Analytics Vidhya

Introduction to Training LLMs from Scratch

Instructor

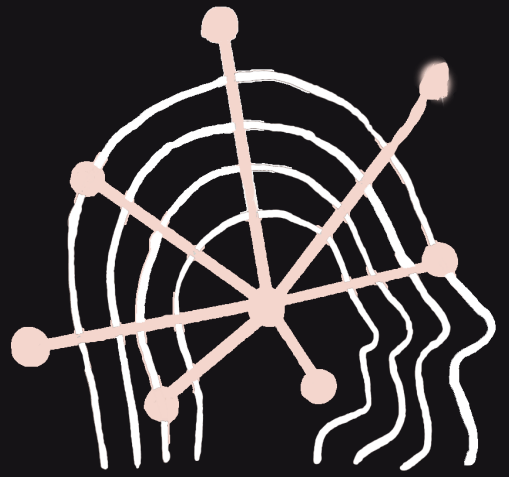
Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Introduction

Bard



CLAUDE 2

Introduction

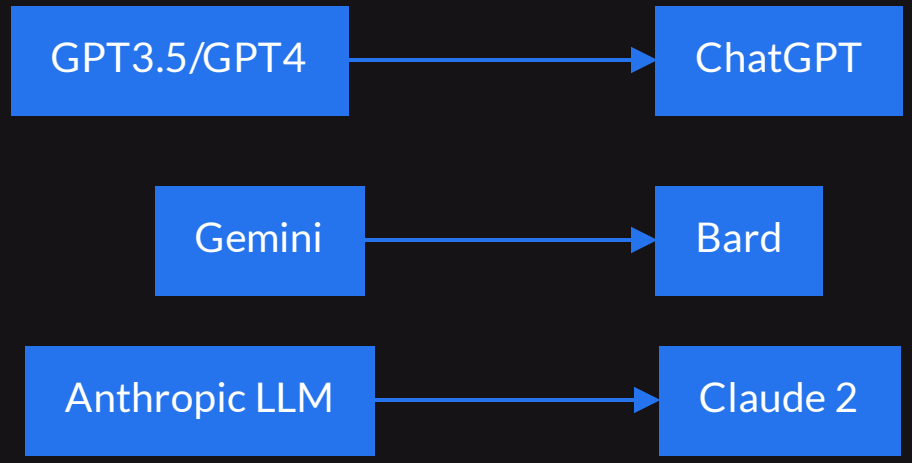
Trip Planning

Business Strategizing

Analytics Vidhya

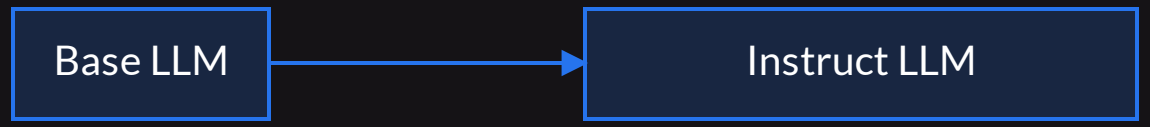
Introduction

Instruction following LLMs are built on top of base LLMs.



Introduction

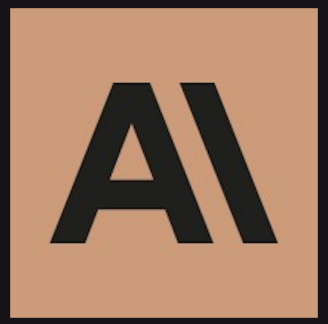
Instruction following LLMs are built on top of base LLMs.



