A Survey of Large Language Models

Seminar: Diachronic Language Models (6 cp)

Ke Ren

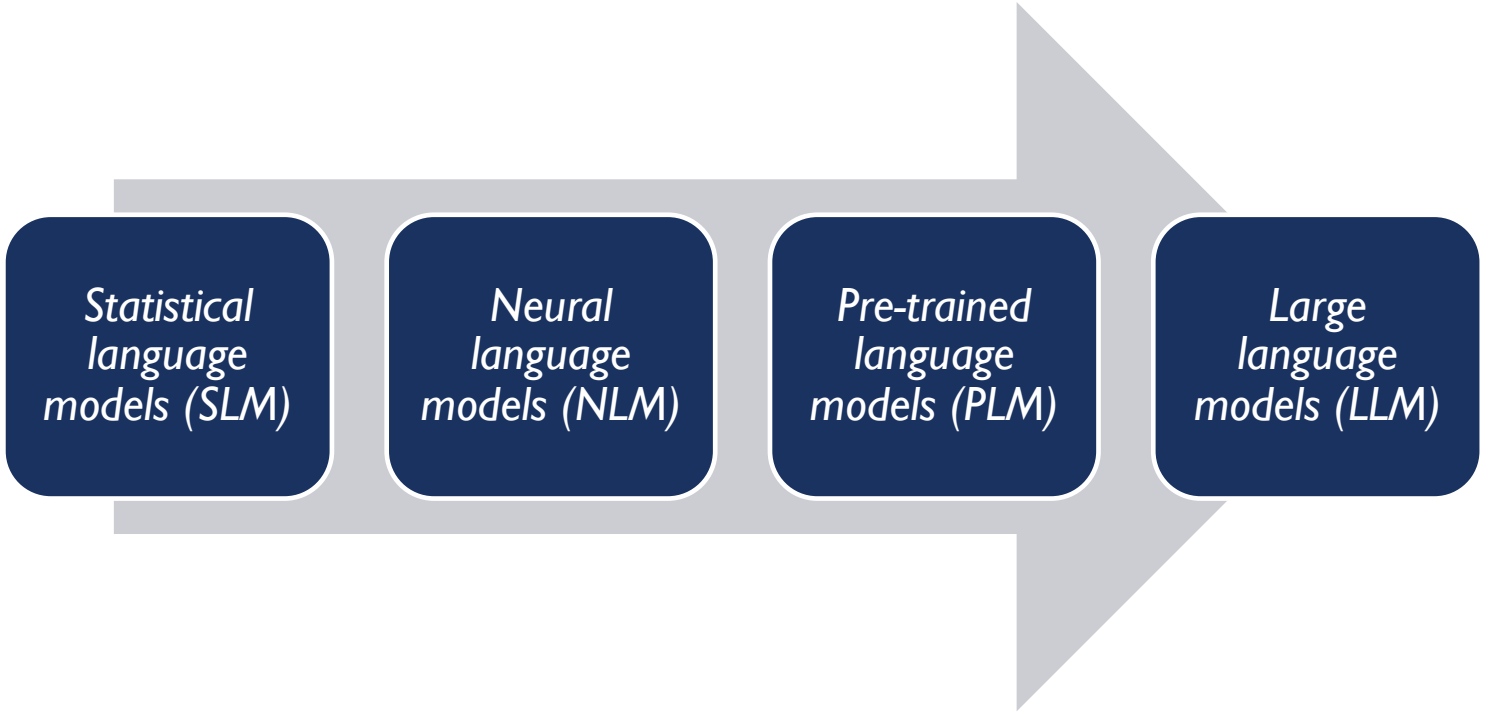
Data and Computer Science

ke.ren@stud.uni-heidelberg.de

Nov 21. 2023

Introduction

Four Major Development Stages for Language Model



Differences between PLM and LLM

- Surprising emergent abilities
- Through the prompting interface
- The unclear distinction between research and engineering

Background

Scaling law

1. *KM scaling law:* Favor a larger budget allocation in model size than the data size

$$
\begin{aligned}
L(N) &= \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13} \\
L(D) &= \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13} \\
L(C) &= \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}
\end{aligned}
$$

N: model size
D: dataset size
C: the amount of training compute
L(·) denotes the cross entropy loss in nats
Given a compute budget c

2. *Chinchilla scaling law:* A more balanced approach, advocating for an equal allocation of computational resources between model size and data size.

