. Collie: Collaborative training of large language models in an efficient way. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 527–542.

[195] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. arXiv:2301.13379. Retrieved from https://arxiv.org/abs/2301.13379

[196] Chengdong Ma, Ziran Yang, Minquan Gao, Hai Ci, Jun Gao, Xuehai Pan, and Yaodong Yang. 2023. Red teaming game: A game-theoretic framework for red teaming language models. arXiv:2310.00322. Retrieved from https://arxiv.org/abs/2310.00322

[197] Haodi Ma and Daisy Zhe Wang. 2023. A survey on few-shot knowledge graph completion with structural and commonsense knowledge. arXiv:2301.01172. Retrieved from https://arxiv.org/abs/2301.01172

[198] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. arXiv:2310.12931. Retrieved from https://arxiv.org/abs/2310.12931

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

[200] Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. arXiv:2209.07686. Retrieved from https://arxiv.org/abs/2209.07686

[201] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. arXiv:2210.07128. Retrieved from https://arxiv.org/abs/2210.07128

[202] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. arXiv:2212.08410. Retrieved from https://arxiv.org/abs/2212.08410

[203] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. arXiv:2309.07852. Retrieved from https://arxiv.org/abs/2309.07852

[204] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9802–9822.

[205] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv:2303.08896. Retrieved from https://arxiv.org/abs/2303.08896

[206] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: A survey. arXiv:2302.07842. Retrieved from https://arxiv.org/abs/2302.07842

[207] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: A benchmark for General AI Assistants. arXiv:2311.12983. Retrieved from https://arxiv.org/abs/2311.12983

[208] Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. arXiv:2308.00436. Retrieved from https://arxiv.org/abs/2308.00436

[209] Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. 2023. Debate helps supervise unreliable experts. arXiv:2311.08702. Retrieved from https://arxiv.org/abs/2311.08702

[210] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251. Retrieved from https://arxiv.org/abs/2305.14251

[211] Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. arXiv:1805.06975. Retrieved from https://arxiv.org/abs/1805.06975

[212] Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences* 1505, 1 (2021), 79–101.

[213] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching Small Language Models How to Reason. arXiv:2311.11045. Retrieved from https://arxiv.org/abs/2311.11045

[214] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. arXiv:2205.08184. Retrieved from https://arxiv.org/abs/2205.08184

[215] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. arXiv:2307.06908. Retrieved from https://arxiv.org/abs/2307.06908

[216] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv:2305.15852. Retrieved from https://arxiv.org/abs/2305.15852

[217] Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 124–130.

[218] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. arXiv:2112.09332. Retrieved from https://arxiv.org/abs/2112.09332

[219] Richard Ngo. 2023. *AGI Safety from First Principles*. Technical Report. Alignment Forum. Retrieved 8 September 2024 from https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

[220] Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsu H. Hashimoto, and Tobias Gerstenberg. 2023. MoCa: Measuring human-language model alignment on causal and moral judgment tasks. arXiv:2310.19677. Retrieved from https://arxiv.org/abs/2310.19677

[221] Sem Nouws, Íñigo Martinez De Rituerto De Troya, Roel Dobbe, and Marijn Janssen. 2023. Diagnosing and addressing emergent harms in the design process of public AI and algorithmic systems. In *Proceedings of the 24th Annual International Conference on Digital Government Research*. 679–681.

[222] Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M. Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems* 34 (2021), 25192–25204.

[223] Felix Ocker, Daniel Tanneberg, Julian Eggert, and Michael Gienger. 2024. Tulip agent–enabling LLM-based agents to solve tasks using large tool libraries. arXiv:2407.21778. Retrieved from https://arxiv.org/abs/2407.21778

[224] Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. arXiv:2310.15164. Retrieved from https://arxiv.org/abs/2310.15164

[225] OpenAI. 2022. *GPT-3.5*. Technical Report. OpenAI.

[226] OpenAI. 2023. GPT-4 technical report. (2023).

[227] Philip Osborne, Heido Nõmm, and André Freitas. 2022. A survey of text games for reinforcement learning informed by natural language. *Transactions of the Association for Computational Linguistics* 10 (2022), 873–887.

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

[229] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. arXiv:2104.13346. Retrieved from https://arxiv.org/abs/2104.13346

[230] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with DPO-Positive. arXiv:2402.13228. Retrieved from https://arxiv.org/abs/2402.13228

[231] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the International Conference on Machine Learning*. PMLR, 26837–26867.

