LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, Maram Hasanain
Qatar Computing Research Institute, HBKU, Doha, Qatar
{fialam, shchowdhury, sboughorbel, mhasanain}@hbku.edu.qa

Abstract

The recent breakthroughs in Artificial Intelligence (AI) can be attributed to the remarkable performance of Large Language Models (LLMs) across a spectrum of research areas (e.g., machine translation, question-answering, automatic speech recognition, text-to-speech generation) and application domains (e.g., business, law, healthcare, education, and psychology). The success of these LLMs largely depends on specific training techniques, most notably instruction tuning, RLHF, and subsequent prompting to achieve the desired output. As the development of such LLMs continues to increase in both closed and open settings, evaluation has become crucial for understanding their generalization capabilities across different tasks, modalities, languages, and dialects. This evaluation process is tightly coupled with prompting, which plays a key role in obtaining better outputs. There has been attempts to evaluate such models focusing on diverse tasks, languages, and dialects, which suggests that the capabilities of LLMs are still limited to medium-to-low-resource languages due to the lack of representative datasets. The tutorial offers an overview of this emerging research area. We explore the capabilities of LLMs in terms of their performance, zero- and few-shot settings, fine-tuning, instructions tuning, and close vs. open models with a special emphasis on low-resource settings. In addition to LLMs for standard NLP tasks, we will focus on speech and multimodality.[1]

1 Tutorial Content Description

Large Language Models (LLMs) are prominent examples of Foundation Models (FMs), based on the Transformer network architecture (Vaswani et al., 2017). Trained to predict the subsequent token in a sequence, LLMs capture implicit and intricate information contained in the data. Moreover, when created using multilingual training data, the models capture linguistic nuances, phonological patterns, and semantic relationships across languages, strengthening its multilingual capabilities. However, understanding how their capabilities generalize across tasks and languages requires a systematic evaluation approach.

1.1 Benchmarking LLMs for different tasks and languages

The HELM project (Liang et al., 2022) assessed English LLMs across various metrics and scenarios. BIG-Bench (Srivastava et al., 2022) introduced a large-scale evaluation with 214 tasks, considering low-resource languages as well. Other efforts included evaluations of ChatGPT, GPT2.5, BLOOMZ, and OpenAI GPT as in Bang et al. (2023); Ahuja et al. (2023); Hendy et al. (2023); Abdelali et al. (2023); Scao et al. (2022).

For speech, OpenAI's Whisper (Radford et al., 2022), Google's USM (Zhang et al., 2023), and other speech models are explored by the speech community. They are general-purpose speech models with multilingual capabilities, designed for speech recognition (ASR) and other tasks. The benchmarking efforts include Speech Processing Universal PERformance Benchmark (SUPERB) initiative (Yang et al., 2021) which includes a collection of benchmarking tools, resources, and a leader board for 10 tasks from six domains.

1.2 LLMs and lower-resources languages

These LLMs have been trained on datasets from the internet, ingesting many resources in different languages. For close models (e.g., ChatGPT) the coverage and the distribution of the content for medium-to-low-resource languages are unknown. Most of the open-sourced models uses commoncrawl dataset, which is skewed for many languages. For example, Bloom, that is trained on 46 natural

[1]The content of the tutorial will be available at the following website: https://llm-low-resource-lang.github.io/.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Tutorial Abstracts, pages 27–33
March 21, 2024 ©2024 Association for Computational Linguistics

languages and 13 programming languages [2], has only 4.6%, 0.02% and 0.70% language coverage for Arabic, Swahili and Hindi respectively (Scao et al., 2022).

With models trained on such distribution of data, this raises questions on their capabilities on medium-to-low-resource languages in a variety of language processing tasks. To understand the capabilities of LLMs, there has been several research efforts. Bang et al. (2023) reports that ChatGPT fails to generalize to low and extremely low resources languages (e.g., Marathi, Sundanese, and Buginese). Lai et al. (2023) reports that ChatGPT generally performs better for English than other languages. Ahuja et al. (2023) evaluate 8 different tasks with 33 languages and report that LLMs perform better on high-resource languages and languages that are in Latin scripts. In our work for Arabic, we evaluate ChatGPT on 33 tasks, 59 datasets with 96 test setups using zero-shot setting. Performances are significantly lower on 88 test setups (Abdelali et al., 2023). This study also focused on tasks covering different Arabic dialects and reports that models perform comparably for MSA than other dialects such as Egyptian, Gulf, Levantine, and Maghrebi.

In the realm of speech technology, OpenAI's recent Whisper model has demonstrated that the performance in low-resource languages is still relatively poor, a trend that correlates with the size of the pre-training dataset. Subsequently, Google's USM models have shown further improvements in performance, achieving an average word error rate (WER) of less than 30% across 73 languages.

