



Foundation Models for Medical NLP

Borchert, Dr. Schapranow
Data Management for Digital Health
Winter 2023

Agenda

Pillars of the Lecture

Medical Use Cases



Biology Recap



Oncology



Nephrology



Infectious
Diseases

Technology Foundation



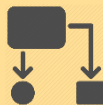
Data
Sources



Data
Formats



Processing and
Analysis



Software
Architectures

Machine Learning

Data



Refine

Evaluate



Prediction +
Probability

Foundation Models

Data Management for
Digital Health, Winter
2023
2

Agenda

Pillars of the Lecture

Medical Use Cases



Biology Recap



Oncology



Nephrology



Infectious
Diseases

Technology Foundation



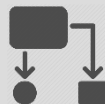
Data
Sources



Data
Formats



Processing and
Analysis



Software
Architectures

Machine Learning

Data



Refine

Evaluate



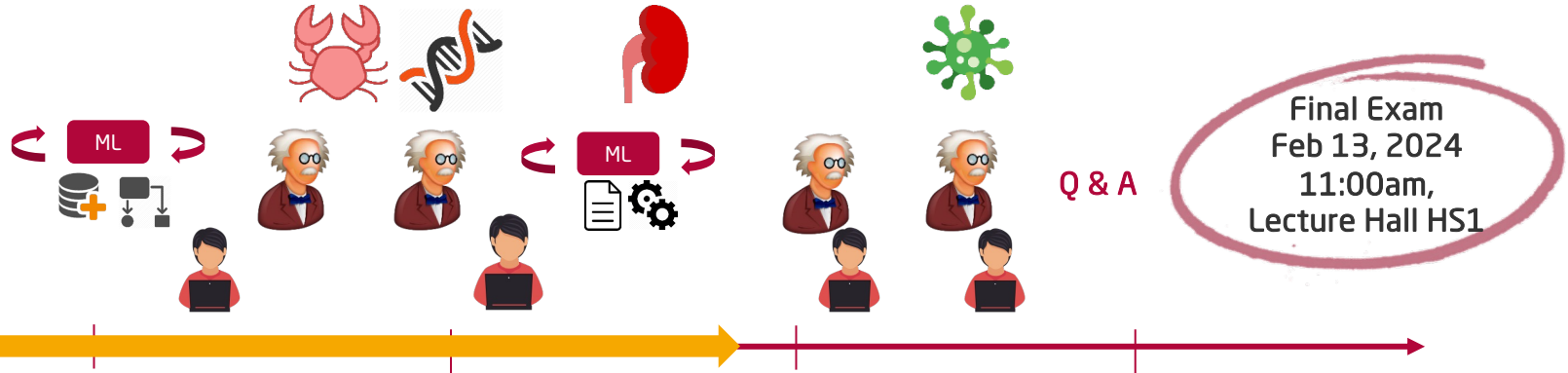
Prediction +
Probability

Foundation Models

Data Management for
Digital Health, Winter
2023

3

Lecture Schedule



Nov

Dec

Jan

Feb

- Lecture Kickoff
- Actors in Healthcare
- Digital Health Data

- Machine Learning (ML) Foundations
- Use Case Oncology
- Biology Recap

- Natural Language Processing
- Use Case Nephrology & Intensive Care
- Supervised ML & Deep Learning

- Use Case Infectious Diseases
- Unsupervised ML

Foundation Models

Data Management for
Digital Health, Winter
2023

Agenda

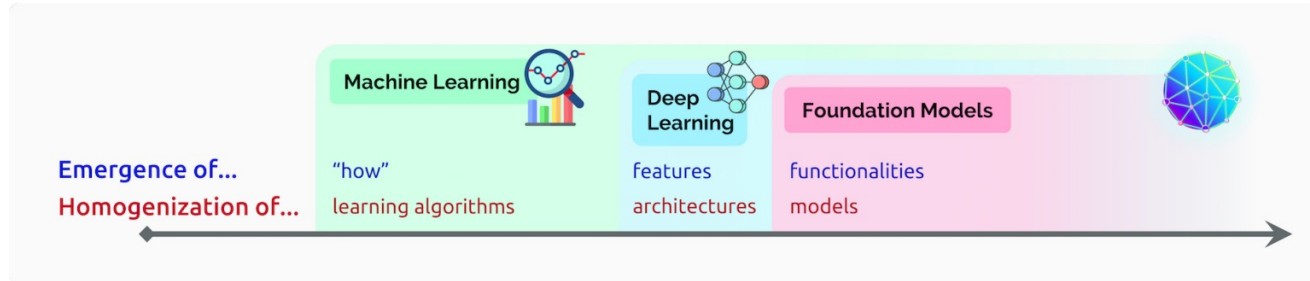
- Foundation Models
- Language Modelling
- Pre-Trained Language Models (BERT, GPT, and co.)
- Zero-shot Inference, In-context Learning
- Problems of LLMs

Foundation Models

Data Management for
Digital Health, Winter
2023

5

Foundation Models

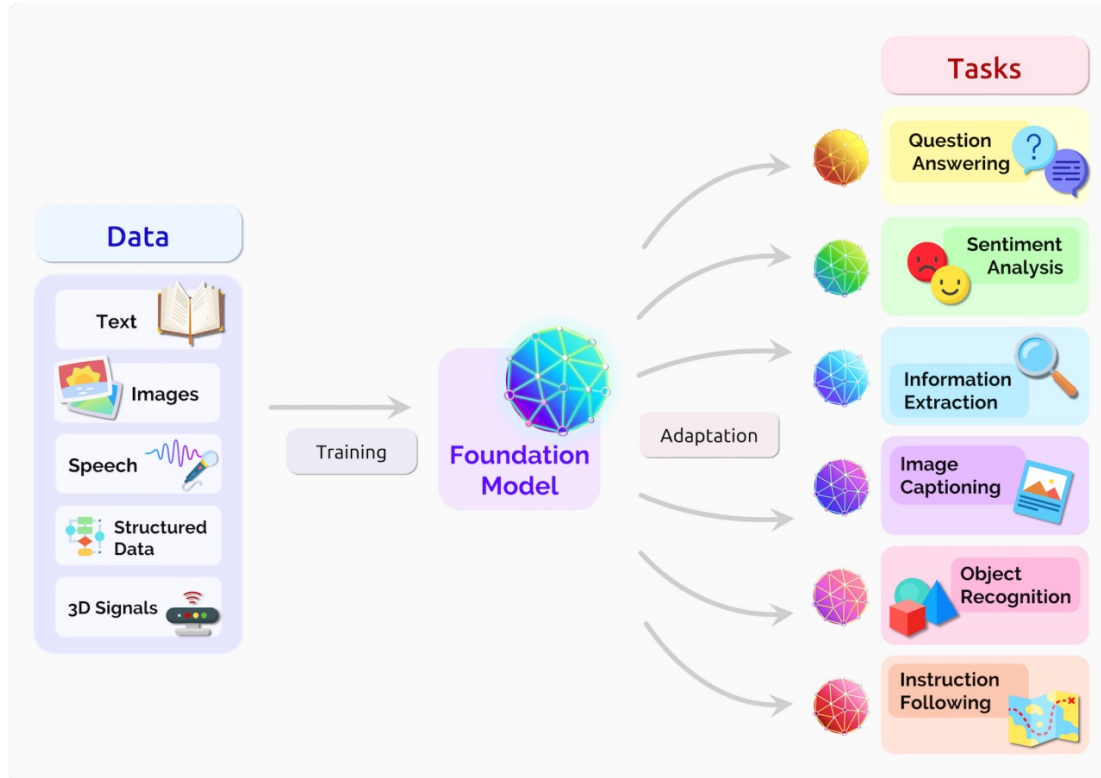


Foundation Models

Data Management for
Digital Health, Winter
2023

6

Adaptation of Foundation Models (Transfer Learning)



