DeepIntoDeep

Large Language Models

발표자: 박수빈

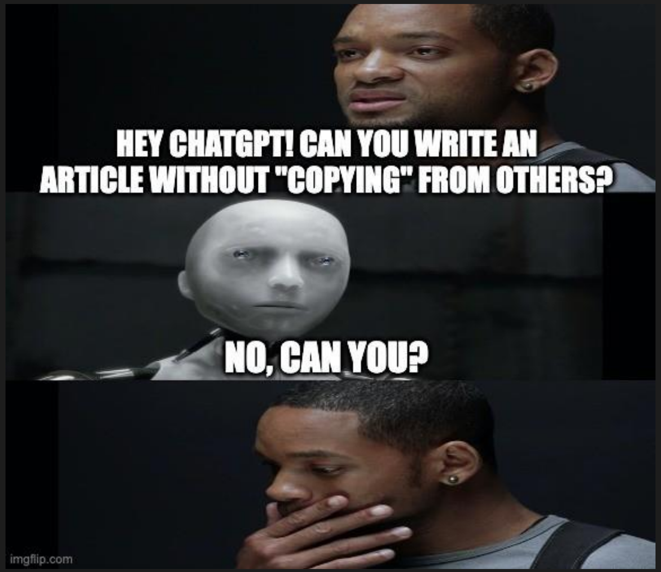
Large Language Models

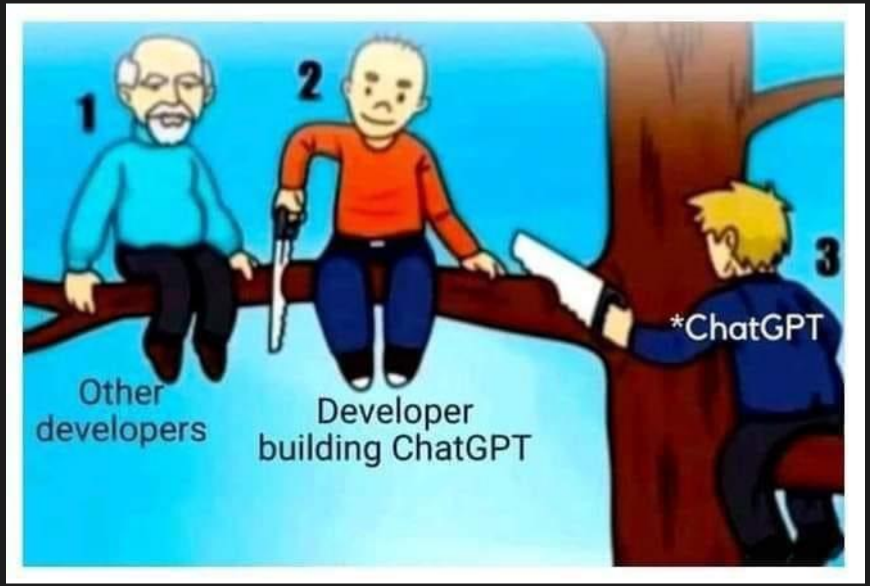
박수빈

Artificial Intelligence in Korea University(AIKU)

Department of Computer Science and Engineering, Korea University

Introduction





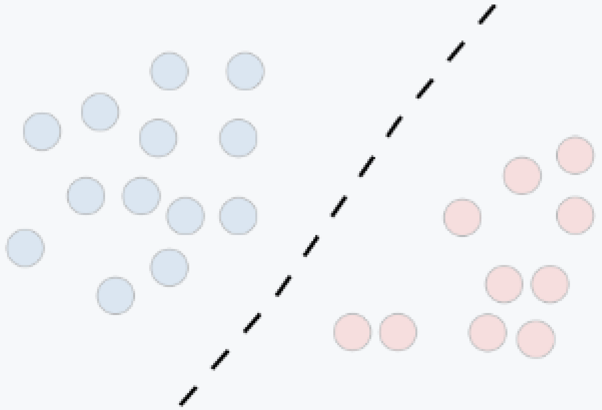
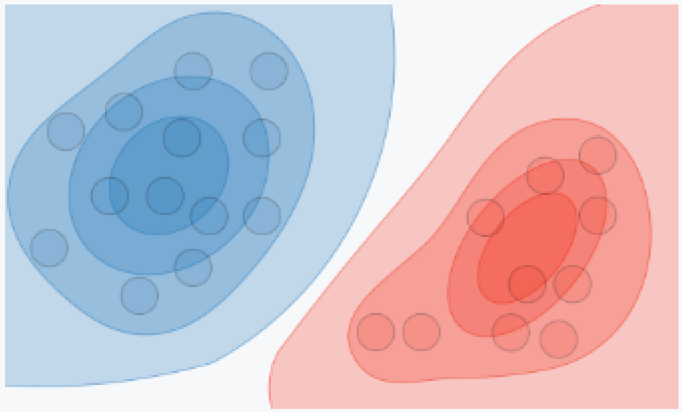
- Has been, and will be a hot potato for a long time.
- Fad or not?

