A PhD Student's Perspective on Research in NLP in the Era of Very 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
LANGUAGE AND INFORMATION TECHNOLOGIES (LIT)
UNIVERSITY OF MICHIGAN
{oignat,jinzhi,mihalcea}@umich.edu

Abstract

Recent progress in large language models 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 career – wonder about what NLP research area they should focus on. This document is a compilation of NLP research directions that are rich for exploration, reflecting the views of a diverse group of PhD students in an academic research lab. While we identify many research areas, many others exist; we do not cover those areas that are 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

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 to the early research on statistical machine translation (Brown et al., 1988; Wilkes, 1994) and statistical speech processing (Jelinek, 1976). While these models have always been used as an integral part of broad application categories such as text classification, information retrieval, or text generation, only in recent years they found a "life of their own" with widespread use and deployment.

The impressive advancements we have witnessed in current "large" and "very large" language models 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 large language models (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 Graphics Processing Units (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 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. For instance, a search on Google Scholar for "language models" leads to 50,000 publications over the past five years, a third of the roughly 150,000 papers published during the past 25 years.

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

*Oana Ignat and Zhijing Jin contributed equally to the manuscript. Rada Mihalcea initiated and guided the work, and provided overall supervision. All other authors are listed in alphabetical order.

*https://apply.knowinnovation.com/darpaaiforward/

applications in NLP is much more vast 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 research in this field, 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 document is a compilation of ideas from PhD students, building upon their initial expertise and existing interests and brainstorming around 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 PhD, all the way to students who recently completed a PhD degree. The LIT students come from a wide variety of backgrounds, including China, Iran, Japan, Mexico, Nigeria, Romania, Russia, South Korea, 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.

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 that one could write a PhD 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. Finally, third, we targeted research directions that can find solutions with reasonable computational costs achievable with setups more typically available in academic labs.

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 behind 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. The following sections provide brief descriptions of fourteen research areas rich in exploration, each of them with 2–4 research directions. These areas could be broadly divided into areas that cannot be addressed by LLMs for being too data-hungry or for lacking reasoning or grounding abilities (Sections 2–6, 8, 12); areas for which we cannot use LLMs because of not having the right data (Sections 9, 13, 14); or areas that could contribute to improving the abilities and quality of LLMs (Sections 7, 10, 11, 15).

2 Multilinguality and Low-Resource Languages

Background. Multilingual models are designed to handle multiple languages, whether for the task of machine translation (MT) or other tasks. A major challenge is handling low-resource languages, for which there is limited availability of training data, which can result in poor translation quality and poor performance on these languages. The research community has proposed several techniques to overcome this challenge, such as data augmentation, including synthetic data generation through back-translation (Sennrich et al., 2015; Edunov et al., 2018), parallel-corpus mining (Artetxe and Schwenk, 2018), or OCR (Rijhwani et al., 2020; Ignat et al., 2022); and multilingual models, which are pre-trained models that can handle multiple languages and can be fine-tuned on low-resource languages to improve translation quality. Recent efforts to develop multilingual models for low-resource languages include NLLB-200 (NLLB Team et al., 2022), a state-of-the-art Mixture of Experts (MoE) model trained on a dataset containing more than 18 billion sentence pairs. The same team also created and open-sourced an expanded benchmark dataset, FLORES-200 (Goyal et al., 2021), for evaluating MT models in 200 languages and over 40k translation directions.

Gaps. State-of-the-art MT models such as NLLB-200 (NLLB Team et al., 2022) still perform poorly on many low-resource languages, such as African languages. For instance, recent work tested the ChatGPT MT performance on low-resource languages (e.g., Marathi, Sundanese, and Buginese) and found an overall poor performance, especially in non-Latin scripts (Bang et al., 2023b). They also found that ChatGPT can perform reasonably well on low-resource language to-English translation, but it cannot perform English to low-resource language translation. In addition, MT systems do not exist for the vast majority of the world's roughly 7,000 languages.

Research Directions.

1. **Improve MT performance on current low- and very low-resource language benchmarks.** There is still plenty of room for improving the state-of-the-art models on current benchmarks such as FLORES-200. This benchmark has spurred recent interest in creating other benchmarks for low-resource languages, such as African languages (Vegi et al., 2022; Reid et al., 2021). Extremely low-resource languages do not have a significant web presence and thus do not have adequate bitext for training MT systems. These languages may have a translation of the Bible (the most translated document in the world), which can serve as a starting point for developing MT systems (McCarthy et al., 2020; Mueller et al., 2020). There is also recent interest in manually creating parallel corpora, such as for the languages Amis (Zheng et al., 2022) and Minankabau (Koto and Koto, 2020), but this process is expensive and time-consuming. In the absence of bilingual and even monolingual training corpora, one line of research is developing and expanding translation dictionaries using models of word formation (Wu and Yarowsky, 2018, 2020).

2. **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 when it comes to tasks such as prediction, classification, or generation. Some models are trained in part on web text like Common Crawl (Smith et al., 2013), which contains predominantly English text. Open questions include how much data, and what combination of languages, are necessary to enable similar performance on multiple languages. Additionally, cross-lingual projection continues to be a potential source of data for models in other languages, by leveraging the data available in major languages along with existing MT systems, to transfer model architectures onto other languages.

3. **Code-switching.** Code-switching (CS) is a phenomenon in which a speaker alternates between languages while adhering to the grammatical structure of at least one language. CS data presents a unique set of challenges for NLP tasks. The nature of CS leads speakers to create "new" words, meaning that models designed to accommodate CS data must be robust to out-of-vocabulary tokens (Çetinoğlu et al., 2016). Training data is difficult to obtain, also making it difficult to learn CS-specific models. An active area of research is in determining to what extent LLMs can generate synthetic CS data; previous methods commonly use parallel corpora to substitute tokens with grammatical rules as constraints (Xu and Yvon, 2021; Lee and Li, 2020). Additional areas of research include exploring to what extent models can be generalizable

across different language combinations, and learning models that can effectively distinguish between highly similar languages, such as dialects of the same parent language (Aguilar et al., 2020). Recently, benchmarks such as LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020) have been established to evaluate performance on classic NLP tasks on a number of common languages, but these benchmarks are not all-encompassing in regards to tasks and language combinations.

3 Reasoning

Background. Reasoning is a fundamental aspect of human intelligence, playing a critical role in problem-solving or decision-making by drawing inferences from premises, facts, and knowledge using logical principles and cognitive processes. There are various reasoning types, including deductive, inductive, abductive, quantitative, causal, and moral reasoning. Improving reasoning skills in NLP is vital for tasks such as question answering, reading comprehension, and dialogue systems, as it can enhance a model's generalization ability in unseen scenarios. NLP research has evolved significantly, from early rule-based and symbolic approaches to statistical methods in the 1990s, which utilized probabilistic models and machine learning algorithms. In recent years, deep learning and neural networks have revolutionized the field, achieving state-of-the-art performance on various tasks. However, challenges remain in attaining human-like reasoning and generalization abilities, driving continued research for more sophisticated and robust NLP models.