Foundation Models

Data Management for
Digital Health, Winter
2023

"Natalie ran out of ... " \rightarrow

P("flour")	= 0.01
P("the")	= 0.01
P("plutonium")	= 0.0001
P("or")	= 0.0000001

- Assigning probabilities to sequences of words
- Applications: Spelling correction, machine translation, speech recognition

Using **chain rule of probability**:

$P(\text{"Natalie ran out of flour"}) =$
 $P(\text{"flour"} \mid \text{"Natalie ran out of"}) \cdot P(\text{"Natalie ran out of"})$

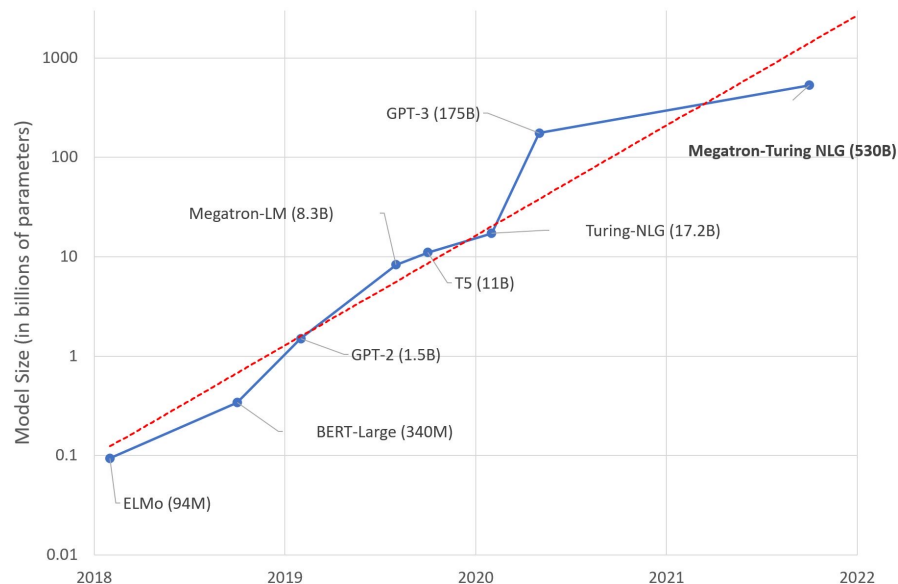
$P(\text{"Natalie ran out of flour"}) > P(\text{"Natalie ran out of plutonium"})$
 $P(\text{"Natalie ran out of flour"}) > P(\text{"Natalie ran out of floor"})$
 $P(\text{"Natalie ran out of flour"}) > P(\text{"flour Nathalie out of ran"})$

Perplexity (of a test set) to evaluate LM performance:

$$W = w_1 w_2 \dots w_n \quad PPL(W) = P(w_1 w_2 \dots w_n)^{-\frac{1}{N}}$$

Language Models as Foundation Models

- Language Models do not need labelled data
 - Leverage vast amounts (trillions of tokens) of diverse texts from websites, textbooks, ...
- Insight: predicting the next word requires an understanding of language (syntax) and probably a model of the world (semantics)
- A ML model trained for LM needs to compress this complex knowledge in its parameters, which can be leveraged through transfer learning
- Problems:
 - Unknown words → Subword tokenization
 - Large scale training → Self-Attention Networks (aka Transformers)



Foundation Models

Data Management for
Digital Health, Winter
2023

- Problems of representing each word individually:
 - **Vocabulary size** can get huge due to morphological richness, compound words, multilinguality
 - Words unseen in the training data can not be represented (**OOV**)
- Idea: use a fixed number of **subtokens** (one or more characters), which represent the input text efficiently
- Basic Algorithm (Byte Pair Encoding):
 - Start with vocabulary with single characters
 - Add most frequent pairs of characters to the vocabulary
 - Repeat until maximum vocabulary size is reached
- Different variants implement other criteria for splitting and merging (e.g., SentencePiece)

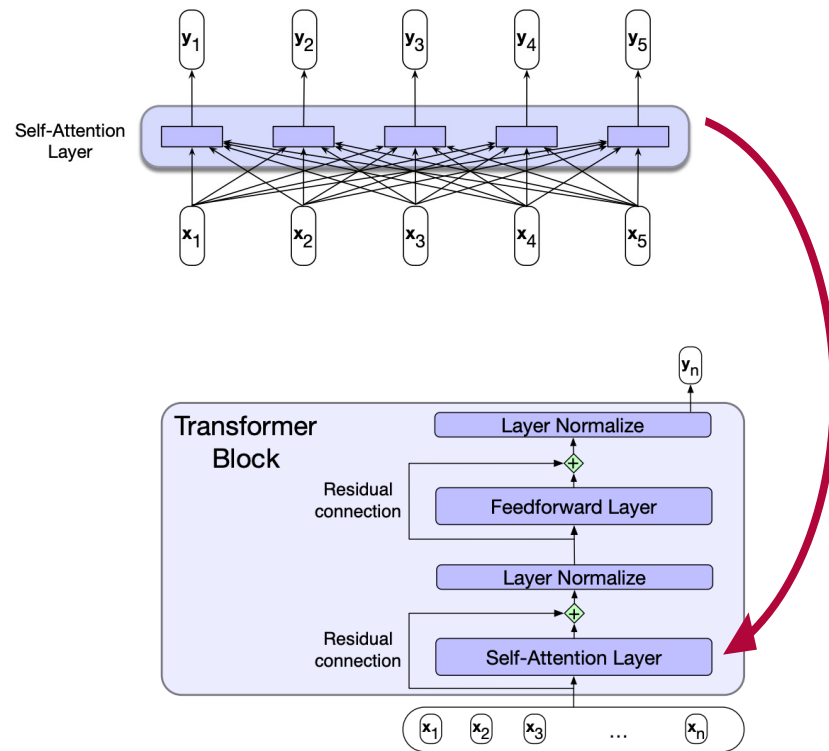
Prüfungsvorleistung



Pr- ü - fung - sv - or - le - istung

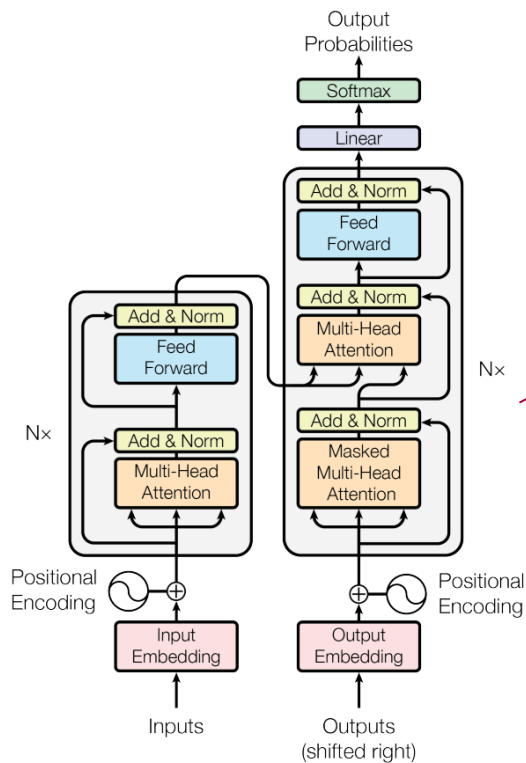
Attention is All you Need (Vaswani et al., 2017)