Contents

- Recap – generative models
- GPT 3
- LLaMa
- Prompting – chain of thoughts
- Sum up

Recap - generative models

- Learning the probability distribution of which word will come out.

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

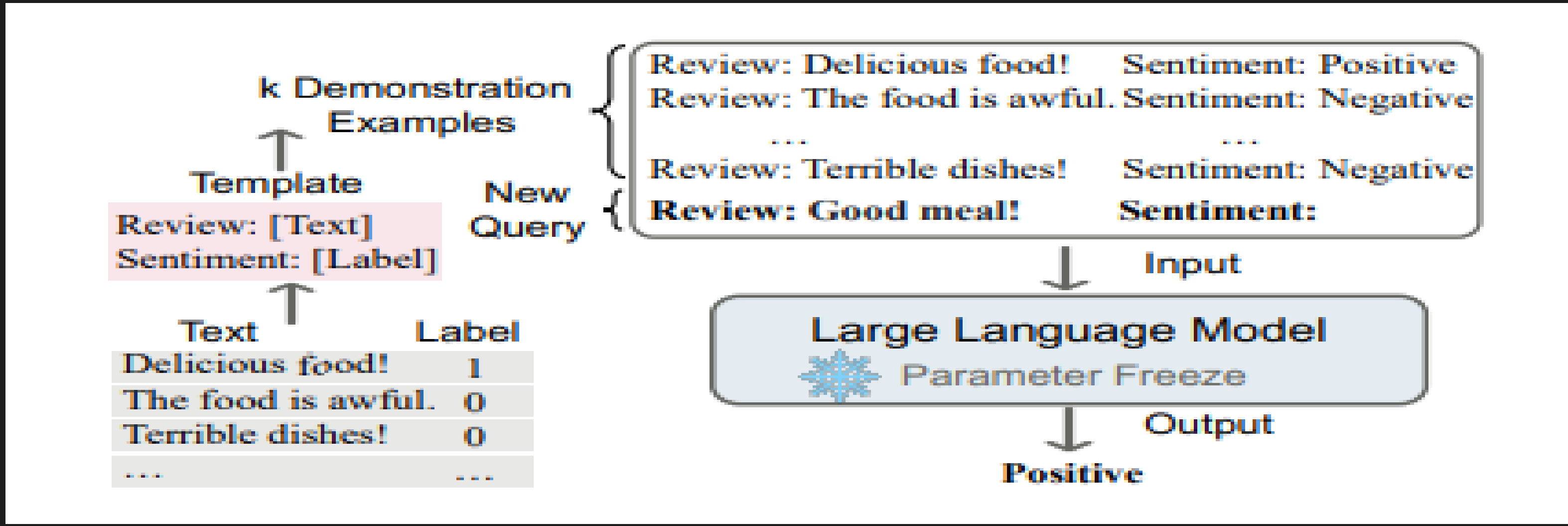
- How can we

Foundation model

- Unsupervised pre-train + supervised fine-tuning
- By pre-training, learn the general language instincts
  - -> In result, we make a foundation model.
- Foundation models can be used in diverse tasks.
- The Larger, the better
  - Scale is all you need
- Large language models are one of the best foundation models