Gaps. Although LLMs have shown impressive performance on many reasoning benchmarks (Brown et al., 2020b; Ouyang et al., 2022; Zhang et al., 2022; Touvron et al., 2023a; OpenAI, 2023), there are still several directions that remain challenging. They struggle to robustly manage formal reasoning (Jin et al., 2022b; Stolfo et al., 2023; Jin et al., 2023a), as we often see LLMs prone to errors that a formal or symbolic system would not make. Additionally, since most of their training interacts with a world of text, NLP models still lack grounding in real-world experiences when reasoning (Ignat et al., 2021). Lastly, more fundamental questions remain to be answered, such as distinguishing empirical knowledge and rational reasoning, and unveiling how LLMs reason.

Research Directions.

1. **Robust formal reasoning.** Formal reasoning has long been a challenging task for neural networks. LLMs are far from complete mastery of formal tasks such as numerical reasoning (Stolfo et al., 2023; Miao et al., 2020), logical reasoning (Jin et al., 2022b), and causal inference (Jin et al., 2023a,c), often making obvious mistakes (Goel et al., 2021; Jin et al., 2020). To this end, a robust model should know how to generalize. To robustly manage formal reasoning, one could explore a variety of directions, such as combining the strengths of neural networks and symbolic AI. A popular line of work has been to integrate external reasoning systems, such as calculators, python interpreters, knowledge retrieval from databases, or search engines (Schick et al., 2023; Mialon et al., 2023).

2. **Grounded reasoning in the physical real world.** While current models generate coherent and contextually relevant responses, they often lack an understanding of the physical world and its constraints. This can lead to linguistically plausible responses that are nonsensical or unrealistic in practice. To address this issue, one direction is to explore ways to incorporate external knowledge sources, multimodal data, or simulated world scenarios to ground the reasoning skills of the models.

3. **Responsible reasoning in social contexts.** With increasing numbers 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 towards certain subpopulations, 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., 2021). We foresee that interdisciplinary collaboration with domain experts and policymakers will be needed.

4. **Formally defining reasoning and designing proper evaluation framework.** There is a rising need to refine the definition of reasoning, because LLMs start to make the difference between knowl-

edge and reasoning blurry – when a model memorizes a reasoning pattern, does it count as the mastery of reasoning or knowledge? Models already start to show an increasing mastery of templated solutions by pattern matching, which seems to be the reasoning that many want. Fundamentally, it leads to a question about what are the sparkles of intelligence that humans excel at, and how different are these from empirically learning how to do template matching. Beyond redefining reasoning, then the other open question is how to test models' reasoning skills. We face problems such as data contamination, Goodhart's law (a dataset failing to reflect the skill once it is exploited), and a lack of reliable metrics to evaluate multi-step reasoning.

5. **Analyzing how prompts help reasoning.** There are two types of prompting whose effect on LLMs are worth inspection: in-context learning and chain of thought. Recent work shows that conditioning on in-context examples has a similar effect to finetuning the model (Akyürek et al., 2022), and researchers start to decode the mechanisms that models start to pick up from the given context, such as induction heads (Olsson et al., 2022). Apart from the in-context instructions, we can also prompt LLMs with intermediate steps using chain-of-thought prompting. This approach breaks down reasoning tasks into smaller sub-problems, similar to human problem-solving. However, it is debatable whether language models truly reason or just generate statistically-alike sequences, and to what extent AI systems can learn to reason from few-shot exemplars.

4 Knowledge Bases

Background. A knowledge base is a collection of facts about real-world objects, abstract concepts, or events. The knowledge inside a knowledge base is usually represented as a triplet consisting of a head entity, a tail entity, and their relationships. For instance `(Barack Obama, birthPlace, Honolulu)` is an example of a triplet indicating a place-of-birth relationship. Some knowledge bases focus more on factual knowledge, such as DBPedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007), while others focus more on commonsense, such as ConceptNet (Speer et al., 2017) and ASER (Zhang et al., 2020).

Knowledge bases have found use in many downstream applications, including relation extraction (Weston et al., 2013), machine reading (Yang and Mitchell, 2017), and reflection generation in counseling dialogues (Shen et al., 2022). Many have found that integrating external knowledge improves performance on such knowledge-intensive tasks (Yu et al., 2022). Moreover, knowledge bases are often structured in a well-defined ontology of relationships and entities, allowing humans to more easily interpret the inferences grounded on knowledge bases.

Gaps. Although LLMs are trained on extensive datasets and demonstrate the capacity to tackle a wide variety of tasks (Brown et al., 2020a; Bubeck et al., 2023a), their internal knowledge remains limited in many respects, both with respect to general knowledge, as well as domain-specific (Ofek et al., 2016) or culture-specific knowledge (Yin et al., 2022). Additionally, LLMs frequently hallucinate, generating claims based on false facts. Although reinforcement learning from human feedback (RLHF) can mitigate this phenomenon, the problem of hallucination is inherent to the model. Grounding the model's output on an explicit knowledge base would likely reduce hallucination and enable users to verify the correctness of an assertion more easily. It also opens up the possibility for performing logical reasoning with the large body of existing works.

Research Directions.

1. **Knowledge-guided LLM.** The integration of knowledge into LLMs is a promising research direction for solving the hallucination problem by grounding the model's response on a verified resource of knowledge. ChatGPT seeks to address this through plugins, which indicates that the problem is not going to be solved by the LLM itself, but depends on the individual use case. There have been attempts to retrieve or generate knowledge for enhanced response generation with systems like DialogGPT (Zhang et al., 2019). Search engines such as Bing also conduct a web query for factual questions before composing a response. However, how LLMs should most efficiently and effectively interact with customized external knowledge bases remains an open problem.

2. **Automatic knowledge base construction.** Many applications can benefit from specialized knowledge bases, whether for improved human interpretability or to serve as a stand-alone resource. The automatic construction of such knowledge

bases is an interesting direction and requires many challenges to be addressed, such as knowledge coverage, factuality of the knowledge, knowledge linking, and so on. These challenges are amplified when the knowledge bases are constructed for specialized domains such as healthcare or chemistry. However, once these problems are addressed, researchers will be able to utilize LLMs to dynamically curate a knowledge base from up-to-date raw text and an ontology for complex applications such as tracking medication interactions from articles from PubMed.