- Attention is a way to include context when making predictions
- In processing each element of the sequence, the model attends:
 - to all inputs, both before and including the current one (causal), or
 - all the inputs in the sequence (bi-directional)
- Unlike RNNs, computations at each time step are independent and can be performed in parallel

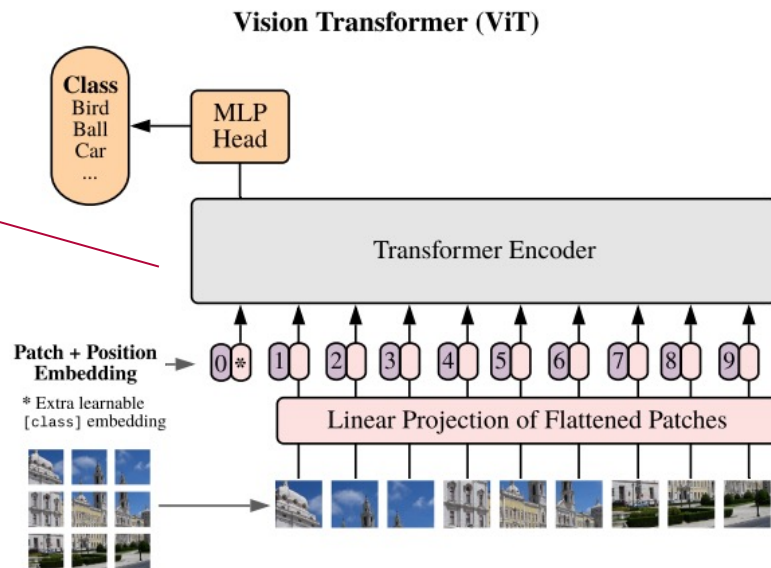


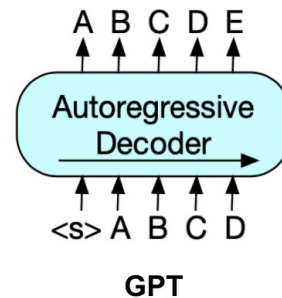
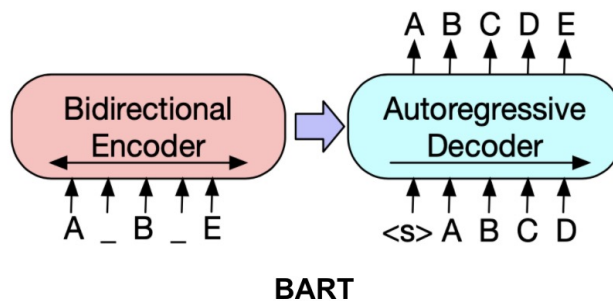
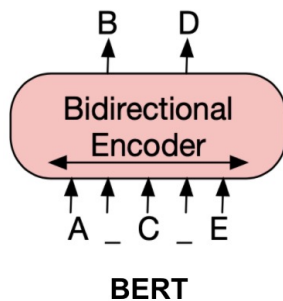
Transformers

General-Purpose Building Blocks



Transformer
Block

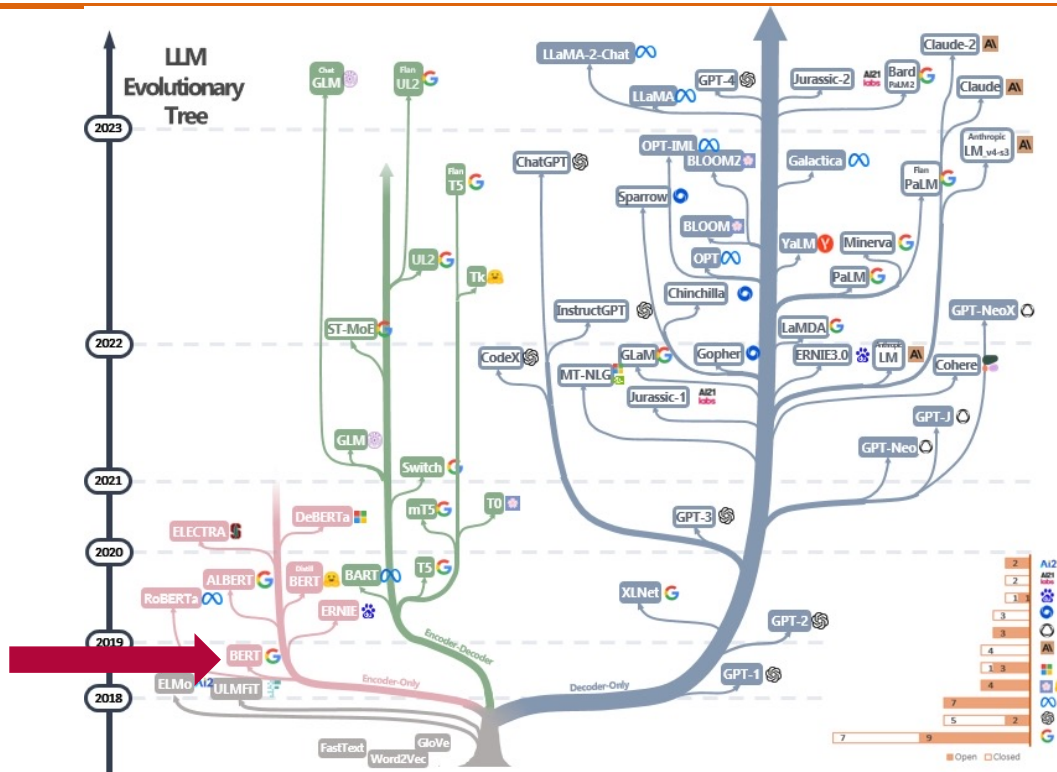




Foundation Models

Data Management for
Digital Health, Winter
2023
13

Families of Language Models

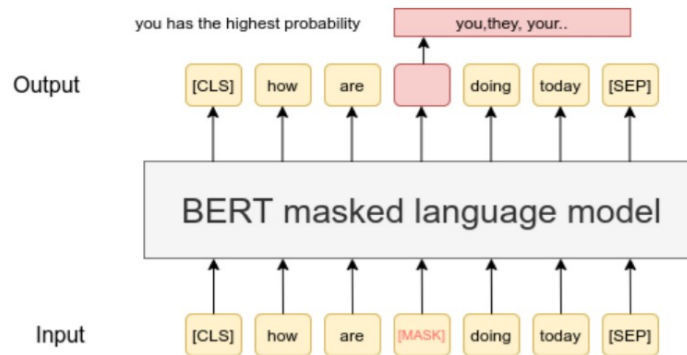


Foundation Models

Data Management for Digital Health, Winter 2023
14

Masked Language Modelling (MLM)

- Original **BERT** (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019):
 - 30,000 token subword vocabulary
 - 12 layers of Transformer blocks
 - 110M (small) to 340M (large) parameters
- Trained with objective to predict:
 - Masked words
 - Next sentence
- Trained on English text only:
 - BooksCorpus (800M words)
 - English Wikipedia (2,500M words)



Foundation Models

Data Management for
Digital Health, Winter
2023
15

Hugging Face Demo

<https://huggingface.co/bert-base-uncased>

⚡ Inference API ⓘ

📄 Fill-Mask

Examples ▾

Mask token: [MASK]

NLP in healthcare is useful for [MASK].

