Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models

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, Shinka Mori, Joan Nwatu,
Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, Rada Mihalcea
University of Michigan
LANGUAGE AND INFORMATION TECHNOLOGIES (LIT)
{oignat,jinzhi,mihalcea}@umich.edu

Abstract

Recent progress in large language models (LLMs) has enabled the deployment of many generative NLP applications. At the same time, it has also led to a misleading public discourse that "it's all been solved." Not surprisingly, this has, in turn, made many NLP researchers – especially those at the beginning of their careers – worry about what NLP research area they should focus on. *Has it all been solved, or what remaining questions can we work on regardless of LLMs?* To address this question, this paper compiles NLP research directions rich for exploration. We identify fourteen different research areas encompassing 45 research directions that require new research and are not directly solvable by LLMs. While we identify many research areas, many others exist; we do not cover areas currently addressed by LLMs, but where LLMs lag behind in performance or those focused on LLM development. We welcome suggestions for other research directions to include: https://bit.ly/nlp-era-llm

Keywords: Large language models, challenges for NLP, open questions, applied NLP, responsible NLP, fundamental NLP

1. Background

Language models represent one of the fundamental building blocks in NLP, with their roots traced back to 1948 when Claude Shannon introduced Markov chains to model sequences of letters in English text (Shannon, 1948). They were then heavily used in connection with early research on statistical machine translation (Brown et al., 1988; Wilkes, 1994) and statistical speech processing (Jelinek, 1976). While these models have always been an integral part of broad application categories such as text classification, information retrieval, or text generation, only in recent years have they found a "life of their own" with widespread use and deployment.

The impressive advancements we have witnessed in current large language models (LLMs) directly result from those earlier models. They build on the same simple yet groundbreaking idea: given a series of previous words or characters, we can predict what will come next. The new LLMs benefit from two main developments: (1) the proliferation of Web 2.0 and user-generated data, which has led to a sharp increase in the availability of data; and (2) the growth in computational capabilities through the introduction of GPUs. Together, these developments have facilitated the resurgence of neural networks (or deep learning) and the availability of very large training datasets for these models.
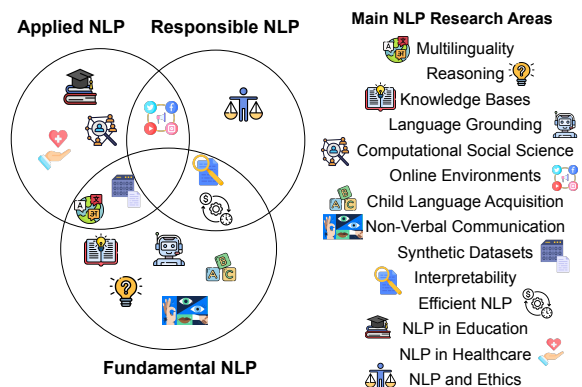
Current LLMs have output quality comparable to

Figure 1: Research areas in NLP that are rich for exploration, spanning Fundamental, Responsible, and Applied research.

human performance, with the added benefit of integrating information from enormous data sources, far surpassing what one individual can accumulate in their lifetime. The number of applications that benefit from using LLMs is continuously growing, with many cases where the LLMs are used to replace entire complex pipelines. LLMs becoming "lucrative" has led to a surge in industry interest and funding, alongside a sharp increase in the number of research publications on LLMs.

While these advances in LLMs are very real and truly exciting, and give hope for many new generative language applications, LLMs have also "sucked the air out of the room." A recent funding call from DARPA has completely replaced the term

---

*Oana Ignat and Zhijing Jin contributed equally to the manuscript.

NLP with LLM: in their listing of experts sought for the program, we see the fields of "Computer Vision" and "Machine Learning" listed alongside "Large Language Models" (but not "Natural Language Processing").* Replacing NLP with LLMs is problematic for two main reasons. First, the space of language insights, methods, and broad applications in NLP is much vaster than what can be accomplished by simply predicting the next word. Second, even if not technologically new, LLMs still represent an exclusionary space because of the amount of data and computation required to train.

This public discourse that often reduces the entire field of NLP to the much smaller space of LLMs is not surprisingly leading to a dilemma for those who have dedicated their careers to advancing NLP research, and especially for junior PhD students who have only recently embarked on the path of becoming NLP researchers. "***What should I work on?***" is a question we hear now much more often than before, often as a reaction to the misleading thought that "it's been all solved."

The reality is that there is much more to NLP than just LLMs. This paper aims to answer the question: "What are rich areas of exploration in the field of NLP that could lead to a PhD thesis and cover a space that is not within the purview of LLMs." Spoiler alert: there are many such research areas!

About This Document. This document reflects the ideas about "the future of NLP research" from the members of an academic NLP research lab in the United States. The Language and Information Technologies (🔥LIT) lab at the University of Michigan includes students at various stages in their degree, starting with students who are about to embark on a Ph.D., all the way to students who recently completed a Ph.D. degree. The LIT students come from a wide variety of backgrounds, including China, Iran, Japan, Mexico, Nigeria, Romania, Russia, South Korea, the United States, and Uruguay, reflecting a very diverse set of beliefs, values, and lived experiences. Our research interests cover a wide range of NLP areas, including computational social science, causal reasoning, misinformation detection, healthcare conversation analysis, knowledge-aware generation, commonsense reasoning, cross-cultural models, multimodal question answering, non-verbal communication, visual understanding, and more.

We provide a list of open research questions that are not solved by LLMs. As showed in Figure 1, we cover three major categories, from fundamental NLP (Section 2), to responsible NLP (Section 3), and applied NLP (Section 4). Spanning across the

three categories, we cover 14 open research topics, each with three to four specific research directions. Note that when a research topic touches multiple categories, for convenience, we list them under the major one.

When compiling the ideas in this document, we followed three main guiding principles. First, we aimed to identify areas of NLP research that are rich for exploration, e.g., areas one could write a Ph.D. thesis on. Second, we wanted to highlight research directions that do not have a direct dependency on a paid resource; while the use of existing paid APIs can be fruitful for certain tasks, such as the construction of synthetic datasets, building systems that cannot function without paid APIs is not well aligned with academic core research goals. Third, we targeted research directions that can find solutions with reasonable computational costs achievable with setups more typically available in academic labs. Finally, we found inspiration in the ACL list of research areas, from which we selected the ones not in the purview of LLMs.*

Our brainstorming process started with ideas written on sticky notes by all the authors of this document, followed by a "clustering" process where we grouped the initial ideas and identified several main themes. These initial themes were then provided to small groups of 2–3 students, who discussed them, expanded or merged some of the themes, and identified several directions worthy of exploration. The final set of themes formed the seed of this document. Each research area has then had multiple passes from multiple students (and Rada) to delineate the background of each theme, the gaps, and the most promising research directions.

Disclaimer. The research areas listed in this document are just a few of the areas rich in exploration; many others exist. In particular, we have not listed the numerous research directions where LLMs have been demonstrated to lag in performance (Bang et al., 2023a), including information extraction, question answering, text summarization, and others. We have also not listed the research directions focused on LLM development, as that is already a major focus in many current research papers, and our goal was to highlight the research directions other than LLM development. We welcome suggestions for other research areas or directions to include: https://bit.ly/nlp-era-llm

Document Organization. We provide a list of open research questions that are not solved by LLMs. As showed in Figure 1, we cover three major categories, from fundamental NLP (Section 2),

---

to responsible NLP (Section 3), and applied NLP (Section 4). Spanning across the three categories, we cover 14 open research topics, each with three to four specific research directions. Note that when a research topic touches multiple categories, for convenience, we list them under the major one.

The following sections provide brief descriptions of fourteen research areas rich in exploration, each with 3–4 research directions. These areas can be divided into areas that cannot be solved by LLMs for being too data-hungry or for lacking reasoning or grounding abilities (subsections 2.1–2.5, 3.3, 4.3); areas for which we cannot use LLMs because of not having the right data (subsections 2.6, 4.1, 4.2); or areas that could contribute to improving the abilities and quality of LLMs (subsections 3.1, 3.2, 3.4, 4.4). When compiling the research directions, we follow several guiding principles. First, we aim to identify areas that are rich for exploration, that one could write a PhD thesis on. Second, we want to highlight research directions that do not directly depend on a paid resource, such as a paid API. Third, we target research directions that can find solutions with reasonable computational costs, achievable in academic labs. Finally, we find inspiration from ACL 2018 list of research areas, from which we select the ones not in the purview of LLMs (15/21 areas). The mapping of our research areas and the ACL 2018 tracks can be found in Appendix Table 1.

2. Fundamental NLP

Fundamental NLP tasks represent a significant subset of NLP research, as illustrated in Figure 2. Among these, we first consider different "L's in "NLP," namely different choices of languages. Although most NLP tasks and datasets use English as a medium, there is a growing trend to extend NLP to more non-English languages (Section 2.1), child language (Section 2.5), and non-verbal communication (Section 2.6). Moreover, there are different "P"s in "NLP" too, where we consider different types of processing tasks on text, such as reasoning (Section 2.2), knowledge bases (Section 2.3), and language grounding (Section 2.4).

In the following, we present the main research directions for each of these research areas.

2.1. Multilinguality

Low-resource machine translation. Despite the impressive performance of machine translation (MT) on major languages (Hassan et al., 2018a; Liu et al., 2020), there is a big gap when it comes to low-resource languages. There has been a rise in small benchmarks dedicated to low-resource languages (Vegi et al., 2022; Reid et al., 2021; Goyal

et al., 2021), but there is also a big need for large training corpora. As many low-resource languages do not have a significant web presence, alternative solutions are needed, such as manually curated parallel corpora (Zheng et al., 2022; Koto and Koto, 2020), OCR (Rijhwani et al., 2020; Ignat et al., 2022), or translation dictionaries using models of word formation (Wu and Yarowsky, 2018, 2020).

