Harnessing Large Datasets for Large Language Models

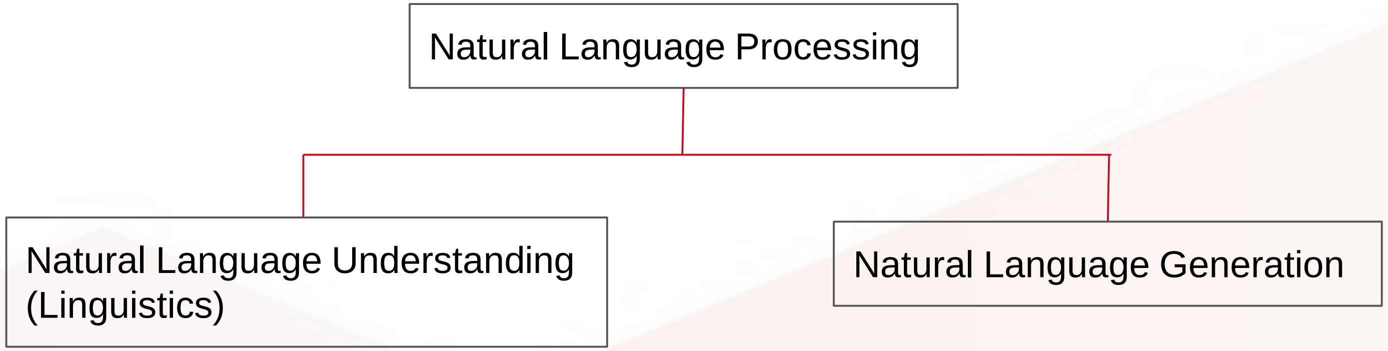
Presented at: 2nd European Summer School on Artificial Intelligence (ESSAI 2024), Athens, Greece
Date of event: 23rd July 2024
Presented by: Jennifer D'Souza (Junior AI Research Group Lead)
Contact: https://www.linkedin.com/in/jennifer-l-dsouza/

Natural Language Processing

- Natural Language Processing (NLP) is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages. It came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language.
- It can be classified into two parts:
    - Natural Language Understanding or Linguistics and
    - Natural Language Generation which evolves the task to understand and generate the text.

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing



NLU enables machines to understand natural language and analyze it by extracting concepts, entities, emotion, keywords etc. However such functions are predicated on basic building blocks of Linguistics which we will gloss over next.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing

Natural Language Processing

Natural Language Understanding (Linguistics)

Natural Language Generation

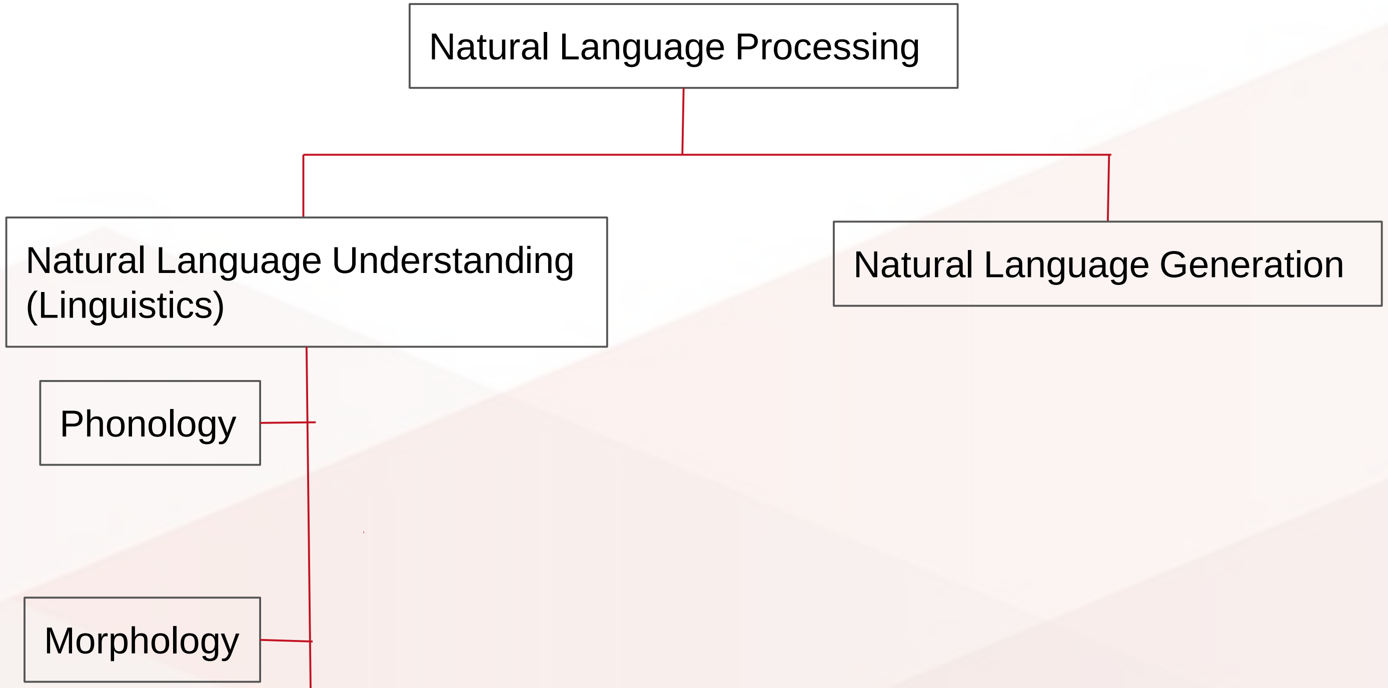
Phonology

Phonology is the study of how the sounds in a language work, including their rules, patterns, and how they change in different situations. E.g., consider the sound of “c” in different contexts

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing



Morphology is like the "building blocks" of words, where we study how words are formed from smaller parts (called morphemes) to understand their meanings and how they can change. E.g., the word "unhappiness" is made up of three morphemes: "un-" (a prefix meaning "not"), "happy" (the root word), and "-ness" (a suffix that turns an adjective into a noun).

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing

Natural Language Processing

Natural Language Understanding (Linguistics)

Natural Language Generation

Phonology

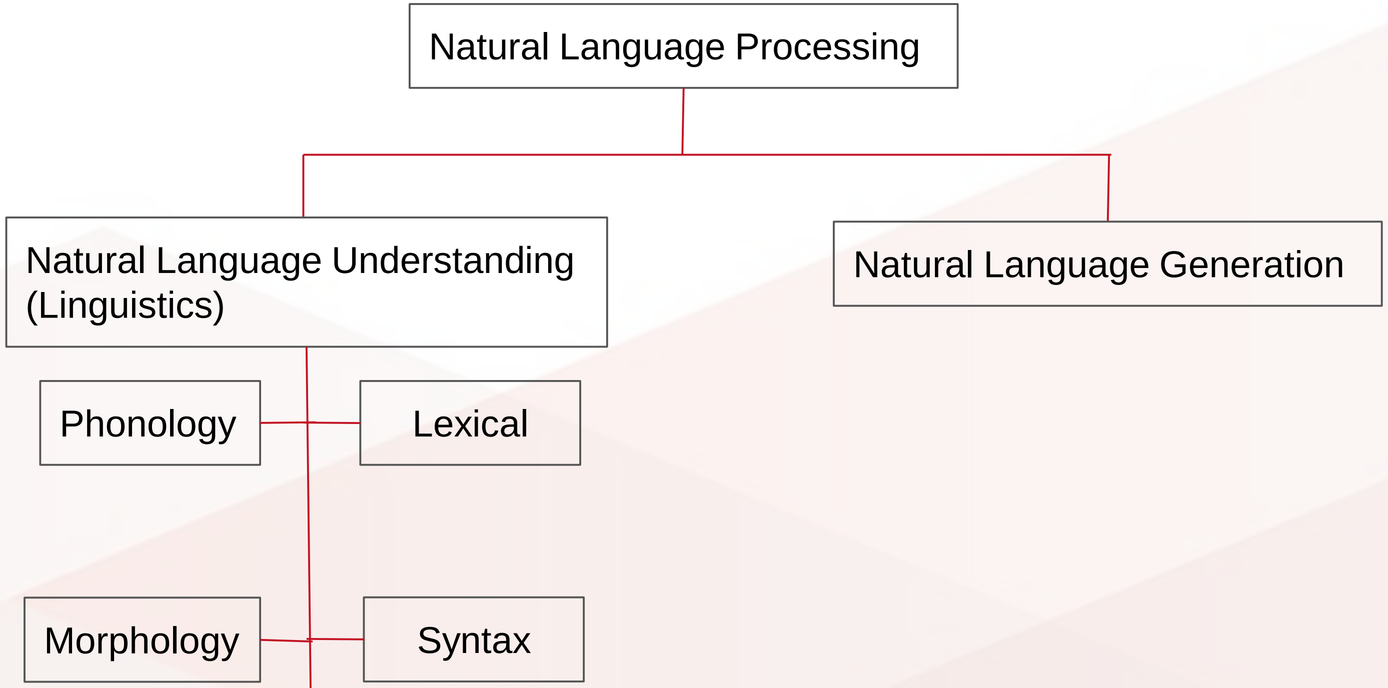
Lexical

Lexical, in simple terms, refers to the words or vocabulary of a language, including their meanings and how they are used in sentences.

Morphology

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

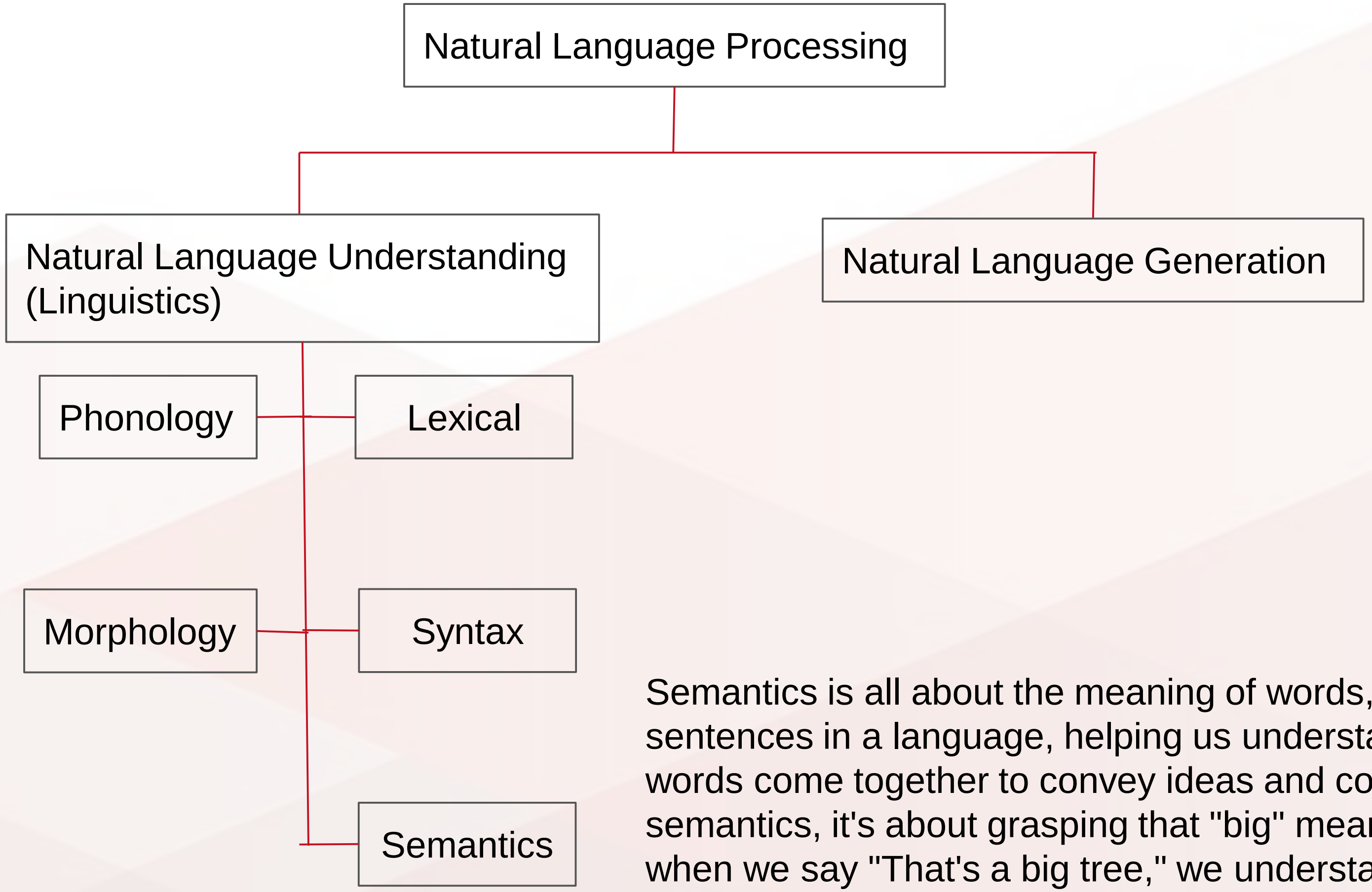
Natural Language Processing



Syntax is like the grammar of a language; it's the set of rules that determines how words are organized in a sentence to make it meaningful and coherent. E.g., In English syntax, the sentence "The cat chased the mouse" is considered correct, while "Chased mouse the cat" is not, illustrating how the order of words matters for making sense in a sentence.

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing

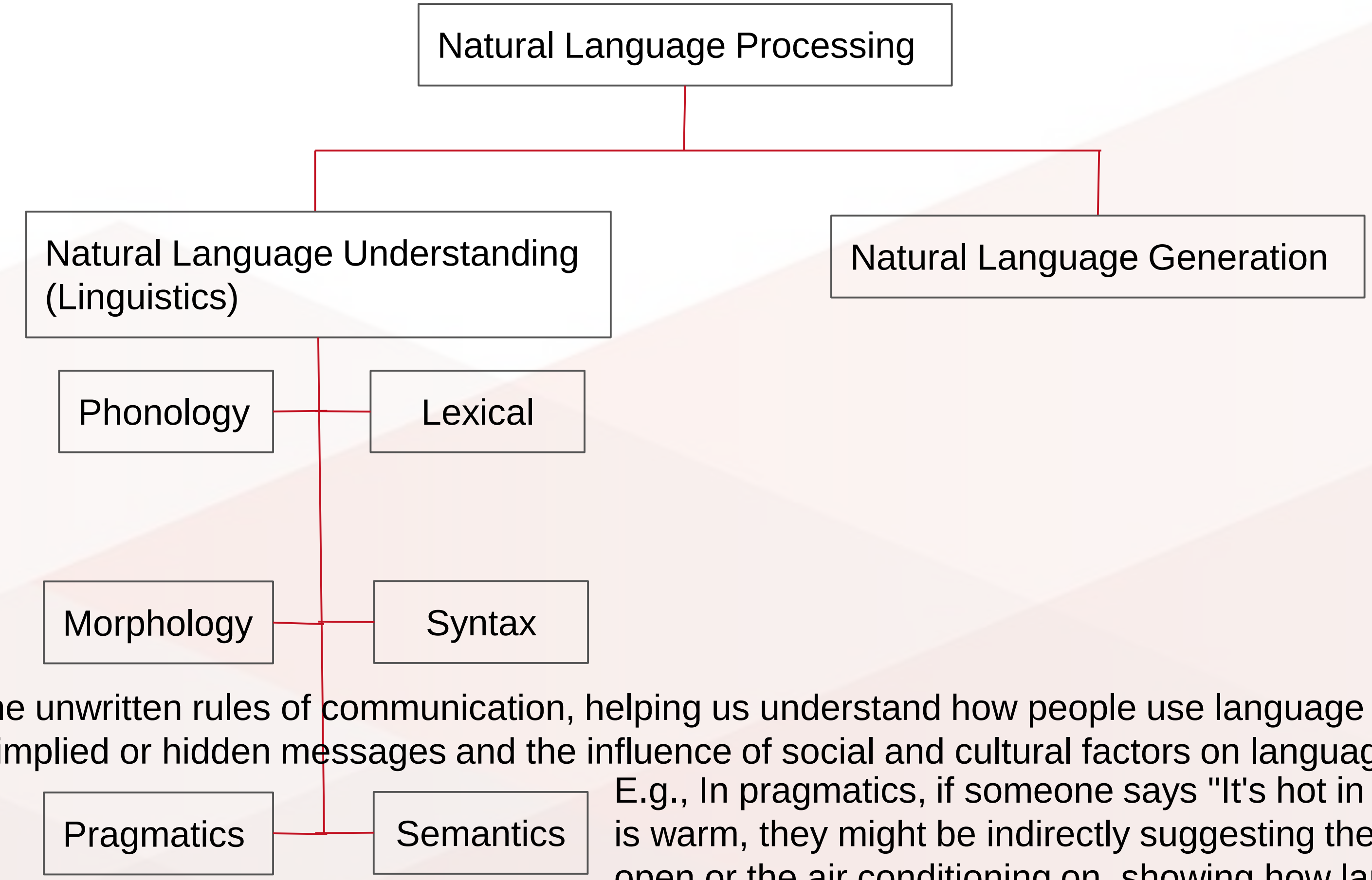


Semantics is all about the meaning of words, phrases, and sentences in a language, helping us understand how different words come together to convey ideas and concepts. E.g., In semantics, it's about grasping that "big" means large in size, so when we say "That's a big tree," we understand that the tree is substantial in its dimensions.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing



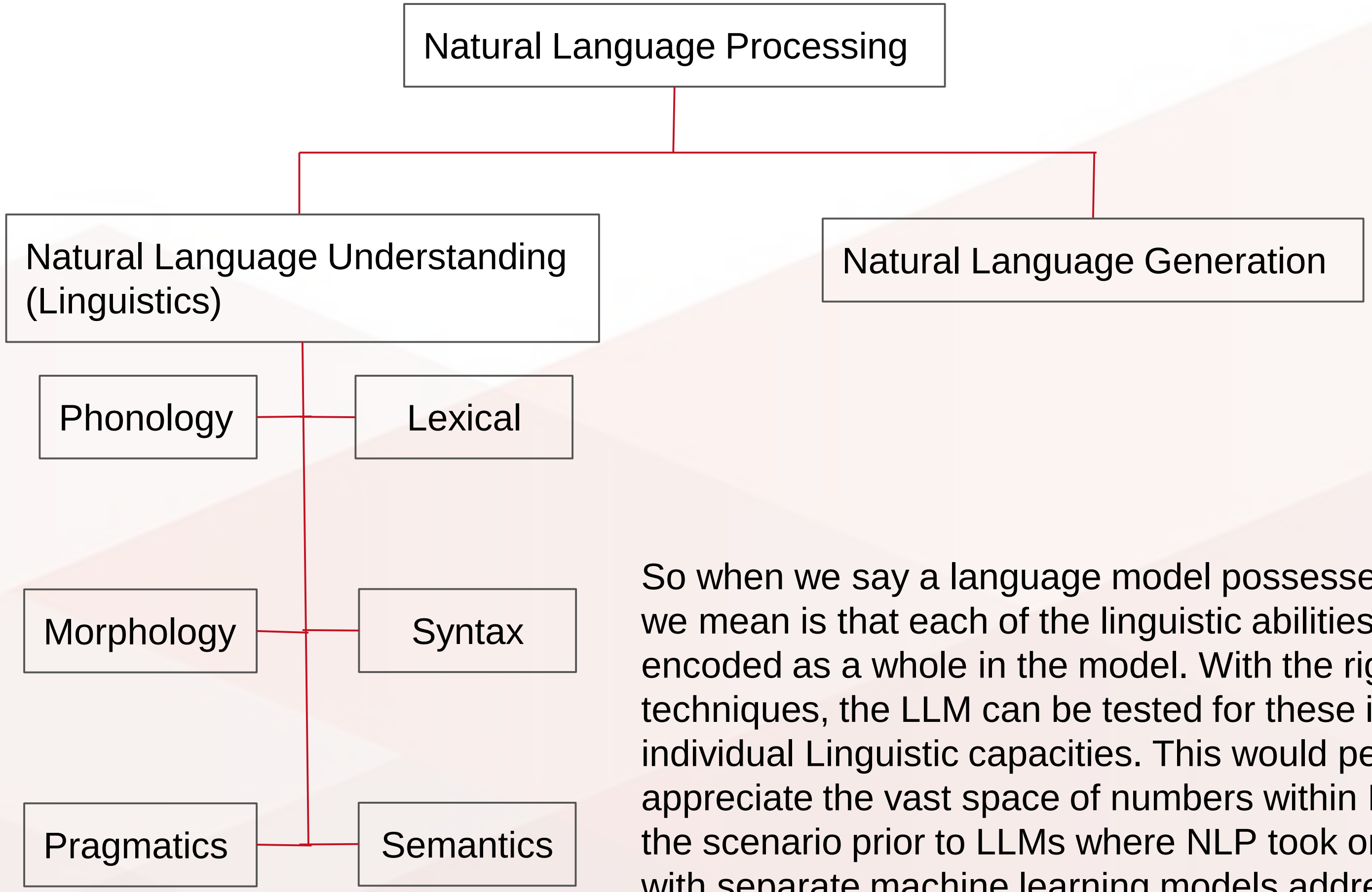
Pragmatics is like the unwritten rules of communication, helping us understand how people use language in context to convey meaning, including implied or hidden messages and the influence of social and cultural factors on language use.

E.g., In pragmatics, if someone says "It's hot in here" when the room is warm, they might be indirectly suggesting they want a window open or the air conditioning on, showing how language can convey more than just the words spoken.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

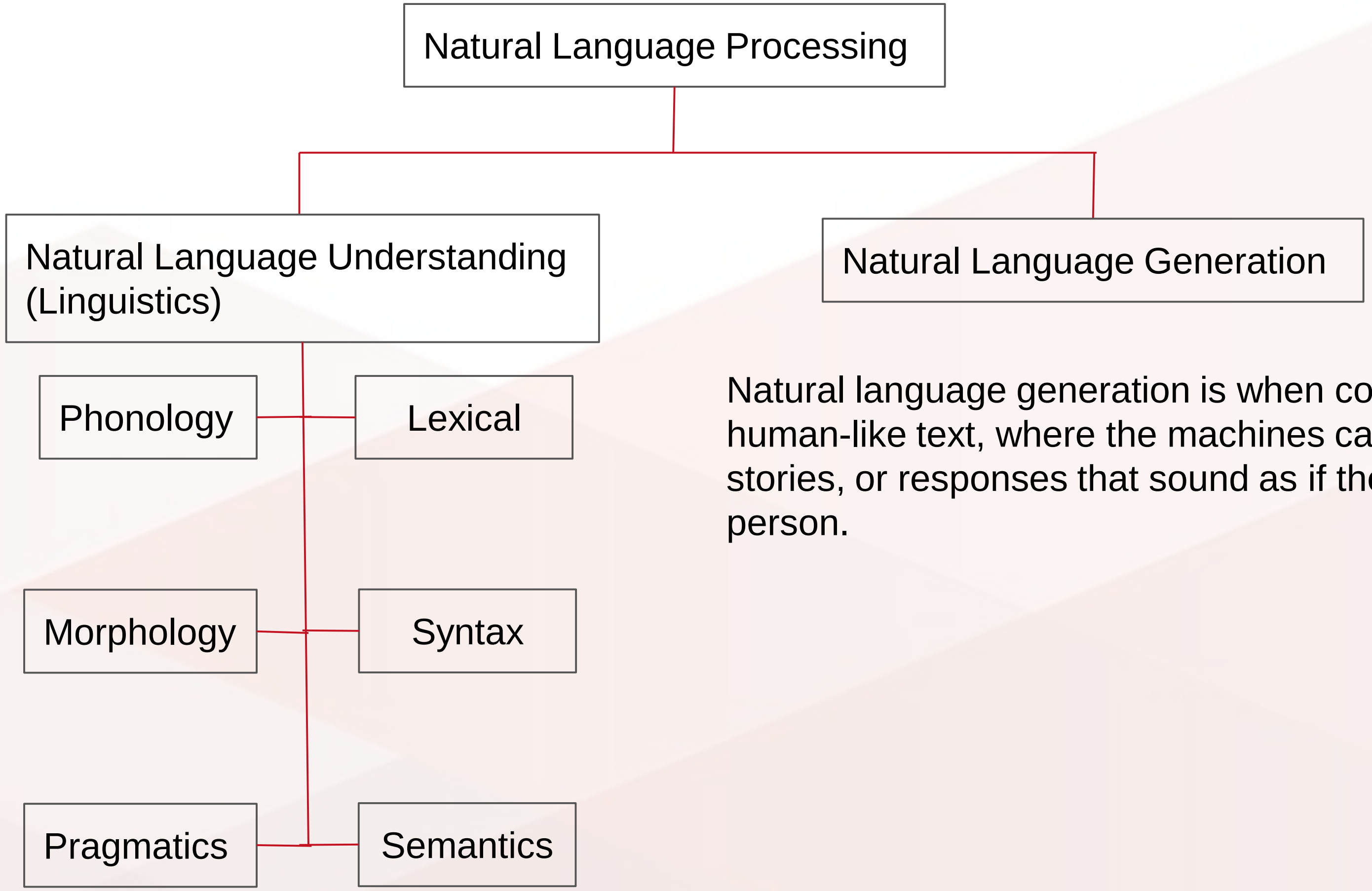
Natural Language Processing



So when we say a language model possesses NLU abilities, what we mean is that each of the linguistic abilities are implicitly encoded as a whole in the model. With the right probing techniques, the LLM can be tested for these implicitly encoded individual Linguistic capacities. This would perhaps make better appreciate the vast space of numbers within LLMs. Contrast this to the scenario prior to LLMs where NLP took on a modular nature with separate machine learning models addressing different Linguistic aspects for NLP.

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

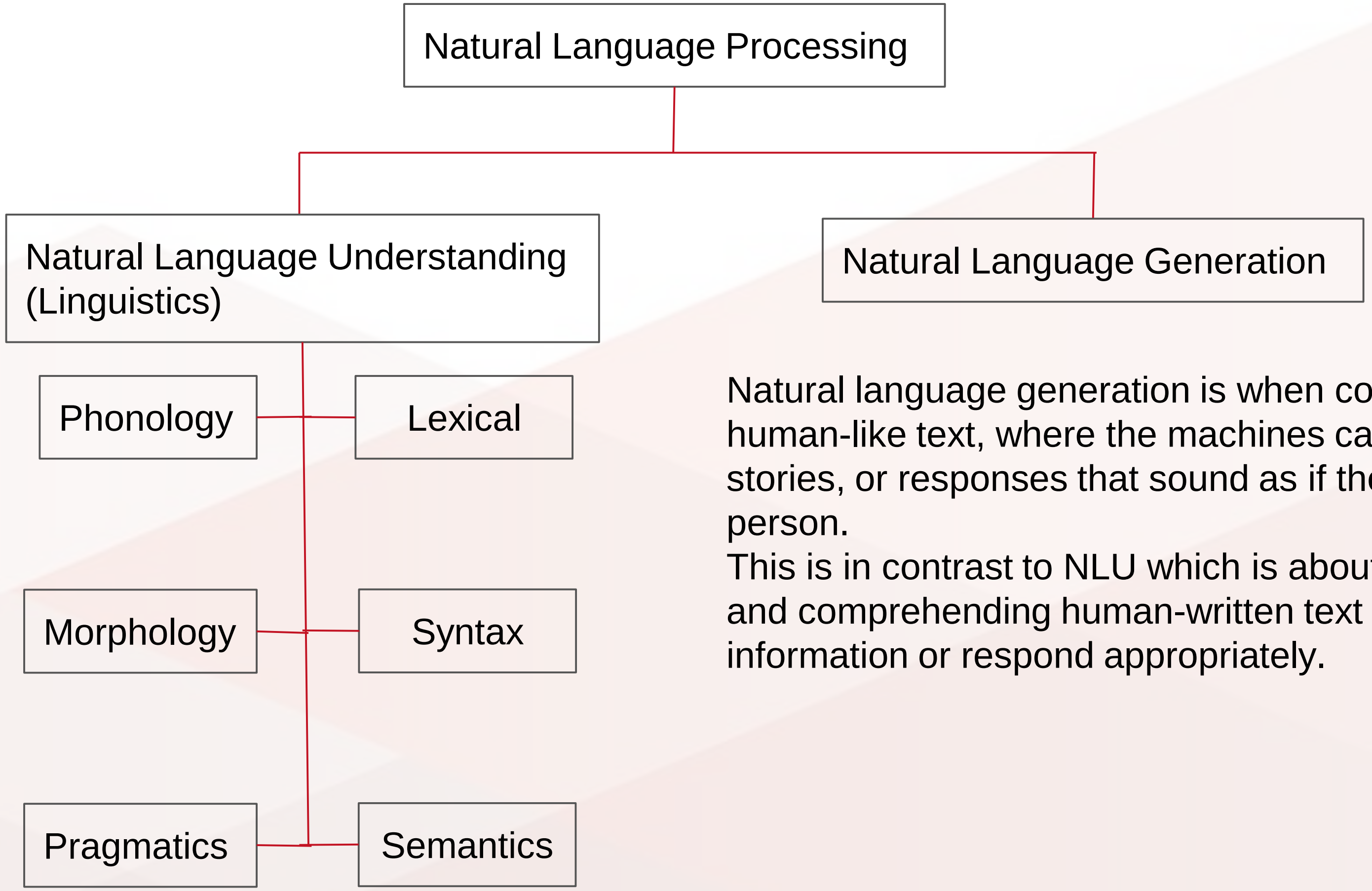
Natural Language Processing



Natural language generation is when computer's create human-like text, where the machines can write articles, stories, or responses that sound as if they were written by a person.

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing

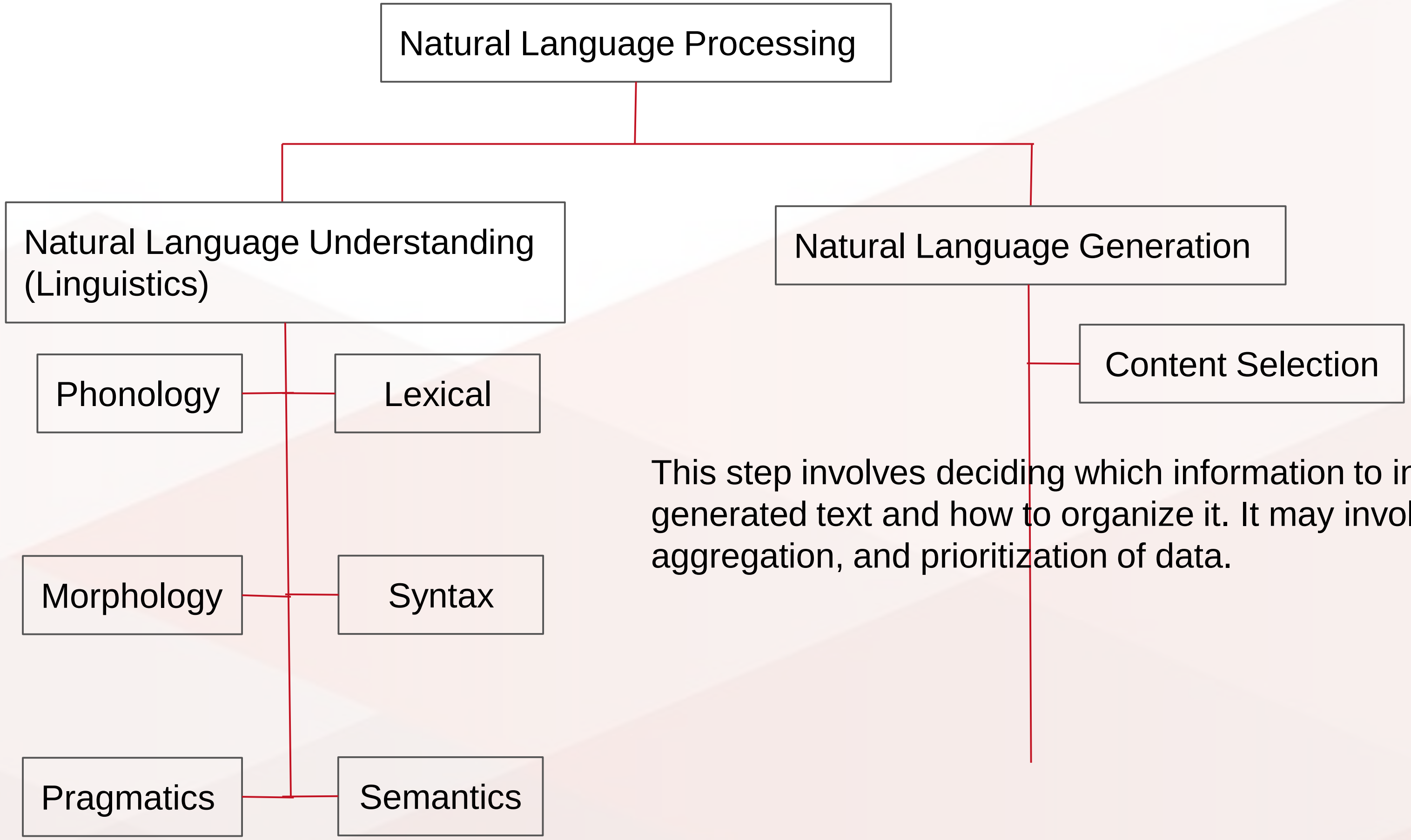


Natural language generation is when computer's create human-like text, where the machines can write articles, stories, or responses that sound as if they were written by a person.
This is in contrast to NLU which is about computers reading and comprehending human-written text to extract information or respond appropriately.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing

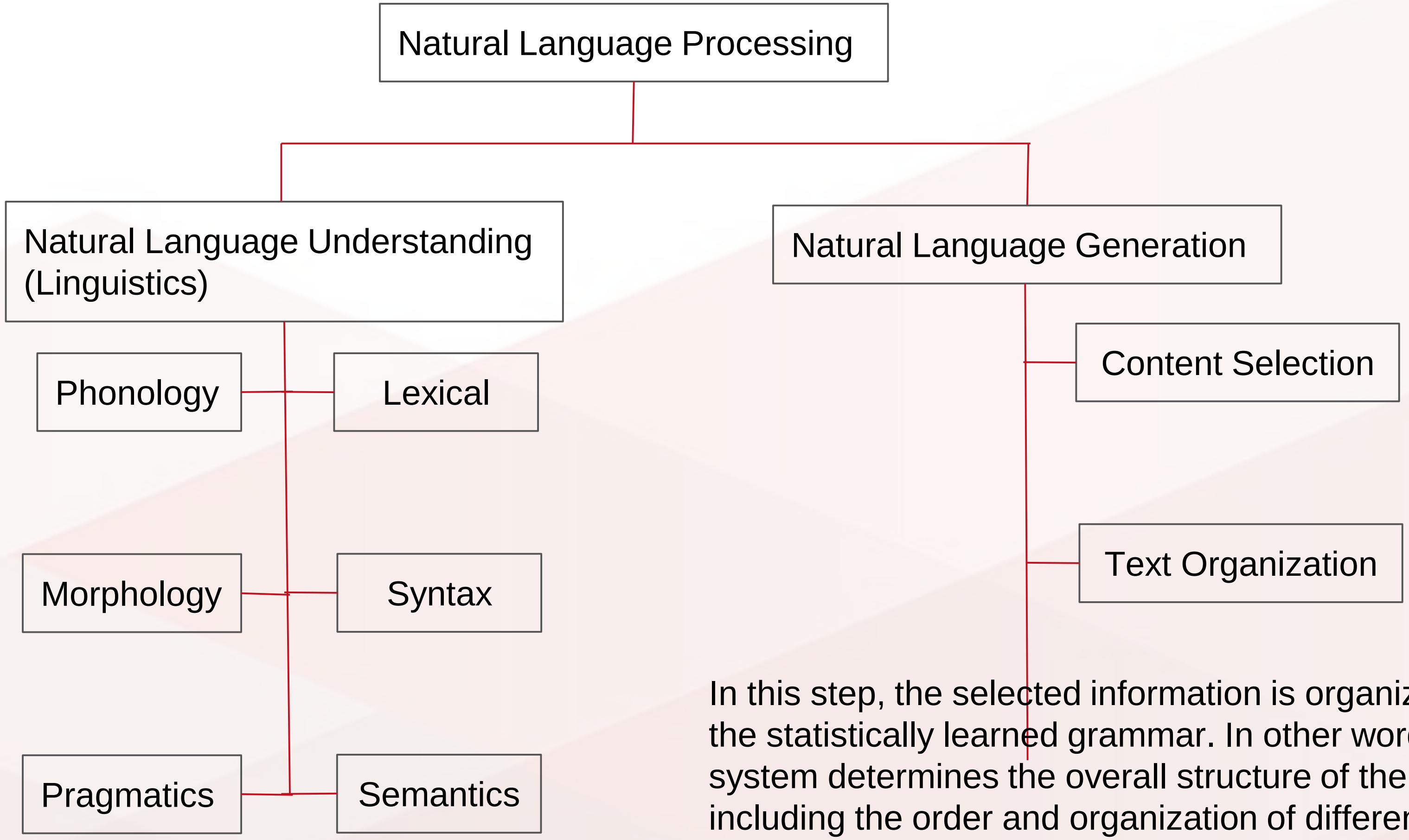


This step involves deciding which information to include in the generated text and how to organize it. It may involve filtering, aggregation, and prioritization of data.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

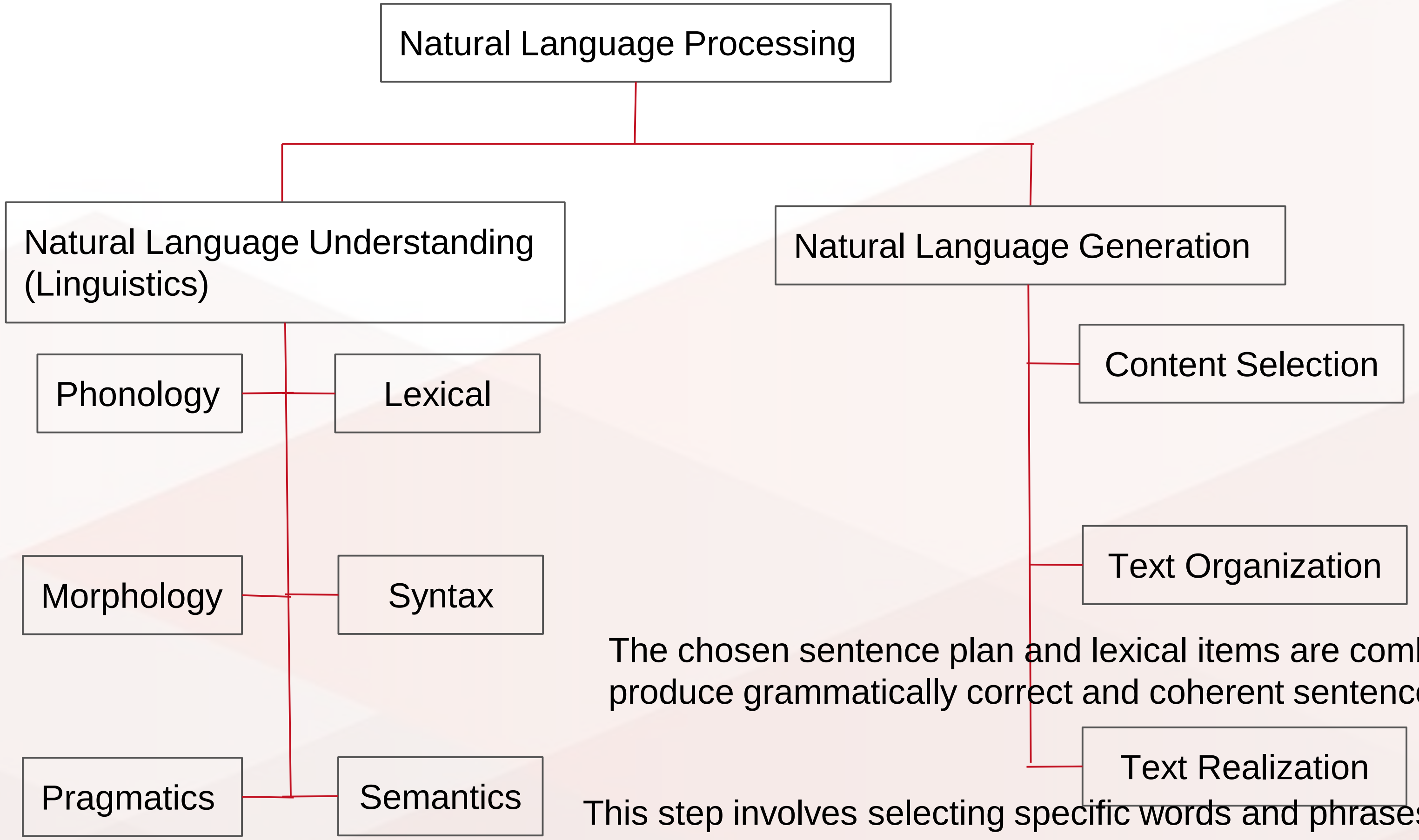
Natural Language Processing



In this step, the selected information is organized according to the statistically learned grammar. In other words, the NLG system determines the overall structure of the generated text, including the order and organization of different sections or paragraphs.