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.066 s



</> JSON Output

📄 Maximize

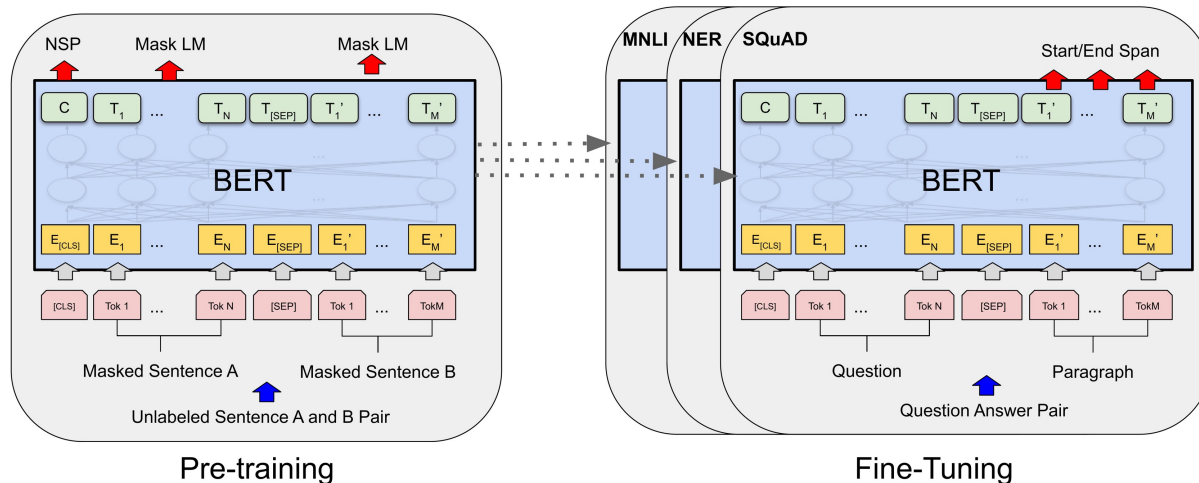
Foundation Models

Data Management for
Digital Health, Winter
2023

16

Pre-Trained Language Model (PLM)

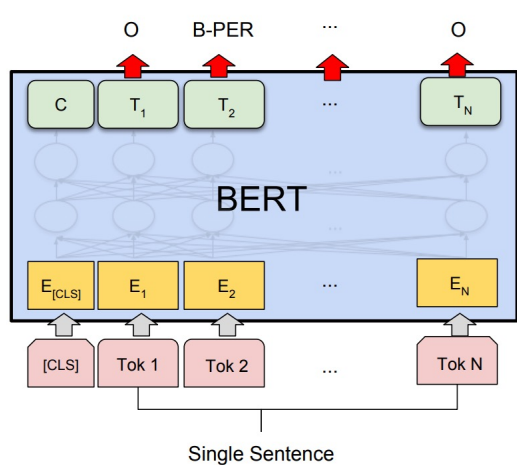
- MLM is usually not a useful task in itself
- For other tasks BERT needs **fine-tuning**
- Encoder of BERT computes a **contextual embedding** for each (sub)token
- Just like static word embeddings, these can be used as an input for other models



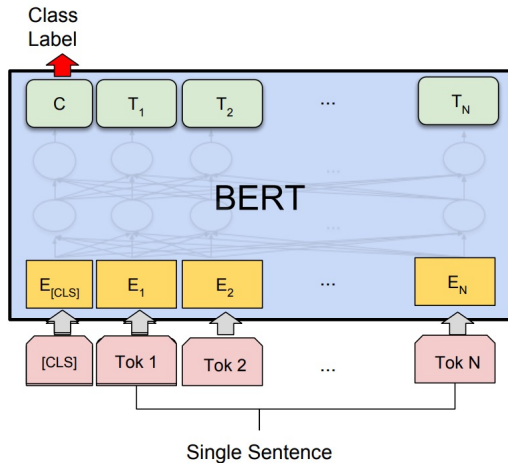
Foundation Models

Data Management for
Digital Health, Winter
2023
17

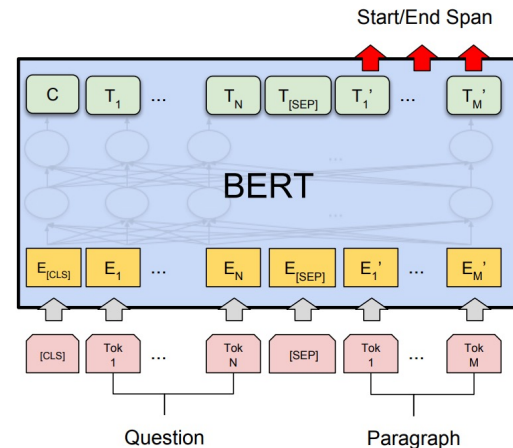
Fine-Tuning BERT for Different Tasks



Sequence Tagging
(like NER)



Text Classification



... many other tasks

Foundation Models

Data Management for
Digital Health, Winter
2023
18

- Instead of training on English, general-domain text, we can use:
 - Text from other languages (GBERT, GottBERT, CamemBERT, BERTje)
 - Multilingual text (mBERT, XLM-RoBERTa)
 - Domain-specific text (BioBERT, PubMedBERT, BioGottBERT, medBERT.de)
- Trade-off between pre-training from scratch and continuing training
 - Training from scratch can result in better presentation of domain-specific terminology (e.g., PubmedBERT)
 - Continual pre-training can leverage existing models



Foundation Models

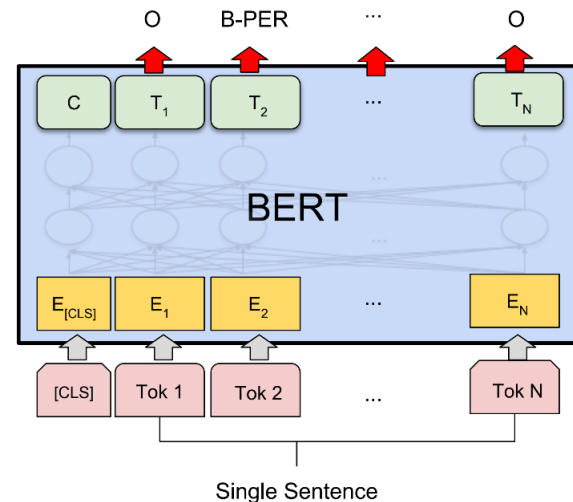
Data Management for
Digital Health, Winter
2023

Example: GGPONC NER Models

- Fine-tuning pre-trained BERT models + token classification head with approx. 200k manual NER annotations for findings, procedures and substances
- Domain-specific models outperform general-domain ones

Model	AUROC	AUROC _{tok}	F1 _{mac}	F1 _{tok}
GGPonc				
GottBERT	98.01	98.07	75.18	73.43
BioGottBERT	97.96	98.05	76.07	74.62
Multilingual BERT	97.81	97.95	73.75	72.20
German-MedBERT	97.55	97.66	74.48	72.45
medBERT.de	98.20	98.35	75.93	75.12
medBERT.de _{dedup}	98.10	98.22	75.57	74.92