3. **General and Cultural Commonsense.** Cultural knowledge available in NLP models is often limited to a handful of Western cultures and does not account for the vast diversity of the cultural views of the world (Arora et al., 2023). With the increasingly wide spread of NLP applications, this limitation may result in direct adverse impact on the users of these applications, by not accounting for their values, believes, and world views. More work is needed to understand the limitations of NLP models, including LLMs, with respect to their knowledge of different cultural groups. Further, 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.

5 Language Grounding

Background. Language grounding is the ability to tie verbal expressions to their referents in the non-linguistic world (Patel and Pavlick, 2022). The non-linguistic world can be physical or non-physical, e.g., TextWorld (Côté et al., 2018). Significant research advancements are due to leveraging sensory data to build datasets and tasks for teaching ML models how to perform language grounding. Popular tasks include visual question answering (Agrawal et al., 2015; Singh et al., 2019), image and video captioning (Mokady et al., 2021; Zhou et al., 2019), text to image retrieval (Wang et al., 2022; Fang et al., 2021), and text to image/video generation (Ramesh et al., 2021; Villegas et al., 2022). Models like CLIP (Radford et al., 2021) demonstrate that large-scale image-text pre-training can benefit transformer-based visual-language models. Following the trend, more multi-modal models such as GPT-4 significantly increased their training corpus (OpenAI, 2023) and added new modalities such as audio (Zellers et al., 2022).

Gaps. Even though recent multimodal models like GPT-4 exhibit impressive zero-shot performance, as they outperform most fine-tuned but smaller multi-modal models, they come with a cost. First, they lack a true understanding of the world (Hendricks and Nematzadeh, 2021; Thrush et al., 2022), they lack domain knowledge, and cannot generalize to real-life settings (e.g., personalized situations, in-the-wild data). Second, these models are very difficult or even impossible to interpret. They occasionally exhibit unreliable behaviors like hallucinations when generating new data (e.g., image/video generation, image/video captioning). Finally, only a few universities and institutions can afford the resources to use them properly. The cost of GPUs is constantly on the rise, and working with diverse modalities, visual in particular, is significantly more expensive in terms of both computer memory and computation.

Research Directions.

1. **How to best combine 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 (e.g., gestures can be used to express confidence in what is being expressed verbally), thus reducing the need for relying on billions of data points. However, in some cases, the modalities end up competing with each other, and thus many uni-modal models outperform multi-modal models (Wang et al., 2019a; Huang et al., 2021).

2. **Grounding with less studied modalities.** Most work on grounding revolves around visual, textual, or audio modalities. However, less studied modalities in the context of grounding, such as physiological, sensorial, or behavioral, are valuable in diverse applications such as 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). These modalities raise interesting questions across the entire pipeline, starting with data collection and representation, all the way to evaluation and deployment.

3. **Grounding "in the wild" and for diverse domains.** Most research around grounding is per-

formed on data collected in lab settings, or on images and videos from indoor activities such as movies (Lei et al., 2019) or cooking (Zhou et al., 2018). More realistic settings and outdoors "in the wild" data are much less studied (Castro et al., 2022). Such data poses new challenges with respect to availability, quality, distribution, and so on, which opens up new research directions. Moreover, the application of these models to diverse domains (e.g., robotics, medicine, navigation, education, accessibility) requires adapting to working with fewer data points or different types of data, alongside with the need for in-domain expertise to better understand the problem setup.

6 Computational Social Science

Background. Computational social science (CSS), the study of social sciences using computational methods, remains at least partly untouched by LLMs. While they can automate some of the languages tasks related to CSS such as sentiment analysis and stance detection (Liang et al., 2022), questions such as "how humans share news in social networks" or "the cultural differences in language use during catastrophic social events" are considered largely out of scope for generative models. With the success and impact of AI in social science in the past decade, computational and data-driven methods have penetrated major areas of social science (Lazer et al., 2009, 2020) giving rise to new interdisciplinary fields such as computational communication studies, computational economics, and computational political science.

Gaps. While NLP continues to have a large impact on shaping research in CSS, large foundation models are underutilized in hypothesizing and evaluating ideas in the field. Generative models are designed to serve users end-to-end through natural language, and often the need for customizing these large models is not addressed due to high fine-tuning costs or proprietary technology. In the absence of expert or fine-tuned LLMs, the applications of such models in CSS remain limited to generic data labeling and processing such as stance detection or sentiment analysis.

Research Directions.

1. **Population-level data annotation and labeling.** CSS researchers already apply less-than-perfect models on large datasets of human interactions to help them narrow down social concepts and study them. While some annotations can be handled by LLMs (Gilardi et al., 2023), the need for human crowdworkers will unlikely go away. This is particularly true in CSS, as researchers are mostly interested in population-level trends rather than precision at the individual level.

2. **Development of new CSS-aiding abstractions, concepts, and methods.** Word and sentence-level embeddings have had a large impact on CSS in recent years. Topic modeling, such as LDA (Blei et al., 2003), and keyword extraction have been prevalent in CSS prior to the introduction of embeddings. These are examples of methods that encapsulate generic capabilities at a high abstraction level in CSS, as they are frequently used in studies across several subfields of CSS. As CSS researchers transition to using more powerful AI technologies, the concepts and algorithms that unlock new capabilities for them are yet to be developed.

3. **Multicultural and multilingual CSS.** Most CSS studies focus on English or a handful of 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. For instance, how do languages evolve, or how do values vary across cultures? This is an area for future work that can lead to compounding impacts on the social sciences.

7 NLP for Online Environments

Background. The impact of NLP on online environments can be observed through two adversarial phenomena: content generation and moderation. The rapid generation of content, such as LLM-generated articles and social media updates, can be supported by various stakeholders. It is very likely that many can achieve high click-through rates to their websites by generating fake news and disinformation, which raises concerning social issues that necessitate timely regulation. Conversely, moderation is a form of gate-keeping. By using NLP to monitor and analyze user-generated content on digital platforms (Nakov et al., 2021; Kazemi et al., 2021a) to remove policy-violating materials, content moderation can maintain balance in the online ecosystem (Thorne et al., 2018; Nakov et al., 2021; Gillespie, 2020; Kazemi et al., 2021a; Shaar et al.,

2020).

Gaps. There are several concerns about content generation and moderation. For generation, it is of top priority to identify the underlying purpose of the generation and avoid malicious manipulation of users. For moderation, a concern is that the current moderation models are still opaque, unprecised, unaccountable, and poorly understood (Gorwa et al., 2020). Additionally, there are several existing challenges in building models to detect undesired content, including the difficulty in designing taxonomy for undesired content, the time-consuming nature of data labeling, and the inadequacy of academic datasets in revealing the real-world data distribution (Markov et al., 2023).

