Deep learning with textual sequences

Artificial Intelligence for Digital Health (AID)
M.Sc. in Digital Health – University of Pisa
Davide Bacciu (davide.bacciu@unipi.it)

Lecture outline

- Representing textual information
  - Word embeddings
  - Skip-grams
- Tackling textual modelling tasks
  - Masked language modelling
  - Relevant language model architectures
  - Pretraining and fine tuning
- Language models in healthcare
- Foundation models

Text sequences (in Healthcare)

Much valuable healthcare information is “hidden” in natural language text

Mr. Blind is a 79-year-old white white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center.

The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient’s hematocrit was stable and he was given no further intervention.

The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.

On the morning of the 22nd the patient developed tachypnea congestive heart failure. A medical consult was obtained at t

Hospital. The patient was given intravenous Lasix.

Credits: MIT HST.956

orange=demographics
blue=patient condition, diseases, etc.
brown=procedures, tests
magenta=results of measurements
purple=time

EHR language has its own challenges

Code and acronyms matter

3/11/98 IPN	(date of) Intern Progress Note,
SOB & DOE ↓	the patient's shortness of breath and dyspnea on exertion are decreased,
VSS, AF	the patient's vital signs are stable and the patient is afebrile,
CXR ⊕ LLL ASD no Δ	a recent new chest xray shows a left lower lobe air space density that is unchanged from the previous radiograph,
WBC 11K	a recent new white blood cell count is 11,000 cells per cubic milliliter,
S/B Cx ⊕ GPC c/w PC, no GNR	the patient's sputum and blood cultures are positive for gram positive cocci consistent with pneumococcus, no gram negative rods have grown,
D/C Cef →PCN IV	so the plan is to discontinue the cefazolin and then begin penicillin treatment intravenously.

Telegraphic language

Prototypical tasks in healthcare natural language processing (NLP)

- Assign a meaning (in any) from some taxonomy/ontology/terminology to words or phrases (entity recognition)
  - e.g., “rheumatoid arthritis” ==> 714.0 (ICD9)
- Determine if a word/phrase represents protected health information
  - e.g., “Mr. Bill suffers from Huntington’s Disease”
- Determine aspects of each entity: time, location, certainty, ...
- Determine the relationship between relevant entities
  - e.g., precedes, causes, treats, prevents, indicates, ...
- Identify document fragments that are relevant for answering to a specific medical question;
  - e.g., where is the patient’s exercise regimen discussed?
- Summarize documents or parts of them

Traditional Information Extraction Tasks in NLP

• Named Entity Recognition (NER): Identify and classify entities (e.g., diseases, drugs, organizations) in text.

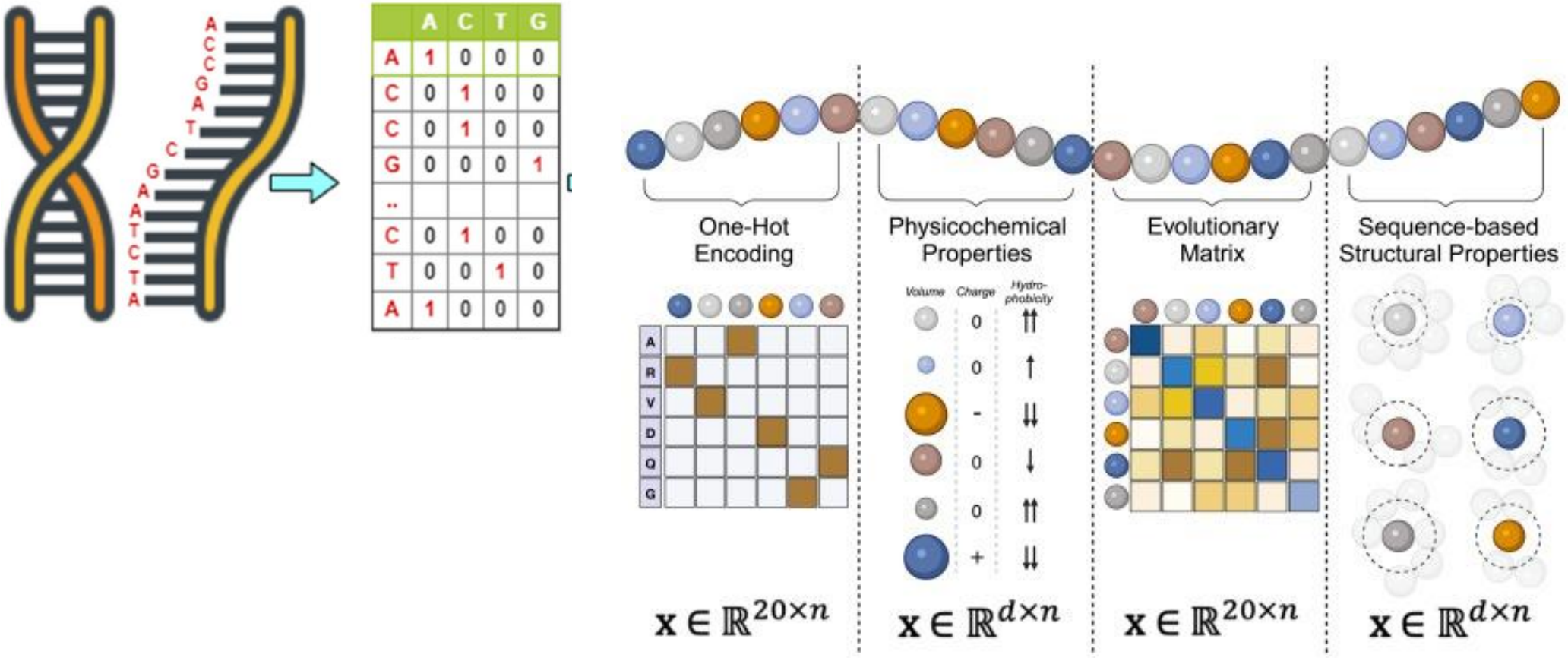
• Relation Extraction (RE): Find relationships between entities (e.g., 'causes', 'treats').