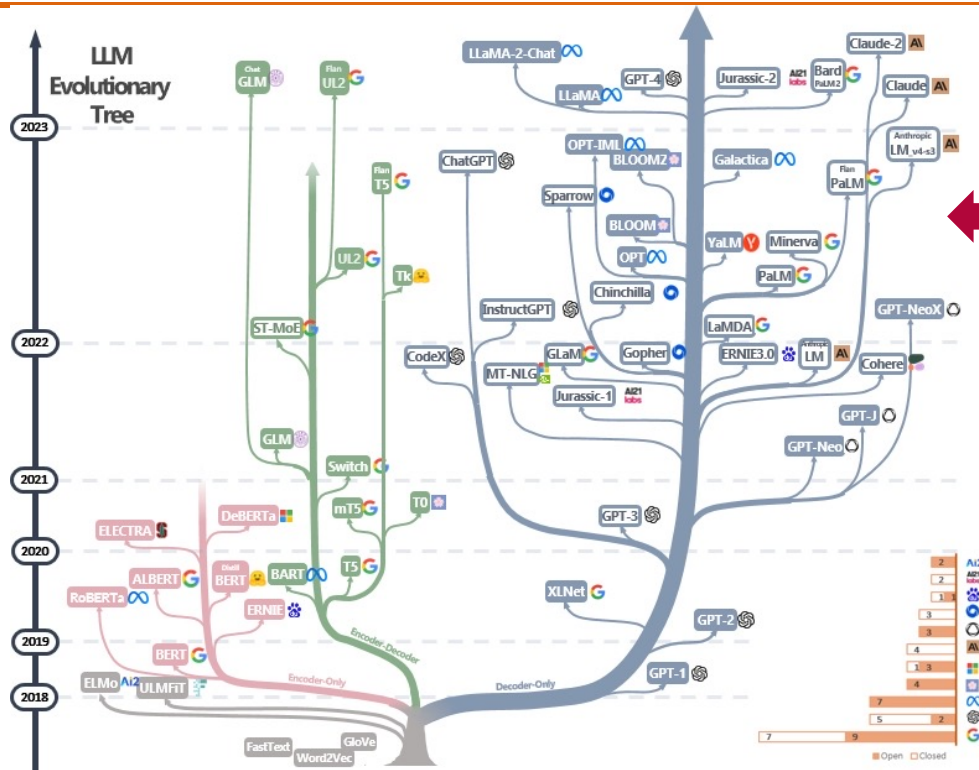
Test set performance



ML and Corpora

Data Management for
Digital Health, Winter
2022
20

Families of Language Models

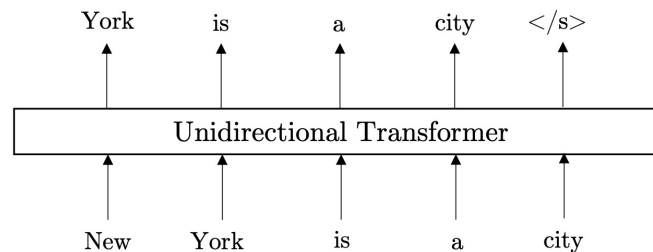


Foundation Models

Data Management for
Digital Health, Winter
2023
21

Autoregressive (Large) Language Models

- Learn to predict the next token
- GPT-3 (Generative Pre-trained Transformer) has 176B (!) parameters and was trained on 570GB of text
- Many tasks can be framed as text generation problems, but output can be quite unpredictable / harmful



$$\log p(\mathbf{x}) = \sum_{t=1}^T \log p(x_t | \mathbf{x}_{<t})$$

Foundation Models

Data Management for
Digital Health, Winter
2023
22

What is in the Pre-training Data of LLMs?

- It's hard to figure out the actual training data used for LLM development, but we know it for some examples, e.g., Llama-1 (Meta)

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

Foundation Models

Data Management for
Digital Health, Winter
2023
23

Large Language Models are

large artificial neural networks that are trained on a large **corpus** of text. They are supervised models are typically trained using unsupervised

amount = 8.39%

body = 1.14%

corpus = 73.91%

dataset = 6.3%

text = 5.24%

Total: -0.3 logprob on 1 tokens

(94.98% probability covered in top 5 logits)

Foundation Models

Data Management for
Digital Health, Winter
2023
24

English: Large Language Models are great.

German: Große Sprachmodelle sind großartig.

Original:

Welcome to the class: We are very excited that you are interested in learning more about the principles of data management for digital health and

Specific examples from selected fields of digital health to understand where and how data is/needs to be acquired,
Known challenges in acquiring and processing these types of data in their specific digital health domain,
How to deal with and address specific requirements and limitation of accessing and using digital health data,
and how the complex analysis of high-dimensional multi-modal digital health data can be facilitated through the use of latest soft- and hardware learning approaches.

After participating in the course, you will be equipped with the ability to:

Assess requirements of selected real-world use cases from the medical field (oncology, nephrology, infectious diseases),
Select latest technology building blocks to create viable healthcare software solutions, and
Analyze requirements for data analysis and processing, e.g. for machine learning (supervised and unsupervised learning, large language models)
In the course, we will have invited guest speakers with medical background sharing their real-world experience with you in interactive presentations.

Further details about the structure of the lecture will be shared with you in the kickoff lecture (please check the date on the top of this page).

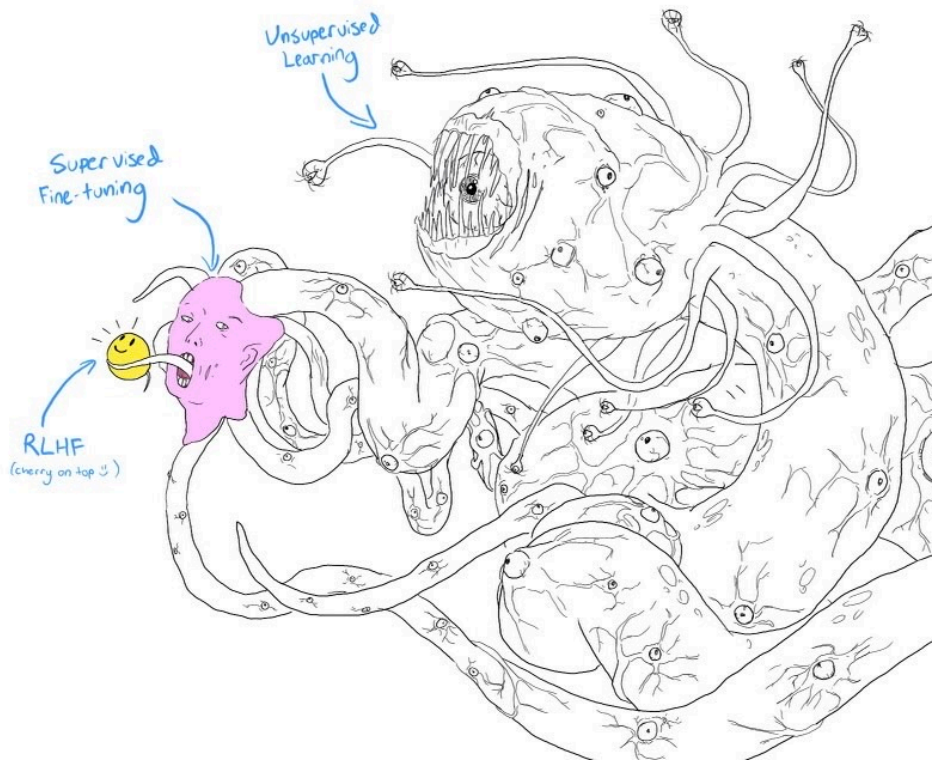
Summary:

This lecture will cover the principles of data management for digital health and how it differs from other topics. We will provide examples of digital health requirements for medical use cases, select technology building blocks, and analyze data for machine learning. There will be guest speakers to answer your questions.