Research Directions.

1. **Detecting and debunking online misinformation.** Misleading content on the internet is growing in abundance, and an increase in volume in the upcoming years due to the rise in popularity of AI generated content is likely unavoidable. NLP can be used on several fronts to slow down the spread of misleading content. In the interest of extending help to fact-checkers and journalists, NLP systems remain underutilized, leaving a golden opportunity for building fact-checking technology that empowers fact-checkers to scale up their efforts (Kazemi et al., 2022). Additionally, NLP assisted fact-checking is often built in English, and therefore there is an increasing need for low-resource and cross-lingual NLP to help address misinformation in less resourceful parts of the world. Detecting and debunking misinformation also involves multimodal processing, since misinformation spreads in various formats. Network signals, such as who likes or reposts content, also encode rich information that can be attached alongside other modalities to help improve misinformation detection. Additionally, NLP for fact-checking can largely benefit from focusing on retrieval and knowledge augmented methods, since in order to check factuality of claims, one needs to search through and find the relevant context around the claim.

2. **Ensuring diverse representations.** With the prevalence of LLM-generated contents, the voice of majority 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, lack of diversity and especially representations for marginalized groups' voices will be a concerning problem as LLM-generated content will be increasingly used online.

3. **Avoiding mis-moderation and detecting 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 culture and social environments. It is important to make the moderation algorithms fair to all groups.

Conversely, due to various political interests (e.g., Iran wanting to limit the discussion of women freedom), governments are likely to limit the set of topics discussed online. It does become an important direction to trace what topics and opinions are filtered or demoted on the internet, and reflect on the freedom of speech in the political environment.

4. **Identifying stakeholders behind the generated contents.** As machine-generated content proliferates, it will be increasingly challenging to judge which information to trust. One promising direction is to develop NLP models to identify the stakeholders behind the generated contents, and their interest types, such as commercial profits (e.g., from advertisements or customer attraction) or political interests (e.g., to affect more people to hold certain opinions that would largely benefit an interest group).

8 Child Language Acquisition

Background. While some claim that LLMs "show sparks of AGI" (Bubeck et al., 2023b), they do not mimic the path followed by humans when acquiring language (Bowerman and Levinson, 2001). Ideally, we want smaller, more efficient models of language that are tightly paired with environment grounding (Lazaridou et al., 2017). On the path to efficient AGI, we have a hard-to-beat baseline: language acquisition in children. Most children can acquire up to three languages through often limited interactions and observations of language. While it is not completely understood how children learn language exactly, we know they do not require terabytes of text training instances.

There is also a growing body of research exploring the connection between LLMs and child language acquisition, specifically in the context of statistical learning (Wilcox et al., 2022), with recent research exploring how LLMs can be used to

model and simulate the statistical learning mechanisms that children use to acquire language (Contreras Kallens et al., 2023). Developments in this area have broader implications for low-resource and endangered languages as sample-efficient language modeling algorithms can unlock LLM-level capabilities to entirely new languages and cultures.

Gaps. Achieving the data efficiency of such an efficient baseline – children – is exciting, but there are no silver bullets: psychologists, neuroscientists, and linguists are among the scientists who have been studying language acquisition in children for decades and despite achieving greater understanding of the process in human children, we have not yet developed a working theory that reproduces the same process computationally with comparable data efficiency. This lack of progress can be attributed to the difficulty of studying children, as both recruitment and IRB approval for such studies rightly impose limitations on the types of data that can be collected. Among others, the little data that is collected is often limited in expressibility, as children who have not learned a language yet cannot communicate effectively, which limits experiment design. In a wide range of child language studies, parents are present to make sure the children can stay focused on the experiment and follow guidelines. Additionally, it is difficult to control confounding variables when you have no control over the subjects of the experiment.

Research Directions.

1. **Sample-efficient language learning.** This is an area ripe with opportunities to advance our understanding of language and develop more data efficient NLP tools. There is a great need for fundamental and theoretical research into sample-efficient language learning. Computational theories and algorithms for achieving state-of-the-art on smaller data regimes are an exciting area for researchers interested in core NLP, and the pursuit of the state-of-the-art performance may soon be rerouted to data-efficiency scores. Related to this direction is the goal of establishing baselines for sample-efficient lamnguage learning. Having a lower-bound goal (e.g. X hours of interaction achieving Y score) can enable the NLP community to have a more accurate understanding of progress in terms of data efficiency. While such estimates might already exist, getting more precision and depth will further advance our knowledge of language learning.

2. **Benchmark development in child language acquisition.** With the advancement of large language and multimodal systems, there are 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 language over a long period of time. Additionally, such datasets could be used to train models that are specifically tailored to the way that children learn language, which could enable new ways to understand child language use, as well as the development of models that are able to learn from fewer examples, similar to how humans learn language.

3. **Language models as biological models for child language acquisition**. A biological model refers to the study of a particular biological system, believed to possess crucial similarities with a specific human system, in order to gain insights and understanding about the human system in question. McCloskey has famously advocated for utilizing neural models as biological models to investigate human cognitive behavior and consequently develop theories regarding that behavior (McCloskey, 1991). With NLP models that have started to exhibit some similarities to human language use, we now have the opportunity to explore theories regarding how human infants acquire languages. For example, (Chang and Bergen, 2021) investigated the process of word acquisition in language models by creating learning curves and age of acquisition for individual words. Leveraging existing datasets such as WordBank (Frank et al., 2016) and CHILDES (MacWhinney, 1992), as well as new benchmarks, alongside with increasingly powerful language models, we now have the ability to conduct experiments to analyze language acquisition (e.g., phoneme-level acquisition, intrinsic rewards), and gain new insights into child language acquisition.

9 Non-Verbal Communication

Background. Non-verbal communication includes, among others, gestures, facial expressions, body language, and posture. A particular form of non-verbal communication consists of sign language, which represents the primary medium of communication used by people who are deaf. Sev-

eral studies have shown the importance of nonverbal communication in everyday interactions (McNeill, 1992; Alibali et al., 2000). Recent work in NLP has highlighted the importance of integrating non-verbal information into existing language representations as a way to obtain richer representations, including for instance language models (Wang et al., 2019b) or visual models (Fan et al., 2021); other previous work has shown that nonverbal communication such as facial expressions or gestures are aligned with the verbal channel and that different cultural or language contexts can be associated with different interpretations of these non-verbal expressions (Abzaliev et al., 2022; Matsumoto and Assar, 1992). There is also an entire body of research focused on the understanding and generation of sign language (Joze, 2019; Bragg et al., 2019), as well as the communication across different communities of sign language speakers (Camgoz et al., 2020).

