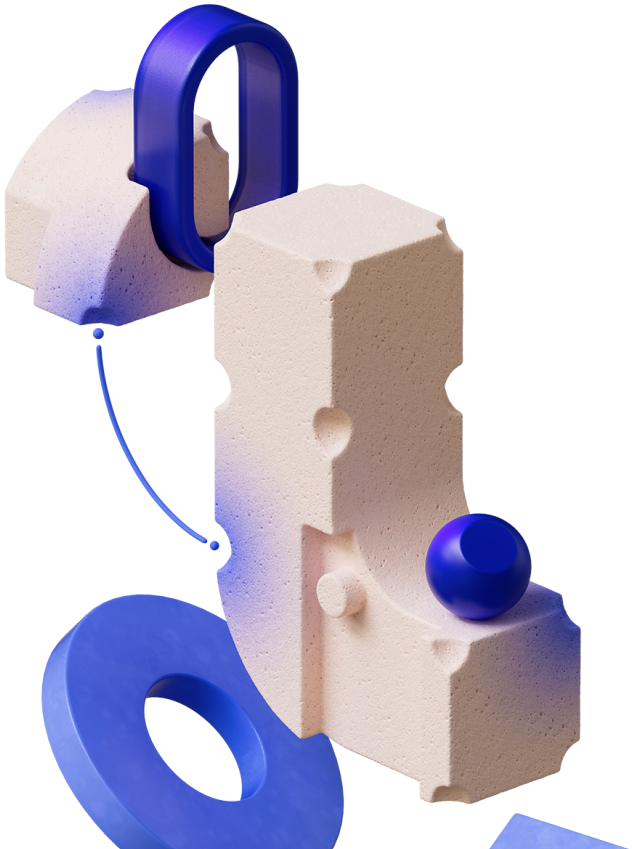
Google DeepMind



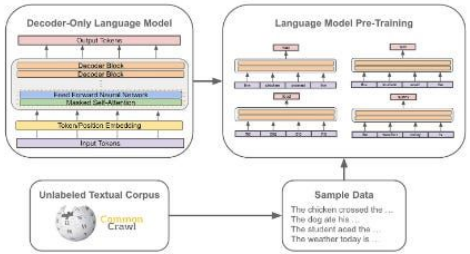
Introduction to Pre-training

Nikita Saxena
Research Engineer

Agenda

Alignment

Pre-Training



SFT

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

RLHF

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

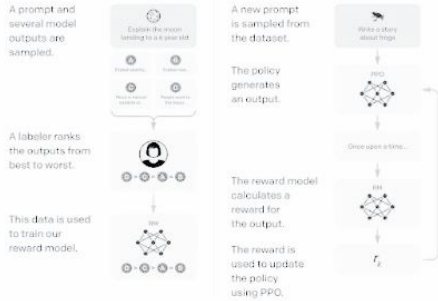
This data is used to train our reward model.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Quick Recap

Goal

- Enable LLMs to learn **general features and patterns** from large-scale datasets.
- Improve their generalization across tasks while **reducing reliance on labeled data.**

Internet-Scale Data

- Sources: *Common Crawl, C4 (Colossal Clean Crawled Corpus), The Pile, BooksCorpus, Wikipedia, GitHub, ArXiv.*
- Scale: *Terabytes to Petabytes of raw text.*

Data Sources & Preprocessing

1

01 Data Sources

Data Sources

Feature	Common Crawl (web scrapes)	Curated Sources (Books, Code)
Scale	**Massive** (Petabytes)	Smaller (Gigabytes to Terabytes)
Diversity	Extremely high	More focused, higher average quality
Quality	Contains noise, boilerplate, spam, low-quality text	Generally higher, professionally edited/reviewed
Structure	Often unstructured (raw HTML dumps)	Structured (e.g., book chapters, papers)
Cost (Acquisition)	Relatively low (publicly available)	Can be high (licensing, digitization)
Cleaning Effort	**Significant** - requires extensive filtering & deduplication	Lower
Primary Benefit	Breadth of knowledge, linguistic diversity, capturing "the long tail"	Depth in specific domains, factual accuracy, coherent long-form text
Key Challenge	Filtering out harmful/low-quality content without losing valuable data	Representativeness, potential for narrower scope

Sources


Papers
PubMed Central
PubMed Abstract
ArXiv
PhilPapers
S2ORC Full Text
S2ORC Abstract
Internet
Wikipedia
Legal/Formal
Freelaw
USPTO
Conversational
StackExchange
Reasoning
DM Maths
Books
PG19

Source Distribution of TxT360

TxT360 was the **first dataset to globally deduplicate** 99 CommonCrawl snapshots and 14 high-quality data sources from diverse domains (e.g., FreeLaw, PG-19, etc.)

Data Source	Raw Data Size	Token Count	Information Cut-Off Date
CommonCrawl	9.2 TB	4.83T	2024-30
Papers	712 GB	154.96B	Q4 2023
Wikipedia	199 GB	35.97B	-
Freelaw	71 GB	16.7B	Q1 2024
DM Math	22 GB	5.23B	-
USPTO	45 GB	4.95B	Q3 2024
PG-19	11 GB	2.63B	-
HackerNews	4.1 GB	1.08B	Q4 2023
Ubuntu IRC	4.7 GB	1.54B	Q3 2024
Europarl	6.1 GB	1.96B	-
StackExchange	79 GB	27B	Q4 2023

Source Distribution of OLMo2

OLMo2 is AI2's fully **open language models**, matching open-weight only models like Llama 3.1 and Qwen 2.5 while using fewer FLOPs and with fully transparent training data, code, and recipe.

Source	Type	Tokens	Words	Bytes	Docs
Pretraining ✦ OLMo 2 1124 Mix					
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

Table 1 Composition of the pretraining data for OLMo 2. The OLMo 2 1124 Mix is composed of StarCoder (Li et al., 2023b; Kocetkov et al., 2022), peS2o (Soldaini and Lo, 2023), web text from DCLM (Li et al., 2024) and Wiki come from Dolma 1.7 (Soldaini et al., 2024). arXiv comes from Red-Pajama (Together AI, 2023), while OpenWebMath (Paster et al., 2023) and Algebraic Stack come from ProofPile II (Azerbayev et al., 2023).

Samples of Subsets

Subset	Sample
DM-Math	the ninth root of 1961195424 to the nearest integer?\n11\nWhat is 188698044 to the power of 1/9, to the nearest integer?\n8 ...
Law	Compensation Plan is an unfunded, non-qualified deferred compensation\narrangement for non-employee members of the Board of Directors of Eastman\nChemical Company (the \"Company\"). Under this Plan ...
PubMed	Red cell associated IgG in patients suffering from Plasmodium falciparum malaria.\nQuantitation of red cell associated IgG in 62 Gambian patients with P. falciparum malaria and 23 normal adult controls ...
Paper	\\section{Introduction} \n \\subsection{Overview} \n\\label{sec:intro} \n \n\\IEEEPARstart{C}{ontent-based} image retrieval (CBIR) is a special \ncase of image classification. It can be viewed as the process of ...
Books	A Division of Simon & Schuster, Inc. \n1230 Avenue of the Americas \nNew York, NY 10020 \nwww.SimonandSchuster.com\n\nCopyright \u00a9 2002 by Charles Rosen\n\nAll rights reserved ...
StackExchange	Q: How to put link on a slideshow gallery? I am having trouble putting a link on a slideshow gallery. Slideshow block's menu doesn't have any options to put a link. \n\nSeeking a way out, I installed MetaSlider ...
Wikipedia	International Atomic Time (abbreviated TAI, from its French name) is a high-precision atomic coordinate time standard based on the notional passage of proper time on Earth's geoid. TAI is a weighted average ...
WebText	Three Ways An Search engine optimization Company Can En - 27 Nov 2018 05:51\n[[image\"/>\n<h1>Unraveling The Thriller Of Search Engine Optimisation For Your corporation Needs</h1>\n<p>I\u2019ve ...
StarCoder	#include <iostream>\n#include <unistd.h>\n#include <sys/wait.h>\n\nusing namespace std;\n\nint main()\n{\n int pid, ppid;\n\n fork();\n fork();\n fork();\n\n pid = getpid();\n ppid = getppid();\n\n ...
Patent	Graphene is an aromatic conducting polymer comprising a monolayer of sp2-bonded carbon atoms in a planar honeycomb network. Due to its properties of electrical and thermal conductivity, mechanical ...

Table 2: Samples of all subsets of the pretraining data of K2 DIAMOND.

02

Data Filtering

Comparison Across Different Sources

Dataset	Data Reading	Text Extraction	URL Filtering	Language Identification	Line Removal	PII Filtering	Exact Deduplication	Fuzzy Deduplication
TxT360	warc	trafilatura	Yes	fastText	Yes	Yes	Bloom Filter	Global
FineWeb	warc	trafilatura	Yes	fastText	Yes	Yes	n/a	Local
RefinedWeb	warc	trafilatura	Yes	fastText	Yes	No	ExactSubStr	Local
RedPajamaV2	wet	n/a	Yes	fastText	Yes	No	Bloom Filter	Local
C4	wet	n/a	No	langdetect	Yes	No	n/a	Local
Dolma	warc	?	No	fastText	Yes	Yes	Bloom Filter	Local
RedPajamaV1	wet	n/a	No	fastText	No	No	n/a	Local
The Pile	warc	jusText	No	pycld2	No	No	n/a	Global

Text Extraction

Extract **high-quality text from HTML source documents** with the help of specialized parsers

Remove boilerplate content, such as navigation menus, advertisements, and other irrelevant text.

