11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 11: Multimodal Language Models and the Future

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Aditya Veerubhotla, Neehar Peri*

Edited by Paul Liang *Scribes: Aditya Rathod, Yihan Cao*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources. In week 11's discussion session, the class discussed GPT-like models - we define challenges in multi-modal learning that are on the rise and challenges that are less relevant due to GPT-like models, speculate on the future on large pre-trained GPT-like models in the next 5 years, and consider product ideas and research agenda's that can leverage GPT-like models. The following was a list of provided research probes:

- Think of multimodal research problems and technical challenges that are becoming more important and possibly enabled by GPT-like models. Out of these, which problems are academia particularly well suited to work on?
- Think of multimodal research problems and technical challenges that became less relevant or maybe even solved by GPT-like models.
- At a high-level (1-minute elevator pitch), describe one specific multimodal research project that you could embark on enabled by GPT-like models. Describe the key research questions, technical challenges, evaluation criteria, and broader impact.
- At a high-level (1-minute elevator pitch), describe one real-world 'product' idea enabled by GPT-like models. Prepare a 'sales pitch': current shortcomings, motivation, broad impact, potential technical challenges, any real-world deployment issues you could face, and evaluating success and impact. Who are the stakeholders who might use the product? How do you think this product will help them?
- What could the future of pretrained models look like? More modalities, more generative capabilities, more personalization, more efficiency, or what else? Which fundamental multimodal technical challenges will arise as more multimodal pretrained models are created?
- How can we, in an academic environment, do impactful research in multimodal given the success of these pretrained models? What would be your proposed 10-year research agenda in multimodal ML, assuming you had access to funding and researchers?

As background students read the following papers:

- (Required) Bubeck et al. [2023]. Most in-depth evaluation of GPT4 so far. As you read the paper, think about both its current capabilities but also try to extrapolate- what could the future bring, and how are we best prepared to take advantages of its future capabilities?
- (Required) Eloundou et al. [2023]. The potential social impact of GPT-like models on society and the workforce.
- (Optional) OpenAI [2023]. Original GPT4's report from OpenAI.
- (Optional) Haase and Hanel [2023]. How can generative AI help human in creative tasks? As you read this paper, think about how other similar interactive user studies can be designed to evaluate pretrained models for various qualities.
- (Optional) Piantadosi [2023]. How pretrained models can influence our fundamental study of linguistics. When you read this paper, think about how pretrained models can contribute to our understanding of

fundamental science, not just downstream applications.
- (Optional) Marcus [2023]. Discussion of GPT successes and failures from a famous LLM pessimist.
- (Optional) Neumann et al. [2023]. The impact of GPT-like models on education.
- (Relevant) Ye et al. [2023]. More evaluation of language models.
- (Relevant) Harsha Nori [2023]. More evaluations of GPT-like models on medical tasks.
- (Relevant) Sanderson [2023]. Perception of language models by scientists.
- (Relevant) Augustin Lecler [2023]. Potential impact of GPT-like models on medical tasks.

We summarize several main takeaway messages from group discussions below:

1 Multimodal Challenges from GPT-like Models

Increasingly important challenges
Interpretability
Bias Mitigation
Compositional Reasoning
Logical Reasoning
Counterfactual Reasoning
Interaction and Embodiment
Training Efficiency with Additional Modalities
Capturing Subtle Cues
Co-Learning on Small Scale Data
Fine-Tuning on Small Scale Data
Embodiment
Planning

Table 1: Although multi-modal LLMs are able to produce many reasonable outputs and can understand human intent to solve well defined tasks, we posit that these models must still improve in their reasoning capabilities. In particular, auto-regressive models may not be well suited for reasoning tasks because the model does not know how the sentence will end when predicting the next token.

Recent advances in GPT-like models show remarkable results on tasks like diverse generation, text summarization, information extraction, and text-image captioning. Moreover, it is becoming increasingly easier to interact with these models, through prompting or chatting with them. However, these models still struggle with compositional, logical, and counterfactual reasoning. We posit that emergent reasoning is unlikely to result from current multimodal LLMs and will need a different approach.

Do we need a different approach of reasoning?