References
Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Natural Language Processing



The chosen sentence plan and lexical items are combined to produce grammatically correct and coherent sentences

This step involves selecting specific words and phrases to fit into the sentence plan, taking into account vocabulary, tense, voice, and other linguistic factors.

References

Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744.

Large Language Models (LLMs)

- A large language model (LLM) is a type of machine learning model that is trained on massive amounts of text data to generate natural language text.

References
https://www.fiddler.ai/blog/an-intro-to-llms-and-generative-ai

Large Language Models (LLMs)

- A large language model (LLM) is a type of machine learning model that is trained on massive amounts of text data to generate natural language text.
- LLMs are neural network-based models that use deep learning techniques to analyze patterns in language data, and they can learn to generate text that is grammatically correct and semantically meaningful.
  - As LLMs are scaled, they can unlock new capabilities, such as translating foreign languages, writing code, and more. All they have to do is observe recurring patterns in language during model training.

References
https://www.fiddler.ai/blog/an-intro-to-llms-and-generative-ai

Large Language Models (LLMs)

- A large language model (LLM) is a type of machine learning model that is trained on massive amounts of text data to generate natural language text.
- LLMs are neural network-based models that use deep learning techniques to analyze patterns in language data, and they can learn to generate text that is grammatically correct and semantically meaningful.
    - As LLMs are scaled, they can unlock new capabilities, such as translating foreign languages, writing code, and more. All they have to do is observe recurring patterns in language during model training.
- LLMs can be quite large, with billions of parameters, and they require significant computing power and data to train effectively.

References
https://www.fiddler.ai/blog/an-intro-to-llms-and-generative-ai

Large Language Models (LLMs)

- A large language model (LLM) is a type of machine learning model that is trained on massive amounts of text data to generate natural language text.
- LLMs are neural network-based models that use deep learning techniques to analyze patterns in language data, and they can learn to generate text that is grammatically correct and semantically meaningful.
  - As LLMs are scaled, they can unlock new capabilities, such as translating foreign languages, writing code, and more. All they have to do is observe recurring patterns in language during model training.
- LLMs can be quite large, with billions of parameters, and they require significant computing power and data to train effectively.
- The most well-known LLMs include OpenAI’s GPT (Generative Pre-trained Transformer) models and Google’s BERT (Bidirectional Encoder Representations from Transformers) models. These models have achieved impressive results in various NLP tasks, including language translation, question-answering, and text generation.

References
https://www.fiddler.ai/blog/an-intro-to-llms-and-generative-ai

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)
ELMo	AI2	Feb 2018	94,000,000
GPT	OpenAI	Jun 2018	110,000,000
BERT	Google	Oct 2018	340,000,000
XLM	Facebook	Jan 2019	655,000,000
GPT-2	OpenAI	Mar 2019	1,500,000,000
RoBERTa	Facebook	Jul 2019	355,000,000
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000
T5	Google	Oct 2019	11,000,000,000
Turing-NLG	Microsoft	Feb 2020	17,000,000,000
GPT-3	OpenAI	May 2020	175,000,000,000
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000
Gopher	DeepMind	Dec 2021	280,000,000,000

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)
ELMo	AI2	Feb 2018	94,000,000
GPT	OpenAI	Jun 2018	110,000,000
BERT	Google	Oct 2018	340,000,000
XLM	Facebook	Jan 2019	655,000,000
GPT-2	OpenAI	Mar 2019	1,500,000,000
RoBERTa	Facebook	Jul 2019	355,000,000
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000
T5	Google	Oct 2019	11,000,000,000
Turing-NLG	Microsoft	Feb 2020	17,000,000,000
GPT-3	OpenAI	May 2020	175,000,000,000
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000
Gopher	DeepMind	Dec 2021	280,000,000,000

94M

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)	
ELMo	AI2	Feb 2018	94,000,000	**94M**
GPT	OpenAI	Jun 2018	110,000,000	**110M**
BERT	Google	Oct 2018	340,000,000	
XLM	Facebook	Jan 2019	655,000,000	
GPT-2	OpenAI	Mar 2019	1,500,000,000	**1.5B**
RoBERTa	Facebook	Jul 2019	355,000,000	
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000	
T5	Google	Oct 2019	11,000,000,000	
Turing-NLG	Microsoft	Feb 2020	17,000,000,000	
GPT-3	OpenAI	May 2020	175,000,000,000	**175B**
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000	
Gopher	DeepMind	Dec 2021	280,000,000,000	

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)	
ELMo	AI2	Feb 2018	94,000,000	94M
GPT	OpenAI	Jun 2018	110,000,000	110M
BERT	Google	Oct 2018	340,000,000	340M
XLM	Facebook	Jan 2019	655,000,000	
GPT-2	OpenAI	Mar 2019	1,500,000,000	1.5B
RoBERTa	Facebook	Jul 2019	355,000,000	
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000	
T5	Google	Oct 2019	11,000,000,000	11B
Turing-NLG	Microsoft	Feb 2020	17,000,000,000	
GPT-3	OpenAI	May 2020	175,000,000,000	175B
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000	
Gopher	DeepMind	Dec 2021	280,000,000,000	

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)	
ELMo	AI2	Feb 2018	94,000,000	94M
GPT	OpenAI	Jun 2018	110,000,000	110M
BERT	Google	Oct 2018	340,000,000	340M
XLM	Facebook	Jan 2019	655,000,000	655M
GPT-2	OpenAI	Mar 2019	1,500,000,000	1.5B
RoBERTa	Facebook	Jul 2019	355,000,000	355M
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000	
T5	Google	Oct 2019	11,000,000,000	11B
Turing-NLG	Microsoft	Feb 2020	17,000,000,000	
GPT-3	OpenAI	May 2020	175,000,000,000	175B
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000	
Gopher	DeepMind	Dec 2021	280,000,000,000	

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples

Model	Organization	Date	Size (# params)	
ELMo	AI2	Feb 2018	94,000,000	**94M**
GPT	OpenAI	Jun 2018	110,000,000	**110M**
BERT	Google	Oct 2018	340,000,000	**340M**
XLM	Facebook	Jan 2019	655,000,000	**655M**
GPT-2	OpenAI	Mar 2019	1,500,000,000	**1.5B**
RoBERTa	Facebook	Jul 2019	355,000,000	**355M**
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000	**8.3B**
T5	Google	Oct 2019	11,000,000,000	**11B**
Turing-NLG	Microsoft	Feb 2020	17,000,000,000	**17B**
GPT-3	OpenAI	May 2020	175,000,000,000	**175B**
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000	**530B**
Gopher	DeepMind	Dec 2021	280,000,000,000	**280B**

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples - A Few Insights

- **Increase in size.** First, what do we mean by large? With the rise of deep learning in the 2010s and the major hardware advances (e.g., GPUs), the size of neural language models has skyrocketed. The examples we saw showed that the model sizes increased by an order of 5000x over just 4 years.

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples - A Few Insights

- **Increase in size.** First, what do we mean by large? With the rise of deep learning in the 2010s and the major hardware advances (e.g., GPUs), the size of neural language models has skyrocketed. The examples we saw showed that the model sizes increased by an order of 5000x over just 4 years.
- **Emergence.** What difference does scale make? Even though much of the technical machinery is the same, the surprising thing is that “just scaling up” these models produces new **emergent** behavior, leading to qualitatively different capabilities and qualitatively different societal impact.
    - For more information on research discovering emergence please see the “Beyond the Imitation Game” project here https://github.com/google/BIG-bench

References
https://stanford-cs324.github.io/winter2022/lectures/introduction/#a-brief-history

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)
Chinchilla	DeepMind	Apr 2022	70,000,000,000
PaLM	Google	Apr 2022	540,000,000,000
OPT	MetaAI	May 2022	175,000,000,000
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000
Flan-T5	Google	Nov 2022	11,000,000,000
ChatGPT	OpenAI	Nov 2022	175,000,000,000
LLaMA	MetaAI	Feb 2023	65,200,000,000
Alpaca	Stanford	March 2023	65,000,000,000
GPT-4	OpenAI	March 2023	170,000,000,000,000
Falcon	TII	May 2023	40,000,000,000

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	70B
PaLM	Google	Apr 2022	540,000,000,000	
OPT	MetaAI	May 2022	175,000,000,000	
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	
Flan-T5	Google	Nov 2022	11,000,000,000	
ChatGPT	OpenAI	Nov 2022	175,000,000,000	
LLaMA	MetaAI	Feb 2023	65,200,000,000	
Alpaca	Stanford	March 2023	65,000,000,000	
GPT-4	OpenAI	March 2023	170,000,000,000,000	
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	70B
PaLM	Google	Apr 2022	540,000,000,000	540B
OPT	MetaAI	May 2022	175,000,000,000	
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	
Flan-T5	Google	Nov 2022	11,000,000,000	11B
ChatGPT	OpenAI	Nov 2022	175,000,000,000	
LLaMA	MetaAI	Feb 2023	65,200,000,000	
Alpaca	Stanford	March 2023	65,000,000,000	
GPT-4	OpenAI	March 2023	170,000,000,000,000	
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	70B
PaLM	Google	Apr 2022	540,000,000,000	540B
OPT	MetaAI	May 2022	175,000,000,000	175B
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	
Flan-T5	Google	Nov 2022	11,000,000,000	11B
ChatGPT	OpenAI	Nov 2022	175,000,000,000	
LLaMA	MetaAI	Feb 2023	65,200,000,000	65B
Alpaca	Stanford	March 2023	65,000,000,000	
GPT-4	OpenAI	March 2023	170,000,000,000,000	
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	70B
PaLM	Google	Apr 2022	540,000,000,000	540B
OPT	MetaAI	May 2022	175,000,000,000	175B
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	176B
Flan-T5	Google	Nov 2022	11,000,000,000	11B
ChatGPT	OpenAI	Nov 2022	175,000,000,000	
LLaMA	MetaAI	Feb 2023	65,200,000,000	65B
Alpaca	Stanford	March 2023	65,000,000,000	
GPT-4	OpenAI	March 2023	170,000,000,000,000	
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	70B
PaLM	Google	Apr 2022	540,000,000,000	540B
OPT	MetaAI	May 2022	175,000,000,000	175B
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	176B
Flan-T5	Google	Nov 2022	11,000,000,000	11B
ChatGPT	OpenAI	Nov 2022	175,000,000,000	176B
LLaMA	MetaAI	Feb 2023	65,200,000,000	65B
Alpaca	Stanford	March 2023	65,000,000,000	
GPT-4	OpenAI	March 2023	170,000,000,000,000	170T
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	**70B**
PaLM	Google	Apr 2022	540,000,000,000	**540B**
OPT	MetaAI	May 2022	175,000,000,000	**175B**
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	**176B**
Flan-T5	Google	Nov 2022	11,000,000,000	**11B**
ChatGPT	OpenAI	Nov 2022	175,000,000,000	**176B**
LLaMA	MetaAI	Feb 2023	65,200,000,000	**65B**
Alpaca	Stanford	March 2023	65,000,000,000	**65B**
GPT-4	OpenAI	March 2023	170,000,000,000,000	**170T**
Falcon	TII	May 2023	40,000,000,000	

References
https://orkg.org/comparison/R609337/

Large Language Models (LLMs) Examples - 2022 & 2023

Model	Organization	Date	Size (# params)	
Chinchilla	DeepMind	Apr 2022	70,000,000,000	**70B**
PaLM	Google	Apr 2022	540,000,000,000	**540B**
OPT	MetaAI	May 2022	175,000,000,000	**175B**
BLOOM	Big Science, HuggingFace	Jul 2022	176,000,000,000	**176B**
Flan-T5	Google	Nov 2022	11,000,000,000	**11B**
ChatGPT	OpenAI	Nov 2022	175,000,000,000	**176B**
LLaMA	MetaAI	Feb 2023	65,200,000,000	**65B**
Alpaca	Stanford	March 2023	65,000,000,000	**65B**
GPT-4	OpenAI	March 2023	170,000,000,000,000	**170T**
Falcon	TII	May 2023	40,000,000,000	**40B**

References
https://orkg.org/comparison/R609337/

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

(I) Datasets for Pretraining

Jennifer D'Souza
Technische Informationsbibliothek (TIB)
Welfengarten 1B // 30167 Hannover

Leibniz
Gemeinschaft

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

(I) Datasets for Pretraining

Jennifer D'Souza
Technische Informationsbibliothek (TIB)
Welfengarten 1B // 30167 Hannover

Leibniz
Gemeinschaft

Pretraining?

Pre-training is the process where models are trained on huge quantities of data, which helps them comprehend a broad range of language patterns and constructs.

Pretraining Datasets?

Pretraining datasets are foundational to creating optimal LLMs for the following reasons:

1. **Broad Language Understanding**
2. **Shared Knowledge Base**
3. **Efficient Transfer Learning**
4. **Context and World Knowledge**
5. **Generalization and Scalability**

Plan for the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

Plan for the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

Pretraining Architectures

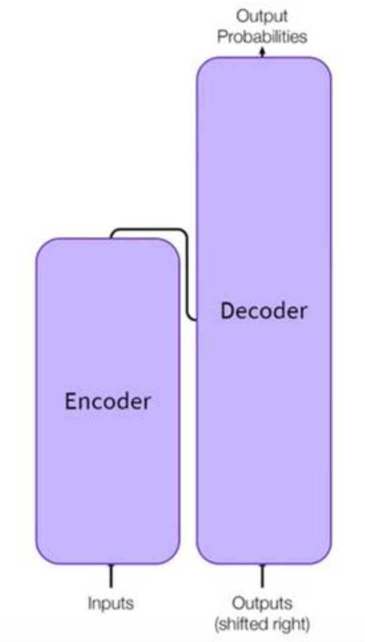
- **Transformer - Encoder only architecture**
  - Bi-directional: context from the left, and the right
  - Good at extracting meaningful information
  - Sequence classification, question answering, masked language modeling
  - NLU: Natural Language Understanding
  - Examples of encoders: BERT, RoBERTa, ALBERT

Pretraining Architectures

- **Transformer - Decoder only architecture**
  - Unidirectional: access to their left (or right!) context
  - Great at causal tasks; generating sequences
  - NLG: Natural language generation
    - Note this is a traditional application sense of a decoder. Recent work on LLMs leverage various input representations for wide variety of downstream tasks that are effectively addressed via the text generation objective.
  - Examples: GPT-series, Falcon, LLaMA

Pretraining Architectures

- **Transformer - Encoder-Decoder architecture**
    - Combines the functionality of both the encoder and decoder architectures. Encoders produce models that are very good at natural language understanding. Decoders produce models that are very good at text generation. Encoder-decoders combined produce models that generate text based on functionalities to encode bidirectional contextual sequence representations.

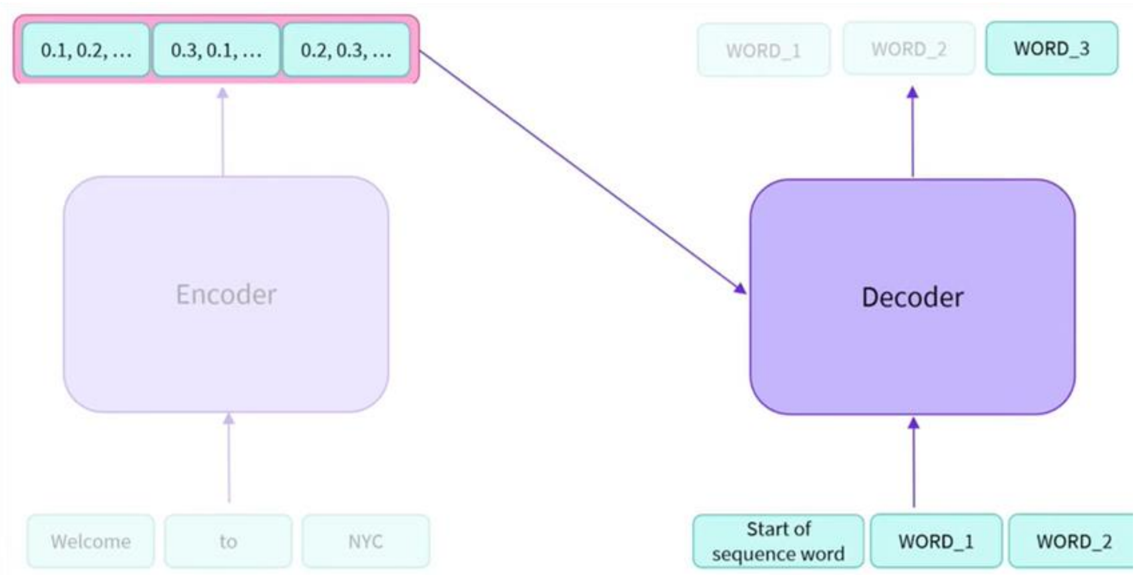


Source

HuggingFace Tutorials https://www.youtube.com/watch?v=0_4KEb08xrE

Pretraining Architectures

- **Transformer - Encoder-Decoder architecture**
  - The decoder’s autoregressive behavior allows it to add words that it just generated as output and allows it to include it as part of the generation input sequence.



Source
HuggingFace Tutorials https://www.youtube.com/watch?v=0_4KEb08xrE

Pretraining Architectures

- **Transformer - Encoder-Decoder architecture**
  - Sequence to sequence tasks; many-to-many: translation, summarization
  - Examples: BART, T5, mT5, Pegasus, mBART …

Plan for the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

Plan for the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence-to-sequence generation.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence-to-sequence generation.

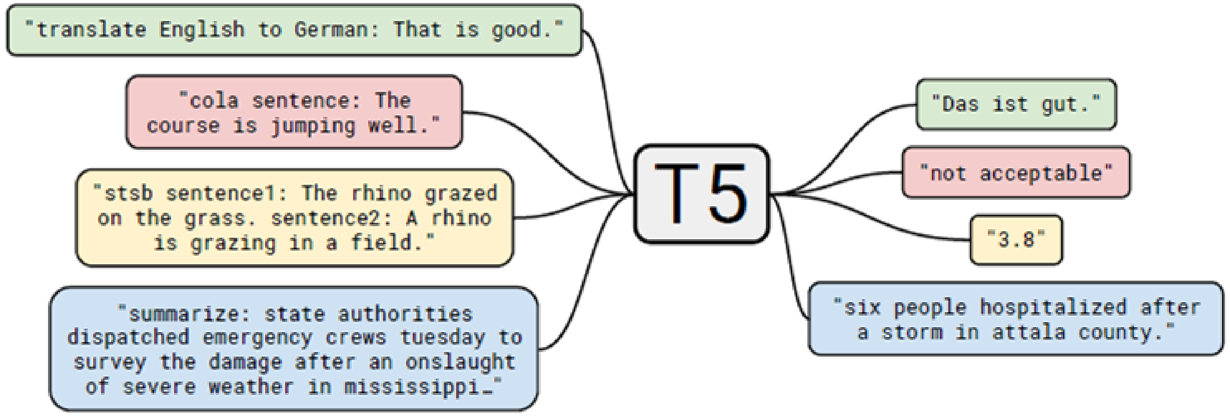


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence-to-sequence generation.

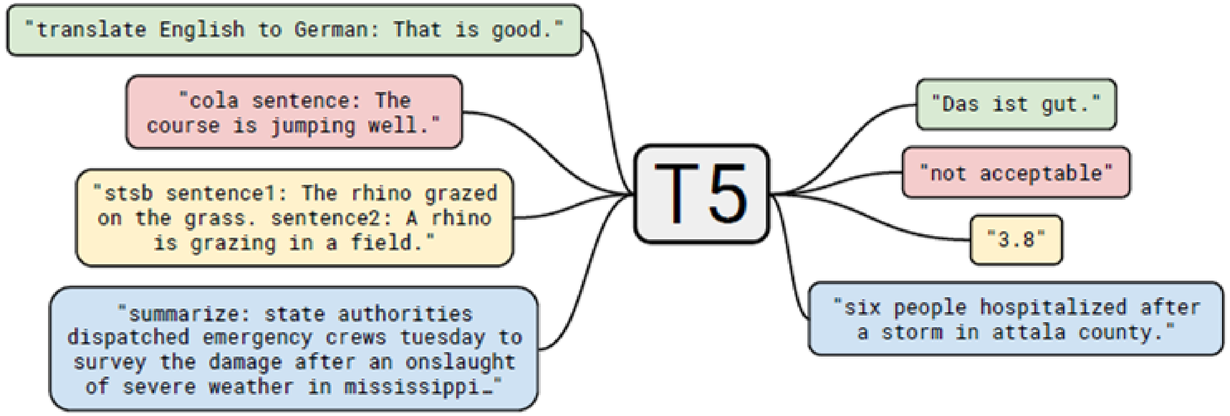


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

Since the same model is used to perform many tasks, the way the model is told which task to perform is by prepending the input with the task.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence to sequence generation.

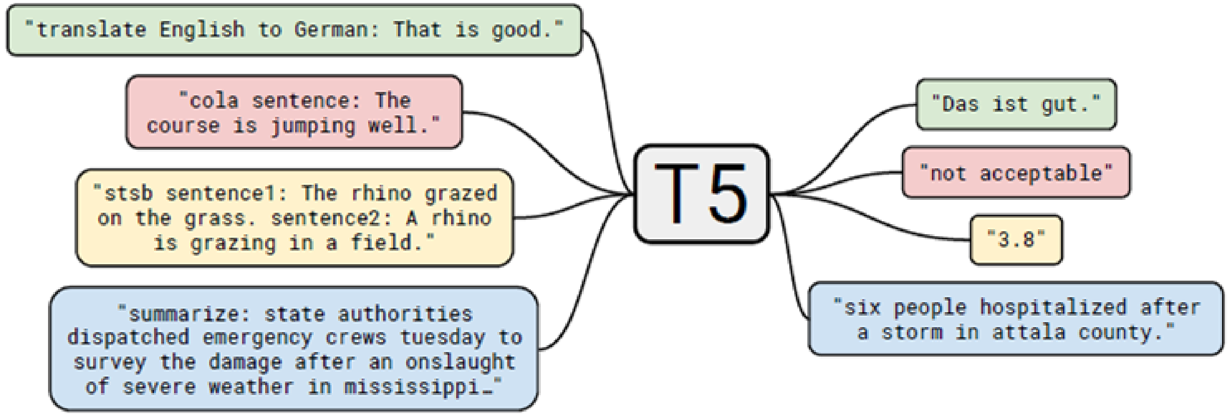


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

While machine translation as a text-to-text task is straightforward, the same method is also applied to classification tasks, as seen in the red box.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence to sequence generation.

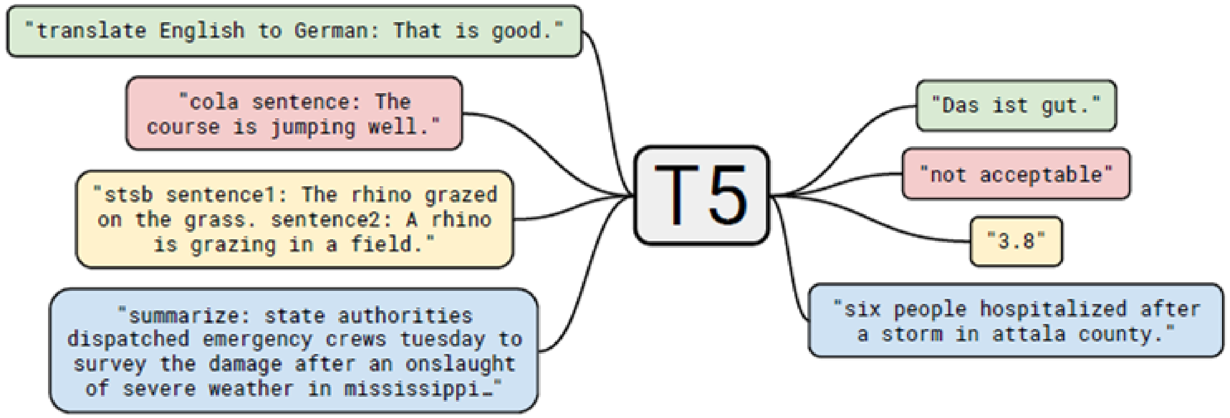


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

For the classification task, the model is trained to predict the text. Note in traditional paradigms, we usually have a softmax layer which predict the probabilities that are then mapped to a label.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Key idea**
  - Modeling every NLP problem as a text-to-text task or sequence to sequence generation.

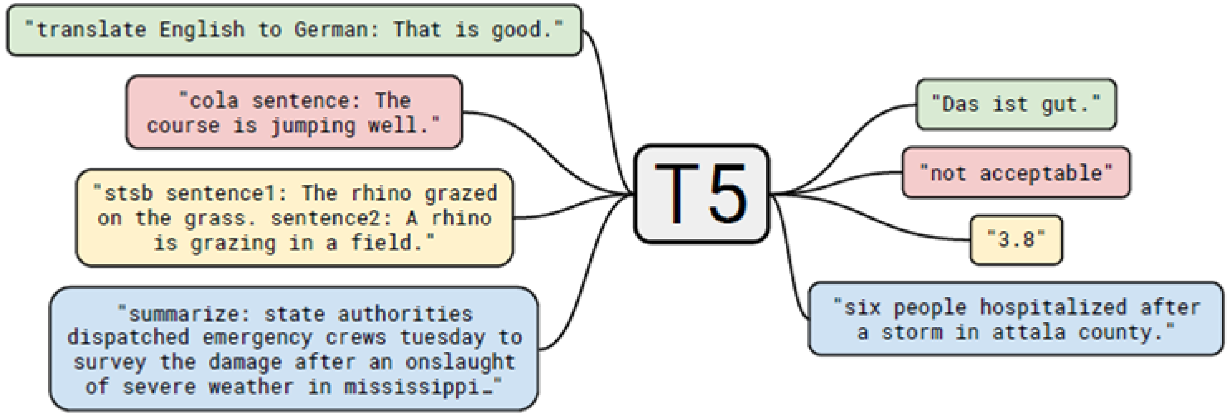


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

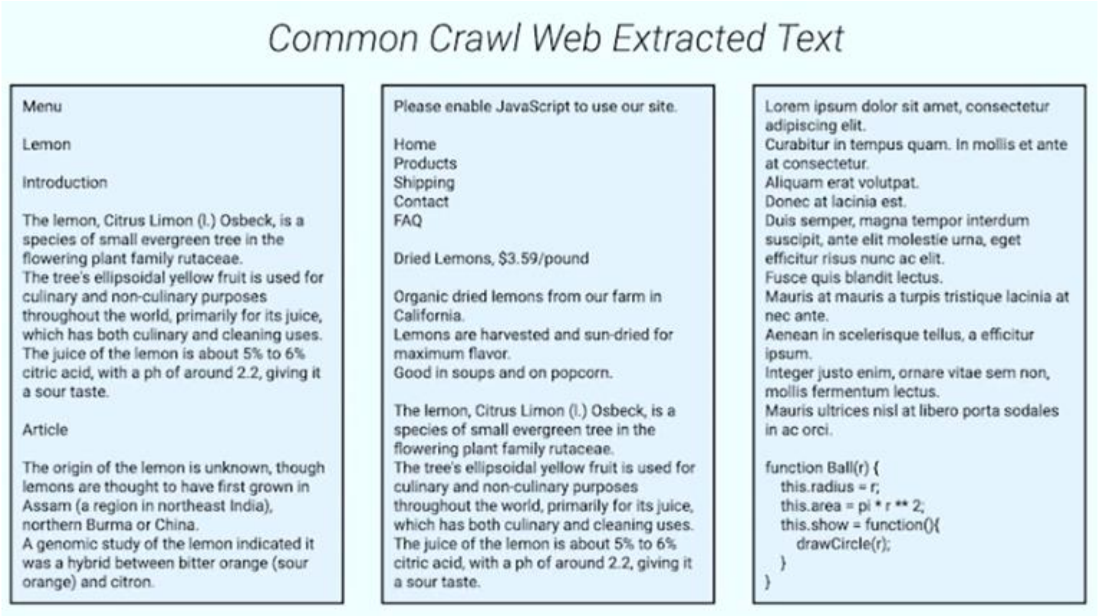
Interestingly the same text-to-text framework is also applied to regression problems.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

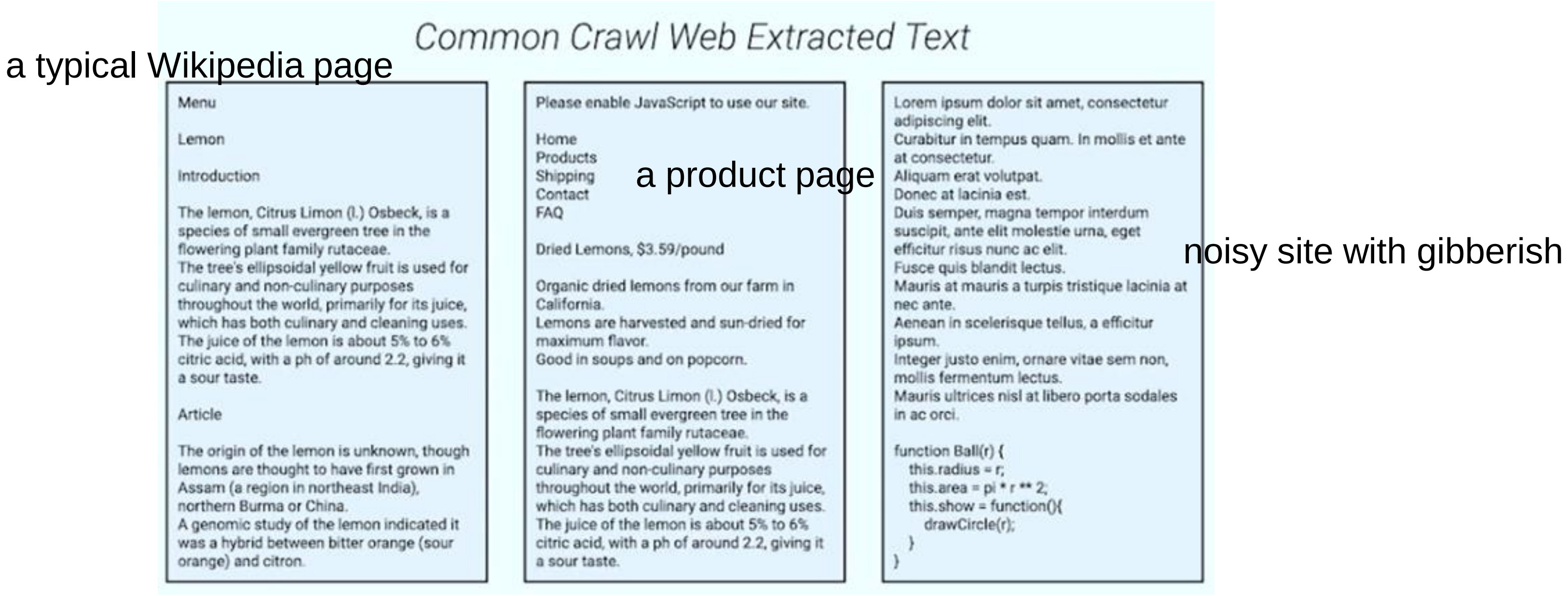


References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.



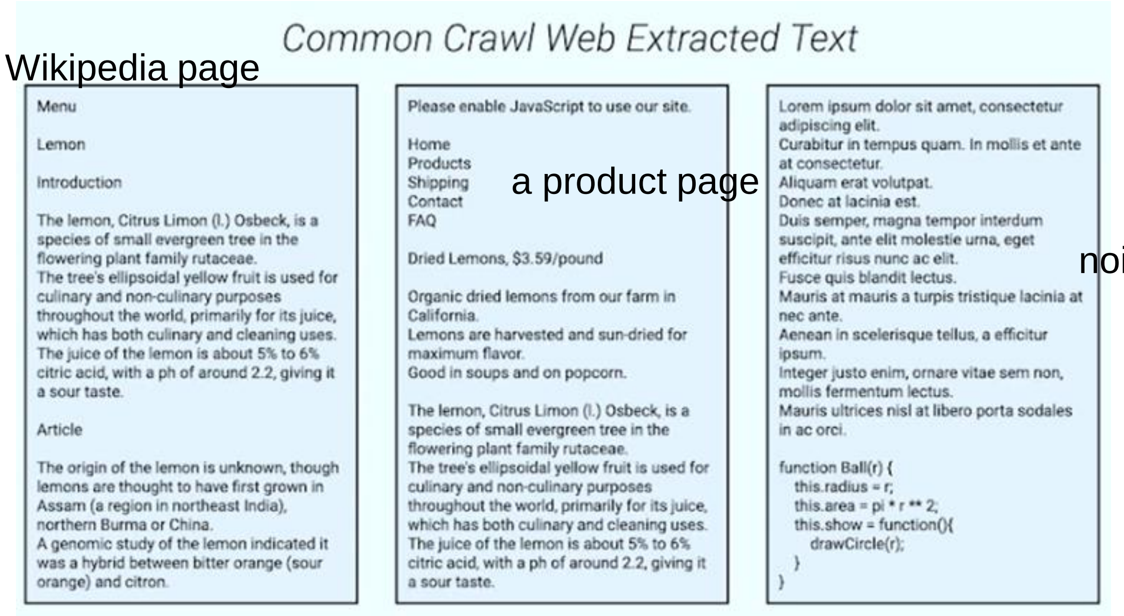
References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



a product page

noisy site with gibberish

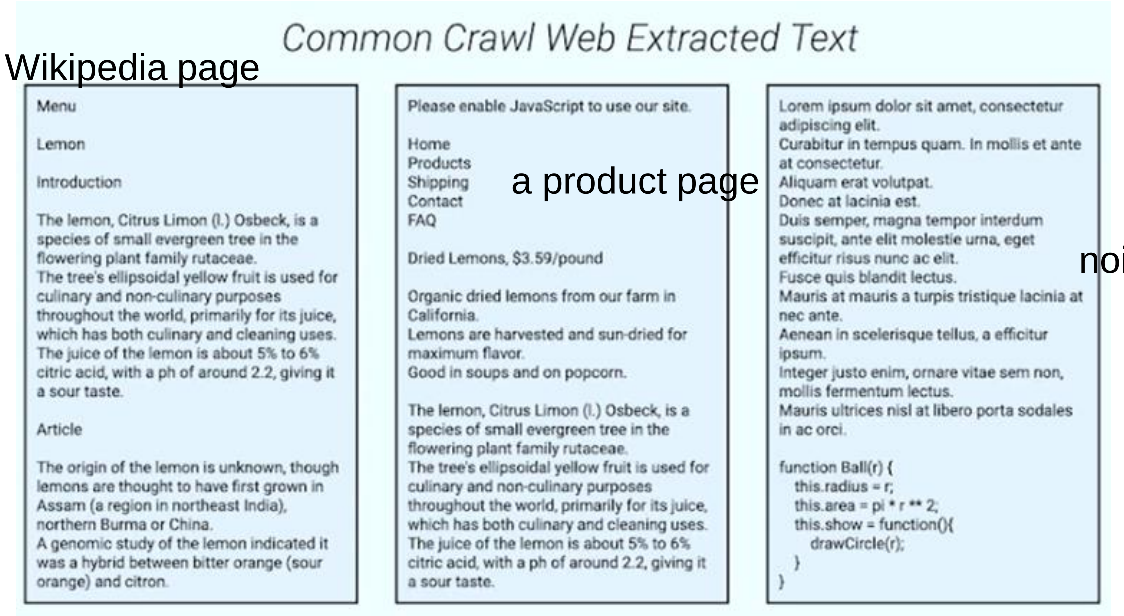
This reflects the typical state of web-crawled text. It is noisy and includes a lot of stuff that isn't natural language text such as menus or code.

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



a product page

noisy site with gibberish

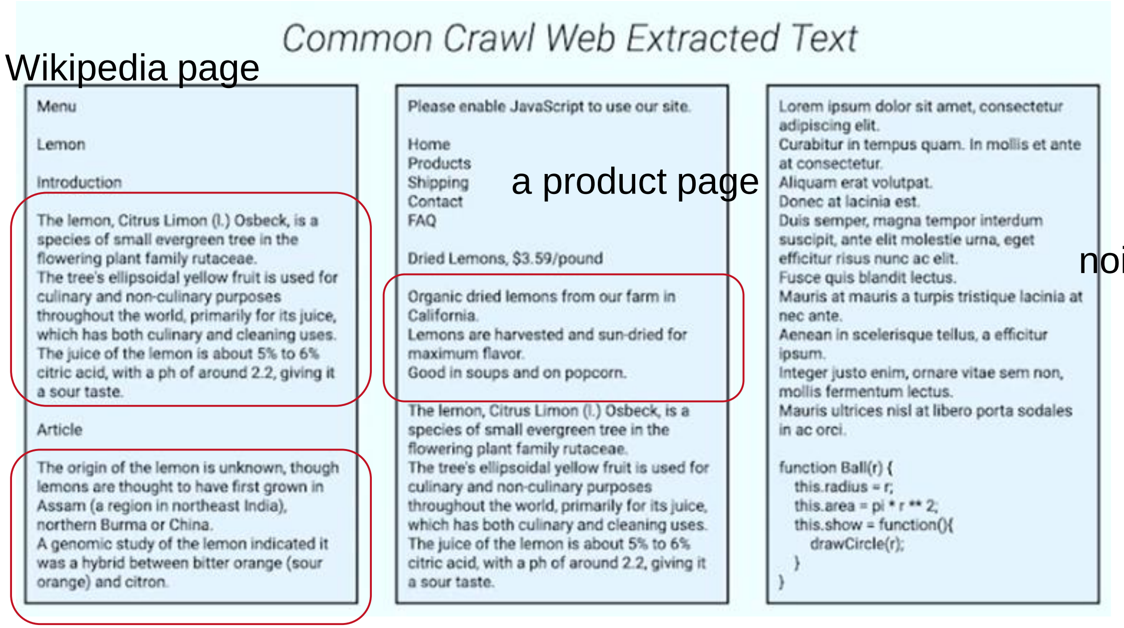
The authors of this work hypothesized that one could obtain a more effective pretrained model if the web text could be cleaned in some way.

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



a product page

noisy site with gibberish

They came up with some lightweight heuristics that resulting filtered unwanted chunks and retained only valid text chunks shown in the red boxes resulting in a version of commoncrawl called filtered commoncrawl.