Gaps. Understanding the alignment between non-verbal modalities and verbal language remains an open problem, especially given the challenges that some of these modalities use different spectrums (continuous vs discrete). Correspondingly, the discretization and interpretation of these signals can be difficult, leading to challenges regarding their joint use, or the integration of such non-verbal information into existing large language-based models. In sign language research, there are still many open problems in understanding and generating sign languages, encompassing both the compilation of representative sign language datasets and the development of effective computational models.

Research Directions.

1. **Non-verbal language interpretation.** Since many subareas of non-verbal communication require non-verbal information, the representation, discretization, and interpretation of this information is a rich direction of exploration. For instance, while previous work has identified a potential "code-book" of facial expressions (Song et al., 2013), more work is needed to find the ideal set of representations that can be used across modalities, contexts and cultures. The interpretation of these expressions and gestures, and their alignment across modalities, also remain an open problem. In particular, the increasing use of LLMs has the potential to open up new paradigms for understanding non-verbal communication through textual descriptions. For instance, when an LLM is prompted with "Please answer which gesture I am describing: a person puts her arms wide away, smiling and moving towards the other person," it replies with "The gesture you are describing is likely a hug, indicating a friendly or affectionate greeting or farewell. . . ," which can be used as a textual representation of the hugging gesture.

2. **Sign language understanding, generation, and translation.** An open research problem is the development of sign language lexicons (Athitsos et al., 2008) and corpora (Li et al., 2020) that can be used to train and evaluate computational models. These resources are essential for developing and testing recognition and interpretation models, but they are often difficult and expensive to create. In sign language understanding, one of the biggest challenges is the development of effective models that can accurately recognize and interpret sign language gestures. This is difficult because sign languages exhibit a relatively high degree of variability in manual gestures, including differences in handshape, movement, and orientation; additionally, other non-manual features such as facial expressions, body pose and eye gaze often play a role in sign languages, which can further complicate the recognition process. Finally, sign language generation is also an open research area, with challenges residing in the development of generation models that can lead to sign language communication that is fluent and expressive. Such models are needed to enable or enrich communication between speakers of the same sign language; speakers of different sign languages; or speaker of verbal and sign languages.

3. **Effective joint verbal and non-verbal communication.** Ultimately, both verbal and non-verbal signals should be considered during communication. We want AI systems to 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. Open research problems encompass not only the development of language models for each of these modalities, but effective fusion methodologies, which will enable large joint models for simultaneous verbal and non-verbal communication.

10 Synthetic Datasets

Background. In NLP research, synthetic data is typically needed when the more traditional human data collection is infeasible, expensive, or has privacy concerns (Mattern et al., 2022). With the advancement of generative models (Tang et al., 2023), synthetic data generation has seen applicability in various domains. Examples include back-translation for low-resource languages (Sennrich et al., 2015; Edunov et al., 2018), semantic parsing (Rosenbaum et al., 2022a), intent classification (Rosenbaum et al., 2022b), structured data generation (Borisov et al., 2022), or medical dialogue generation (Chintagunta et al., 2021a; Liednikova et al., 2020). The process typically involves pre-training the model if domain adaptation is necessary (Chintagunta et al., 2021b), prompting the model to generate the dataset, and evaluating its quality automatically or via expert validation.

Gaps. The use of synthetic data faces challenges such as difficulty in data quality control (Kim et al., 2022) (due to lack of evaluation metrics for text generation), lack of diversity, potential bias in the data-generating model, and inherent limitations of the data-generating model such as hardship to capture of long-range dependency (Orbach and Goldberg, 2020; Guan et al., 2020).

Research Directions.

1. **Knowledge distillation.** Knowledge distillation is the task of transferring knowledge from teacher models to typically smaller student models. For example, Kim et al. (2022) frame their synthetic dialog dataset as having been distilled from InstructGPT. While earlier methods for distillation involved learning from the soft output logits of teacher models (Hinton et al., 2015), this signals a move toward directly utilizing LLM outputs as synthetic examples (West 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, such as in the case of Alpaca (Taori et al., 2023).

2. **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.

3. **Transforming existing datasets.** Given an existing dataset, we 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 news articles from HTML to plain text format), modality transfer (e.g., generating textual descriptions of images or videos or generating captions or subtitles for audio-visual content), or style transfer (Jin et al., 2022a) (e.g., translating the writing style of the text from verbose to concise).

11 Interpretability

Background. Interpretability is the task of understanding and explaining the decision-making processes of machine learning models, making them more transparent and justifiable (Danilevsky et al., 2020). Interpretable NLP systems can foster thrust by enabling end-users, practitioners, and researchers to understand the model's prediction mechanisms, and ensure ethical NLP practices. Historically, traditional NLP systems, such as rule-based methods (Woods, 1973), Hidden Markov models (Ghahramani, 2001; Rabiner, 1989), and logistic regression (Cramer, 2002), were inherently interpretable, known as white-box techniques. However, recent advancements in NLP, most of which are black-box methods, come at the cost of a loss in interpretability. To address this issue, interpretability has emerged as a research direction, focusing on developing techniques that provide insight into the inner workings of NLP models (Mathews, 2019; Danilevsky et al., 2020). Key research findings include attention mechanisms, rule-based systems, and visualization methods that help bridge the gap between complex language models and human interpretability, ultimately contributing to the responsible deployment of NLP systems.

Gaps. The current state of interpretability research in NLP focuses on understanding model predictions, feature importance, and decision-making processes. Techniques like attention mechanisms (Vaswani et al., 2017), LIME (Ribeiro et al., 2016), and SHAP (Lundberg and Lee, 2017) have emerged to provide insights into model behavior.

However, gaps remain in areas like robustness, generalizability, and ethical considerations. Additionally, interpretability methods often lack standardization and struggle to address complex, large-scale models like transformers, limiting their applicability in real-world scenarios.

Research Directions.

1. **Probing.** One promising direction is to investigate the internal representations 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, 2022). This can help understand the reasoning capabilities of models and identify potential biases (Li et al., 2022; Meng et al., 2022).

2. **Mechanistic Interpretability.** While probing mostly looks at the attributes of the features learned by the model, mechanistic interpretability aims to uncover the underlying *mechanisms and algorithms* within a model that contribute to its decision-making process (Nanda et al., 2023; Conmy et al., 2023). It extracts computational subgraphs from neural networks (Conmy et al., 2023; Wang et al., 2023; Geiger et al., 2021), and its high-level goal is to reverse engineer the entire deep neural network (Chughtai et al., 2023).

3. **Improving interpretability by human-in-the-loop.** Human-in-the-loop interpretability research in NLP focuses on incorporating human feedback and expertise to enhance model interpretability. This approach aims to 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., 2023).