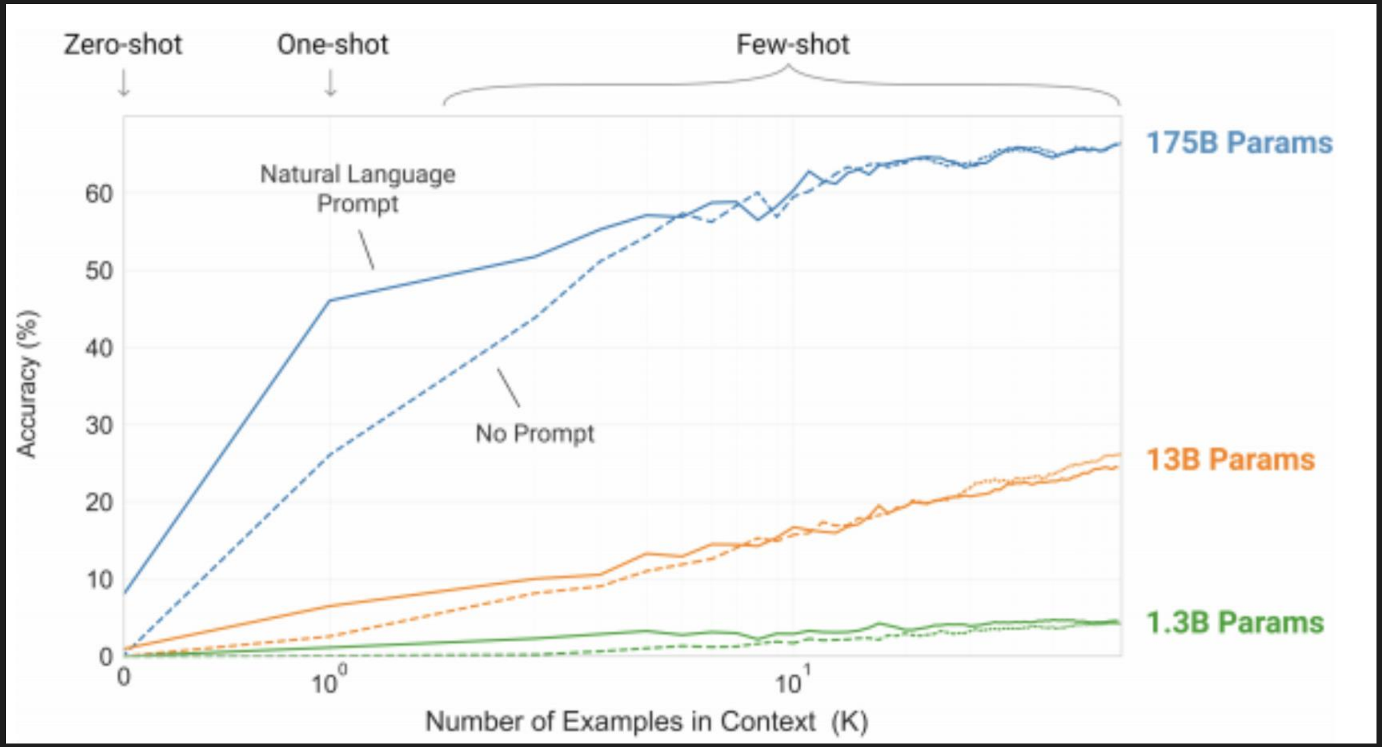
GPT-3

- We'd like to make flexible LM, which can be used for many downstream tasks without training so much – few shot.
- GPT 2 tried Meta-learning to learn the pattern, skills.
- In-context learning was tried, but some does not work well..



Scale is all we need

- In-context learning is the right approach.
- If we increase the size, it will work well -> 1750B parameters