• Event Extraction: Detect complex events involving multiple entities.

• Coreference Resolution: Link different expressions referring to the same entity.

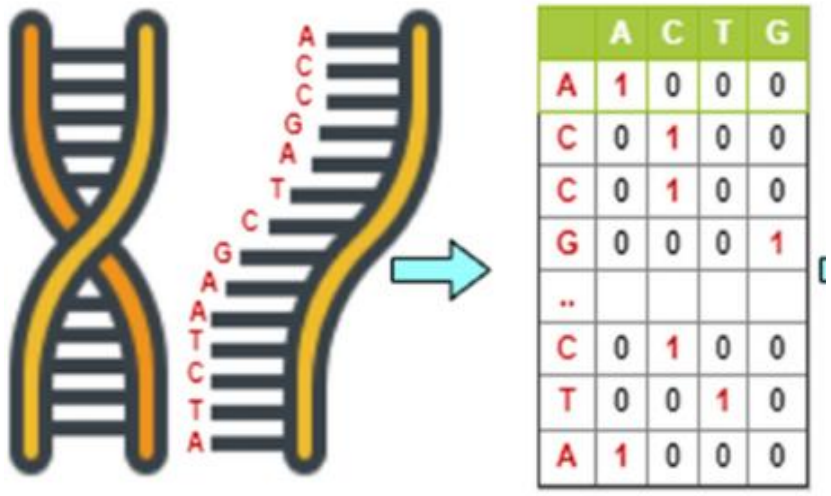
• Question Answering (QA): Extract specific answers from text given a question.

Value is not only on natural language text



Representing text (of any nature)

We have already encountered a way to represent the symbolic information of textual data as numerical features



	A	C	T	G
A	1	0	0	0
C	0	1	0	0
C	0	1	0	0
G	0	0	0	1
..				
C	0	1	0	0
T	0	0	1	0
A	1	0	0	0

What could possibly go wrong?

$$a = [1, 0, \ldots, 0, \ldots, 0, 0]$$
$$abaco = [0, 1, \ldots, 0, \ldots, 0, 0]$$
$$\ldots = \ldots$$
$$modello = [0, 0, \ldots, 1, \ldots, 0, 0]$$
$$\ldots = \ldots$$
$$zuppiera = [0, 0, \ldots, 0, \ldots, 1, 0]$$
$$zuppo = [0, 0, \ldots, 0, \ldots, 0, 1]$$

Issues of one-hot representations

Issue n. 1: vector size can be very large in natural languages

- 100k+ for the largest vocabularies (e.g., Chinese)

Issue n. 2 : one-hot vectors are orthogonal

- I would like a glass of orange juice
- I would like a glass of apple juice
- I would like a glass of eggplant juice

It's clear that semantically, *orange juice* and *apple juice* are much more closely related than *orange juice* and *eggplant juice*.