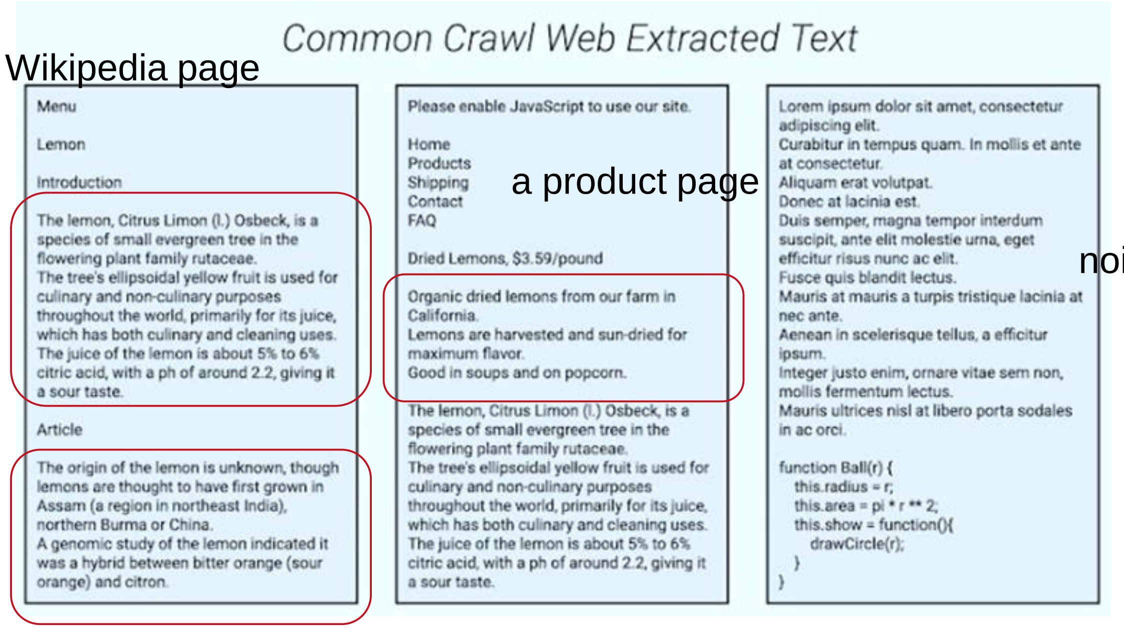
References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



noisy site with gibberish

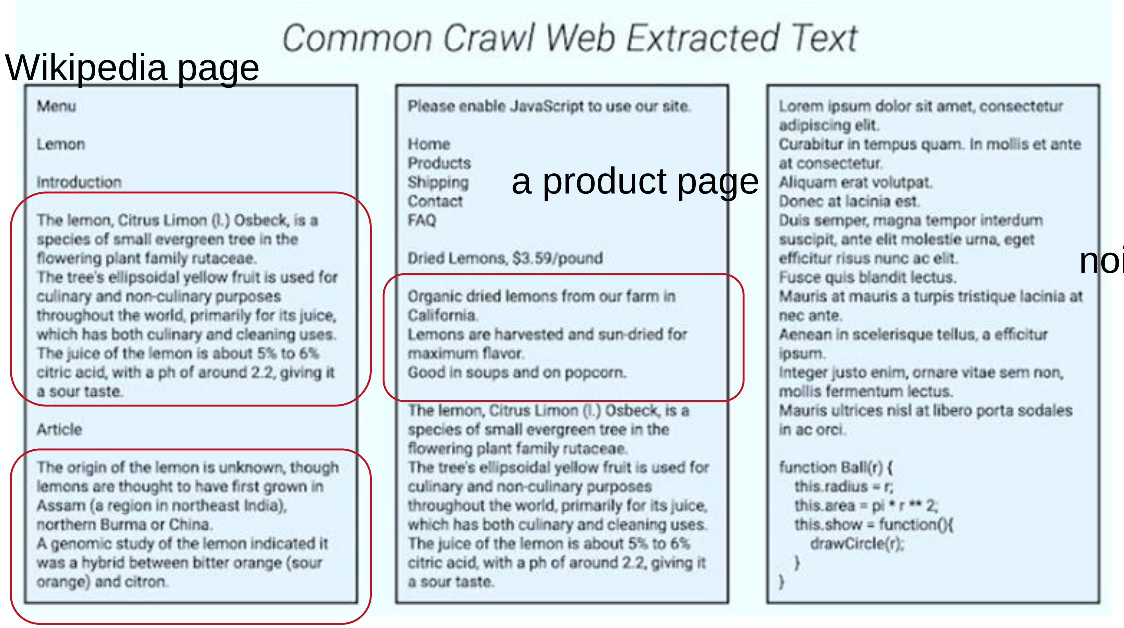
Filtering heuristics:
1. remove lines that didn't end in a punctuation.

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



noisy site with gibberish

Filtering heuristics:
2. remove any lines that contain the text Javascript because the authors found websites contained a lot of redundant text which said “activate javascript on your browser”

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551

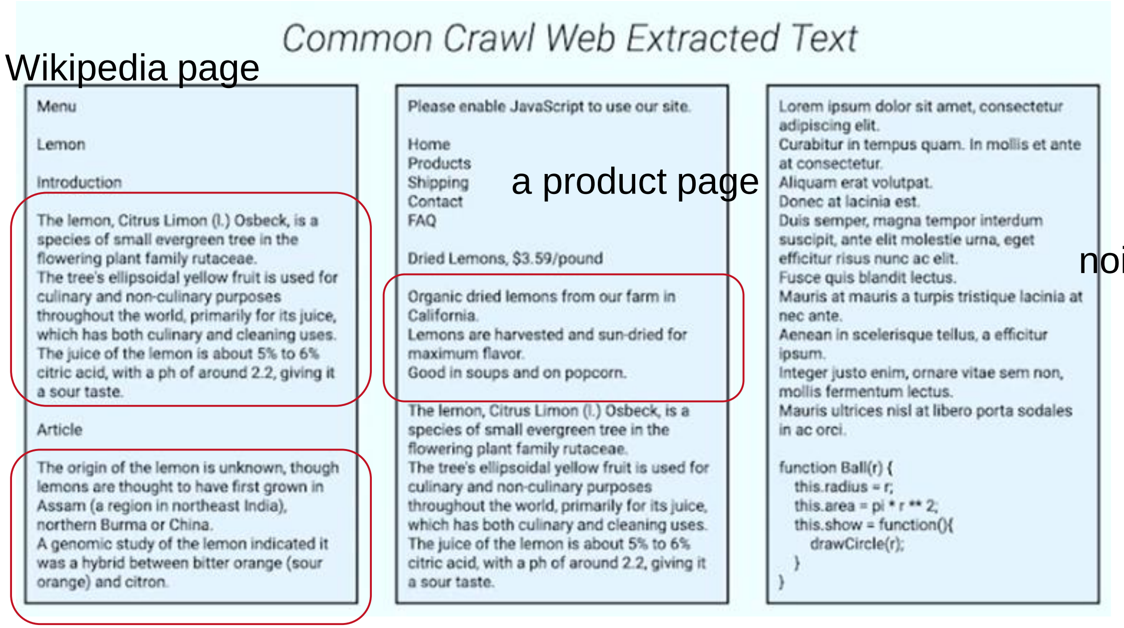
T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page

a product page

noisy site with gibberish



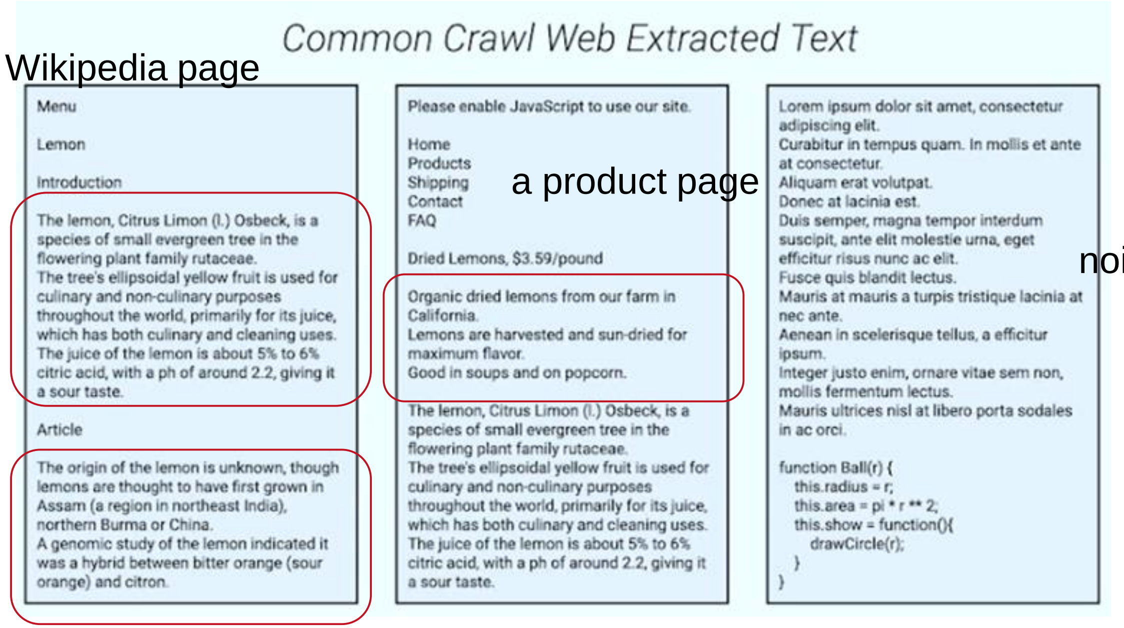
Filtering heuristics:
3. remove all text with a curly bracket, because a curly bracket often appears in code and not in natural language

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.



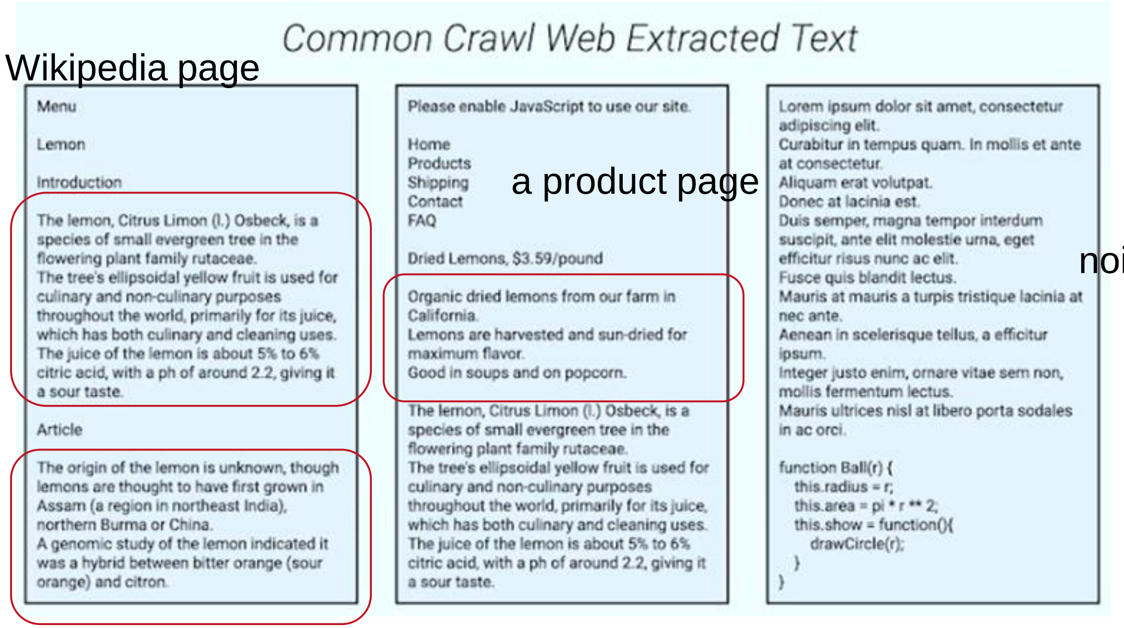
Filtering heuristics:
4. also deduplicated the dataset using the method of a sliding window to ensure that a chunk appeared only once in the whole corpus.

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



a product page

noisy site with gibberish

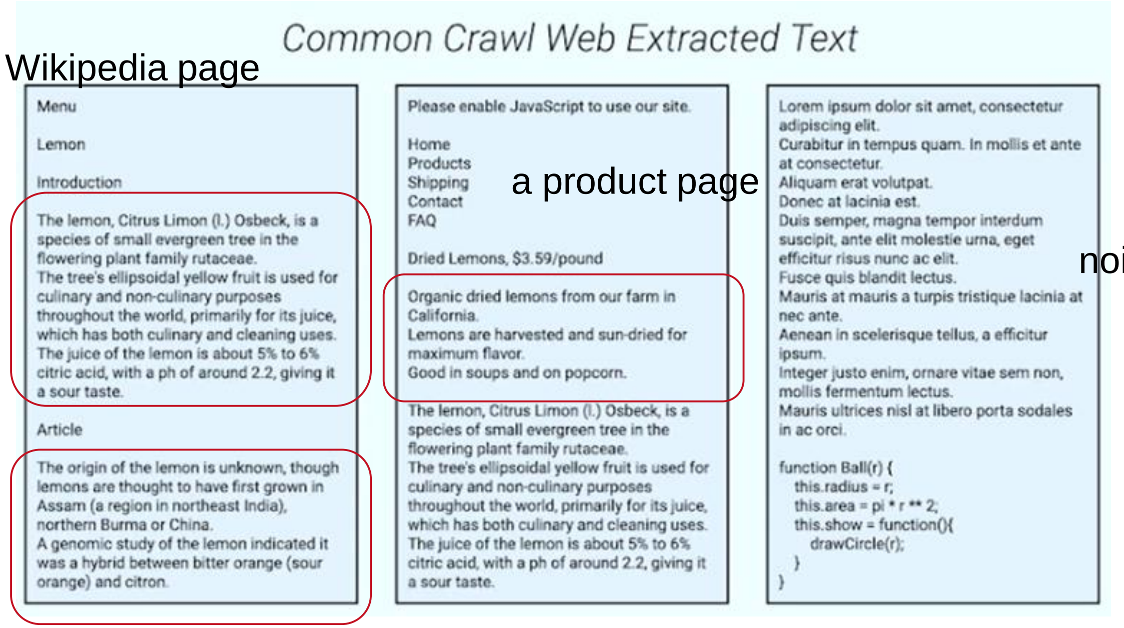
Filtering heuristics:
5. finally langdetect was used to retain only sentences in English.

References
1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training data sources**
  1. Common Crawl
     - The Common Crawl initiative is a nonprofit organization that conducts large-scale web crawling to create and maintain an openly accessible archive of web content for research and public use. https://commoncrawl.org/
     - A typical dump of common crawl looks as shown in the Figure below.

a typical Wikipedia page



a product page

noisy site with gibberish

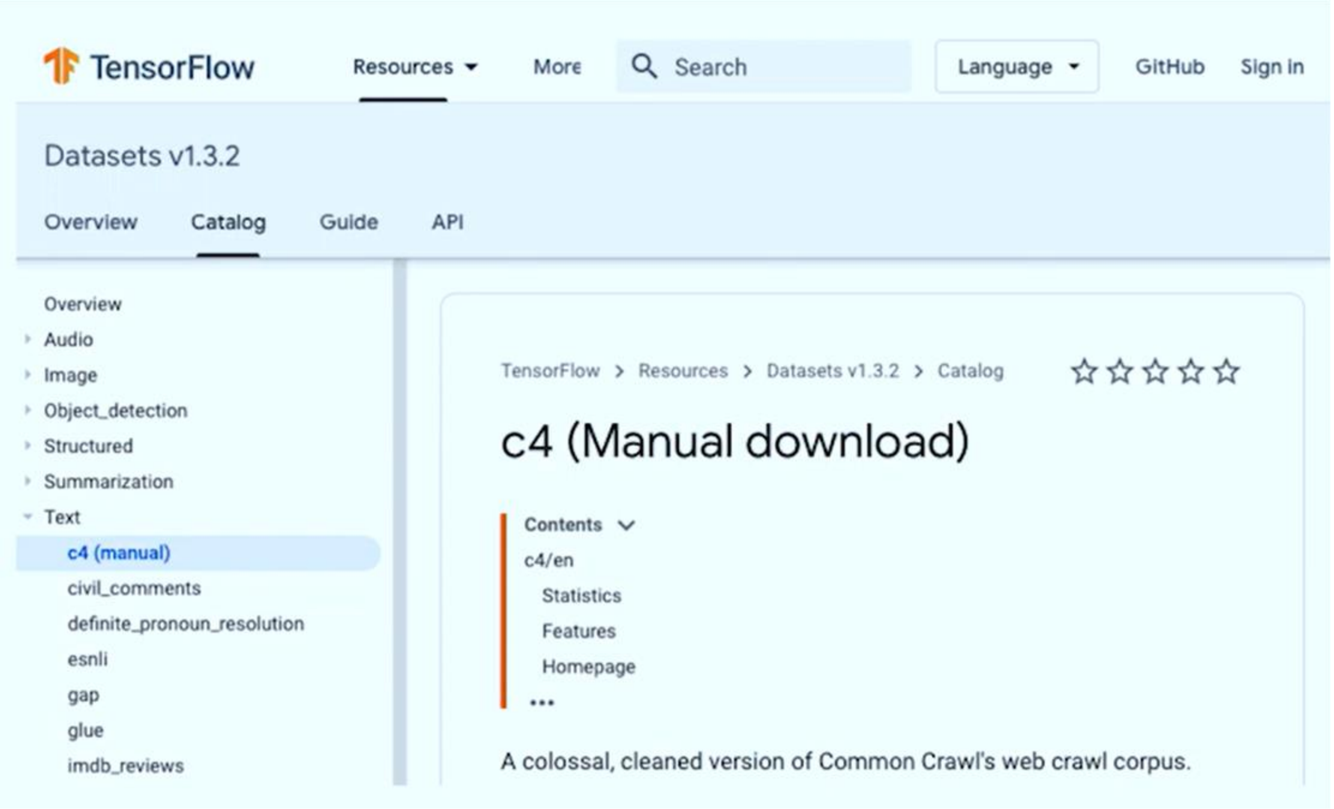
Filtered version of CommonCrawl comprised roughly 700 GB of English text from terabytes of original data.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training dataset**
    1. Colossal Cleaned Crawled Corpus (C4)



1. Publicly available on Tensorflow datasets https://www.tensorflow.org/datasets/catalog/c4
2. Presents a great academic exercise to recreate a cleaned version of the web.
    a. Common Crawl dataset is publicly available
    b. The code for the filtering heuristics is publicly available.

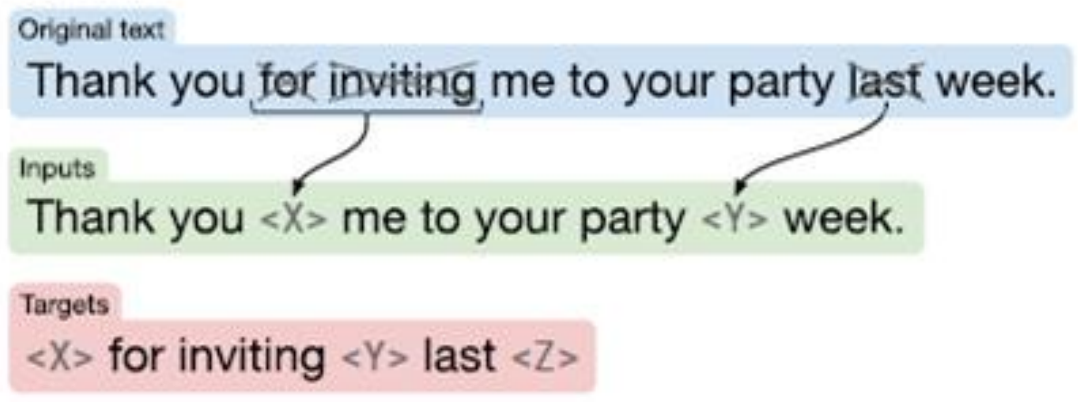
Great example of open-source, replicable research accessible in community facilitating LLM developments.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **Pre-training Objective**



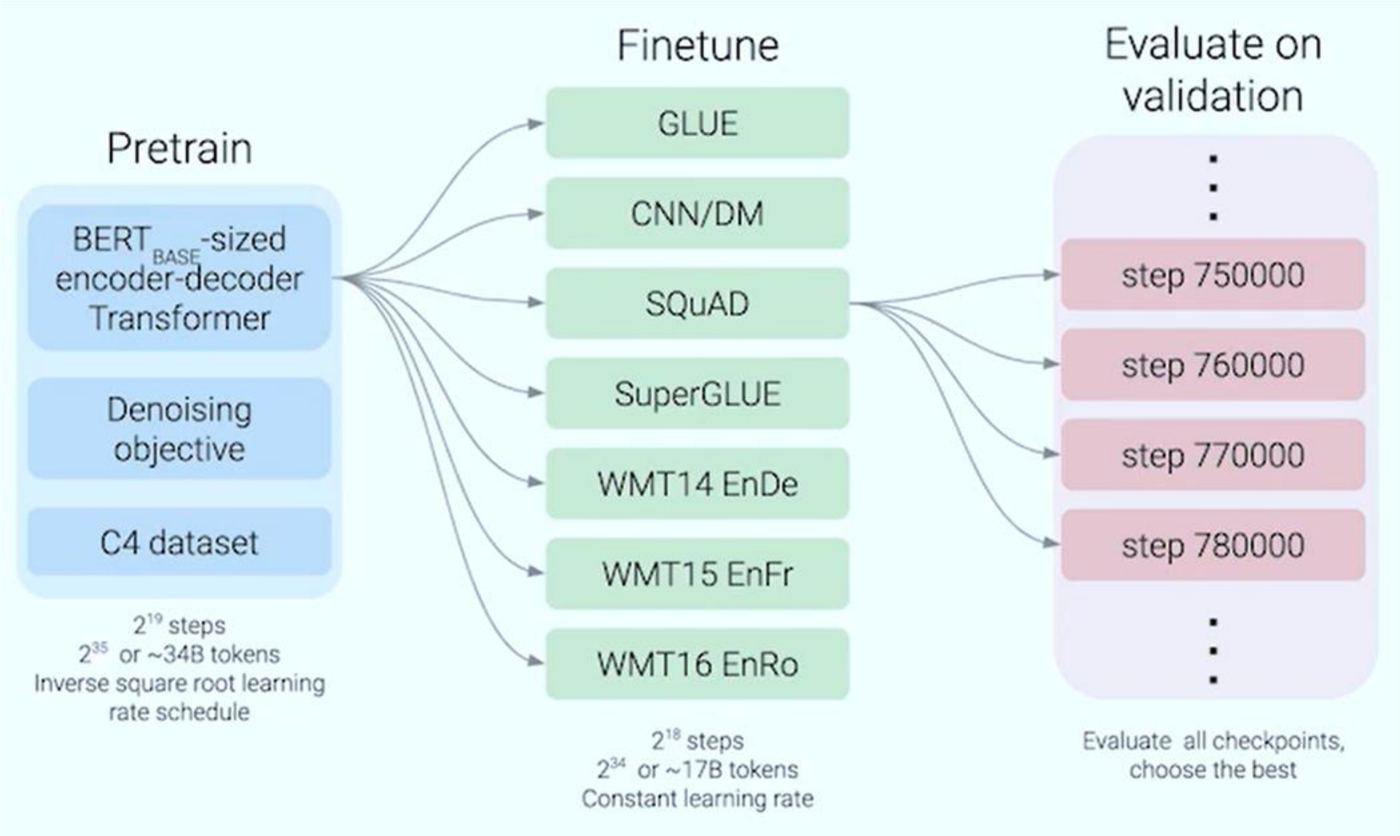
Span Corruption in the context of a standard encoder-decoder transformer architecture.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

T5 – Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

- **T5**



The final T5 model is obtained by finetuning the pretrained model on each individual task encoded by the dataset selected. Each finetuning dataset is represented in the text-to-text objective seen earlier.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

Plan for Part I of III of the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

Plan for Part I of III of the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

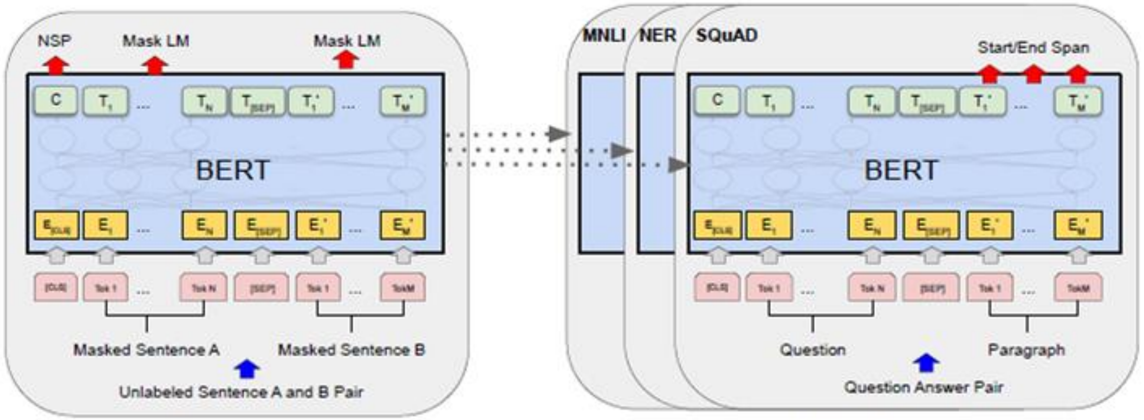
BERT – Pre-training of deep bidirectional transformers for language understanding

- **Key idea**
  - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**)
    - BERT is for pretraining Transformers Encoder.

References
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)..

BERT – Pre-training of deep bidirectional transformers for language understanding

- **Key idea**
  - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**)
    - BERT is for pretraining Transformers Encoder.



Pre-training Fine-Tuning

Two stages:
1. pre-training and
2. fine-tuning

References
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)..

BERT – Pre-training of deep bidirectional transformers for language understanding

- **Pretraining datasets**
  - BooksCorpus
    - A corpus of fiction books from various genres comprising 800M words.
  - English Wikipedia
    - Extracted only the text passages and ignore lists, tables, and headers resulting in 2,500M words.

References
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)..

BERT – Pre-training of deep bidirectional transformers for language understanding

- **Pretraining datasets**
  - BooksCorpus
    - A corpus of fiction books from various genres comprising 800M words.
  - English Wikipedia
    - Extracted only the text passages and ignore lists, tables, and headers resulting in 2,500M words.

Image taken from DeepMind's blog post "A new model and dataset for long-range memory" (link in description)

	Average context (words)	Format	Source	Training Set Size	Vocabulary Size
1B Word	27	Sentences	News	4.15GB	793K
Penn Treebank	355	Articles	WSJ News	5.1MB	10K
WikiText-103	3.6K	Articles	Wikipedia	515MB	267K
PG-19	69K	Books	Books	10.9GB	Open vocabulary

State-of-the-art corpus was 1B words.

References
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)..

BERT – Pre-training of deep bidirectional transformers for language understanding

- **Evaluations**
  - At the time of its release in October 2018, finetuned BERT-large (340M parameters) could outperform both the state-of-the-art models as well as GPT-1.
  - The BERT models were the first breakthrough showing the power of pretraining objectives to obtain strong downstream model performances.

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

References

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)..

Plan for Part I of II of the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

The success of Decoder-only models

The success of Decoder-only models

- Autoregressive vs. Autoencoder Architectures:
    - Decoder-only models, such as GPT (Generative Pre-trained Transformer) models, are autoregressive, which means they generate text or sequences one token at a time from left to right. This autoregressive nature is well-suited for a wide range of language generation tasks like text completion, generation, and translation.

The success of Decoder-only models

- Autoregressive vs. Autoencoder Architectures:
  - Decoder-only models, such as GPT (Generative Pre-trained Transformer) models, are autoregressive, which means they generate text or sequences one token at a time from left to right. This autoregressive nature is well-suited for a wide range of language generation tasks like text completion, generation, and translation.
  - Encoder-only models, like BERT (Bidirectional Encoder Representations from Transformers), are designed as autoencoders, focusing on capturing bidirectional contextual embeddings. While this is beneficial for various downstream tasks, it doesn't directly lend itself to text generation tasks like autoregressive models.

The success of Decoder-only models

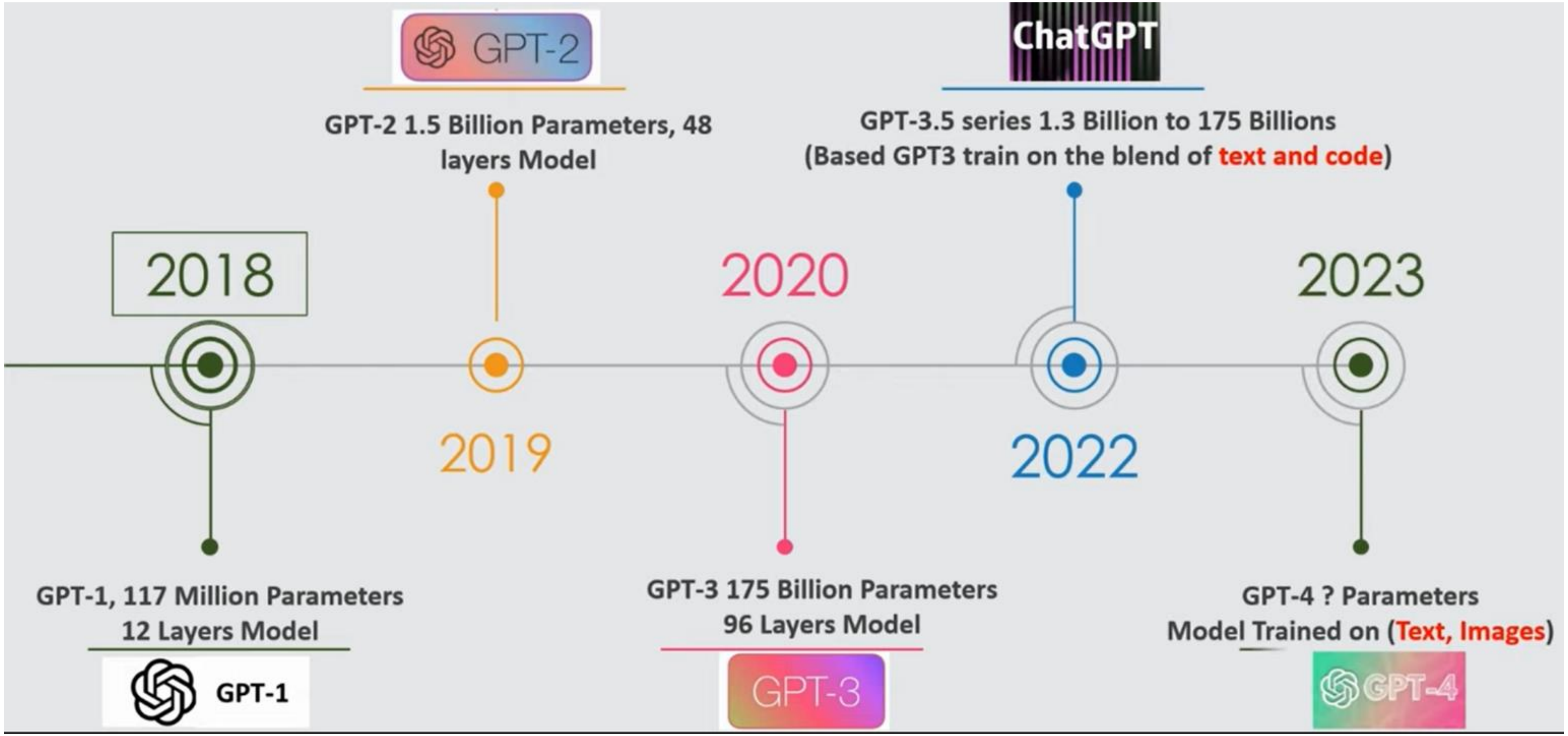
- Autoregressive vs. Autoencoder Architectures:
  - Decoder-only models, such as GPT (Generative Pre-trained Transformer) models, are autoregressive, which means they generate text or sequences one token at a time from left to right. This autoregressive nature is well-suited for a wide range of language generation tasks like text completion, generation, and translation.
  - Encoder-only models, like BERT (Bidirectional Encoder Representations from Transformers), are designed as autoencoders, focusing on capturing bidirectional contextual embeddings. While this is beneficial for various downstream tasks, <u>it doesn't directly lend itself to text generation tasks like autoregressive models</u>.

Decoder-only models because of their text-generation objective can perform with more versatility in the case of downstream tasks compared to encoder-only models.

Plan for the Talk

- Browse pre-training datasets for models categorized by their pre-training architectures
  - Decoder-only
  - Encoder-only
  - Encoder-Decoder

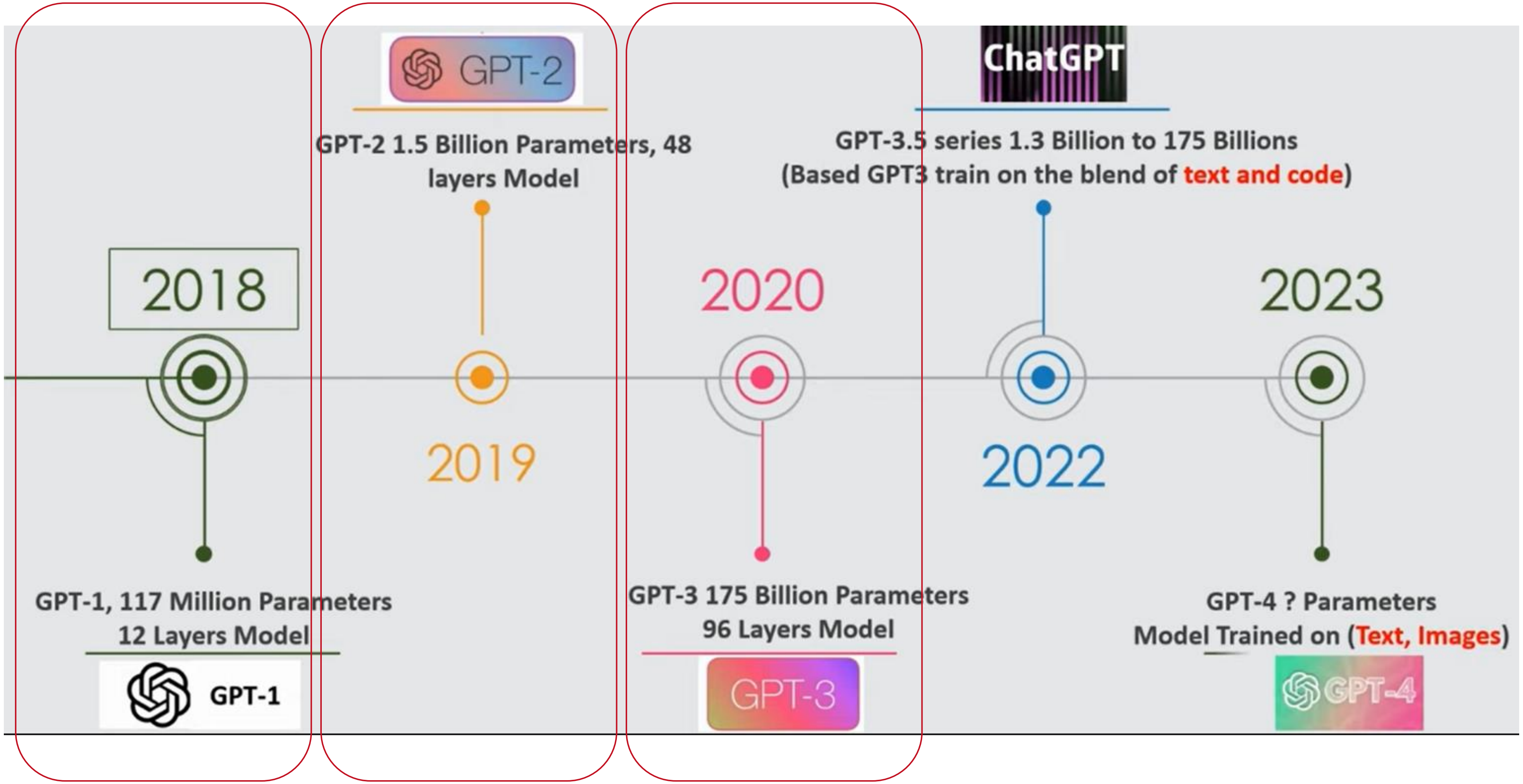
Decoder-only Pretraining Architecture: GPT-series



References

1. GPT-1 to GPT-4: The Evolution of AI Language Models https://www.youtube.com/watch?v=dNFC57Bz10c

Decoder-only Pretraining Architecture: GPT-series



References

1. GPT-1 to GPT-4: The Evolution of AI Language Models https://www.youtube.com/watch?v=dNFC57Bz10c

GPT-1 – Language Models are Few-Shot Learners

- **Key idea**
  - Application of the **generative pre-training (or decoder) strategy** to produce a general “in a sense” task-agnostic model.
    - *generative pre-training* of a language model on a diverse corpus of unlabeled text and *discriminative fine-tuning* on each specific task
      - task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Key idea**
  - Specific choice of a pre-training corpus that allowed modeling long-range dependencies

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - BookCorpus dataset
    - 7000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - BookCorpus dataset
    - 7000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance

Image taken from DeepMind's blog post "A new model and dataset for long-range memory" (link in description)

	Average context (words)	Format	Source	Training Set Size	Vocabulary Size
1B Word	27	Sentences	News	4.15GB	793K
Penn Treebank	355	Articles	WSJ News	5.1MB	10K
WikiText-103	3.6K	Articles	Wikipedia	515MB	267K
PG-19	69K	Books	Books	10.9GB	Open vocabulary

Shows how the choice of the dataset has an impact on long-range context modeling.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - BookCorpus dataset
    - 7000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance

Image taken from DeepMind's blog post "A new model and dataset for long-range memory" (link in description)

	Average context (words)	Format	Source	Training Set Size	Vocabulary Size
1B Word	27	Sentences	News	4.15GB	793K
Penn Treebank	355	Articles	WSJ News	5.1MB	10K
WikiText-103	3.6K	Articles	Wikipedia	515MB	267K
PG-19	69K	Books	Books	10.9GB	Open vocabulary

Previous language models were looking at 1B Word dataset.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - BookCorpus dataset
    - 7000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance

Image taken from DeepMind's blog post "A new model and dataset for long-range memory" (link in description)

	Average context (words)	Format	Source	Training Set Size	Vocabulary Size
1B Word	27	Sentences	News	4.15GB	793K
Penn Treebank	355	Articles	WSJ News	5.1MB	10K
WikiText-103	3.6K	Articles	Wikipedia	515MB	267K
PG-19	69K	Books	Books	10.9GB	Open vocabulary

Note that it has only 27 words average context length.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - BookCorpus dataset
    - 7000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance
    - The Books Corpus reportedly has 20k average words in context and thus the LLM is able to practice long-range dependency modeling resulting in a more effective downstream model.

Image taken from DeepMind's blog post "A new model and dataset for long-range memory" (link in description)

	Average context (words)	Format	Source	Training Set Size	Vocabulary Size
1B Word	27	Sentences	News	4.15GB	793K
Penn Treebank	355	Articles	WSJ News	5.1MB	10K
WikiText-103	3.6K	Articles	Wikipedia	515MB	267K
PG-19	69K	Books	Books	10.9GB	Open vocabulary

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
Zhu, Yukun, et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." *Proceedings of the IEEE international conference on computer vision*. 2015.

