LLM-driven Instruction Following: Progresses and Concerns

Wenpeng Yin[†], **Qinyuan Ye**[‡], **Pengfei Liu**[◇], **Xiang Ren**[‡] and **Hinrich Schütze**[♯]
[†]Penn State; [‡]USC; [◇]SJTU; [♯]LMU Munich
wenpeng@psu.edu; {qinyuany, xiangren}@usc.edu
stefanpengfei@gmail.com; hinrich@hotmail.com

Abstract

The progress of natural language processing (NLP) is primarily driven by machine learning that optimizes a system on a large-scale set of task-specific labeled examples. This learning paradigm limits the ability of machines to have the same capabilities as humans in handling new tasks since humans can often solve unseen tasks with a couple of examples accompanied by task instructions. In addition, we may not have a chance to prepare task-specific examples of large-volume for new tasks because we cannot foresee what task needs to be addressed next and how complex to annotate for it. Therefore, task instructions act as a novel and promising resource for supervision.

This tutorial targets researchers and practitioners who are interested in AI and ML technologies for NLP generalization in a low-shot scenario. In particular, we will present a diverse thread of instruction-driven NLP studies that try to answer the following questions: (i) What is task instruction? (ii) How is the process of creating datasets and evaluating systems conducted? (iii) How to encode task instructions? (iv) When and why do some instructions work better? (v) What concerns remain in LLM-driven instruction following? We will discuss several lines of frontier research that tackle those challenges and will conclude the tutorial by outlining directions for further investigation.

1 Introduction

This proposal is driven by a fundamental question of task generalization in NLP: how to comprehend a new task if labeled examples are pretty limited? One goal of AI is to build a system that can continually understand and solve new tasks. Labeled examples, as the mainstream task representation, are unlikely to be available in large numbers or even do not exist. Then, is there any other task representation that can contribute to task comprehension? Task instructions provide another dimension of supervision for expressing the task semantics. Instructions often contain more abstract and comprehensive knowledge of the target task than individual labeled examples. With the availability of task instructions, systems can be quickly built to handle new tasks, especially when task-specific annotations are scarce (Wang et al., 2022; Yin et al., 2022). Instruction following is inspired by the typical human learning for new tasks, e.g., a little kid can well solve a new mathematical task by learning from its instruction and a few examples. This new learning paradigm has recently begun to attract the attention of the machine learning and NLP communities.

Despite the importance, frontier research in instruction following is still struggling with the following questions. First, should instructions be constructed to express the target task as detailed as possible (e.g., MTurk instructions (Mishra et al., 2022)) or to align with the format of supervising tasks (e.g., natural language inference (Yin et al., 2019) or language modeling (Brown et al., 2020)) as well as possible? Second, how to effectively encode instructions that may consist of some specific requirements such as “maximal output length 5”, and “do not generate anything else apart from one of the following $\cdots$”? Third, what are the factors (e.g., model size, task numbers) that influence a system’s generalization, robustness, etc.? Fourth, how to evaluate instruction-following systems? Last, what is the future for academia and industry in this ChatGPT era?

In this tutorial, we will systematically review several lines of frontier research on developing systems that are supervised by task instructions. Beyond introducing pioneering work that parsed instructions to cope with individual tasks, such as soccer game (Kuhlmann et al., 2004), software control (Branavan et al., 2009, 2011), etc., we will focus on recent LLM-based approaches for cross-task generalization given task instructions. Specifically, in light of the heterogeneous formats and dis-

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 19–25
December 6-10, 2023 ©2023 Association for Computational Linguistics

parate rationales underlying instructions, we shall endeavor to establish a unified lens for interpreting the essence of various instructions. Subsequently, a structured exposition and critical analysis will be undertaken, encompassing a spectrum of aspects such as diverse instruction-following datasets, rigorous evaluation methodologies, multifaceted performance-influencing factors, and lingering concerns within this domain.

Participants will learn about recent trends and emerging challenges in this topic, representative tools and learning resources to obtain ready-to-use models, and how related technologies benefit end-user NLP applications.

2 Outline of Tutorial Content

This **half-day** tutorial presents a systematic overview of recent advancements in NLP with supervision from task instructions. The detailed contents are outlined below.

2.1 Background and motivation [20min]

We will define the main research problem and motivate the topic by presenting several real-world NLP and instruction-driven AI applications, as well as several key challenges that are at the core of classic machine learning.

2.2 What is the essence of instructions? [30min]

Various researchers may hold differing viewpoints on the nature of instructions, with some specializing in particular types of instructions while overlooking the interconnections among various instruction categories. In this section, we aim to establish a unified perspective for understanding the essence of instructions.