4. **Basing the generated text on references.** Explainability relates to understanding why a certain generative NLP model output is provided and evaluating its correctness, possibly through calibration (Naeini et al., 2015). Being factually correct is not a restriction that generative models have to follow; rather, they are generally trained to imitate human-written text by predicting the most likely text that comes next. This predicted text, in turn, is prone to hallucinations (Ji et al., 2022) that causes a lack of trust from the users. A promising solution is to provide reliable sources for the facts output by a model, by attaching references and showing any additional reasoning steps. For example, citations can be included along with its bibliography, or pointers to documents in the training data (or a document database) can be attached to the output. Such a system should evaluate the extent to which these sources back up the claims made by the model.

12 Efficient NLP

Background. Efficient NLP is a research direction aiming to optimize the use of resources for NLP models. This objective arises from the need to address the challenges posed by the increasing scale of language models and their growing resource consumption present new challenges for NLP advances (Touvron et al., 2023b; Zhang et al., 2023). Indeed, it is widely acknowledged that scaling up is an essential approach for achieving state-of-the-art performance on NLP tasks, especially those skills emerged with the scaling law (Wei et al., 2022; Bowman, 2023). However, developing LLMs requires substantial energy and financial resources for training and inference, which raises concerns about the AI carbon footprint and the economic burden on NLP product development (Strubell et al., 2019). In light of these concerns, prior research has underscored the critical need for effectively reducting CO2 equivalent emissions (CO2e) and Megawatt hours (MWh), and increase of Power Usage Effectiveness (Patterson et al., 2022; Thompson et al., 2020).

Gaps. There is significant scope for improving the efficiency of NLP across various dimensions, including data curation, model design, and training paradigms, presenting numerous research opportunities. Addressing data efficiency involves tackling challenges like enhancing data deduplication techniques, assessing data quality, and curating vast amounts of data. When it comes to refining model design, key challenges include improving the attention mechanism efficiency, developing alternative no-parameters modules for parameter reduction, and optimizing the model depth or efficiency. Lastly, in the realm of training paradigms, there is potential for advancements in promoting engineering, fine-tuning, and prompt-tuning techniques.

Research Directions.

1. **Data efficiency.** Data efficiency can be enhanced through data deduplication, where redundant or noisy data is removed, thereby improving performance with fewer data items. Although there is existing work that aims to boost model performance with fewer data points by removing noisy examples and deduplicating useless data (Lee et al., 2022; Mishra and Sachdeva, 2020; Hoffmann et al., 2022), there is a lack of effective methods for data deduplication for vast corpora (>700B Tokens) or raw web data curation.

2. **Model design.** A large body of methods increases model efficiency by improving attention mechanisms (Tay et al., 2020, 2022; Dao et al., 2022; Ma et al., 2022). However, challenges remain in handling extremely long context modeling in transformer architectures. Sparsing models can scale up the width of models for increased expressiveness while reducing theoretical FLOPs. Notable practices include applying mixture of experts architectures in a feed-forward layer of a transformer-based model (Fedus et al., 2021, 2022; Du et al., 2022). Engineering such models requires architecture-specific implementation and costs many trials to get the optimal architecture. It is also not unstable in performance (Mustafa et al., 2022).

3. **Efficient downstream task adaptation.** Efficient fine-tuning aims to adapt the pre-trained model to downstream tasks by updating a small part of the parameters (Pfeiffer et al., 2020; Moosavi et al., 2022; Schick and Schütze, 2021). Prompt-tuning/prefix tuning modifies activations with additionally learned vectors without changing model parameters (Valipour et al., 2022; Lester et al., 2021). However, it is necessary to find a way for efficient automatic prompt construction.

13 NLP in Education

Background. There is a rich history of NLP applications for education, including dedicated workshops such as the yearly ACL Workshop on Innovative Use of NLP for Building Educational Applications organized by the Special Interest Group for Building Educational Applications. These applications include tools to aid learners (e.g., language learning applications such as Duolingo[*], or grammar correction tools such as Grammarly[*]), tools to assist teachers and organizations in grading (e.g., the e-rater system that is used to grade GRE essays (Burstein et al., 1997)), tools to assist curriculum and assessment development (e.g., systems for developing multiple-choice questions (Kurdi et al., 2020)) and tools for education researchers (e.g., systems to build representations of classroom interactions (Alic et al., 2022)). Researchers have been testing the application of models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) in these areas since their release, and are now beginning to incorporate larger models.

Gaps. Many of the deployed NLP applications in the education space have been developed prior to wide spread use of LLMs, and we are likely to see large-scale deployment of task-specific models based on LLMs soon. While much of the prior work includes standalone applications, developing models that can easily be incorporated into existing education pipelines, e.g., by integrating what students have learned thus far, is an area open for further exploration. Importantly, a long-standing goal in education is to personalize materials and assessments to the needs of individual students, and NLP has the potential to contribute towards that goal.

Research Directions.

1. **Controllable text generation.** Dialog systems and more generally text generation have been previously used in education applications. Within this space, controllable text generation can be used for a more personalized experience, for instance to introduce students to new terms using automatically generated stories related to their interests or to modify stories to be accessible to grade school students with different reading levels. Similarly, while we have seen extensive work in reading comprehension, we can now start to imagine applications where the comprehension of a text will be tested based on a student's prior experience, as well as previous tests that they have been exposed to, for a more adaptable learning experience.

2. **Educational explanation generation.** Personalized classroom material could also include the generation of explanations for students, grounded in their understanding (or lack thereof) of the material. For example, an NLP system could be used

[*] https://www.duolingo.com

[*] https://www.grammarly.com

to help a student understand a tricky sentence in an academic paper, or to rephrase an answer given by their instructor in hopes of discovering an explanation that connects to the student's body of knowledge. Automatic grading is also an area where NLP has made many contributions (Mohler and Mihalcea, 2009), but it still encompasses open research problems such as providing an explanation for a less than perfect grade.

3. **Intelligent tutoring systems.** Intelligent tutoring systems show significant promise for personalized education (Mousavinasab et al., 2021). NLP methods can be developed to generate targeted practice questions and explain students' mistakes in a wide range of fields, all the way from English or History to Physics or Computer Science. These systems will likely improve as NLP evolves to mimic human reasoning more reliably; currently, it is necessary to be careful when deploying NLP in education without a human-in-the-loop, as even when given simple math problems, NLP models (including the most recent LLMs (OpenAI, 2023)) can often confidently give incorrect answers and explanations.

It is worth mentioning that the reception of LLMs in the education community has largely been one of fear due to the possibility of increased academic dishonesty. This has led to courses and universities adopting policies regulating how AI can be used in their courses, such as the Yale policy.[*] Whether the overall curriculum will be adjusted to incorporate LLMs in a positive manner is yet to be seen, but we are optimistic that this recent progress can have a positive impact on education if deployed in appropriate circumstances.