Multilingual models that work well for all languages. Although most recent LLMs claim to be multilingual, they do not perform equally well in all languages (OpenAI, 2023; Huang et al., 2023a; Ahuja et al., 2023). This inequality stems from the different proportions of text of different languages in the training corpora, as well as the annotators with demographics focused in a few countries in the RLHF process to finetune the models (Ouyang et al., 2022b). At present, general-purpose LLMs do not perform as well as models trained specifically for translation; future research can explore incorporating off-the-shelf LLMs into MT systems.

Code-switching. Code-switching refers to text involving expressions in several languages while adhering to the grammatical structure of at least one language. Challenges include the large variation of code-switching phenomena, lack of training data, and a large number of out-of-vocabulary tokens (Çetinoğlu et al., 2016). Open research directions include synthetic data generation (Xu and Yvon, 2021; Fang and Wu, 2022; Lee and Li, 2020), evaluating existing LLMs on code-switched text across language combinations (Aguilar et al., 2020; Khanuja et al., 2020), and distinguishing highly similar languages, such as dialects of the same parent language (Aguilar et al., 2020).

2.2. Reasoning

Complex reasoning. Complex and multi-step reasoning has proven to be challenging for LLMs. For instance, LLMs still fall short in numerical reasoning (Stolfo et al., 2023; Miao et al., 2020), logical reasoning (Jin et al., 2022c; Eisape et al., 2023), grounded reasoning (Ignat et al., 2021), and causal inference (Jin et al., 2023a,c), often making obvious mistakes (Goel et al., 2021; Jin et al., 2020b). One reason for that is that the next-word prediction objective can easily encourage the LM to assign a high likelihood to invalid reasoning (Khalifa et al., 2023). Even fully-supervised training over correct reasoning demonstrations does not solve the issue (Uesato et al., 2022). While scaling seems to help, careful prompt engineering is still needed to tease out correct reasoning (Wei et al., 2022c; Fu et al., 2022; Zhou et al., 2022b; Zhang et al., 2022c). To build LLMs that are robust at reasoning, one could explore a variety of directions, such as

combining the strengths of neural networks and symbolic AI. Another growing direction is the integration of LLMs with external reasoning tools, such as calculators, interpreters, database interfaces, or search engines (Schick et al., 2023; Mialon et al., 2023a).

Responsible reasoning in social contexts. With an increasing number of applications that use NLP models, it is foreseeable that models will need to make complicated decisions that involve moral reasoning as intermediate steps. For example, when creating a website, there may be moral choices to consider such as catering to certain sub-populations, or overly optimizing for user attention or click-through rates. These decision principles are pervasive in our daily life, across small and large tasks. We believe there is much to be studied in understanding or improving the ability of AI systems to reason over socially-complicated and morally-charged scenarios given different social contexts and cultural backgrounds (Jin et al., 2023b; Hendrycks et al., 2021a; Liu et al., 2021). We foresee that interdisciplinary collaboration with domain experts and policymakers will be needed.

Formally defining reasoning and designing a proper evaluation framework. There is a rising need to refine the definition of reasoning, as LLMs start showing an increasing mastery of templated solutions through pattern matching – when a model memorizes a reasoning pattern, does it count as reasoning or knowledge? Fundamentally, this leads to questions about what are the domains of intelligence that humans excel at, and how different these are from empirically learning how to do template matching. Beyond redefining reasoning, other open questions include how to test a model's reasoning skills in the face of data contamination, Goodhart's law (a dataset failing to reflect the skill once exploited) (Goodhart, 1984), and a lack of reliable metrics to evaluate multi-step reasoning (Golovneva et al., 2022).

2.3. 📖 Knowledge Bases

Automatic knowledge base construction. Specialized knowledge bases are helpful resources for domain-specific applications. Successful automatic knowledge base construction can take up-to-date text in free forms (Maedche and Staab, 2000), and adapt an ontology for complex applications, such as tracking medication interactions from articles from PubMed (Xu et al., 2020).* However, this task faces many challenges, such as knowledge coverage, factuality of the knowledge, and knowledge linking, which are rich, open areas of research. Specifically, Wang et al. (2023a)

---

*https://pubmed.ncbi.nlm.nih.gov/

shows that ChatGPT performs poorly on out-of-distribution data, such as new medical diagnosis and product review datasets. Also, the training data cutoff limits the coverage of new concepts. Specifically concerning factuality, KG completion framed as a text generation task also suffers from hallucination from the LLM in various tasks (Ji et al., 2022).

Knowledge-guided NLP. As NLP models become more powerful through exposure to massive pretraining corpora (Hoffmann et al., 2022a; Wei et al., 2022a), researchers start to question whether mere pretraining is sufficient, as models suffer heavily from hallucination (Xiao and Wang, 2021; Dziri et al., 2022a). A rising research question is how to efficiently and effectively interact with external knowledge bases (Zhang et al., 2019), such as through web browsing (Nakano et al., 2021; Komeili et al., 2022; Schick et al., 2023) and customized knowledge base lookup (Wilmot and Keller, 2021; Mialon et al., 2023b).

Culture-specific knowledge and common sense. Knowledge and common sense in NLP models are usually dominated by a few Western cultures, and do not account for the vast diversity of the cultural views in the world (Arora et al., 2023). The first step is to understand the limitations of NLP models, including LLMs, with respect to their knowledge of different cultural groups (Hovy and Yang, 2021; Hershcovich et al., 2022; Arora et al., 2023). Once these limitations are better understood, a major open research direction is how to acquire and represent the knowledge that encodes these cultural views, as well as how and when to invoke this cultural knowledge.

2.4. 🤖 Language Grounding

Fusing multiple modalities. Efficiently and effectively combining different modalities, i.e., audio, video, text, and others, is still an open problem. Different modalities often complement each other, thus potentially reducing the need for billions of data points. However, in some cases, modalities end up competing with each other (Yao and Mihalcea, 2022), and thus many uni-modal models outperform multi-modal models (Wang et al., 2019b; Huang et al., 2021).

Grounding for less studied modalities. Most work on grounding revolves around visual, textual, or acoustic modalities. However, less-studied modalities, such as physiological, sensorial, or behavioral, have been found valuable in diverse applications, including measuring driver alertness (Jie et al., 2018; Riani et al., 2020), detecting depression (Bilalpur et al., 2023), or detecting deceptive

behaviors (Abouelenien et al., 2016). Current multimodal LLMs are restricted to textual, audio, and visual domains (Zhang et al., 2023a; Lyu et al., 2023a), requiring much effort to integrate the less-studied modalities.

Grounding "in the wild" and for diverse domains. Most research around grounding uses data collected indoors in the lab, or on images and videos of indoor activities from sources such as movies (Lei et al., 2019a) or online vlogs (Ignat et al., 2019). There are fewer studies on outdoor activities in more realistic "in the wild" settings (Castro et al., 2022). Collecting such data poses new challenges related to availability, quality, or distribution, which opens up new research directions. Moreover, applying these models to diverse domains (e.g., robotics, medicine, and education) requires adapting to fewer data points or different types of data, and adding in-domain expertise to understand the problem setup better. As shown in Yin et al. (2023), the multimodal LLMs are currently not equipped to tackle these challenges.

2.5. Child Language Acquisition

Sample-efficient language learning. Child language acquisition is both an important research topic in psycholinguistics (McNeill, 1970), and also a valuable source of inspiration for sample-efficient language learning for NLP (Linzen, 2020). By mimicking the learning strategies of children, models can achieve better generalization with limited training data (Barak et al., 2016). Research in this area brings hope to improve the performance of NLP models while reducing the amount of training data required (Gulordava et al., 2018; Warstadt et al., 2023a,b). LLMs require far more data than children to acquire language, and LLMs can be improved in sample efficiency by learning how children acquire language.

Language models as biological models for child language acquisition. Since the last century, there has been research using neural models as biological models to develop theories of human cognitive behavior (McCloskey, 1991). Combining powerful LLMs with psycholinguistic studies, researchers can gain inspiration for various processes in child language acquisition. For instance, insights into word acquisition can be gained by comparing the models' learning curves and children's age of acquisition for different words (Chang and Bergen, 2021). Other phenomena, such as phoneme-level acquisition (Christiansen and Chater, 1999; Martin et al., 2023) or intrinsic rewards (Gibson et al., 2019; Mu et al., 2022), can also be explored by using computational models

on existing benchmarks such as WordBank (Frank et al., 2016) or CHILDES (MacWhinney, 1992)

Benchmark development in child language acquisition. While there are currently only very few language acquisition benchmarks, NLP and multimodal systems bring opportunities to ease and scale child language benchmark construction. For example, controlled experiments on carefully-constructed supervised benchmarks can be augmented by large video datasets of children learning a language over a long period of time.

2.6. Non-Verbal Communication

Non-verbal language interpretation. Non-verbal language interpretation analyzes non-verbal cues such as facial expressions, gestures, and body language to enhance the performance of NLP systems (Mavridis, 2015; Schuller, 2018). For instance, while previous work has identified a potential "code-book" of facial expressions (Song et al., 2013), it remains an open research direction how to determine the set of expressions and gestures that can be used across modalities, contexts, and culture (Matsumoto and Assar, 1992; Abzaliev et al., 2022). Currently, there are no LLMs that combine the gesture modality with the text.

Sign language. As a visual-gestural communication system, sign language has gained increasing attention in NLP due to its unique challenges and wide applications (Koller et al., 2016, 2018; Camgoz et al., 2020a). There are many research directions in sign language, such as data curation and evaluation addressing the high variability in manual gestures (Athitsos et al., 2008b; Li et al., 2020a), incorporation of additional information, i.e., facial expressions, body pose, eye gaze (Cao et al., 2018; Baltruaitis et al., 2018); and sign language generation for various scenarios, such as speakers of the same sign language, across different sign languages, and with a combination of verbal and sign languages (Adaloglou et al., 2022). Current systems use separate models for translating sign language into the text (Lim et al., 2023), which is then provided to LLMs. Directly providing sign language to LLM might be more efficient.