Existing multi-modal LLMs already train on a large percentage of available data on the internet, making it difficult to trivially scale up the amount of training data. This suggests that reasoning cannot emerge by simply training on *more* data. Further, the training objective of current LLMs may not be well suited for reasoning because the model does not know how a sentence will end when predicting the next token. Current attempts at eliciting reasoning from LLMs leverage chain-of-thought prompting. However, we argue that chain-of-thought prompting isn't truly zero-shot, but rather a one-shot prompt which allows the model to pattern match to the input. Existing chain-go-thought prompting methods are brittle, and need considerable effort to elicit the expected response from the model, further suggesting that current multi-modal models are not well suited for human-like reasoning.

Moreover, it is unclear how to empirically define reasoning in the context of large multi-modal LLMs, particularly when LLMs are capable of reproducing the training data. Determining the right evaluation protocol for evaluating reasoning is incredibly challenging, and is a key question that remains unanswered. One potential protocol may be to ask LLMs to explain a novel concept and answer followup questions, similar

to how children can demonstrate understanding by "showing their work".

What will large pre-trained models look like in 5 years?

The next-generation of large pre-trained models will undoubtedly leverage more multi-modal data, particularly audio, text, and videos from YouTube. YouTube remains an untapped source of nearly infinite unconstrained data that provides paired multi-modal data for free. In addition, ego-centric videos Grauman et al. [2022] will likely play an important role in fine-tuning large pre-trained models for robotics tasks. Lastly, future pre-trained models will embrace continually learning to adapt based on user queries, feedback, and new information created on the internet. This will ensure that pre-trained models continue to update as new information is created.

2 Potential Research Agenda with GPT-like Models

- Improved Benchmarking. Given that it is difficult to quantify reasoning in GPT-like models, it is important that we invest in multi-modal benchmarks for vision-language models that can attempt to address these challenges. An improvement in this skill would have wide ranging applications, such as in embodiment.
- Integrating GPT-like with embodied agents. By combining LLMs with embodied agents, we can achieve more accurate language understanding and instruction following, as well using the world knowledge in these models to improve performance.
- Using LLMs to Generate Training Data. As LLMs generate more realistic multi-modal data, we can self-train [wang2022self] to improve larger models or distill training data to improve smaller models.
- Identifying Multi-Modal DeepFakes. As GPT-like models are released for public use, managing and mitigating the damage of generated content requires being able to identify multimodal deep fakes.

3 Potential Product Ideas with GPT-like Models

- Game Generation. GPT-like models are able to generate 3D assets from text data. and can generate RPG games from a prompt. We can leverage GPT-like models to enable more randomness is game asset, story-lines, and scene generation.
- Help Desk Support. We can use LLMs to quickly address frequently asked questions **?** and categorize nuanced questions to be addressed by specific teams, making it easier to deal with a large volume of help requests.
- Sound Track Generation. We can use GPT-like models to synthesize sound tracks conditioned on a movie scene.
- Teaching Assistant. We can use GPT-like models to create customized lesson plans and explain specific concepts in greater detail, acting like a personalized tutor.

4 Role of Academia and Industry

The development of large multimodal language models like GPT-4 is a collaborative effort between industry and academia. Industry invests heavily in the computing infrastructure and resources needed to train and fine-tune these models. This allows them to leverage the models for commercial applications such as chatbots, virtual assistants, and other NLP tasks that improve efficiency and productivity. Industry also benefits from cutting-edge research in areas such as transfer learning and reinforcement learning, which can be applied to improve the performance of these models. However, there are limitations to the industry's approach, and this is where academia plays a critical role. Academia can provide a more future-focused research approach, and address the limitations of these models, such as the lack of high-quality data for certain domains, and propose new research directions to overcome these challenges.

References

Philippe Soyer Augustin Lecler, Loic Duron. Revolutionizing radiology with gpt-based models: Current applications, future possibilities and limitations of chatgpt. 2023.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

Jennifer Haase and Paul HP Hanel. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *arXiv preprint arXiv:2303.12003*, 2023.

Scott Mayer McKinney Dean Carignan Eric Horvitz Harsha Nori, Nicholas King. Capabilitis of gpt-4 on medical challenge problems. 2023.

Gary Marcus. Gpt-4's successes, and gpt-4's failures. 2023.

Michael Neumann, Maria Rauschenberger, and Eva-Maria Schön. "we need to talk about chatgpt": The future of ai and higher education. 2023.

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

Steven Piantadosi. Modern language models refute chomsky's approach to language. 2023.

Katherine Sanderson. Gpt-4 is here: What scientists think. 2023.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.