Defining a New NLP Playground

Sha Li[1], **Chi Han**[1], **Pengfei Yu**[1], **Carl Edwards**[1], **Manling Li**[2], **Xingyao Wang**[1], **Yi R. Fung**[1], **Charles Yu**[1], **Joel R. Tetreault**[3], **Eduard H. Hovy**[4], **Heng Ji**[1]
[1] University of Illinois Urbana-Champaign [2] Northwestern University
[3] Dataminr, Inc. [4] University of Melbourne
{shal2, chihan3, pengfei4, cne2, xingyao6, yifung2, ctyu2, hengji}@illinois.edu
manling.li@northwestern.edu, jtetreault@dataminr.com, eduard.hovy@unimelb.edu.au

Abstract

The recent explosion of performance of large language models (LLMs) has changed the field of Natural Language Processing (NLP) more abruptly and seismically than any other shift in the field's 80-year history. This has resulted in concerns that the field will become homogenized and resource-intensive. The new status quo has put many academic researchers, especially PhD students, at a disadvantage. This paper aims to define a new NLP playground by proposing 20+ PhD-dissertation-worthy research directions, covering theoretical analysis, new and challenging problems, learning paradigms, and interdisciplinary applications.

1 Introduction

It is the best of times. It is the worst of times. We are living in an incredibly exciting yet strange era of Natural Language Processing (NLP) research due to the recent advancements of large language models (LLMs) on various data modalities, from natural language (Brown et al., 2020) and programming language (Chen et al., 2021; Wang et al., 2023a) to vision (Radford et al., 2021; Li et al., 2022a; Wang et al., 2022b) and molecules (Edwards et al., 2022; Zeng et al., 2022; Su et al., 2022).

At the core, LLMs produce text sequences word-by-word by computing conditional probability based on context. At a sufficiently large scale, they can answer questions, generate arguments, write poetry, impersonate characters, negotiate contracts and achieve competitive results across a wide variety of standard NLP tasks including entity typing, sentiment analysis, and textual entailment, showcasing "emergent behavior" such as in-context learning (Wei et al., 2022).

However, this "moment of breakthrough" received a polarized response in the NLP research community: while some welcomed the progress, others felt lost. Why is NLP so vulnerable to a single advancement?

In retrospect, when NLP adopted the machine learning paradigm in the early 1990s it started along a journey that led to increased homogeneity. The dominant methodology became: (1) Identify a challenge problem or task; (2) Create a dataset of desired input-output instances; (3) Select or define one or more evaluation metrics; and (4) Develop, apply, and refine machine learning models and algorithms to improve performance.

If a challenge did not support the creation of a dataset (e.g., text styles of people in different professions) or metric (e.g., summaries of novels or movies), or worse yet if it was not amenable to a machine learning solution, then mainstream NLP simply did not address it. For a long time, NLG was in this position because its starting point —semantic representations— were neither standardized, nor easy to produce at scale, nor amenable to direct evaluation. No dataset, no metric — little attention. Yet multi-sentence NLG starting with deep semantic input, and with output tailored to different audiences, is arguably the most complex task in NLP, since it involves so many aspects of linguistic communication together. As such, it surely deserved the concentrated effort that NLP has bestowed on MT, Speech Recognition, QA, and other major challenges in the past.

Suddenly, within the space of a few months, the landscape changed. NLP encountered an engine that seemingly could do everything the field had worked on for decades. Many subtasks in NLP seemed to become irrelevant overnight: Which grammar formalism to parse into? Which rhetorical structure and focus control model for multi-sentence coherence? Which neural architecture is optimal for information extraction or summarization? None of that matters if the magical engine can do the entire end-to-end language-to-language task seamlessly (Sanh et al., 2022; OpenAI, 2023).

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11932–11951
December 6-10, 2023 ©2023 Association for Computational Linguistics

Dozens of Ph.D. theses lost their point, because their point was a small step in the process that no longer seemed needed. The dominant paradigm is also challenged: instead of setting up benchmarks and then developing models accordingly, people started discovering new abilities of such models (Bubeck et al., 2023) (who knew that LLMs could draw unicorns using TikZ?).

An important constraint is the practicality of the goal. This newer generation of LLMs is beyond the practical reach of all but a small number of NLP researchers. Unless one of the organizations building LLMs provides free access for research —an unlikely occurrence given the estimated six-figure monthly expense to run one— or a procedure is developed to construct university-sized ones cheaply, the academic NLP community will have to be quite creative in identifying things that either generative LLMs cannot do *in principle* or applications that can be built without re-training them and at the same time are important and doable *in practice*.

Inspired by the efforts of a group of PhD students (Ignat et al., 2023), we believe it would be a valuable exercise to define new research roadmaps. We believe that while LLMs seemingly close research avenues, they also open up new ones. Current LLMs remain somewhat monolithic, expensive, amnesic, delusional, uncreative, static, assertive, stubborn, and biased black boxes. They still have a surprising deficiency (near-random performance) in acquiring certain types of knowledge (Wang et al., 2023f), knowledge reasoning and prediction. In this paper, we aim to define a new NLP playground by proposing a wide range of PhD-dissertation-worthy research directions to democratize NLP research again. In particular, we cover observations and suggestions along the perspectives of LLM theory (Section 2), challenging new tasks (Section 3), important but understudied learning paradigms (Section 4), proper evaluation (Section 5), and interdisciplinary applications (Section 6).

2 Theoretical Analysis of LLMs

There is a growing necessity to open the black box of machine learning models through theoretical analysis. In this section, we advocate for both **mathematical** (by mathematical analysis) and **experimental** (inducing rules and laws such as Ghorbani et al. (2021); Hoffmann et al. (2022) from extensive experimental observations) theories of LLMs.