14 NLP in Healthcare

Background. Applications for NLP in healthcare can be classified by their use and impact on key stakeholders such as providers, patients, and public health officials (Zhou et al., 2022; Olaronke and Olaleke, 2015). When focusing on health providers, NLP is often used to support clinical decision making by (1) aggregating and consolidating available data and research, and (2) extracting relevant information from data. These tasks involve important challenges such as standardization of healthcare data, accurate labeling, extraction and retrieval of health concepts as well as categorization of patient conditions (Dash et al., 2019). Similarly, NLP is used to address patient requests for information on applications such as question answering for health-related questions, and retrieval of information relevant to medical treatments or illnesses. Recent work in this area has focused on the analysis of language in the mental health space covering both professional therapy (Sharma et al., 2020; Pérez-Rosas et al., 2017; Min et al., 2022) and social media conversations (Tabak and Purver, 2020; Lee et al., 2021; Biester et al., 2020). Regarding assisting public health officials, NLP is being used for surveillance of public health to identify diseases and risk factors or at-risk populations (Naseem et al., 2022; Jimeno Yepes et al., 2015; Yates et al., 2014) and also to moderate aspects such as misinformation or public sentiment online (Hou et al., 2019; Kazemi et al., 2021b).

Gaps. One of the most glaring limitations of NLP in healthcare is the scarcity of high-quality, annotated clinical data. Although social media data can be useful in some contexts, clinical data is essential in developing tools for clinical decision making, and often not publicly available due to privacy and ethics concerns. Another shortcoming is the lack of language diversity as work to date has primarily focused on English or other high-resource languages (Mondal et al., 2022) but devoted less efforts towards minority languages. Additionally, the lack of human evaluation of NLP-based health systems has made it challenging to measure their effectiveness in the real world. Current automatic evaluation metrics do not necessarily speak to patient outcomes. Hence, human-centric studies must be conducted in evaluating the efficacy of NLP-powered tools in healthcare.

Research Directions.

1. **Healthcare benchmark construction.** Although the documentation of recent LLMs reports very high performance for various medical question answering benchmarks, or medical licensing texts, there are many other tasks in healthcare that lack the data required to achieve similarly good performance. Access to medical datasets is often limited because of privacy issues, and therefore other approaches may be required to compile such benchmarks. Synthetic datasets are one such option (Chintagunta et al., 2021a; Liednikova et al., 2020). Other options including paraphrasing of existing

[*] https://poorvucenter.yale.edu/AIguidance

datasets as a form of data augmentatiom; or using LLMs as a starting point to bootstrap datasets. Another open research direction is the evaluation of the quality of the benchmarks. Additionally, research is needed to find effective ways to produce new health datasets in low-resource languages or low-resource domains.

2. **NLP for clinical decisions.** NLP systems can used as brainstorming or decision making tools that can assist experts in their evaluation and decision process. They can be used to synthesize new knowledge (e.g., the latest research papers on a medical finding), and make that available to the medical practitioners. Further, bringing together general medical knowledge and personal patient information requires new strategies for knowledge integration. Since clinical diagnoses and treatments are high-stake decisions, it is crucial that the NLP systems be reliable and interpretable, to provide clear reasoning behind their predictions. Such processes also require the interdisciplinary collaboration with medical experts to make sure that the system aligns with their domain knowledge and clinical practice.

3. **Drug discovery.** Drug discovery is a critical research area that has often been considered in relation to biomedical and chemical research, but more recently has gained the attention of NLP researchers. NLP methods can enable the efficient extraction and analysis of information from large amounts of scientific literature, patents, sovical media, clinical records, and other biomedical sources. Open research directions include the identification and prioritization of drug-target interactions, the discovery of new drug candidates, the prediction of compound properties, and the optimization drug design. New NLP methods can also contribute to the identification of novel drug-target associations and can enable more effective drug repurposing efforts.

15 NLP and Ethics

Background. Recognition for the role of ethics in NLP is on the rise, especially with the development of increasingly powerful models with potentially far-reaching societal implications. There are important ethical considerations when developing NLP models (Bender et al., 2020), and there is ongoing research work that aims to address critical ethical aspects such as dual use, fairness, and privacy.

Gaps. Aside from the issues described above, other ethical concerns surrounding the use and applications of recent LLMs include: lack of attribution, poor model explainability, skill degradation, disruption of the labor market, model misuse, and model disuse. In addition to educating people about ethics, we need further investigation into the extent of these concerns and determining how NLP techniques can reduce their impact.

Research Directions.

1. **Dual use.** Many NLP applications that have positive impact can at the same time be used in harmful ways. Identifying possible harm from NLP models and applications can be achieved through discussions before deployment and data surveys after deployment to identify potentially harmful applications. Additionally, developing NLP systems that help detect, discourage, and prevent harmful use, such as fact-checkers, is crucial. Adversarial NLP can also be used to explore the limitations and loopholes of NLP systems and improve their robustness.

2. **Fairness.** There is a need for methods that evaluate the fairness of NLP models, and detect and mitigate bias. This includes investigating dataset creation practices and their correlation with model bias (Wang et al., 2020). Such research should examine whether stricter requirements for dataset creation can reduce bias and inequalities that might be exacerbated by models trained on or evaluated on biased data.

3. **Privacy.** As personalized NLP applications (including in fields such as education or healthcare) require an understanding of the user, privacy protection in NLP systems has become an essential research direction. New techniques are needed to identify and anonymize sensitive user information while maintaining the utility of the data for analysis and decision-making. This includes methods such as differential privacy, federated learning, and secure multi-party computation to ensure the confidentiality and security of patient data in NLP-driven healthcare applications. Additionally, an area where NLP systems can make an impact is data policy, where NLP methods can be developed for summarizing data policies of digital products in understandable formats for users, and ensuring model alignment with such policies (Carlini et al., 2021).

4. **Attribution and detection of machine-generated data.** Developing standard approaches for attribution that NLP models can use while generating content is essential (i.e., can we teach AI models to attribute content using membership inference or other approaches?) (Collins, 2023). Domains such as programming or creative writing (Swanson et al., 2021) have already begun incorporating LLMs into the workflow, which requires the determination of the ownership and rights to such creations.

5. **Integrating NLP models as human assistants rather than human replacements.** This can be achieved using NLP models for human training applications. Models could be used to improve human spelling, writing, and reading comprehension abilities. However, it is essential to note that LMs have shown excellent ability in masquerading wrong answers as correct. These answers can be delivered to a student whose job is to find the loopholes in the argument or choose the incorrect answer. It also has the potential to widen the existing inequalities in society. It also raises concerns about the ethical implications of relying on machines to augment human performance and how this could affect our perception of what it means to be human (Eloundou et al., 2023).

