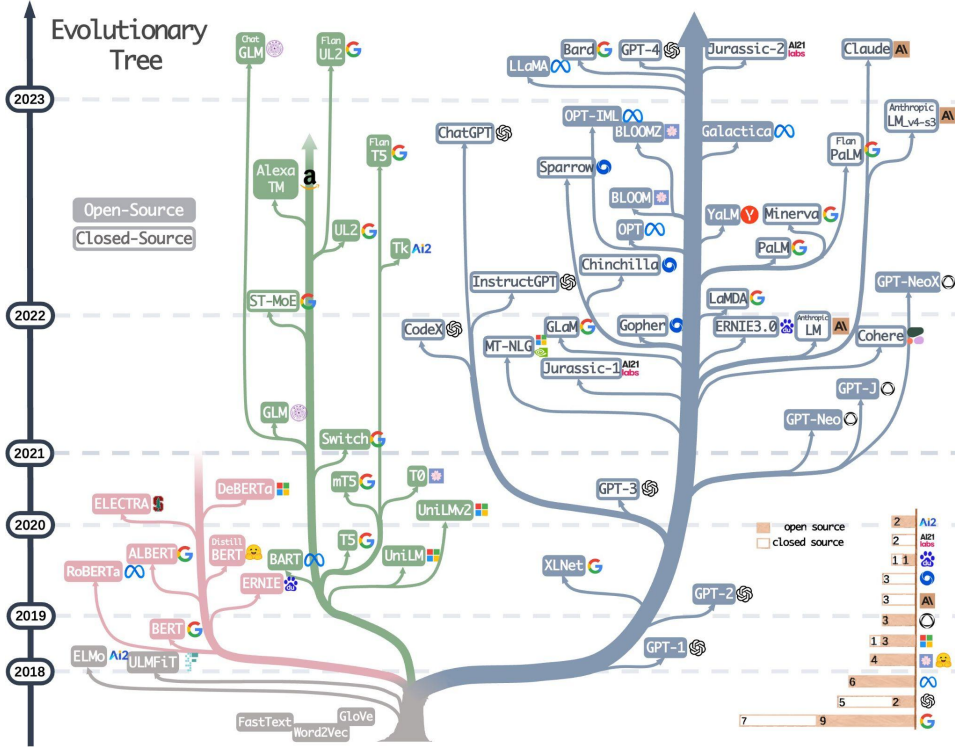


LLMs: Practices, Paradigm Shifts, Remaining Challenges

Jingfeng Yang
Applied Scientist, Amazon

Evolution of LLMs



Yang J, Jin H, Tang R, et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond[J]. arXiv preprint arXiv:2304.13712, 2023.

Harnessing LLMs in Practice

Practical Guides – Data

- LLMs generalize better than fine-tuned models in downstream tasks facing out-of-distribution data, such as adversarial examples and domain shifts.
- LLMs are preferable to fine-tuned models when working with limited annotated data, and both can be reasonable choices when abundant annotated data is available, depending on specific task requirements.
- It's advisable to choose models pre-trained on fields of data that are similar to downstream tasks.

Practical Guides – Traditional NLU Tasks

- Fine-tuned models generally are a better choice than LLMs in traditional NLU tasks, but LLMs can provide help while requiring strong out-of-distribution generalization ability.

Practical Guides – Generation Tasks

- Due to their strong generation ability and creativity, LLMs show superiority at most generation tasks, including both code and text generation.

Practical Guides – Knowledge-intensive tasks

- LLMs excel at knowledge-intensive tasks due to their massive real-world knowledge.
- LLMs struggle when the knowledge requirements do not match their learned knowledge, or when they face tasks that only require contextual knowledge, in which case fine-tuned models can work as well as LLMs.

Practical Guides – Abilities Regarding Scaling

- With the exponential increase of model scales, LLMs become especially capable of reasoning like arithmetic reasoning and commonsense reasoning.
- Emergent abilities become serendipity for uses that arise as LLMs scale up, such as ability in word manipulation and logical ability.
- In many cases, performance does not steadily improve with scaling due to the limited understanding of how large language models' abilities change as they scale up.