Raw format


```
{
    "text": "Charles III Was Crowned King. But Can He Ever Be the Star? - Betteridge's Law\nBetteridge's Law\nNo.\nMay 22, 2023\nCharles III Was Crowned King. But Can He Ever Be the Star?\n",
    "url": "http://betteridgeslaw.com/2023/05/charles-iii-was-crowned-king-but-can-he-ever-be-the-star/"
}
```


Extracted format


```
{
    "text": "Betteridge's Law\nNo.\nMay 22, 2023\nCharles\nIII\nWas Crowned King. But Can He Ever Be the Star?",
    "url": "http://betteridgeslaw.com/2023/05/charles-iii-was-crowned-king-but-can-he-ever-be-the-star/"
}
```

URL Filtering and Blocklist

Potentially **harmful content** such as adult content.
Digital version of the curated data (e.g. wikipedia.org) to avoid duplication


```
{
    "text": "Enter the username or e-mail you used in your profile. A password reset link will be sent to you by e
mail.\nLoading...",
    "meta": {
        "lang": "en",
        "lang_score": 0.8463284969329834,
        "url": "https://hentaiio.com/category/xxx-anime-twitter/",
        "timestamp": "2023-01-26T21:24:07Z",
        "cc-path": "crawl-data/CC-MAIN-2023-06/segments/1674764494826.88/warc/CC-MAIN-20230126210844-2023012700084
4-00000.warc.gz"
    }
}
```

Line-level Removal

Remove lines

- do not end with a terminal punctuation mark (i.e., “.”, “?”, “!”, or “"”).
- only composed of uppercase characters or only numeric characters
- contain only one word.


```
{
    "text": "Search\nAdd filters:\nUse filters to refine the search results.\nResults 1-1 of 1 (Search time: 0.0 seconds).\n- previous\n- 1\n- next\nItem hits:\n|Issue Date\n|Title\n|Author(s)\n|2008\n|The Spaghetti House Delivery\n|Dew, Wai Leung (刁偉樑); Heung, Shun Wai (香信瑋); Hui, Hiu Ching (許曉貞); Lam, Hoi Yin (林愷然); Lung, Fei Wan (龍飛雲)\nDiscover\nAuthor\nHas File(s)\n- 1 true",
    "meta": {
        "lang": "en",
        "lang_score": 0.6708550453186035,
        "url": "http://dspace.cityu.edu.hk/handle/2031/24/simple-search?query=&sort_by=score&order=desc&rpp=10&filter_field_1=subject&filter_type_1=equals&filter_value_1=Food+service+--+China+--+Shenzhen+Shi+--+Marketing&filter_field_2=author&filter_type_2=equals&filter_value_2=Lung%2C+Fei+Wan+%28%E9%BE%8D%E9%A3%9B%E9%9B%B2%29&filter_field_3=author&filter_type_3=equals&filter_value_3=Lam%2C+Hoi+Yin+%28%E6%9E%97%E6%84%B7%E7%84%B6%29&filter_field_4=author&filter_type_4=equals&filter_value_4=Hui%2C+Hiu+Ching+%28%E8%A8%B1%E6%9B%89%E8%B2%9E%29&etal=0&filtername=dateIssued&filterquery=2008&filtertype=equals",
        "timestamp": "2023-11-28T09:25:19Z",
        "cc-path": "crawl-data/CC-MAIN-2023-50/segments/1700679099281.67/warc/CC-MAIN-20231128083443-20231128113443-00000.warc.gz"
    },
```

Document-Level Filtering

Remove documents

- with excessive line, paragraph, or n-gram repetitions.
- containing multiple, short duplicate passages, as well as those with few, but longer duplicate passages.


```
{
    "text": "This article is made up of亚博体育，AI learns through the Internet and automatically writes, does not represent our position, reprinted, contact the author and indicate the source: http://www.afconthefield.com/292619lk.html\nThis article is made up of亚博体育，AI learns through the Internet and automatically writes, does not represent our position, reprinted, contact the author and indicate the source: http://www.afconthefield.com/292619lk.html",
    "meta": {
        "lang": "en",
        "lang_score": 0.726877748966217,
        "url": "http://afconthefield.com/292619lk.html",
        "timestamp": "2023-01-26T22:53:49Z",
        "cc-path": "crawl-data/CC-MAIN-2023-06/segments/1674764494826.88/warc/CC-MAIN-20230126210844-20230127000844-00000.warc.gz"
    },
```

Statistics-based Heuristics

Remove any document which satisfies any of the following criteria:

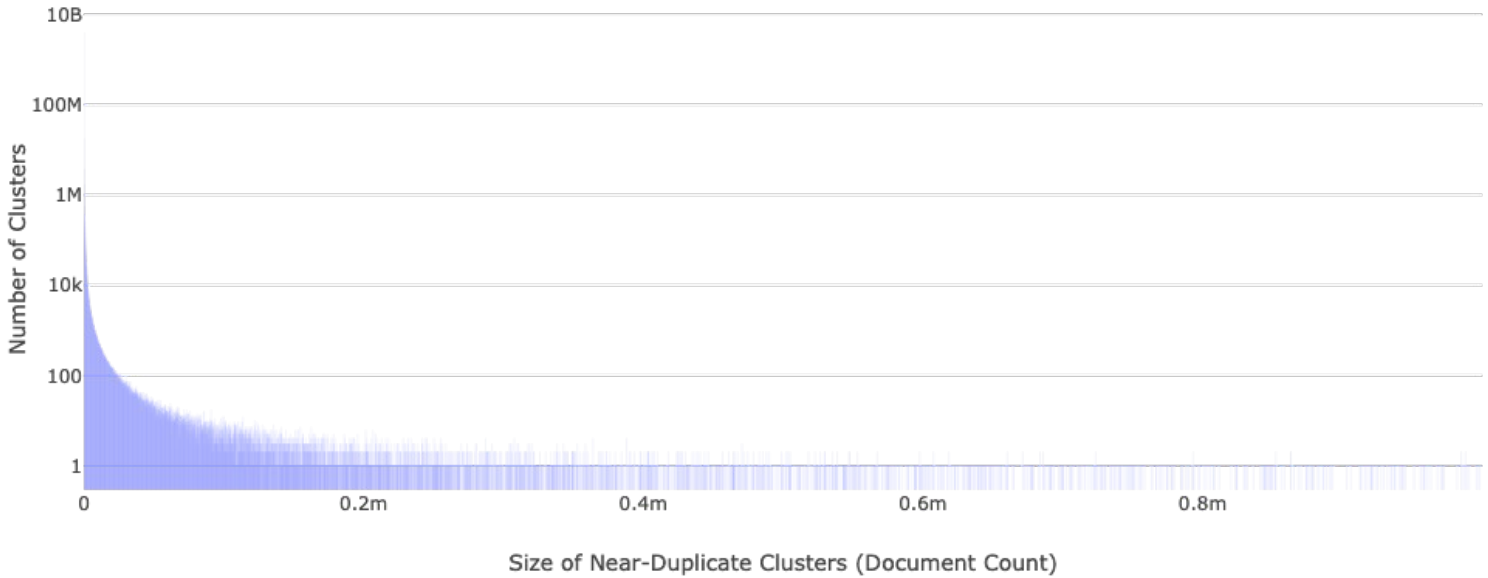
- less than 50 words or more than 100,000 words
- its mean word length is outside the range of 3 to 10
- it contains less than 3 sentences
- its symbol-to-word ratio is greater than 0.1
- the words that contain at least one alphabetic character are less than 80% of the whole words
- it contains less than two of the stop words (the, be, to, of, and, that, have, with


```
{
    "text": "You should open your own Breakfast-Diner. I would be your everyday-guest, even If I had to move away from germany for that :D\nmums mums mums!\nIvy: Hahaha! Good idea. :)Melwa: :)\nSkicka en kommentar",
    "meta": {
        "lang": "en",
        "lang_score": 0.8643032908439636,
        "url": "http://365daysofbreakfast.blogspot.com/2011/07/monday.html",
        "timestamp": "2023-11-28T09:40:02Z",
        "cc-path": "crawl-data/CC-MAIN-2023-50/segments/1700679099281.67/warc/CC-MAIN-20231128083443-20231128113443-00000.warc.gz",
        "url_score": 0.0
    },
```

02

Deduplication

Deduplication



By implementing deduplication and selective upsampling, we gain control over the pretraining data distribution, rather than relying on the inherent distribution of the source.

Deduplication

The near-deduplication process involves

1. generating signatures for every document,
2. matching these signatures to identify near-duplicates,
3. clustering the near-duplicate documents to select all but one for deletion.

1 Questions about this plan? Visit Houseplans.com today or call 1-800-913-2350\nView this plan at https://www.houseplans.com/plan/4284-square-feet-4-bedrooms-4-bathroom-european-house-plans-3-garage-11520\nView this plan at https://www.houseplans.com/plan/4284-square-feet-4-bedrooms-4-bathroom-european-house-plans-3-garage-11520\nIn addition to the house plans you order, you may also need a site plan that shows where the house is going to be located on the property. You might also need beams sized to accommodate roof loads specific to your region. Your home builder can usually help you with this. You may also need a septic design unless your lot is served by a sanitary sewer system. Many areas now have area-specific energy codes that also have to be followed. This normally involves filling out a simple form providing documentation that your house plans are in compliance.\nTo find out what documents you should expect with your house plans, see https://www.houseplans.com/whats-included.\nIn some regions, there is a second step you will need to take to insure your house plans are in compliance with local codes. Some areas of North America have very strict engineering requirements. Examples of this would be earthquake-prone areas of California and the Pacific Coast, hurricane risk areas of the Florida, Gulf & Carolina Coasts. New York, New Jersey, Nevada, and parts of Illinois require review by a local professional as well. If you are building in these areas, it is most likely you will need to hire a state licensed structural engineer to analyze the design and provide additional drawings and calculations required by your building department. If you aren't sure, building departments typically have a handout they will give you listing all of the items they require to submit for and obtain a building permit.\nAdditionally, stock plans do not have a professional stamp attached. If your building department requires one, they will only accept a stamp from a professional licensed in the state where you plan to build. In this case, you will need to take your house plans to a local engineer or architect for review and stamping. In addition, plans which are used to construct homes in Nevada are required to be drawn by a licensed Nevada architect.\nNote: All sales on house plans are final. No refunds or exchanges can be given once your order has been fulfilled or once we have begun to customize a home plan to your specifications.