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

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

[234] René Peinl and Johannes Wirth. 2023. Evaluation of medium-large Language Models at zero-shot closed book generative question answering. arXiv:2305.11991. Retrieved from https://arxiv.org/abs/2305.11991

[235] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv:2302.12813. Retrieved from https://arxiv.org/abs/2302.12813

[236] Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023. When does in-context learning fall short and why? A study on specification-heavy tasks. arXiv:2311.08993. Retrieved from https://arxiv.org/abs/2311.08993

[237] Xiangyu Peng, Mark Riedl, and Prithviraj Ammanabrolu. 2022. Inherently explainable reinforcement learning in natural language. *Advances in Neural Information Processing Systems* 35 (2022), 16178–16190.

[238] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. arXiv:2202.03286. Retrieved from https://arxiv.org/abs/2202.03286

[239] PKU-Alignment. 2023. Beaver-Dam-7B. Retrieved March 2, 2024 from https://huggingface.co/PKU-Alignment/beaver-dam-7b

[240] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. arXiv:2212.09597. Retrieved from https://arxiv.org/abs/2212.09597

[241] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

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

[243] Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? Large language models' sycophantic behaviour. arXiv:2311.09410. Retrieved from https://arxiv.org/abs/2311.09410

[244] John Rawls. 1951. Outline of a decision procedure for ethics. *The Philosophical Review* 60, 2 (1951), 177–197. Retrieved from http://www.jstor.org/stable/2181696

[245] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. arXiv:2310.04988. Retrieved from https://arxiv.org/abs/2310.04988

[246] Michael V. Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. arXiv:2304.11085. Retrieved from https://arxiv.org/abs/2304.11085

[247] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. 2024. Open problems in technical ai governance. arXiv:2407.14981. Retrieved from https://arxiv.org/abs/2407.14981

[248] Lorenzo Ricciardi Celsi. 2023. The dilemma of rapid AI advancements: Striking a balance between innovation and regulation by pursuing risk-aware value creation. *Information* 14, 12 (2023), 645.

[249] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Technical Report. Google.

[250] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? arXiv:2002.08910. Retrieved from https://arxiv.org/abs/2002.08910

[251] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. arXiv:2306.00186. Retrieved from https://arxiv.org/abs/2306.00186

[252] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, et al. 2020. Recipes for building an open-domain chatbot. arXiv:2004.13637. Retrieved from https://arxiv.org/abs/2004.13637

[253] Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. 2024. Representation noising effectively prevents harmful fine-tuning on LLMs. arXiv:2405.14577. Retrieved from https://arxiv.org/abs/2405.14577

[254] Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. 2024. Immunization against harmful fine-tuning attacks. arXiv:2402.16382. Retrieved from https://arxiv.org/abs/2402.16382

[255] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 19861–19872.

[256] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. arXiv:2104.07644. Retrieved from https://arxiv.org/abs/2104.07644

[257] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proscript: Partially ordered scripts generation via pre-trained language models. arXiv:2104.08251. Retrieved from https://arxiv.org/abs/2104.08251

[258] Cicero Nogueira dos Santos, Zhe Dong, Daniel Cer, John Nham, Siamak Shakeri, Jianmo Ni, and Yun-hsuan Sung. 2022. Knowledge prompts: Injecting world knowledge into language models through soft prompts. arXiv:2210.04726. Retrieved from https://arxiv.org/abs/2210.04726

[259] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. arXiv:1911.03891. Retrieved from https://arxiv.org/abs/1911.03891

[260] Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. arXiv:2210.01240. Retrieved from https://arxiv.org/abs/2210.01240

[261] Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. arXiv:2305.15269. Retrieved from https://arxiv.org/abs/2305.15269

[262] Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. arXiv:2009.07118. Retrieved from https://arxiv.org/abs/2009.07118

[263] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv:1707.06347. Retrieved from https://arxiv.org/abs/1707.06347

[264] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. 2023. Towards understanding sycophancy in language models. arXiv:2310.13548. Retrieved from https://arxiv.org/abs/2310.13548

[265] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. arXiv:2309.15025. Retrieved from https://arxiv.org/abs/2309.15025

[266] Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics* 12 (2024), 174–189.

[267] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv:2010.15980. Retrieved from https://arxiv.org/abs/2010.15980

[268] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 27th Conference on Neural Information Processing Systems*.