However, with one-hot vectors, there would be no difference between them, and it becomes difficult to generalize the concept of *fruit juice*

Word Embeddings

The starting point: context matters

Mr. Huntington was treated for Huntington's Disease at Huntington Hospital, located on Huntington Avenue

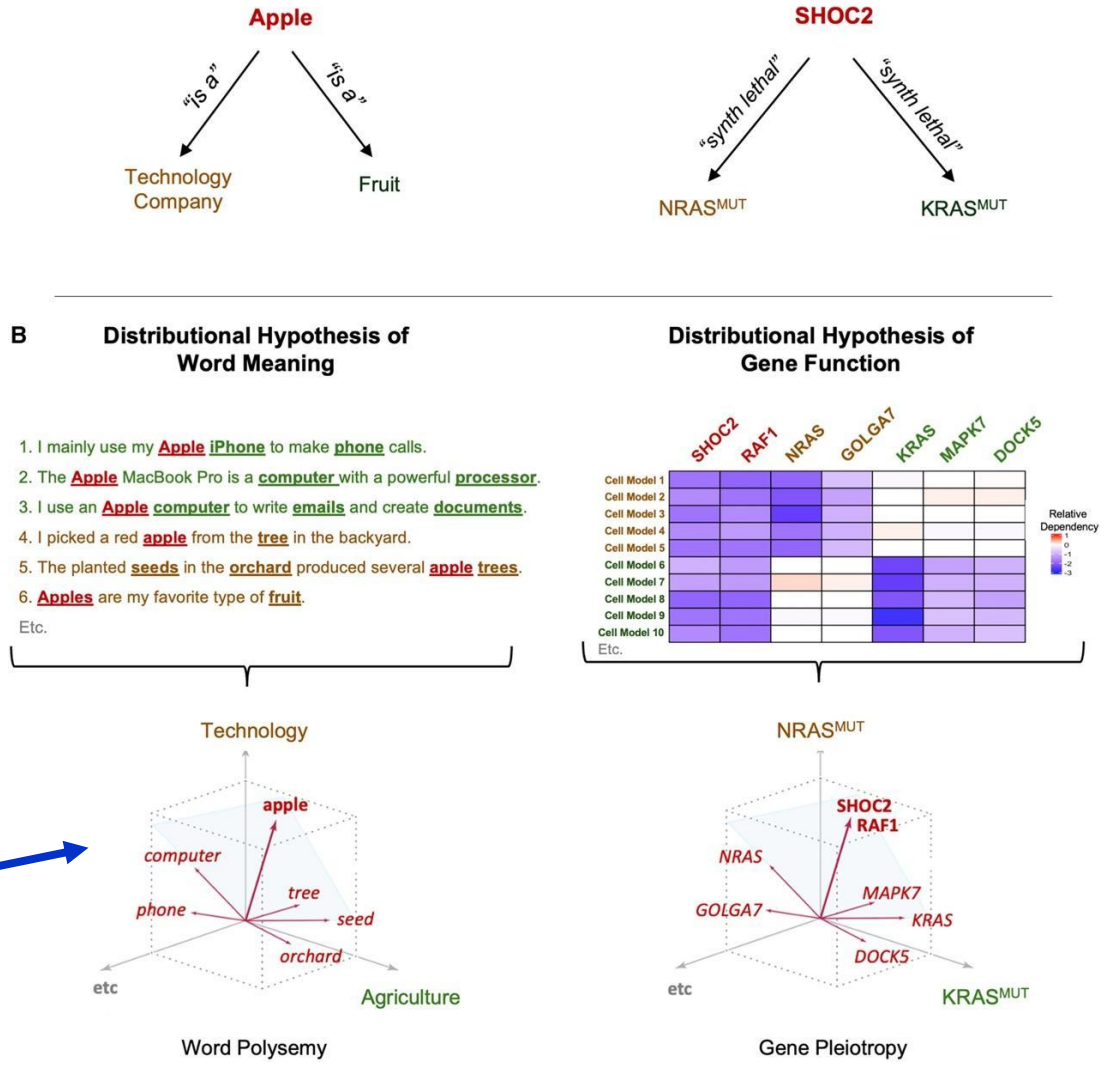
- Same word, different role, interpretation and relevance for the task at hand
- Yet we are able to clearly identify the meaning of a word thanks to the context in which a word occurs

Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings (Harris, 1954)

A word is characterized by the company it keeps (Firth, 1957)

We need a vectorial representation of text, consistent with the distributional hypothesis



Credits: M. Zitnik. AIM2, 2025

Word embedding (distributed word representation)

Dense vectorial word representation

- From: *model* = $[0 \ldots 0\ 1\ 0 \ldots 0]$
- To: *model* = $[0.23\ 0.61 \ldots 0.12]$

where

- The size of the dense representation << size of one-hot
- Semantic similarity can be mapped to vector metrics, e.g. cosine similarity

$$sim(e_1, e_2) = \frac{e_1 \cdot e_2}{\|e_1\| \|e_2\|}$$

Cosine similarity of one-hot vectors is zero!