1 View plan at https://www.dreamhomesource.com/plan/1622-square-feet-2-bedroom-1-50-bathroom-0-garage-adobe-southwestern-mediterranean-sp123639\nView this plan at https://www.dreamhomesource.com/plan/1622-square-feet-2-bedroom-1-50-bathroom-0-garage-adobe-southwestern-mediterranean-sp123639\nIn addition to the house plans you order, you may also need a site plan\nthat shows where the house is going to be located on the property. You\nmight also need beams sized to accommodate roof loads specific to your\nregion. Your home builder can usually help you with this. You may also\nneed a septic design unless your lot is served by a sanitary sewer system.\nMany areas now have area-specific energy codes that also have to be\nfollowed. This normally involves filling out a simple form providing\ndocumentation that your house plans are in compliance.\nTo find out what documents you should expect with your house plans, see\nIn some regions, there is a second step you will need to take to insure\nyour house plans are in compliance with local codes. Some areas of North\nAmerica have very strict engineering requirements. Examples of this would\nbe earthquake-prone areas of California and the Pacific Coast, hurricane\nrisk areas of the Florida, Gulf & Carolina Coasts. New York, New Jersey,\nNevada, and parts of Illinois require review by a local professional as\nwell. If you are building in these areas, it is most likely you will need\nto hire a state licensed structural engineer to analyze the design and\nprovide additional drawings and calculations required by your building\ndepartment. If you aren't sure, building departments typically have a\nhandout they will give you listing all of the items they require to submit\nfor and obtain a building permit.\nAdditionally, stock plans do not have a professional stamp attached. If\nyour building department requires one, they will only accept a stamp from\na professional licensed in the state where you plan to build. In this\ncase, you will need to take your house plans to a local engineer or\narchitect for review and stamping. In addition, plans which are used to\nconstruct homes in Nevada are required to be drawn by a licensed Nevada\nNote: All sales on house plans are final. No refunds or exchanges can\nbe given once your order has been fulfilled or once we have begun to\ncustomize a home plan to your specifications.

Personally Identifiable Information Removal

- **Information that can be used to identify an individual,** such as names, addresses, phone numbers, email addresses, and social security numbers.
- **Reduce the risk of data breaches** and unauthorized access to sensitive information.
- Prevents the models generating that specific PII during inference time.

PII Type	Examples	Target
Email	john.doe@llm360.ai	firstname.lastname@example.com
IP Address	172.217.164.110	[22.214.171.124 , ...]

Pretraining Architecture

I. The New Tokenizer: Byte Pair Encoding
II. Core LLM Architecture
III. Pretraining Workflow
IV. Advanced Pretraining: Mixture of Experts

01

The New Tokenizer

Byte Pair Encoding

Initial vocabulary:
characters

Split each word
into characters

Words in the data:

word	count
c a t	4
m a t	5
m a t s	2
m a t e	3
a t e	3
e a t	2

Current merge table:

(empty)

Special Tokens

Special Token	Usage
<phone_number> <email_address> <ip_address>	Mask personal identifiable information.
<jupyter_code> <issue_start>	Handle diverse data from GitHub.
<fim_suffix> <fim_middle>	For fill-in-the-middle (FIM) code generation tasks.

02

Core LLM Architecture

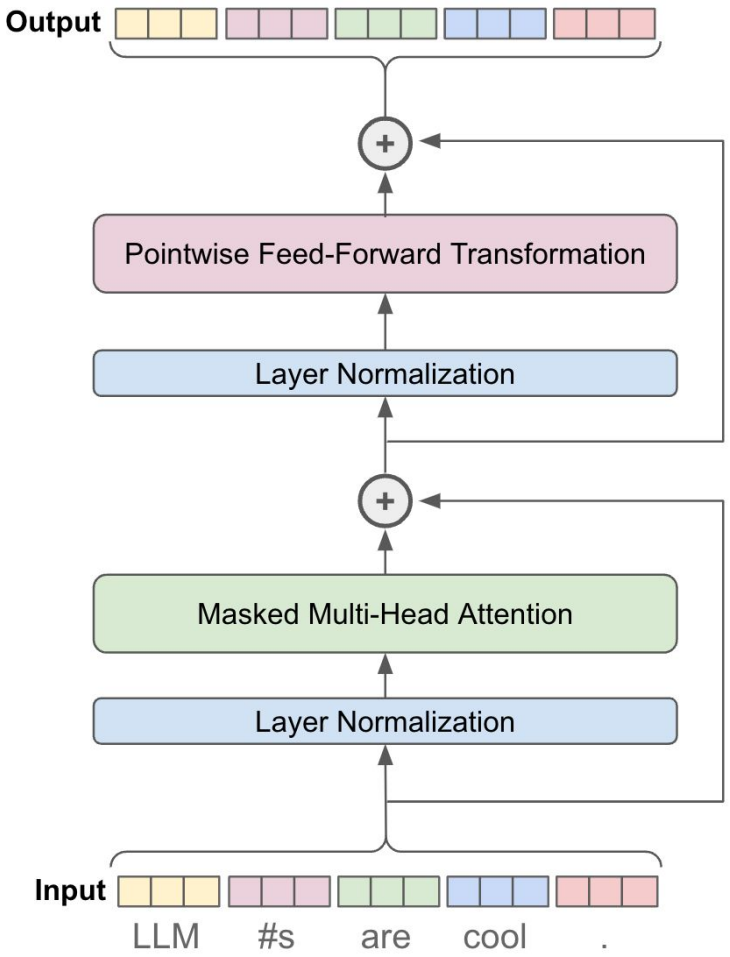
Decoder-Only Transformer Block

Masked Self-Attention

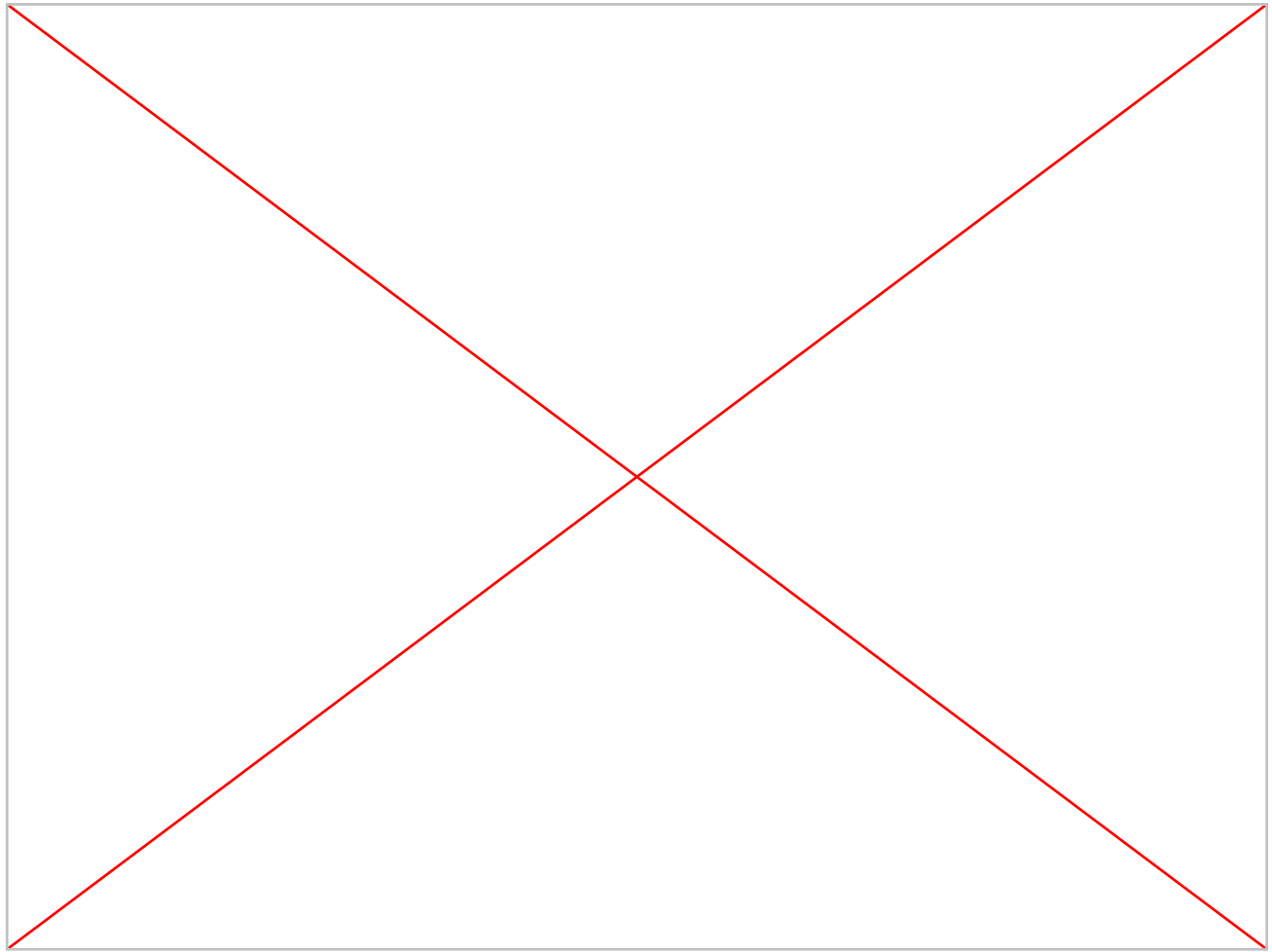
Let's consider our token sequence ["LLM", "#s", "are", "cool", "."] and assume we are trying to compute attention scores for the token "are".