2.1 Mechanism Behind Emergent Abilities

LLMs have displayed impressive emergent capabilities such as instruction following, chain-of-thought reasoning, and in-context learning (Brown et al., 2020; Wei et al., 2022; Min et al., 2022; Wei et al.; Logan IV et al., 2022; Wei et al., 2021). For example, the ability of **instruction following** enables models to follow novel instructions. For guidance on prompting beyond heuristics, we need a comprehensive understanding of how instructions work. Some initial theories suggest an explanation through Bayesian inference (Jiang, 2023), which relies on strong assumptions without practical insights. Here we advocate for theories on the feasibility of constraining or measuring models' deviation from instructions. A multi-player setting is also important, where one user's prompt is composed with another player's prompt (such as OpenAI's hidden meta instruction) before being fed into LLMs, where additional security issues might arise for the first user.

Chain-of-thought (CoT) reasoning is where LLMs tackle complex tasks by generating solutions in a sequential, step-by-step manner. CoT theoretically enhances the computational capacity of Transformer-based models to solve problems exceeding $\mathcal{O}(n^2)$ complexity. While some constructive explanations have been suggested (Feng et al., 2023a), they are not fully validated as the underlying mechanism. Importantly, it is worth investigating the verifiability problem of the reasoning chain (whether CoT can be trusted as a valid logic chain) and its calibration (whether LLMs formulate ad-hoc CoTs for arbitrary conclusions).

In-context learning (ICL), where LLMs learn from demonstration examples in-context without parameter updates, has seen explanations based on gradient-descent (Akyürek et al., 2022; von Oswald et al., 2022), kernel regression (Han et al., 2023a) or Bayesian inference (Xie et al.; Jiang, 2023; Wang et al., 2023d). Important challenges remain and necessitate more comprehensive explanations, such as sensitivity to example order and robustness to perturbed input-output mapping. We hypothesize that a deeper understanding of how LLMs balance algorithmic solutions with implicit language inference can help clarify these questions, which might be approachable by exploring how LLMs disentangle semantic and functional information.

Model-specific vs. Model-agnostic is a persistent gap among explanations, raising the question of whether the emergent abilities depend on the Transformer architecture or simply fitting the pre-training data. With some recent work suggesting that other architectures achieve comparable performance in some domains (Peng et al., 2023; Zhai et al., 2021), this open question is important for prioritizing among model design (including other architectures), prompting engineering, and simply carefully collecting larger datasets. To bridge this gap, we also advocate for theoretical frameworks beyond (mixture) of HMMs to better model language data properties.

2.2 Theoretical Robustness and Transparency

Robustness is to ensure that no backdoor designs or adversarial usages can be easily implemented in the model. Although not a novel problem by definition, this issue has new implications and formulations in the LLM era. In a situation where most users do not have access to the pre-training and model-editing details, we call for research into robustness diagnosis *for arbitrary given LLM*. Despite negative evidence suggesting it may be nearly impossible to prevent adversarial prompting under certain conditions (Wolf et al., 2023), we maintain a positive outlook and hope that it can be potentially overturned under more realistic conditions, such as high computational complexity in searching for adversarial prompts.

Transparency in LLMs is concerned with alignment between the model's self-explanations and its internal computational rationale. With empirical studies suggesting that LLMs may not always accurately express their "thoughts" (Turpin et al., 2023), computational modeling of LLM intentions becomes essential. The quest for transparency is important for preventing LLMs from generating misleading rationales to humans. We advocate for establishing both positive and negative theorems on counteracting false rationales under different conditions, along with examining associations between "faithfulness" modes and neuron activities in specific architectures.

3 New and Challenging Tasks

3.1 Knowledge Acquisition and Reasoning

Knowledge inside LLMs The black box property of LLMs poses a significant challenge when it comes to evaluating implicit knowledge within the model. Initial studies have been conducted to elicit/identify (Cohen et al., 2023; Shin et al., 2020; Petroni et al., 2019, 2020; Fung et al., 2023; Gudibande et al., 2023; Li et al., 2023c) and localize/edit knowledge (Dai et al., 2021; Meng et al., 2022a,b; Zhu et al., 2020; Mitchell et al., 2022a; De Cao et al., 2021; Hase et al., 2023; Meng et al., 2022a; Mitchell et al., 2022b). However, our understanding of the knowledge organization within language models (*where* and *how* knowledge is stored) is still limited, and it remains uncertain whether full comprehension is achievable. Moreover, existing studies primarily focus on factual or commonsense knowledge, overlooking more complex knowledge such as rules of inference (Boolos et al., 2002).

Large-Scale Knowledge Reasoning LLMs have demonstrated promising performance across various reasoning tasks (Dua et al., 2019; Miao et al., 2020; Cobbe et al., 2021; Yu et al., 2020; Bhagavatula et al., 2020; Talmor et al., 2019) when appropriately prompted, such as through the use of Chain-of-Thought and improved Chain-of-Thought (Wei et al.; Chowdhery et al., 2022; Xue et al., 2023; Diao et al., 2023; Wang et al., 2023e; Paul et al., 2023) or Program-of-Thought (Chen et al., 2022). However, current reasoning benchmarks (Cobbe et al., 2021; Ling et al., 2017; Patel et al., 2021; Hosseini et al., 2014; Miao et al., 2020; Koncel-Kedziorski et al., 2016; Talmor et al., 2019; Geva et al., 2021) focus on reasoning with small-scale context, typically consisting of hundreds of words. This level of reasoning falls short when tackling complex tasks, such as scientific research, which demands knowledge from extensive volumes of related literature and domain-specific knowledge bases. Retrieval-augmentation (Guu et al., 2020; Khandelwal et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Lai et al., 2023b) serves as a powerful tool for integrating large-scale contextual knowledge into language models. However, current retrieval methods predominantly rely on semantic similarities, while humans possess the *accommodative* learning (Illeris, 2018) ability to draw inspirations from semantically dissimilar knowledge and transfer it to the target task. To achieve this, we not only need to extend the input context length, but also understand how models organize knowledge and develop more effective knowledge representations and evaluation metrics (Section 5).