GPT-1 – Language Models are Few-Shot Learners

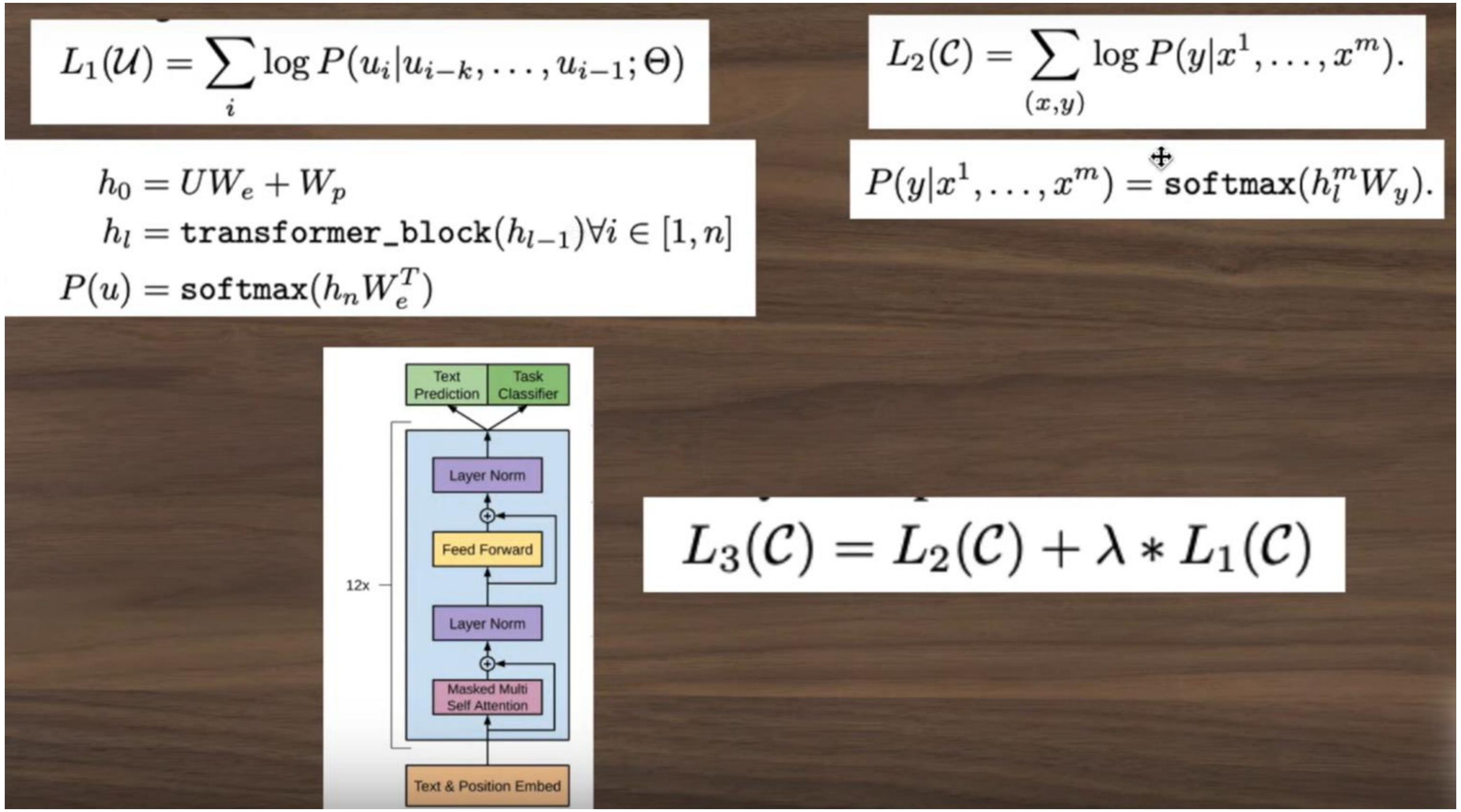
- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning in one pass

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together

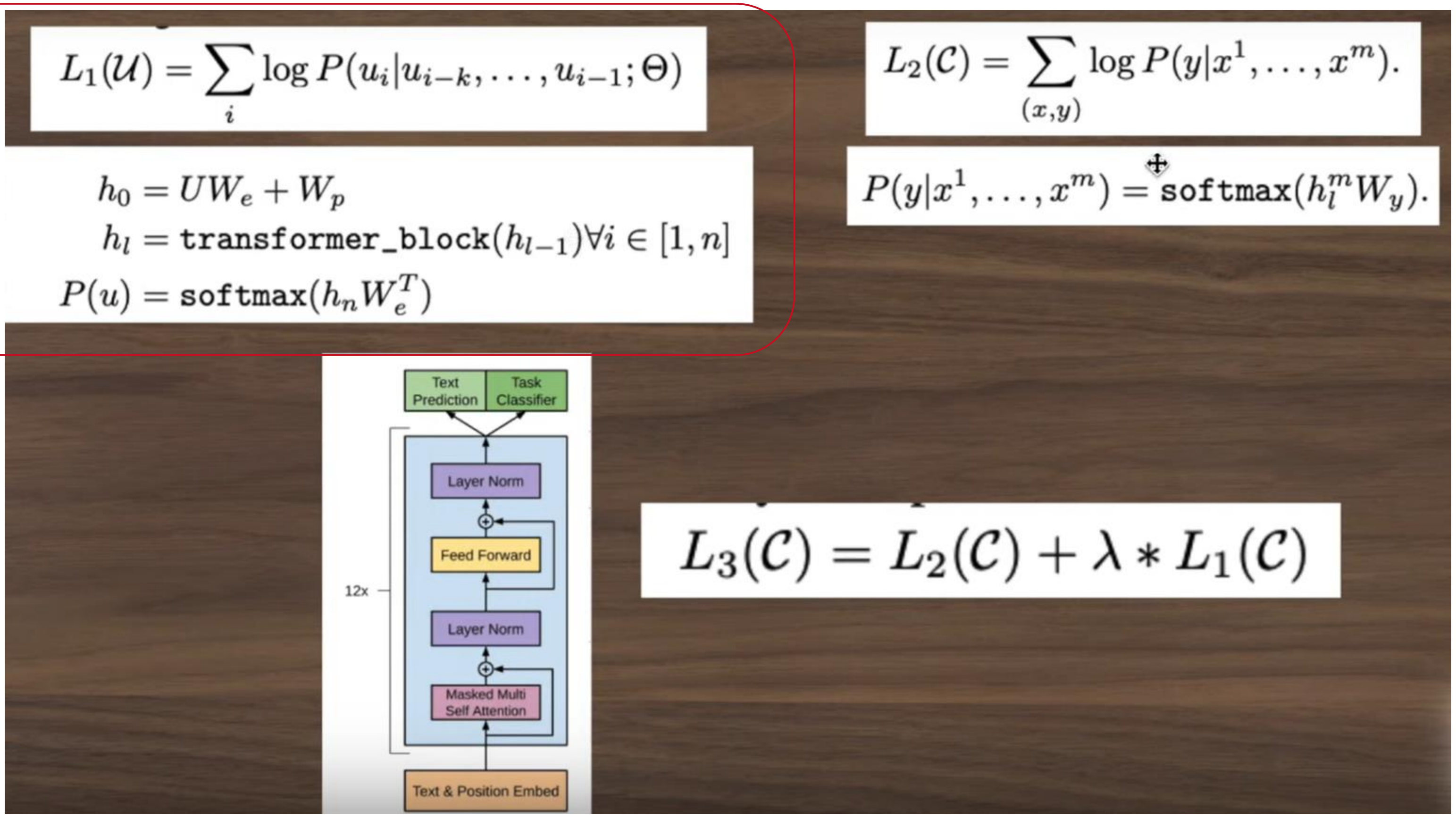


References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together

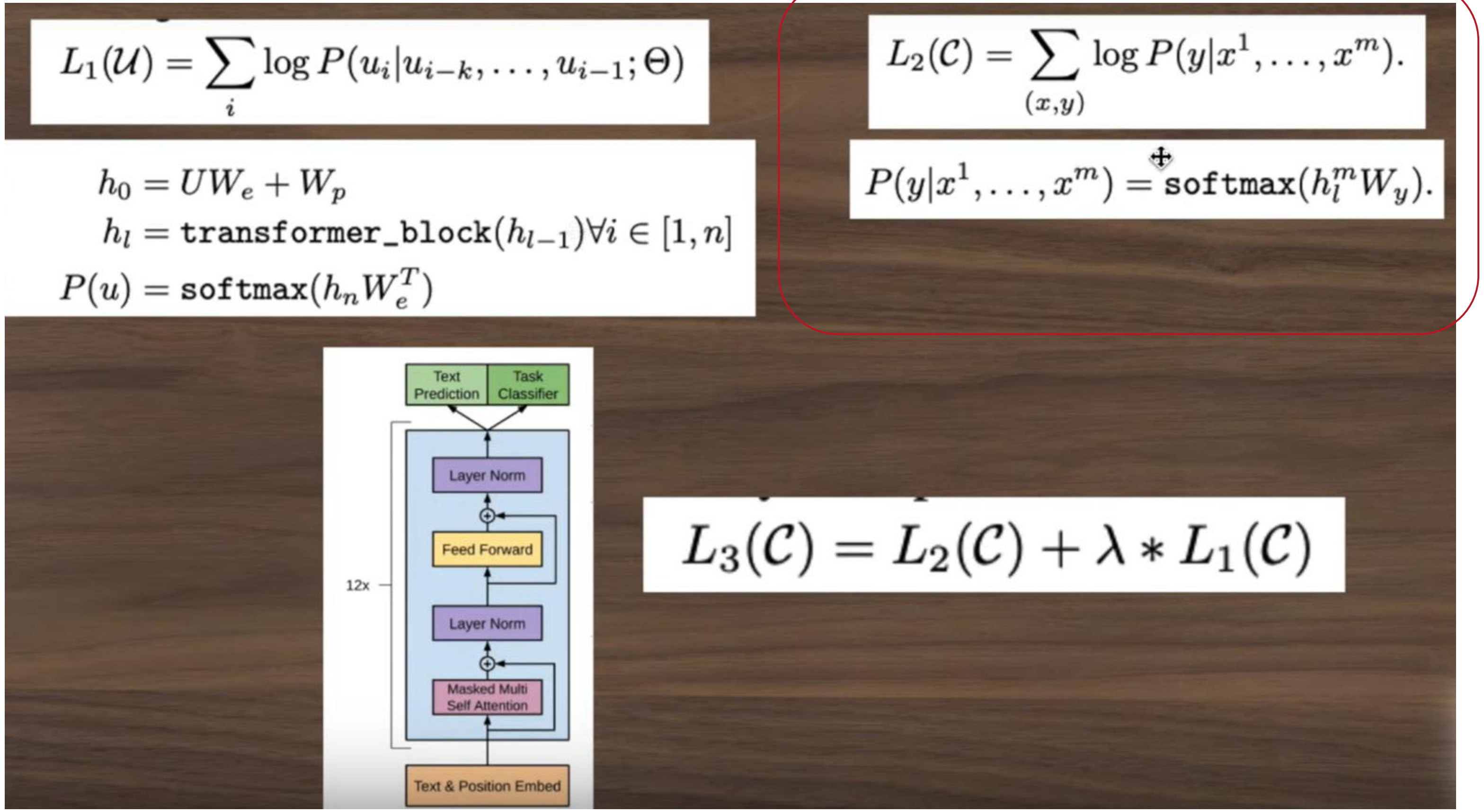


The generative pretraining or the text prediction objective where the next token is predicted given a context,

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together



The discriminative finetuning or the classification task objective.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

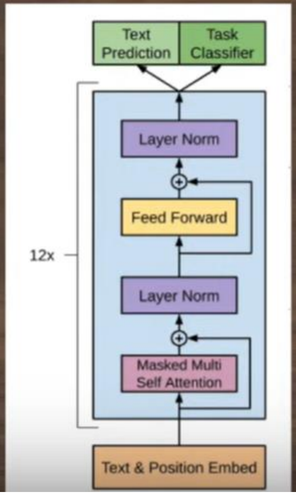
- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y | x^1, \ldots, x^m).$$

$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \texttt{softmax}(h_n W_e^T) \end{aligned}$$

$$P(y | x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$



$$\boxed{L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})}$$

Both objectives are then combined into a joint objective where their parameters are respectively learnt together.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together
    - In the input data, doesn't the classification task need to be represented different than the pretraining task?

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

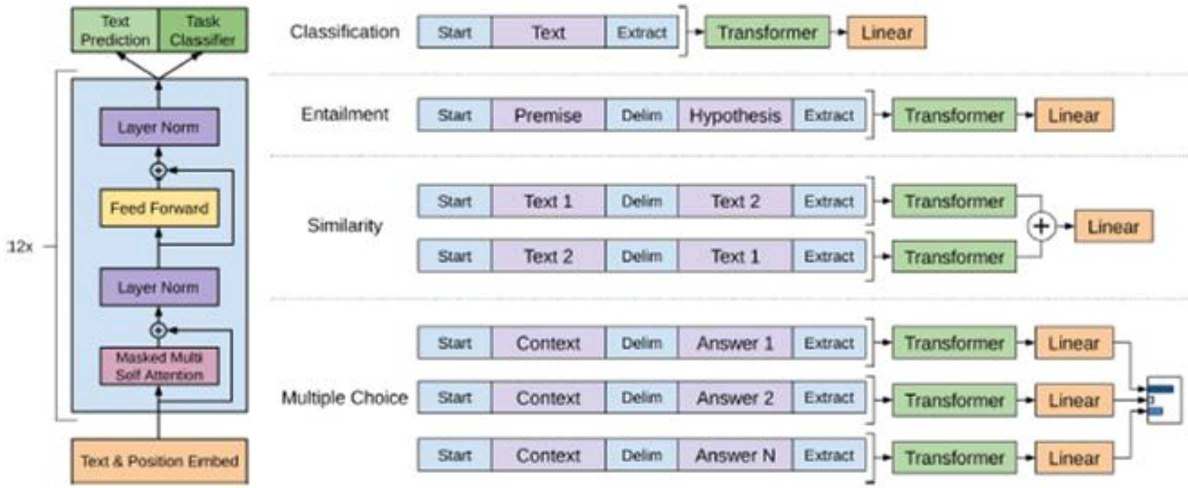
- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together
    - In the input data, doesn’t the classification task need to be represented different than the pretraining task?
      - This is thus a specific contribution of the paper, where the classification tasks are modeled specifically such that they are also a seamless representation of a text generation task.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

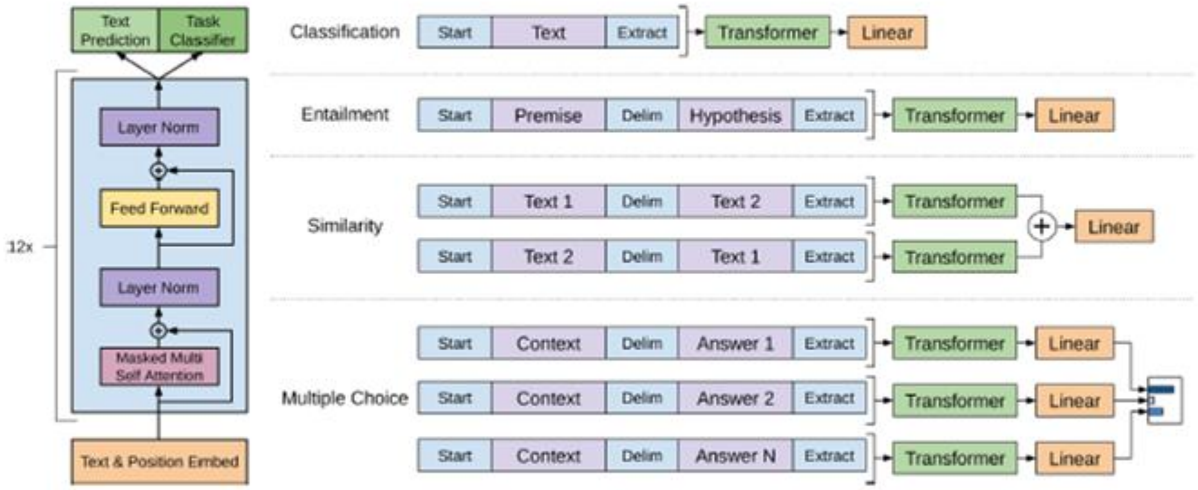
- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together
    - In the input data, doesn't the classification task need to be represented different than the pretraining task?
      - This is thus a specific contribution of the paper, where the classification tasks are modeled specifically such that they are also a seamless representation of a text generation task.



References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together
    - In the input data, doesn't the classification task need to be represented different than the pretraining task?
      - This is thus a specific contribution of the paper, where the classification tasks are modeled specifically such that they are also a seamless representation of a text generation task.



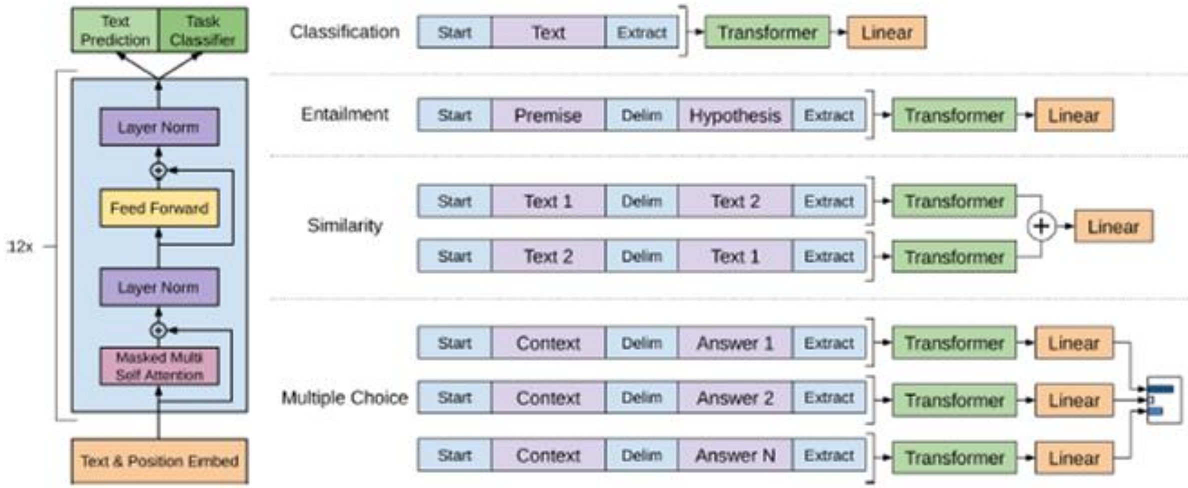
Special structured tokens such as the dollar sign or delimiters are indicators of the classification tasks. Otherwise the model reads the input as a generation task.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-1 – Language Models are Few-Shot Learners

- **GPT-1: Pretraining and Finetuning**
  - Generative Pretraining and Discriminative Finetuning together
    - In the input data, doesn't the classification task need to be represented different than the pretraining task?
      - This is thus a specific contribution of the paper, where the classification tasks are modeled specifically such that they are also a seamless representation of a text generation task.



By utilizing task-specific input adaptations, the pretraining model still processes the structured text input as a <u>single contiguous sequence of tokens.</u>

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Key idea**

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Key idea**
  - Bigger is better!
    - One of the seminal works that demonstrate that language models begin to learn tasks like question answering, machine translation, reading comprehension, and summarization without any explicit supervision when trained on a large-scale–in the order of millions–generic dataset of web pages.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining data sources**
  - Some notes
    - Up until this point, most prior work trained language models on a single domain of text, such as news articles, Wikipedia, or fiction books.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining data sources**
  - Some notes
    - Up until this point, most prior work trained language models on a single domain of text, such as news articles, Wikipedia, or fiction books.
    - The GPT-2 work is one the seminal works to explore a large-scale dataset of a generic nature as web pages.
      - This set the precedent for almost all following work on LLMs that have all relied on web pages.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining data sources**
  - **Web pages**
    - Requirements:
      - There is a lot of messy data on the web, how to retrieve web links that point to legitimate web pages?

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining data sources**
  - **Web pages**
    - Created a web scrape that emphasized web page or document quality per the following strategy:
      - Only scrape web pages curated by humans
        - To do this manually would be forbidding
        - Instead they looked at upvoted reddit posts (maximum 3 karma) with web links and scraped the text from only those web links.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining dataset**
  1. **WebText corpus**
     - Contains slightly over 8 million documents for a total of 40 GB of text

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

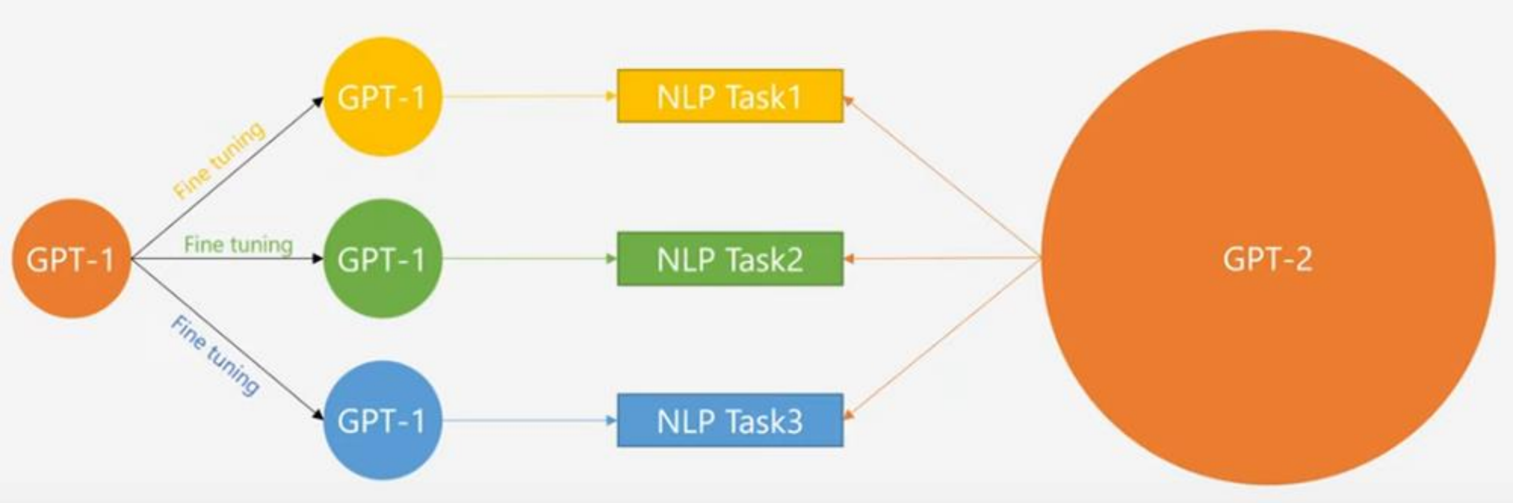
- **Pretraining GPT-2**
  - Only generative pretraining applied

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

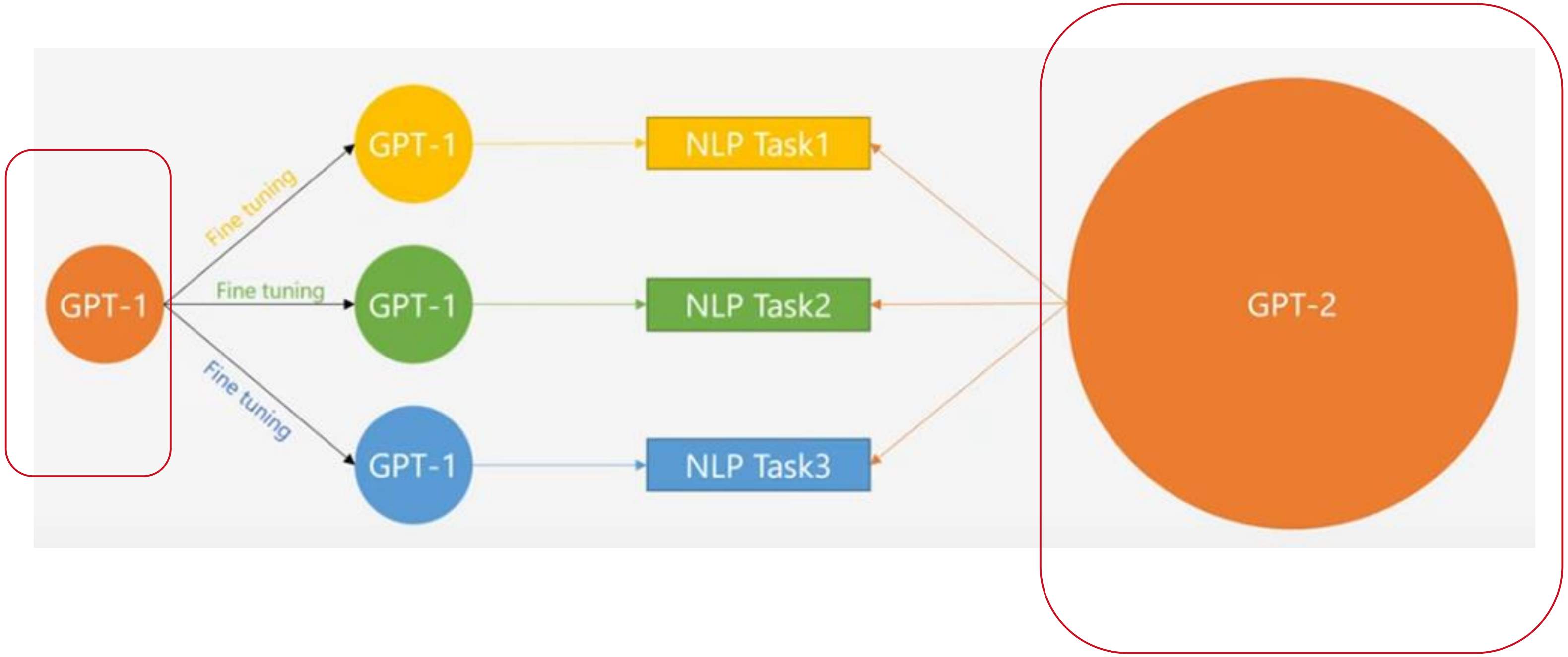
- **Pretraining GPT-2**
  - Only generative pretraining applied



References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied

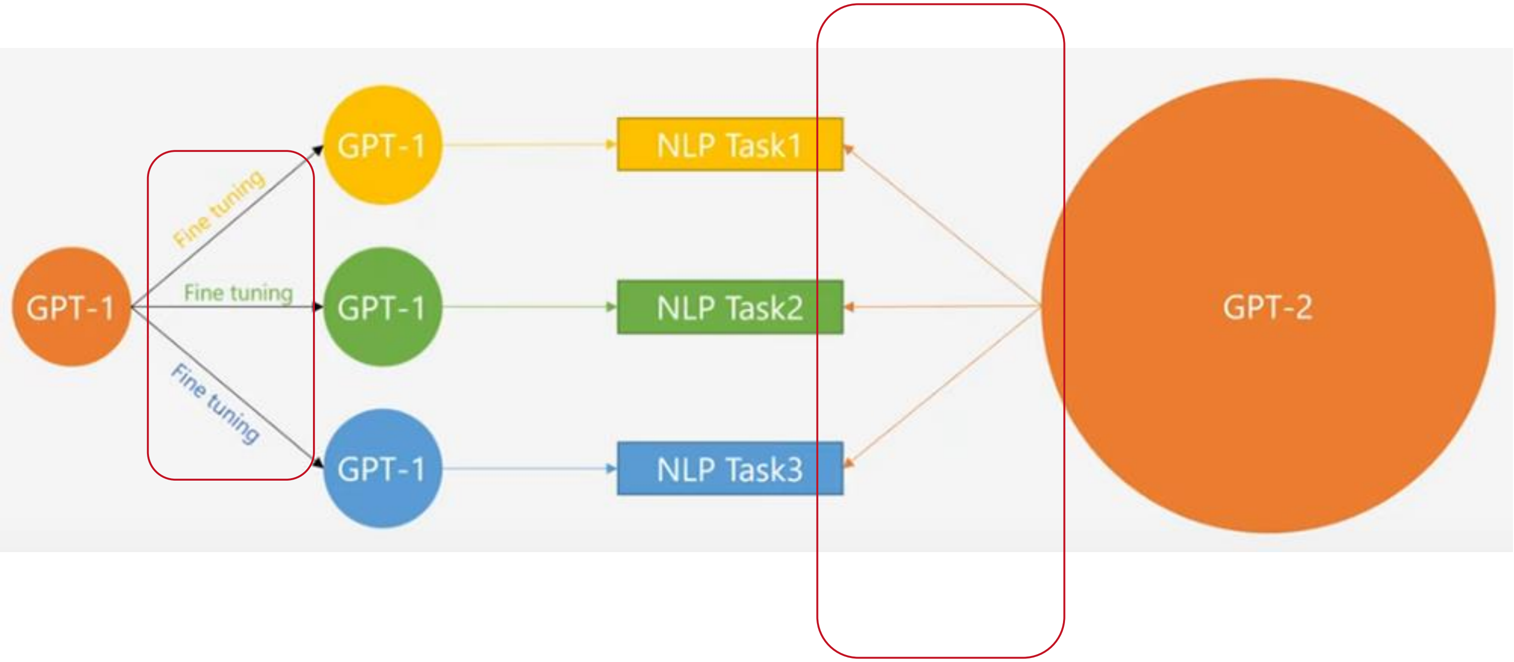


Size Comparison between GPT-2 and GPT-1
- GPT-2 has 10x more parameters than GPT-1

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied



No finetuning in GPT-2

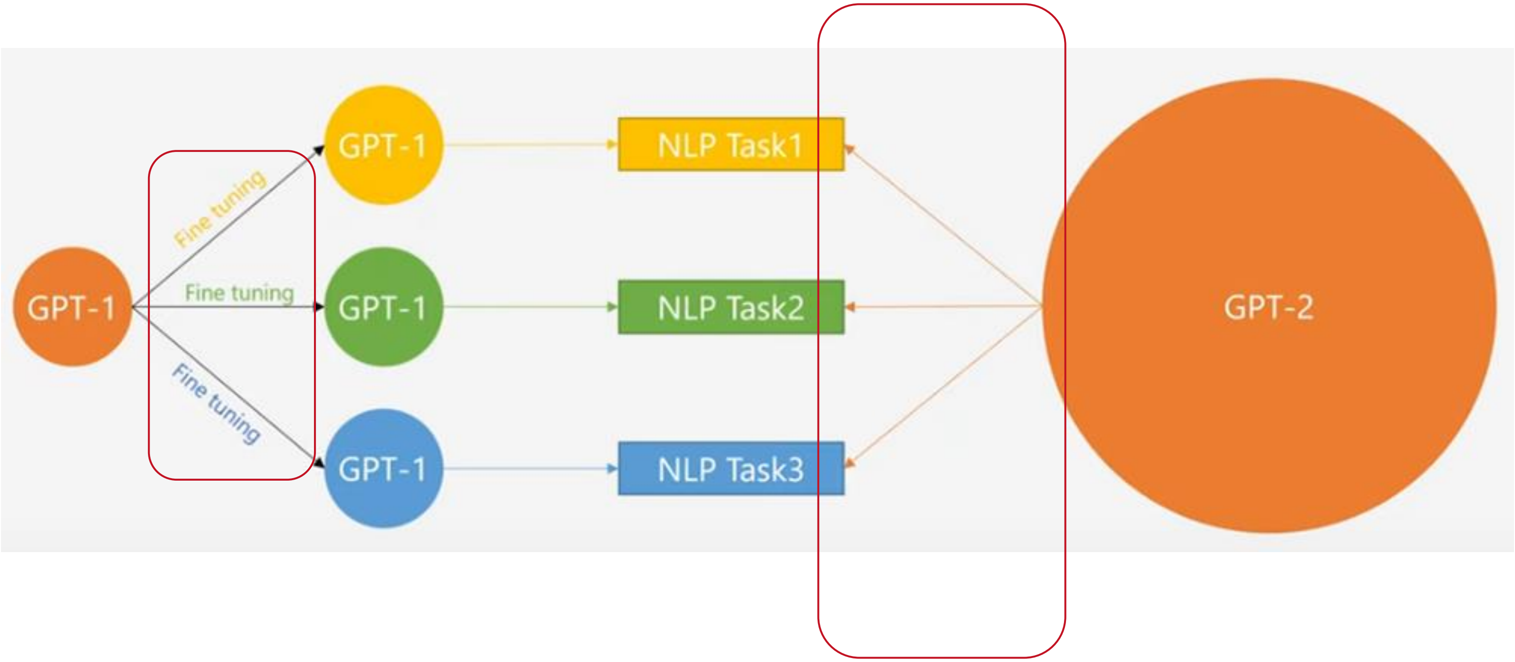
- While GPT-1 was finetuned for downstream tasks, there is no finetuning in GPT-2

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied



Thus GPT-2 is a pure language model.

- How can it perform multiple tasks?

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?

GPT-1	GPT-2
P(output | input)	P(output | input , task)

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?

GPT-1	GPT-2
P(output | input)	P(output | input , task)

- Language modeling objective in GPT-1 emphasizes maximizing the probability of the next token (i.e. output) given the input. To learn tasks, GPT-1 is further fine tuned.

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?

GPT-1	GPT-2
P(output | input)	P(output | input , task)

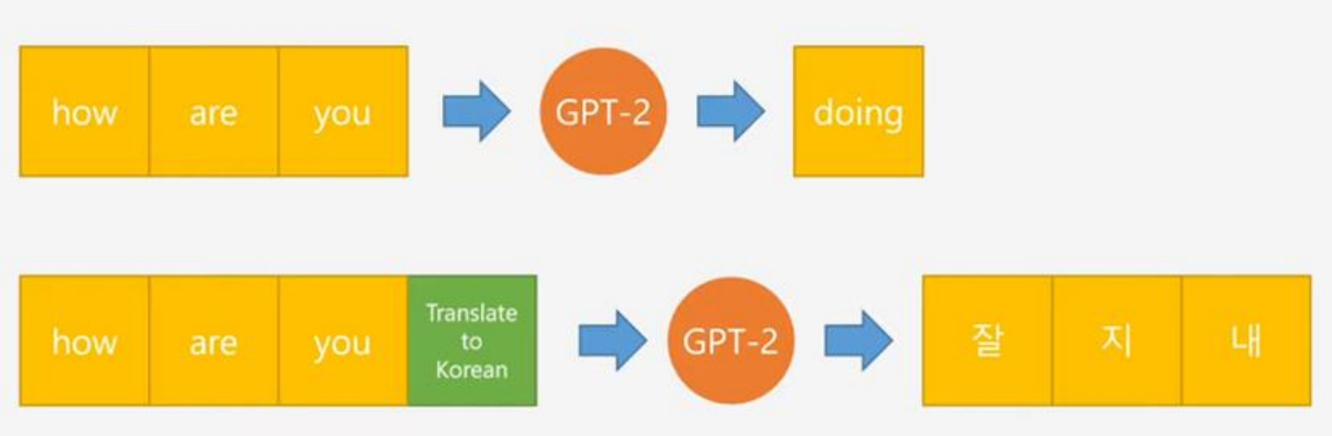
- Language modeling objective in GPT-1 emphasizes maximizing the probability of the next token (i.e. output) given the input. To learn tasks, GPT-1 is further fine tuned.
- In contrast, the language modeling objective in GPT-2 works on maximizing the probability of the next token (i.e. output) given input tokens as well as task specific tokens.

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)



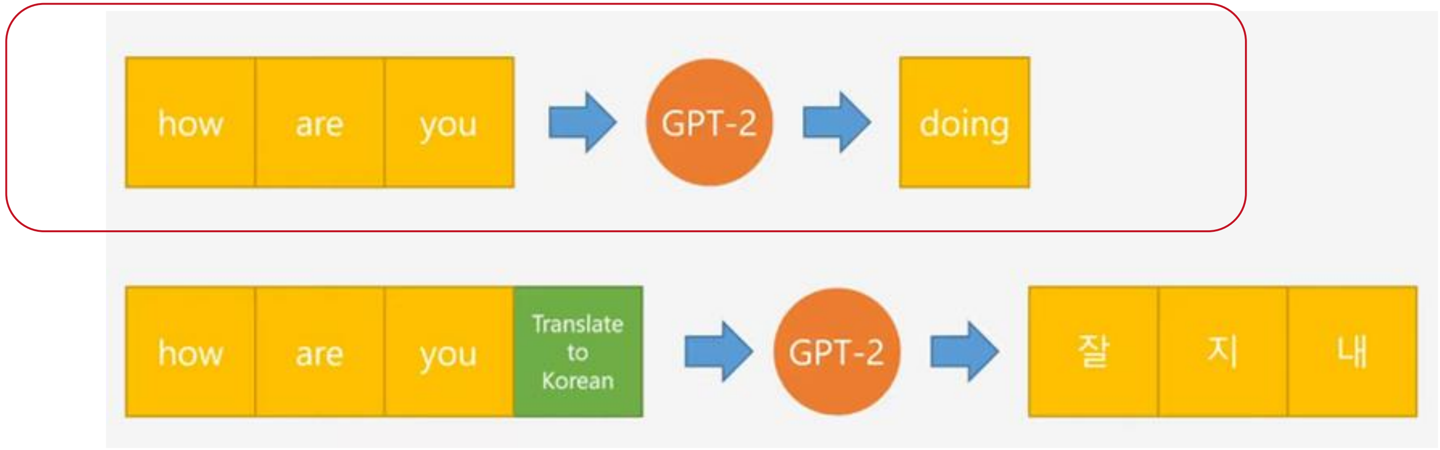
Expected input representation

References

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)

1.



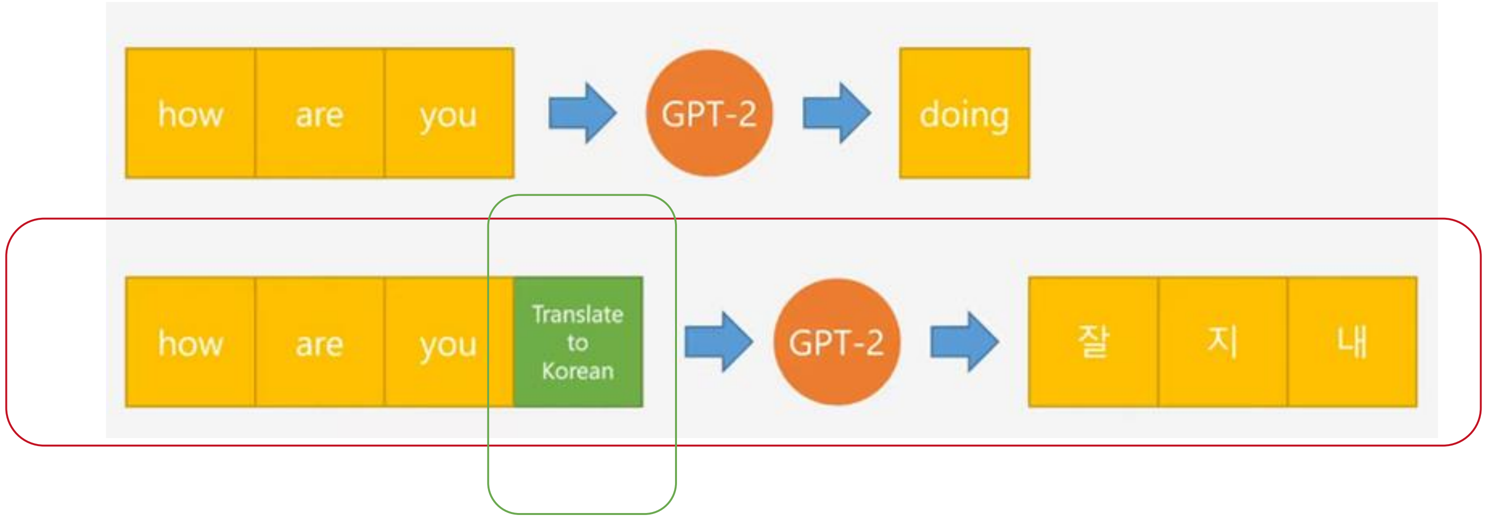
In the first case, the model is expected to interpret the desired task as next token prediction.

References

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)

2.



In the second case, the model is expected to interpret the desired task per the natural language specification highlighted in green.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)

Strategy: unlike GPT-1 there is no task-specific dataset used.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)

Strategy: unlike GPT-1 there is no task-specific dataset used.

Instead the learning is entirely dependent on the patterns in natural language text present in the large-scale dataset of web pages.