So far, we have learned that self-attention will compute an attention score between "are" and every other token within the sequence. With masked self-attention, however, we only compute attention scores for "LLM", "#s", and "are".

Masked self-attention prohibits us from looking forward in the sequence during self-attention.



Masked Self-Attention



Context Length: Rotary Positional Embeddings

Rotary Positional Embeddings has been utilized to extend long-context capability.

Let us watch a quick video to understand Rotary Positional Embeddings.

03

Pre-training Workflow

Training Setup

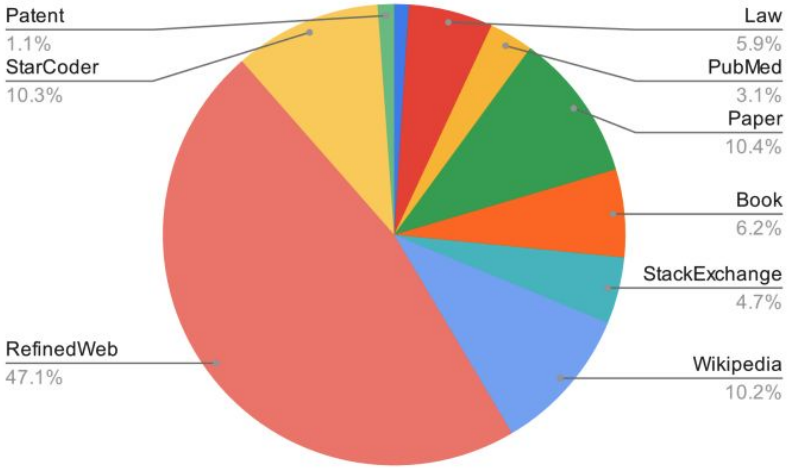
- **Warm-up** at the beginning, gradually increasing the learning rate to the target level
- **Middle stage**: use cos / cos_decay / constant / constant_decay, with a relatively large learning rate. Whether to apply decay depends; could refer to others' technical reports or experiment on small-scale training
- **Later stage:** extend training sequence length, and change the frequency of RoPE to make the model adapt to longer texts

In summary, pretraining usually follows a two-stage or multi-stage training process: First train on full corpus, then train on small-scale long-text data, and finally conduct annealing on small-scale high-quality data.

Training Setup

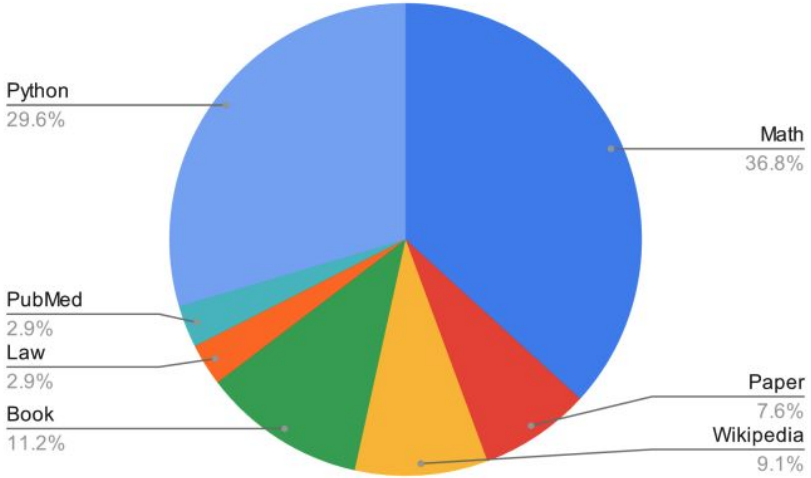
Major Stage

Involves training on a diverse corpus of trillion tokens drawn from a variety of sources, including web texts, academic papers, books, code, as well as mathematical, medical, and legal documents.



Long-context Stage

Enhance the model's generation abilities, such as arithmetic reasoning and coding, as well as expanding the context length.



04

Advanced Pre-training: Mixture of Experts

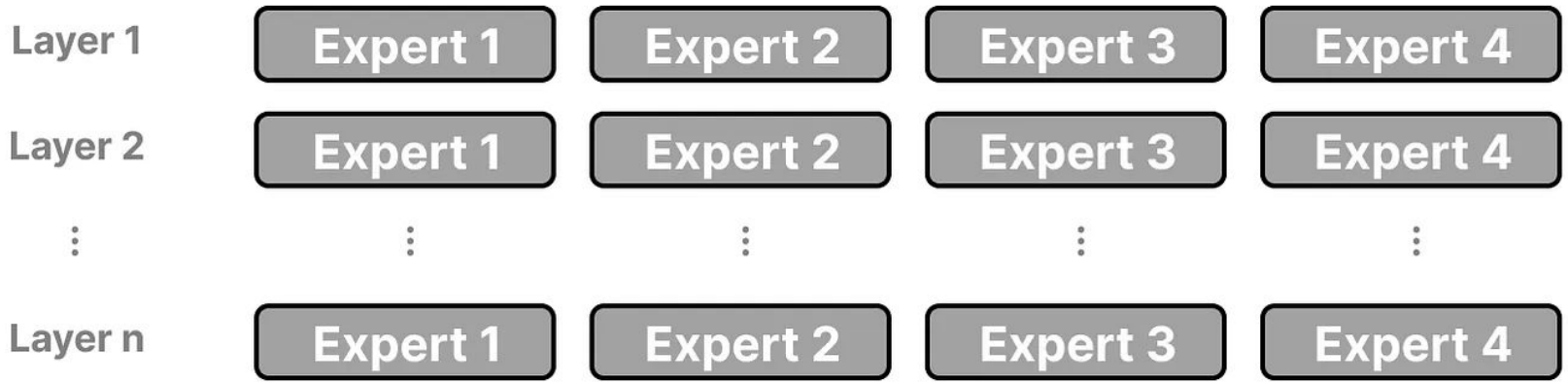
Mixture of Experts

Experts

Each Feed Forward Network (FFN) layer now has a set of “experts” of which a subset can be chosen. These “experts” are typically FFNs themselves.

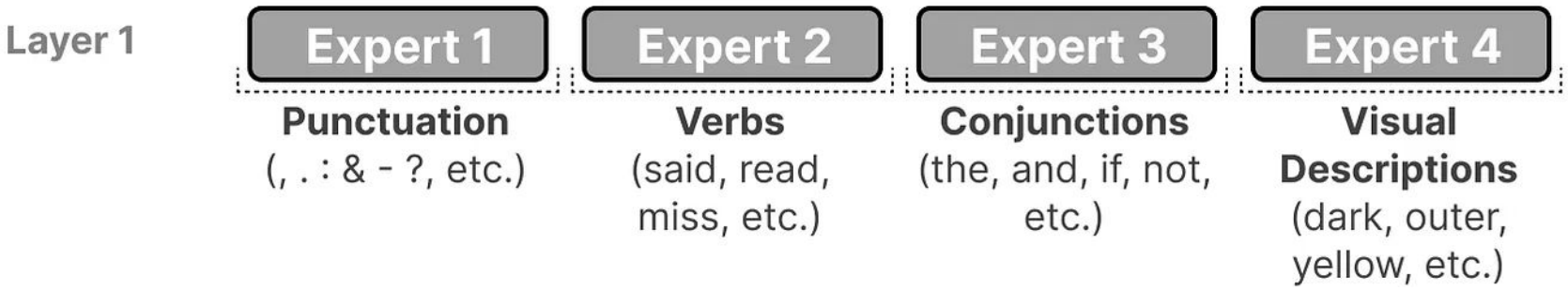
Router or Gate Network

Determines which tokens are sent to which experts.