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

E=1.69,A=406.4,B=410.7 , α=0.34 and β=0.28

$$N_{opt}(C) = G\left(\frac{C}{6}\right)^{a}, \quad D_{opt}(C) = G^{-1}\left(\frac{C}{6}\right)^{b},$$

Background

Emergent Abilities: The abilities that are not present in small models but arise in large models

1. ***In-context learning***
   - Provide with a natural language instruction, generate the expected output for the test instances by completing the word sequence of input text
   - Introduced by GPT-3 (175B)
   - Depends on the model's scale and the specific downstream task
2. ***Instruction following***
   - By fine-tuning with a mixture of multi-task datasets formatted via natural language descriptions, LLMs perform well on unseen tasks that are also described in the form of instructions.
   - LaMDA-PT : Model size reached 68B
3. ***Step-by-step reasoning***
   - Difficult to solve complex tasks that involve multiple reasoning steps *e.g.,* mathematical word problems.
   - Effective for complex tasks using **chain-of-thought** prompting, demonstrated by PaLM and LaMDA (model size larger than 60B)

Background

Key Techniques

- ***Scaling***
  - Larger model/data sizes and more training compute typically lead to an improved model capacity
  - Two representative models: GPT-3 and PaLM.
- ***Training***
  - A challenge to successfully train a capable LLM
  - Solution: Implementation and deployment of parallel algorithms e.g. DeepSpeed and Megatron-LM
- ***Ability Eliciting***
  - Potential abilities might not be explicitly exhibited when LLMs perform some specific tasks.
  - Solution: Design suitable task instructions or specific in-context learning strategies

Background

Key Techniques

- ***Alignment Tuning***
  - Likely to generate toxic, biased, or even harmful content for humans
  - Necessary to align LLMs with human values
  - Solution: Utilize reinforcement learning with human feedback *e.g. InstructGPT and ChatGPT*
- ***Tools Manipulation***
  - Perform less well on the tasks that are not best expressed in the form of text
  - Solution: Employ external tools to compensate for the deficiencies of LLMs e.g. ChatGPT

Capacity and Evaluation

Capacity

Basic Ability Evaluation

- *Language Generation*
  - **Language Modeling**
    - Predict next token, datasets like Penn Treebank, WikiText-103, the Pile, and LAMBADA.
  - **Conditional Text Generation**
    - Machine translation, summarization, QA, with examples of GPT-4's translation performance.
  - **Code Synthesis**
    - Generating computer programs, benchmarks like APPS, HumanEval, MBPP
    - Evaluate: calculate the pass rate against the test cases, i.e., pass@k

Capacity

Basic Ability Evaluation

- **Knowledge Utilization**: An important ability of intelligent systems to accomplish knowledge-intensive tasks
  - **Closed-Book QA**
    - Test the acquired factual knowledge of LLMs from the pre-training corpus
    - Datasets: Natural Questions, Web Questions and TriviaQA
    - Show a scaling law pattern in terms of both model size and data size
  - **Open-Book QA**
    - LLMs can extract useful evidence from the external knowledge base or document collections.
    - Datasets: Natural Questions, OpenBookQA and SQuAD
      - Have overlap with closed-book QA datasets but incorporate external data sources
    - Evaluation: The metrics of accuracy and F1 score
  - **Knowledge Completion**
    - In knowledge completion tasks, LLMs might be (to some extent) considered as a knowledge base.
    - Can be leveraged to complete or predict the missing parts of knowledge units
    - Categories: Knowledge graph completion tasks (e.g., FB15k- 237 and WN18RR) and fact completion tasks
    - Difficult for existing LLMs to accomplish knowledge completion tasks related to specific relation types

Major Issues

- Hallucination
  - Solution: Alignment Tuning and Tool Utilization
- Knowledge recency
  - Solution: Augmenting LLMs with external knowledge sources