Faithfulness and Factuality Ensuring the truthfulness of generation output requires optimal utilization of internal knowledge within the model and external knowledge, which includes the input context, knowledge bases, and open web resources. Access to external knowledge typically relies on the success of information retrieval (Lewis et al., 2020; He et al., 2023; Yu et al., 2023c,b), information extraction (Wen et al., 2021; Huang et al., 2023), grounded generation (Li et al., 2021, 2022b; Gao et al., 2023a; Weller et al., 2023; Lai et al., 2023a) and knowledge-augmented generation (Petroni et al., 2020; Geva et al., 2023). Internal knowledge involves the implicit parametric knowledge stored within the model, the correction and refinement of which is limited to the inference stage (Lee et al., 2022; Meng et al., 2022a,b; Chen et al., 2023a). To effectively minimize hallucination and correct factual errors, it is crucial to not only decipher how knowledge is interpreted through model parameter patterns, but to understand how the model pieces knowledge together and governs the underlying logic during generation. A significant challenge in knowledge-guided generation is defining an appropriate knowledge representation that supports both complex structures and distributed representations. We believe this representation should combine the strength of symbolic-based reasoning to minimize unwarranted inferences, and the flexibility of distributed representations to encode any semantic granularity. Drawing insights from misinformation detection and knowledge comparative reasoning systems could also be one useful dimension of signals for improving faithfulness and factuality (Liu et al., 2021a; Fung et al., 2021; Wu et al., 2022, 2023).

3.2 Creative Generation

Although people have long envisioned using models for creative writing, this has only become a reality recently, when language generation models could reliably produce fluent text. Compared to previous sections where generated text is a vehicle for knowledge, creative use cases focus more on the style or form of language and encourage open-ended output. [1]

Creative Writing Assistance Since language models offer conditional generation ability out-of-the-box, they have been adopted by many people in the creative industry for brainstorming or research tools (Kato and Goto, 2023; Gero et al., 2023; Halperin and Lukin, 2023). One key challenge for such tools is promoting creative generation, instead of generating the most probable continuation, which was what language models were trained for. Current LMs have been observed by writers to overrely on cliques or tropes and produce overly moralistic and predictable endings (Chakrabarty et al., 2024). While the plot should be unexpected, details in the story should not go against commonsense (unless it is part of the setting), and maintain consistency within the story. This requires a model that enables controllability over the level of creativity in its output. Do we need to train a more creative model, or can we fix the problem at the inference stage? On the other hand, the focus on detoxification of LMs through RLHF (reinforcement learning with human feedback) might have led to the incompetency of the model in navigating deeper and morally challenging themes.

Another direction for exploration is how to build better writing tools that work together with humans. Some attempts have been made to allow users to interact through instructions (Chakrabarty et al., 2022) or use editing sequences to improve writing quality (Schick et al., 2022). These could serve as critical building blocks toward the goal of developing a model that supports different types of input and can improve itself and personalize through interaction. In addition, models can also assist in different stages of writing, such as world-building and reviewing drafts. It remains to be explored where the model is most effective and where human writers should step in and make decisions.

Interactive Experiences Text generation models can not only be assistants for writing static scripts but also open up an opportunity to create dynamic and personalized experiences for the user by conditioning on their input. These interactive experiences can be used for education, therapy, game design, or filmmaking. More recently, there have been attempts to connect conversational models with other components such as speech recognition, text-to-speech, and audio-to-face rendering to create an end-to-end immersive experience of interacting with non-playable characters[2][3]. Another related open area for exploration is to create

[1] In this section we limit our scope to applications of text generation, however, we fully acknowledge the potential of multi-modal creative generation, such as generating personal avatars, movie clips, and 3D scenes.

[2] NVIDIA blog

[3] https://charisma.ai/

emotion-oriented experiences, which is one of the key goals of storytelling (Lugmayr et al., 2017). We should consider creating narratives based on the desired emotional response and the reader's feedback (Brahman and Chaturvedi, 2020; Ziems et al., 2022; Mori et al., 2022).

4 New and Challenging Learning Paradigms

4.1 Multimodal Learning

In light of the remarkable progress of the language world, we are now poised to venture into a multitude of modalities that were previously beyond consideration. Some learning signals stem from reading static data, such as images, videos, speech, and more, which will be discussed in this section; while other signals require interacting with the physical world, which will be detailed in Section 4.2.2.