We begin by introducing various typical forms of instructions. For instance, some instructions serve to elucidate the output labels in classification tasks, as exemplified by **NLI-oriented task instructions** (Yin et al., 2019; Xu et al., 2022; Li et al., 2022; Xia et al., 2021; Sainz et al., 2021, 2022). These instructions treat the outputs as hypotheses and transform the target problems into natural language inference (NLI) to leverage the supervision available in existing NLI datasets. Other instructions aim to enhance the input text, such as prompts, which are designed to leverage the rich supervision from pretrained language models (Radford et al., 2019; Schick and Schütze, 2021b,a, 2022). Thus, they are referred to as **LM-oriented instructions**. Additionally, there are more natural instructions contributed by end-users who lack expertise in machine learning or LLMs. These instructions attempt to convey the task's semantics regardless of the specific technique to be employed. We categorize these as **human-oriented instructions** (Efrat and Levy, 2020; Mishra et al., 2022; Wang et al., 2022; Lou et al., 2023). To adhere to human-oriented instructions, LLMs are frequently trained on a diverse array of instruction-following tasks. Consequently, we consolidate these distinct types of instructions under the umbrella term *instructions as supervision-oriented textual expressions*.

2.3 Instruction-following datasets and evaluations [30min]

Initially, we introduce a range of **crowdsourced datasets**, which include P3 (Sanh et al.), Big-bench (Srivastava et al., 2022), Dolly (Conover et al., 2023), Natural-Instructions (Mishra et al., 2022; Wang et al., 2022), Multi-Instruct (Xu et al., 2023b), etc. Nevertheless, human-crafted datasets have inherent limitations due to the constraints of human effort, making it challenging to expand the diversity and complexity of tasks. Consequently, recent efforts have turned to **LLM-generated datasets**, as exemplified by Self-Instruct (Wang et al., 2023), Unnatural-Instruct (Honovich et al., 2023), Dynosaur (Yin et al., 2023), WizardLM (Xu et al., 2023a), LongForm (Köksal et al., 2023), Muffin (Lou et al., 2023), and others. Irrespective of the datasets' origin, this tutorial will elucidate their objectives and distinctions from a scaling perspective.

Regarding the evaluation, we commence with automated assessments conducted on a selection of high-quality crowdsourced datasets, including Natural-Instructions (Mishra et al., 2022; Wang et al., 2022), T0 (Sanh et al.), Big-bench (Srivastava et al., 2022), etc. Subsequently, we introduce Vicuna system (Chiang et al., 2023), which employed GPT-4[1] for automated evaluations. Finally, we proceed to human assessments, which take into account various criteria, as demonstrated in works such as (Wang et al., 2023; Yin et al., 2023; Askell et al., 2021).

[1] https://openai.com/research/gpt-4

2.4 Methodology for instruction tuning [30min]

An established experimental framework for instruction tuning entails initially training a model on a set of provided instructions and subsequently assessing its performance on unseen instructions. In this context, we will present three distinct methodologies for modeling instructions: (i) The **Concatenation** method, which involves the straightforward concatenation of elements from the instruction and task input to form a lengthy textual sequence. This composite sequence is then fed into an LLM to generate the desired output. Representative works include (Mishra et al., 2022; Wang et al., 2022; Yin et al., 2022). (ii) **Hypernetwork-based approaches** (Ye and Ren, 2021; Ivison et al., 2022), where a hypernetwork (Ha et al., 2017) is trained to generate instruction-specific model parameters, which are subsequently integrated into a primary network. (iii) **Reinforcement learning with human feedback** methods (Bai et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020), which involve the utilization of reinforcement learning techniques guided by human-provided comparison data.

2.5 When and why it works [30min]

Most instruction-driven systems assume that each task has a single instruction. We can imagine that different users can convey a task with instructions of distinct textual expressions. Some prompt-based LLMs also show varying performance in dealing with prompts of different templates (Schick and Schütze, 2022; Kojima et al., 2022). A question arises: how to predict and explain an instruction' behavior? To the end, we first introduce the work by Gu et al. (2023) that explored the robustness of pretrained instruction learning system in handling (i) the same task with distinct instructions written by different MTurkers, and (ii) instruction of varying degrees of abstractions. Then, we present a series of works that i) explain prompts performance by LLM-oriented perplexity (Gonen et al., 2022), the model bias (Zhao et al., 2021), or ii) improve instructions by reformulating them into more effective ones (Khashabi et al., 2022).