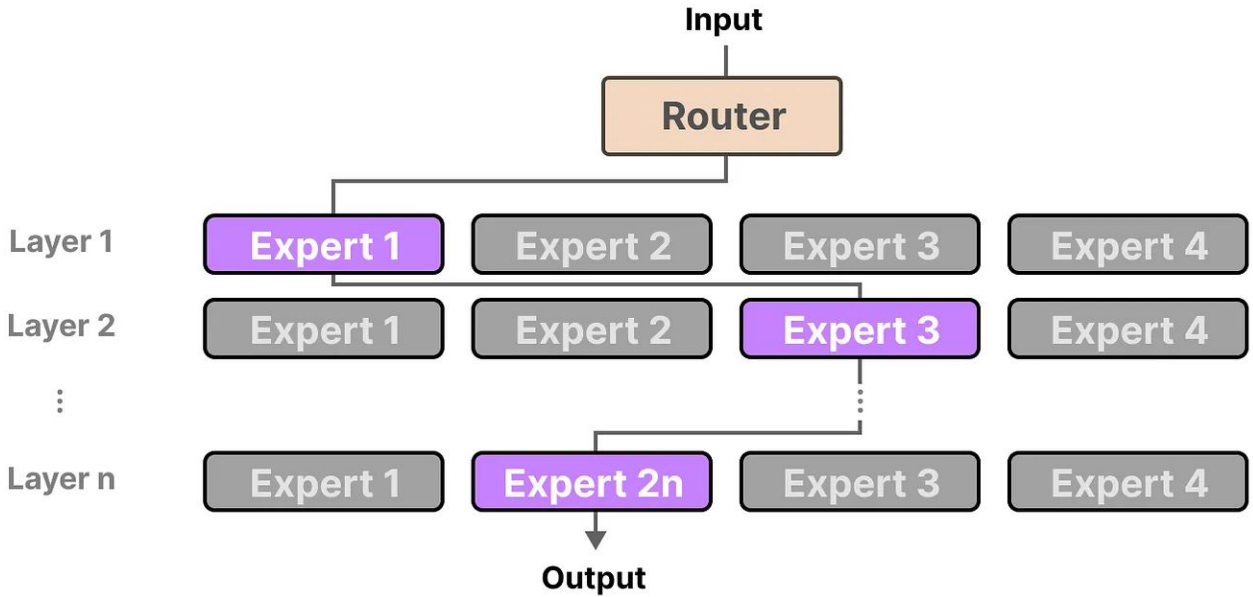
Experts

An **expert** is not specialized in a specific domain like “Psychology” or “Biology”.
At most, it learns **syntactic information on a word level** instead.

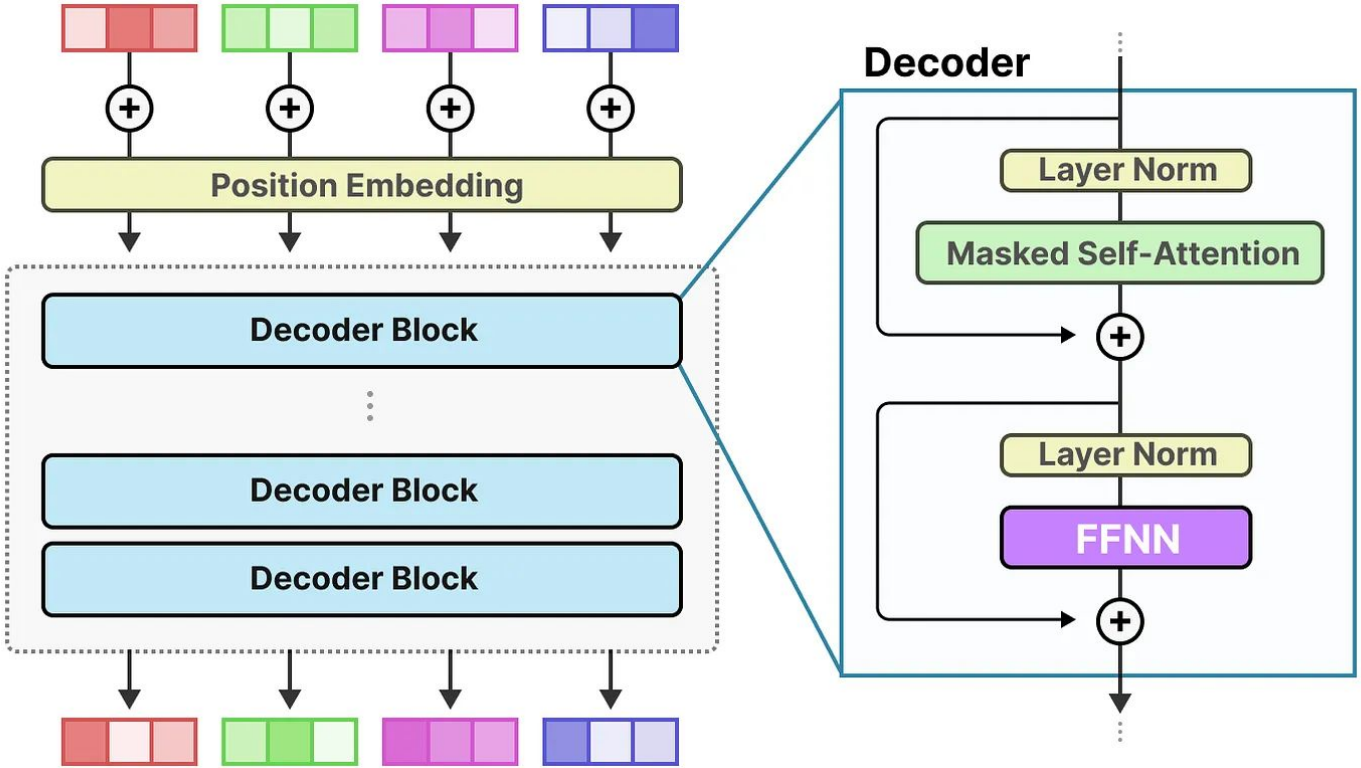


Router

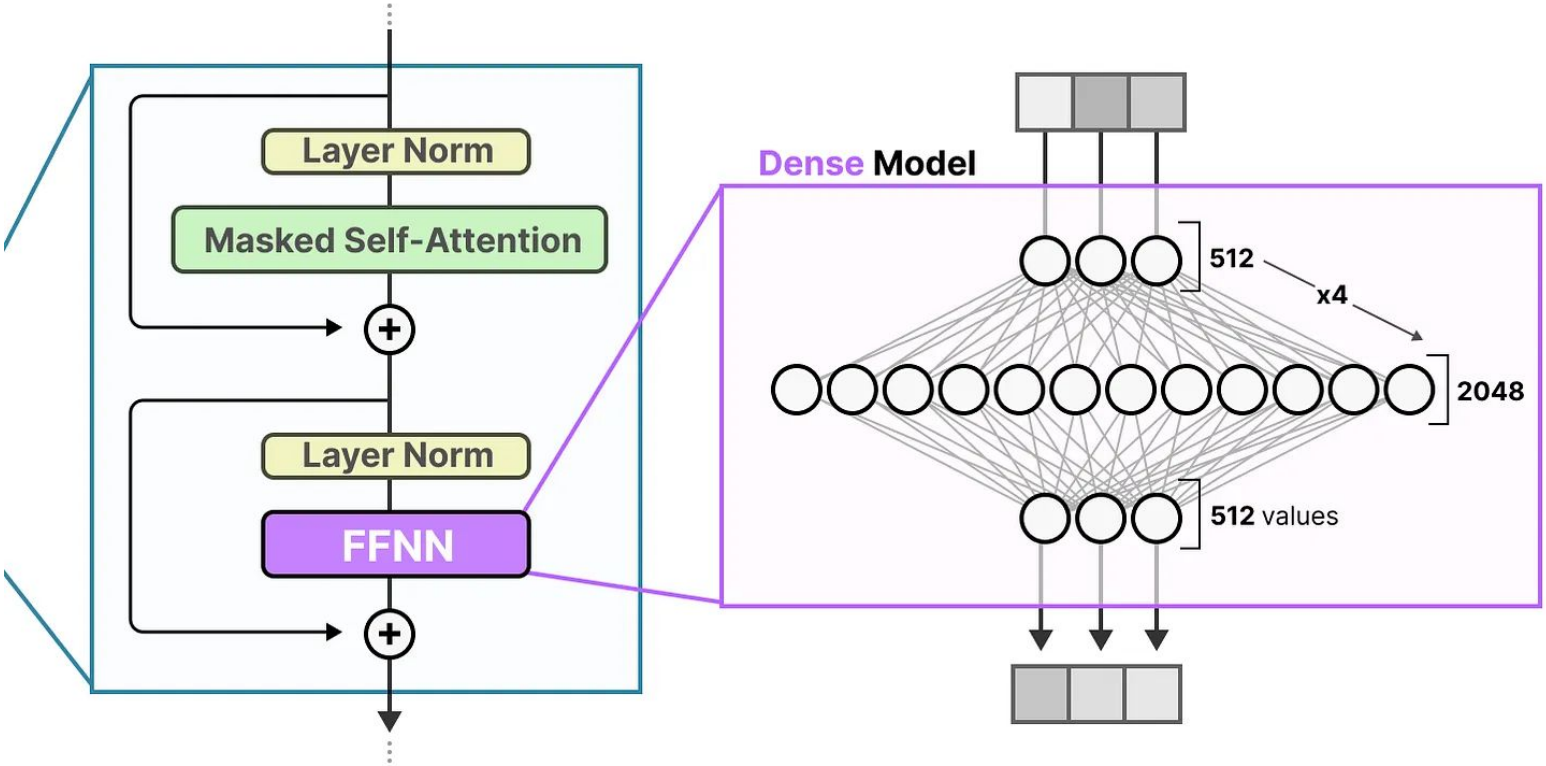
The router (gate network) **selects the expert(s) best suited** for a given input.



What do the Experts Replace?



What do the Experts Replace?