Embedding matrix

- We need a simple way to compute the embedding $\boldsymbol{e}_w$ of a word w starting from its one-hot representation $\boldsymbol{I}_w$

$$\boldsymbol{e}_w = \boldsymbol{E}\boldsymbol{I}_w$$

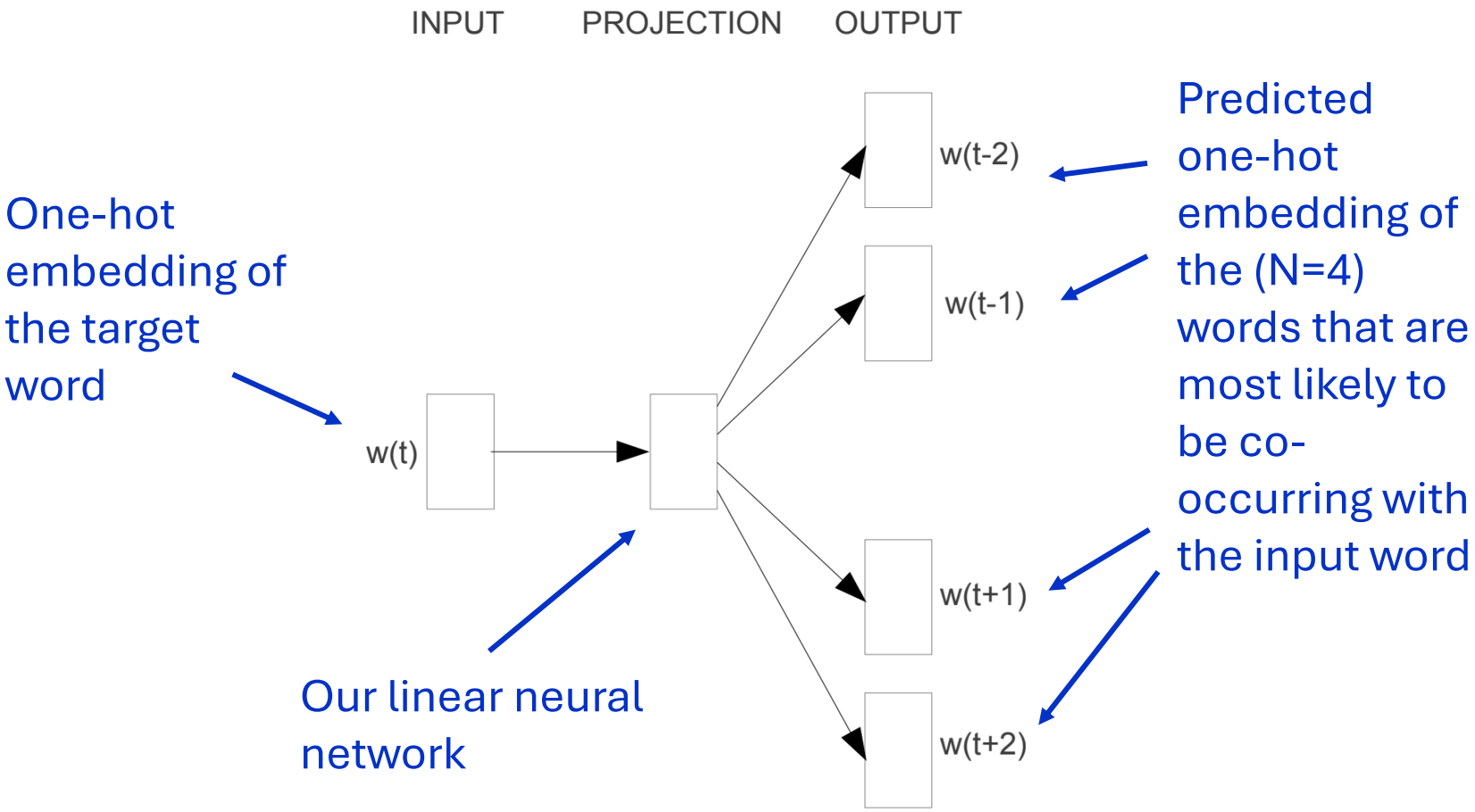
- $\boldsymbol{E}$ is the embedding matrix
- Multiplication selects a specific column from the matrix (corresponding to the target word)

Now we are only missing a way to learn the elements of the embedding matrix in a way consistent with the distributional hypothesis (spoiler: we use neural networks)