Practical Guides – Miscellaneous Tasks

- Fine-tuned models or specified models still have their space in tasks that are far from LLMs' pretraining objectives and data.
- LLMs are excellent at mimicking human, data annotation and generation. They can also be used for quality evaluation in NLP tasks and have bonuses like interpretability

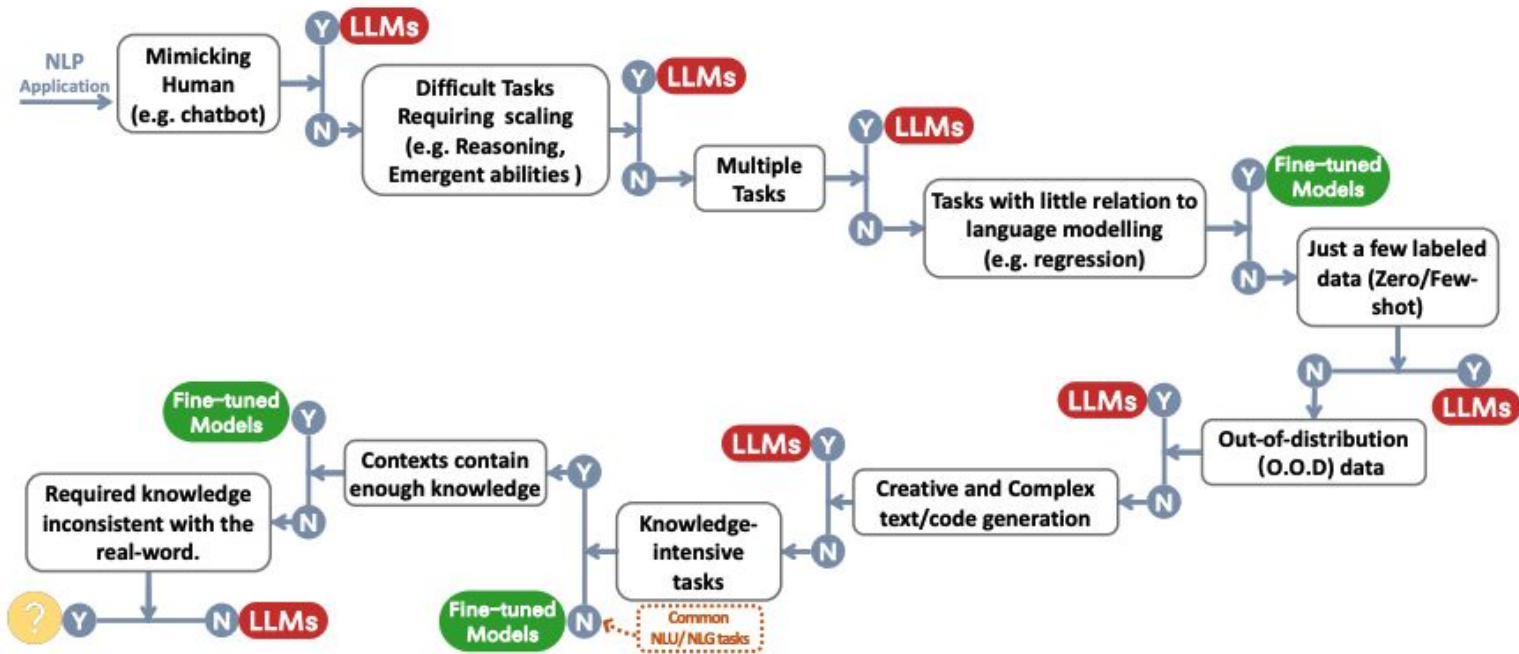
Practical Guides – Real World "Tasks"

- LLMs are better suited to handle real-world scenarios compared to fine-tuned models. However, evaluating the effectiveness of models in the real world is still an open problem.
- This should be due to two factors:
 - During the pretraining stage, LLMs have seen web-scale real world data, giving it potential to mimic real-world distribution.
 - Instruction Tuning and Alignment data contains diverse real-world user instructions, which makes it handle real-world user inputs well.

Practical Guides – Other Considerations

- Light, local, fine-tuned models should be considered rather than LLMs, especially for those who are sensitive to the cost or have strict latency requirements. Parameter-Efficient tuning can be a viable option for model deployment and delivery.
- The zero-shot approach of LLMs prohibits the learning of shortcuts from task-specific datasets, which is prevalent in fine-tuned models. Nevertheless, LLMs still demonstrate a degree of shortcut learning issues.
- Safety concerns associated with LLMs should be given utmost importance as the potentially harmful or biased outputs, and hallucinations from LLMs can result in severe consequences. Some methods such as human feedback have shown promise in mitigating these problems.