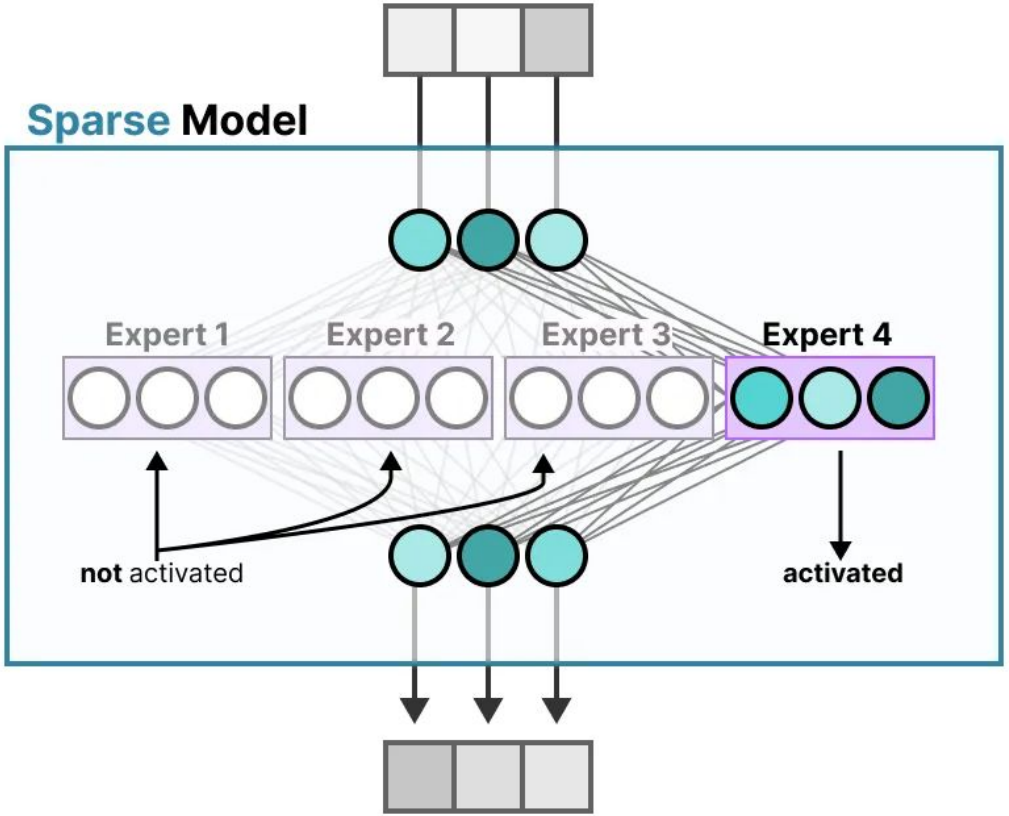


Sparse Model

We can chop up our dense model into pieces (so-called experts), retrain it, and only activate a subset of experts at a given times.

Sparse Models only **activate a portion of their total parameters.**

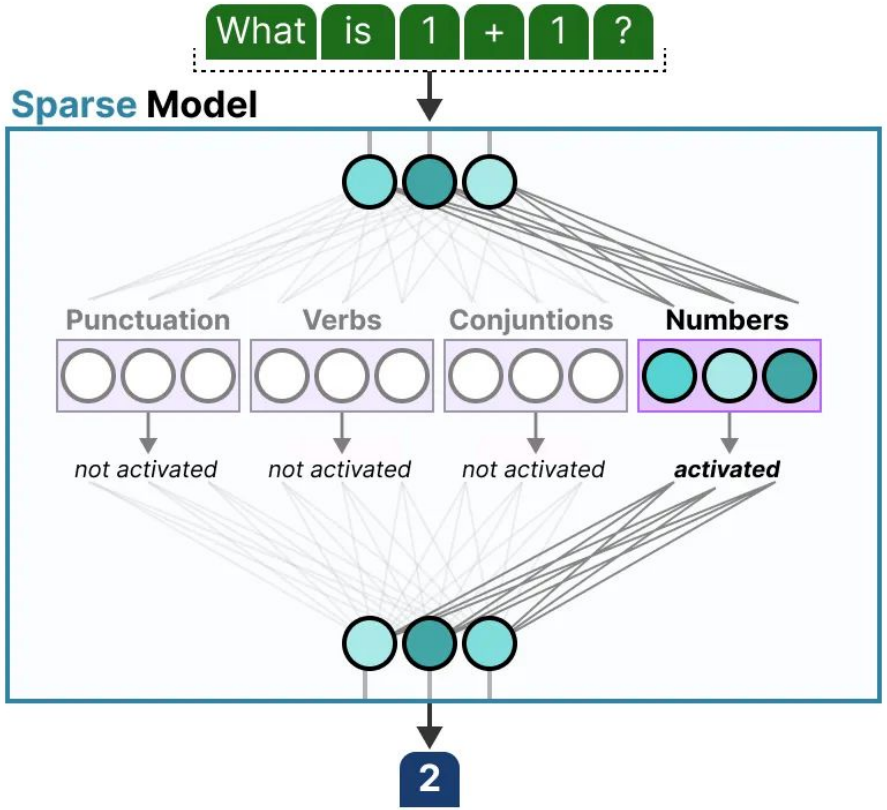
They are closely related to Mixture of Experts.



Sparse Model

The underlying idea is that each expert learns different information during training.

Then, **when running inference, only specific experts are used as they are most relevant** for a given task.



When asked a question, we can select the expert best suited for a given task.

What do Experts Learn?

Expert specialization	Expert position	Routed tokens
Punctuation	Layer 2	, , , , , , , , ,- , , , , ,).)
	Layer 6	, , , , , : . : , & , & & ? &-,,? ,,, .
Conjunctions and articles	Layer 3	The the the the the the the the The...
	Layer 6	a and and and and and and or and ...
Verbs	Layer 1	died falling identified fell closed left posted lost felt left said read miss place struggling falling signed died...
Visual descriptions *color, spatial position*	Layer 0	her over her know dark upper dark outer center upper blue inner yellow raw mama bright bright over open your dark blue
Counting and numbers *written and numerical forms*	Layer 1	after 37 19. 6. 27 I I Seven 25 4, 54 I two dead we Some 2012 who we few lower

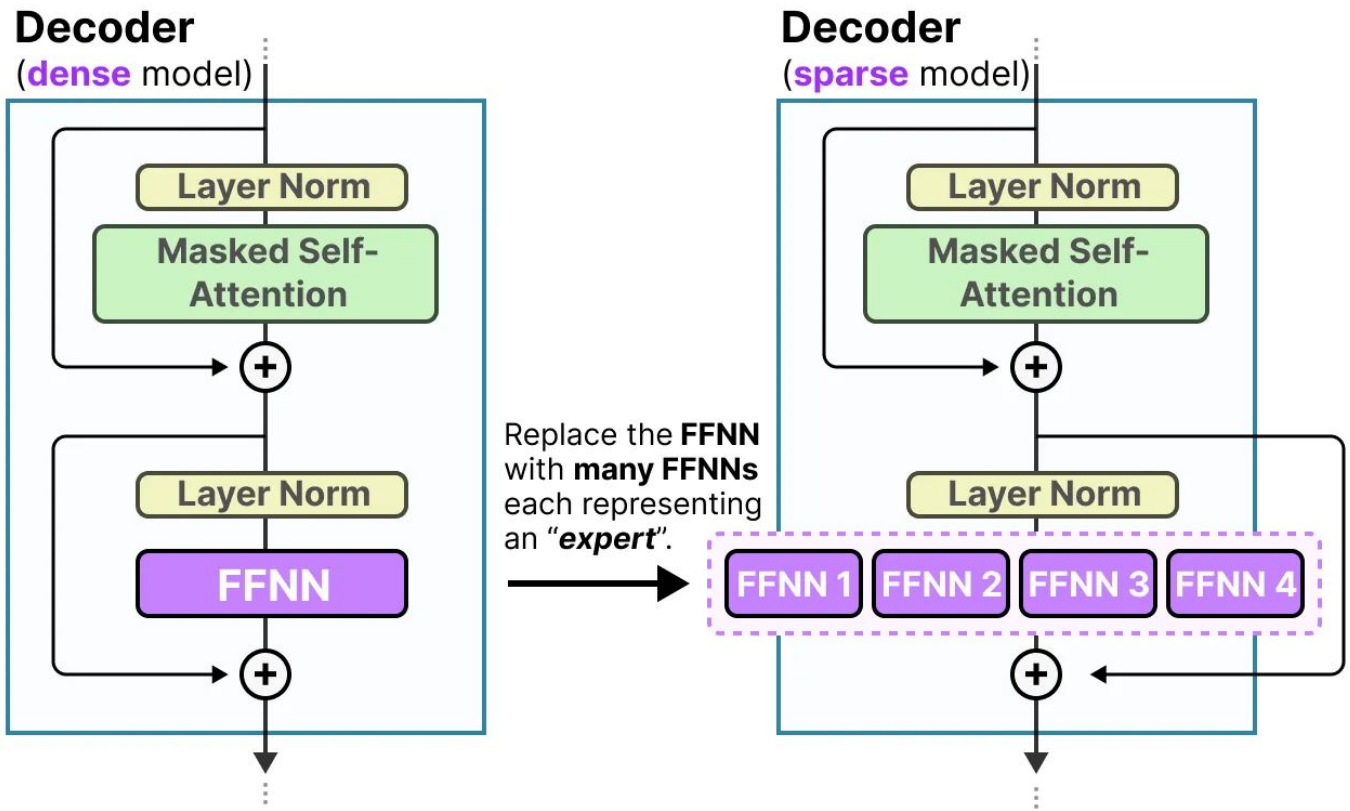
Expert specialization of an encoder model in the ST-MoE paper.

What do Experts Learn?

Example from Mixtral 8x7B paper where each token is colored with the first expert choice.