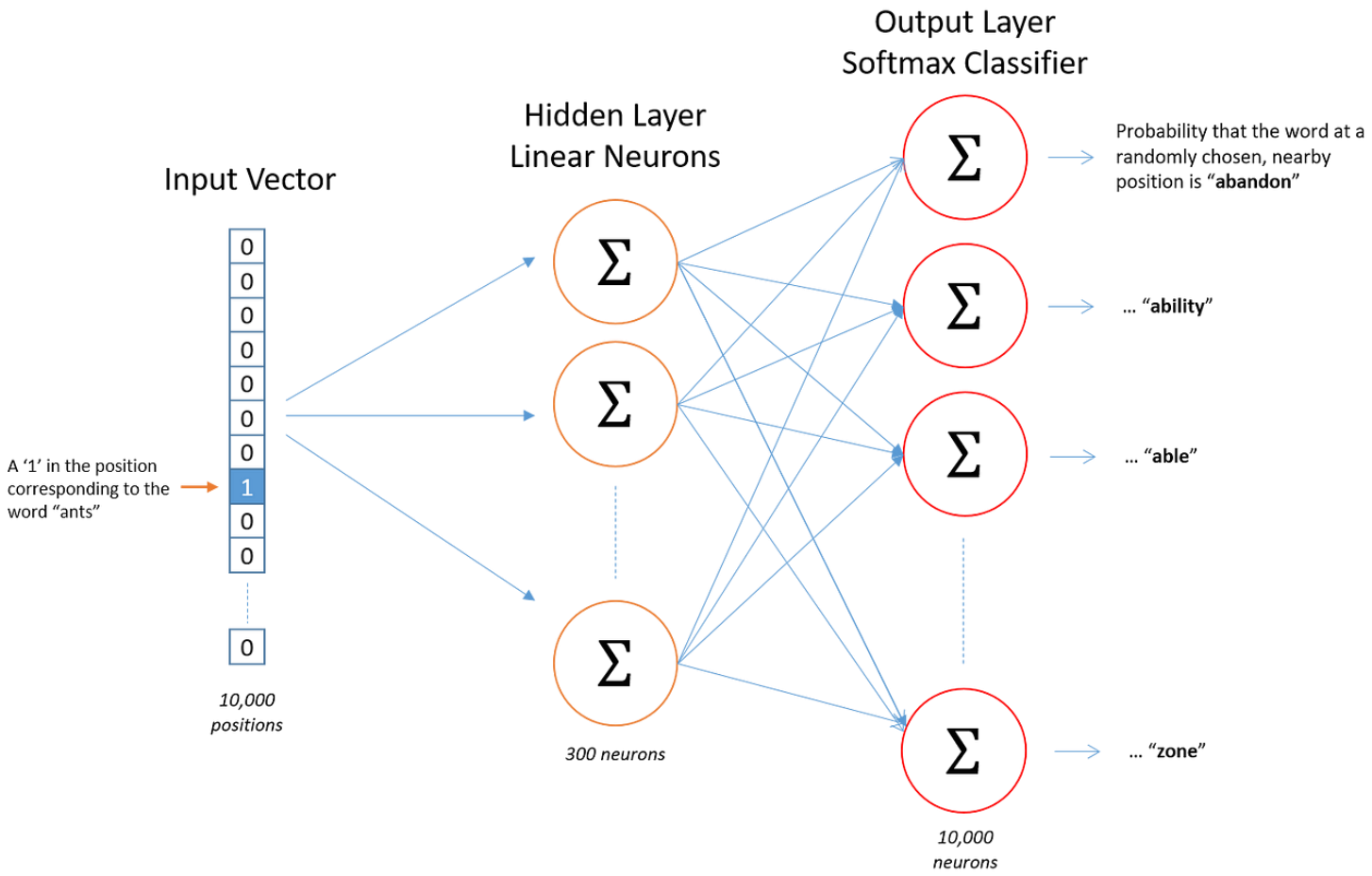
Learning word embeddings

- Take a corpus of sentences in our target language (e.g. download Wikipedia in English) and encode the words with one-hot
- Create a dataset containing pairs consisting of a
  - A word
  - Contextually associated words (e.g. words that occur nearby in the corpus)
- We create a linear neural network with:
  - Input layer: V units (equal to the vocabulary size)
  - Hidden layer: d units (equal to the embedding size)
  - Output layer: V units (equal to the vocabulary size)
- Our embedding matrix **E** is the weight matrix of the hidden layer, with dimensions **d × V**, initialized randomly

Skip-Gram approach



Skip-Gram Network (Explicitly)