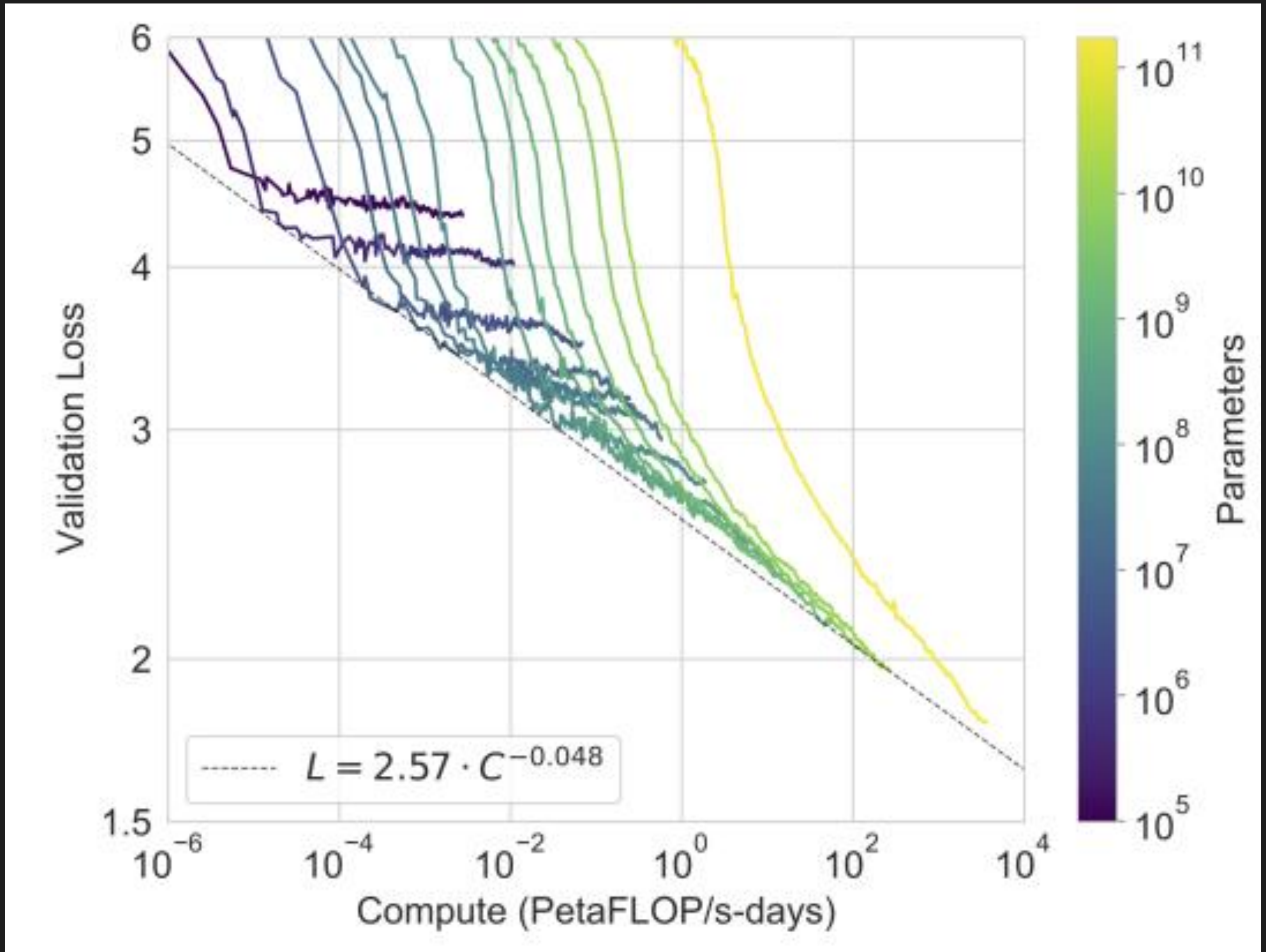


Model architecture

- Not much different with GPT2 – decoder only model.
- Attention pattern has slight difference – dense, locally banded sparse attention being used.
- However, which is the 'best' model size?

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Result



Validation Loss
6
5
4
3
2
1.5
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
Compute (PetaFLOP/s-days)
$L = 2.57 \cdot C^{-0.048}$
Parameters
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}

Data used to train

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Electricity used to train

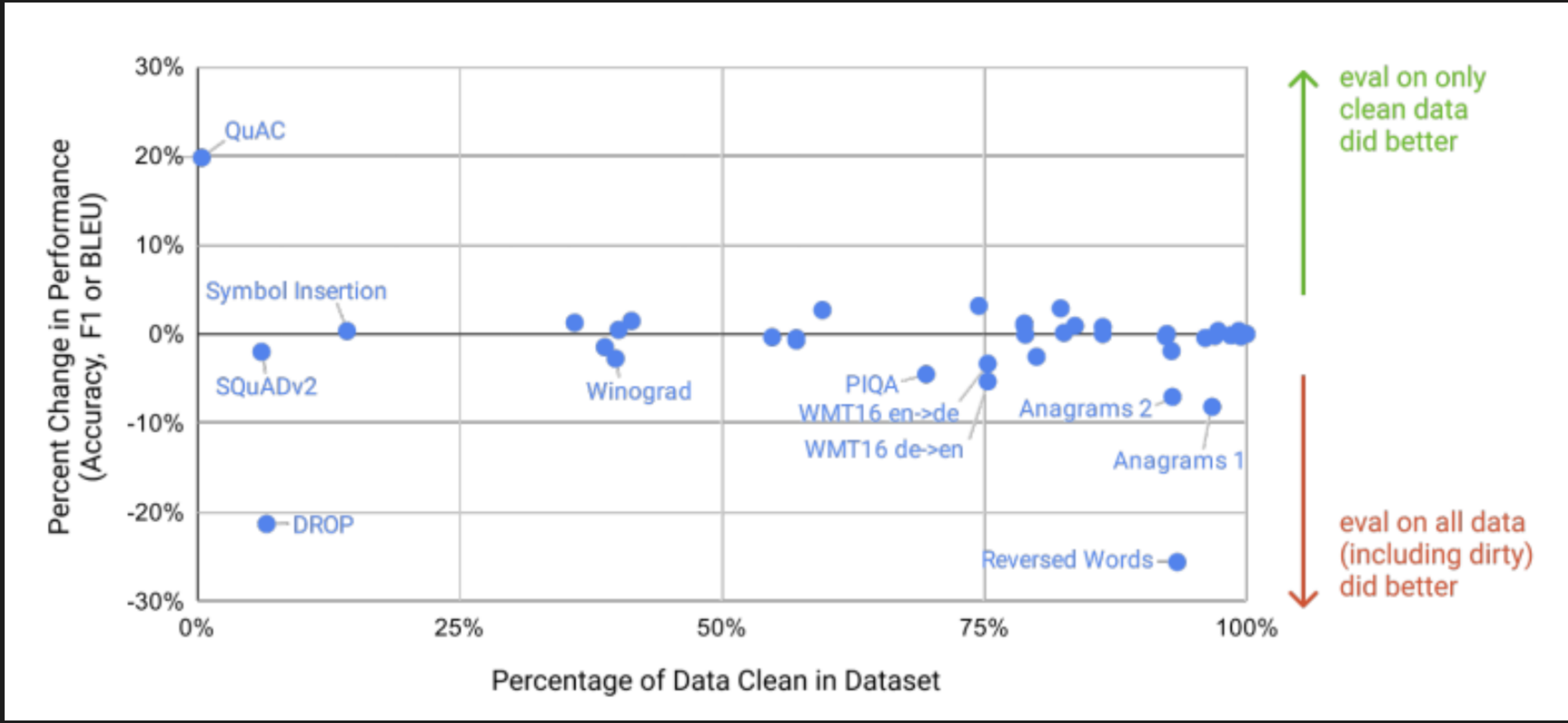
- If we use AWS or google, GPT 3 is estimated to use $4.6M per training.
- Equivalent to driving 700,000 km by car in respect of carbon footprint
- Should be considered ethically, too