Decision of Using LLMs or Finetuning a Smaller Model



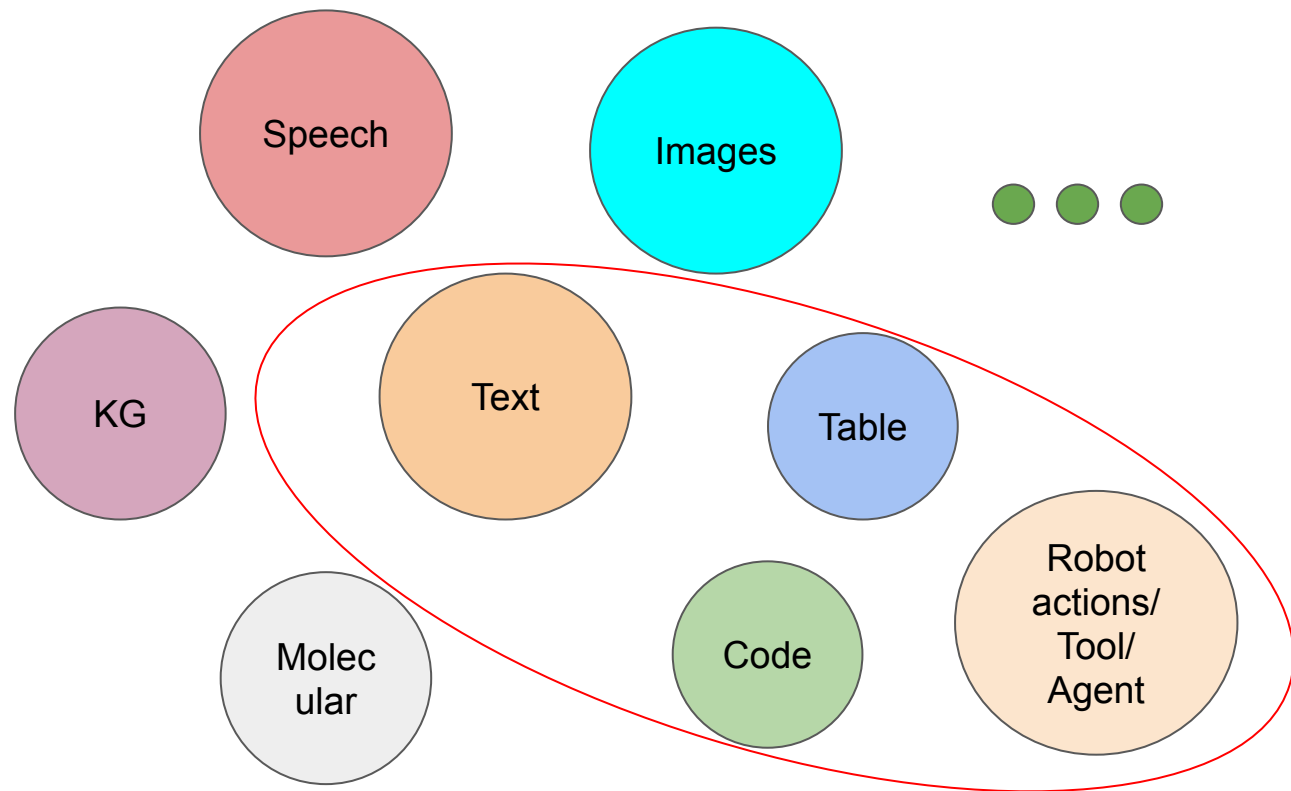
Paradigm Shifts and Remaining Challenges Towards AGI

Remaining Challenges Towards AGI

- Multimodality and Embodied AI
 - Using intermediate abstractions for grounding.
 - Direct modeling: Inductive biases v.s. scaling of data and model size?
- Planning and Reasoning
 - Pre-LLM Era: Neural-symbolic models, multi-stage and modular models, etc.
 - LLM Era: Scaling + CoT, interface generation, what else?
- Human-centered AGI
 - Alignment
 - AI safety

Multimodality and Embodied AI

The Multimodality World



Two Approaches to Multimodality AGI

- End2end Modeling
 - Table-text encoding / decoding
 - Visual-language encoding / decoding
 - Text-code encoding / decoding
- Using abstractions to bridge LLM and other modalities
 - Long-standing goal of Semantic Parsing
 - Transforming Natural Language to Formal Language (e.g. SQL to be executed on tables)
 - Using LLM to generate functions and APIs, and then execute them (e.g. Binder, ToolFormer, ChatGPT Plugins)
 - Robots relying on low-level policy or planner that can translate LM decisions into low-level actions (e.g. PaLM-E)

Intermediate abstractions as inductive biases still play an important role to bridge LLMs and some modalities

LLM Era: Conclusion

- Effect of architectural inductive biases is decreasing after scaling.
- However, some inductive biases could encourage “early emergence or emergent abilities at a much smaller scale than purely scale-induced emergence.”
- Architectural Inductive biases -> prompting as inductive biases

Planning and Reasoning

Reasoning in LLM Era: Conclusion

- Scaling + CoT (Advanced Prompting Techniques to generate reasoning paths)
- Interface/function generation and reasoning execution (Binder, ToolFormer etc.)
- Personally, I still think there should be some fundamental model changes to reach 100% reasoning accuracy with one model, although traditional reasoning schema could not match the performance of Scaling + CoT.