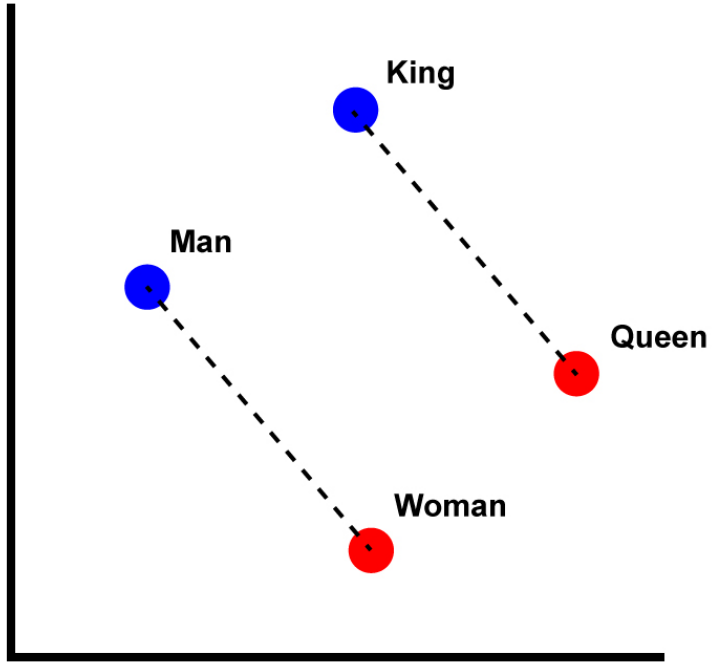


Pretrained work embeddings

- There are many variants of the to the word embedding methodology (Word2Vec) which modify the task, the learning strategy or the pretraining pipeline (e.g. CBOW, GloVe, …)
- Available as open pretrained embeddings learned on very large datasets
- Some have been specialized to the biomedical domain (e.g. from PubMed); they capture out-of-vocabulary words common in clinical language
  - BioWordVec
  - PubMed2Vec
  - BioNLP Embeddings

Embeddings spaces hint at a semantic organization

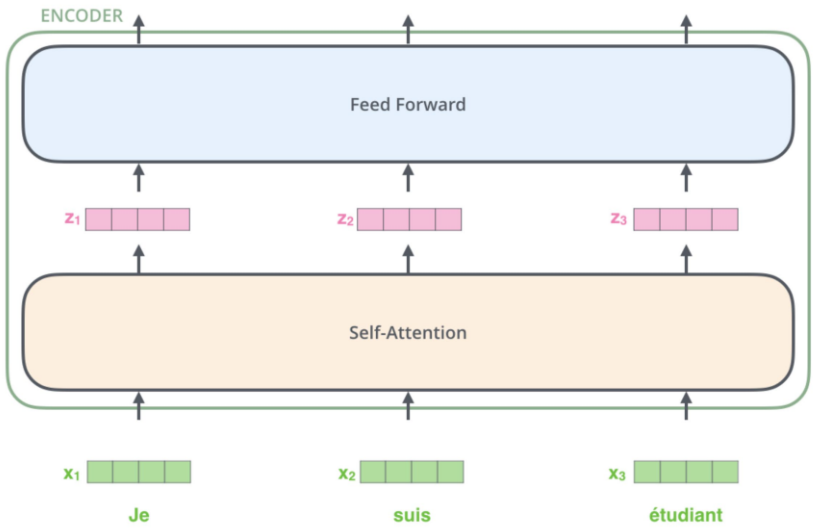
emb(‘*Paris*’) - emb(‘*France*’) + emb(‘*Italy*’) ≈ emb(‘Rome’)



Source: Wikipedia

Using Word2vec embeddings

- We can use Word2vec to learn embedding of textual data other than natural language! E.g. genomic or protein sequences
- Encode words or whole sentences (aggregated embeddings) in input to ML models to solve medical NLP tasks
  - Entity recognition
  - Role labeling
  - Identify relevant/similar document fragments
- Represent textual data in input to RNNs and sequence-to-sequence models
- ..and of course to get in the input embedding to be fed to Transformers!