Multimodal encoding, at its core, involves learning the "correspondence" or "alignment" among various modalities, which always facing the challenges of **Granularity Difference** across modalities. This is a new and growing area with several solutions proposed to align across modalities: (1) a hard alignment that enables granularity-aware fusion (Tan and Bansal, 2020; Li et al., 2022a; Momeni et al., 2023; Wang et al., 2022c, 2023f); (2) a soft alignment to project the text space with the vision space (Zhou et al., 2023; Li et al., 2023b; Zhu et al., 2023; Lin et al., 2023). Beyond these semantic alignment challenges, there are further difficulties when it comes to non-semantic abstractions:

Geometric Reasoning: Recognizing spatial relationships, such as "*left*", "*right*", "*beside*", "*above*", or "*behind*", requires comprehensive geometric mental simulation, which existing models consistently making errors (Kamath et al., 2023). Maintaining transformation invariance, regardless of position, rotation, or scale, remains a core challenge. Besides, current models, predominantly trained on 2D images, inherently miss out on the intricacies of 3D spatial configurations, inhibiting understanding of depth and relative object sizes based on distance. To address these challenges, existing efforts augment existing large models with an agent view to infer spatial layouts, predicting possible navigations from visual and textual cues (Liu et al., 2022; Berrios et al., 2023; Feng et al., 2023b). However, we believe the underlying challenge lies in the missing objective of geometric reasoning. Existing pretraining paradigms predominantly focus on semantic alignment between image/video-language pairs, while features (e.g., low-level edges, lines) are largely omitted in the encoded image representation.

Context Ambiguity: Accurate understanding should factor in the wide context of temporal dynamics, social dynamics, emotional dynamics, and more. The temporal dimension presents a unique challenge in understanding vision and speech. Existing methods only focus on temporal ordering (Zellers et al., 2021, 2022) and forward/backward generation (Seo et al., 2022; Yang et al., 2023a; Cheng et al., 2023). However, temporal dynamics is much more complicated. For instance, a video gesture (like a nod) may correspond to a later affirmation in the speech (Li et al., 2019). Such ambiguity requires reasoning over a wider context with various constraints. Emotion, another yet-underexplored abstract dimension, is conveyed through tone, pitch, speed in speech, and through expressions or body language in vision. Besides, social norm understanding is challenging as the same word or facial expression can convey different emotions depending on the context. Thus, potential solutions require to take into account various contexts, including preceding conversations or events, along with casual reasoning.

Hierarchical Perception: Human cognition is inherently hierarchical. When processing visual signals, our attention is not uniformly distributed across every pixel but focus on salient regions that carry the most information, allowing us to quickly identify key features and make sense of our surroundings (Hochstein and Ahissar, 2002; Eickenberg et al., 2017). However, existing models overlook such attention hierarchy and tend to lose focus when asking about visual details (Gao et al., 2023b). To address this challenge, interpreting natural scenes requires hierarchical recognition, from broader contexts down to detailed attribute abstraction. Besides, aligning visual hierarchies with linguistic structures is important. Further, it requires the ability to perform abstraction over details, balancing between an abstracted scene understanding and intricate recognition is an ongoing challenge.

4.2 Online Learning

Trained on static corpora, existing models are incapable of keeping themselves updated on new information or learning from interaction history for

self-improvement. To alleviate these issues, this section discusses the need for next-generation models to learn in an *online* setting.

4.2.1 Updating Information Within Models

A straightforward approach to updating models is to continue training on new data. This is however not efficient, since we only care about new information which accounts for a small fraction of the data, nor effective, as fine-tuning on new data might interfere with learned information in models. To achieve efficient updates, we would like the model to automatically identify notable information in new data (Yu and Ji, 2023) instead of relying on heavy human selection or preprocessing as in knowledge editing tasks (Dai et al., 2021; Meng et al., 2022a,b; Zhu et al., 2020; De Cao et al., 2021; Hase et al., 2023; Mitchell et al., 2022b). Effectively updating the model requires overcoming the bias toward (Yu and Ji, 2023; Wei et al., 2023) as well as avoiding catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990) of learned prior information. This might be achieved by changing the training paradigm to increase model capacity over time (e.g. progressive training (Gong et al., 2019), MoE (Shen et al., 2023)) or better understanding of knowledge organization within models (as detailed in Section 3.1) so that edits can be performed with minimal interference.

4.2.2 Learning from Continuous Interactions

Interaction is essential in human learning (Jarvis, 2006). Humans learn how to best tackle different tasks by interacting with the **environment**, and they learn social norms from their interactions with other **humans**. Moreover, such interactions are **multi-turn** in nature, allowing humans to iteratively refine their actions for the task at hand *and* continuously improve their mental model's capability of performing similar tasks in the future.

Interaction with Environments We consider environments a broad category of systems that provide feedback upon actions. The world we live in can be regarded as a typical environment: the law of physics would decide the world state change and provide sensor stimuli to the actor (e.g., Ahn et al. (2022)). Training a model (i.e., Embodied AI) that can interact with the physical world through multi-modal input (Driess et al., 2023; Jiang et al., 2023) poses challenges related to multi-modal learning (Section 4.1) as well as unique challenges due to long-horizon planning requirements and dynamic environments. The concept of environments also extends to human-crafted environments (e.g., programming language interpreters (Wang et al., 2023b), embodied simulators (Shridhar et al., 2020)) that provide automated feedback for any input by rules. Such artificial environments allow easy collection of automatic feedback which could prepare models for deployment in the physical world.

Interaction with Humans Beyond learning from generic human preference towards building generalist agents (Ouyang et al., 2022), real-world applications typically require customizable solutions (e.g., personalized agents) to be created efficiently. We advocate for a new learning paradigm where models can be taught through (multi-modal) interactions with humans, including natural language feedback (Padmakumar et al., 2022; Wang et al., 2023c) and physical demonstration (Lee, 2017). Such complex problem nature may also involve customized retrieval from a large toolset of specialized models and effective action planning (Qin et al., 2023; Yuan et al., 2023).

5 Evaluation

As models become increasingly powerful and multi-purpose, their evaluation has become a growing bottleneck for advancing NLP. We first discuss the question of "what should be evaluated" followed by "how should we measure performance."