Joint verbal and non-verbal communication. Ultimately, both verbal and non-verbal signals should be considered during communication. Future AI systems should be equally capable of understanding "I don't know", shrugging the shoulders, or ¯_(ツ)_/¯ . Representing, fusing, and interpreting these signals jointly is ultimately the long-term goal of AI-assisted communication (Mavridis, 2014). Open research problems encompass not only the development of language models for each

of these modalities but effective fusion methodologies that enable large joint models for simultaneous verbal and non-verbal communication.

3. Responsible NLP

With NLP models in more applications, it is crucial to promote responsible NLP via ethical considerations (Section 3.1), interpretability (Section 3.2), green/efficient NLP (Section 3.3), and careful use of NLP in online environments (Section 3.4).

3.1. ⚖ NLP and Ethics

Dual use. Many NLP applications that have a positive impact can at the same time be used in harmful ways (Hovy and Spruit, 2016; Wu et al., 2023b). Donker (2023) highlights an instance where LLMs were used to generate erroneous review reports, thereby disrupting the standards for peer review. The research community should foster interdisciplinary collaboration to fight against malicious applications of NLP technologies, such as deceptive text generation, automated disinformation campaigns, LLM-powered cybersecurity attacks (Kang et al., 2023), and arms racing.

Fairness. There is a need for methods that evaluate the fairness of NLP models, and mitigate their bias. Generative language models have been found to manifest harmful stereotypes in downstream tasks such as automated reference letter writing (Wan et al., 2023). While fairness in LLMs is an active area of research, much existing work focuses on limited aspects such as binary gender. Mitigating bias will involve investigating dataset creation practices and their correlation with model bias (Wang et al., 2020). Such research should examine whether stricter requirements for data creation can reduce bias and inequalities that might be exacerbated by models trained on or evaluated on biased data (Anonymous, 2024).

Privacy. With the increasing use of LLMs for personalized NLP applications, concerns have been raised regarding access to user data through LLMs, especially in sensitive areas like healthcare (Meskó and Topol, 2023; Marks and Haupt, 2023). In response, researchers are investigating privacy-preserving methods such as differential privacy (Dwork, 2006), federated learning (McMahan et al., 2017), and secure multi-party computation to ensure the confidentiality and security of user data (Lindell, 2021).

Attribution of machine-generated data. The use of generative LLMs in the creative industry has led to issues such as lack of copyright, plagiarism, and profit shifting. Text generated by LLMs can reveal sensitive or copyrighted contents from their training data. Therefore, it is essential to develop standard approaches (e.g., membership inference (Shokri et al., 2017)) for attribution that NLP models can use while generating content (Collins, 2023), especially for domains such as programming or creative writing (Swanson et al., 2021), where LLM-generated content is on the rise.

3.2. 🔍 Interpretability

Probing. One promising direction to investigate is the internal representation of NLP models, including LLMs, by designing probing tasks that can reveal the linguistic (Hewitt and Manning, 2019; Hewitt and Liang, 2019) and world knowledge captured by the models (Elhage et al., 2022; Geva et al., 2021, 2022b). This can help understand the reasoning capabilities of models and identify potential biases (Li et al., 2022a; Meng et al., 2022a).

Mechanistic interpretability. While probing mostly considers the attributes of the features learned by the model, there are currently several open research questions around mechanistic interpretability, which aims to uncover the underlying *mechanisms and algorithms* within a model that contributes to its decision-making process (Nanda et al., 2023; Conmy et al., 2023). These models aim to extract computational subgraphs from neural networks (Conmy et al., 2023; Wang et al., 2023b; Geiger et al., 2021), and their high-level goal is to reverse engineer the entire deep neural network (Chughtai et al., 2023).

Human-in-the-loop to improve interpretability. Incorporating human feedback to enhance model interpretability can improve model transparency, facilitate better decision-making, and foster trust between AI systems and users. By involving humans, researchers can identify and address biases, ensure ethical considerations, and develop more reliable and understandable NLP models. There are various promising directions, such as active learning and interactive explanation generation (Mosca et al., 2023; Mosqueira-Rey et al., 2023a).

Basing the generated text on references. As model-generated text is prone to hallucinations (Ji et al., 2022), a promising way to improve its reliability is to explain its conclusion step by step and supply references or sources to back up the claims (Wei et al., 2022d; Izacard et al., 2022).

3.3. 🌱 Green/Efficient NLP

Model efficiency. The trend of scaling up NLP models has accentuated the need of increasing model efficiency (Strubell et al., 2020; Hessenthaler et al., 2022). Researchers have proposed efficiency enhancement techniques from various

aspects, such as improving the attention mechanisms (Tay et al., 2020, 2022; Dao et al., 2022; Ma et al., 2022), sparsing models to scale up the width of models for increased expressiveness while reducing theoretical FLOPs, and applying mixture-of-experts architectures (Fedus et al., 2021, 2022; Du et al., 2022). However, there are open challenges in how to develop the optimal architectures that balance economics, efficiency, and performance (Mustafa et al., 2022).

Efficient downstream task adaptation. Increased applicability of pre-trained models requires efficient fine-tuning methods that adapt to downstream tasks (Lee et al., 2020; Liu et al., 2023; Wu et al., 2023a) by updating a small subset of the parameters (Pfeiffer et al., 2020; Moosavi et al., 2022; Schick and Schütze, 2021; Hu et al., 2023). For example, prompt-tuning/ prefix-tuning modifies activations with additionally learned vectors without changing model parameters (Valipour et al., 2022; Lester et al., 2021).

Data efficiency. Another method to improve efficiency is to remove redundant or noisy data in the first place. Despite existing efforts on removing noisy examples and deduplicating data on smaller scales (Lee et al., 2022b; Mishra and Sachdeva, 2020; Hoffmann et al., 2022b), there is a lack of effective methods for data deduplication and curation for vast corpora (>700B Tokens) or raw web data used for training very big models.

3.4. NLP for Online Environments

Combating misinformation. With the rising capability of text generation models and strong commercial and political interests, it is inevitable to see an increase in online misinformation. To fight back against powerful generative models to spread misinformation, we need to use powerful discriminative models to detect them. There is a large need for fact-checking technology (Thorne et al., 2018; Nakov et al., 2021; Kazemi et al., 2022), across different languages (Das et al., 2023), different modality (Abdelnabi et al., 2022), and by utilizing techniques from other areas, such as network analysis to track who likes or reposts false contents (Guarino et al., 2020), and retrieval and knowledge-augmented methods (Ciampaglia et al., 2015; Markov et al., 2023) to search through and find the relevant context around the claim. One caution is that LLMs are prone to hallucinations (Dziri et al., 2022b; Raunak et al., 2021) and factual inconsistencies (Tam et al., 2023), so they might not be self-sufficient to combat misinformation reliably.

Ensuring content diversity. With the prevalence of LLM-generated content, the majority's voice may end up amplified on the web, since data-driven models such as LLMs tend to remember the type of data that is the most represented in its corpus. Thus, the lack of diversity and especially representation of marginalized groups' voices will be a concerning problem as LLM-generated content will be increasingly used online (Field et al., 2021a).

Preventing mis- and over-moderation. Similar to the heterogeneity issue in content generation, content moderation techniques might also overlook the nuances of expressions in under-represented groups, or specific social environments, making them unfairly delete safe speech by minority groups (Sap et al., 2019a; Xia et al., 2020). Apart from mis-moderation, there is also over-moderation. Due to various political interests (e.g., Florida aiming to curtail discussions about race or queer identities), governments are likely to limit the set of topics discussed online, so it is important to trace what topics and opinions are filtered or demoted on the internet and reflect on the freedom of speech in the political environment (Wright, 2006; Gorwa et al., 2020).

4. Applied NLP

After discussing tasks in fundamental NLP and responsible NLP, we now look into the wide applications of NLP in various domains, with a few selected discussions on NLP for healthcare (Section 4.1), education (Section 4.2), computational social science (CSS) (Section 4.3), and synthetic data generation (Section 4.4).

4.1. NLP for Healthcare

Healthcare benchmark construction. Healthcare is a domain that heavily suffers from data scarcity, which is usually due to data unavailability (typically for low-resource domains), or inaccessibility (due to privacy and ethics concerns). Potential strategies to create and scale-up health datasets include synthetic data generation (Chintagunta et al., 2021a; Liednikova et al., 2020; Mattern et al., 2022b) or data augmentation from existing data (Dai et al., 2023). These strategies can improve the distribution of biased datasets, help ensure data privacy protections, and reduce the cost of data collection. However, data generation by LLMs also brings concerns of bias propagation and information leakage (Arora and Arora, 2022). Furthermore, researchers need metrics to measure the fidelity of synthetic data compared with real data (Chen et al., 2021).

Improving clinical communication. NLP has shown great potential in enhancing communication in healthcare, such as simplifying the medical jargon for laymen (Jin et al., 2022a), developing

educational tools for healthcare professionals (Min et al., 2022), and providing personalized healthcare recommendations (Choi et al., 2016; Roehrs et al., 2018). New research directions include developing advanced NLP models for medical dialogue systems and exploring the ethical implications of NLP-driven communication in healthcare (Ravi et al., 2016; Jakesch et al., 2019). Current LLMs may only be useful in limited settings, as trust in LLMs has been shown to depend on the health-related complexity of questions (Nov et al., 2023).

Drug discovery. Since the hypothesis space for drug designs is exponential (Ruddigkeit et al., 2012), NLP methods have been explored to assist clinicians to efficiently extract and analyze information from large amounts of scientific literature, patents, clinical records, and other biomedical sources. Open research directions in this domain include identifying and prioritizing the drug-target interactions, discovering new drug candidates, predicting compound properties, and optimizing drug designs (Brown et al., 2020a). Despite their great potential, the use of LLMs still face many challenges, such as the lack of transparency in the model decision-making process, which limits the applicability and reliability (Thirunavukarasu et al., 2023).