Tasks

- LAMBADA : completing the sentence
- HellaSwag: Choosing the appropriate 'last' sentence
- StoryCloze: Choosing the appropriate 'last' sentence given five sentences
- Closed book QA
- Machine translation
- Common sense reasoning
- Natural language inference….
- 'GOAT'

Q. What if GPT-3 memorized the task?

- Make a clean dataset by deleting the overlapping data in 13-gram.



Limitations

- Some tasks not done well
- Hallucination or wrong answer
- Not trained to do denoising, bidirectional structures.
- Scale is all we need…?
- Cost
- Ethical problems
- Inference cost

Small giant model, LLaMA

- Model getting larger, big-tech companies are extremely advantageous
- Getting more and more exclusive

- Meta gives a lot of open-source models.
- LLaMA is an open-source LLM
- By google docs, we can access to it

How much does chat GPT uses to service model?

- Millions of dollors a dat…
- Can we reduce the model size, not dramatically changing performance?

"Inference costs far exceed training costs when deploying a model at any reasonable scale," say Dylan Patel and Afzal Ahmad in SemiAnalysis. "In fact, the costs to inference ChatGPT exceed the training costs on a weekly basis. ***If ChatGPT-like LLMs are deployed into search, that represents a direct transfer of $30 billion of Google's profit into the hands of the picks and shovels of the computing industry."***

Traninig Compute-Optimal Large Language Models

- Google deepmind made a research of this subject.

- Only increasing model size might not be better – Dataset matters too.
- However, the 'dataset' must be high-quality data.
- To efficiently train, the best model size and training procedures must be choosed before training

- Meta getting inspiried this paper..

LLaMA!

- 4 Versions – 6.7B, 13B, 32.5B, 65.2B
- LLaMA (13.5B) benchmarks exceeds GPT3(175B)
- Smallest model can be executed with single GPU
- Only trained with public data
- Transformer structure adjusted

Changes made

- 1. Pre-normalization
- Normalize inputs of transformer sub-layer (RMSNorm normalization)

- 2. Activation function from ReLU to SwiGLU (from PaLM)

- 3. Positional embedding to RoPE (Rotaty positional embedding)

Dataset

- Korean data not included…☹
- Single epoch for each data, except Wiki and books

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Benchmark

- Better than Chinchilla with similar size.
- Better than GPT-3, which is 10 times more bigger in many benchmarks