Language Modelling

Notable language modelling tasks and architectures

- Now we know how to effectively encode data that is originally in textual/symbolic form (word embeddings)
- We also have available a powerful architecture which can use to further refine the word embeddings into richer neural representations capturing contextual information (Transformers)
- We can now turn our attention to how the two can be used to build language models which can be used and re-used across a variety of specialized tasks on sequential data

Language Models – Basic Terminology

- Tokens: atomic (i.e. irreducible) elements in our sequence modelling problem
  - Words of the vocabulary
  - Syllabes or phonemes
  - Amminoacids or k-mers
  - …
- Language model (LM): a probabilistic model of how tokens can be combined to obtain valid and meaningful sequences from our language
  - Trained LMs are often used for tasks other than generating sequences
  - General idea: train the LM to reconstruct the full sequence from a noisy version provided in input

Training LMs

- Typically involves two-phases
  - Pretraining (self-supervised/unsupervised) over a very large and general dataset
  - Fine-tuning (supervised or self-supervised) over a specialized data set for a particular end-task/domain
- Types of end-tasks
  - next-sentence prediction (or completion)
  - translation, summarization or paraphrasing
  - question answering

  ⇒ encoder-decoder
  - language modeling ⇒ decoder-only
  - token/word classification or regression ⇒ encoder-only
  - text/sequence/document classification or regression ⇒ encoder-only
- Different task types are associated with different configurations of the encoder-decoder architecture

Relevant Architectures

- We focus on encoder-only and decoder-only approaches as they are the most commonly used both in general purpose domains as well as for specialized healthcare applications
- Two champions
  - BERT (encoder-only)
  - GPT (decoder-only)

BERT - Bidirectional Encoder Representations from Transformers

Key innovations

- Introduced masked language modelling as a powerful, general and low-annotation effort approach to perform pretraining scaling up size and performance of LMs
- Used bidirectional self-attention layers (the equivalent of BiLSTM) for obtaining more contextualized embeddings
- Enabled the pretrain-once and fine-tune-as-needed scheme
  - Inspired a whole lot of follow-ups *BERTs (RoBERTa, ALBERT, BART, …)

Masked Language Modelling

- A fill-in-the-gap training task
- You are all perfectly able to complete the following sentence:

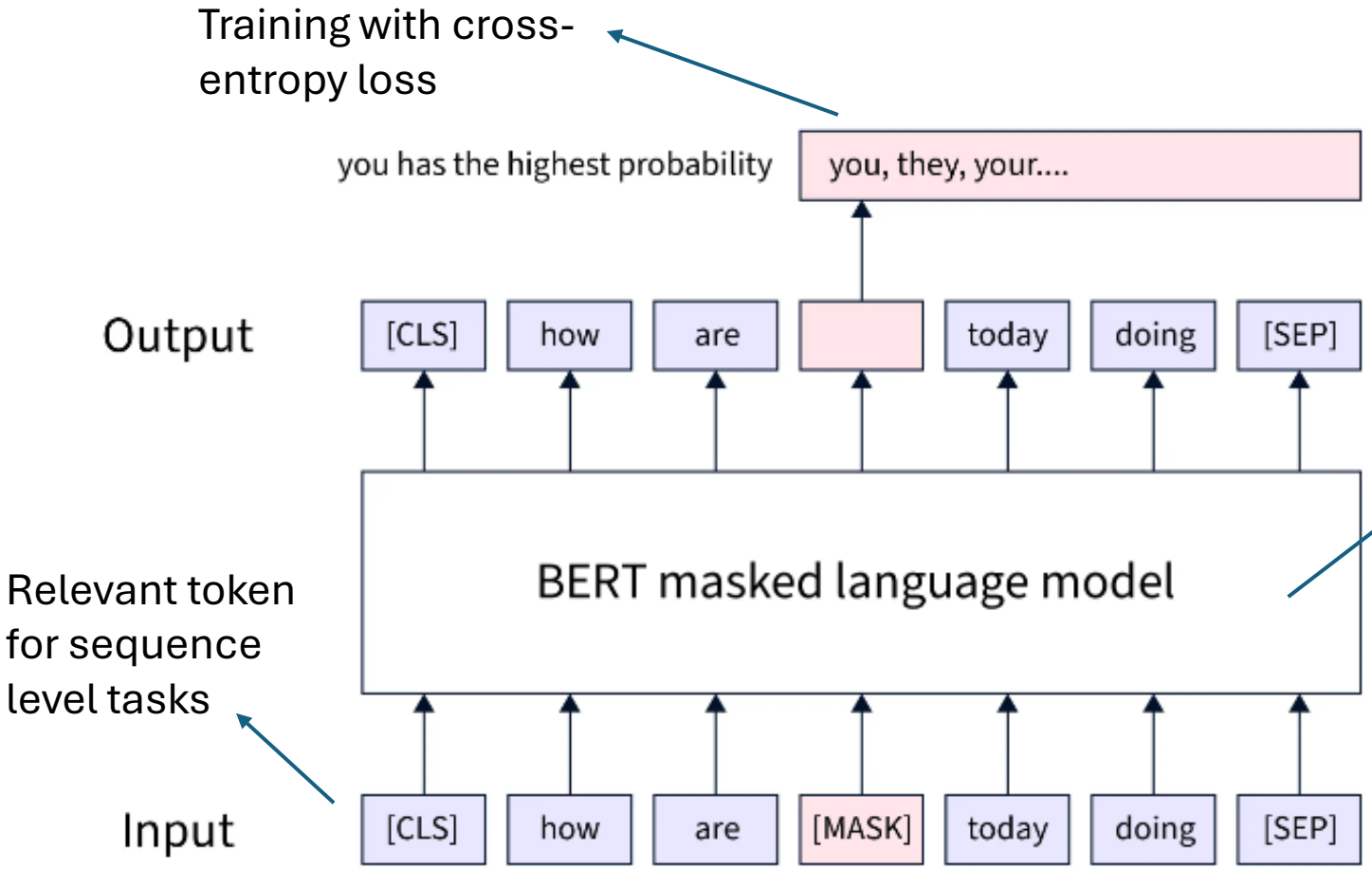
How are _____ doing?