References
Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Pretraining GPT-2**
  - Only generative pretraining applied
    - How can GPT-2 work on multiple tasks without fine tuning?
      - P(output | input, task)

The screen highlights several examples of the machine translation task encoded in the running text obtained from web pages.

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: **"Mentez mentez, il en restera toujours quelque chose,"** which translates as, **"Lie lie and something will always remain."**

"I hate the word **'perfume,'"** Burr says. 'It's somewhat better in French: **'parfum.'**

If listened carefully at 29:55, a conversation can be heard between two guys in French: **"-Comment on fait pour aller de l'autre coté? -Quel autre coté?"**, which means **"- How do you get to the other side? - What side?"**.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"Brevet Sans Garantie Du Gouvernement", translated to English: **"Patented without government warranty"**.

Strategy: unlike GPT-1 there is no task-specific dataset used.

Instead the learning is entirely dependent on the patterns in natural language text present in the large-scale dataset of web pages.

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Evaluation**
  - Zero-shot task performance tests

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

This so-called task agnostic model outperforms the finetuned SOTA on most tasks.

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-2 – Language Models are Unsupervised Multitask Learners

- **Evaluation**
  - Zero-shot task performance tests

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Thus they arrive at the finding that "Large Language Models are Unsupervized Multitask Learners"

References

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT-3 – Language Models are Few-Shot Learners

- **Key idea**

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Key idea**
  - Scaling up language models greatly improves **task-agnostic**, **few-shot performance**, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches.
    - One of the first papers to formally introduce the in-context learning strategy via few-shot demonstrations

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Key idea**
  - Scaling up language models greatly improves **task-agnostic**, **few-shot performance**, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches.
    - One of the first papers to formally introduce the in-context learning strategy via few-shot demonstrations
  - Pretraining model architecture is the same as GPT-2. The training dataset size is significantly further increased. Also multiple data genres are introduced.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Pretraining data sources**
  - Web pages
    - Colossal Cleaned Common Crawl Dataset consisting of nearly a trillion words.
      - This dataset was introduced in the T5 paper which was a model released after GPT-2.
  - Other diverse sources
    - added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.
      - expanded version of the WebText dataset first introduced in the GPT-2 paper.
      - two internet-based books corpora (Books1 and Books2)
      - English language Wikipedia
        - Wikipedia is classified better in the encyclopedia text genre rather than web pages.

GPT-3 – Language Models are Few-Shot Learners

- **Pretraining datasets**

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Pretraining datasets**

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

Total available tokens

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Pretraining datasets**

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

Percentage of the total available tokens sampled.

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Pretraining datasets**

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Resulting in a total of 300 billion tokens used to pretrain the model.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Central thesis of the work:** Is this significantly larger model able to perform a task better, not only when it sees task-specific tokens (like GPT-2), but also successful task examples.
  - This is the seminal work to introduce the concept of in-context learning.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Central thesis of the work:** Is this significantly larger model able to perform a task better, not only when it sees task-specific tokens (like GPT-2), but also successful task examples.
  - This is the seminal work to introduce the concept of in-context learning.

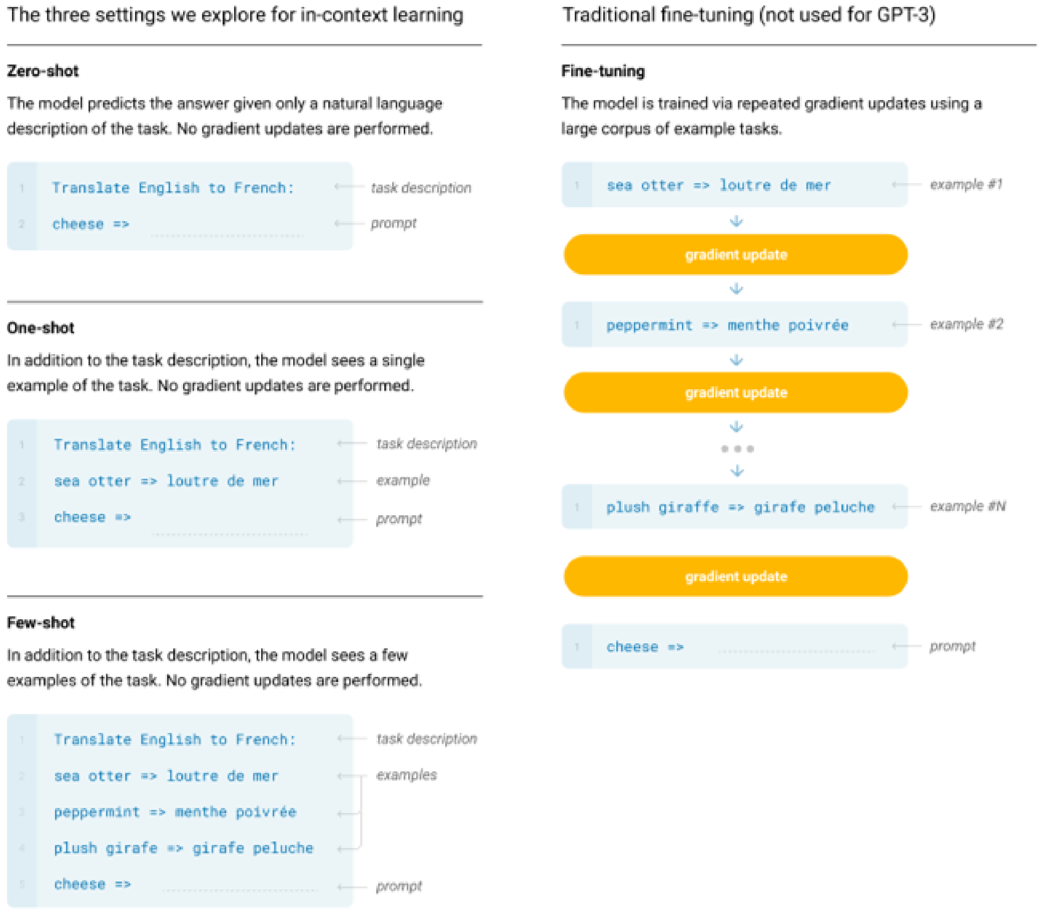
Let’s take a look at the evaluation settings.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- Evaluation Settings

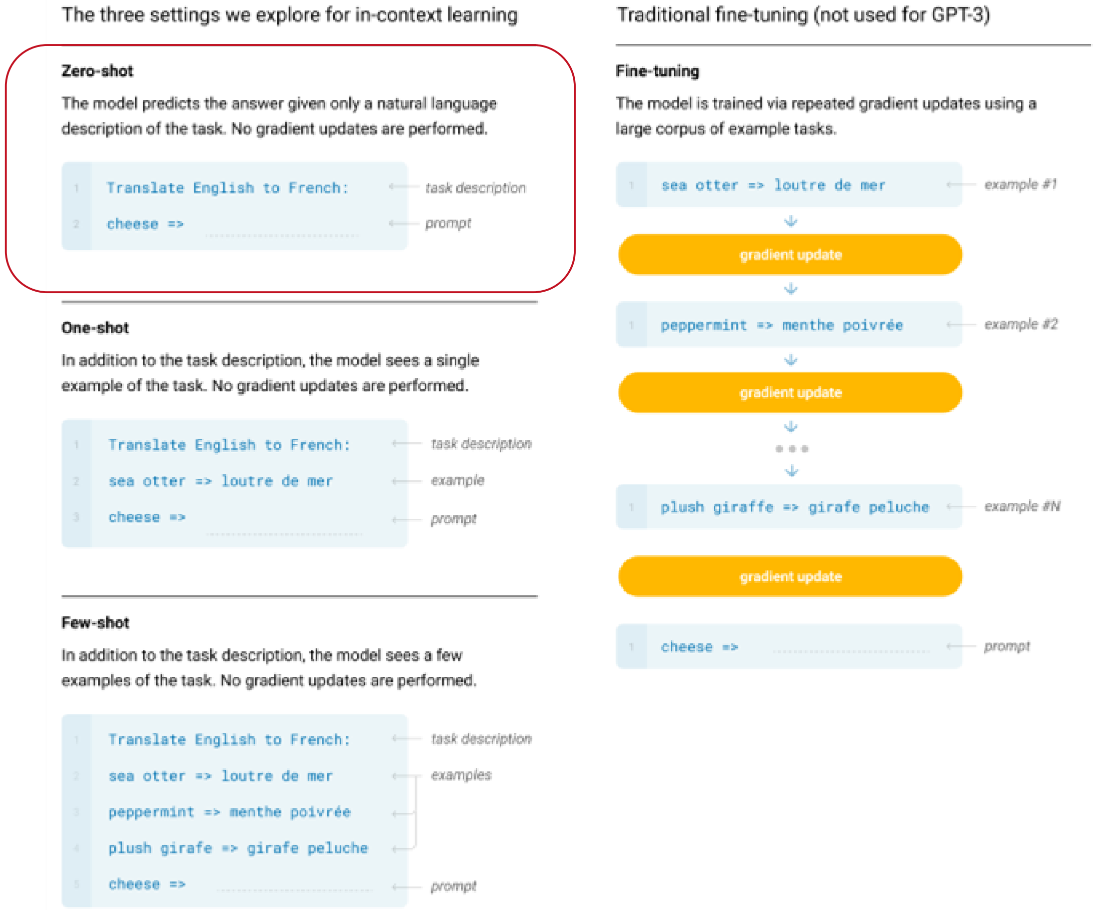


References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- Evaluation Settings

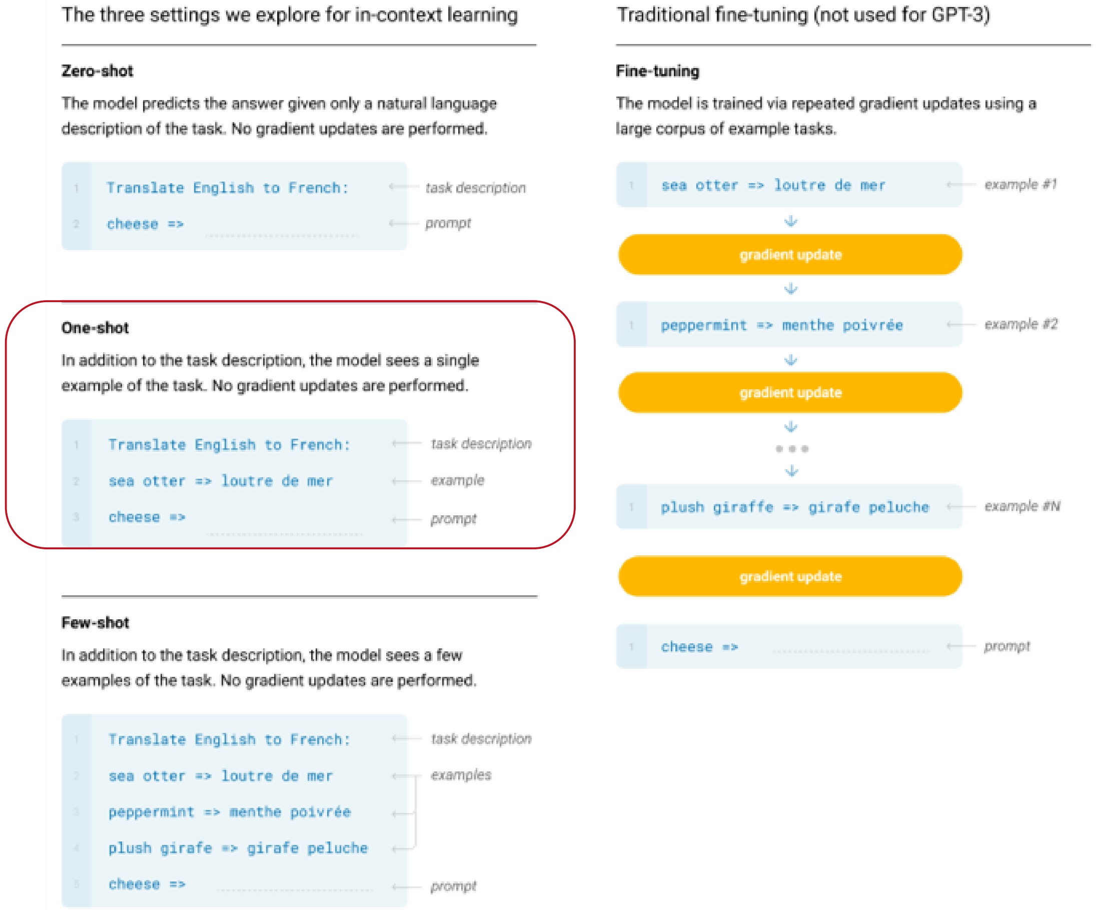


References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- Evaluation Settings

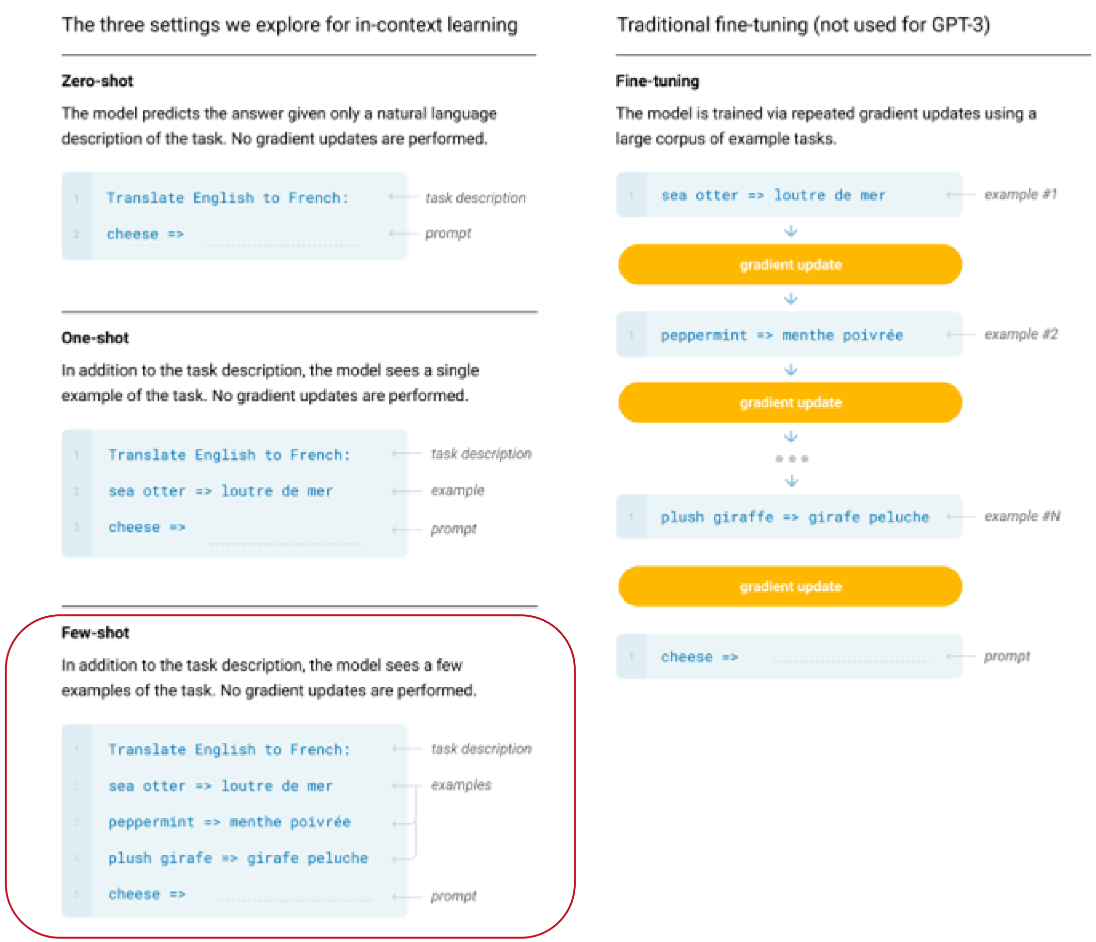


References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluation Settings**

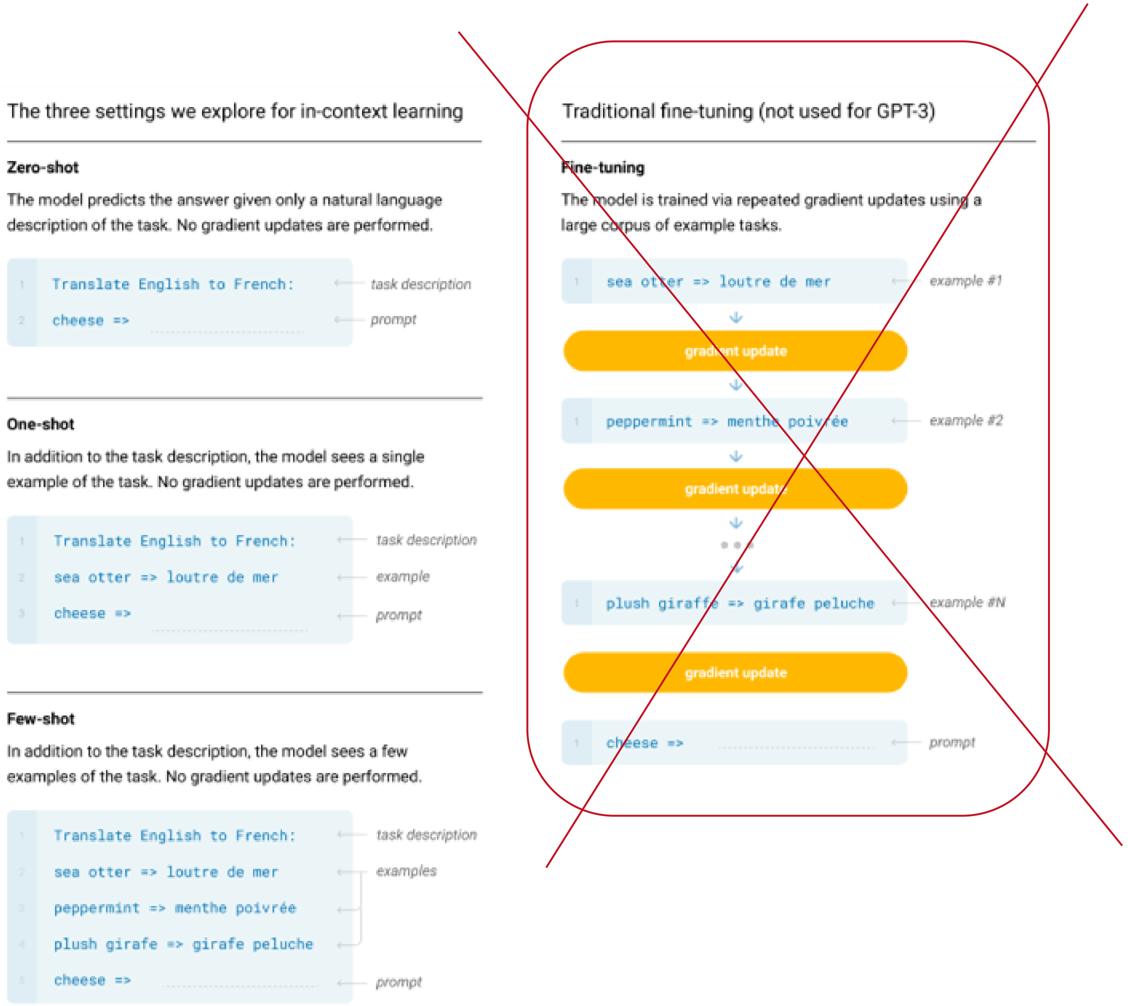


References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluation Settings**



References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluated Models**

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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluated Models**

8 models tested at different parameter sizes and with different layers

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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluated Models**

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}

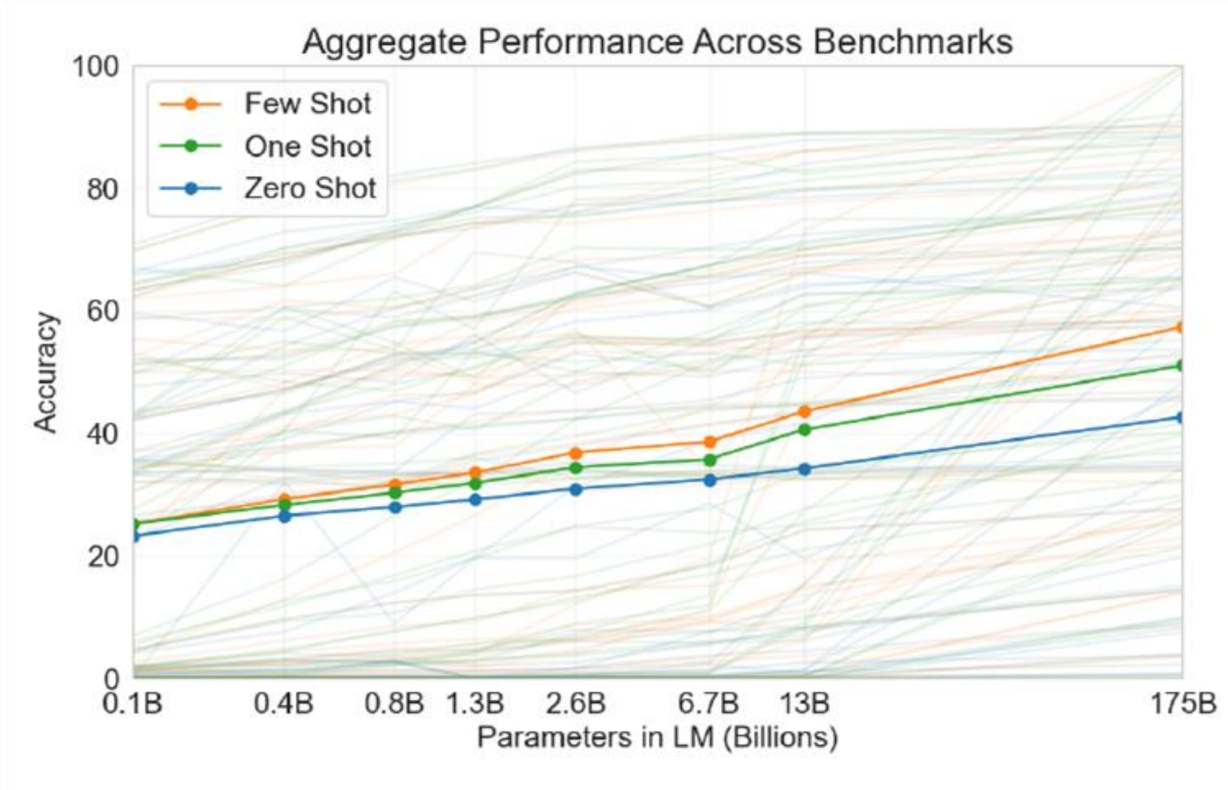
Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Note: 175B GPT-3 is 100x large than 1.5B GPT-2

References
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluation**
  - Are large language models indeed few-shot learners?



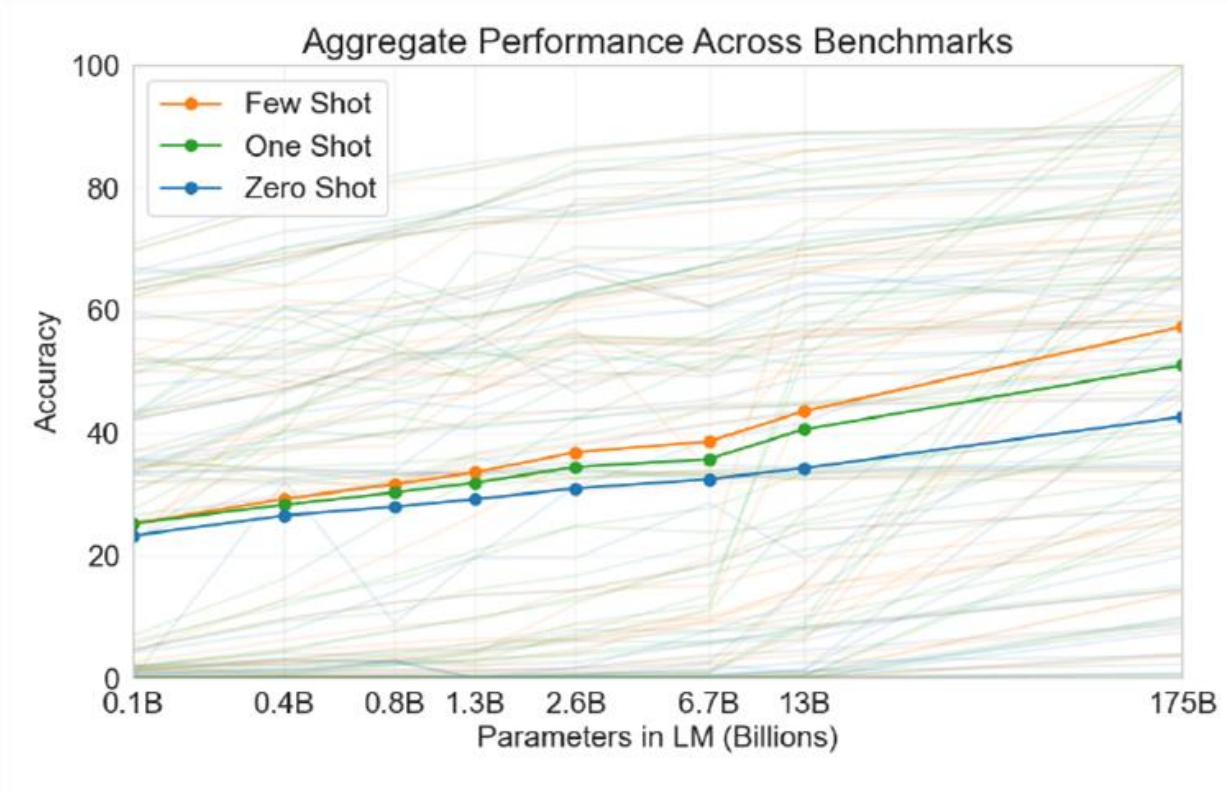
The three different evaluation settings lend themselves to the central thesis of the paper

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluation**
  - Are language models indeed few-shot learners?



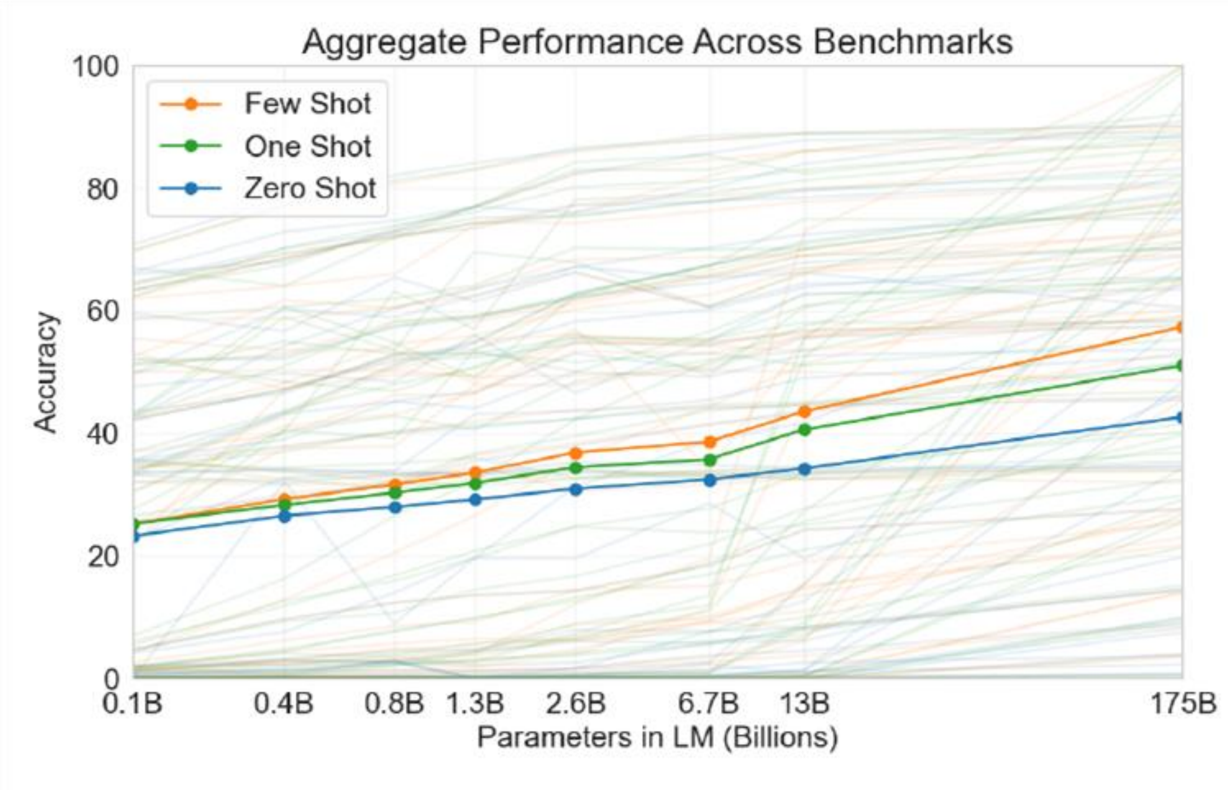
Aggregate performance over 42 accuracy-denominated benchmarks.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

GPT-3 – Language Models are Few-Shot Learners

- **Evaluation**
  - Are language models indeed few-shot learners?



Aggregate performance over 42 accuracy-denominated benchmarks. While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning.

References

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Takeaways from Decoder-only Pretraining Architecture: GPT-series

- Research trajectory: Bigger seems to be better
  - underlying hypothesis: Larger language models encode more parameters that can be interpreted as actual and diverse representations of human language

GPT-J-6B: A 6 billion parameter autoregressive language model

References
https://en.wikipedia.org/wiki/GPT-J
https://huggingface.co/EleutherAI/gpt-j-6b

GPT-J-6B: A 6 billion parameter autoregressive language model

- **Key idea gist**
  - Given a set compute budget, where one cannot obtain a 100B parameter model, can effective LLMs still be produced?
    - Yes! If the underlying pretraining dataset is diverse enough.

References
https://en.wikipedia.org/wiki/GPT-J
https://huggingface.co/EleutherAI/gpt-j-6b

GPT-J-6B: A 6 billion parameter autoregressive language model

- **Key idea**
  - GPT-J was trained on a diverse range of internet text called the **Pile corpus** and is known for its ability to generate high-quality text completions and understand a wide variety of prompts.
    - The central idea for subsequent discussion is the **open-source Pile corpus that modeled diverse text genres presenting itself as a unique, open-accessible resource** for pretraining small-scale LLMs.
  - GPT-J gained popularity as an accessible alternative for researchers and developers who want to experiment with LLMs without the high computational costs typically associated with them.

References
https://en.wikipedia.org/wiki/GPT-J
https://huggingface.co/EleutherAI/gpt-j-6b

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Key idea**
  - ***increased training dataset diversity*** improves general cross-domain knowledge and downstream generalization capability for LLMss.
  - the Pile is an 825 GiB English text corpus targeted constructed from 22 diverse high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**
  - PubMed Central, ArXiv, GitHub, the FreeLaw Project, Stack Exchange, the US Patent and Trademark Office, PubMed, Ubuntu IRC, HackerNews, YouTube, PhilPapers, and NIH ExPorter. Furthermore, as extensions of original corpora i.e. OpenWebText and BookCorpus datasets, they release OpenWebText2 and BookCorpus2, respectively.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

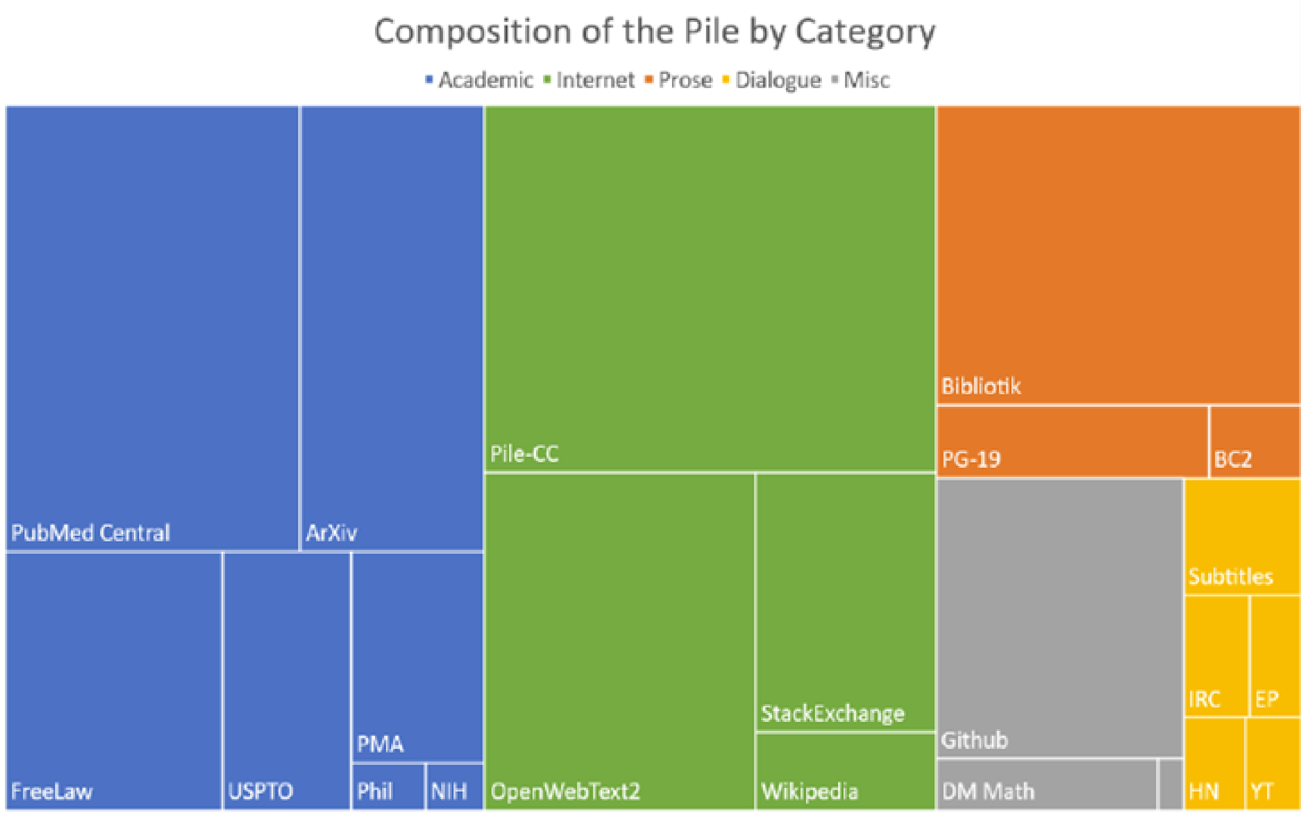


Figure 1: Treemap of Pile components by effective size.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

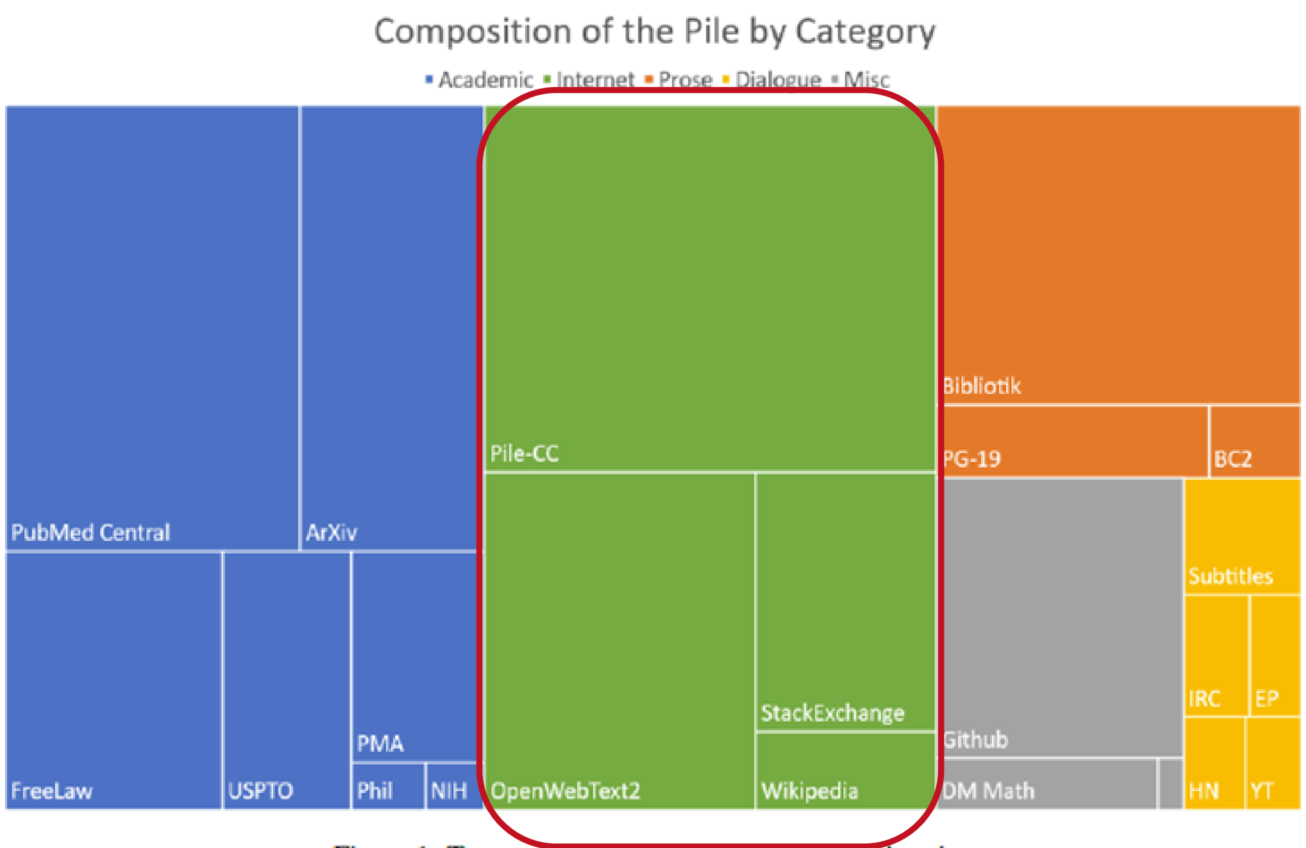


Figure 1: Treemap of Pile components by effective size.