2.6 Concerns of instruction following [30min]

In this section, we will address concerns related to instruction following across four distinct dimensions: (i) The “inverse scaling law” observed in LLMs when dealing with negation (Mishra et al., 2022; Jang et al., 2022; Hossain et al., 2022). (ii) Unanticipated behavior arising in the realm of instruction comprehension, drawing from human capabilities in following instructions (Webson and Pavlick, 2022). (iii) The issue of task-hungry models. Despite shifting our research focus from cross-example generalization to cross-task generation, the creation of large-scale instruction-following datasets presents another challenge. To enhance LLMs’ instruction-following abilities for new tasks, the collection of extensive training tasks becomes a necessity. (iv) The emergence of adversarial instruction attacks (Shu et al., 2023; Wan et al., 2023; Kang et al., 2023; Li et al., 2023).

2.7 Future directions [10min]

In the last section, we will discuss some critical and foreshadowing research directions, such as scalable oversight and alignment (Hendrycks et al., 2021; Bowman et al., 2022), explainable instruction learning, and how to encode instructions without the help of labeled examples, etc.

3 Specification of the Tutorial

The proposed tutorial is considered a **cutting-edge** tutorial that introduces new frontiers in instruction-driven NLP. The presented topic has not been covered by any ACL/EMNLP/EACL/NAACL/AACL/COLING tutorials in the past 4 years. A tiny overlap exists between our section “LM-oriented task instructions” and the ACL tutorial (Beltagy et al., 2022), which presented LLM techniques for NLP. But Beltagy et al. (2022) focused on various training techniques, such as self-training, meta-training, etc., our tutorial has a broader scope of instruction learning, in which prompt-based LLM is merely a sub-area.

Audience and Prerequisites Based on the level of interest in this topic, we expect around 150 participants. While no specific background knowledge is assumed of the audience, it would be best for the attendees to know about basic deep learning technologies, pre-trained language models (e.g., BERT). A **reading list** that could help provide background knowledge to the audience before attending this tutorial is given in Appendix A.1.

Breadth We estimate that at least 60% of the work covered in this tutorial is from researchers other than the instructors of the tutorial.

Diversity Considerations This tutorial will cover instruction learning for NLP as well as non-NLP problems, such as instruction-driven navigation, software control, etc. We will also cover content applying instruction supervision for individual tasks as well as cross-task generation. Our presenter team has a diverse background regarding geography and gender. Our team will promote our tutorial on social media to diversify our audience participation.

Material Access Online All the materials are openly available at www.wenpengyin.org/publications

4 Tutorial Instructors

The following are biographies of the speaker.

Wenpeng Yin is an Assistant Professor in the Department of Computer Science and Engineering at Penn State University. His research focuses on NLP with three sub-areas: (i) learning from task instructions; (ii) information extraction; (iii) NLP for education, bioinformatics, etc. Dr. Yin has presented the tutorial "Indirectly Supervised Natural Language Processing" at ACL'23, and tutorial "Learning from Task Instructions" at KONVENS'23. Additional information is available at www.wenpengyin.org.

Qinyuan Ye is a fifth-year Ph.D. student at the University of Southern California, advised by Prof. Xiang Ren. Her research interest lies in natural language processing. In particular she is interested in approaches that reduce human annotation efforts, including methods leveraging distant supervision, high-level human supervision (e.g., explanations, instructions), and meta-learning. Additional information is available at yeqy.xyz.

Pengfei Liu is an associate professor at Shanghai Jiaotong University and leads the Generative Artificial Intelligence Research Lab (GAIR). His research topics currently focus on information extraction, text generation, language pre-training, and NLP system evaluation. He won the Best Demo Paper award in ACL 2021 and the Outstanding Demo Paper award in ACL 2022. Homepage: http://pfliu.com.

Xiang Ren is an Associate Professor in Computer Science and the Andrew and Erna Viterbi Early Career Chair at USC. Ren's research seeks to build generalizable NLP systems that can handle a wide variety of language tasks and situations. He works on new algorithms and datasets to make NLP systems cheaper to develop and maintain, arm machine models with common sense, and improve model's transparency and reliability to build user trust. His research work has received several best paper awards in top NLP and AI conference venues. Ren has been awarded an NSF CAREER Award, multiple faculty research awards from Google, Facebook, Amazon, JP Morgan and Sony, and the 2018 ACM SIGKDD Doctoral Dissertation Award. He was named Forbes' Asia 30 Under 30 in 2019. Ren has presented a number of tutorials, such as Knowledge-Augmented Methods for Natural Language Processing at ACL 2022, Scalable Construction and Reasoning of Massive Knowledge Bases at NAACL 2018, and other related tutorials at WWW'18, CIKM'17, etc. Homepage: https://shanzhenren.github.io.