4.2. 🎓 NLP for Education

Intelligent tutoring systems. The rising capability of NLP systems has given rise to intelligent tutoring systems to generate targeted practice questions and explain students' mistakes in a wide range of subjects, from English or History to Physics or Computer Science (Mousavinasab et al., 2021). Responsible development of these systems requires human-in-the-loop checks to ensure their reliability, as NLP models are still lacking when it comes to more challenging reasoning and grounding tasks (Kanda et al., 2004). Other challenges include lack of diverse data, both in terms of population and time, privacy concerns and trustworthiness, and the need for better evaluation mechanisms (Lin et al., 2023).

Educational explanation generation. To enrich teaching materials, NLP models can also help generate explanations for complicated questions or reading materials, as well as for automatic grading systems, since students improve more easily when grading is justified by corresponding explanations (Mohler and Mihalcea, 2009). However, some concerns include overreliance on the model, lack of expertise among educators (Redecker and Punie, 2017), and between real knowledge and convincingly written but unverified model output (Kasneci et al., 2023). Therefore, it is important to under-stand the limitations of LLMs and use them only as a tool to support and enhance learning, but not as a replacement for human teachers (Pavlik, 2023).

Controllable text generation. In education, there is a growing need for controllable text generation (Lee et al., 2011; Zhang et al., 2022a). This is helpful, for instance, for applications aiming to introduce students to new terms by generating memorable stories corresponding to their academic skill levels, interests, and prior experience. However, it is often difficult for LLMs to ensure domain diversity of the generated text while pursuing controllability, which leads to the catastrophic forgetting problem in LLMs (Zhai et al., 2023). Additionally, we lack reliable evaluation techniques, as well as dedicated benchmarks and datasets for text generation with diverse control requirements (Zhang et al., 2023b).

4.3. 🕵️ Computational Social Science

Development of new abstractions, concepts, and methods. NLP enables automatic analyses of massive text for the study of computational social science (CSS), which has been benefited by the evolution of NLP methods from topic modeling (Blei et al., 2003), keyword extraction (Onan et al., 2016), to word embeddings (Pennington et al., 2014), and LLMs (Brown et al., 2020c). It is foreseeable that further advancement in NLP models will unlock the possibilities of more customized, high-level text analyses for CSS. Evaluation paradigms need to evolve to capture the validity of LLMs as language generators, since human evaluation also can be unreliable in CSS (Karpinska et al., 2021). Moreover, many CSS tasks contain large target label spaces (Grudin, 2006), which is a challenge for current LLMs that have limited memory and quadratic space complexity (Ziems et al., 2023).

Population-level data annotation and labeling. CSS research shows a large interest in using LLMs to annotate data to simulate human interactions (Gilardi et al., 2023a). However, human studies will be unlikely to go away, as LLMs' effectiveness in annotation remains partial. Ollion et al. (2023) show that few-shot and zero-shot models are often outperformed by models fine-tuned with human annotations. Additionally, ChatGPT usually yields higher recall than precision, showing a tendency to output more false positives.

Multicultural and multilingual CSS. Most CSS studies focus on English or other major languages, and address mostly Western cultures. However, there are many important questions in social science that require large-scale, multilingual, and multicultural analyses (Shen et al., 2019), such as how languages evolve, and how values vary across

cultures (Garimella et al., 2016). This area for future work can lead to compounding impacts on the social sciences. However, the data-driven nature of LLMs makes them limited by the under-representation of minority communities and low-resource languages in the training data. Collecting more data related to this can help minimize the data disparity. Additionally, CSS researchers study cultures, norms, and beliefs that change across time, hence LLMs will need a high level of temporal grounding (Ziems et al., 2023).

4.4. 📑 Synthetic Datasets

Knowledge distillation. Knowledge distillation transfers knowledge from larger, more complex models (the teacher) to typically smaller, simpler models (the student). Knowledge distillation allows the knowledge and capability to be compressed into much smaller models, reducing the computational and memory requirements of NLP systems. While earlier methods in knowledge distillation often learn from the soft output logits of teacher models (Hinton et al., 2015), more recent ones utilize LLM outputs as synthetic examples (West et al., 2022; Kim et al., 2022). This allows practitioners to transform or control the generated data in different ways, such as using finetuned models to filter for quality. Moreover, synthetic data can be used to directly emulate the behavior of LLMs with much smaller, focused models (Taori et al., 2023).

Control over generated data attributes. Currently, the predominant method is to provide natural text specifications with instructions and examples, but optimizing these prompts often relies on a simple trial-and-error approach. Additionally, specifying attributes through instructions or examples can be imprecise or noisy. The development of robust, controllable, and replicable pipelines for synthetic data generation remains an open research question (Kim et al., 2022).

Transforming existing datasets. Given an existing dataset, one can apply various changes to create a semantically preserving new dataset, but with a new style. Common approaches include format change (e.g., converting a dataset of HTML news articles to plain text), modality transfer (e.g., generating textual descriptions of images or videos or generating captions or subtitles for audiovisual content), or style transfer (Chintagunta et al., 2021b; Jin et al., 2022a) (e.g., translating the writing style of the text from verbose to concise).

5. So "What Should *I* Work On?"

The future of NLP research is bright ✨. As illustrated by the 45 research directions spanning the
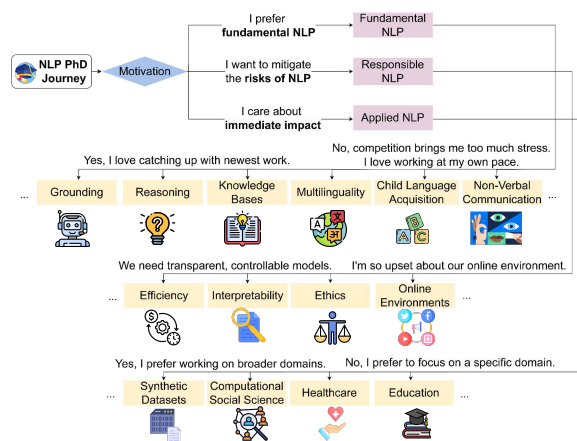

Figure 2: So "What should I work on?" Based on your *motivation*, you can choose from at least three main NLP areas: Fundamental, Responsible, and Applied. Furthermore, your choice could depend on whether or not you like fast-paced research, whether you want to make models transparent and trustworthy, or, whether you prefer broader or more specific domains.

fourteen research areas overviewed in this paper, the rapid progress we are currently witnessing in LLMs does not mean that "it's all been solved." On the contrary, numerous research directions within NLP are not solved by the current LLMs. They add to the many existing tasks in NLP where LLMs' performance is limited (Bang et al., 2023a), as well as the growing number of new areas enabled by the new LLM capabilities.

More broadly, as a field, we now have the opportunity to move away from performance-focused technology development and acknowledge that NLP is about language *and* people. This brings about a new focus on enabling technologies that are culture and demographic aware, that are robust, interpretable, efficient, and aligned with solid ethical foundations — ultimately, technologies that make a lasting positive impact on society.

How to choose a research direction to work on? As suggested in Figure 2, start with your *motivation* and *interests*: consider your previous experiences, look around at your community, explore your curiosities about language and about people, and try to find what resonates with you the most. Building on this foundation, identify the tasks that connect to your motivations. This paper serves as a starting point to inspire this exploration.

Broader Impact

We believe this work and the open research directions we identified can have an overall positive impact on the NLP research community, especially for junior students facing the challenge of re-orienting

their research directions in the era of LLMs.

We conclude by highlighting what we foresee as the main role of this paper. First, we did not aim to cover the entire rich space of NLP, which is impossible for any research lab to enumerate exhaustively. Instead, we provided a starting point for students and researchers to regain their hope in NLP research, and find a direction they can contribute to that is not solved by LLMs. Second, this overview paper did not aim to solve any of the tasks we listed, but rather to identify the open space for future work. We thus did not provide full details for the research directions; instead, we introduced each research direction with a brief description, its broad application, and highlight the remaining challenges and open questions, especially those that are not addressed by LLMs. Our main goal is to inspire future researchers to deepen their exploration on the topics.

We welcome suggestions for other research areas or directions to include: `https://bit.ly/nlp-era-llm`.

Acknowledgments

6. Bibliographical References

a. ChatGPT plugins. `https://openai.com/blog/chatgpt-plugins`. Accessed: 2023-4-24.

b. Duolingo - learn a language for free @. `https://www.duolingo.com/`. Accessed: 2023-4-24.

c. Grammarly: Free writing AI assistance. `https://www.grammarly.com/`. Accessed: 2023-4-24.

d. OpenAI codex. `https://openai.com/blog/openai-codex`. Accessed: 2023-5-4.

e. Stanford CRFM. `https://crfm.stanford.edu/2023/03/13/alpaca.html`. Accessed: 2023-5-4.

f. Stanford CRFM. `https://crfm.stanford.edu/2023/03/13/alpaca.html`. Accessed: 2023-5-4.

2023. AI guidance. `https://poorvucenter.yale.edu/AIguidance`. Accessed: 2023-4-24.

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949.

M Abouelenien, V Pérez-Rosas, and others. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions*.

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. Cebab: Estimating the causal effects of real-world concepts on NLP model behavior. In *NeurIPS*.

Artem Abzaliev, Andrew Owens, and Rada Mihalcea. 2022. Towards understanding the relation between gestures and language. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5507–5520.

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th. Papadopoulos, Vassia Zacharopoulou, George J. Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2022. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Trans. Multim.*, 24:1750–1762.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual question answering.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed,

Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

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

Martha W Alibali, Sotaro Kita, and Amanda J Young. 2000. Gesture and the process of speech production: We think, therefore we gesture. *Lang. Cogn. Process.*, 15(6):593–613.