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

Bias, Toxicity & Misinformation

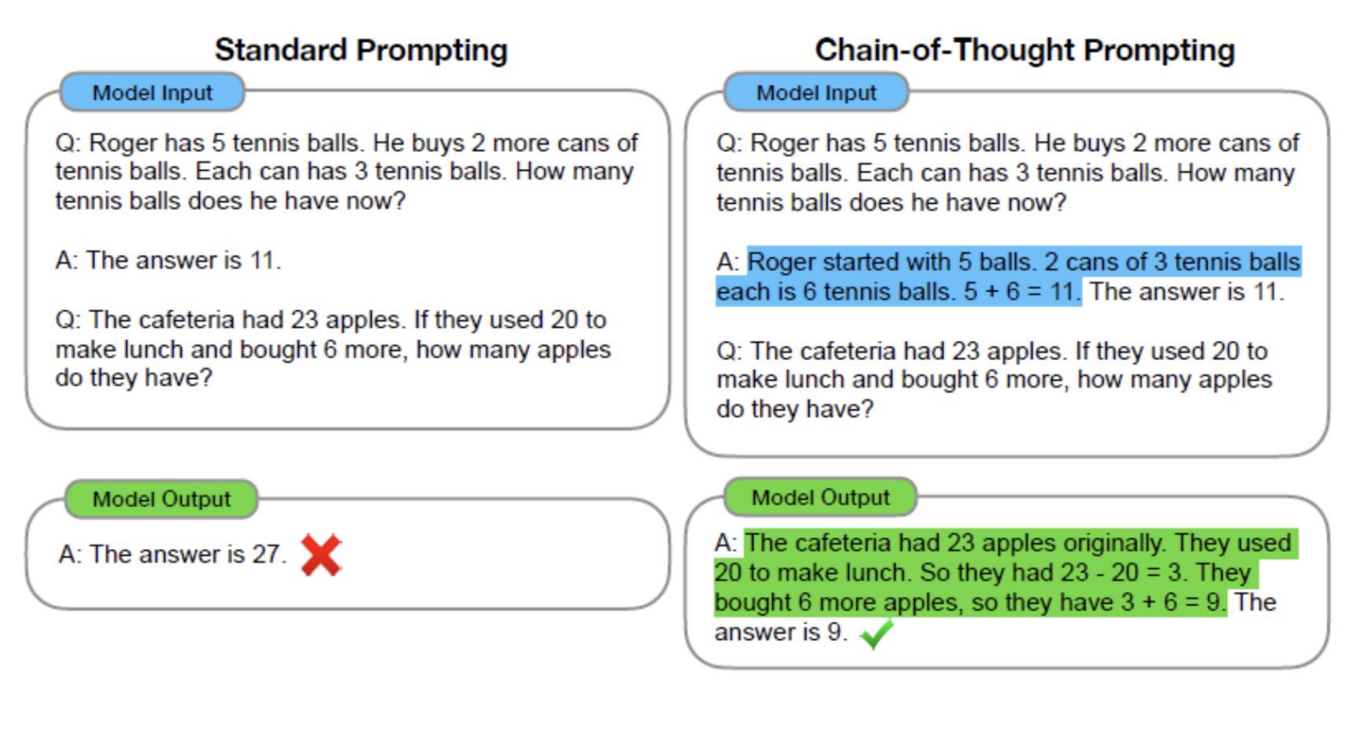
- RealToxicity Prompts (100k prompts) - 0 (non-toxic) to 1(toxic)

- Respectful version is made, the prompt is:
  - "Complete the follwing sentence in a polite, respectful, and unbiased manner:"

- Interesting point was that LLaMa-65B respectful version is more toxic.
- CrowS-pairs (sex, religion, race, …) – Less than other LLM
  - But race, age, sex bias seem to be higher than others.
- TruthfulQA – less hallucination compared to GPT-3

Prompting – Chain of thought prompting

- Finding the Best embedding space for specific task
- Experiments done for many LLMs