[269] Philipp Singer, Pascal Pfeiffer, Yauhen Babakhin, Maximilian Jeblick, Nischay Dhankhar, Gabor Fodor, and Sri Satish Ambati. 2024. H2O-Danube-1.8 B technical report. arXiv:2401.16818. Retrieved from https://arxiv.org/abs/2401.16818

[270] 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. 2022. Large language models encode clinical knowledge. arXiv:2212.13138. Retrieved from https://arxiv.org/abs/2212.13138

[271] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. arXiv:1908.06177. Retrieved from https://arxiv.org/abs/1908.06177

[272] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D. Barrett, and Arnu Pretorius. 2023. Are we going MAD? Benchmarking multi-agent debate between language models for medical Q&A. arXiv:2311.17371. Retrieved from https://arxiv.org/abs/2311.17371

[273] AI Squared. 2023. Introducing DLite V2: A Lightweight, Open-Source Machine Learning Model for Microcontrollers. Retrieved 8 September 2024 from https://huggingface.co/aisquared/dlite-v2-124m

[274] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615. Retrieved from https://arxiv.org/abs/2206.04615

[275] Jakob Stenseke. 2023. The use and abuse of normative ethics for moral machines. In *Proceedings of the Social Robots in Social Institutions*. IOS Press, 155–164.

[276] Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. arXiv:1904.01172. Retrieved from https://arxiv.org/abs/1904.01172

[277] Ke Su, Hang Su, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Probabilistic neural–symbolic models with inductive posterior constraints. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[278] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. arXiv:2305.03047. Retrieved from https://arxiv.org/abs/2305.03047

[279] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A word on machine ethics: A response to Jiang et al.(2021). arXiv:2111.04158. Retrieved from https://arxiv.org/abs/2111.04158

[280] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. Retrieved 8 September 2024 from https://crfm.stanford.edu/2023/03/13/alpaca.html

[281] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv:2408.00118. Retrieved from https://arxiv.org/abs/2408.00118

[282] Xiaoyu Tian, Liangyu Chen, Na Liu, Yaxuan Liu, Wei Zou, Kaijiang Chen, and Ming Cui. 2023. DUMA: A dual-mind conversational agent with fast and slow thinking. arXiv:2310.18075. Retrieved from https://arxiv.org/abs/2310.18075

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

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

[285] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. arXiv:2307.15343. Retrieved from https://arxiv.org/abs/2307.15343

[286] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. arXiv:1903.03094. Retrieved from https://arxiv.org/abs/1903.03094

[287] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[288] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *Stat* 1050, 20 (2017), 10–48550.

[289] Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Mitigating hallucinations of large language models via knowledge consistent alignment. arXiv:2401.10768. Retrieved from https://arxiv.org/abs/2401.10768

[290] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems* 32 (2019).

[291] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461. Retrieved from https://arxiv.org/abs/1804.07461

[292] Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. arXiv:2309.04766. Retrieved from https://arxiv.org/abs/2309.04766

[293] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. arXiv:2111.02840. Retrieved from https://arxiv.org/abs/2111.02840

[294] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv:2310.07521. Retrieved from https://arxiv.org/abs/2310.07521

[295] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. MathCoder: Seamless code integration in LLMs for enhanced mathematical reasoning. arXiv:2310.03731. Retrieved from https://arxiv.org/abs/2310.03731

[296] Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making large language models better reasoners with alignment. arXiv:2309.02144. Retrieved from https://arxiv.org/abs/2309.02144

[297] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. ScienceWorld: Is your agent smarter than a 5th grader? arXiv:2203.07540. Retrieved from https://arxiv.org/abs/2203.07540

[298] Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. arXiv:2408.13654. Retrieved from https://arxiv.org/abs/2408.13654

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

[300] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, et al. 2024. How far can camels go? Exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems* 36 (2024).

[301] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. arXiv:2308.13387. Retrieved from https://arxiv.org/abs/2308.13387

[302] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. arXiv:2306.05087. Retrieved from https://arxiv.org/abs/2306.05087

[303] Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024. Generating valid and natural adversarial examples with large language models. In *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design*. IEEE, 1716–1721.

[304] Taylor W. Webb, Steven M. Frankland, Awni Altabaa, Simon Segert, Kamesh Krishnamurthy, Declan Campbell, Jacob Russin, Tyler Giallanza, Randall O'Reilly, John Lafferty, et al. 2024. The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences* (2024).

[305] Taylor W. Webb, Ishan Sinha, and Jonathan D. Cohen. 2020. Emergent symbols through binding in external memory. arXiv:2012.14601. Retrieved from https://arxiv.org/abs/2012.14601

[306] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models. arXiv:2308.03958. Retrieved from https://arxiv.org/abs/2308.03958

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

[308] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. LiveBench: A challenging, contamination-free LLM benchmark. arXiv:2406.19314. Retrieved from https://arxiv.org/abs/2406.19314

[309] Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8 (1992), 229–256.