Foundation Models

Data Management for
Digital Health, Winter
2023
25

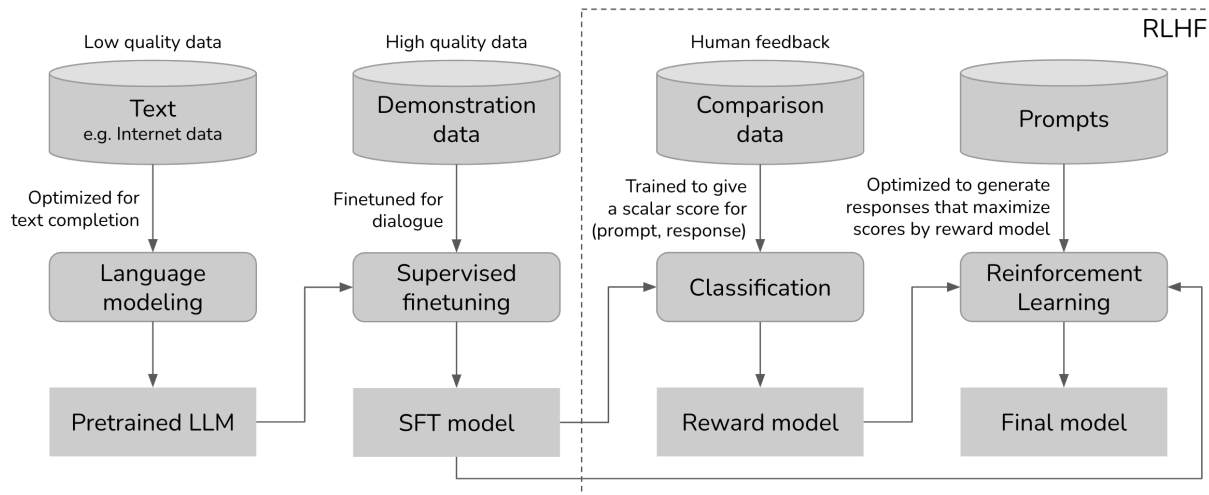
Making LLMs More Useful / Less Harmful



Foundation Models

Data Management for
Digital Health, Winter
2023
26

Fine-Tuning LLMs with High Quality Data and Human Feedback



Scale
May '23

>1 trillion
tokens

10K - 100K
(prompt, response)

100K - 1M comparisons
(prompt, winning_response, losing_response)

10K - 100K
prompts

Examples
Bolded: open
sourced

GPT-x, Gopher, **Falcon**,
LLaMa, **Pythia**, **Bloom**,
StableLM

Dolly-v2, **Falcon-Instruct**

InstructGPT, ChatGPT,
Claude, **StableVicuna**

Foundation Models

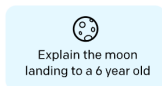
Data Management for
Digital Health, Winter
2023
27

Supervised Fine-Tuning and Reinforcement Learning from Human Feedback

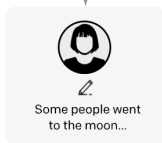
Step 1

Collect demonstration data, and train a supervised policy.

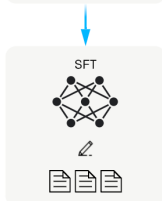
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



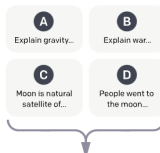
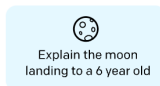
This data is used to fine-tune GPT-3 with supervised learning.



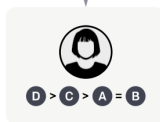
Step 2

Collect comparison data, and train a reward model.

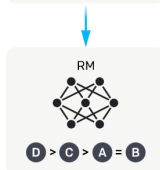
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

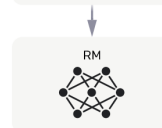
A new prompt is sampled from the dataset.



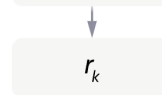
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Foundation Models

Data Management for
Digital Health, Winter
2023
28

Example: Phenotyping

"... SOCIAL HISTORY:
Widowed since 1972, no
tobacco, no alcohol, lives
alone."

non-smoker

"... Social History: No
alcohol use and quit
tobacco greater than 25
years ago with a 10-pack
year smoking history.""

ex-smoker

"..He is a heavy smoker
and drinks 2-3 shots per
day at times."

current smoker

Text Data & NLP

Data Management for
Digital Health, Winter
2023
29

What is the smoking status of the person described in this clinical note:

```
"Social History: No alcohol use and quit tobacco greater than 25 years ago with a 10-pack year smoking history."
```

Prompt Template

Input

- Zero-shot prompting without providing any specific examples or prior context
- Model is expected to understand and execute the task based solely on its pre-existing knowledge and instructions in the prompt

Foundation Models

Data Management for
Digital Health, Winter
2023
30

Your task is to determine the smoking status of the person described in a clinical note.

Here are some examples:

Input: "Smoker until 1999"

Output: ex-smoker

Input: "SOCIAL HISTORY: Widowed since 1972, no tobacco, no alcohol, lives alone."

Output: non-smoker

Input: "He is a heavy smoker and drinks 2-3 shots per day at times."

Output: current smoker

Input:

"Social History: No alcohol use and quit tobacco greater than 25 years ago with a 10-pack year smoking history."

Output:

Prompt Template

Few-Shot Examples

Input

- In-context learning (ICL) refers to the model's ability to adapt and respond based on the immediate context provided within a prompt.
- Few-shot prompting is a type of ICL where the model is given a few examples to illustrate the task
- Examples can be fixed or based on the input

Foundation Models

Data Management for
Digital Health, Winter
2023



https://github.com/hpi-dhc/dm4dh-2023/blob/main/6_LLMs.ipynb

Foundation Models

Data Management for
Digital Health, Winter
2023
32

Solution to All NLP Problems?

- Supervised fine-tuning often outperforms zero-shot / few-shot prompting, in particular for very specific tasks
 - e.g., many information extraction subtasks
- Data contamination in benchmarks is an issue, when test sets are freely available on the internet
- **Additional issues of using a commercial API provider?**
 - Data privacy / IP
 - Reproducibility (Model Updates)
 - Reliability / Availability
 - Transparency (Training Data, Post Filtering)
 - Cost (trade-off)
 - Dependencies / Vendor Lock-In