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 224–233, Seattle, Washington. Association for Computational Linguistics.

Badr AlKhamissi, Siddharth Verma, Ping Yu, Zhijing Jin, Asli Celikyilmaz, and Mona Diab. 2023. OPT-R: Exploring the role of explanations in fine-tuning and prompting for reasoning skills of large language models.

Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-Symbolic language modeling with automaton-augmented retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 468–485. PMLR.

Saleema Amershi, Maya Cakmak, W. Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35:105–120.

Anonymous. 2024. The generation gap: Exploring age bias in large language models.

Stuart Armstrong. 2013. General purpose intelligence: Arguing the orthogonality thesis. *Analysis and Metaphysics*, 12:68–84.

Anmol Arora and Ananya Arora. 2022. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthcare Journal*, 9(2):190–193.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings.

Isaac Asimov. 1942. *Runaround*. Astounding Science Fiction.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021a. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021b. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

V Athitsos, C Neidle, S Sclaroff, J Nash, and others. 2008a. The american sign language lexicon video dataset. *2008 IEEE Computer*.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008b. The american sign language lexicon video dataset. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 0:1–8.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer Berlin Heidelberg.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones,

Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Available at SSRN 4337484*.

Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, Susana Carvalho, and Andrew Zisserman. 2021. Automated audiovisual behavior recognition in wild primates. *Sci Adv*, 7(46):eabi4883.

Tadas Baltruaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V Do, Yan Xu, and Pascale Fung. 2023b. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.

Libby Barak, Adele E. Goldberg, and Suzanne Stevenson. 2016. Comparing computational cognitive models of generalization in a language acquisition task. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 96–106, Austin, Texas. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M Bender, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2015. Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for programming, artificial intelligence, and reasoning*, pages 532–548. Springer.

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 *International Conference on Learning Representations*.

Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao, editors. 2022. *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2020. Quantifying the effects of COVID-19 on mental health support forums. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J Mielke, Wilson Y Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain,

Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access multilingual language model.

Maneesh Bilalpur, Saurabh Hinduja, Laura A Cariola, Lisa B Sheeber, Nick Alien, László A Jeni, Louis-Philippe Morency, and Jeffrey F Cohn. 2023. Multimodal feature selection for detecting mothers' depression in dyadic interactions with their adolescent offspring. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience grounds language.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

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.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators.

Nick Bostrom and Eliezer Yudkowsky. 2014. *The ethics of artificial intelligence*. Cambridge University Press.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of moral judgement on reddit.

George J Bowdery. 1941. Conventions and norms. *Philosophy of Science*, 8(4):493–505.

Melissa Bowerman and Stephen Levinson. 2001a. *Language Acquisition and Conceptual Development*. Language Culture and Cognition. Cambridge University Press.

Melissa Bowerman and Stephen C Levinson. 2001b. *Language Acquisition and Conceptual Development*. Cambridge University Press.

Samuel R Bowman. 2023. Eight things to know about large language models.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosiute, Amanda Askell, Andy Jones, Anna

Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models. *CoRR*, abs/2211.03540.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. Association for Computing Machinery.

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, and Mike Frank. 2016. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *Cognitive Science*.

Adrian M P Braşoveanu and Răzvan Andonie. 2020. Visualizing transformers for NLP: A brief survey. In *2020 24th International Conference Information Visualisation (IV)*, pages 270–279.

Nathan Brown, Peter Ertl, Richard A. Lewis, Torsten Luksch, Daniel Reker, and Nadine Schneider. 2020a. Artificial intelligence in chemistry and drug design. *Journal of Computer-Aided Molecular Design*, 34:709–715.

P Brown, J Cocke, S Della Pietra, V Della Pietra, F Jelinek, R Mercer, and P Roossin. 1988. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*. aclanthology.org.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Others. 2020b. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, 33:1877–1901.

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. 2020c. 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. 2023a. Sparks of artificial general intelligence: Early experiments with GPT-4.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023b. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Jill Burstein, Susanne Wolff, Chi Lu, and Randy M Kaplan. 1997. An automatic scoring system for advanced placement biology essays. In *Fifth Conference on Applied Natural Language Processing*, pages 174–181, Washington, DC, USA. Association for Computational Linguistics.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Sign language transformers: Joint end-to-end sign language recognition and translation. pages 10023–10033.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033. openaccess.thecvf.com.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186.

Dallas Card and Noah A. Smith. 2020. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3:34.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, and Others. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022. In-the-Wild video question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2021. Word acquisition in neural language models.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Alexander Cheves. 2017. 21 words the queer community has reclaimed (and some we haven't). *The Advocate*.

China AI Report. 2020. China AI report 2020.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021a. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021b. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. *Machine Learning for Healthcare Conference*, pages 301–318.

Morten H Christiansen and Nick Chater. 1999. Connectionist natural language processing: The state of the art. *Cognitive science*, 23(4):417–437.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *CoRR*, abs/2303.16779.

Jennifer Chubb, Sondess Missaoui, Shauna Concannon, Liam Maloney, and James Alfred Walker. 2021. Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. *CoRR*, abs/2302.03025.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1811–1824. Association for Computational Linguistics.

Shlomo Cohen. 2011. *The Proto-Ethical Dimension of Moods*, pages 173–184.

Keith Collins. 2023. How ChatGPT could embed a 'watermark' in the text it generates. *The New York Times*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *CoRR*, abs/2304.14997.

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Anna Costello, Ekaterina Fedorova, Zhijing Jin, and Rada Mihalcea. 2023. Editing a woman's voice. In *ICSSI*.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A learning environment for text-based games.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2023. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, pages 1–12.

J S Cramer. 2002. The origins of logistic regression.

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

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing.

Tri Dao, Beidi Chen, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Ré. 2021. Pixelated butterfly: Simple and efficient sparse training for neural network models.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):1–25.

Maartje de Graaf, Somaya Ben Allouch, and Jan van Dijk. 2017. Why do they refuse to use my robot? reasons for Non-Use derived from a Long-Term home study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 224–233, New York, NY, USA. Association for Computing Machinery.

Britni de la Cretaz. 2016. What it's like to chestfeed. *The Atlantic*.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. 2022. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1962–1981, New York, NY, USA. Association for Computing Machinery.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Yiwen Ding, Jiarui Liu, Zhiheng Lyu, Kun Zhang, Bernhard Schoelkopf, Zhijing Jin, and Rada Mihalcea. 2022. Voices of her: Analyzing gender differences in the AI publication world.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning pretrained language models: Weight initializations, data orders, and early stopping.

Serena Does, Belle Derks, and Naomi Ellemers. 2011. Thou shalt not discriminate: How emphasizing moral ideals rather than obligations increases whites' support for social equality. *Journal of Experimental Social Psychology*, 47(3):562–571.

Tjibbe Donker. 2023. The dangers of using large language models for peer review. *The Lancet Infectious Diseases*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani,

Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Saramsha Dotel, Avishekh Shrestha, Anish Bhusal, Ramesh Pathak, Aman Shakya, and Sanjeeb Prasad Panday. 2020. Disaster assessment from satellite imagery by analysing topographical features using deep learning. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, IVSP '20, pages 86–92, New York, NY, USA. Association for Computing Machinery.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. GLaM: Efficient scaling of language models with Mixture-of-Experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

Abhishek Dutta, Natalia Pérez-Campanero, Graham K Taylor, Andrew Zisserman, and Cait Newport. 2023. A robust and flexible deep-learning workflow for animal tracking.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022a. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at scale.

Tiwalayo Eisape, Mh Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. 2023. A systematic comparison of syllogistic reasoning in humans and language models. *CoRR*, abs/2311.00445.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *CoRR*, abs/2209.10652.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models.

Jon Elster. 2006. Fairness and norms. *Social Research: An International Quarterly*, 73:365–376.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Oren Etzioni. 2018. Point: Should AI technology be regulated? Yes, and here's how. *Commun. ACM*, 61(12):30–32.

European Commission. 2019. Ethics guidelines for trustworthy artificial intelligence.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: developing and governing AI that does not lie. *CoRR*, abs/2110.06674.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. 2022. Domino: Discovering systematic errors with Cross-Modal embeddings.

Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. 2021. Learning triadic belief dynamics in nonverbal communication from videos. pages 7312–7321.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. CLIP2Video: Mastering Video-Text retrieval via image CLIP.

Shaohua Fang and Zhiyi Wu. 2022. Syntactic prediction in l2 learners: evidence from english disjunction processing. *International Review of Applied Linguistics in Language Teaching*.

Fabio Fasoli, Peter Hegarty, and Andrea Carnaghi. 2019. Sounding gay, speaking as a "fag": Auditory gaydar and the perception of reclaimed homophobic language. *Journal of language and social psychology*, page 0261927X19852753.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Anjalie Field, Su Lin Blodgett, Zeerak Talat, and Yulia Tsvetkov. 2021a. A survey of race, racism, and anti-racism in nlp. In *Annual Meeting of the Association for Computational Linguistics*.

Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett, editors. 2021b. *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online.

Joseph F Fletcher. 1997. *Situation ethics: The new morality*. Westminster John Knox Press.

Alex Foote, Neel Nanda, Esben Kran, Ionnis Konstas, and Fazl Barez. 2023. N2G: A scalable approach for quantifying interpretable neuron representations in large language models. *CoRR*, abs/2304.12918.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.

Michael Frank, Mika Braginsky, Daniel Yurovsky, and Virginia Marchman. 2016. Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling. *Psychological science*, 24(10):2020–2029.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamile Lukosiute, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models. *CoRR*, abs/2302.07459.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom B. Brown, Jared Kaplan, Sam McCandlish, Chris Olah, and Jack Clark. 2022a. Predictability and surprise in large generative models. *CoRR*, abs/2202.07785.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022b. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differ-

ences in word usage. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 674–683.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586.