1.3 Multimodality

Along side with NLP, speech, and multimodal generative models have also emerged (Liu et al., 2023a; Zhu et al., 2023a; OpenAI, 2023a). ChatGPT has demonstrated multi-modal abilities on variety of tasks. Following that, Zhu et al. (2023a) developed MiniGPT-4, which is trained by combining Vicuna (Chiang et al., 2023) and Blip-2 (Li et al., 2023). Recently, OpenAI, Google, and Meta released GPT-4 Vision (OpenAI, 2023b), Gemini (Team et al., 2023), and AnyMAL (Moon et al., 2023), respectively, each focusing on multimodal aspects. The idea of the these attempts was to train a model by aligning visual information from a pre-trained vision encoder with an LLM. Though their capabilities have not been widely studied across tasks and languages, it is important to explore and understand their capabilities that can enhance future studies.

1.4 Dialects

In our study for Arabic (Abdelali et al., 2023), we observed that the gaps in LLMs' performance between MSA and dialectal datasets (e.g., for machine translation (MT) and speech recognition task) are more pronounced, indicating ineffectiveness of LLMs for under-represented dialects. For example, in both the GPT-models, we noticed a large discrepancy in the POS accuracy of 0.810 versus 0.379 on MSA and dialects respectively. Similarly, for Arabic dialect identification tasks (ADI) we notice a significant difference between the SOTA acoustic and lexical model with respect to LLMs results.

1.5 Prompting for LLMs

Prompt design plays a critical role in influencing the performance of Large Language Models (LLMs), as evidenced in (Reynolds and McDonell, 2021; Dong et al., 2022). These models are highly sensitive to minor variations in the prompts, such as word choice and the order of examples in few-shot settings. Ahuja et al. (2023) have investigated various monolingual and multilingual prompts, discovering that English-language templates generally outperform those in native languages. The performance of a task also depends on native and non-native language prompts. In our study focusing on Arabic (Abdelali et al., 2023) and Bangla (Hasan et al., 2023), we have found that performance can vary considerably depending on whether the prompts are in a native or non-native language. This variability is observed in both zero-shot and few-shot settings. Another point of interest in few-shot settings is the method used for selecting shots and arranging them in a reasonable order. Various approaches have been reported, such as random selection (Khondaker et al., 2023), class-based selection (e.g., Liang et al. (2022) selected examples to ensure class coverage in classification tasks), and Maximal Marginal Relevance-based (MMR) selection (Carbonell and Goldstein, 1998).

1.6 What this tutorial offers

Here, we provide an overview of the capabilities of LLMs for diverse tasks, languages, dialects, and modalities, including text, speech, and multimodality. We start with an introduction to LLMs, includ-

[2] https://huggingface.co/bigscience/bloom

ing a brief history and their significant capabilities in downstream tasks. This is followed by an in-depth examination of various LLMs developed for NLP, speech, and multimodal applications, emphasizing their utility across different tasks.

In the third part of the tutorial, we delve into the intricacies of prompting, which serves as a foundational element for obtaining output from these LLMs. In this part, we will also include a hands-on demonstration of tools that have been developed to further facilitate research on LLMs. The fourth part of the tutorial will focus on a more comprehensive discussion about low-resource languages, addressing both the challenges they present and future directions for research. Finally, we will discuss hallucination, bias, toxicity, and computational resources needed for model training and inference. An outline of the tutorial is reported in Section 3.

2 Type of the Tutorial

The tutorial is both introductory, covering a number of topics related to the capabilities of LLMs, but it is also cutting-edge, covering some latest developments in these areas. Attendees will have an overview of tasks, languages, dialects and modalities related to LLMs, which will put them up to speed to do research in the area. The tutorial targets anyone interested in employing LLMs for NLP, speech and multimodal tasks. We believe researchers working on lower-resource languages will be especially interested. We expect the audience to have intermediate machine learning knowledge.

3 Outline of the Tutorial

Below, we offer an outline of the tutorial. More information and materials will is available online on the tutorial website upon the tutorial acceptance.

3.1 Introduction [30 min]

(i) LLMs

  (a) A brief history of LLMs
  (b) Capabilities in downstream NLP, speech, and multimodal tasks

References: (Mielke et al., 2021; Sennrich et al., 2016; Wu et al., 2016; Kudo and Richardson, 2018; Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020)

3.2 Models and their capabilities for low-resource languages [30 min]

The following are just a few examples of models. They will not be the only ones covered in the tutorial.

(i) Models for NLP tasks
  (a) GPT 3.5 (ChatGPT), GPT-4
  (b) Bloom, LLaMA, mT5, Flan, PaLM

(ii) Models for Speech tasks
  (a) USM
  (b) Whisper

(iii) Models for Multimodality
  (a) Closed models: GPT-4 Vision, Gemini
  (b) Open Models: MiniGPT, LLaVA

References: (Brown et al., 2020; Liu et al., 2023a; Xue et al., 2020; Scao et al., 2022; Touvron et al., 2023; Zhu et al., 2023a)

3.3 Prompt Engineering [50 min]

(i) Zero-shot

(ii) Few-shots and selection methods

(iii) Prompt templates

(iv) Mono/Cross lingual prompting

(v) Prompt programming

(vi) Tools and resources (e.g., LLMeBench (Dalvi et al., 2023), OpenICL (Wu et al., 2023), PromptBench (Zhu et al., 2023b)) and lm-evaluation-harness (Gao et al., 2023).