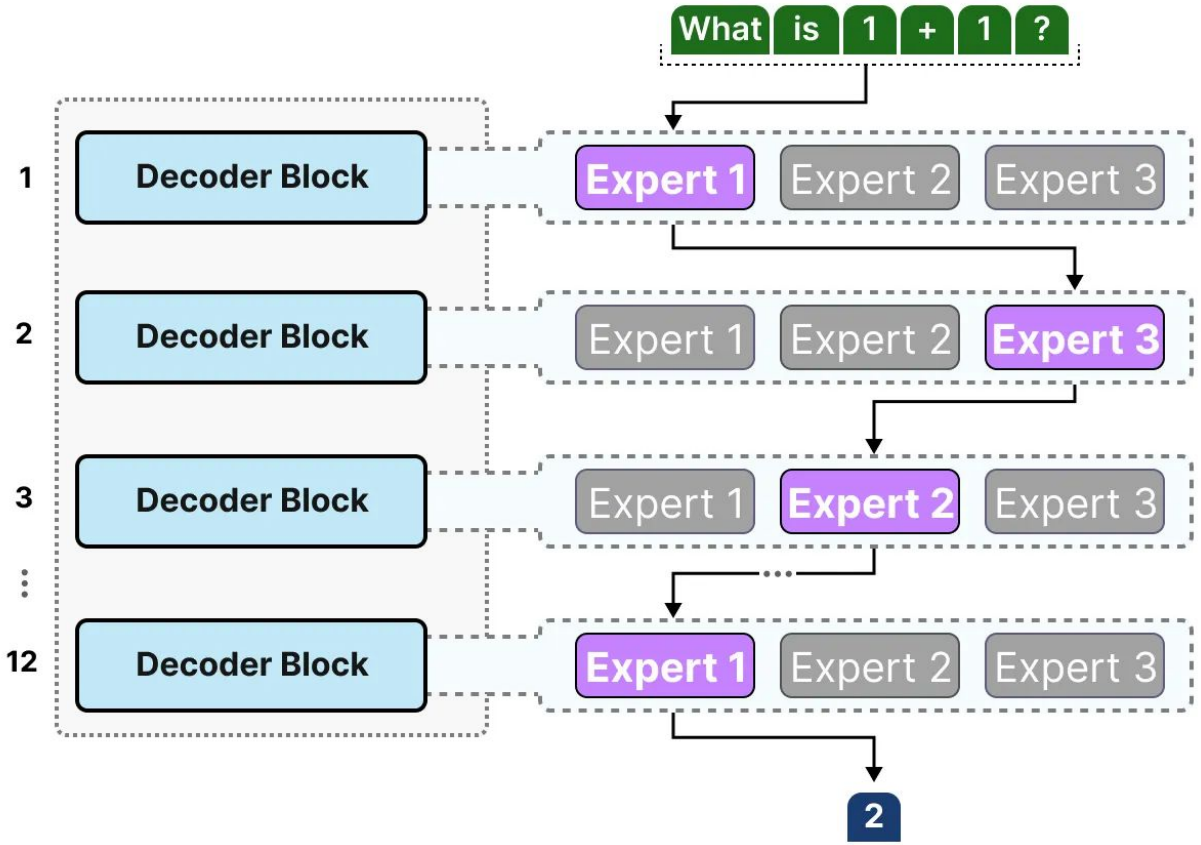
```
class MoeLayer(nn.Module):
    def __init__(self, experts: List[nn.Module],
        super().__init__()
        assert len(experts) > 0
        self.experts = nn.ModuleList(experts)
        self.gate = gate
        self.args = moe_args
```

The Updated Decoder



Architecture of Experts

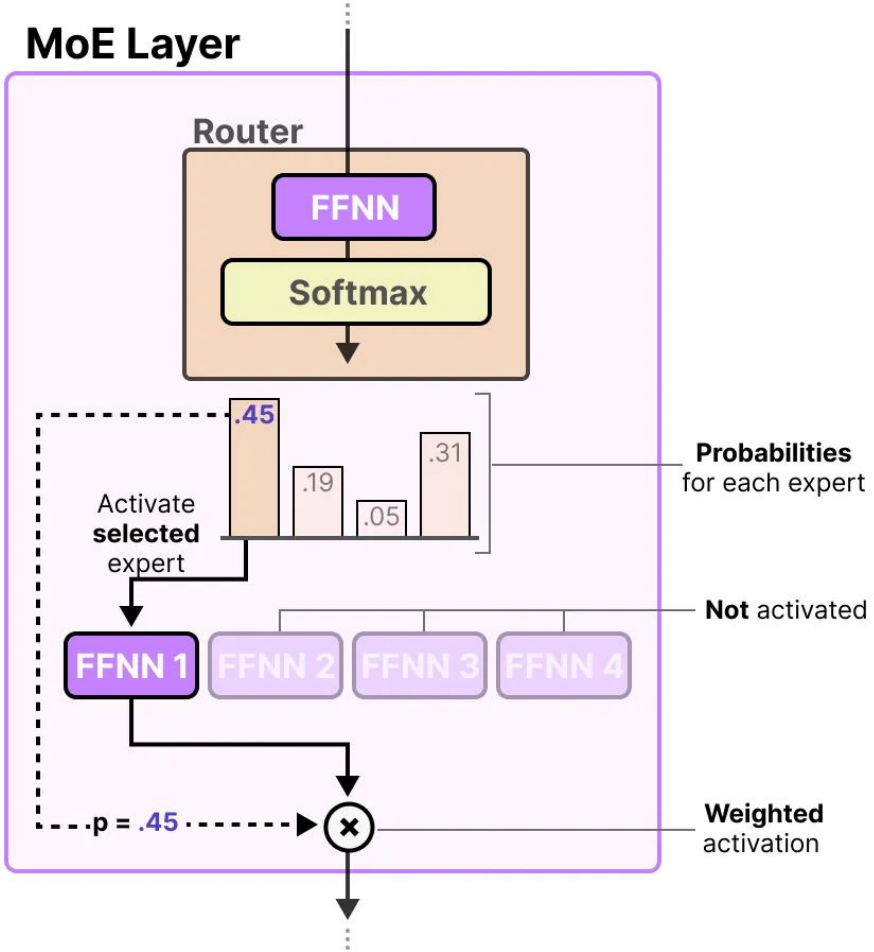
Since most LLMs have several decoder blocks, a **given text will pass through multiple experts** before the text is generated.



The Router

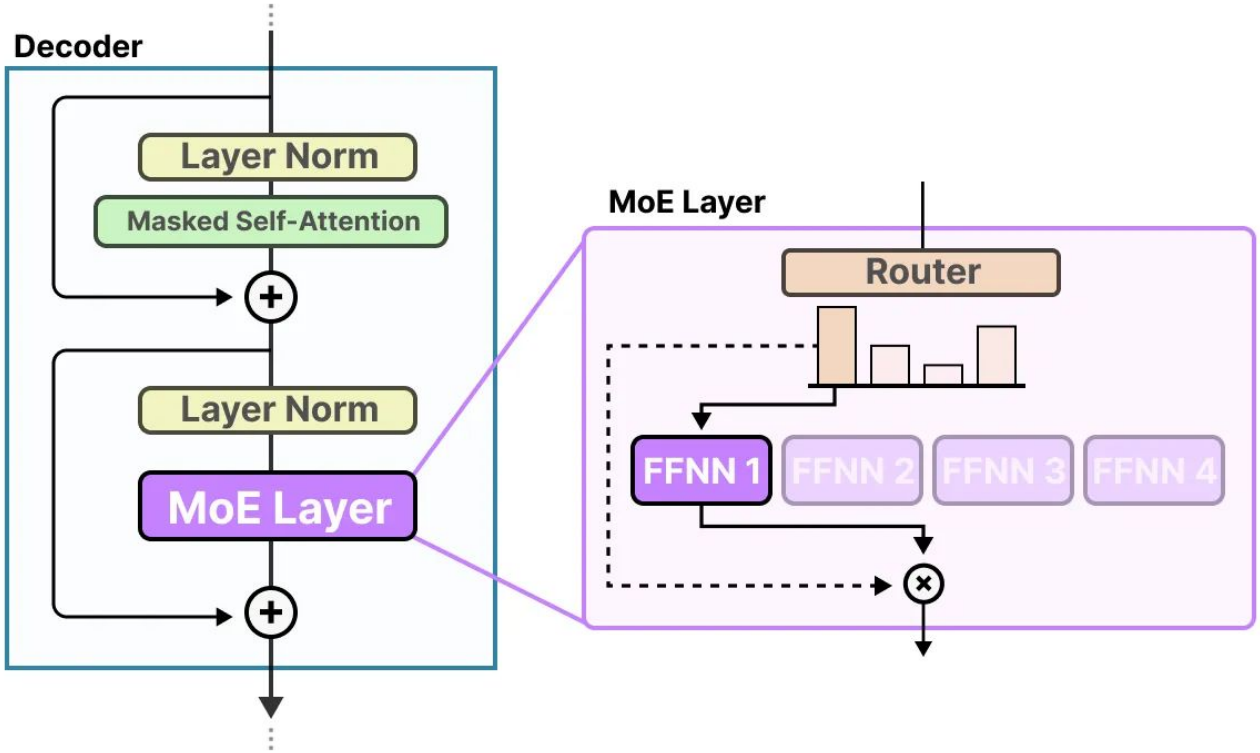
The router is also an FFNN and is used to choose the expert based on a particular input.