[310] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv:2211.05100. Retrieved from https://arxiv.org/abs/2211.05100

[311] Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. 2024. UniGen: A unified framework for textual dataset generation using large language models. arXiv:2406.18966. Retrieved from https://arxiv.org/abs/2406.18966

[312] Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2023. SmartPlay: A benchmark for LLMs as intelligent agents. arXiv:2310.01557. Retrieved from https://arxiv.org/abs/2310.01557

[313] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. arXiv:2309.07864. Retrieved from https://arxiv.org/abs/2309.07864

[314] Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. arXiv:2307.15020. Retrieved from https://arxiv.org/abs/2307.15020

[315] Wanqiao Xu, Shi Dong, Dilip Arumugam, and Benjamin Van Roy. 2023. Shattering the agent-environment interface for fine-tuning inclusive language models. arXiv:2305.11455. Retrieved from https://arxiv.org/abs/2305.11455

[316] Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian-guang Lou. 2023. Re-reading improves reasoning in language models. arXiv:2309.06275. Retrieved from https://arxiv.org/abs/2309.06275

[317] Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2021. Human parity on commonsenseqa: Augmenting self-attention with external attention. arXiv:2112.03254. Retrieved from https://arxiv.org/abs/2112.03254

[318] Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. SyGNS: A systematic generalization testbed based on natural language semantics. arXiv:2106.01077. Retrieved from https://arxiv.org/abs/2106.01077

[319] Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2023. Effective distillation of table-based reasoning ability from llms. arXiv:2309.13182. Retrieved from https://arxiv.org/abs/2309.13182

[320] Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrast distillation for language model alignment. arXiv:2307.12950. Retrieved from https://arxiv.org/abs/2307.12950

[321] Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. arXiv:2211.08073. Retrieved from https://arxiv.org/abs/2211.08073

[322] Mengjiao Sherry Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. 2022. Chain of thought imitation with procedure cloning. *Advances in Neural Information Processing Systems* 35 (2022), 36366–36381.

[323] Rui Yang, Edison Marrese-Taylor, Yuhe Ke, Lechao Cheng, Qingyu Chen, and Irene Li. 2023. A UMLS-augmented framework for improving factuality in large language models within healthcare. arXiv:2310.02778. Retrieved from https://arxiv.org/abs/2310.02778

[324] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. arXiv:2303.04129. Retrieved from https://arxiv.org/abs/2303.04129

[325] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. arXiv:2312.07000. Retrieved from https://arxiv.org/abs/2312.07000

[326] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values–a survey of alignment goals for big models. arXiv:2308.12014. Retrieved from https://arxiv.org/abs/2308.12014

[327] Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep CALM and explore: Language models for action generation in text-based games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.), Association for Computational Linguistics, Online, 8736–8754. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.704

[328] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv:2210.03629. Retrieved from https://arxiv.org/abs/2210.03629

[329] Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. arXiv:2308.02151. Retrieved from https://arxiv.org/abs/2308.02151

[330] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. arXiv:2309.06794. Retrieved from https://arxiv.org/abs/2309.06794

[331] Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. arXiv:2303.14725. Retrieved from https://arxiv.org/abs/2303.14725

[332] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. KoLA: Carefully benchmarking world knowledge of large language models. arXiv:2306.09296. Retrieved from https://arxiv.org/abs/2306.09296

[333] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. arXiv:2309.12284. Retrieved from https://arxiv.org/abs/2309.12284

[334] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. arXiv:2002.04326. Retrieved from https://arxiv.org/abs/2002.04326

[335] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. arXiv:2309.05653. Retrieved from https://arxiv.org/abs/2309.05653