Human-centered AGI

ChatBots: Real-world Alignment (Alignment Tax and Tradeoff)




- Alignment could be harmful to in-context-learning ability without specific tricks.
- Tradeoff between helpfulness and harmless.

ChatGPT

Claude

Anthropic's safety-first LM API is sometimes too safe to be useful

Bard

Rank	Model	Elo Rating	Description	License
1	 GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	 Claude-v1	1224	Claude by Anthropic	Proprietary
3	 GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Weights available; Non-commercial
5	Koala-13B	1022	a dialogue model for academic research by BAIR	Weights available; Non-commercial
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance	Apache 2.0
7	Oasst-Pythia-12B	928	an Open Assistant for everyone by LAION	Apache 2.0
8	ChatGLM-6B	918	an open bilingual dialogue language model by Tsinghua University	Weights available; Non-commercial
9	StableLM-Tuned-Alpha-7B	906	Stability AI language models	CC-BY-NC-SA-4.0
10	Alpaca-13B	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford	Weights available; Non-commercial
11	FastChat-T5-3B	902	a chat assistant fine-tuned from FLAN-T5 by LMSYS	Apache 2.0
12	Dolly-v2-12B	863	an instruction-tuned open large language model by Databricks	MIT

Alpaca

Vicuna

Koala

OpenAssistant

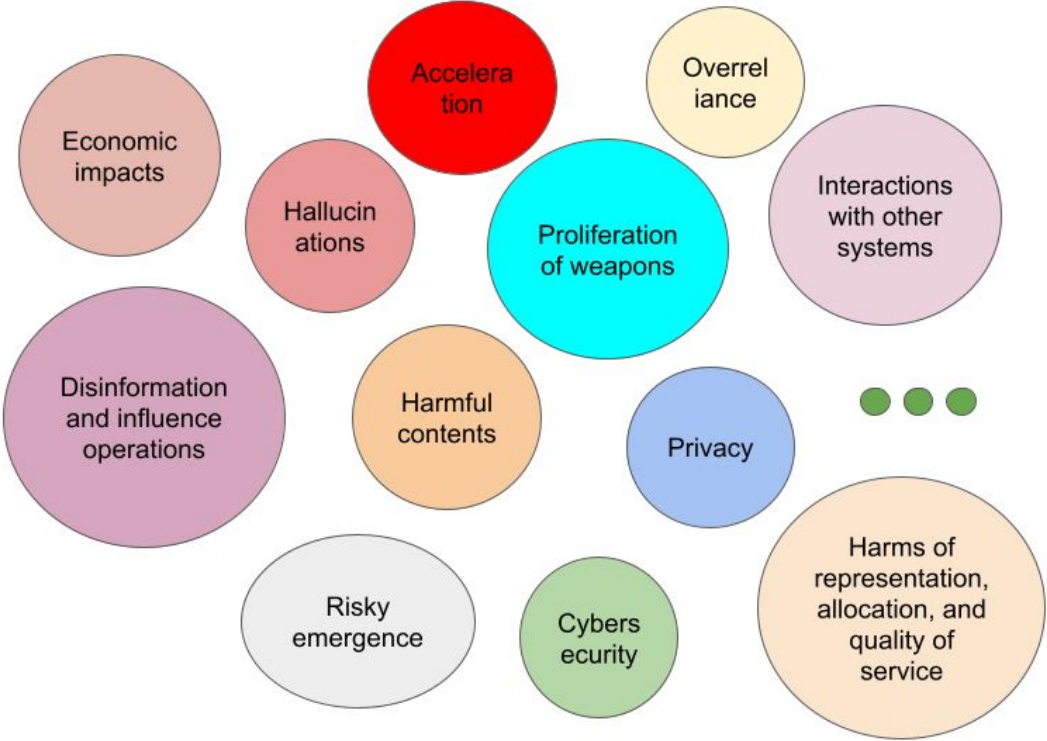
Dolly

<https://lmsys.org/>

Safety Alignment

- Model Mitigations:
 - Pretraining data filtering
 - RLHF or RLAIF
 - Constitutional AI: AI criticism and revision, Learning from AI feedback
 - GPT-4: Rule-based reward model (RBRM)
- Evaluation: Expert Red Teaming, Classifier for automatic quantitative evaluation
- Usage Policy and Monitoring
- Moderation classifier

Existing and Potential AI Safety Issues



Thanks!