Hinrich Schütze is Chair of Computational Linguistics and co-director of the Center of Information and Language Processing at Ludwig-Maximilians-Universität München (LMU Munich), Germany. He was the President of the Association for Computational Linguistics in 2020, and General Chair of ACL 2013. In 2022, Prof. Schütze was elected as ACL Fellow. Prior to joining LMU Munich, he was a Professor of Theoretical Computational Linguistics at the University of Stuttgart. Hinrich holds a Ph.D. in computational linguistics from Stanford University. Additional information is available at https://schuetze.cis.lmu.de.

Ethical Considerations

We do not anticipate any ethical issues particularly to the topics of the tutorial.

References

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

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

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37, Dublin, Ireland. Association for Computational Linguistics.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

S. R. K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of ACL*, pages 82–90.

S. R. K. Branavan, David Silver, and Regina Barzilay. 2011. Learning to win by reading manuals in a monte-carlo framework. In *Proceedings of ACL*, pages 268–277.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *CoRR*, abs/2010.11982.

Dan Goldwasser and Dan Roth. 2011. Learning from natural instructions. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1794–1800. IJCAI/AAAI.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *CoRR*, abs/2212.04037.

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. 2023. Robustness of learning from task instructions. In *Findings of ACL*, pages 13935–13948.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *ICLR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *ICLR*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of ACL*, pages 14409–14428.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of ACL*, pages 716–723.

Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi, and Matthew Peters. 2022. Hint: Hypernetwork instruction tuning for efficient zero-shot generalisation. *arXiv preprint arXiv:2212.10315*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly understand prompts? A case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop, 03 December 2022, New Orleans, Louisiana, USA*, volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR.

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

Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to gptk's language. In *Findings of ACL*, pages 589–612.

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

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. Longform: Optimizing instruction tuning for long text generation with corpus extraction. *CoRR*, abs/2304.08460.

Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *The AAAI workshop on supervisory control of learning and adaptive systems*, pages 30–35.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:607–622.

Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2023. Do you really follow me? adversarial instructions for evaluating the robustness of large language models. *CoRR*, abs/2308.10819.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2023. MUFFIN: Curating multi-faceted instructions for improving instruction following.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of ACL*, pages 3470–3487.

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

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

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of EMNLP*, pages 1199–1212.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of NAACL*, pages 2439–2455.

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

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of EACL*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of EMNLP*, pages 390–402.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts - A real-world perspective. *Trans. Assoc. Comput. Linguistics*, 10:716–731.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *CoRR*, abs/2306.17194.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. In *NeurIPS*.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *Proceedings of ICML*, volume 202, pages 35413–35425.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of ACL*, pages 13484–13508.

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

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020. Learning from explanations with neural execution tree. In *Proceedings of ICLR*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of NAACL*, pages 2300–2344.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip S. Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of NAACL-HLT*, pages 1351–1360.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. In *Proceedings of CoNLL*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023b. Multi-instruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of ACL*, pages 11445–11465.

Qinyuan Ye and Xiang Ren. 2021. Learning to generate task-specific adapters from task description. In *Proceedings of ACL/IJCNLP (Volume 2: Short Papers)*, pages 646–653.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *CoRR*, abs/2305.14327.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of EMNLP-IJCNLP*, pages 3912–3921.

Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. ConTinTin: Continual learning from task instructions. In *Proceedings of ACL*, pages 3062–3072.

Yichi Zhang and Joyce Chai. 2021. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *Findings of ACL/IJCNLP*, pages 4202–4213.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of ICML*, volume 139, pages 12697–12706.

A Appendix

A.1 Recommended Paper List

The following is a reading list that could help provide background knowledge to the audience before attending this tutorial:

- Learning from Natural Instructions (Goldwasser and Roth, 2011)
- Learning from Explanations with Neural Execution Tree (Wang et al., 2020)
- Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach (Yin et al., 2019)
- Textual Entailment for Event Argument Extraction: Zero- and Few-Shot with Multi-Source Learning (Sainz et al., 2022)
- Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (Liu et al., 2021)
- True Few-Shot Learning With Prompts—A Real-World Perspective (Schick and Schütze, 2022)
- The Turking Test: Can Language Models Understand Instructions? (Efrat and Levy, 2020)
- Hierarchical Task Learning from Language Instructions with Unified Transformers and Self-Monitoring (Zhang and Chai, 2021)
- Cross-Task Generalization via Natural Language Crowdsourcing Instructions (Mishra et al., 2022)
- MUFFIN: Curating Multi-Faceted Instructions for Improving Instruction Following (Lou et al., 2023)