The internet datasets.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

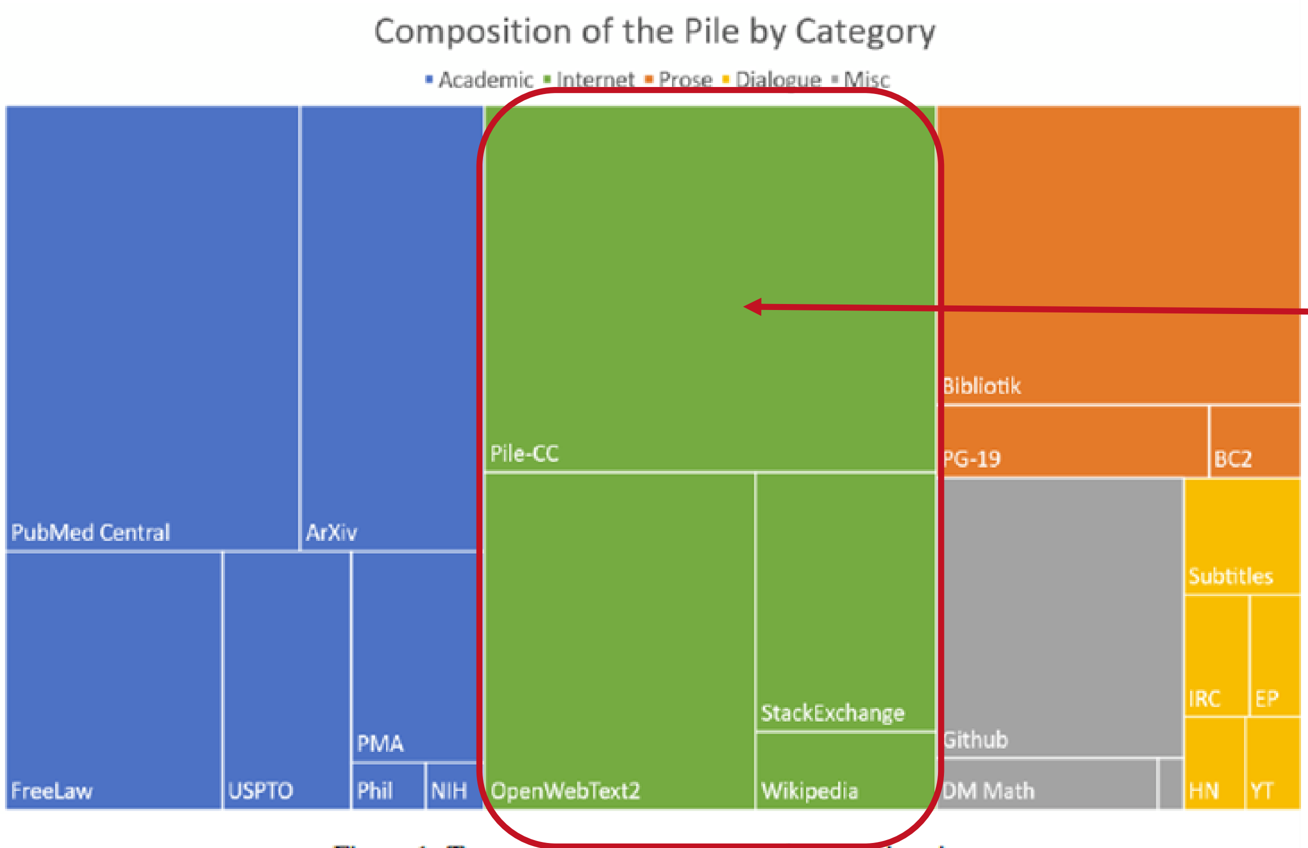


Figure 1: Treemap of Pile components by effective size.

The internet datasets.

Pile-CC: a processed version of Common Crawl using the jusText tool.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

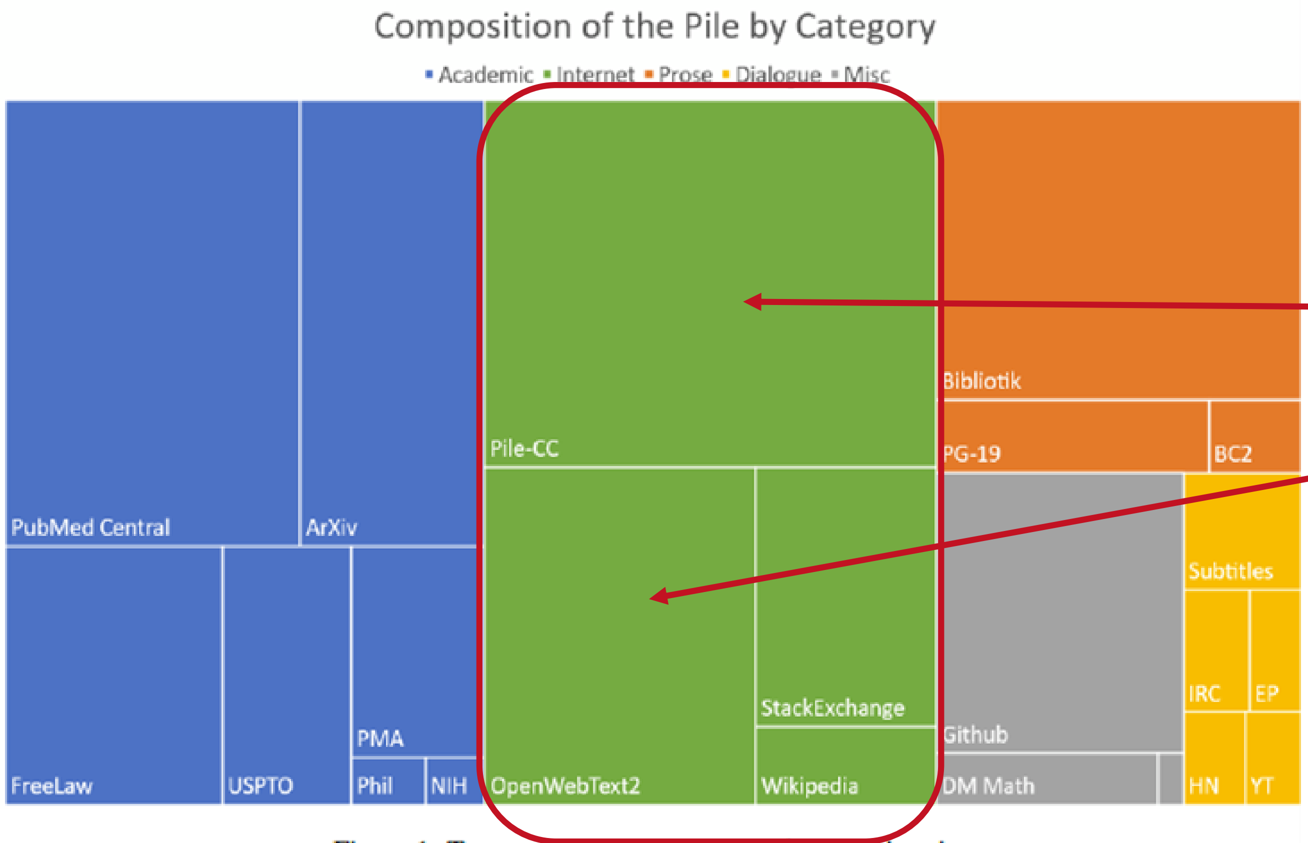


Figure 1: Treemap of Pile components by effective size.

The internet datasets.

Pile-CC: a processed version of Common Crawl using the jusText tool.

OpenWebText2: similar to WebText, scrapes text from web links in upvoted Reddit posts, relies on more recent Reddit submissions.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

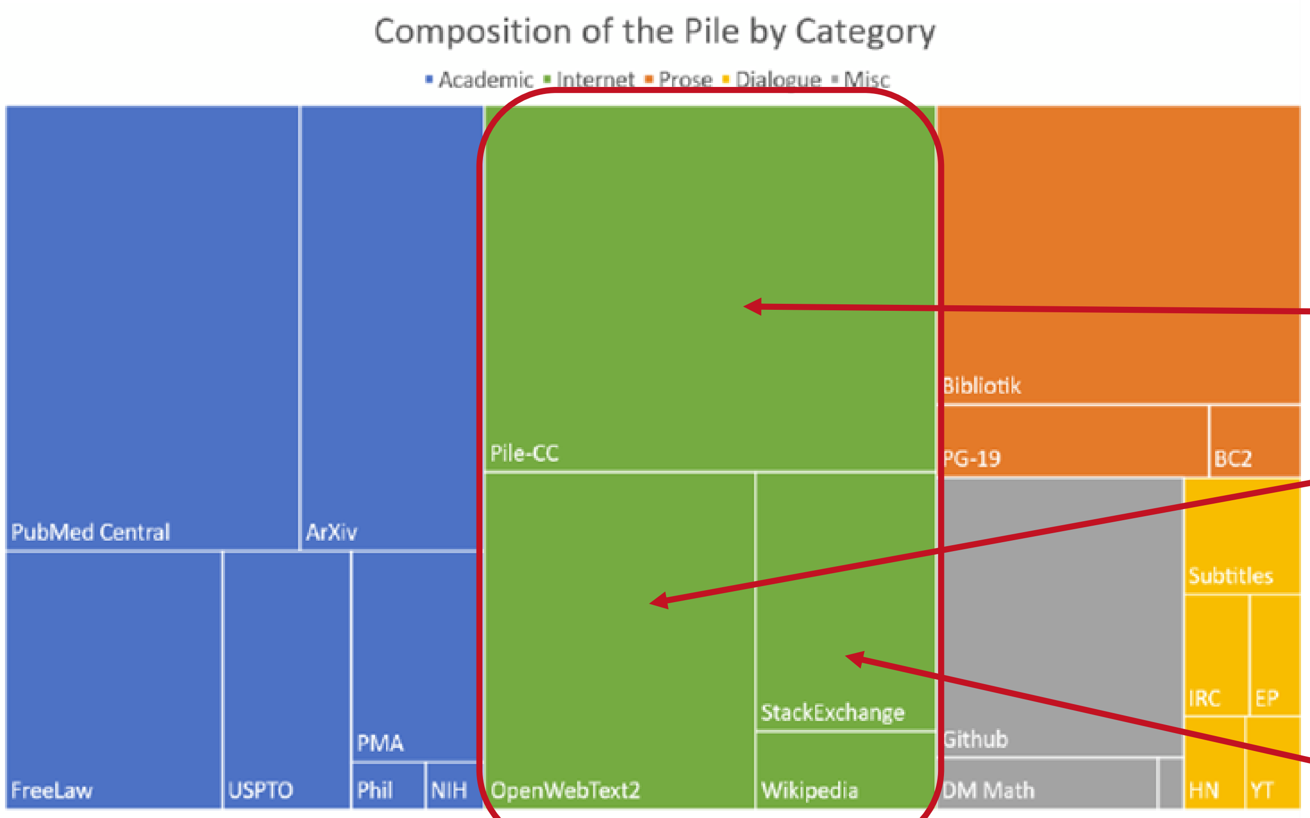


Figure 1: Treemap of Pile components by effective size.

The internet datasets.

Pile-CC: a processed version of Common Crawl using the jusText tool.

OpenWebText2: similar to WebText, scrapes text from web links in upvoted Reddit posts, relies on more recent Reddit submissions.

Stack Exchange: largest publicly available repositories of QA pairs, covering a wide range of subjects—from programming, to gardening, to Buddhism

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

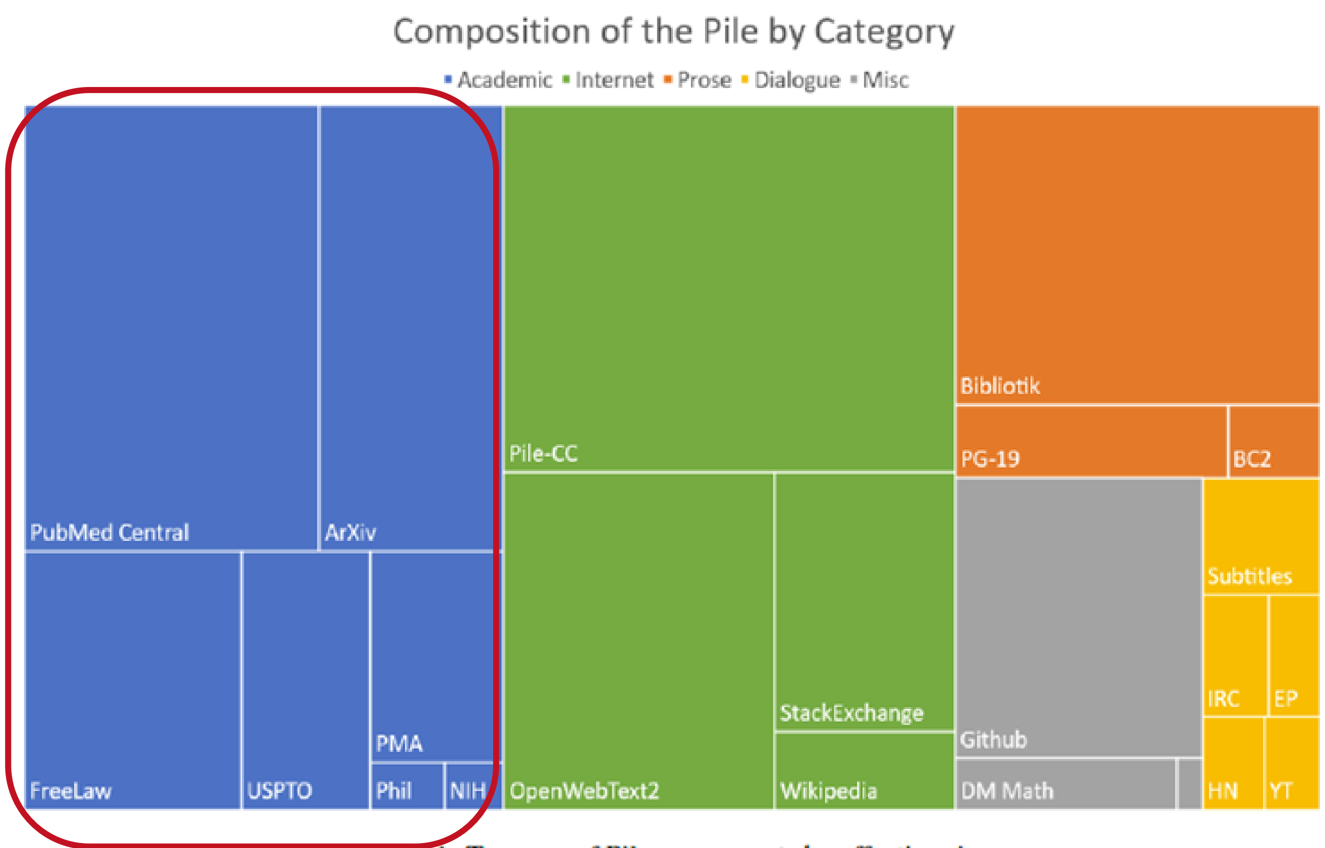


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

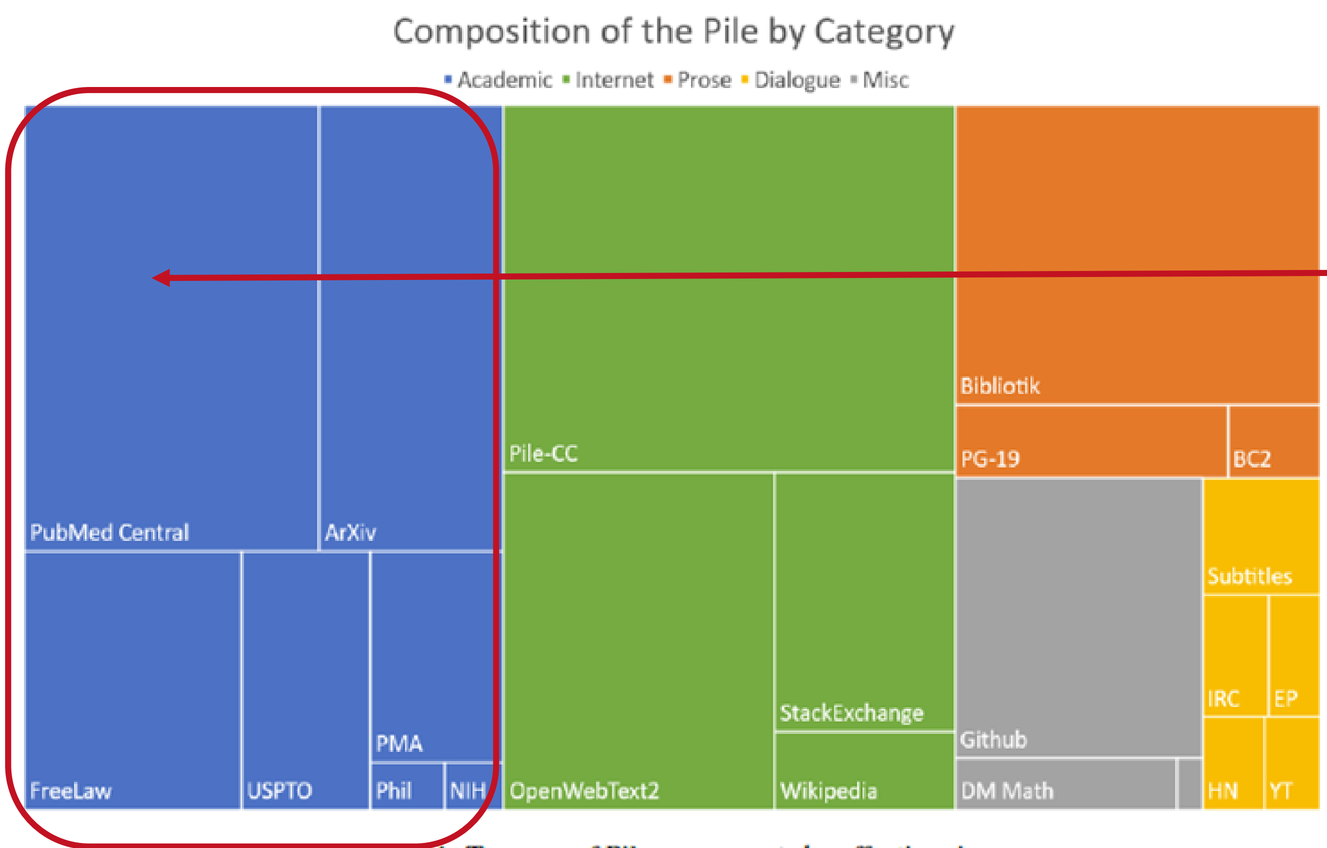


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

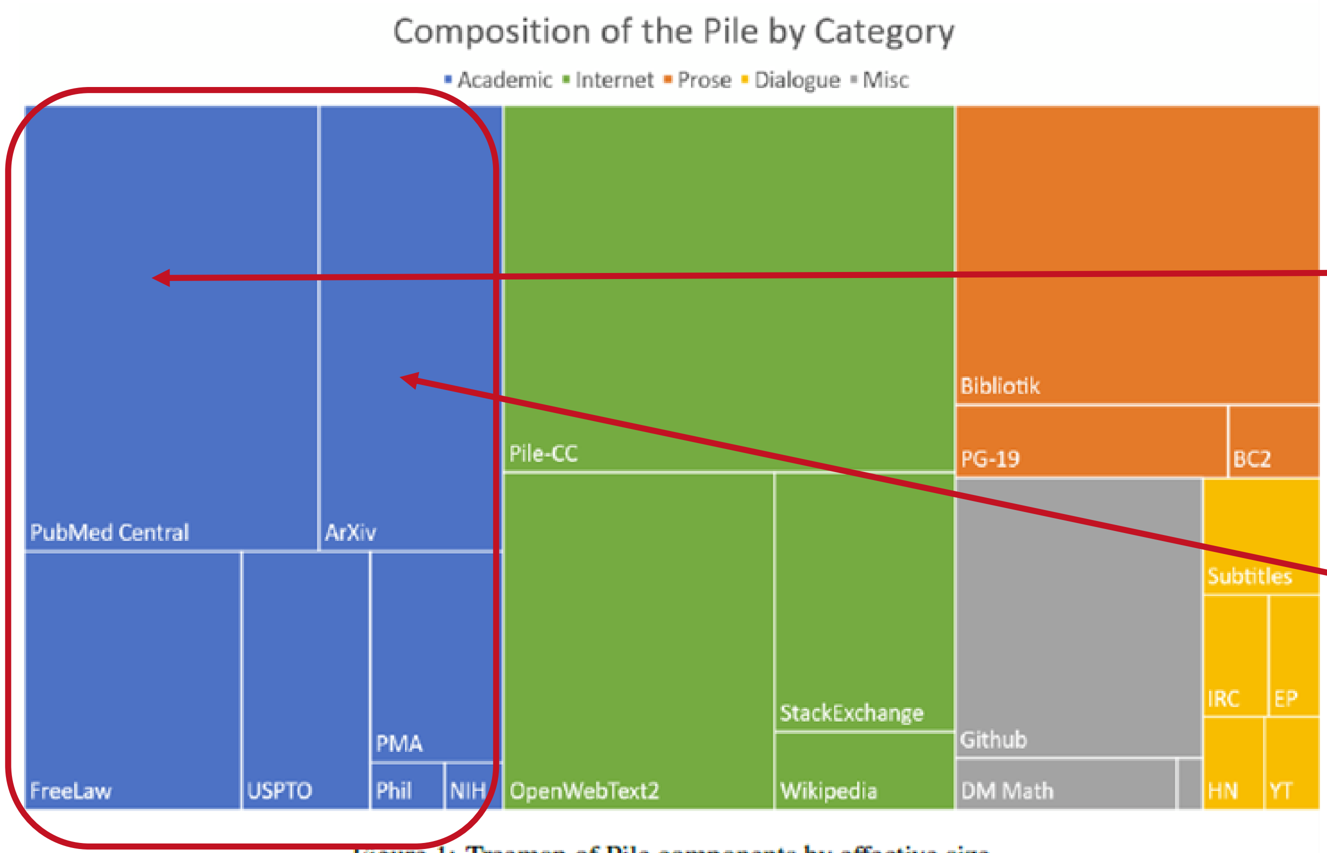


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

arXiv: as a source of high quality text and math knowledge

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

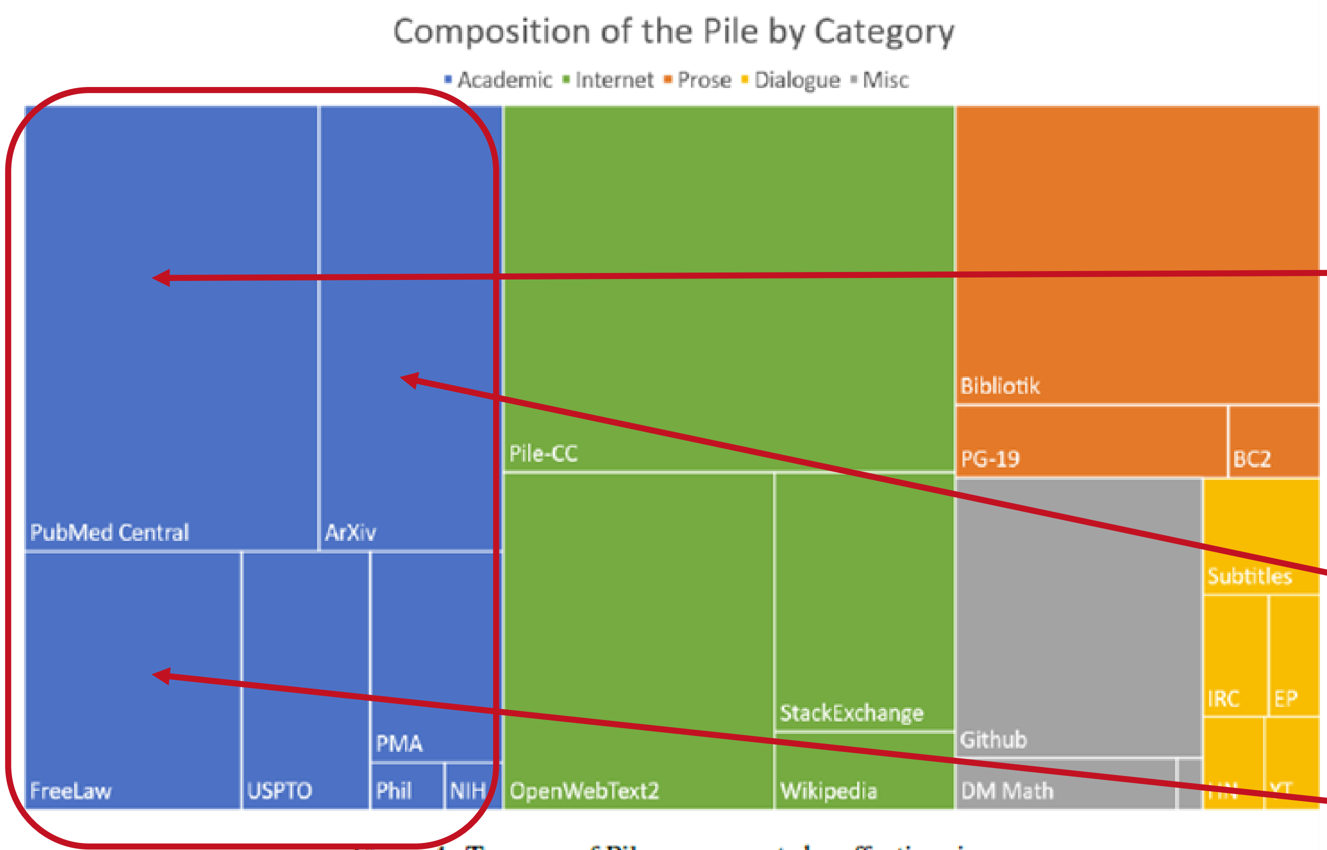


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

arXiv: as a source of high quality text and math knowledge

FreeLaw: as a source supporting academic studies in the legal realm, one of its sources provides bulk downloads for millions of legal opinions from federal and state courts

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

USPTO: background sections of patents granted by the US Patents office.

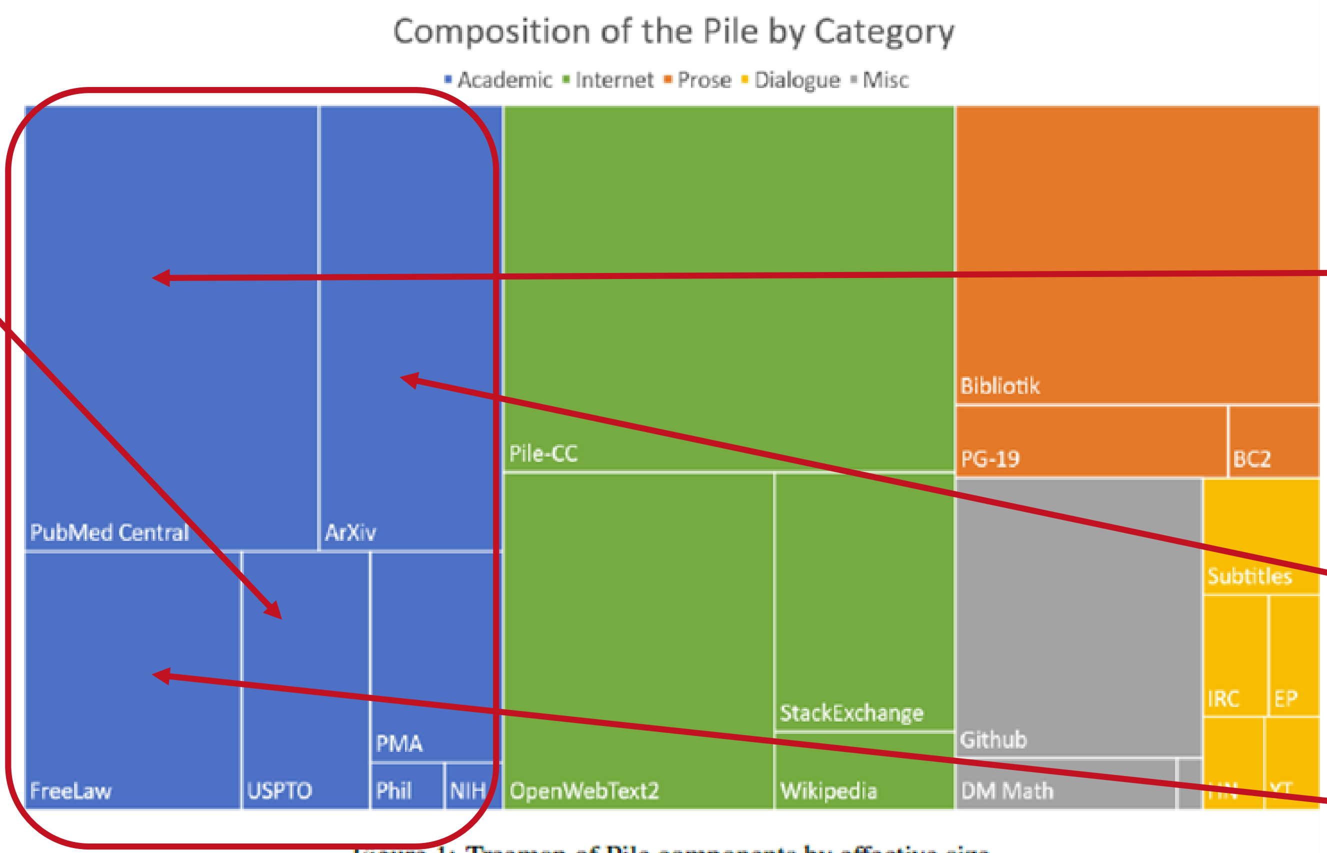


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

arXiv: as a source of high quality text and math knowledge

FreeLaw: as a source supporting academic studies in the legal realm, one of its sources provides bulk downloads for millions of legal opinions from federal and state courts

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

USPTO: background sections of patents granted by the US Patents office.

Phil: OA philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario

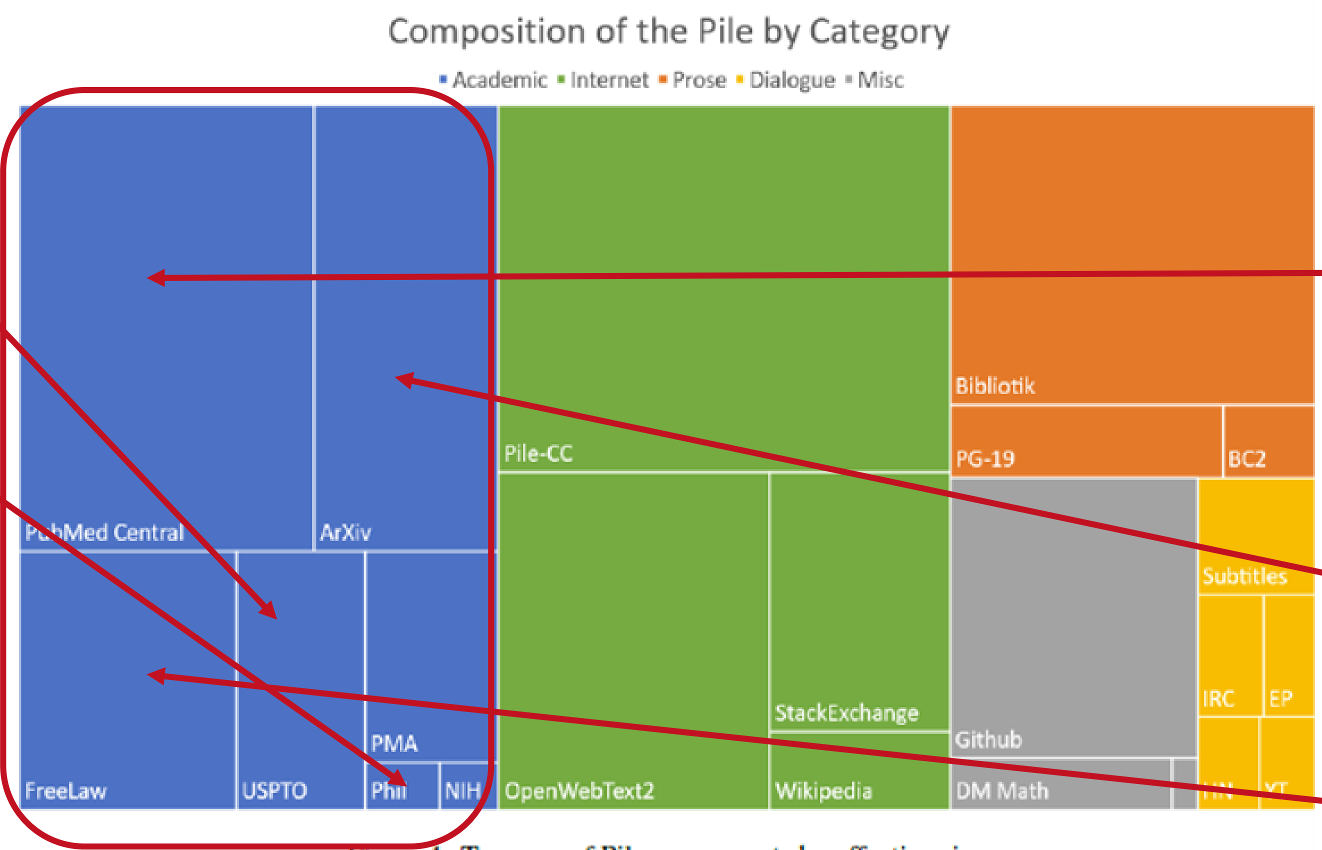


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

arXiv: as a source of high quality text and math knowledge

FreeLaw: as a source supporting academic studies in the legal realm, one of its sources provides bulk downloads for millions of legal opinions from federal and state courts

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

USPTO: background sections of patents granted by the US Patents office.

Phil: OA philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario

The remaining two are still some academic corpora in biomedicine.

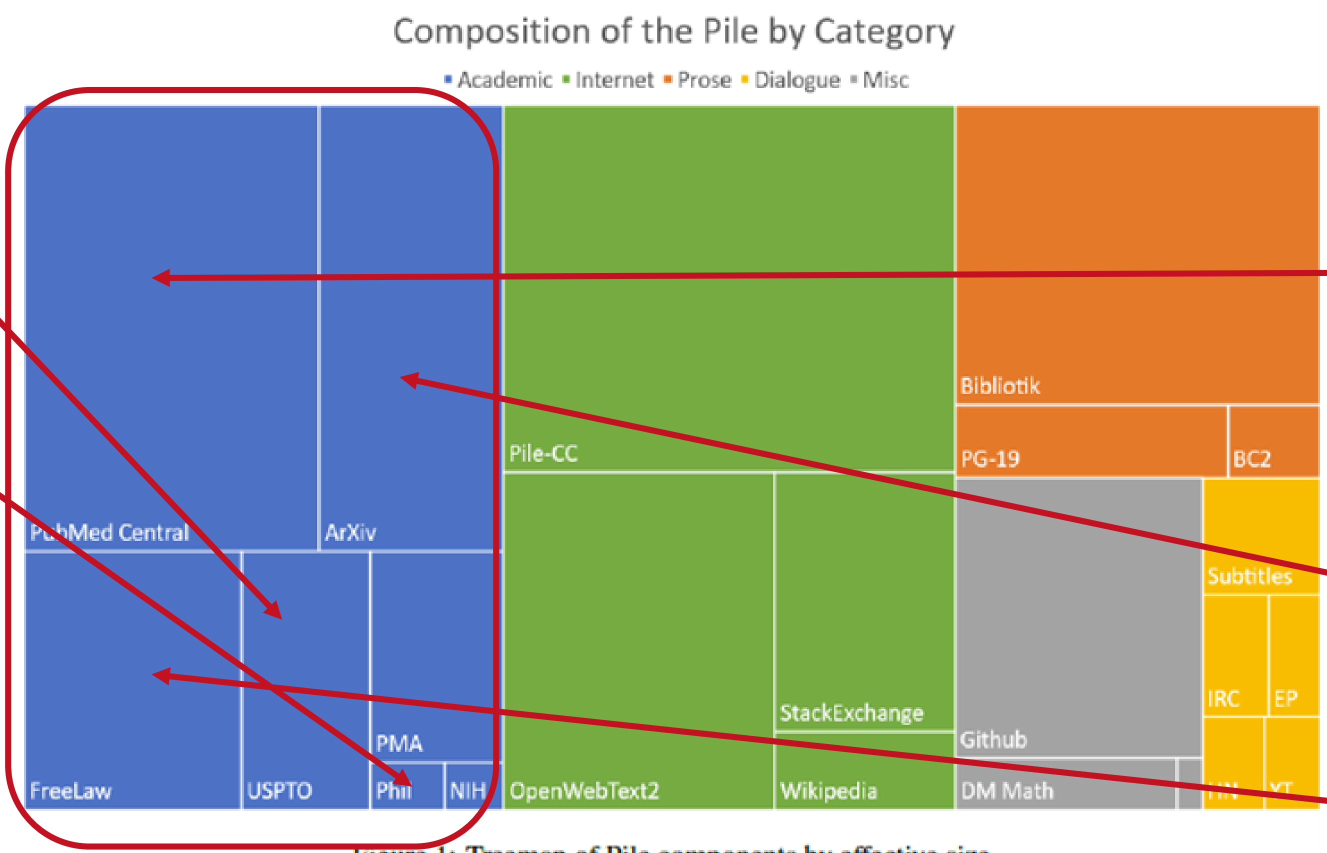


Figure 1: Treemap of Pile components by effective size.

The academic datasets.

PubMed Central: run by the USA's National Center for Biotechnology Information (NCBI), providing open, full-text access to nearly five million biomedical publications.

arXiv: as a source of high quality text and math knowledge

FreeLaw: as a source supporting academic studies in the legal realm, one of its sources provides bulk downloads for millions of legal opinions from federal and state courts

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

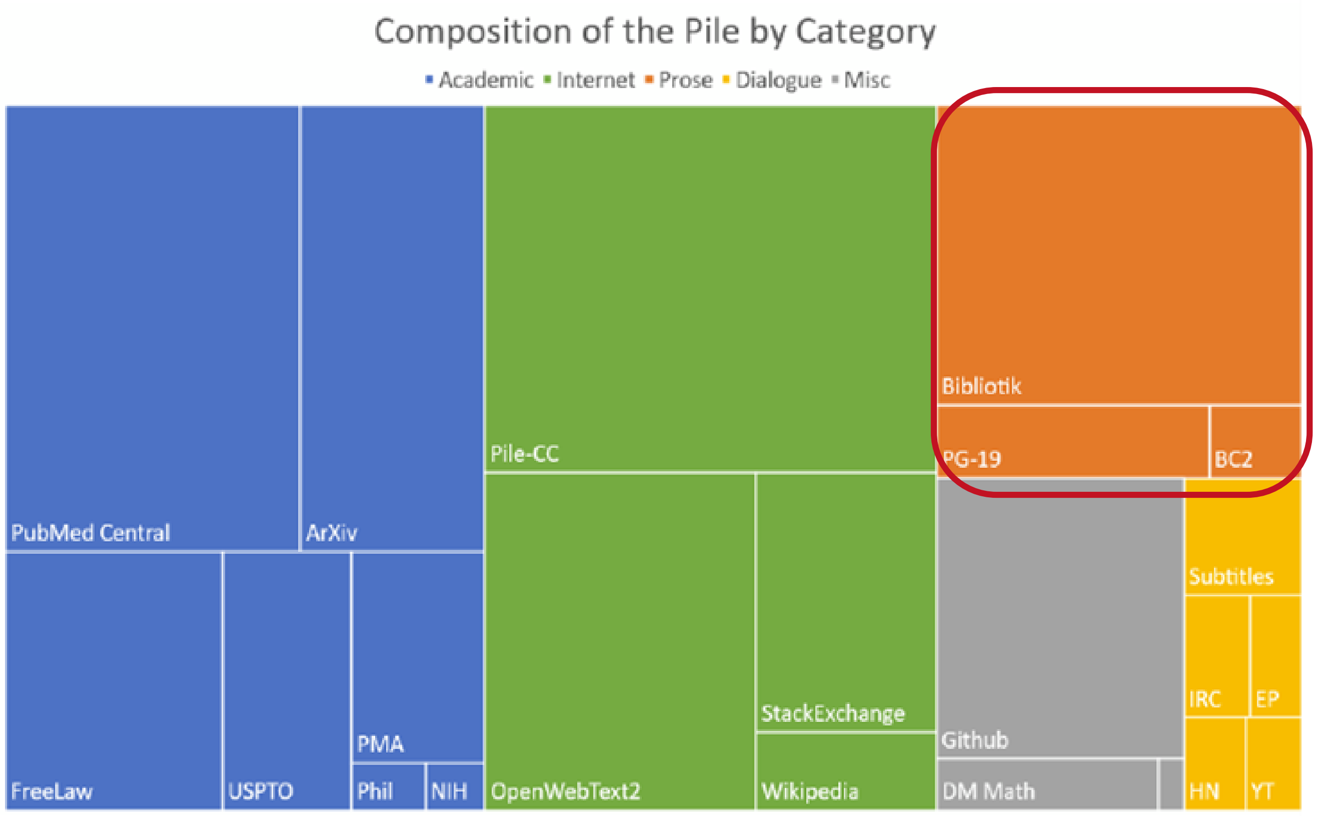


Figure 1: Treemap of Pile components by effective size.