Atticus Geiger, Christopher Potts, and Thomas Icard. 2023a. Causal abstraction for faithful model interpretation. *CoRR*, abs/2301.04709.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2023b. Finding alignments between interpretable causal variables and distributed neural representations. *CoRR*, abs/2303.02536.

Bernard Gert and Joshua Gert. 2020. The Definition of Morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022b. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zoubin Ghahramani. 2001. An introduction to hidden markov models and bayesian networks. *Int. J. Pattern Recognit. Artif. Intell.*, 15:9–42.

Avijit Ghosh and Genoveva Fossas. 2022. Can there be art without an artist? *arXiv preprint arXiv:2209.07667*.

Edward Gibson, Richard Futrell, Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Philip Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23:389–407.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023a. ChatGPT outperforms Crowd-Workers for Text-Annotation tasks.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023b. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. 2021. Explainable AI: current status and future directions.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A

suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Fernando Gonzalez, Zhijing Jin, Jad Beydoun, Bernhard SchÃƒÂ¶lkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond good intentions: Reporting the research landscape of NLP for social good. *CoRR*, abs/2305.05471.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT Press.

Charles A. E. Goodhart. 1975. Problems of monetary management: The u.k. experience. *Papers in Monetary Economics*, 1.

Charles A. E. Goodhart. 1984. Problems of monetary management: The uk experience.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Corentin J Gosling and Bastien Trémolière. 2021. Reliability of moral decision-making: Evidence from the trolley dilemma. *Quarterly Journal of Experimental Psychology*, 74(6):981–990.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Significant Gravitas. 2023. Auto-gpt. https://github.com/Significant-Gravitas/Auto-GPT.

Jonathan Grudin. 2006. Why personas work: The psychological evidence. *The persona lifecycle*, 12:642–664.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A Knowledge-Enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Stefano Guarino, Noemi Trino, Alessandro Chessa, and Gianni Riotta. 2019. Beyond fact-checking: Network analysis tools for monitoring disinformation in social media. In *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 436–447. Springer.

Stefano Guarino, Noemi Trino, Alessandro Chessa, and Gianni Riotta. 2020. Beyond fact-checking: Network analysis tools for monitoring disinformation in social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 436–447. Springer.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Ning Dai, Xipeng Qiu, Xiangyang Xue, David Wipf, and Zheng Zhang. 2020a. P2: A plan-and-pretrain approach for knowledge graph-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 100–106, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020b. CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Ziyu Wang, Xipeng Qiu, Weinan Zhang, Jun Zhu, Zheng Zhang, and David Wipf. 2021. Fork or fail: Cycle-consistent training with many-to-one mappings. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1828–1836. PMLR.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *CoRR*, abs/2305.01610.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Association for Computational Linguistics (ACL)*, pages 107–112.

Jonathan Haidt. 2013. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Richard Mervyn Hare. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

Harvard Business Review. 2020. AI can make bank loans more fair.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018a. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018b. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Taylor Hatmaker. 2020. Twitter broadly bans any covid-19 tweets that could help the virus spread.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced Bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Melissa Heikkilä. 2023. How OpenAI is trying to make ChatGPT safer and less biased. *MIT Technology Review*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing Image-Language transformers for verb understanding.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021b. Unsolved problems in ML safety. *CoRR*, abs/2109.13916.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. 2022. Scaling laws and interpretability of learning from repeated data. *CoRR*, abs/2205.10487.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Measuring and manipulating knowledge representations in language models. *CoRR*, abs/2304.00740.

Daniel Hershcovich, Stella Frank, Heather C. Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6997–7013. Association for Computational Linguistics.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7817–7836. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

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. 2022a. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

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. 2022b. Training Compute-Optimal large language models.

Joseph Hoover, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon, and Morteza Dehghani. 2019. Bound in hatred: The role of group-based morality in acts of hate.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *2019 International conference on multimodal interaction*, pages 235–243.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Annual Meeting of the Association for Computational Linguistics*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 588–602. Association for Computational Linguistics.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *ArXiv*, abs/2304.01933.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *CoRR*, abs/2305.07004.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP/IJCNLP*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023b. Language is not all you need: Aligning perception with language models.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably).

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*, abs/2210.03050.

Oana Ignat, Laura Burdick, Jia Deng, and Rada Mihalcea. 2019. Identifying visible actions in lifestyle vlogs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6406–6417, Florence, Italy. Association for Computational Linguistics.

Oana Ignat, Santiago Castro, Hanwen Miao, Weiji Li, and Rada Mihalcea. 2021. Whyact: Identifying action reasons in lifestyle vlogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4770–4785.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzman. 2022. OCR improves machine translation for Low-Resource languages. *arXiv preprint arXiv*.

Geoffrey Irving, Paul F. Christiano, and Dario Amodei. 2018. AI safety via debate. *CoRR*, abs/1805.00899.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

F Jelinek. 1976. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64(4):532–556.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1–38.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021a. Scaling up visual and Vision-Language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021b. Scaling up visual and Vision-Language representation learning with noisy text supervision.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Can machines learn morality? The Delphi experiment. *CoRR*, abs/2110.07574.

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

Zhuoni Jie, Marwa Mahmoud, Quentin Stafford-Fraser, Peter Robinson, Eduardo Dias, and Lee Skrypchuk. 2018. Analysis of yawning behaviour in spontaneous expressions of drowsy drivers. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 571–576.

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using Twitter. In *Proceedings of BioNLP 15*, pages 164–170, Beijing, China. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022a. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020a. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020b. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*.

Zhijing Jin. 2021. Causality for NLP reading list. https://github.com/zhijing-jin/Causality4NLP_Papers.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021a. How good is NLP? A sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023a. Causal Benchmark: A benchmark of 10,000+ causal inference questions.

Zhijing Jin, Amir Feder, and Kun Zhang. 2022b. CausalNLP tutorial: An introduction to causality for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–22, Abu Dubai, UAE. Association for Computational Linguistics.

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020c. GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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 2019, Hong Kong, China, November 3-7, 2019, pages 3095–3107. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022c. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180Ã¢â‚¬â€œ–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022d. When to make exceptions: Exploring language models as accounts of human moral judgment. In *NeurIPS*.

Zhijing Jin, Sydney Levine, Max Kleiman-Weiner, Jiarui Liu, Francesco Ortu, Fernando Gonzalez Adauto, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2023b. Trolley problems for large language models across 100+ languages.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2023c. Can large language models infer causation from correlation?

Zhijing Jin and Rada Mihalcea. 2023. Natural language processing for policymaking. In Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli, and Michele Vespe, editors, *Handbook of Computational Social Science for Policy*, chapter 7, pages 141–162. Springer Cham.

Zhijing Jin, Zeyu Peng, Tejas Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. 2021b. Mining the cause of political decision-making from social media: A case study of COVID-19 policies across the US states. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 288–301, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021c. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhijing Jin, Yongyi Yang, Xipeng Qiu, and Zheng Zhang. 2020d. Relation of the relations: A new paradigm of the relation extraction problem. *CoRR*, abs/2006.03719.

Hamid Reza Vaezi Joze. 2019. MS-ASL: A Large-Scale data set and benchmark for understanding american sign language. *bmvc2019.org*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, 19:61 – 84.

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. 2023. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. *CoRR*, abs/2302.05733.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Ashkan Kazemi, Artem Abzaliev, Naihao Deng, Rui Hou, Davis Liang, Scott A Hale, Verónica Pérez-Rosas, and Rada Mihalcea. 2022. Adaptable

claim rewriting with offline reinforcement learning for effective misinformation discovery. *arXiv preprint arXiv:2210.07467*.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021a. Claim matching beyond English to scale global fact-checking. 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 4504–4517, Online. Association for Computational Linguistics.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021b. Extractive and abstractive explanations for fact-checking and evaluation of news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–50.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *CoRR*, abs/2103.14659.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Grace: Discriminator-guided chain-of-thought reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. 2016. A survey on the state-of-the-art machine learning models in the context of NLP. *KJS*, 43(4).

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. SODA: Million-scale dialogue distillation with social commonsense contextualization.

Bret Kinsella. 2022. YouChat is like ChatGPT with real-time search built in. https://synthedia.substack.com/p/youchat-is-like-chatgpt-with-real. Accessed: 2023-4-1.

Bret Kinsella. 2023. Perplexity offers a new conversational search experience challenging google and ChatGPT. https://synthedia.substack.com/p/perplexity-offers-a-new-conversational. Accessed: 2023-3-31.

Lawrence Kohlberg. 1976. Moral stages and moralization. *Moral development and behavior*, page 31– 53.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916.

Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3793–3802.

Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *International Journal of Computer Vision*, 126:1311–1325.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin,

Ireland. 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.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Adv. Neural Inf. Process. Syst.*, 23.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. *Communication theory*, 15(2):127–147.

Anne Lauscher, Federico Bianchi, Samuel Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics.

Angeliki Lazaridou, Grzegorz Chrupała, Raquel Fernández, and Marco Baroni. 2016. Multimodal semantic learning from child-directed input. In *Knight K, Nenkova A, Rambow O, editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12-17; San Diegio, California. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 387–92.* repositori.upf.edu.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Social science. computational social science. *Science*, 323(5915):721–723.

David M J Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Andrew Lee, Jonathan K Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4257–4272.

Andrew Lee, David Wu, Emily Dinan, and Mike Lewis. 2022a. Improving chess commentaries by combining language models with symbolic reasoning engines.

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL*, 23:25 – 58.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, compositional video question answering.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019a. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019b. TVQA+: Spatio-Temporal grounding for video question answering.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Richard Leiter, Enrico Santus, Zhijing Jin, Katherine Lee, Miryam Yusufov, Edward Moseley, Yujie Qian, Jiang Guo, and Charlotta Lindvall. 2020a. An Artificial Intelligence algorithm to identify documented symptoms in patients with heart failure who received Cardiac Resynchronization Therapy (s717). *Journal of Pain and Symptom Management*, 59(2):537–538.