5.1 Benchmarks

Language models are known to be multi-task learners, and the new generation of LLMs can achieve impressive performance under few-shot or even zero-shot conditions. This has led to the creation of many general benchmarks such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), Super-NaturalInstructions (Wang et al., 2022a), HELM (Liang et al., 2022), and AGIEval (Zhong et al., 2023). While setting up comprehensive benchmarks is useful, current benchmarks still have the following limitations: (1) lack diverse and difficult tasks that are important for real-world applications; (2) only contain static data sets that are not sufficient for applications that require multi-turn context-dependent input such as situation-grounded dialog; (3) robustness deficiencies, and (4) lack of support for performance analysis.

Although some benchmarks extend to thousands of NLP tasks, most of them are variants of sentence-level tasks, while ignoring more challenging tasks such as structured prediction and cross-document reasoning. For example, Li et al. (2023a) reported that LLMs methods obtained 25.2%-68.5% lower performance than state-of-the-art methods based on much smaller models for nearly all of the Information Extraction tasks. Task design should also aim to assist with human users' daily tasks, as exemplified by the most popular tasks being related to planning and seeking advice by the ChatGPT users at ShareGPT [4]. Another issue is that benchmarks quickly saturate due to the development of newer models, and thus "live" benchmarks that can be updated over time (Kiela et al., 2021) might be worth pursuing.

To move beyond static data, we believe that simulated environments such as large-scale multi-player game environments can serve as an efficient solution. Games have been used as a way of benchmarking progress of reinforcement learning algorithms (Silver et al., 2018; Guss et al., 2021) and also used to collect static datasets in NLP (Urbanek et al., 2019; Bara et al., 2021; Lai et al., 2022). Game worlds provide a cheap way to explore different environments and situations, which is necessary for grounded language learning and learning through interaction. Humans can interact with models playing as characters in the game to evaluate their performance, or we can let models interact with each other (Park et al., 2023) and evaluate their interaction behavior as a whole.

Finally, we advocate for work on model diagnosis beyond the current brittle paradigm of case study through manual inspection: methods that help identify which parts of the input the model underperform on (Liu et al., 2021b), what are the model's behavior patterns and what data this performance could be attributed to (Ilyas et al., 2022).

5.2 Metrics

Automatic evaluation metrics have been an accelerant for NLP progress in the last 20 years. Heuristic-based metrics (Papineni et al., 2002; Lin, 2004; Lavie and Agarwal, 2007) have been found to correlate weakly with human preferences (Liu et al., 2016). As a result, the field has pivoted to model-based metrics which have shown better alignment with human judgment (Lowe et al., 2017; Zhang et al., 2020; Sellam et al., 2020; Yuan et al., 2021; Zhong et al., 2022). However such metrics might allow for shortcut approaches or come with biases embedded in the scoring model (Sun et al., 2022).

Automatic metrics struggle with open-ended natural language generation problems such as conversation and creative writing tasks due to the absence of ground truth. LLMs present an opportunity to tackle this problem (Zheng et al., 2023; Fu et al., 2023; Liu et al., 2023b), but they also suffer from certain biases including position, verbosity, and self-enhancement biases (models prefer themselves) that users should be cautious about. We need to develop metrics beyond accuracy and evaluate aspects such as robustness (Chen et al., 2023b), bias, consistency (Chan et al., 2023), informativeness, truthfulness, and efficiency.

On the other hand, human evaluation has traditionally been perceived as the more trustworthy evaluation method and a better indicator of the model utility. However, as models improve, it is questionable whether crowdworkers are adequate to serve as assessors (or annotators), particularly in fields such as science, healthcare, or law. Annotator bias (Geva et al., 2019; Sap et al., 2022) and disagreement (Fornaciari et al., 2021) should also be taken into consideration. If we design our models to be "assistants", a more useful human evaluation might not be to identify which output is more correct, but which output can help the human complete the task more efficiently.

6 NLP+X Interdisciplinary Applications

6.1 Human-Centered NLP

As LLMs become ubiquitous in both the research and public spheres, mitigating potential harms, both allocation and representation (Blodgett et al., 2020), to social groups using these models must be a core consideration. Social bias and stereotypes are a common way for LLMs to materialize these internal defects, so debiasing these models is important for fairness and robustness. Furthermore, LLMs must be aware of the extra-contextual requirement of abiding by the sociocultural norms expected by the user (Fung et al., 2023), especially when used as chatbots directly interacting with humans.

Post-hoc debiasing and improving the social awareness of pretrained LLMs are important to this end. Though modern approaches have made great advances in democratizing LLM training, most

[4] https://sharegpt.com/

builders don't have a need to pretrain their own LLMs, opting to, at most, fine-tune them. Rather than hope that an LLM is unbiased after pretraining, many researchers have discussed the utility in having a separate general debiasing step to account for any unintended associations stemming from pretraining (Yu et al., 2023a; Omrani et al., 2023; Yang et al., 2023b). Relatively less explored is the complementary requirement of augmenting LLMs with the awareness and ability to abide by sociocultural norms. The crux of the problem is training the model to recognize *what* behaviors in its training data are the results of sociocultural norms, discover *why* and *when* those norms should be followed, and *how* those norms can be followed (i.e., is it only in a specific way or is this a behavior that can be generalized across situations?).

Another important direction is personalization based on the user, particularly for chatbots. LLMs have an amazing ability to multiplex behavior based on the language context provided in the prompt (Section 2.1), but they do not have the ability to account for the audience apart from what's inferred from text. This poses a problem for personalization because the same context or conversation can have differing levels of appropriateness depending on the audience (e.g., something that one finds relatively harmless may be incredibly offensive to someone else). Thus, we must improve LLMs' ability to infer the personal norms and appropriate behaviors in each individual context independently and act accordingly. This may, in part, involve bridging the gap between distant users who share similar beliefs to decode latent representations (Sun et al., 2023). In parallel, we can also provide users with multi-dimensional controls for generation (Han et al., 2023b), including their sentiment, political stance, and moral values, so that they can directly influence the model's language usage.