Foundation Models

Data Management for
Digital Health, Winter
2023
33

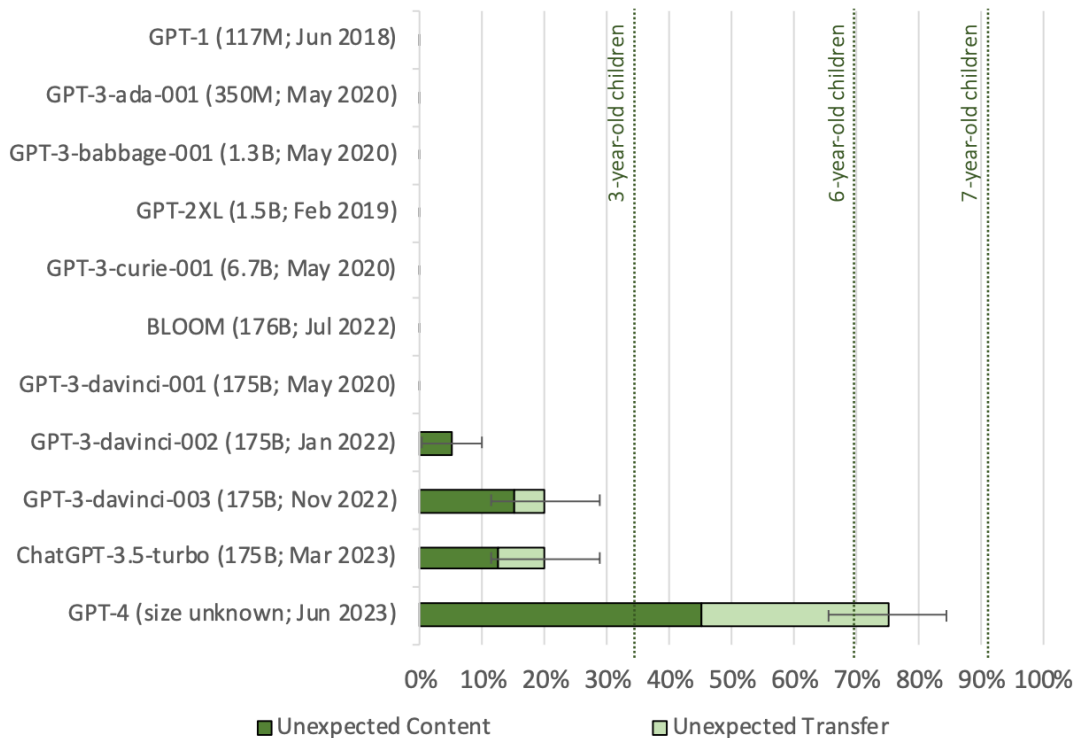
LLM Benchmarks

Rank	Size	Model	Total	MT-Bench	CoT	Python Code	LM-Eval
1		GPT-4-0613	77.78	8.89	0.71	0.66	
2		GPT-3.5-Turbo-0613	65.17	8.22	0.52	0.53	
3		GPT-3.5-Turbo-0301	63.67	8.11	0.48	0.55	
4	70B	WizardLM 70B V1.0	56.38	7.68	0.47	0.27	72.23
5	70B	Stable Beluga 2	55.60	7.42	0.47	0.29	
6	70B	OpenAssistant Llama-2 70B V10	55.01	7.13	0.47	0.31	74.89
7	70B	WizardMath 70B V1.0	51.78	6.86	0.49	0.16	
8	70B	LLaMA-2 70B Chat	48.18	7.10	0.36	0.20	71.54
9	34B	WizardCoder Python 34B V1.0	47.15	6.30	0.29	0.49	
10	65B	Guanaco-65B	44.32	6.63	0.33	0.15	73.92
11	13B	Open-Assistant Code-Llama 13B V10	43.98	6.00	0.29	0.37	62.03
12	33B	Vicuna 33B V1.3	43.04	6.97	0.28	0.11	70.35
13	7B	Mistral-7B-Instruct-v0.1	42.66	6.69	0.28	0.15	

Foundation Models

Data Management for
Digital Health, Winter
2023
34

Capabilities of LLMs "Theory of Mind"



Foundation Models

Data Management for
Digital Health, Winter
2023
35

Path to AGI?

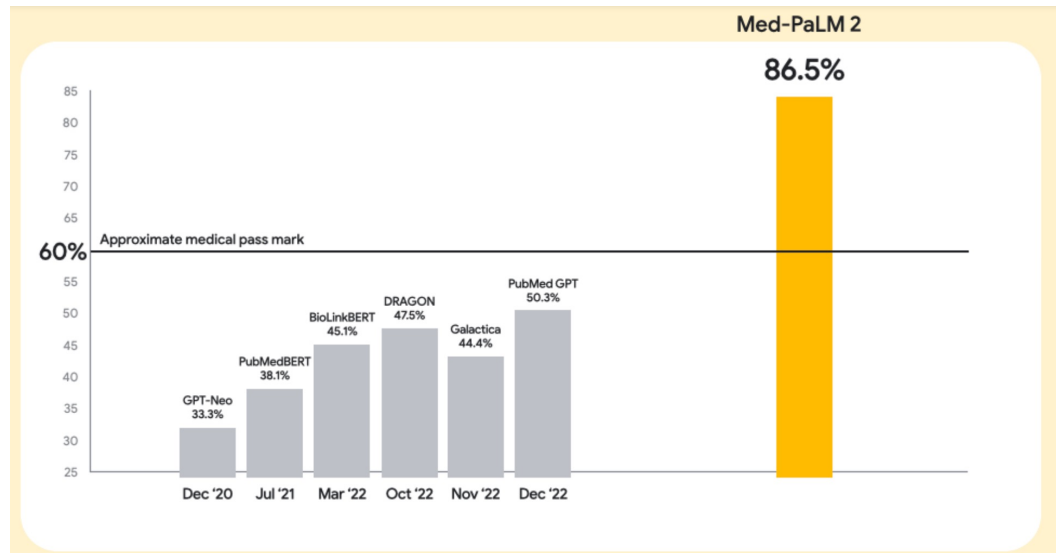
Performance (rows) x Generality (columns)	Narrow <i>clearly scoped task or set of tasks</i>	General <i>wide range of non-physical tasks, including metacognitive abilities like learning new skills</i>
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging <i>equal to or somewhat better than an unskilled human</i>	Emerging Narrow AI GOFAI ⁴ ; simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023)
Level 2: Competent <i>at least 50th percentile of skilled adults</i>	Competent Narrow AI toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert <i>at least 90th percentile of skilled adults</i>	Expert Narrow AI spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso <i>at least 99th percentile of skilled adults</i>	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman <i>outperforms 100% of humans</i>	Superhuman Narrow AI AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved

Morris et al. (2023), Levels of AGI: Operationalizing Progress on the Path to AGI, <https://arxiv.org/abs/2311.02462>

Foundation Models

Data Management for Digital Health, Winter 2023
36

- BioGPT (Luo et al., 2022):
 - GPT-2 Model trained on PubMed articles
 - Supervised Fine-Tuning on Narrow Tasks (NER, RE, QA, ...)
- PubMedGPT / BioMedLM (Stanford, 2022):
 - 2.7B Parameter GPT Model trained on PubMed articles
- Med-PaLM 2 (Singhal et al., 2023)
 - PaLM 2 (Google, general-purpose)
 - Instruction Finetuning on Medical QA datasets

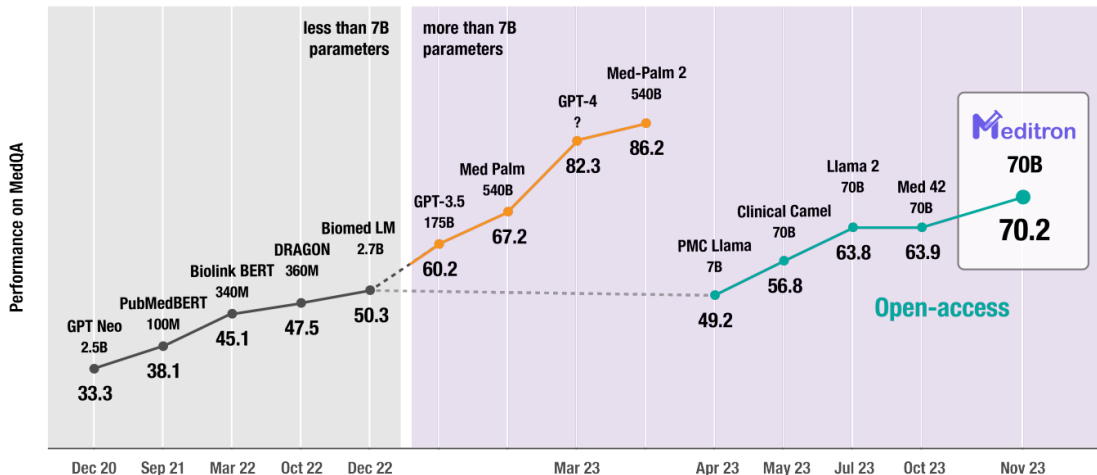


Foundation Models

Data Management for
Digital Health, Winter
2023
37

MediTron

- Based on Llama-2 (70B)
- Fine-tuned on 48.1B tokens of:
 - PubMed Articles
 - Clinical Guidelines
 - Small “replay” dataset from general domain to prevent catastrophic forgetting
- For Evaluation: task-specific fine-tuning on Medical QA dataset



Redistribution allowed
 → Data released
 → Scrapers and cleaners released



Redistribution forbidden
 → Scrapers and cleaners released

Foundation Models

Data Management for Digital Health, Winter 2023
 38

Hallucinations

- LLMs often generate text, that sounds reasonable, but is factually incorrect (aka confabulations)
- Potential reasons:
 - Internet is full of wrong information
 - Human raters in RLHF / instruction-tuning might prefer confident answers
 - "Just text" - no grounding in the real world
 - Ambiguous Queries
- Potential solutions (none of them is perfect):
 - Output Filters
 - Fine-Tuning
 - Give LLM access to "world knowledge"