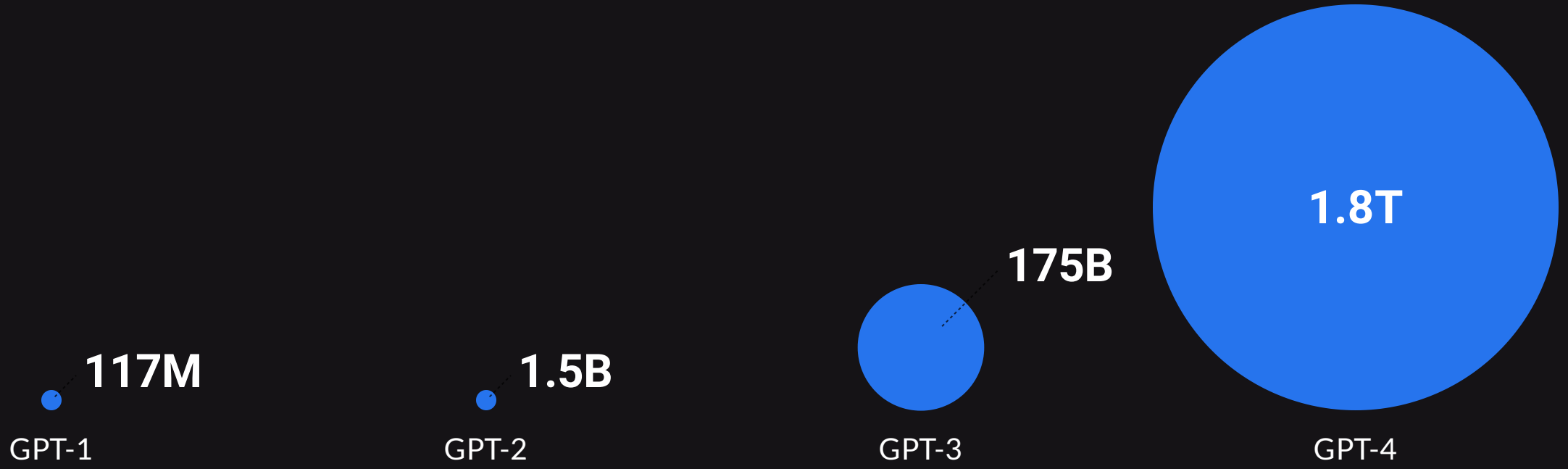
Experiments

- Training on domain specific datasets
- Varying Model architectures
- Scaling LLMs

Major Contributions



The rise of GPT-4



The rise of Llama 2

Llama

65 billion parameters & 1.4 trillion tokens

Llama 2

70 billion parameters & 2 trillion tokens

The rise of Gemini

PaLM

540 billion parameters & trillion+ tokens

PaLM 2

3.6 trillion tokens

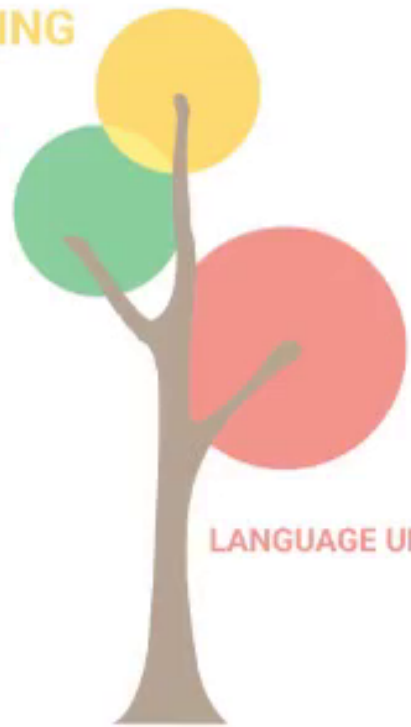
The rise of Gemini

The rise of Gemini

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	95.3% 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	94.4% maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	53.2% 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)