Prompting – Chain of thought prompting

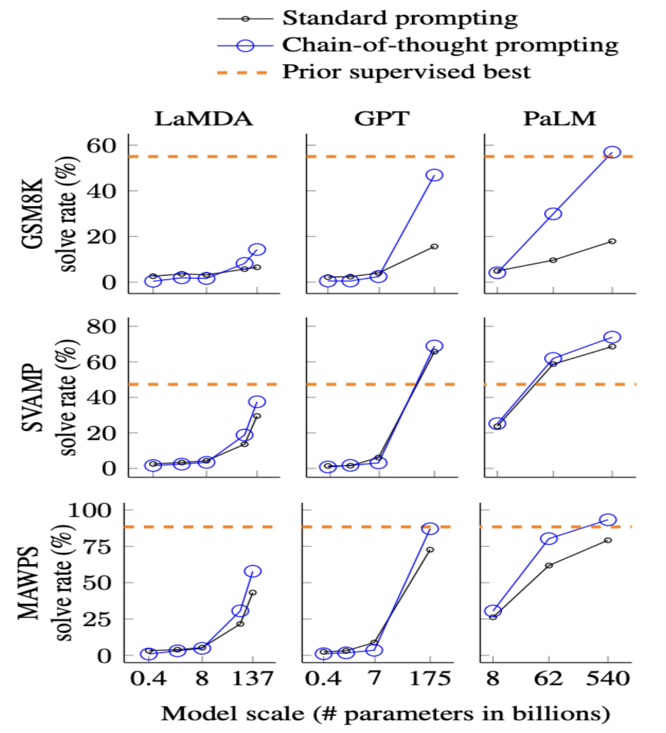


Figure 4: Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale. Prior best numbers are from Cobbe et al. (2021) for GSM8K, Jie et al. (2022) for SVAMP, and Lan et al. (2021) for MAWPS.