6.2 NLP for Science

One area with the most potential impact from NLP is science (Hope et al., 2022; Zhang et al., 2023). Although researchers have long been interested in extracting actionable information from the literature (Hersh and Bhupatiraju, 2003; Griffiths and Steyvers, 2004; Li et al., 2016; Wang et al., 2021), this has been challenging due to the variety and complexity of scientific language. With the growing capabilities of NLP techniques, intensified focus is now deserved because of both the potential impacts and the challenges that will need to be overcome.

One exciting emerging area is jointly learning natural language and other data modalities in the scientific domain (Edwards et al., 2021; Zeng et al., 2022; Edwards et al., 2022; Taylor et al., 2022), and one of the largest problems in current LLMs–hallucination–becomes a strength for discovering new molecules (Edwards et al., 2022), proteins (Liu et al., 2023a), and materials (Xie et al., 2023).

Another noteworthy application is NLP for Medicine. As a particular motivating example, there are an estimated 10^{33} realistic drug-like molecules (Polishchuk et al., 2013). Within these drugs, there are substructures which confer beneficial drug properties, and the knowledge about these properties are reported in millions of scientific papers. However, existing LLMs are pretrained only from unstructured text and fail to capture this knowledge, in part due to inconsistencies in the literature.

Recent solutions for domain-knowledge-empowered LLMs include development of a lightweight adapter framework to select and integrate structured domain knowledge into LLMs (Lai et al., 2023b), data augmentation for knowledge distillation from LLMs in general domain to scientific domain (Wang et al., 2023g), and tool learning frameworks leveraging foundation models for more complicated sequential actions problem solving (Qin et al., 2023; Qian et al., 2023). Overall, future research can explore bespoke architectures, data acquisition techniques, and training methodologies for comprehending the diverse modalities, domain-specific knowledge, and applications within science.

6.3 NLP for Education

LLMs readily capture a vast knowledge of many subjects, and augmenting LLMs with external knowledge naturally leads to improved abilities for eliciting that knowledge to generate lesson plans and materials. However, there are also applications in education which seem distinct from general NLP tasks. In particular, personalizing education and the educational experience with LLMs would allow educators to focus on the more general efforts of high-level teaching. Then, the utility of using language models to educate comes not from the language model's ability to "learn" the appropriate

knowledge but in its ability to find associations. One facet of this challenge comes from identifying and analyzing gaps in a student's understanding or learning. For example, apart from simply scoring essays or responses across discrete dimensions such as fluency or sentence structure or by identifying keyspans (Mathias and Bhattacharyya, 2020; Takano and Ichikawa, 2022; Fiacco et al., 2022), one could use LLMs to determine which parts of a freeform submission indicate a gap and associate it with a learning goal provided by the teacher, without using specific (and costly to create) gold-labeled responses, so that the student has actionable feedback and can work on self-improvement. As part of this work, we need to accurately identify which portions of the response are written by the student as opposed to copied from an AI assistant. This would ensure that gaps aren't hidden, but would require a longitudinal view of the student's ability. Also, we must be able to ensure that the LLM's recommendations are based on actual details of the student and the text rather than being general predictions with high priors or based on hallucinations. Furthermore, rather than simplifying original lesson materials (Mallinson et al., 2022; Omelianchuk et al., 2021), we should invest in using LLMs to generate or retrieve materials or scaffolding that *help* to advance the students' learning rate.

7 What We Need

Our overall aim is to combat both the stultification of NLP as a mere evaluation optimization endeavor and to dispel fears that LLMs and generative AI will shut down the field. As an old saying goes, frequent moves make a tree die but a person prosperous. Just as NLP researchers in the 1980s had to learn about machine learning and then embrace it as a core technique in the field, so we now must explore and embrace LLMs and their capabilities. Machine learning did not 'solve' the challenges of NLP: it did not produce an engine that could learn languages, translate, answer questions, create poetry, and do all the things a child can do. Some people claim that LLMs can do all this, and more. But we are in the first flush of engagement, and have not yet have time to discover all their shortcomings.

Central is the challenge of scale. No child needs to read or hear more than half the internet's English text in order to use language. What reasoning and sensory capabilities do people have that LLMs lack? How can NLP research evolve to model and encompass those? We urgently need global infrastructures to dramatically scale up computing resources, because the open-source models still cannot achieve performance comparable to GPT variants (Gudibande et al., 2023). But we also urgently need deeper thinking about the foundational conceptual models driving our field.

During this unique period when NLP researchers feel uncertain regarding which research problems to pursue, we as a community need a collective effort to systematically change and refine our paper review system and academic success measurements, in order to establish a more inclusive research environment and encourage researchers (particularly those in junior positions) to explore long-term, high-risk topics that are crucial for the entire field. The new challenges also require us to be more open-minded to close collaboration with researchers from other fields, including social science, natural science, computer vision, knowledge representation and reasoning, and human-computer interaction.

Limitations

In this paper we describe some new or under-explored NLP research directions that remain dissertation-worthy. We propose a wider and exciting version of NLP that encourages people to focus on a wider range of more challenging and difficult problems with exciting potential impacts for social good. These problems may not always admit of easy datasets and pure machine learning solutions. Our list is not meant to be exhaustive, and we choose these directions as examples. It is up to NLP researchers to uncover the problems and develop novel solutions.