Richard E Leiter, Enrico Santus, Zhijing Jin, Katherine C Lee, Miryam Yusufov, Isabel Chien, Ashwin Ramaswamy, Edward T Moseley, Yujie Qian, Deborah Schrag, et al. 2020b. Deep Natural Language Processing to identify symptom documentation in clinical notes for patients with heart failure undergoing Cardiac Resynchronization Therapy. *Journal of Pain and Symptom Management*, 60(5):948–958.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for Parameter-Efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Belinda Li, Jane Yu, Madian Khabsa, Luke Zettlemoyer, Alon Halevy, and Jacob Andreas. 2022a. Quantifying adaptability in pre-trained language models with 500 tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 4696–4715, Seattle, United States. Association for Computational Linguistics.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469. openaccess.thecvf.com.

Han Li, Robert E Kraut, and Haiyi Zhu. 2021. Technical features of asynchronous and synchronous community platforms and their effects on community cohesion: a comparative study of forum-based and chat-based online mental health communities. *Journal of Computer-Mediated Communication*, 26(6):403–421.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022b. Emergent world representations: Exploring a sequence model trained on a synthetic task.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-Semantics aligned pre-training for Vision-Language tasks.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Ré, Diana Acosta-Navas, Drew A Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard,

Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning health-bots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 638–648, Barcelona, Spain (Online). International Committee on Computational Linguistics.

JongYoon Lim, Inkyu Sa, Bruce MacDonald, and Ho Seok Ahn. 2023. A sign language recognition system with pepper, lightweight-transformer, and llm.

Chien-Chang Lin, Anna YQ Huang, and Owen HT Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *CoRR*, abs/2205.14334.

Yehuda Lindell. 2021. Secure multiparty computation. *Commun. ACM*, 64(1):86–96.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Angli Liu, Jingfei Du, and Veselin Stoyanov. 2019a. Knowledge-augmented language model and its application to unsupervised named-entity recognition. *CoRR*, abs/1904.04458.

Siyang Liu, Naihao Deng, Sahand Sabour, Yilin Jia, Minlie Huang, and Rada Mihalcea. 2023. Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15264–15281, Singapore. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483,

Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pre-training approach. *ArXiv*, abs/1907.11692.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021a. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *AAAI*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021b. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *AAAI*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic visiolinguistic representations for Vision-and-Language tasks.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *ArXiv*, abs/1705.07874.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023a. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *ArXiv*, abs/2306.09093.

Zhiheng Lyu, Zhijing Jin, Justus Mattern, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2023b. Psychologically-inspired causal prompts. *CoRR*, abs/2305.01764.

Zhiheng LYU, Zhijing Jin, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Can large language models distinguish cause from effect? In *UAI 2022 Workshop on Causal Representation Learning*.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: Moving average equipped gated attention.

Brian MacWhinney. 1992. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8(2):217–218.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *CoRR*, abs/2210.07128.

Alexander Maedche and Steffen Staab. 2000. Mining ontologies from text. In *International Conference Knowledge Engineering and Knowledge Management*.

João Carlos Magalhães and Christian Katzenbach. 2020. Coronavirus and the frailness of platform governance. *Internet Policy Review*.

Nikolay Malkin, Sameera Lanka, Pranav Goel, Sudha Rao, and Nebojsa Jojic. 2021. GPT perdetry test: Generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. 2014. A theory of blame. *Psychological Inquiry*, 25(2):147–186.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world.

Mason Marks and Claudia E Haupt. 2023. Ai chatbots, health privacy, and challenges to hipaa compliance. *JAMA*.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Philip Levy. 2023. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. 2022. Causal abstraction with soft interventions. *CoRR*, abs/2211.12270.

Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. In *Intelligent Computing*, pages 1269–1292. Springer International Publishing.

David Matsumoto and Manish Assar. 1992. The effects of language on judgments of universal facial expressions of emotion. *Journal of Nonverbal Behavior*, 16:85–99.

Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022a. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *CoRR*, abs/2212.10678.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022b. Differentially private language models for secure data sharing. *ArXiv*, abs/2210.13918.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022c. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schölkopf, and Mrinmaya Sachan. 2022d. Differentially private language models for secure data sharing.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Nikolaos Mavridis. 2014. A review of verbal and non-verbal human-robot interactive communication. *Robotics Auton. Syst.*, 63:22–35.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Michael McCloskey. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6):387–395.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep

networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

David McNeill. 1970. The acquisition of language: The study of developmental psycholinguistics.

David McNeill. 1992. Hand and mind: What gestures reveal about thought. 416.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022b. Locating and editing factual associations in GPT. In *NeurIPS*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.

Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *npj Digital Medicine*, 6(1):120.

Cade Metz. 2016. Self-driving cars will teach themselves to save lives—but also take them | wired. https://www.wired.com/2016/06/self-driving-cars-will-power-kill-wont-conscience/.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023a. Augmented language models: A survey. *CoRR*, abs/2302.07842.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023b. Augmented language models: a survey.

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.

Do June Min, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra and Bhavdeep Singh Sachdeva. 2020. Do we need to create big datasets to learn a task? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 169–173, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP prefix for image captioning.

Ishani Mondal, Kabir Ahuja, Mohit Jain, Jacki O'Neill, Kalika Bali, and Monojit Choudhury. 2022. Global readiness of language technology for healthcare: What would it take to combat the next pandemic? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4320–4335, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

James Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21:18–21.

Nafise Moosavi, Quentin Delfosse, Kristian Kersting, and Iryna Gurevych. 2022. Adaptable adapters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3742–3753, Seattle, United States. Association for Computational Linguistics.

Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, and Georg Groh. 2023. IFAN: An Explainability-Focused interaction framework for humans and NLP models.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023a. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023b. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Jesse Mu, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and Edward Grefenstette. 2022. Improving intrinsic exploration with language abstractions. *arXiv preprint arXiv:2202.08938*.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Basil Mustafa, Carlos Riquelme Ruiz, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with LIMoE: the Language-Image mixture of experts.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *CoRR*, abs/2301.05217.

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G. Dunn. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model.

New York Times. 2021. Résumé-writing tips to help you get past the a.i. gatekeepers.

Ansong Ni, Jeevana Priya Inala, Chenglong Wang, Alex Polozov, Christopher Meek, Dragomir Radev, and Jianfeng Gao. 2023a. Learning math reasoning from self-sampled correct and partially-correct solutions. In *The Eleventh International Conference on Learning Representations*.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? A causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Qian Wang, Mrinmaya Sachan, and Markus Leippold. 2023b. When does aggregating multiple skills with multi-task learning work? A case study in financial NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp

Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling Human-Centered machine translation.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *CoRR*, abs/2303.13375.

Oded Nov, Nina Singh, and Devin Mann. 2023. Putting chatgpt's medical advice to the (turing) test: Survey study. *JMIR Med Educ*, 9:e46939.

NPR. 2020. Researchers warn against 'autonomous weapons' arms race.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114.

Nir Ofek, Soujanya Poria, Lior Rokach, Erik Cambria, Amir Hussain, and Asaf Shabtai. 2016. Unsupervised commonsense knowledge enrichment for domain-specific sentiment analysis. *Cognitive Computation*, 8:467–477.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Iroju Olaronke and J. Olaleke. 2015. A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, 08:44–50.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype!

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023. GPT-4 technical report.

Eyal Orbach and Yoav Goldberg. 2020. Facts2Story: Controlling text generation by key facts.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. VALSE: A Task-Independent benchmark for vision and language models centered on linguistic phenomena.

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.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink.

John V Pavlik. 2023. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1):84–93.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the*

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016a. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Gonçalo Pereira, Rui Prada, and Pedro A. Santos. 2016b. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Luís Moniz Pereira and Ari Saptawijaya. 2007. Modelling morality with prospective logic. In *Portuguese Conference on Artificial Intelligence*, pages 99–111. Springer.

Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022a. Red teaming language models with language models. *CoRR*, abs/2202.03286.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022b. Discovering language model behaviors with model-written evaluations. *CoRR*, abs/2212.09251.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

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.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case study: Deontological ethics in NLP.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. GraphIE: A graph-based framework for information extraction. 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 751–761, Minneapolis, Minnesota. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Lawrence R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286.

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

Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nazneen Rajani, Weixin Liang, Lingjiao Chen, Meg Mitchell, and James Zou. 2022. SEAL : Interactive tool for systematic error analysis and labeling.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image generation.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Danushka Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. 2016. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christine Redecker and Yves Punie. 2017. european framework for the digital competence of educators: digcomp-edu. *Luxembourg, Luxembourg: Publications Office of the European Union. https://doi. org/10.2760/159770*.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Reid. 1788. *Essays on the active powers of man*. Edinburgh University Press.

Reuters. 2018. Amazon scraps secret AI recruiting tool that showed bias against women.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, pages 314:1–314:7. ACM.

Kais Riani, Michalis Papakostas, Hussein Kokash, M Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Towards detecting levels of alertness in drivers using multiple modalities. *Petra*.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR post correction for endangered language texts.

Alex Roehrs, Cristiano André da Costa, Rodrigo da Rosa Righi, Sandro José Rigo, and Matheus Henrique Wichman. 2018. Toward a model for personal health record interoperability. *IEEE journal of biomedical and health informatics*, 23(2):867–873.

A Rogers, O Kovaleva, and A Rumshisky. A primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput. Linguist.*

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Marco Damonte, and Isabel Groves. 2022a. CLASP: Few-Shot Cross-Lingual data augmentation for semantic parsing.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52 11:2864–75.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4661–4675.

Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1).

Jackson Ryan. 2022. Meta trained an AI on 48M science papers. it was shut down after 2 days. https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-paper Accessed: 2023-3-31.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *EMNLP 2019*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also Few-Shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-music generation with long-context latent diffusion. *CoRR*, abs/2301.11757.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2020. The moral choice machine. *Frontiers in artificial intelligence*, 3:36.

Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via Gradient-Based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

C E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021a. How much can CLIP benefit Vision-and-Language tasks?

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. ChatGPT and other large language models are double-edged swords. *Radiology*, page 230163.

Yiting Shen, Steven R Wilson, and Rada Mihalcea. 2019. Measuring personal values in cross-cultural user-generated content. In *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings 11*, pages 143–156. Springer.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021b. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624.

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.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Richard A Shweder. 1990. In defense of moral realism: Reply to gabennesch. *Child Development*, 61(6):2060–2067.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video generation without Text-Video data.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Yale Song, Louis-Philippe Morency, and Randall Davis. 2013. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, ICMI '13, pages 237–244, New York, NY, USA. Association for Computing Machinery.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2021. Analyzing stereotypes in generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. *AAAI*, 31(1).

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher

Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13693–13696. AAAI Press.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. Association for Computing Machinery.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Online. Association for Computational Linguistics.

Tom Tabak and Matthew Purver. 2020. Temporal mental health dynamics on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *North American Chapter of the Association for Computational Linguistics*.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Suzanne Tamang, Zhijing Jin, and Vyjeyanthi S. Periyakoil. 2023. Revelations from a machine learning analysis of the most downloaded articles published in journal of palliative medicine 1999Ã¢â¬â€œ2018. *Journal of Palliative Medicine*, 26(1):13–16. PMID: 36607778.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023a. Does synthetic data generation of LLMs help clinical text mining?

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023b. Does synthetic data generation of LLMs help clinical text mining?

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

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):1–28.

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.

Max Tegmark. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf Publishing Group.

The Washington Post. 2020. The u.s. says humans will always be in control of AI weapons. but the age of autonomous war is already here.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning.

Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for Visio-Linguistic compositionality.

George Tolkachev, Stephen Mell, Steve Zdancewic, and Osbert Bastani. 2022. Counterfactual explanations for natural language interfaces.

Michael Tomasello and Amrisha Vaish. 2013. Origins of human cooperation and morality. *Annual review of psychology*, 64:231–255.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Marcos Treviso, António Góis, Patrick Fernandes, Erick Fonseca, and Andre Martins. 2022a. Predicting attention sparsity in transformers. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 67–81, Dublin, Ireland. Association for Computational Linguistics.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H Martins, André F T Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2022b. Efficient methods for natural language processing: A survey.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Mark Ungar. 2000. State violence and lesbian, gay, bisexual and transgender (lgbt) rights. *New Political Science*, 22(1):61–75.

United Nations. 2021. Universal declaration of human rights.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2022. DyLoRA: Parameter efficient tuning of pre-trained models using dynamic Search-Free Low-Rank adaptation.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. ¡§kelly is a warm person, joseph is a role model¡¨: Gender biases in llm-generated reference letters. In *EMNLP-Findings*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. 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.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2020. REVISE: A tool for measuring and mitigating bias in visual datasets.

Hongru Wang, Zhijing Jin, Jiarun Cao, Gabriel Pui Cheong Fung, and Kam-Fai Wong. 2021. Inconsistent few-shot relation classification via cross-attentional prototype networks with contrastive learning. *CoRR*, abs/2110.08254.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Weirong Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xingxu Xie. 2023a. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *ArXiv*, abs/2302.12095.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Weiyao Wang, Du Tran, and Matt Feiszli. 2019b. What makes training Multi-Modal classification networks hard?

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022a. Image as a foreign language: BEiT pretraining for all vision and Vision-Language tasks.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019c. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proc. Conf. AAAI Artif. Intell.*, 33(1):7216–7223.

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 Gary 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. 2022b. Supernaturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023a. Findings of the BabyLM challenge: Sampleefficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023b. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Alicia Nicki Washington. 2020. When twice as good isn't enough: The case for cultural competence in computing. In *SIGCSE*, SIGCSE '20, pages 213–219, New York, NY, USA. Association for Computing Machinery.

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. 2022a. Emergent abilities of large language models. *CoRR*, abs/2206.07682.

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. 2022b. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022d. Chainof-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Daniel Weld and Oren Etzioni. 1994. The first law of robotics (a call to arms). In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, page 1042–1047. AAAI Press.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1371, Seattle, Washington, USA. Association for Computational Linguistics.

White House. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88.

Maurice V Wilkes. 1994. *Using Large Corpora*. MIT Press.

David Wilmot and Frank Keller. 2021. Memory and knowledge augmented language models for inferring salience in long-form stories. *CoRR*, abs/2109.03754.

David B Wong. 2009. *Natural moralities: A defense of pluralistic relativism*. Oxford University Press.

W A Woods. 1973. Progress in natural language understanding: an application to lunar geology. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, AFIPS '73, pages 441–450, New York, NY, USA. Association for Computing Machinery.

Scott Wright. 2006. Government-run online discussion fora: Moderation, censorship and the shadow of control1. *The British Journal of Politics and International Relations*, 8(4):550–568.

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

Winston Wu and David Yarowsky. 2018. Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Xiaodong Wu, Ran Duan, and Jianbing Ni. 2023b. Unveiling security, privacy, and ethical concerns of chatgpt. *ArXiv*, abs/2307.14192.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2022a. Causal proxy models for concept-based model explanations. *CoRR*, abs/2209.14279.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. 2022b. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295, Seattle, United States. Association for Computational Linguistics.

M. Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *International Workshop on Natural Language Processing for Social Media*.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2734–2744. Association for Computational Linguistics.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605, Online. Association for Computational Linguistics.

Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, Xin Li, Weijia Xu, Vetle I. Torvik, Yi Bu, Chongyan Chen, Islam Akef Ebeid, Daifeng Li, and Ying Ding. 2020. Building a pubmed knowledge graph. *Scientific Data*, 7.

Jitao Xu and Franccois Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *CALCS*.

Zhichao Xu, Hemank Lamba, Qingyao Ai, Joel Tetreault, and Alex Jaimes. 2023. Counterfactual editing for search result explanation.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.

Charles. Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press.

Yiqun Yao and Rada Mihalcea. 2022. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834.

Andrew Yates, Jon Parker, Nazli Goharian, and Ophir Frieder. 2014. A framework for public health surveillance. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv*, abs/2306.13549.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2013. Nature language reasoning, a survey.

Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissi, Siddharth Verma, Zhijing Jin, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2023. ALERT: adapting language models to reasoning tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.*, 54(11s):1–38.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2020. Turingadvice: A generative and dynamic evaluation of language use. *arXiv preprint arXiv:2004.03607*.

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. *Adv. Neural Inf. Process. Syst.*, 34:23634–23651.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022a. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023b. A survey of controllable text generation using transformer-based pre-trained language models.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW

'20, pages 201–211, New York, NY, USA. Association for Computing Machinery.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023c. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: open pre-trained transformer language models. CoRR, abs/2205.01068.

Xuan Zhang and Kevin Duh. 2020. Reproducible and efficient benchmarks for hyperparameter optimization of neural machine translation systems. Transactions of the Association for Computational Linguistics, 8:393–408.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. DialoGPT: Large-Scale generative pre-training for conversational response generation.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022c. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021a. Ethical-advice taker: Do language models understand natural language interventions? In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4158–4164, Online. Association for Computational Linguistics.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021b. Ethical-advice taker: Do language models understand natural language interventions?

Mingjun Zhao, Haijiang Wu, Di Niu, and Xiaoli Wang. 2020. Reinforced curriculum learning on Pre-Trained neural machine translation models. AAAI, 34(05):9652–9659.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for Amis-Mandarin translation. In Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pages 79–84, Taipei, Taiwan. Association for Computational Linguistics.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2022a. Natural language processing for smart healthcare. IEEE Reviews in Biomedical Engineering, pages 1–17.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022b. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625.

Karen Zhou, Ana Smith, and Lillian Lee. 2021. Assessing cognitive linguistic influences in the assignment of blame. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 61–69, Online. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2019. Unified Vision-Language Pre-Training for image captioning and VQA.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, number Article 930 in AAAI'18/IAAI'18/EAAI'18, pages 7590–7598. AAAI Press.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. CoRR, abs/1909.08593.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

A. Appendix

Our Area	Standard ACL 2018 Tracks
Sec 2.1 Multilinguality	Multilinguality; Machine Translation
Sec 2.2 Reasoning	Question Answering; Textual Inference
Sec 2.3 Knowledge Bases	Information Extraction; Document Analysis
Sec 2.4 Language Grounding	Vision, Robotics, Multimodal, Grounding and Speech
Sec 2.5 Child Language Acquisition	Linguistic Theories, Cognitive Modeling and Psycholinguistics
Sec 2.6 Non-Verbal Communication	-
Sec 3.1 NLP and Ethics	-
Sec 3.2 Interpretability	Machine Learning; Resources and Evaluation
Sec 3.3 Green/Efficient NLP	-
Sec 3.4 NLP for Online Environments	Social Media; Dialogue and Interactive Systems; Sentiment Analysis and Argument Mining
Sec 4.1 NLP for Healthcare	Multidisciplinary; Dialogue and Interactive Systems; Information Extraction; Generation
Sec 4.2 NLP for Education	Multidisciplinary; Dialogue and Interactive Systems; Information Extraction; Generation
Sec 4.3 Computational Social Science	Multidisciplinary; Multilinguality; Dialogue and Interactive Systems; Sentiment Analysis and Argument Mining
Sec 4.4 Synthetic Datasets	Resources and Evaluation; Generation

Table 1: The mapping of our research areas and the ACL 2018 tracks. We address 15/ 21 areas from ACL 2018 list of research areas. The unaddressed domains, are within the purview of LLMs or outside our area of expertise (tagging and parsing): Discourse and Pragmatics; Phonology, Morphology and Word Segmentation; Sentence-level Semantics; Summarization; Tagging, Chunking, Syntax and Parsing; and Word-level Semantics.