Capacity

Basic Ability Evaluation

- ***Complex Reasoning****: The ability of understanding and utilizing supporting evidence or logic to derive conclusions or make decisions*
  - **Knowledge Reasoning**
    - Rely on logical relations and evidence about factual
    - Datasets: CSQA/StrategyQA for commonsense knowledge reasoning and ScienceQA for science knowledge reasoning
  - **Symbolic Reasoning**
    - Manipulate the symbols in a formal rule setting to fulfill some specific goal
    - Datasets for last letter concatenation and coin flip
  - **Mathematical Reasoning**
    - Solve math problems
      - Datasets: SVAMP, GSM8k, and MATH
      - Employ Chain of Thought prompting and ITPs
    - Automated theorem proving
      - Datasets: PISA and miniF2F
      - Evaluation: The proof success rate

Major Issues

- *Reasoning inconsistency*
  - *Solution: Fine-tuning LLMs with process-level feedback*
- *Numerical computation*
  - *Solution: Use mathematical tools and tokenize digits into individual tokens*

Capacity

Advanced Ability Evaluation

- *Human Alignment*
  - LLms should well conform to human values and needs
  - Benchmarks like TruthfulQA, Winogender, CrowS-Pairs, and expert reviews
- *Interaction with External Environment*
  - Utilize AI environments like VirtualHome, ALFRED, BEHAVIOR
  - Proficiency in real-world scenarios and open-world environments.
- *Tool Manipulation*
  - Enhance performance with external tools (search engines, calculators, compilers)
  - Integration of third-party plugins in platforms like ChatGPT.

Benchmark

Comprehensive Evaluation Benchmarks

- MMLU
  - A versatile benchmark for large-scale evaluation of multi-task knowledge understanding
  - Cover a wide range of knowledge domains
- BIG-bench
  - Comprise 204 tasks that encompass a broad range of topics
- HELM
  - Implement a core set of 16 scenarios and 7 categories of metrics
- Human-level test benchmarks
  - Evaluate the comprehensive ability of LLMs with questions designed for testing humans
  - Example: AGIEval , MMCU, M3KE, C-Eval and Xiezhi
- Others
  - TyDiQA, MGSM

Evaluation Approaches

Evaluation of Fine-tuned LLMs.

- *Human-based evaluation*
  - Test tasks are usually in the form of open-ended questions.
  - Invite human evaluators to make judgments on the quality of answers generated by LLMs.
    - Two methods: Pairwise comparison(e.g. Chatbot Arena) and single-answer grading (e.g. HELM)
- *Model-based evaluation*
  - Leverage powerful closed-source LLMs such as ChatGPT and GPT-4 as a surrogate for human evaluators

Evaluation of Specialized LLMs

- Construct specific benchmarks tailored for the target domains or applications
- Combine these domain-specific benchmarks with general benchmarks to conduct evaluation
- Example: MultiMedQA combined with MMLU for healthcare

Empirical Evaluation

- A fine-grained evaluation of the both basic and advanced abilities
- *Experimental Settings*
  - **Evaluation Models**
    - *Open-source models*
      - *Base model:* LLaMA (7B), LLaMA 2 (7B) Pythia (7B) and Falcon (7B)
      - *Instruction-tuned models*: Vicuna (7B and 13B) , Alpaca (7B) , and ChatGLM (6B)
      - LLaMA 2-Chat (7B) for comparison
    - *Closed-source models*
      - text-davinci-002/003 (short as *Davinci002/003*), ChatGPT, Claude, and Claude 2

Empirical Evaluation

Experimental Settings

- **Tasks and Datasets**
  - **Language generation**
    - **LAMBADA (language modeling), WMT'22 (machine translation), XSum [425] (text summarization), and HumanEval [92] (code synthesis)**
  - **Knowledge utilization**
    - **Knowledge reasoning: OpenbookQA, HellaSwag, and SocialIQA**
    - **Symbolic reasoning: Colored Objects and Penguins**
    - **Mathematical reasoning: GSM8k and MATH**
  - **Human alignment**
    - **TruthfulQA**
    - **CrowS-Pairs and WinoGender to assess the stereotypes in LLMs**
    - **RealToxityPromptsto evaluate the extent of generating toxic language**
    - **HaluEval to test the ability of LLMs to recognize hallucination**
  - Interaction with environment
    - ALFWorld (household) and WebShop (e-commerce environments)
  - Tool manipulation
    - HotpotQA: Use search engine
    - Gorilla: Invoke model APIs from three hubs of TorchHub, TensorHub and HuggingFace

Results Analysis and Findings

Analysis of Closed-Source Models
(ChatGPT, Claude, Davinci003 and Davinci002)

- As general-purpose task solvers
  - ChatGPT mostly performs the best.
  - The large performance gap between ChatGPT and other closed-source models
- Interaction with the environment and tool manipulation tasks
  - Claude 2, ChatGPT and Davinci003, perform much better.
- Difficult reasoning tasks
  - On MATH and HotpotQA, all perform not well.
- Machine translation task
  - All models have a weak performance on WMT.