Ethical Considerations

The research areas listed in this document are a few of the main areas ripe for exploration; additional ones exist. We do not intend for our proposed positions to be forcefully pedagogical. We encourage diverse and deeper investigation of worthy research areas. Within these proposed directions, we acknowledge that some require access to users' personal information (e.g. chatbot personalization in Section 6.1), and some applications might have high impact on users (e.g. using models to assess a student's grasp of knowledge for targeted education

in Section 6.3). The use of LLMs for creative work has also led to concerns about copyright and regulations over whether AI can be credited as authors. We do not support the use of LLMs for screening or resource allocation purposes without safeguarding measures. Even for lower risk use cases, we opt for more research on the robustness, transparency, and fairness of systems. Finally, we must evaluate the compliance of prompting LLMs with laws and regulations. For instance in education applications, if we require information about the student, we must refer to laws such as FERPA/DPA/GDPR, especially in an online learning setting.

Acknowledgements

This work is based upon work supported by U.S. DARPA KAIROS Program No. FA8750-19-2-1004, U.S. DARPA CCU Program No. HR001122C0034, U.S. DARPA ECOLE Program No. HR00112390060, U.S. DARPA ITM FA8650-23-C-7316, U.S. DARPA SemaFor Program No. HR001120C0123 and U.S. DARPA INCAS Program No. HR001121C0165. The opinions, views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

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

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *ArXiv preprint*, abs/2211.15661.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards Language Models That Can See: Computer Vision Through the LENS of Natural Language. ArXiv:2306.16410 [cs].

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

George S Boolos, John P Burgess, and Richard C Jeffrey. 2002. *Computability and logic*. Cambridge university press.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg,

Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving emerging writers. In *CHI*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023) Findings*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *ArXiv preprint*, abs/2305.14908.

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 preprint*, abs/2107.03374.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. In *arXiv*, volume abs/2211.12588.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Lifan Yuan, Dehan Kong, Hanlu Wu, Ning Shi, Bo Yuan, Longtao Huang, Hui Xue, Zhiyuan Liu, Maosong Sun, and Heng Ji. 2023b. From adversarial arms race to model-centric evaluation: Motivating a unified automatic robustness evaluation framework. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. *ArXiv preprint*, abs/2301.12810.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *ArXiv preprint*, abs/2104.08696.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *ArXiv preprint*, abs/2303.03378.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. 2017. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194.

Guhao Feng, Yuntian Gu, Bohang Zhang, Haotian Ye, Di He, and Liwei Wang. 2023a. Towards revealing the mystery behind chain of thought: a theoretical perspective. *ArXiv preprint*, abs/2305.15408.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023b. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*.

James Fiacco, Shiyan Jiang, David Adamson, and Carolyn Rosé. 2022. Toward automatic discourse parsing of student writing motivated by neural interpretation. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 204–215, Seattle, Washington. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Yi R. Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. *ArXiv preprint*, abs/2305.14627.

Zhifan Gao, Yifeng Guo, Jiajing Zhang, Tieyong Zeng, and Guang Yang. 2023b. Hierarchical perception adversarial learning framework for compressed sensing mri. *IEEE Transactions on Medical Imaging*.

Katy Ilonka Gero, Tao Long, and Lydia Chilton. 2023. Social dynamics of ai support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *ArXiv preprint*, abs/2304.14767.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

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*, 9:346–361.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. In *International Conference on Learning Representations*.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Efficient training of bert by progressively stacking. In *International Conference on Machine Learning*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

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 preprint*, abs/2305.15717.

William H. Guss, Mario Ynocente Castro, Sam Devlin, Brandon Houghton, Noboru Sean Kuno, Crissman Loomis, Stephanie Milani, Sharada Mohanty, Keisuke Nakata, Ruslan Salakhutdinov, John Schulman, Shinya Shiroshita, Nicholay Topin, Avinash Ummadisingu, and Oriol Vinyals. 2021. The minerl 2020 competition on sample efficient reinforcement learning using human priors. In *NeurIPS2021*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Brett A. Halperin and Stephanie M. Lukin. 2023. Envisioning narrative intelligence: A creative visual storytelling anthology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023a. In-context learning of large language models explained as kernel regression. *ArXiv preprint*, abs/2305.12766.

Chi Han, Jialiang Xu, Manling Li, Yi R. Fung, Chenkai Sun, Tarek Abdelzaher, and Heng Ji. 2023b. Lm-switch: Lightweight language model conditioning in word embedding space. *ArXiv preprint*, abs/2305.12798.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *ArXiv preprint*, abs/2301.00303.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

William R Hersh and Ravi Teja Bhupatiraju. 2003. Trec genomics track overview. In *TREC*, volume 2003, pages 14–23.

Shaul Hochstein and Merav Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804.

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

Tom Hope, Doug Downey, Oren Etzioni, and Weld. 2022. A computational inflection for scientific discovery. *ArXiv preprint*, abs/2205.02007.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023. Zero-shot faithful factual error correction. *ArXiv preprint*, abs/2305.07982.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, and Shink Mori. 2023. A phd student's perspective on research in nlp in the era of very large language models. *ArXiv preprint*, abs/2305.12544.

Knud Illeris. 2018. A comprehensive understanding of human learning. In *Contemporary theories of learning*, pages 1–14. Routledge.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *ArXiv preprint*, 2208.

Peter Jarvis. 2006. *Towards a comprehensive theory of human learning*, volume 1. Psychology Press.

Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *ArXiv preprint*, abs/2304.09960.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2023. Vima: Robot manipulation with multimodal prompts. *ICML2023*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's up with vision-language models? investigating their struggle to understand spatial relations. In *EMNLP 2023*.