The prose datasets.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- Data sources

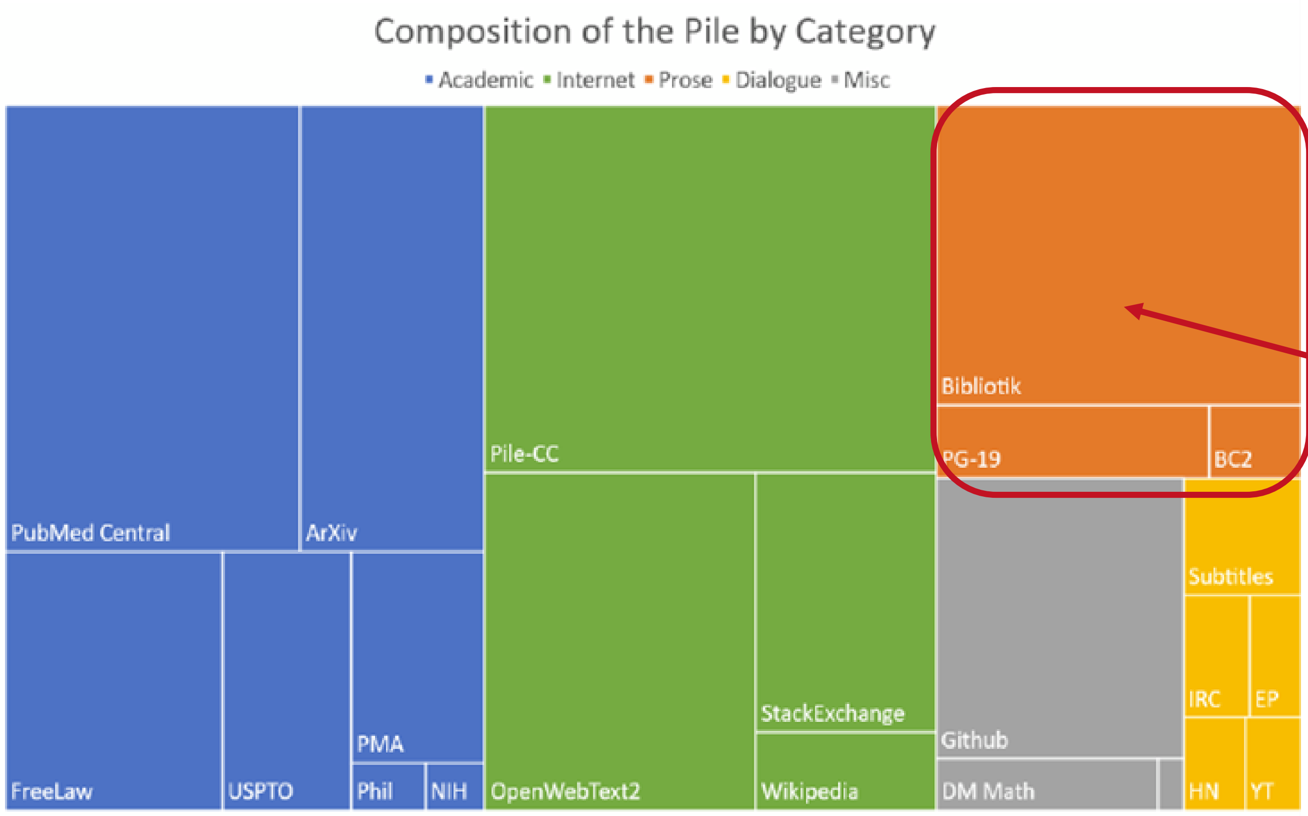


Figure 1: Treemap of Pile components by effective size.

The prose datasets.

Bibliothek or Books3: consists of a mix of fiction and nonfiction books. Books are invaluable for long-range context modeling research and coherent storytelling.

References
Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

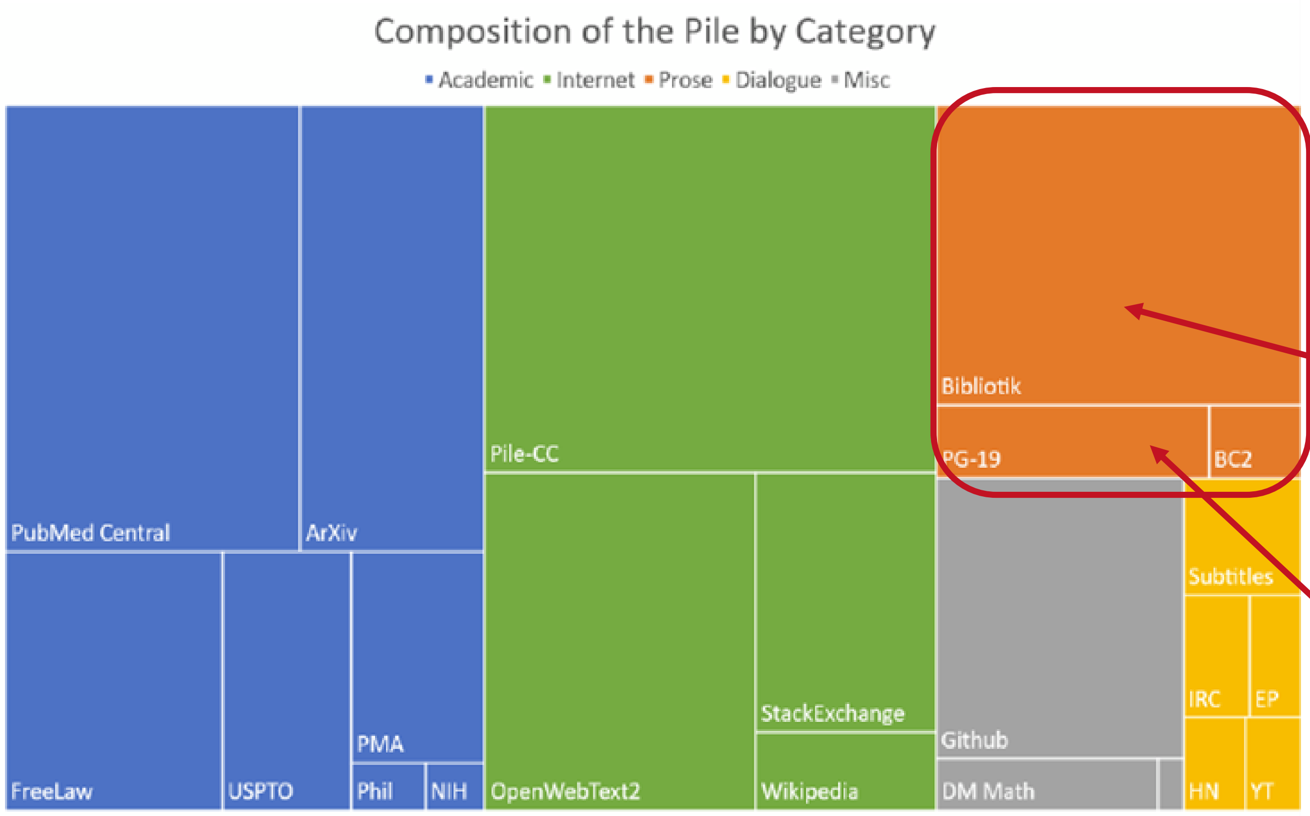


Figure 1: Treemap of Pile components by effective size.

The prose datasets.

Bibliothek or Books3: consists of a mix of fiction and nonfiction books. Books are invaluable for long-range context modeling research and coherent storytelling.

PG-19: Project Gutenberg dataset of classic Western literature from before 1919. Represent distinct styles from the more modern Books3.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

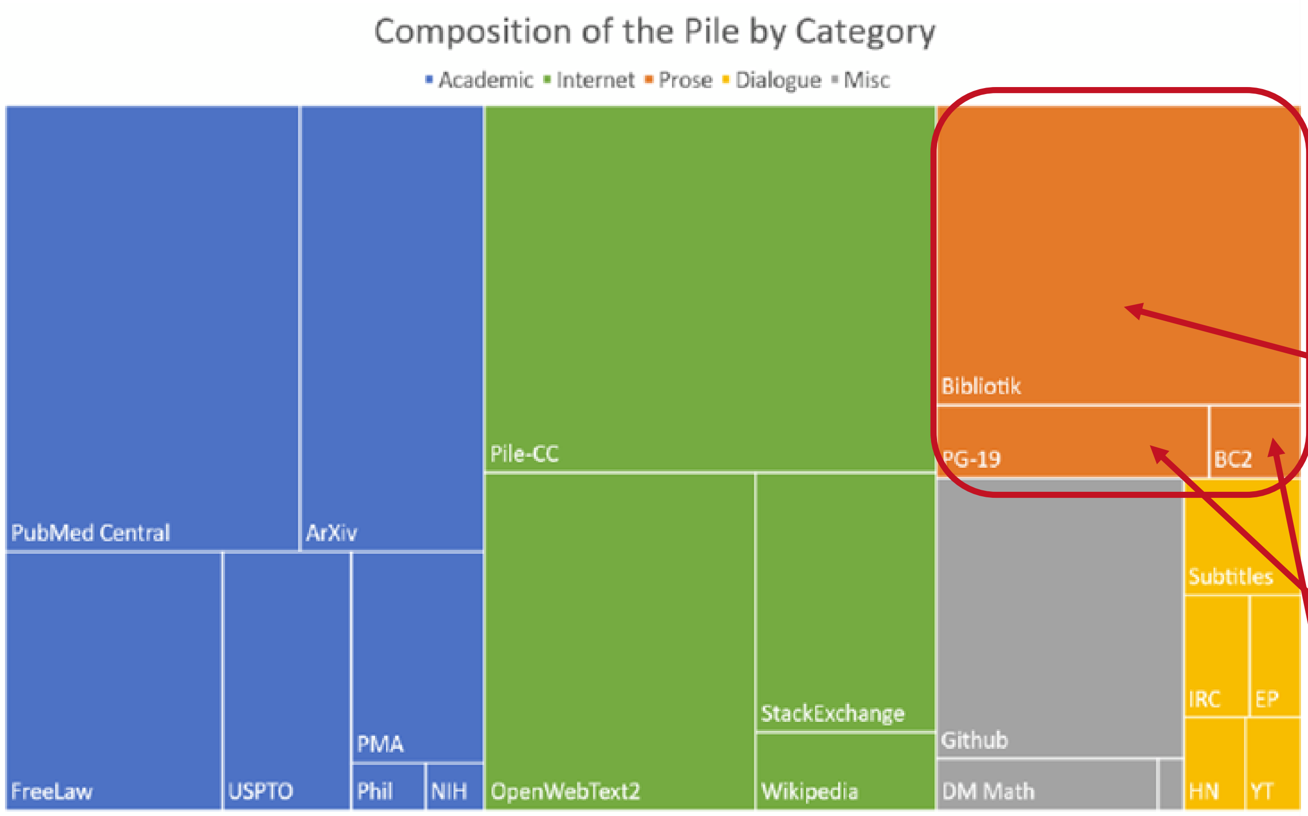


Figure 1: Treemap of Pile components by effective size.

The prose datasets.

Bibliothek or Books3: consists of a mix of fiction and nonfiction books. Books are invaluable for long-range context modeling research and coherent storytelling.

PG-19: Project Gutenberg dataset of classic Western literature from before 1919. Represent distinct styles from the more modern Books3.

BC2: BooksCorpus2 consisting of books written by “as of yet unpublished authors.”

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

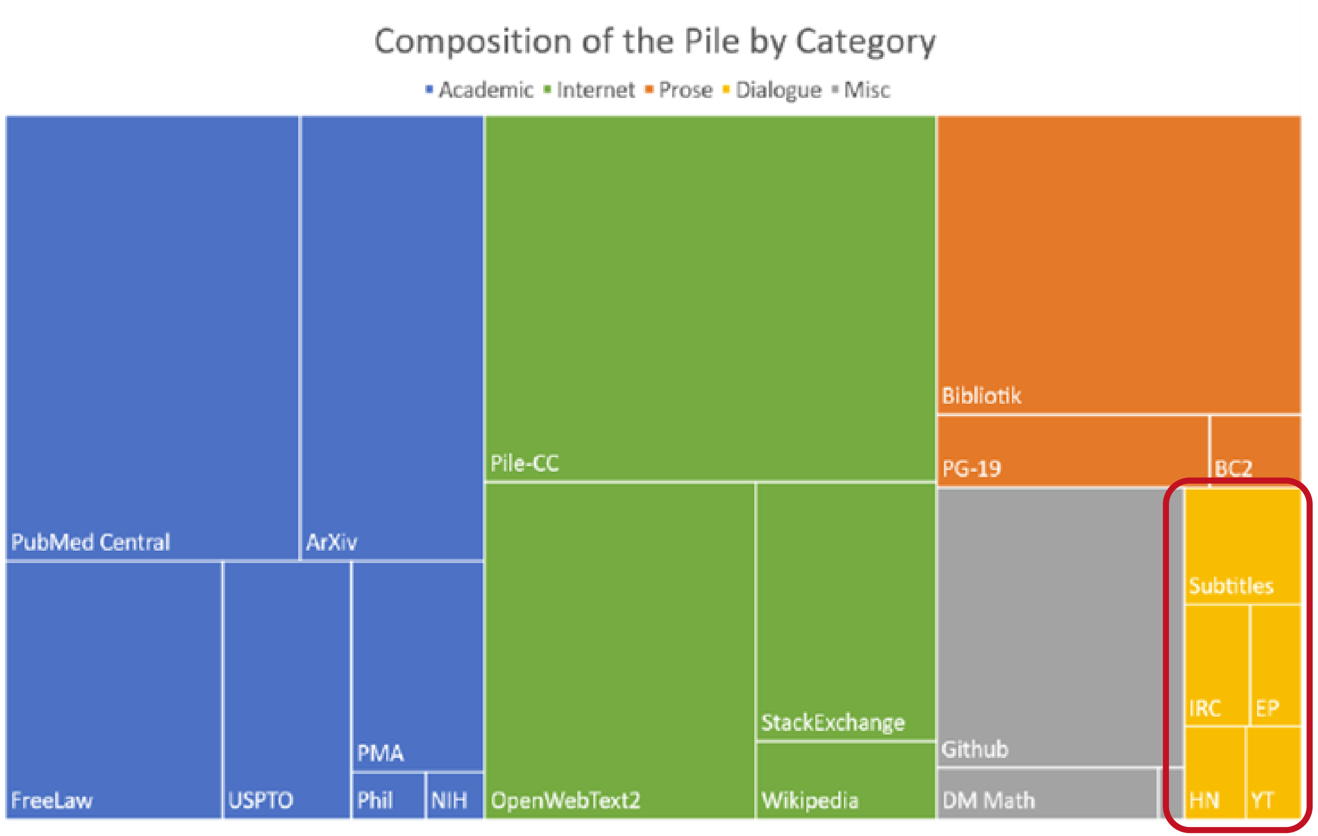


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

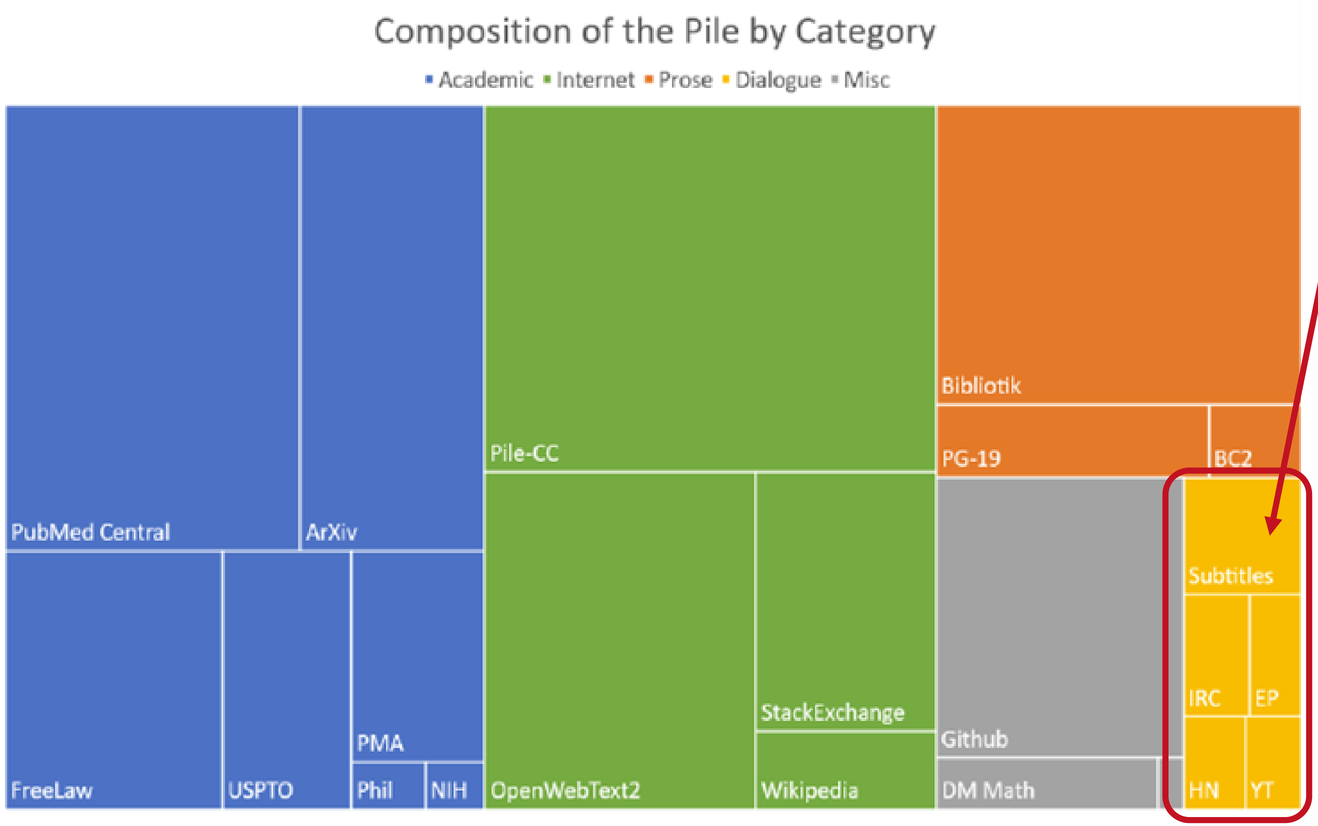


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

Subtitles: parallel corpus of text gathered from human generated closed captions on YouTube. It is a source of educational content, popular culture, and natural dialog.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

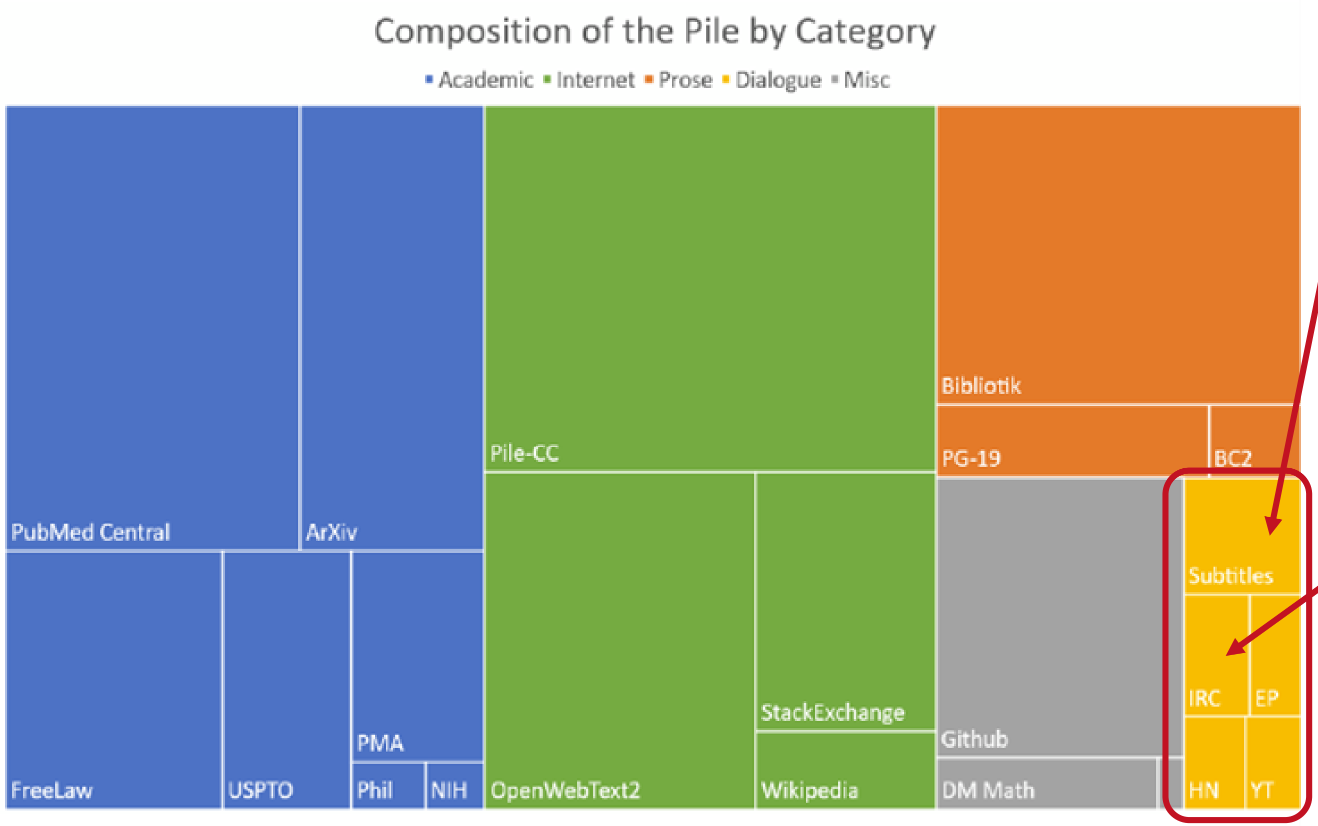


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

Subtitles: parallel corpus of text gathered from human generated closed captions on YouTube. It is a source of educational content, popular culture, and natural dialog.

Ubuntu IRC: publicly available chatlogs of all Ubuntu-related channels on the Freenode IRC chat server. Unique since it models real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

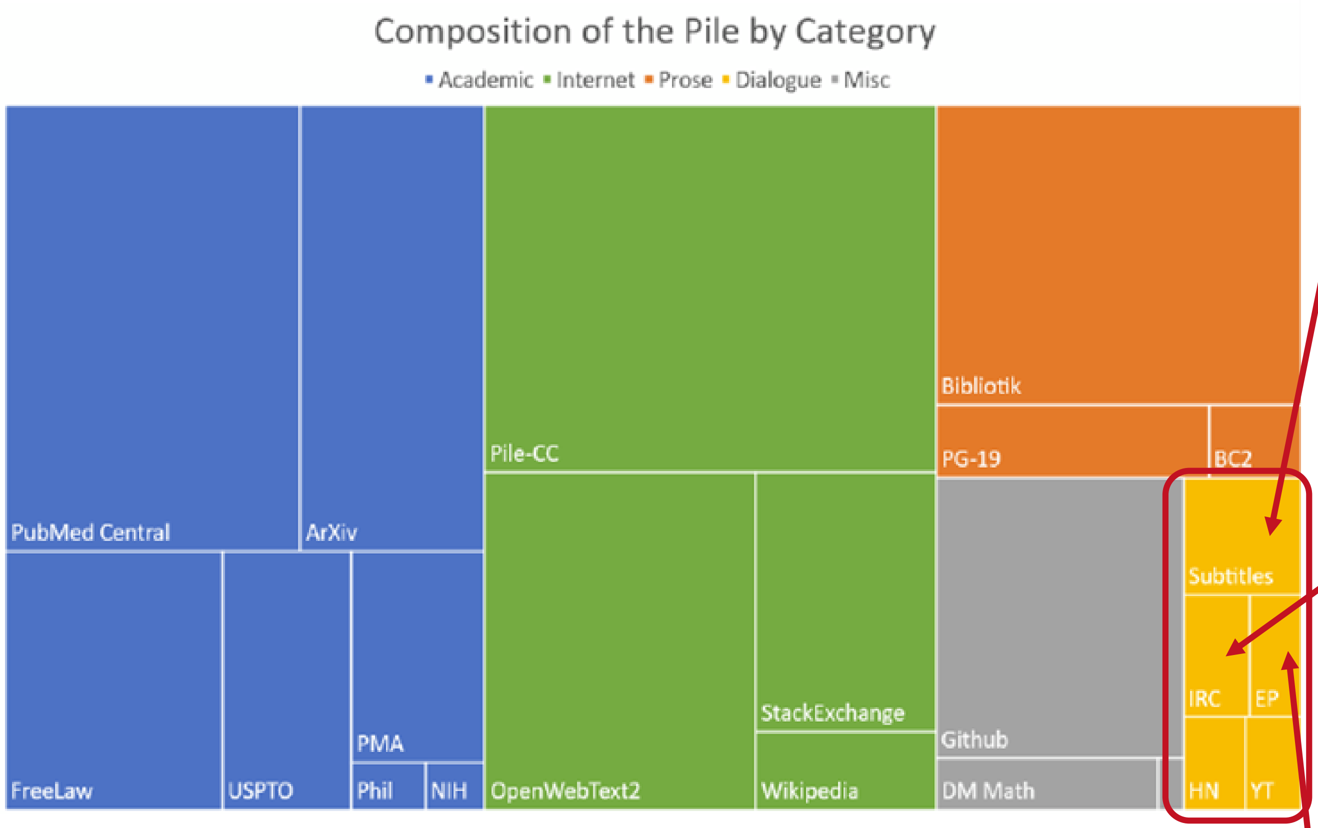


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

Subtitles: parallel corpus of text gathered from human generated closed captions on YouTube. It is a source of educational content, popular culture, and natural dialog.

Ubuntu IRC: publicly available chatlogs of all Ubuntu-related channels on the Freenode IRC chat server. Unique since it models real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

EuroParl: parallel corpus of proceedings of the Euro Parliament in 21 Euro languages 1996 - 2012.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

Hacker News: comment trees on user submitted stories on topics in CS and entrepreneurship.

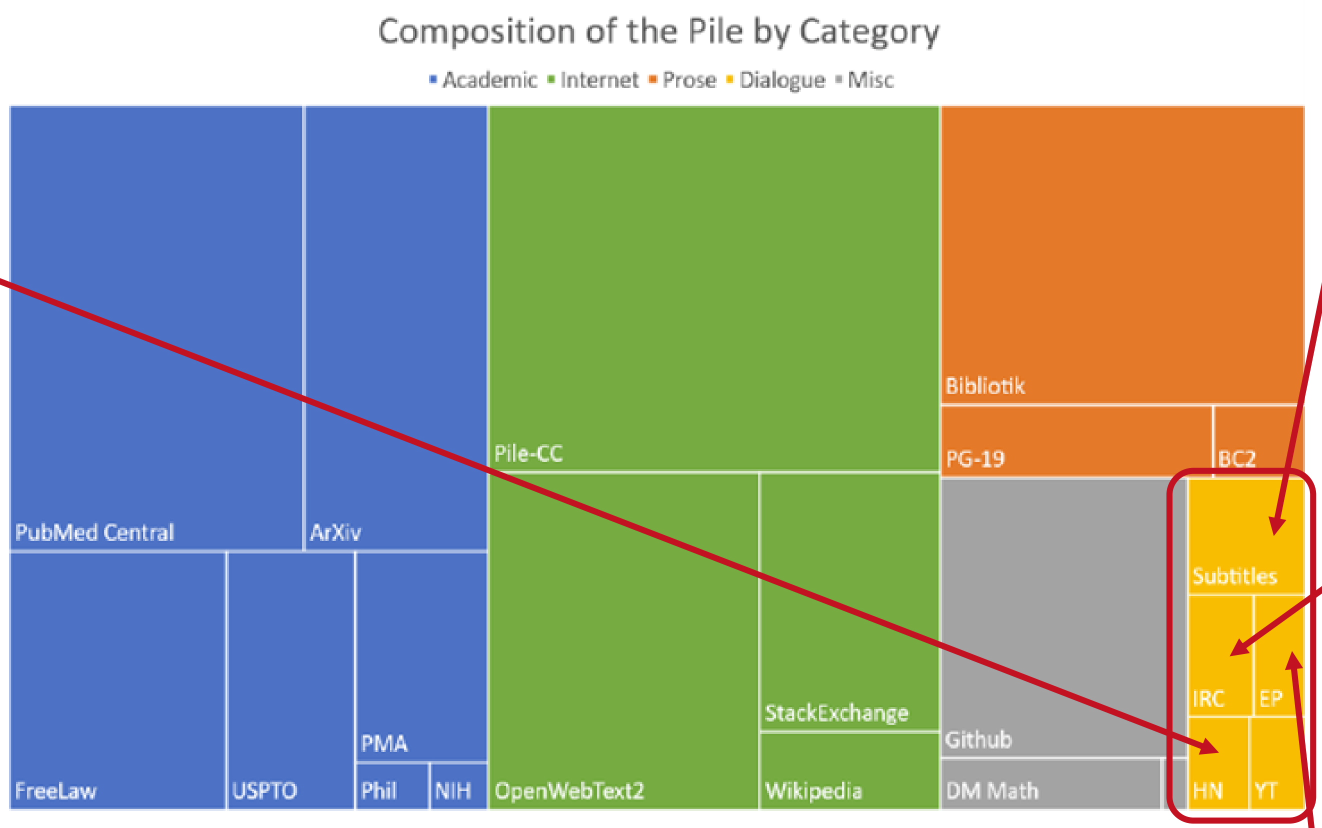


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

Subtitles: parallel corpus of text gathered from human generated closed captions on YouTube. It is a source of educational content, popular culture, and natural dialog.

Ubuntu IRC: publicly available chatlogs of all Ubuntu-related channels on the Freenode IRC chat server. Unique since it models real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

EuroParl: parallel corpus of proceedings of the Euro Parliament in 21 Euro languages 1996 - 2012.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

Hacker News: comment trees on user submitted stories on topics in CS and entrepreneurship.

Enron Emails: a valuable corpus commonly used for research about the usage patterns of email.

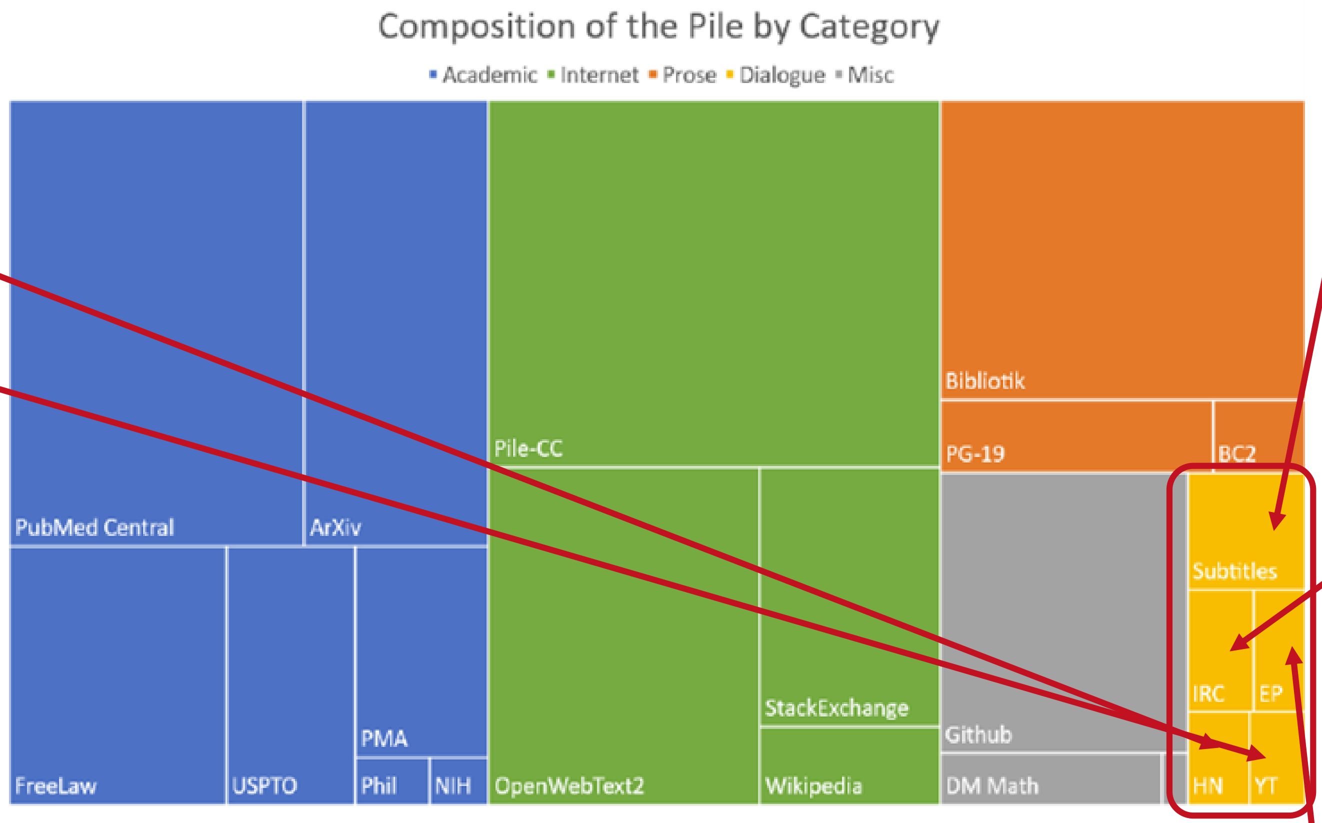


Figure 1: Treemap of Pile components by effective size.

The dialogue datasets.

Subtitles: parallel corpus of text gathered from human generated closed captions on YouTube. It is a source of educational content, popular culture, and natural dialog.

Ubuntu IRC: publicly available chatlogs of all Ubuntu-related channels on the Freenode IRC chat server. Unique since it models real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

EuroParl: parallel corpus of proceedings of the Euro Parliament in 21 Euro languages 1996 - 2012.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources**

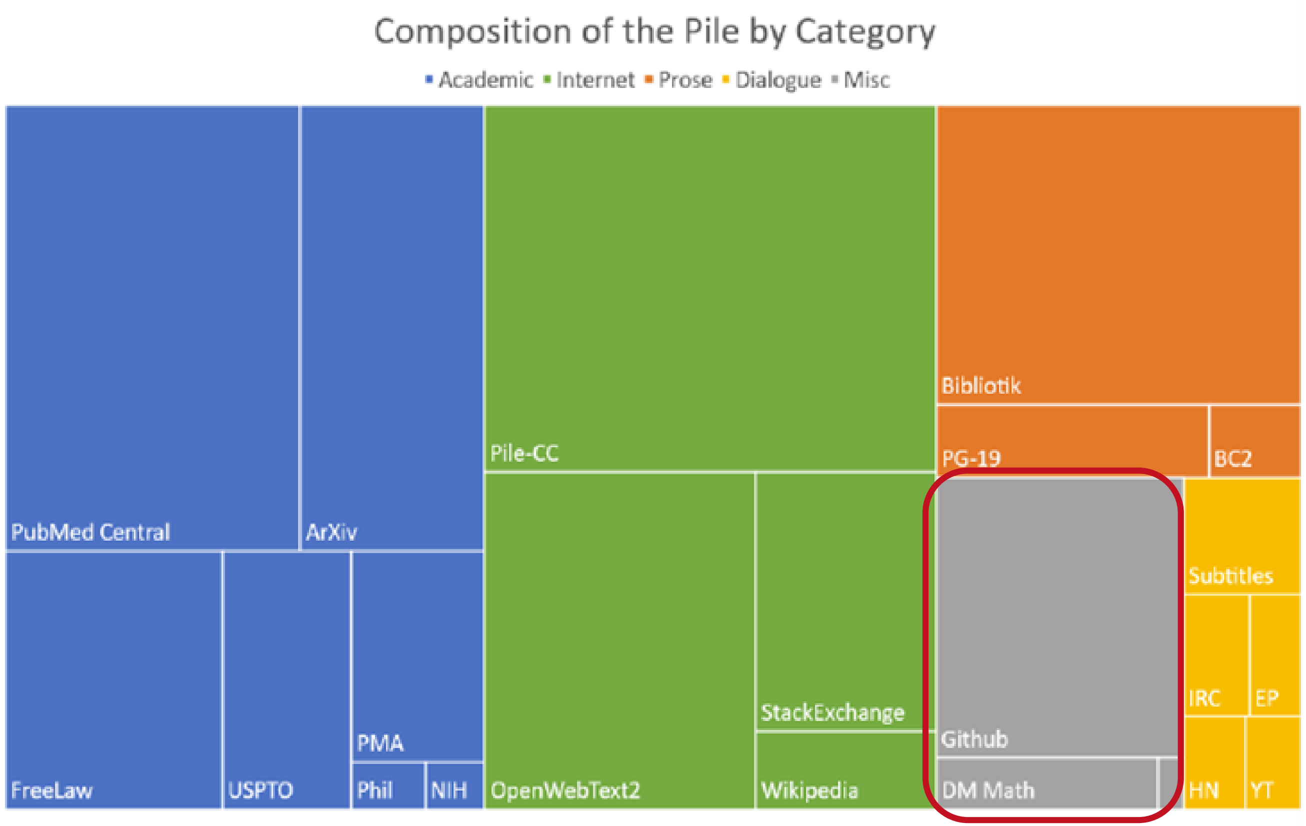


Figure 1: Treemap of Pile components by effective size.

The misc datasets.

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

- **Data sources overview**

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

Link to the dataset:
https://pile.eleuther.ai/

References

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

GPT-J-6B: A 6 billion parameter autoregressive language model

- **Key idea**
  - Given a set compute budget, where one cannot obtain a 100B parameter model, can effective LLMs still be produced?
    - Yes! If the underlying pretraining dataset is diverse enough.

- **A brief discussion of results:**

References
https://en.wikipedia.org/wiki/GPT-J
https://huggingface.co/EleutherAI/gpt-j-6b

GPT-J-6B: A 6 billion parameter autoregressive language model

- **Key idea**
  - Given a set compute budget, where one cannot obtain a 100B parameter model, can effective LLMs still be produced?
    - Yes! If the underlying pretraining dataset is diverse enough.

- **A brief discussion of results:**
  - When neither is fine-tuned, GPT-J-6B performs almost as well as the 6.7 billion parameter GPT-3 (Curie) on a variety of tasks. It even outperforms the 175 billion parameter GPT-3 (Davinci) on code generation tasks. With fine-tuning, it outperforms an untuned GPT-3 (Davinci) on a number of tasks.