QUESTION ANSWERING

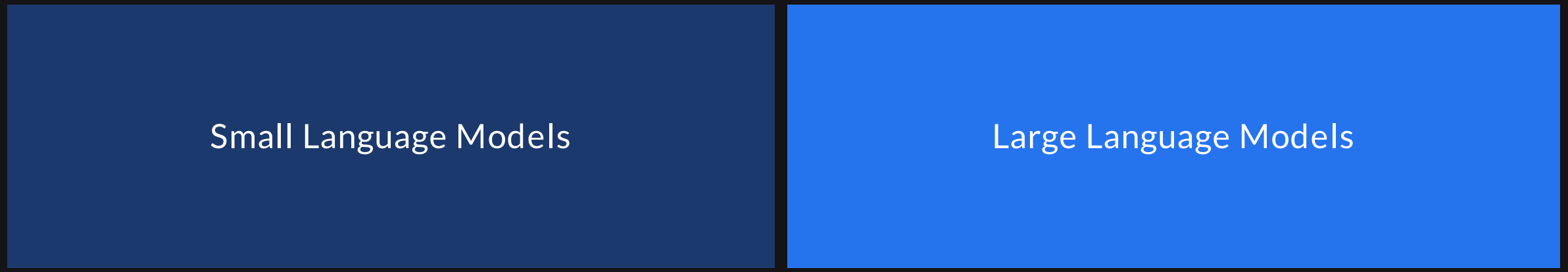
ARITHMETIC



LANGUAGE UNDERSTANDING

8 billion parameters

The rise of Small Language Models



Analytics Vidhya

The rise of Phi

Microsoft released Small Language Models:

- Phi-1
- Phi-1.5
- Phi-2

Phi-1

Date	Model	Model size (Parameters)	Dataset size (Tokens)	HumanEval (Pass@1)	MBPP (Pass@1)
2021 Jul	Codex-300M [CTJ+21]	300M	100B	13.2%	-
2021 Jul	Codex-12B [CTJ+21]	12B	100B	28.8%	-
2022 Mar	CodeGen-Mono-350M [NPH+23]	350M	577B	12.8%	-
2022 Mar	CodeGen-Mono-16.1B [NPH+23]	16.1B	577B	29.3%	35.3%
2022 Apr	PaLM-Coder [CND+22]	540B	780B	35.9%	47.0%
2022 Sep	CodeGeeX [ZXZ+23]	13B	850B	22.9%	24.4%
2022 Nov	GPT-3.5 [Ope23]	175B	N.A.	47%	-
2022 Dec	SantaCoder [ALK+23]	1.1B	236B	14.0%	35.0%
2023 Mar	GPT-4 [Ope23]	N.A.	N.A.	67%	-
2023 Apr	Replit [Rep23]	2.7B	525B	21.9%	-
2023 Apr	Replit-Finetuned [Rep23]	2.7B	525B	30.5%	-
2023 May	CodeGen2-1B [NHX+23]	1B	N.A.	10.3%	-
2023 May	CodeGen2-7B [NHX+23]	7B	N.A.	19.1%	-
2023 May	StarCoder [LAZ+23]	15.5B	1T	33.6%	52.7%
2023 May	StarCoder-Prompted [LAZ+23]	15.5B	1T	40.8%	49.5%
2023 May	PaLM 2-S [ADF+23]	N.A.	N.A.	37.6%	50.0%
2023 May	CodeT5+ [WLG+23]	2B	52B	24.2%	-
2023 May	CodeT5+ [WLG+23]	16B	52B	30.9%	-
2023 May	InstructCodeT5+ [WLG+23]	16B	52B	35.0%	-
2023 Jun	WizardCoder [LXZ+23]	16B	1T	57.3%	51.8%
2023 Jun	**phi-1**	1.3B	7B	50.6%	55.5%

Phi-1.5

	WinoGrande	ARC-Easy	ARC-Challenge	BoolQ	SIQA
Vicuna-13B (v1.1)	0.708	0.754	0.432	**0.835**	0.437
Llama2-7B	0.691	**0.763**	0.434	0.779	0.480
Llama-7B	0.669	0.682	0.385	0.732	0.466
MPT-7B	0.680	0.749	0.405	0.739	0.451
Falcon-7B	0.662	0.719	0.363	0.685	0.452
Falcon-rw-1.3B	0.607	0.633	0.282	0.632	0.405
OPT-1.3B	0.610	0.570	0.232	0.596	–
GPT-Neo-2.7B	0.577	0.611	0.274	0.618	0.400
GPT2-XL-1.5B	0.583	0.583	0.250	0.618	0.394
phi-1.5-web-only (1.3B)	0.604	0.666	0.329	0.632	0.414
phi-1.5-web (1.3B)	**0.740**	**0.761**	**0.449**	0.728	**0.530**
phi-1.5 (1.3B)	0.734	0.756	0.444	0.758	0.526