Jun Kato and Masataka Goto. 2023. Lyric app framework: A web-based framework for developing interactive lyric-driven musical applications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4110–4124, Online. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. Werewolf among us: A multimodal dataset for modeling persuasion behaviors in social deduction games. *ArXiv preprint*, abs/2212.08279.

Tuan M. Lai, Giuseppe Castellucci, Saar Kuzi, Heng Ji, and Oleg Rokhlenko. 2023a. External knowledge acquisition for end-to-end document-oriented dialog systems. In *Proc. The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL2023)*.

Tuan Manh Lai, ChengXiang Zhai, and Heng Ji. 2023b. Keblm: Knowledge-enhanced biomedical language models. *Journal of Biomedical Informatics*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Jangwon Lee. 2017. A survey of robot learning from demonstrations for human-robot collaboration. *arXiv preprint arXiv:1710.08789*.

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*, 35:34586–34599.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv preprint*, abs/2304.11633.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022a. Clip-event: Connecting text and images with event structures. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR2022)*.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022b. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proc. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022)*.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023c. Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. 2023. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 14846–14855.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Lihui Liu, Boxin Du, Yi R. Fung, Heng Ji, Jiejun Xu, and Hanghang Tong. 2021a. Kompare: A knowledge graph comparative reasoning system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, page 3308–3318. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021b. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. 2022. Mind's eye: Grounded language model reasoning through simulation. In *The Eleventh International Conference on Learning Representations*.

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2023a. A text-guided protein design framework. *ArXiv preprint*, abs/2302.04611.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Artur Lugmayr, Erkki Sutinen, Jarkko Suhonen, Carolina Islas Sedano, Helmut Hlavacs, and Calkin Suero Montero. 2017. Serious storytelling – a first definition and review. *Multimedia Tools and Applications*, 76:15707–15733.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *ArXiv preprint*, abs/2210.07229.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. *ArXiv preprint*, abs/2304.06708.

Yusuke Mori, Hiroaki Yamane, Ryohei Shimizu, and Tatsuya Harada. 2022. Plug-and-play controller for story completion: A pilot study toward emotion-aware story writing assistance. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 46–57, Dublin, Ireland. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

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

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. TEACh: Task-driven Embodied Agents that Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *ArXiv preprint*, abs/2304.03442.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *ArXiv preprint*, abs/2305.13048.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. 2013. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27:675–679.

Cheng Qian, Chi Han, Yi R. Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023. Creator: Tool creation for disentangling abstract and concrete reasoning of large language models.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. *ArXiv preprint*, abs/2208.11663.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Sheng Shen, Le Hou, Yan-Quan Zhou, Nan Du, S. Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *ArXiv*, abs/2305.14705.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. IEEE.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *ArXiv preprint*, abs/2209.05481.

Chenkai Sun, Jinning Li, Yi R. Fung, Hou P. Chan, Tarek Abdelzaher, Chengxiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shunya Takano and Osamu Ichikawa. 2022. Automatic scoring of short answers using justification cues estimated by BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 8–13, Seattle, Washington. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *ArXiv preprint*, abs/2305.04388.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent. *ArXiv preprint*, abs/2212.07677.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. 2021. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online. Association for Computational Linguistics.

Xingyao Wang, Sha Li, and Heng Ji. 2023a. Code4struct: Code generation for few-shot event structure prediction. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.

Xingyao Wang, Hao Peng, Reyhaneh Jabbarvand, and Heng Ji. 2023b. Leti: Learning to generate from textual interactions. *ArXiv preprint*, abs/2305.10314.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023c. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023d. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *ArXiv preprint*, abs/2301.11916.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023e. Self-consistency improves chain of thought reasoning in language models.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022a. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. 2023f. Paxion: Patching action knowledge in video-language foundation models. *ArXiv preprint*, abs/2305.10683.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. 2022b. Language models with image descriptors are strong few-shot video-language learners. In *Proc. The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS2022)*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022c. Language models with image descriptors are strong

few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497.

Ziqi Wang, Chi Han, and Heng Ji. 2023g. Understanding the effect of data augmentation on knowledge distillation. In *arXiv*, volume abs/2305.12565.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. In *Transactions on Machine Learning Research*.

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

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *ArXiv preprint*, abs/2303.03846.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *ArXiv preprint*, abs/2305.13252.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *ArXiv preprint*, abs/2304.11082.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023. WeCheck: Strong factual consistency checker via weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 307–321, Toronto, Canada. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Tong Xie, Yuwei Wa, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, and Bram Hoex. 2023. Large language models as master key: Unlocking the secrets of materials science with gpt. *ArXiv preprint*, abs/2304.02213.

Tianci Xue, Ziqi Wang, Chi Han, Zhenhailong Wang, Pengfei Yu, and Heng Ji. 2023. Combat hallucination in arithmetic reasoning with reverse chain-of-thought. In *arxiv*.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726.

Ke Yang, Charles Yu, Yi R. Fung, Manling Li, and Heng Ji. 2023b. Adept: A debiasing prompt framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10780–10788.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023a. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

Pengfei Yu and Heng Ji. 2023. Self information update for large language models through mitigating exposure bias. *ArXiv preprint*, abs/2305.18582.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback. *ArXiv preprint*, abs/2305.14002.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023c. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *ArXiv preprint*, abs/2305.17331.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R. Fung, Hao Peng, and Heng Ji. 2023. Craft: Customizing llms by creating and retrieving from specialized toolsets.

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

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. 2021. An attention free transformer. *ArXiv preprint*, abs/2105.14103.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. 2023. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2105.14103.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv preprint*, abs/2304.06364.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *ArXiv preprint*, abs/2012.00363.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.