References: (Wei et al., 2021; Zhang et al., 2022; Reynolds and McDonell, 2021)

3.4 Limitations and Challenges for low-resource settings [50 min]

(i) Multitask, multilingual, multimodal evaluation for low-resource languages

(ii) Multi-dialects challenges

(iii) Summary of recent benchmarking efforts

References: (Ahuja et al., 2023; Liang et al., 2022; Srivastava et al., 2022; Bang et al., 2023; Ahuja et al., 2023; Hendy et al., 2023; Yang et al., 2021; Radford et al., 2022; Zhang et al., 2023; Abdelali et al., 2023; Bang et al., 2023; Bubeck et al., 2023)

3.5 Other Related Aspects [30 min]

(i) Hallucination

(ii) Bias, Toxicity and Misinformation in LLMs

(iii) Computational Resources, Carbon footprint

References: (Bang et al., 2023)

4 Prerequisites

We expect attendees to be equipped with basic knowledge of machine learning, including familiarity with recent neural network architectures, particularly Transformers, and an understanding of pre-trained language models. Additionally, attendees should be familiar with standard NLP tasks such as text classification, natural language generation, and question answering.

5 Reading List

In addition to the references cited in Section 3, we recommend several surveys: an overview of LLMs (Zhao et al., 2023), prompt engineering (Liu et al., 2023b; Gu et al., 2023), in-context learning (Dong et al., 2022), and evaluation of LLMs (Liang et al., 2022).

6 Tutorial Instructors

Firoj Alam is a Scientist at the Qatar Computing Research Institute (QCRI), HBKU. He received his PhD from the University of Trento, Italy, and has been working for more than ten years in Artificial Intelligence, Deep/machine learning, Natural Language Processing, Social media content, Image Processing, and Conversation Analysis. His current research interest includes LLMs, fact-checking, multimodal propaganda detection in multiple languages. He previously presented tutorials at WWW-2022 and WSDM-2022 on the topic of "Fact-Checking, Fake News, Propaganda, And Media Bias". He was a co-organizer of different shared tasks CheckThat! 2020-2024 at CLEF, SemEval-2021 task 6 (propaganda detection in memes), SemEval-2024 task (multilingual detection of persuasion techniques in memes), WANLP (Arabic-NLP) shared task (2022-2023) and the NLP4IF-2021 shared task. He is also a co-organizer of the BLP-2023 workshop (collocated with EMNLP-2023).

Shammur Absar Chowdhury is a Scientist at QCRI, HBKU. Her research interest includes designing speech models, and interpretability for atypical phenomena in conversation. Dr. Chowdhury authored more than 60 peer-reviewed publications in tier-top conferences and journals; and actively contributed to the research community by organizing shared tasks, challenges, and workshops like SemEval-2022 (Task 3), QASR-TTS-v1.0 (ASRU2023), SLT2023 (Local Chair), summer workshop JSALT2022 (as a senior mentor) along with serving in the program-committee of top-tier conferences and special interest groups (SIGs).

Sabri Boughorbel is a Scientist at QCRI, HBKU. He received his PhD in Machine Learning from the university of Paris Sud. He has an extensive experience in Machine Learning for industrial and academic research. He authored more than 70 peer-reviewed papers and 7 patents. He was awarded several grants in the intersection of machine learning and health. His current research is on leveraging open-sourced LLMs for low-resource languages and developing multi-modal language models. He serves as PC member of top-tier machine learning conferences. In 2023, he co-organized a workshop on AI for Medicine.

Maram Hasanain is a PostDoctoral researcher at QCRI, HBKU. She received her PhD in Computer Science from Qatar University. Her current research interests are Arabic NLP, applied machine learning, and LLMs. Maram co-authored over 25 peer-reviewed publications in top-tier conferences and journals. She has been a co-organizer in the CheckThat! lab at CLEF 2019-2021, 2023 and 2024. She was also a co-organizer of the BroDyn'18 workshop on analysis of broad dynamic topics over social media co-located with ECIR'18.

7 Ethics Statement

Our tutorial is based on our own work in the area, related studies and public sources. Credit will be given wherever needed. Any biases are unintended.

Acknowledgments

The contributions of M. Hasanain were funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund (a member of Qatar Foundation).

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2023. Benchmarking arabic ai with large language models. *arXiv preprint arXiv:2305.14982*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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. *Advances in Neural Information Processing Systems*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Technical report, Microsoft Research.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2023. LLMeBench: a flexible framework for accelerating LLMs benchmarking. *arXiv 2308.04945*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation. *Zenodo*.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

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

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv:2304.08485*.

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

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 preprint arXiv:1907.11692*.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card. *OpenAI*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023. OpenICL: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 489–498.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023b. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.