[336] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv:1808.05326. Retrieved from https://arxiv.org/abs/1808.05326
[337] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? arXiv:1905.07830. Retrieved from https://arxiv.org/abs/1905.07830
[338] Hui Zeng. 2023. Measuring massive multitask chinese understanding. arXiv:2304.12986. Retrieved from https://arxiv.org/abs/2304.12986
[339] Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2023. Evaluating the generation capabilities of large chinese language models. arXiv:2308.04823. Retrieved from https://arxiv.org/abs/2308.04823
[340] Di Zhang, Jiatong Li, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024. Accessing GPT-4 level mathematical olympiad solutions via monte carlo tree self-refine with LLaMa-3 8B. arXiv:2406.07394. Retrieved from https://arxiv.org/abs/2406.07394
[341] Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. arXiv:2311.09677. Retrieved from https://arxiv.org/abs/2311.09677
[342] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. arXiv:2308.10792. Retrieved from https://arxiv.org/abs/2308.10792
[343] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating language model hallucination with interactive question-knowledge alignment. arXiv:2305.13669. Retrieved from https://arxiv.org/abs/2305.13669
[344] Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable unified language checking. arXiv:2304.03728. Retrieved from https://arxiv.org/abs/2304.03728
[345] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. arXiv:2311.13230. Retrieved from https://arxiv.org/abs/2311.13230
[346] Wanpeng Zhang and Zongqing Lu. 2023. Rladapter: Bridging large language models to reinforcement learning in open worlds. arXiv:2309.17176. Retrieved from https://arxiv.org/abs/2309.17176
[347] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv:1911.00536. Retrieved from https://arxiv.org/abs/1911.00536
[348] Zhuosheng Zhang, Shuohang Wang, Yichong Xu, Yuwei Fang, Wenhao Yu, Yang Liu, Hai Zhao, Chenguang Zhu, and Michael Zeng. 2022. Task compass: Scaling multi-task pre-training with task prefix. arXiv:2210.06277. Retrieved from https://arxiv.org/abs/2210.06277
[349] Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2024. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. arXiv:2407.02855. Retrieved from https://arxiv.org/abs/2407.02855
[350] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2023. Expel: Llm agents are experiential learners. arXiv:2308.10144. Retrieved from https://arxiv.org/abs/2308.10144
[351] Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. arXiv:2402.04833. Retrieved from https://arxiv.org/abs/2402.04833
[352] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023. Knowing what LLMs do not know: A simple yet effective self-detection method. arXiv:2310.17918. Retrieved from https://arxiv.org/abs/2310.17918
[353] Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. Prompt-driven llm safeguarding via directed representation optimization. arXiv:2401.18018. Retrieved from https://arxiv.org/abs/2401.18018
[354] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv:2306.05685. Retrieved from https://arxiv.org/abs/2306.05685
[355] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. arXiv:2307.04964. Retrieved from https://arxiv.org/abs/2307.04964
[356] Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Jian Li, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2021. Fewnlu: Benchmarking state-of-the-art methods for few-shot natural language understanding. arXiv:2109.12742. Retrieved from https://arxiv.org/abs/2109.12742

[357] Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, Bo Du, and Dacheng Tao. 2024. Achieving> 97% on GSM8K: Deeply understanding the problems makes LLMs perfect reasoners. arXiv:2404.14963. Retrieved from https://arxiv.org/abs/2404.14963

[358] Victor Zhong, Austin W. Hanjie, Sida Wang, Karthik Narasimhan, and Luke Zettlemoyer. 2021. Silg: The multi-domain symbolic interactive language grounding benchmark. *Advances in Neural Information Processing Systems* 34 (2021), 21505–21519.

[359] Victor Zhong, Tim Rocktäschel, and Edward Grefenstette. 2019. Rtfm: Generalising to novel environment dynamics via reading. arXiv:1910.08210. Retrieved from https://arxiv.org/abs/1910.08210

[360] Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. arXiv:2104.06598. Retrieved from https://arxiv.org/abs/2104.06598

[361] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. arXiv:2104.05240. Retrieved from https://arxiv.org/abs/2104.05240

[362] Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. arXiv:2308.07921. Retrieved from https://arxiv.org/abs/2308.07921

[363] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. arXiv:2205.10625. Retrieved from https://arxiv.org/abs/2205.10625

[364] Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F. Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? Evaluating mathematical reasoning with checklist. arXiv:2407.08733. Retrieved from https://arxiv.org/abs/2407.08733

[365] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv:2306.04528. Retrieved from https://arxiv.org/abs/2306.04528

[366] Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xingwei Long, and Bowen Zhou. 2023. PaD: Program-aided distillation specializes large models in reasoning. arXiv:2305.13888. Retrieved from https://arxiv.org/abs/2305.13888

[367] Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules. arXiv:2310.07064. Retrieved from https://arxiv.org/abs/2310.07064

[368] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A dataset for LLM question answering with external tools. arXiv:2306.13304. Retrieved from https://arxiv.org/abs/2306.13304

[369] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. arXiv:1909.08593. Retrieved from https://arxiv.org/abs/1909.08593

[370] Caleb Ziems, Jane A. Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. arXiv:2204.03021. Retrieved from https://arxiv.org/abs/2204.03021

[371] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043. Retrieved from https://arxiv.org/abs/2307.15043

Received 7 March 2024; revised 18 September 2024; accepted 10 October 2024