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	3[illegible]44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Claude	64.47	3[illegible]23	18.63	51.22	40.92	13.77	14.57	66.62	34.34
Claude 2	45.20	1[illegible]93	19.13	78.04	54.30	21.30	21.06	79.97	35.83
Davinci003	69.98	3[illegible]46	18.19	67.07	51.51	17.76	16.68	88.47	28.29
Davinci002	58.85	3[illegible]41	19.15	56.70	52.11	20.47	18.45	89.23	29.15
LLaMA 2-Chat (7B)	56.12	12.62	16.00	11.59	38.93	12.96	11.32	72.35	23.37
Vicuna (13B)	62.45	20.49	17.87	20.73	29.04	10.75	11.52	20.69	28.76
Vicuna (7B)	63.90	19.95	13.59	17.07	28.58	9.17	6.64	16.96	26.95
Alpaca (7B)	63.35	21.52	8.74	13.41	17.14	3.24	3.00	49.75	26.05
ChatGLM (6B)	33.34	16.58	13.48	13.42	13.42	4.40	9.20	55.39	16.01
LLaMA 2 (7B)	66.39	11.57	11.57	17.07	30.92	5.15	2.51	24.16	28.06
LLaMA (7B)	67.68	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78
Falcon (7B)	66.89	4.05	10.00	10.37	28.74	10.78	8.46	4.08	23.91
Pythia (12B)	61.19	5.43	8.87	14.63	15.73	1.99	4.72	11.66	20.57
Pythia (7B)	56.96	3.68	8.23	9.15	10.16	1.77	3.74	1.03	15.75

Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Claude	81.80	54.95	73.23	59.95	47.65	70.81	20.18	32.09	50.02/30.40
Claude 2	71.60	50.75	58.34	66.76	74.50	82.87	32.24		34.96/19.20
Davinci003	74.40	62.65	69.70	64.60	61.07	57.16	17.66	65.67	64.08/32.40
Davinci002	69.80	47.81	57.01	62.50	67.11	49.96	14.28	76.87	29.66/15.20
LLaMA 2-Chat (7B)	45.62	74.01	43.84	43.40	38.93	9.63	2.22	11.19	24.51/5.60
Vicuna (13B)	43.65	70.51	45.97	53.55	36.91	18.50	3.72	8.96	22.74/5.00
Vicuna (7B)	43.84	69.25	46.27	44.25	36.24	14.03	3.54	1.49	6.90/1.40
Alpaca (7B)	47.82	69.81	47.55	39.35	40.27	4.93	4.16	4.48	0.00/0.00
ChatGLM (6B)	30.42	29.27	33.18	14.05	14.09	3.41	1.10	0.00	0.00/0.00
LLaMA 2 (7B)	44.81	74.25	41.72	43.95	35.75	10.99	2.64	8.96	0.00/0.00
LLaMA (7B)	42.42	73.91	41.46	39.95	34.90	10.99	3.12	2.24	0.00/0.00
Falcon (7B)	39.46	74.58	42.53	29.80	24.16	1.67	0.94	7.46	0.00/0.00
Pythia (12B)	37.02	65.45	41.53	32.40	26.17	2.88	1.96	5.22	3.68/0.60
Pythia (7B)	34.88	61.82	41.01	29.05	27.32	1.82	1.46	7.46	10.75/1.80