Common Sense Reasoning

	PIQA	Hellaswag	MMLU	OpenbookQA	SQUAD (EM)
Vicuna-13B	0.774	**0.578**	–	0.330	–
Llama2-7B	0.781	0.571	**0.453**	0.314	0.67
Llama-7B	0.779	0.562	0.352	0.284	0.60
MPT-7B	0.789	0.571	0.268	0.314	0.60
Falcon-7B	**0.794**	0.542	0.269	0.320	0.16
Falcon-rw-1.3B	0.747	0.466	0.259	0.244	–
OPT-1.3B	0.690	0.415	–	0.240	–
GPT-Neo-2.7B	0.729	0.427	–	0.232	–
GPT2-XL-1.5B	0.705	0.400	–	0.224	–
phi-1.5-web-only (1.3B)	0.743	0.478	0.309	0.274	–
phi-1.5-web (1.3B)	0.770	0.484	0.379	0.360	**0.74**
phi-1.5 (1.3B)	0.766	0.476	0.376	**0.372**	0.72

Language Understanding

Phi-2

Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

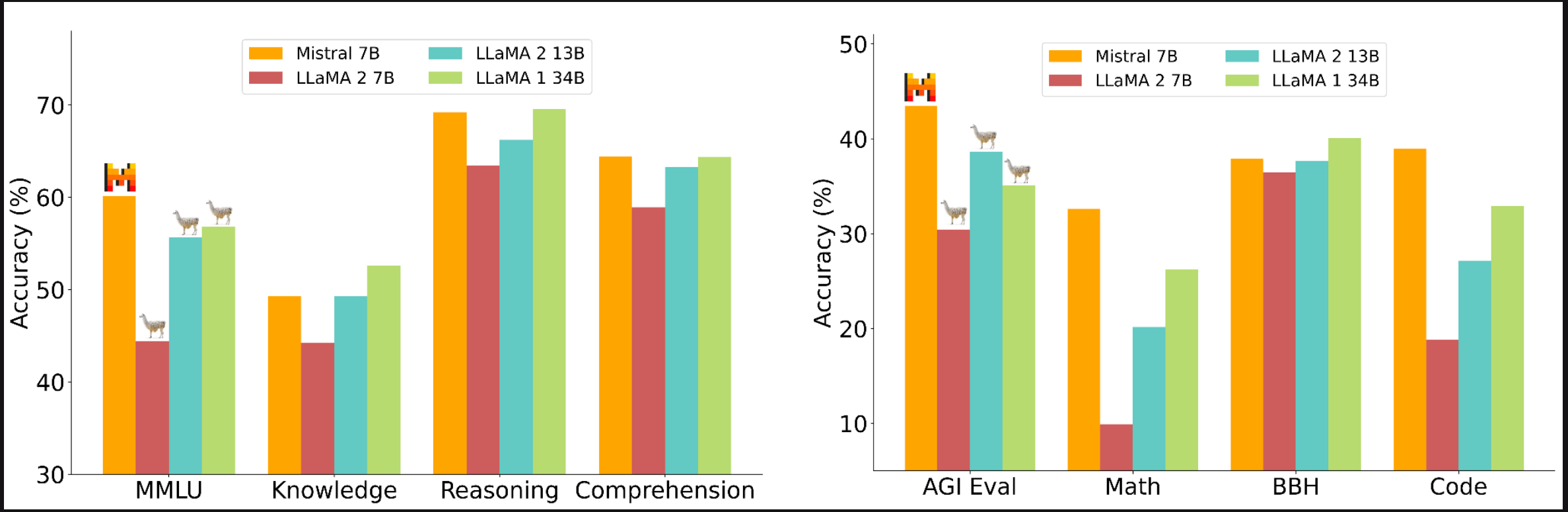
Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.

Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Small Language Models

- Training on domain specific datasets
- High quality training data

The rise of Mistral



How do we build LLMs?

Thank You

Analytics Vidhya