16 So What Should *I* Work On?

The future of NLP research is bright. The rapid progress we are currently witnessing in LLMs does not mean that "it's all been solved." On the contrary, as highlighted in this document, there are numerous unexplored research directions within NLP that are not addressed by the current progress in LLMs. They add to the many existing tasks in NLP where LLMs's performance is limited (Bang et al., 2023a), as well as the growing number of new areas that are 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 and should be fundamentally human-centric. This brings about a new focus on enabling technologies that are culture- and demographic-aware, that are robust, interpretable, and efficient, and that are aligned with a strong ethical foundation — ultimately, technologies that make a lasting positive impact on the society.

How to choose a research direction to work on? 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 and applications in NLP that connect to your motivations and interests. This document hopefully serves as a starting point to guide this exploration.

Acknowledgments

We would like to thank Steve Abney and Rui Zhang for providing feedback and valuable suggestions on earlier versions of this manuscript.

References

M Abouelenien, V Pérez-Rosas, and others. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions*. 6

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. 10

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual question answering. 6

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. 4

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. 5

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. 10

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. 13

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. 6

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. 3

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 0:1–8. 10

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. 5

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. 2, 16

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. 3

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. 15

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. 14

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. 6

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. 7

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. 11

Melissa Bowerman and Stephen Levinson. 2001. *Language Acquisition and Conceptual Development*. Language Culture and Cognition. Cambridge University Press. 8

Samuel R Bowman. 2023. Eight things to know about large language models. 12

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. 10

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. 1

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Others. 2020a. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, 33:1877–1901. 5

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. 2020b. 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*. 4

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. 5

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. 8

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. 13

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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. 10

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. 15

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. 7

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. 3

Tyler A. Chang and Benjamin K. Bergen. 2021. Word acquisition in neural language models. 9

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. 11, 14

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. 11

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. *CoRR*, abs/2302.03025. 12

Keith Collins. 2023. How ChatGPT could embed a 'watermark' in the text it generates. *The New York Times*. 16

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. 12

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. 9

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. 6

J S Cramer. 2002. The origins of logistic regression. 11

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. 11

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. 13

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. 14

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805. 13

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. 13

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at scale. 3, 11

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. 12

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. 16

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. 10

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. CLIP2Video: Mastering Video-Text retrieval via image CLIP. 6

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning. 13

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. 13

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. 9

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. 12

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. 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. 12

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. 12

Zoubin Ghahramani. 2001. An introduction to hidden markov models and bayesian networks. *Int. J. Pattern Recognit. Artif. Intell.*, 15:9–42. 11

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms Crowd-Workers for Text-Annotation tasks. 7

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

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. 4

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. 8

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. 3

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. 11

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing Image-Language transformers for verb understanding. 6

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*. 4

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. 12

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. 12

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. 11

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal large language models. 13

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. 14

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). 6

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. 4

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzman. 2022. OCR improves machine translation for Low-Resource languages. *arXiv preprint arXiv*. 3

F Jelinek. 1976. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64(4):532–556. 1

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. 12

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. 6

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. 14

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. 11

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. 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*. 4

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. 4

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022b. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180Ã¢â‚¬â€œ7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 4

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. 4

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? 4

Hamid Reza Vaezi Joze. 2019. MS-ASL: A Large-Scale data set and benchmark for understanding american sign language. *bmvc2019.org*. 10

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*. 8

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. 7

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. 14

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. 4

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. 11

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. 3

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. 13

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. 8

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. 7

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. 7

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. 14

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Annual Meeting of the Association for Computational Linguistics*. 3

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. 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. 13

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*. 7

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. 13

Belinda Li, Jane Yu, Madian Khabsa, Luke Zettlemoyer, Alon Halevy, and Jacob Andreas. 2022. 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. 12

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. 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. 10

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. 7

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning healthbots 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. 11, 14

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692. 13

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *ArXiv*, abs/1705.07874. 11

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: Moving average equipped gated attention. 13

Brian MacWhinney. 1992. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8(2):217–218. 9

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. 8

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. 11

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. 10

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. 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. 11

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. 3

Michael McCloskey. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6):387–395. 9

David McNeill. 1992. Hand and mind: What gestures reveal about thought. 416. 10

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*. 12

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. 2023. Augmented language models: A survey. *CoRR*, abs/2302.07842. 4

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. 4

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. 14

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. 13

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. 14

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP prefix for image captioning. 6

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. 14

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. 13

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. 12

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054. 12

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. 14

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. 3

Basil Mustafa, Carlos Riquelme Ruiz, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with LIMoE: the Language-Image mixture of experts. 13

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. 12

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. 7

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *CoRR*, abs/2301.05217. 12

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. 14

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. 3

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. 5

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. 14

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. 5

OpenAI. 2023. GPT-4 technical report. 4, 6, 14

Eyal Orbach and Yoav Goldberg. 2020. Facts2Story: Controlling text generation by key facts. 11

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. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155. 4

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. 6

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. 12

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. 14

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. 13

Lawrence R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286. 11

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. 6

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image generation. 6

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. 3

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*. 6

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*. 11

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR post correction for endangered language texts. 3

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Marco Damonte, and Isabel Groves. 2022a. CLASP: Few-Shot Cross-Lingual data augmentation for semantic parsing. 11

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. 11

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. 4

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. 13

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. 3, 11

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. 7

C E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. 1

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. 14

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. 5

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. 6

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. 3

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. 10

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. *AAAI*, 31(1). 5

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. 4

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. 12

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. 5

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. 16

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. 14

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? 11

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`. 11

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. 13

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):1–28. 13

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. 12

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. 7

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. 6

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. 2023a. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971. 4

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. 2023b. LLaMA: Open and efficient foundation language models. 12

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. 13

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*. 11

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. 3

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. 6

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. 15

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*. 12

Weiyao Wang, Du Tran, and Matt Feiszli. 2019a. What makes training Multi-Modal classification networks hard? 6

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a foreign language: BEiT pretraining for all vision and Vision-Language tasks. 6

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019b. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proc. Conf. AAAI Artif. Intell.*, 33(1):7216–7223. 10

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. 12

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. 11

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. 5

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88. 8

Maurice V Wilkes. 1994. *Using Large Corpora*. MIT Press. 1

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. 11

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). 3

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. 3

Jitao Xu and Franccois Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *CALCS*. 3

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. 5

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). 14

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. 5

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. 5

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. 6

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. 5

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. 12

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. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068. 4

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. 5

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. 3

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2022. Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*, pages 1–17. 14

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. 6

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. 7