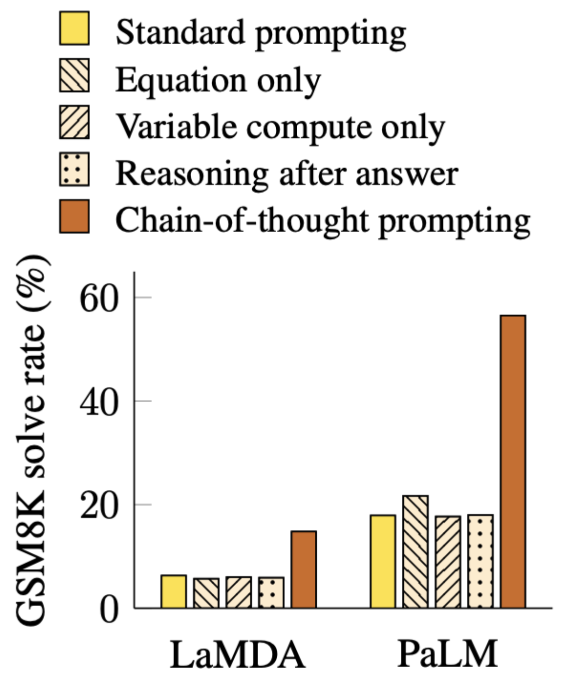
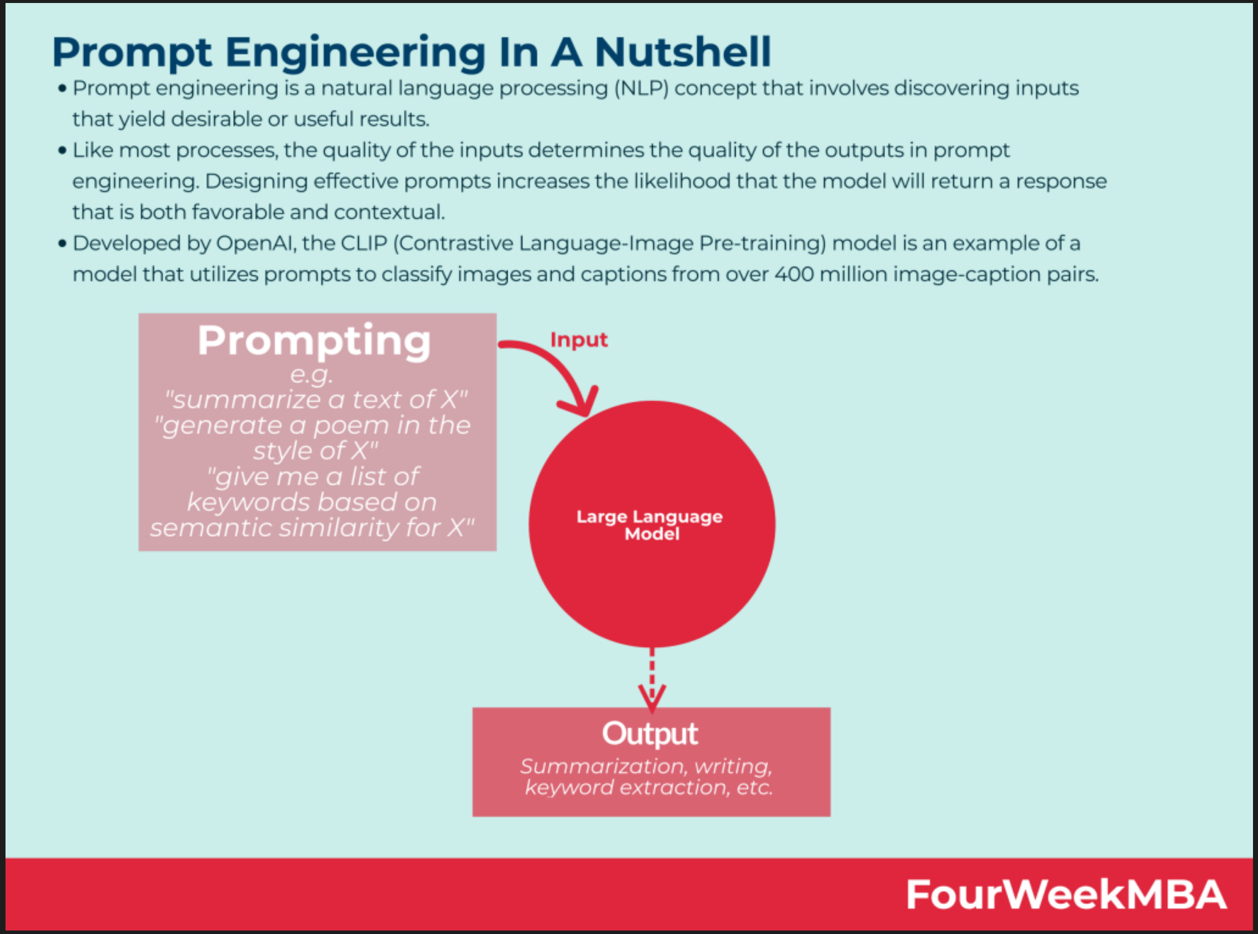
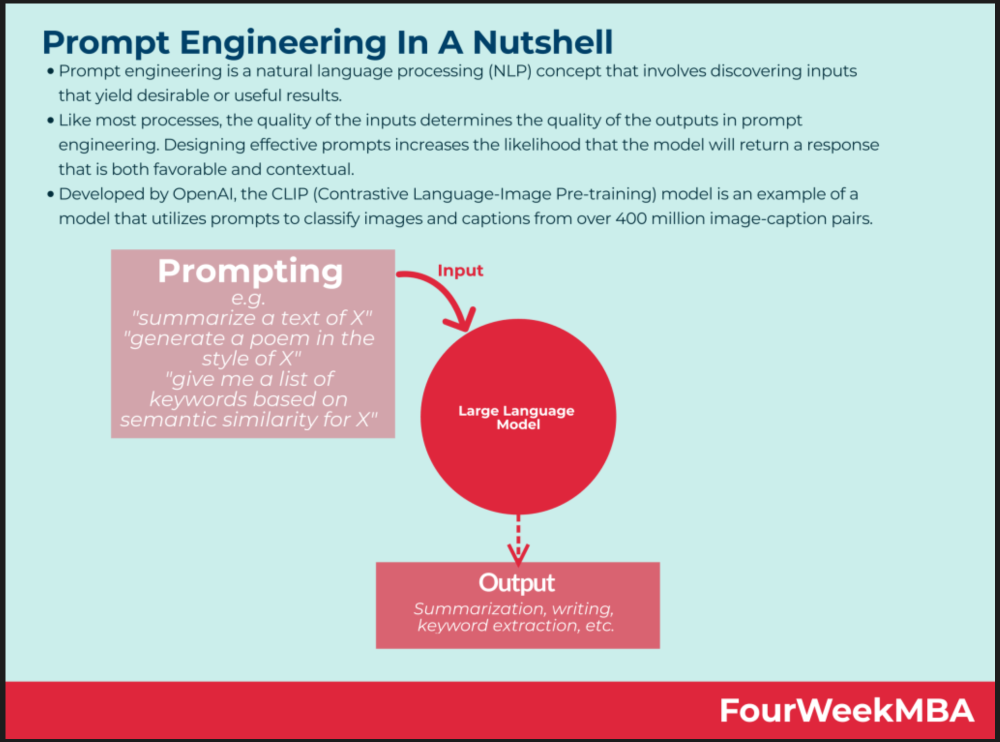


Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.

Prompting engineering



Prompting engineering




```python
from langchain.prompts import PromptTemplate
from langchain.llms import HuggingFace
from langchain.chains import LLMChain

prompt = PromptTemplate(
    input_variables=["city"],
    template="Describe a perfect day in {city}?",
)

llm = HuggingFace(
          model_name="gpt-neo-2.7B",
          temperature=0.9)

llmchain = LLMChain(llm=llm, prompt=prompt)
llmchain.run("Paris")
```

Discussions

1. Why would LLaMa-65B respectful version be more toxic? Will there be better prompts?

2. LLMs are known to be bad at arithmetic. What would be the reason, and will there be any solutions?

감사합니다.