It **outputs probabilities which it uses to select the best matching expert.**

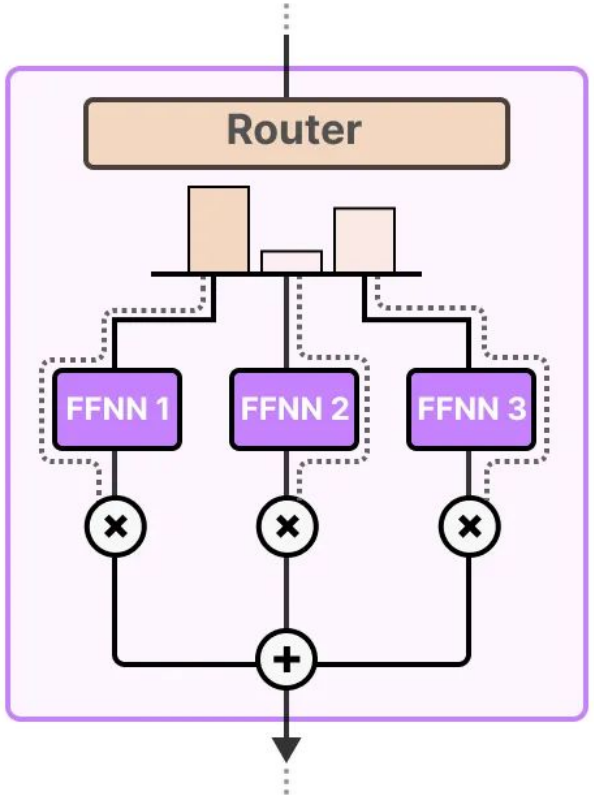


The MoE Layer

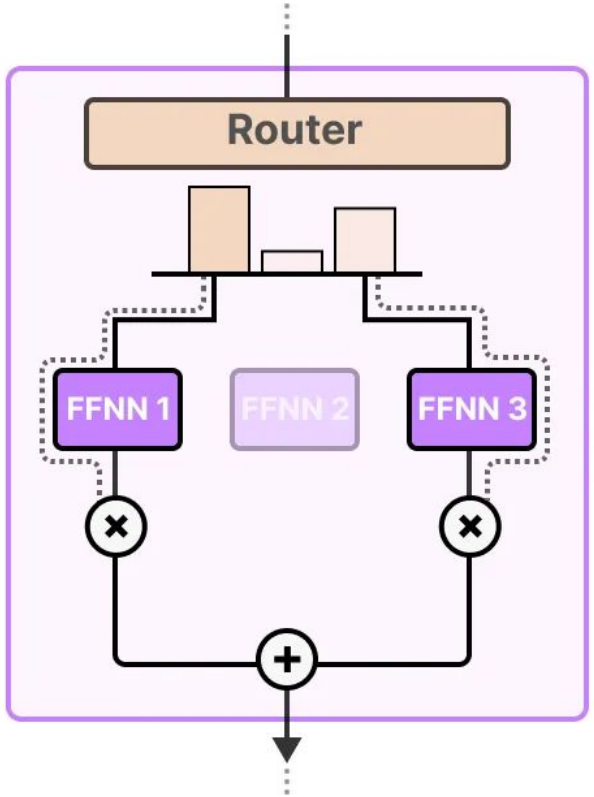
The router together with the experts makes up the **MoE Layer**



Sparse vs. Dense MoE Layer



Dense MoE



Sparse MoE

“

Assume we have a 7B parameter LLM and replace each of its feed-forward sub-layers with an MoE layer comprised of eight experts, where two experts are activated for each token[1].

The full model has about 47B parameters, all of which must be loaded into memory.

But, the model’s inference cost is comparable to a 14B parameter model. Only two experts are used to process each token, which yields ~2 x 7B matrix multiplications.

In this way, we achieve the capacity of a ~50B parameter model without incurring the cost!

[1] This is the exact architecture that was used for Mixtral-8x7B [13], an MoE variant of Mistral-7B [14]

Pretraining Evaluation

I. Perplexity
II. Few Shot Evaluation
III. Completion Ability

01 Perplexity

Perplexity (PPL) is just the **exponential form of the cross entropy loss.**

$$\text{Perplexity}(Model) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_1, \ldots, w_{i-1})\right)$$

PPL only reflects the model's performance of its own kind. Due to differences in tokenizer compression rates, the PPL cannot be compared directly between models if the tokenizers differ. For example, in Chinese, a tokenizer with only single characters as tokens without involving expressions always leads to the lowest loss/PPL.

On the other hand, regardless of the tokenizer compression rate, the l**oss on general knowledge test sets should drop below a certain threshold generally**; otherwise, it indicates the training is insufficient.

02

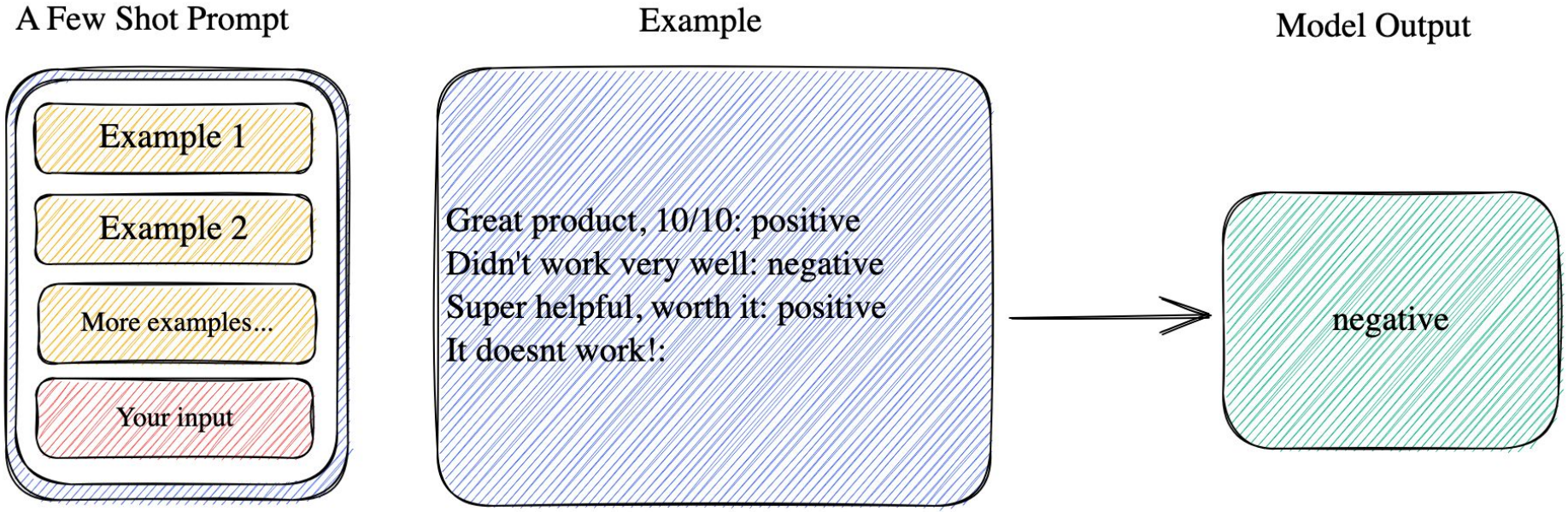
Few Shot Evaluation

Knowledge Benchmarks

Name	Purpose	Paper Link
MMLU	Evaluate LLM's comprehension and reasoning across a wide range of subjects	*Measuring Massive Multitask Language Understanding*
GLUE	Comprehensive assessment of language understanding in various scenarios	*GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*
MultiNLI	Evaluate LLM's ability to correctly perform classification based on premise	*A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference*
SuperGLUE	Evaluate deeper language comprehension and reasoning	*SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*

Few Shot Evaluation

A **pretrained model does not follow the answer format** usually.
We use a few-shot approach to show it some examples, so that it knows what to say next.



03

Completion Ability

Prepare the prompt and ground truth.
Given prompt, ask the base model to generate text.
Calculate the similarity between the generated text and the ground truth.

Metric	Calculation	Function
Rouge-L	Looks at the longest common subsequence between two texts.	Focuses more on sentence structure similarity
BLEU	Calculates whether n-grams (contiguous word sequences) match.	Commonly used in machine translation evaluation
BERTScore	Uses the BERT model to understand semantics first, then calculates similarity.	Semantic level accuracy.

Pretraining Recipes of Different LLMs

4

No training pipeline is quite the same…

Pre-training summary

Qwen 2

- Filtering ✓
- Synthetic data ✓
- Mixing ✓
- Q&A format ✓
- Long-context stage ✓
- Continued pre-training ✓
- High-quality stage ✓
- Knowledge distillation ✗

AFM

- Filtering ✓
- Synthetic data ✓
- Mixing ✓
- Q&A format ✓
- Long-context stage ✓
- Continued pre-training ✓
- High-quality stage ✓
- Knowledge distillation ✓

Gemma 2

- Filtering ✓
- Synthetic data ✗
- Mixing ✓
- Q&A format ✗
- Long-context stage ✗
- Continued pre-training ✗
- High-quality stage ✗
- Knowledge distillation ✓

Llama 3.1

- Filtering ✓
- Synthetic data ✓
- Mixing ✓
- Q&A format ✗
- Long-context stage ✓
- Continued pre-training ✓
- High-quality stage ✓
- Knowledge distillation ✗

References

1. [1701.06538] Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer
2. [2407.06204] A Survey on Mixture of Experts in Large Language Models
3. Seq2seq and Attention
4. TxT360: Trillion Extracted Text - a Hugging Face Space by LLM360
5. A Visual Guide to Mixture of Experts (MoE)
6. Mixture-of-Experts (MoE): The Birth and Rise of Conditional Computation

Thank you.

Nikita Saxena
Research Engineer

Google DeepMind