F You

Give me 10 literature references to include in my lecture "Data Management for Digital Health"

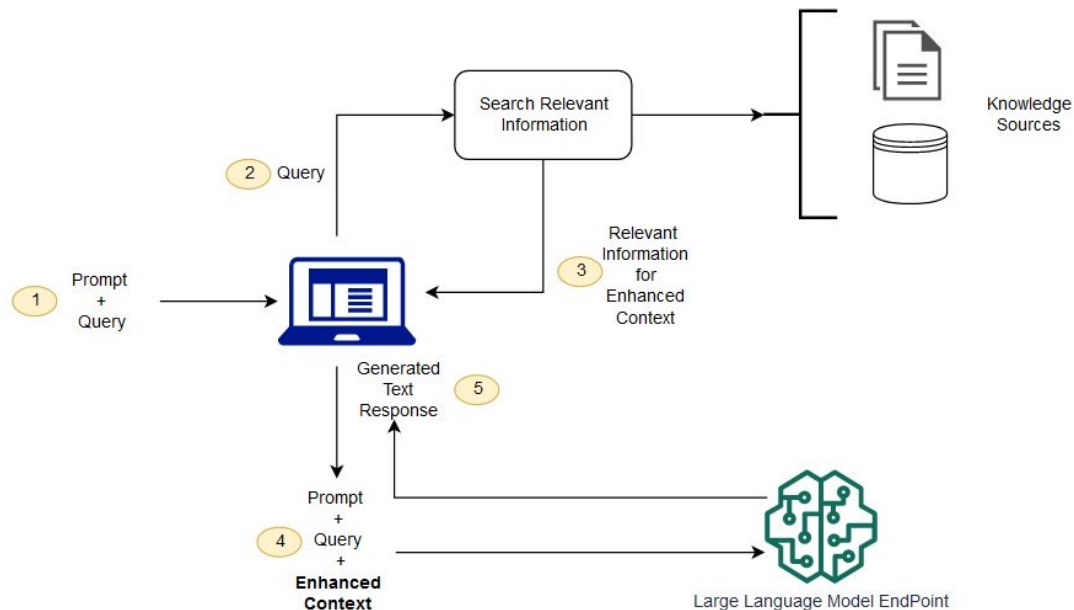
🌀 ChatGPT

For a lecture on "Data Management for Digital Health," you would want to cover various aspects including the fundamentals of data management, specific challenges in the healthcare domain, privacy and security concerns, and the application of modern technologies like AI and Big Data. Here are 10 literature references that span these areas:

1. **"Big Data in Healthcare: Management, Analysis and Future Prospects"** by K. R. Latha and S. J. Vijay.
 - Focuses on big data analytics in healthcare, discussing management and analysis of large datasets.
2. **"Principles of Health Interoperability: HL7 and SNOMED"** by Tim Benson and Grahame Grieve.
 - Offers an in-depth understanding of health data standards, crucial for effective data management.
3. **"Healthcare Data Analytics"** by Chandan K. Reddy and Charu Aggarwal.
 - A comprehensive guide on data analytics applications in healthcare.
4. **"Data Management in Healthcare Systems"** by Lina F. Soualmia and Si-Hung Nguyen.
 - This book explores various aspects of data management in healthcare systems, including data quality and integration.
5. **"Blockchain in Healthcare: Innovations that Empower Patients, Connect Professionals and Improve Care"** by Vikram Dhillon, David Metcalf, and Max Hooper.
 - Discusses the application of blockchain technology in healthcare for secure data management.

Retrieval Augmented Generation

- Combines LLMs with external knowledge retrieval mechanism, using:
 - Standard text search
 - Vector databases for similarity search with word embeddings
- Given a query, the retrieval system retrieves contextually relevant documents or data
- The language model is prompted to generate an answer to the query using the original query + the retrieved context



Retrieval Augmented Generation

you.com

What is retrieval augmented generation?

Retrieval Augmented Generation (RAG) is an AI framework that combines large language models (LLMs) with external knowledge sources to generate responses [1](#) [2](#). It addresses knowledge-intensive tasks where additional background knowledge is required. RAG consists of two phases: retrieval and content generation.

In the retrieval phase, algorithms search for and retrieve relevant information from external knowledge bases or documents based on the user's prompt or question. This retrieved information is then used in conjunction with the original input in the content generation phase to produce an output [3](#). By integrating external knowledge, RAG aims to improve the factual consistency, reliability, and accuracy of the generated responses.

RAG is particularly useful for tasks that require factual knowledge and precise manipulation of information. It allows LLMs to access and ground themselves in up-to-date and accurate information from external sources. This helps mitigate the problem of "hallucination," where LLMs generate random or incorrect information based on their training data.

The integration of retrieval and generation in RAG enables LLMs to produce more contextually relevant and accurate responses by leveraging external knowledge. It provides users with insights into the generative process of LLMs and ensures that the responses are grounded in real-world information.

RAG can be implemented using various techniques and frameworks, such as Azure Machine Learning, Huggingface Transformers, or Amazon SageMaker. These platforms provide tools and resources to build and deploy RAG models effectively.

Please note that the information provided above is based on the search results and may not cover all aspects of retrieval augmented generation. For a more comprehensive understanding, I recommend



Related web results

Retrieval Augmented Generation using Azure Machine Learning ...

learn.microsoft.com

31 Jul 2023 ... Retrieval Augmented Generation (RAG) is a pattern that works with pretrained Large Language Models (LLM) and your own data to generate ...



• Retrieval Augmented Generation (RAG) | Prompt Engineering Guide

promptingguide.ai

7 Aug 2023 ... RAG combines an information retrieval component with a text generator model. RAG can be fine-tuned and its internal knowledge can be modified ...

✗ Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

arxiv.org

22 May 2020 ... Title:Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks ... Abstract: Large pre-trained language models have been shown to store ...



Retrieval Augmented Generation (RAG) - Amazon SageMaker

docs.aws.amazon.com

Retrieval Augmented Generation ChatGPT Plugins

ChatGPT Plugins ▾



You

Give me 10 literature references to include in my lecture "Data Management for Digital Health"



ChatGPT

Used Paperpile ▾

Here are 10 literature references that could be valuable for your lecture on "Data Management for Digital Health":

1. Demirci, H., & Yardan, E. D. (2023). Data management in the digital health environment scale development study. [Link](#)
2. Oliva, A., Grassi, S., Vetrugno, G., Rossi, R., Della Morte, G., Pinchi, V., & Caputo, M. (2022). Management of Medico-Legal Risks in Digital Health Era: A Scoping Review [Link](#)
3. Uslu, D. (2023). Investigation of the Relationship Between Distrust in Health Systems and Data Management in the Digital Health Environment. [Link](#)
4. Ylitalo, A., Laukka, E., Heponiemi, T., & Kanste, O. (2022). Primary healthcare managers' perceptions of management competencies at different management levels in digital health services: secondary analysis. [Link](#)
5. Rychnovská, D. (2021). Anticipatory Governance in Biobanking: Security and Risk Management in Digital Health. [Link](#)

Foundation Models

Data Management for
Digital Health, Winter
2023
42

What to Take Home



- (Large) Language Models and Variants (MLM, Autoregressive)
- Subword Tokenization
- Transformer Architecture
- Fine-Tuning vs. Zero-Shot Prompting vs. Few-Shot Prompting
- Hallucinations and RAG

Foundation Models

Data Management for
Digital Health, Winter
2023
43