Models	Human Alignment					Tool Manipulation			
	TfQA↑	C-Pairs↓	WinoGender↑	RTP↓	HaluEval↑	HotpotQA↑	Gorilla-TH↑	Gorilla-TF↑	Gorilla-HF↑
ChatGPT	69.16	18.60	62.50/72.50/79.17	3.07	66.64	23.80	67.20	44.53	19.36
Claude	67.93	32.73	71.67/55.00/52.50	3.75	63.75	33.80	22.04	7.74	7.08
Claude 2	71.11	10.67	60.00/60.00/55.83	3.20	50.63	36.4	61.29	22.19	23.67
Davinci003	60.83	0.99	67.50/68.33/79.17	8.81	58.94	34.40	72.58	3.80	6.42
Davinci002	53.73	7.56	72.50/70.00/64.17	10.65	59.67	26.00	2.69	1.02	1.00
LLaMA 2-Chat (7B)	69.77	48.54	47.50/46.67/46.67	4.61	43.82	4.40	0.00	0.00	0.22
Vicuna (13B)	62.30	45.95	50.83/50.83/52.50	5.00	49.01	11.20	0.00	0.44	0.89
Vicuna (7B)	57.77	67.44	49.17/49.17/49.17	4.70	43.44	6.20	0.00	0.00	0.33
Alpaca (7B)	46.14	65.45	53.33/51.67/53.33	4.78	44.16	11.60	0.00	0.00	0.11
ChatGLM (6B)	63.53	50.53	47.50/47.50/46.67	2.89	41.82	4.00	0.00	0.00	0.00
LLaMA 2 (7B)	50.06	51.39	48.83/48.83/50.83	6.17	42.23	3.80	0.00	0.00	0.11
LLaMA (7B)	47.86	67.84	54.17/52.50/51.67	5.94	14.18	1.60	0.00	0.00	0.11
Falcon (7B)	53.24	68.04	50.00/50.83/50.00	6.71	37.41	1.00	0.00	0.00	0.00
Pythia (12B)	54.47	65.78	49.17/48.33/49.17	6.59	27.09	0.40	0.00	0.00	0.00
Pythia (7B)	50.92	64.79	51.67/49.17/50.00	13.02	25.84	0.20	0.00	0.00	0.00

TABLE : Evaluation on the eight abilities of LLMs with specially selected tasks.
Orange: closed-source and Blue: open-source

Results Analysis and Findings

Analysis of Open-Source Models (LLaMA 2-Chat, Vicuna, Alpaca, ChatGLM, LLaMA 2, LLaMA, Pythia and Falcon)

- **The instruction-tuned models (LLaMA 2-Chat, Vicuna, Alpaca and ChatGLM) mostly perform better than non- instruction-tuned models (LLaMA 2, LLaMA, Pythia and Falcon).**
- **Small-sized open-source models perform not well on mathematical reasoning, interaction with environment, and tool manipulation tasks.**
- **The top-performing model varies on different human alignment tasks.**
- **Scaling the open-source modesl can improve the performance consistently**

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	3.44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Claude	64.47	3.23	18.63	51.22	40.92	13.77	14.57	66.62	34.34
Claude 2	45.20	1.93	19.13	78.04	54.30	21.30	21.06	79.97	35.83
Davinci003	69.98	3.46	18.19	67.07	51.51	17.76	16.68	88.47	28.29
Davinci002	58.85	3.11	19.15	56.70	52.11	20.47	18.45	89.23	29.15
LLaMA 2-Chat (7B)	56.12	12.62	16.00	11.59	38.93	12.96	11.32	72.35	23.37
Vicuna (13B)	62.45	20.49	17.87	20.73	29.04	10.75	11.52	20.69	28.76
Vicuna (7B)	63.90	19.95	13.59	17.07	28.58	9.17	6.64	16.96	26.95
Alpaca (7B)	63.35	21.52	8.74	13.41	17.14	3.24	3.00	49.75	26.05
ChatGLM (6B)	33.34	16.58	13.48	13.42	13.42	4.40	9.20	55.39	16.01
LLaMA 2 (7B)	66.39	11.57	11.57	17.07	30.92	5.15	2.51	24.16	28.06
LLaMA (7B)	67.68	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78
Falcon (7B)	66.89	4.05	10.00	10.37	28.74	10.78	8.46	4.08	23.91
Pythia (12B)	61.19	5.43	8.87	14.63	15.73	1.99	4.72	11.66	20.57
Pythia (7B)	56.96	3.68	8.23	9.15	10.16	1.77	3.74	11.03	15.75

Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Claude	81.80	54.95	73.23	59.95	47.65	70.81	20.18	32.09	50.02/30.40
Claude 2	71.60	50.75	58.34	66.76	74.50	82.87	32.24		34.96/19.20
Davinci003	74.40	62.65	69.70	64.60	61.07	57.16	17.66	65.67	64.08/32.40
Davinci002	69.80	47.81	57.01	62.55	67.11	49.96	14.28	76.87	29.66/15.20
LLaMA 2-Chat (7B)	45.62	74.01	43.84	43.40	38.93	9.63	2.22	11.19	24.51/5.60
Vicuna (13B)	43.65	70.51	45.97	53.55	36.91	18.50	3.72	8.96	22.74/5.00
Vicuna (7B)	43.84	69.25	46.27	44.25	36.24	14.03	3.54	1.49	6.90/1.40
Alpaca (7B)	47.82	69.81	47.55	39.35	40.27	4.93	4.16	4.48	0.00/0.00
ChatGLM (6B)	30.42	29.27	33.18	14.05	14.09	3.41	1.10	0.00	0.00/0.00
LLaMA 2 (7B)	44.81	74.25	41.72	43.95	35.75	10.99	2.64	8.96	0.00/0.00
LLaMA (7B)	42.42	73.91	41.46	39.95	34.90	10.99	3.12	2.24	0.00/0.00
Falcon (7B)	39.46	74.38	42.53	29.80	24.16	1.67	0.94	7.46	0.00/0.00
Pythia (12B)	37.02	65.45	41.53	32.40	26.17	2.88	1.96	5.22	3.68/0.60
Pythia (7B)	34.88	61.82	41.01	29.05	27.52	1.82	1.46	7.46	10.75/1.80

Models	Human Alignment					Tool Manipulation			
	TfQA↑	C-Pairs↓	WinoGender↑	RTP↓	HaluEval↑	HotpotQA↑	Gorilla-TH↑	Gorilla-TF↑	Gorilla-HF↑
ChatGPT	69.16	18.60	62.50/72.50/79.17	3.07	66.64	23.80	67.20	44.53	19.36
Claude	67.93	32.73	71.67/55.00/52.50	3.75	63.75	33.80	22.04	7.74	7.08
Claude 2	71.11	10.67	60.00/60.00/55.83	3.20	50.63	36.4	61.29	22.19	23.67
Davinci003	60.83	0.99	67.50/68.33/79.17	8.81	58.94	34.40	72.58	3.80	6.42
Davinci002	53.73	7.56	72.50/70.00/64.17	10.65	59.67	26.00	2.69	1.02	1.00
LLaMA 2-Chat (7B)	69.77	48.54	47.50/46.67/46.67	4.61	43.82	4.40	0.00	0.00	0.22
Vicuna (13B)	62.30	45.95	50.83/50.83/52.50	5.00	49.01	11.20	0.00	0.44	0.89
Vicuna (7B)	57.77	47.44	49.17/49.17/49.17	4.70	43.44	6.20	0.00	0.00	0.33
Alpaca (7B)	46.14	65.45	53.33/51.67/53.33	4.78	44.16	11.60	0.00	0.00	0.11
ChatGLM (6B)	63.53	50.53	47.50/47.50/46.67	2.89	41.82	4.00	0.00	0.00	0.00
LLaMA 2 (7B)	50.06	51.39	48.83/48.83/50.83	6.17	42.23	3.80	0.00	0.00	0.11
LLaMA (7B)	47.86	67.84	54.17/52.50/51.67	5.94	14.18	1.60	0.00	0.00	0.11
Falcon (7B)	53.24	68.04	50.00/50.83/50.00	6.71	37.41	1.00	0.00	0.00	0.00
Pythia (12B)	54.47	65.78	49.17/48.33/49.17	6.59	27.09	0.40	0.00	0.00	0.00
Pythia (7B)	50.92	64.79	51.67/49.17/50.00	13.02	25.84	0.20	0.00	0.00	0.00

TABLE : Evaluation on the eight abilities of LLMs with specially selected tasks. Orange: closed-source and Blue: open-source

Thank you for your listening!

Discussion & Questions

- How have LLMs revolutionized AI development?
- What are the remaining issues and future directions for LLMs?