References
https://en.wikipedia.org/wiki/GPT-J
https://huggingface.co/EleutherAI/gpt-j-6b

LLaMA: Open and Efficient Foundation Language Models

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Key idea**
  - given two existing lines of thought: on the one hand, is bigger better? and on the other hand, for a given compute budget, are smaller models optimally trained on more data better?
    - on a given inference budget say distributed in the range of 7B, 13B, 33B, and 65B, how to obtain the optimal model? – for the given inference budget, train on more tokens than typically used. In other words, their hypothesis tests if a smaller model (given inference budget but trained on more tokens) can outperform a larger model.
      - LLaMA-13B outperforms 10x larger GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B
      - specifically, surpassing all preceding models, LLaMA is trained on 1.4T tokens.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Key idea**
  - given two existing lines of thought: on the one hand, is bigger better? and on the other hand, for a given compute budget, are smaller models optimally trained on more data better?
    - on a given inference budget say distributed in the range of 7B, 13B, 33B, and 65B, how to obtain the optimal model? – for the given inference budget, train on more tokens than typically used. In other words, their hypothesis tests if a smaller model (given inference budget but trained on more tokens) can outperform a larger model.
      - LLaMA-13B outperforms 10x larger GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B
      - specifically, surpassing all preceding models, LLaMA is trained on 1.4T tokens.
  - show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets.
    - The best models at the time i.e. GPT-3, Chinchilla, or PaLM were trained on proprietary datasets.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl
      - processed by the CCNet pipeline

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **CCNet: Processing Pipeline used for English CommonCrawl**

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **CCNet: Processing Pipeline used for English CommonCrawl**
  1. Deduplication
     i. normalization - lowercase all characters, replace numbers by a placeholder (i.e. 0) and remove all Unicode punctuation and
     ii. hashing for deduplication - dividing the data into smaller shards, assigning a unique code to each paragraph in each shard using SHA-1 hashing, and then comparing these codes to identify and eliminate duplicates.
        - Other than removing web copies, this step gets rid of a lot boilerplate such as navigation menus, cookie warnings and contact information. It also removes significant amount of English content from non-English webpages. This makes the subsequent language identification step more robust.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **CCNet: Processing Pipeline used for English CommonCrawl**
    1. Deduplication
        i. normalization - lowercase all characters, replace numbers by a placeholder (i.e. 0) and remove all Unicode punctuation and
        ii. hashing for deduplication - dividing the data into smaller shards, assigning a unique code to each paragraph in each shard using SHA-1 hashing, and then comparing these codes to identify and eliminate duplicates.
            - Other than removing web copies, this step gets rid of a lot boilerplate such as navigation menus, cookie warnings and contact information. It also removes significant amount of English content from non-English webpages. This makes the subsequent language identification step more robust.
    2. Language identification – split the data per language using the language classifier from fastText.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **CCNet: Processing Pipeline used for English CommonCrawl**
  1. Deduplication
     i. normalization - lowercase all characters, replace numbers by a placeholder (i.e. 0) and remove all Unicode punctuation and
     ii. hashing for deduplication - dividing the data into smaller shards, assigning a unique code to each paragraph in each shard using SHA-1 hashing, and then comparing these codes to identify and eliminate duplicates.
        - Other than removing web copies, this step gets rid of a lot boilerplate such as navigation menus, cookie warnings and contact information. It also removes significant amount of English content from non-English webpages. This makes the subsequent language identification step more robust.
  2. Language identification – split the data per language using the language classifier from fastText.
  3. N-gram Language model – categorize texts as high, medium, and low quality based on perplexity computations with Wikipedia texts. However this was not applied in their evaluations because they deemed text different from Wikipedia which was cast as low quality based on their language model could still be useful.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*, 2020.

LLaMA: Open and Efficient Foundation Language Models

- **<u>Results</u> – CCNet: Processing Pipeline used for English CommonCrawl**
  - On the Feb 2019 snapshot of CommonCrawl, the application of CCNet produced 3.2TB of compressed documents in 174 languages.

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **Results – CCNet: Processing Pipeline used for English CommonCrawl**
  - On the Feb 2019 snapshot of CommonCrawl, the application of CCNet produced 3.2TB of compressed documents in 174 languages.
  - In terms of tokens, using the SentencePiece tokenizer, the 3 largest languages were English (en) with 532B tokens, Russian (ru) with 101B tokens and Chinese (zh) with 92B tokens. 11 languages with more than 10B tokens, and 27 languages with more than 1B tokens.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **Results – CCNet: Processing Pipeline used for English CommonCrawl**
  - On the Feb 2019 snapshot of CommonCrawl, the application of CCNet produced 3.2TB of compressed documents in 174 languages.
  - In terms of tokens, using the SentencePiece tokenizer, the 3 largest languages were English (en) with 532B tokens, Russian (ru) with 101B tokens and Chinese (zh) with 92B tokens. 11 languages with more than 10B tokens, and 27 languages with more than 1B tokens.
  - In terms of documents, the 3 largest languages were English (en) with 706M documents, Russian (ru) with 167M and German (de) with 105M. 12 languages had more than 10M documents and 29 languages had more than 1M documents.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **Results – CCNet: Processing Pipeline used for English CommonCrawl**
  - On the Feb 2019 snapshot of CommonCrawl, the application of CCNet produced 3.2TB of compressed documents in 174 languages.
  - In terms of tokens, using the SentencePiece tokenizer, the 3 largest languages were English (en) with 532B tokens, Russian (ru) with 101B tokens and Chinese (zh) with 92B tokens. 11 languages with more than 10B tokens, and 27 languages with more than 1B tokens.
  - In terms of documents, the 3 largest languages were English (en) with 706M documents, Russian (ru) with 167M and German (de) with 105M. 12 languages had more than 10M documents and 29 languages had more than 1M documents.
  - As a side-note, Common Crawl is also a good source for lower resource languages. E.g., Afrikaans (af), Gujarati (gu), Khmer (km) and Burmese (my) contains respectively 160MB, 190MB, 154MB and 440MB of data.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
Wenzek, Guillaume, et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." *Proceedings of the Twelfth LREC*. 2020.

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl: from 2017 to 2020
      - processed by the CCNet pipeline

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl: from 2017 to 2020
      - processed by the CCNet pipeline, selected only the English subset
      - additionally, web pages not used as references in Wikipedia were filtered out

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl: from 2017 to 2020
      - processed by the CCNet pipeline, selected only the high-quality English subset
      - additionally, web pages not used as references in Wikipedia were filtered out
    - Colossal Cleaned Common Crawl (C4) Dataset consisting of nearly a trillion words.
      - this dataset was introduced in the T5 paper which was a model released after GPT-2.
      - including diverse **pre-processed** Common Crawl improved performance; C4 preprocessing mostly relies on heuristics such as presence of punctuation marks or the number of words and sentences in a webpage.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining data sources**
  - Web pages
    - English CommonCrawl: from 2017 to 2020
      - processed by the CCNet pipeline
      - additionally, web pages not used as references in Wikipedia were filtered out
    - Colossal Cleaned Common Crawl (C4) Dataset consisting of nearly a trillion words.
      - this dataset was introduced in the T5 paper which was a model released after GPT-2.
      - including diverse **pre-processed** Common Crawl improved performance; C4 preprocessing mostly relies on heuristics such as presence of punctuation marks or the number of words and sentences in a webpage.
  - Other diverse sources
    - Code - Github - projects distributed under the Apache, BSD and MIT licenses. Additionally filtered low quality files with heuristics based on the line length or proportion of alphanumeric characters, etc.
    - Encyclopedia - Wikipedia - dumps from Jun-Aug 2022 covering 20 languages, which use either the Latin or Cyrillic scripts: bg, ca, cs, da, de, en, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sl, sr, sv, uk
    - Books - Gutenberg and Books3
      - Gutenberg Project contains books that are in the public domain.
      - Books3 section of The Pile publicly available dataset
    - Scholarly articles from arXiv from the raw Latex files
    - QA corpus - StackExchange - QA covering diverse domains from CS to Chemistry.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining datasets**

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

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining datasets**

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

Entire training dataset consists of roughly 1T tokens after tokenization using the Byte Pair Encoding (BPE) algorithm.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining datasets**

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

For most of the training data, each token is used only once during training.

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Pretraining datasets**

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

For most of the training data, each token is used only once during training. The Wikipedia and Books corpora are the exceptions over which roughly 2 epochs are performed.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Evaluations**
  - Has training on more tokens helped? Furthermore, is it possible to obtain high-performing models on only open source datasets?

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Evaluations**
  - Has training on more tokens helped? Furthermore, is it possible to obtain high-performing models on only open source datasets?

		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.**

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Evaluations**
  - Has training on more tokens helped? Furthermore, is it possible to obtain high-performing models on only open source datasets?

		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.**

LLaMA 13B outperforms GPT-3 175B on most datasets.

References

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

LLaMA: Open and Efficient Foundation Language Models

- **Evaluations**
  - Has training on more tokens helped? Furthermore, is it possible to obtain high-performing models on only open source datasets?

		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.**

The largest model is better than Chinchilla and 10x larger PaLM 540B on most datasets.

References
Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

Falcon – An open large language model with state-of-the-art performance

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- **Key idea**
  - Falcon LLM **places further emphasis on the pre-training dataset quality** to obtain a more effective downstream model.
    - It employs a custom data pipeline and codebase specifically tailored to extract high-quality content from the web. This custom pipeline allows Falcon LLM to curate and process data from diverse online sources, ensuring it is exposed to a wide range of relevant information during the training phase. By extracting high-quality content, Falcon LLM aims to enhance the accuracy and richness of the language generated by the model.

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- **Key idea**
    - Falcon LLM **places further emphasis on the pre-training dataset quality** to obtain a more effective downstream model.
        - It employs a custom data pipeline and codebase specifically tailored to extract high-quality content from the web. This custom pipeline allows Falcon LLM to curate and process data from diverse online sources, ensuring it is exposed to a wide range of relevant information during the training phase. By extracting high-quality content, Falcon LLM aims to enhance the accuracy and richness of the language generated by the model.
    - Also, Falcon is trained on trillions of tokens of text compared to GPT-3’s billions.
        - Ultimately GPT-4 will compete on the size of the training set. However, the largest Falcon model was a more compact 40B parameter model, while GPT-4 is said to have trillions of parameters.

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- **Key idea**
  - Falcon LLM places further emphasis on the pre-training dataset quality to obtain a more effective downstream model.
    - It employs a custom data pipeline and codebase specifically tailored to extract high-quality content from the web. This custom pipeline allows Falcon LLM to curate and process data from diverse online sources, ensuring it is exposed to a wide range of relevant information during the training phase. By extracting high-quality content, Falcon LLM aims to enhance the accuracy and richness of the language generated by the model.
  - Also, Falcon is trained on trillions of tokens of text compared to GPT-3's billions.
    - Ultimately GPT-4 will compete on the size of the training set. However, the largest Falcon model was a more compact 40B parameter model, while GPT-4 is said to have trillions of parameters.
  - Falcon models support democratized access to LLMs: they are all available under the Apache 2.0 license, while the GPT models are all closed-sourced.
    - Variants of Falcon models released such as Falcon-Instruct or Falcon-Chat

Falcon – An open large language model with state-of-the-art performance

- At the time of its release, i.e. May 2023, finetuned Falcon model beat all other models such as LLaMA-1 or GPT-3 and was the state-of-the-art on the HuggingFace leaderboard.

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- As noted before, the key ingredient for the high quality of the Falcon models is their training data, predominantly based (>80%) on RefinedWeb — a novel massive web dataset based on CommonCrawl.

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- As noted before, the key ingredient for the high quality of the Falcon models is their training data, predominantly based (>80%) on RefinedWeb — a novel massive web dataset based on CommonCrawl.

So what is this RefinedWeb corpus? Let’s take a closer look next.

References
https://huggingface.co/blog/falcon

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ● REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			MASSIVE WEB DATASETS		
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact**: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
			CURATED DATASETS		
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			OURS		
● REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (~ 50% removed)

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ●REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
MASSIVE WEB DATASETS					
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact**: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
CURATED DATASETS					
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
OURS					
●REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (~ 50% removed)

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ●REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			MASSIVE WEB DATASETS		
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact**: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
			CURATED DATASETS		
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			OURS		
●REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (~ 50% removed)

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ● REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
MASSIVE WEB DATASETS					
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact**: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
CURATED DATASETS					
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
OURS					
● REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (~ 50% removed)

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ●REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			MASSIVE WEB DATASETS		
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact**: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
			CURATED DATASETS		
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			OURS		
●REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (~ 50% removed)

The closest to RefinedWeb is the C4 dataset which was introduced in the context of T5.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			MASSIVE WEB DATASETS		
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact:** spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact:** per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact:** per line (optional, not used for results in this paper)
			CURATED DATASETS		
GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy:** MinHash (~ 10% removed)
The Pile	~ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy:** MinHash (~ 26% removed)
PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			OURS		
REFINEDWEB	~ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy:** exact substring+MinHash (~ 50% removed)

RefinedWeb goes beyond C4 in the size of the resulting dataset measured in giga tokens.

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated "high-quality" corpora, such as social media conversations, books, or technical papers.

Table 1. ● REFINEDWEB **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			MASSIVE WEB DATASETS		
C4	∼ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact**: spans of 3 sentences
OSCAR-21.09	∼ 370GT	Public	100%	Built at the line-level	**Exact**: per line (∼ 55% removed)
OSCAR-22.01	∼ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact**: per line (optional, not used for results in this paper)
			CURATED DATASETS		
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy**: MinHash (∼ 10% removed)
▼ The Pile	∼ 340GT	Public	18%	`jusText` for extraction, content filter trained on curated data	**Fuzzy**: MinHash (∼ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			OURS		
● REFINEDWEB	∼ 5,000GT	Public (600GT)	100%	`trafilatura` for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy**: exact substring+MinHash (∼ 50% removed)

RefinedWeb like C4 also relies on CommonCrawl. However they implement new processing methods.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.

Table 1. ●RefinedWeb **improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale.** For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
			Massive web datasets		
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	**Exact:** spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	**Exact:** per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	**Exact:** per line (optional, not used for results in this paper)
			Curated datasets		
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	**Fuzzy:** MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	jusText for extraction, content filter trained on curated data	**Fuzzy:** MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
			Ours		
●RefinedWeb	~ 5,000GT	Public (600GT)	100%	trafilatura for text extraction, document and line-level rules, NSFW URL blocklist	**Exact & fuzzy:** exact substring+MinHash (~ 50% removed)

RefinedWeb when compared to C4 also implements a more sophisticated web page deduplication strategy.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.
  - Central thesis behind the Falcon model team of researchers to create RefinedWeb was the following:
    - Can properly filtered and deduplicated web data alone lead to powerful models?

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.
  - Central thesis behind the Falcon model team of researchers to create RefinedWeb was the following:
    - Can properly filtered and deduplicated web data alone lead to powerful models?
  - RefinedWeb comprises 5 trillion tokens from CommonCrawl:
    - Publicly released a 600B tokens dataset.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **What is Refined Web?**
  - LLMs are trained on a mixture of filtered web data and curated “high-quality” corpora, such as social media conversations, books, or technical papers.
  - Central thesis behind the Falcon model team of researchers to create RefinedWeb was the following:
    - Can properly filtered and deduplicated web data alone lead to powerful models?
  - RefinedWeb comprises 5 trillion tokens from CommonCrawl:
    - Publicly released a 600B tokens dataset.

Let’s take a look at the scalable data processing pipeline introduced in this work…

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**

DOCUMENT PREPARATION			FILTERING		DEDUPLICATION	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across Common-Crawl dumps Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:

DOCUMENT PREPARATION			FILTERING		DEDUPLICATION	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked	From WARC using `warcio`, `trafilatura` for extraction	`fastText` classifier from CCNet, thresholding on top language score	In-document repetition removal and quality heuristics from MassiveWeb	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.)	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays	Remove URLs revisited across CommonCrawl dumps
Appendix G.1	Barbaresi (2021)	Wenzek et al. (2020)	Rae et al. (2021)	Appendix G.2	Lee et al. (2022)	Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation

Document preparation			Filtering		Deduplication	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering

Document preparation			Filtering		Deduplication	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering
    - Deduplication

DOCUMENT PREPARATION			FILTERING		DEDUPLICATION	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering
    - Deduplication

DOCUMENT PREPARATION			FILTERING		DEDUPLICATION	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classi-fier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repe-tition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplica-tion w/ suffix arrays Lee et al. (2022)	Remove URLs revis-ited across Common-Crawl dumps Section 3.3

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering
    - Deduplication

Document preparation			Filtering		Deduplication	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering
    - Deduplication

Document preparation			Filtering		Deduplication	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

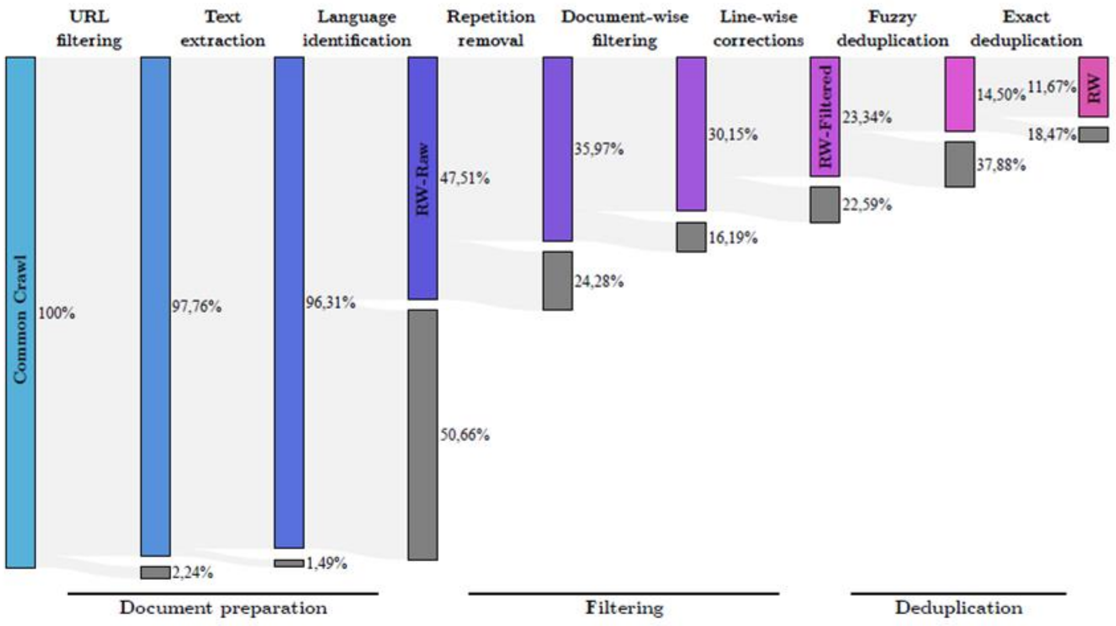
- **Refined Web Scalable and High-Quality Web Text Processing Pipeline**
  - Three main parts:
    - Document preparation
    - Filtering
    - Deduplication

Document preparation			Filtering		Deduplication	
URL filtering	**Text extraction**	**Language identification**	**Document-wise filtering**	**Line-wise filtering**	**Deduplication**	**URL deduplication**
Aggregated block-list, URL scoring, common HQ sources blocked Appendix G.1	From WARC using `warcio`, `trafilatura` for extraction Barbaresi (2021)	`fastText` classifier from CCNet, thresholding on top language score Wenzek et al. (2020)	In-document repetition removal and quality heuristics from MassiveWeb Rae et al. (2021)	Remove undesirable lines (call to actions, navigation buttons, social counters, etc.) Appendix G.2	Fuzzy deduplication w/ MinHash + exact substring deduplication w/ suffix arrays Lee et al. (2022)	Remove URLs revisited across CommonCrawl dumps Section 3.3

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Application of the Processing Pipeline – Arriving at RefinedWeb**

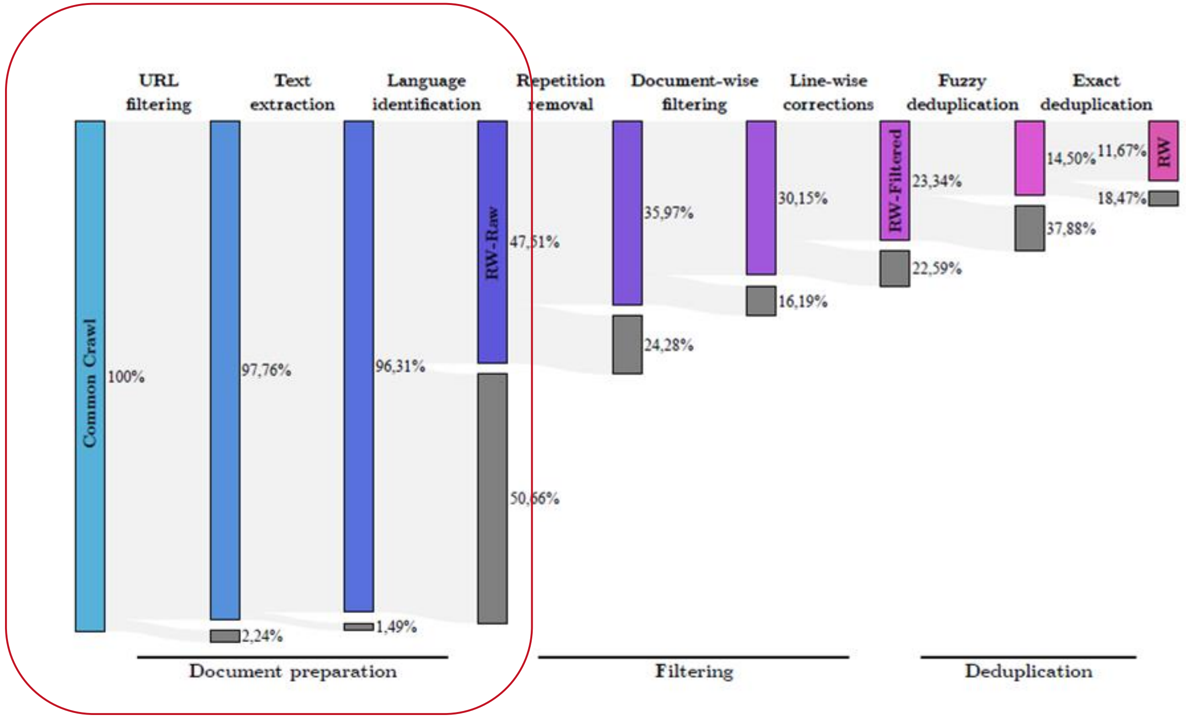


References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Application of the Processing Pipeline – Arriving at RefinedWeb**



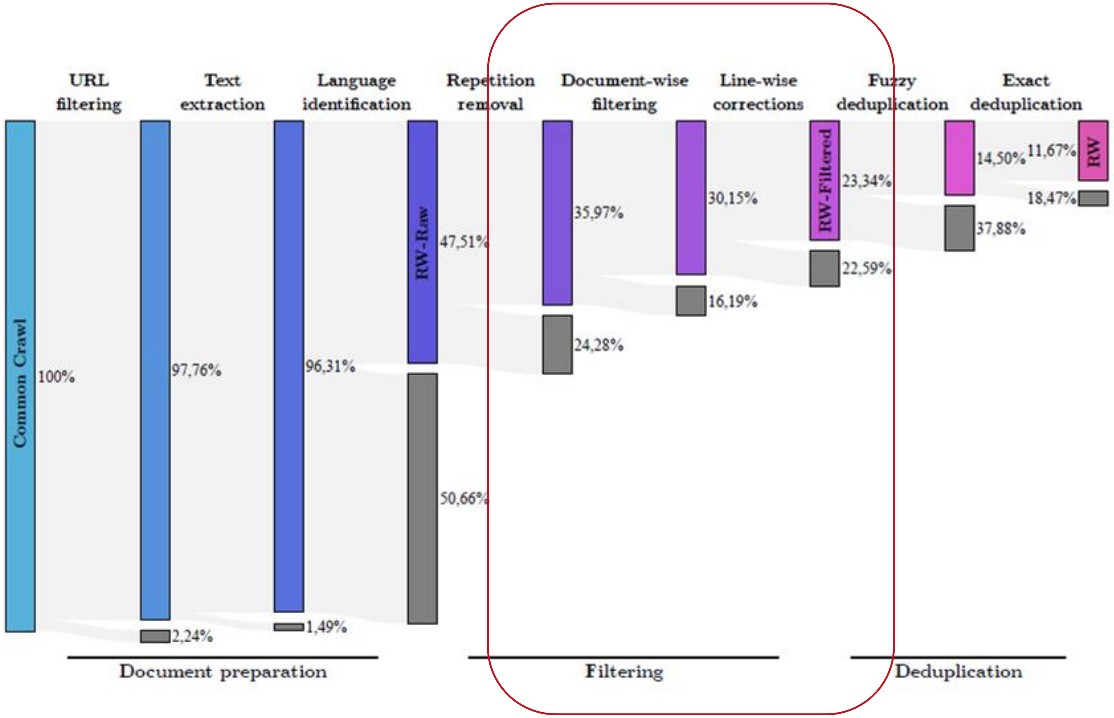
The initial web corpus is reduced by a small proportion after "document preparation." And reduces by 50% after language identification.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Application of the Processing Pipeline – Arriving at RefinedWeb**

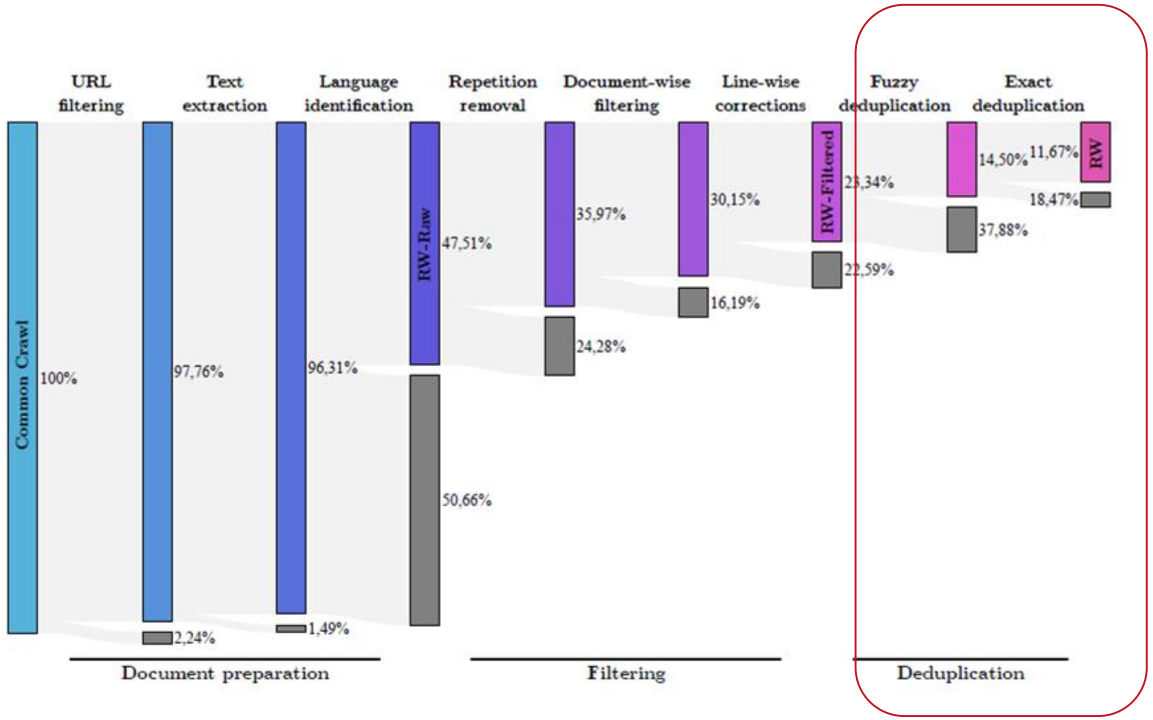


After the filtering stage, only 23% of the original set of web page text is retained.

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Application of the Processing Pipeline – Arriving at RefinedWeb**



After the deduplication stage, only 11% of the original set of web page text are retained as the resulting high-quality corpus.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Is the RefinedWeb strategy effective? In other words, can an effective LLM be obtained from only heuristics-based pipeline to create a high-quality web dataset?

References
Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - The pretrained model was tested on various evaluation datasets in the community.

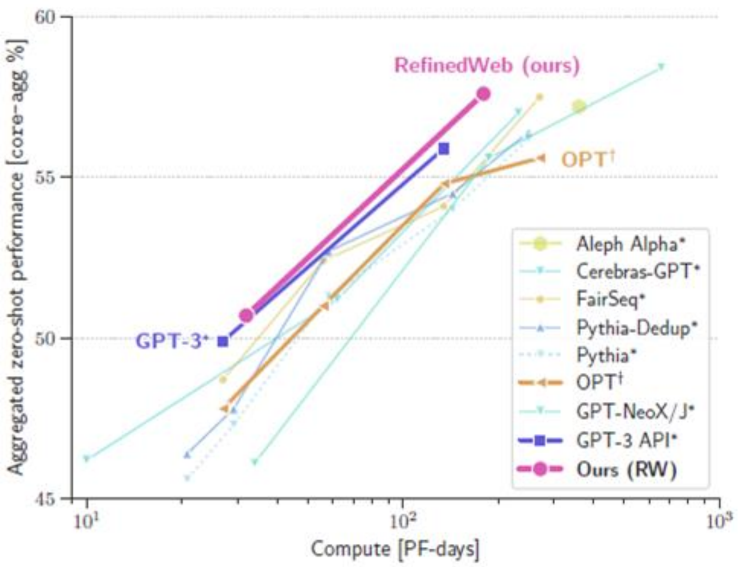
Tasks	Type	Random	small	core	main	ext
HellaSwag (Zellers et al., 2019)	Sentence completion	25.0	✓	✓	✓	✓
LAMBADA (Paperno et al., 2016)	Sentence completion	0.0		✓	✓	✓
Winogrande (Sakaguchi et al., 2021)	Coreference resolution	50.0	✓	✓	✓	✓
PIQA (Bisk et al., 2020)	Multiple-choice question answering	50.0	✓	✓	✓	✓
ARC (Clark et al., 2018)	Natural language inference	25.0	✓	✓	✓	✓
OpenBookQA (Mihaylov et al., 2018)	Multiple-choice question answering	25.0		✓	✓	✓
BoolQ (Clark et al., 2019)	Multiple-choice question answering	50.0	✓		✓	✓
COPA (Gordon et al., 2012)	Sentence completion	50.0			✓	✓
CB (De Marneffe et al., 2019)	Natural language inference	33.3			✓	✓
RTE (Dagan et al., 2010)	Natural language inference	50.0			✓	✓
ReCoRD (Zhang et al., 2018)	Question answering	0.0			✓	
ANLI (Nie et al., 2019)	Natural language inference	33.3			✓	
LogiQA (Liu et al., 2021)	Multiple-choice question answering	25.0				✓
HeadQA (Vilares & Gómez-Rodríguez, 2019)	Multiple-choice question answering	20.0				✓
MathQA (Amini et al., 2019)	Multiple-choice question answering	20.0				✓
PROST (Aroca-Ouellette et al., 2021)	Paraphrase identification	50.0				✓
PubMedQA (Jin et al., 2019)	Multiple-choice question answering	50.0				✓
SciQ (Welbl et al., 2017)	Multiple-choice question answering	25.0	✓			✓

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Can an effective LLM be obtained from only heuristics-based pipeline to create a high-quality web dataset?



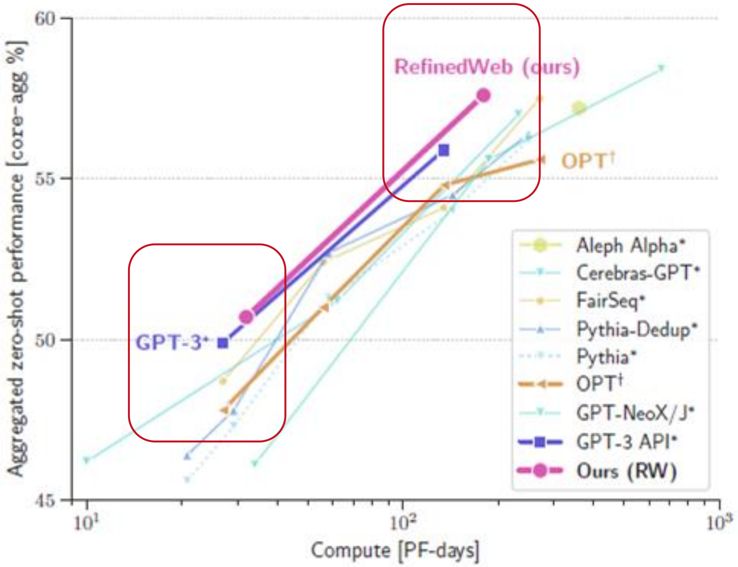
Plot of averaged evaluation scores versus compute budget.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Can an effective LLM be obtained from only heuristics-based pipeline to create a high-quality web dataset?



At similar compute budgets, a model trained purely on RefinedWeb surpases models trained on both Web and highly-curated data. Thus their research question was ascertained.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Generalizability of the RefinedWeb web text data processing pipeline
    - The same data processing pipeline was applied to other existing datasets

	MASSIVE WEB DATASETS			CURATED	OURS
	OSCAR-21.09	OSCAR-22.01	C4	▼ Pile	●RefinedWeb
Base	55.0%	52.7%	**55.7%**	53.4%	52.7%
Filtered	55.4% [+.4]	52.3% [-.4]	**56.2%** [+.5]	54.2% [+.8]	54.3% [+1.6]
removal rate	*-25.0%*	*-39.8%*	*-16.4%*	*-27.1%*	*-50.8%*
Deduplicated	55.6% [+.6]	55.6% [+2.9]	**55.9%** [+.2]	54.5% [+1.1]	
removal rate	*-10.8%*	*-60.8%*	*-7.59%*	*-45.3%*	
Filt.+Dedup.	55.5% [+.5]	55.4% [+2.7]	**56.4%** [+.7]	55.2% [+1.8]	56.2% [+3.5]
removal rate	*-28.2%*	*-62.2%*	*-17.9%*	*-66.0%*	*-75.4%*

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Generalizability of the RefinedWeb web text data processing pipeline
    - The same data processing pipeline was applied to other existing datasets
      - Each processing step eliminated a proportion of the data in other datasets

	Massive web datasets			Curated	Ours
	OSCAR-21.09	OSCAR-22.01	C4	▼ Pile	●RefinedWeb
Base	55.0%	52.7%	**55.7%**	53.4%	52.7%
Filtered	55.4% [+.4]	52.3% [-.4]	**56.2%** [+.5]	54.2% [+.8]	54.3% [+1.6]
removal rate	*-25.0%*	*-39.8%*	*-16.4%*	*-27.1%*	*-50.8%*
Deduplicated	55.6% [+.6]	55.6% [+2.9]	**55.9%** [+.2]	54.5% [+1.1]	
removal rate	*-10.8%*	*-60.8%*	*-7.59%*	*-45.3%*	
Filt.+Dedup.	55.5% [+.5]	55.4% [+2.7]	**56.4%** [+.7]	55.2% [+1.8]	56.2% [+3.5]
removal rate	*-28.2%*	*-62.2%*	*-17.9%*	*-66.0%*	*-75.4%*

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116*

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Generalizability of the RefinedWeb web text data processing pipeline
    - The same data processing pipeline was applied to other existing datasets
      - Each processing step eliminated a proportion of the data in other datasets
      - Furthermore, for the models pretrained on the cleaned datasets by the RefinedWeb pipeline in each stage and across all stages improvements were seen.

	MASSIVE WEB DATASETS			CURATED	OURS
	OSCAR-21.09	OSCAR-22.01	C4	▼ Pile	● RefinedWeb
Base	55.0%	52.7%	**55.7%**	53.4%	52.7%
Filtered	55.4% [+.4]	52.3% [-.4]	**56.2%** [+.5]	54.2% [+.8]	54.3% [+1.6]
removal rate	*-25.0%*	*-39.8%*	*-16.4%*	*-27.1%*	*-50.8%*
Deduplicated	55.6% [+.6]	55.6% [+2.9]	**55.9%** [+.2]	54.5% [+1.1]	
removal rate	*-10.8%*	*-60.8%*	*-7.59%*	*-45.3%*	
Filt.+Dedup.	55.5% [+.5]	55.4% [+2.7]	**56.4%** [+.7]	55.2% [+1.8]	56.2% [+3.5]
removal rate	*-28.2%*	*-62.2%*	*-17.9%*	*-66.0%*	*-75.4%*

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **Evaluations**
  - Generalizability of the RefinedWeb web text data processing pipeline
    - The same data processing pipeline was applied to other existing datasets
      - Each processing step eliminated a proportion of the data in other datasets
      - Furthermore, for the models pretrained on the cleaned datasets by the RefinedWeb pipeline in each stage and across all stages improvements were seen.

	MASSIVE WEB DATASETS			CURATED	OURS
	OSCAR-21.09	OSCAR-22.01	C4	▼ Pile	● RefinedWeb
Base	55.0%	52.7%	**55.7%**	53.4%	52.7%
Filtered	55.4% [+.4]	52.3% [-.4]	**56.2%** [+.5]	54.2% [+.8]	54.3% [+1.6]
removal rate	*-25.0%*	*-39.8%*	*-16.4%*	*-27.1%*	*-50.8%*
Deduplicated	55.6% [+.6]	55.6% [+2.9]	**55.9%** [+.2]	54.5% [+1.1]	
removal rate	*-10.8%*	*-60.8%*	*-7.59%*	*-45.3%*	
Filt.+Dedup.	55.5% [+.5]	55.4% [+2.7]	**56.4%** [+.7]	55.2% [+1.8]	56.2% [+3.5]
removal rate	*-28.2%*	*-62.2%*	*-17.9%*	*-66.0%*	*-75.4%*

The resulting conclusion is that the quality of the pretraining dataset is critical to downstream LLM performance.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **RefinedWeb Summary**
  - Stringent filtering and deduplication could result in a 5T tokens web only dataset suitable to produce models competitive with SOTA, even outperforming LLMs trained on curated corpora.
  - Publicly released a 600GT extract of RefinedWeb
    - 968M individual web pages
    - 2.8TB uncompressed data.
  - Used to train Falcon 7B/40B combined with curated corpora.

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Falcon – An open large language model with state-of-the-art performance

- **In the realm of works creating pre-training datasets, the RefinedWeb pipeline can be considered the current state-of-the-art in terms of producing high-quality downstream web page text data.**

References

Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

Datasets for pretraining

Model	Organization	Date	Training data genre	Training data size	# parameters
GPT-1	OpenAI	Jun 2018	Fiction Books	higher millions	117M
BERT	Google	Oct 2018	Fiction Books, Encyclopedia		340M
GPT-2	OpenAI	Mar 2019	Internet		117M, 345M, 762M, 1.5B
T5	Google	Oct 2019	Internet	34B	60M, 220M, 770M, 3B, 11B
GPT-3	OpenAI	May 2020	Internet w/ encyclopedia, Prose	300B	125M to 175B
GPT-J-6B	EleutherAI	May 2021	Internet w/ encyclopedia & QA, Academic, Prose, Dialogue, Code, Math	402B	6B
LLaMA	MetaAI	Feb 2023	Internet w/ encyclopedia & QA, Academic, Prose, Code	1.4 trillion	7B to 65B
Falcon	TII	May 2023	Internet	in trillions	7B, 40B

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Datasets for Pretraining

Jennifer D'Souza
Technische Informationsbibliothek (TIB)
Welfengarten 1B // 30167 Hannover

Leibniz Gemeinschaft

Conclusion: Takeaways

- As shown in the seminal GPT-1 work, the choice of the pretraining corpus should take into consideration the average context length that the pretraining dataset presents to the transformer architecture. Longer is better. This was not the status quo prior to LLMs.

- Bigger doesn’t always equal better: researchers have found that ultimately smaller, more optimally trained models outshine their behemoth counterparts and require less energy and fewer resources.
  - Refer to the Chinchilla paper by Hoffman et al. “Training Compute-Optimal Large Language Models” on optimal model size and number of tokens for training a transformer language model under a given compute budget. They find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant.

Conclusion: Takeaways

- As shown in the RefinedWeb work for the Falcon models, the quality of the pretraining data is highly conducive to obtaining effective downstream LLMs. In other words, the performance of LLMs on downstream tasks can be improved by pretraining it on a high-quality dataset.

- Effective downstream LLMs also rely on diverse information represented in the pretraining datasets. In this context, web pages work best and are a pretraining data source in almost LLMs.

- Cautionary notes: Self-supervised training on a large corpora of information leads to the model inadvertently learning unsafe content and then sharing it with users. This could be incorrect or misleading medical information and encouraging self-harm. In this context guidelines for safeguards for LLMs are evolving.

TIB

Thank you for your attention!

Questions/Discussion

Find the pre-recorded version of this talk on Youtube!

Find me on LinkedIn!

Creative Commons Namensnennung 3.0 Deutschland
http://creativecommons.org/licenses/by/3.0/de