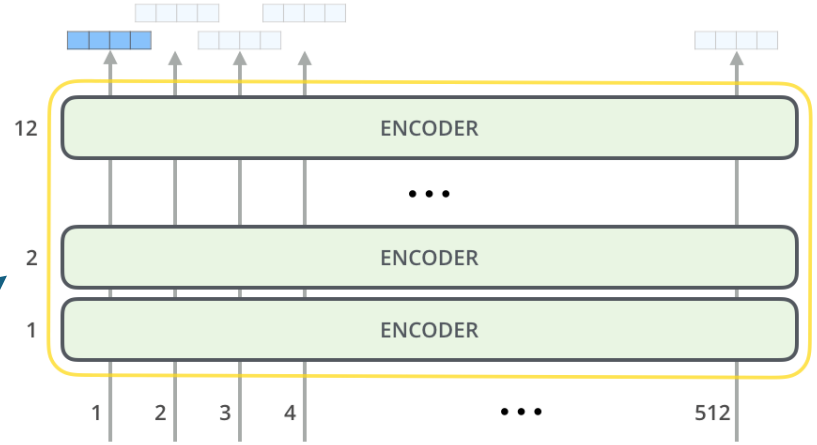
- Because you can infer the missing piece from the context
- If we train the model to solve this task it will become implicitly good at assigning the right meaning to the tokens depending on the context
- When training we will signal the model that the true token is missing by using a dedicated token

How are [MASK] doing?

BERT Pretraining

In fairness BERT is also pretrained on a next sentence binary classification task

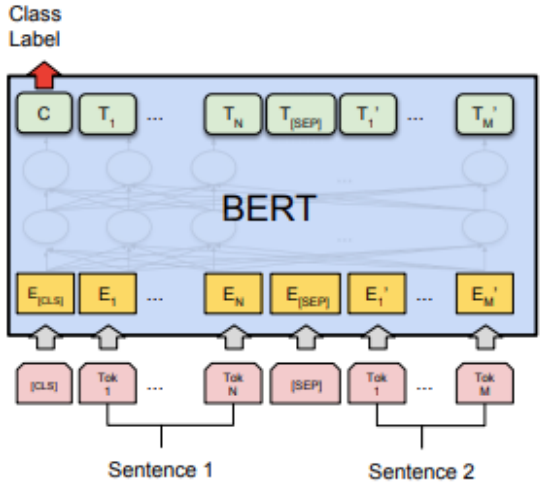




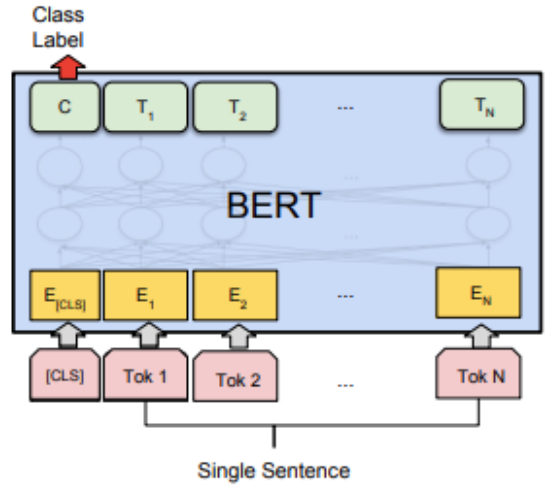
A transformer with 12 (base) to 24 (large) self-attention layers pretrained on Wikipedia and Books corpus

BERT Fine-tuning

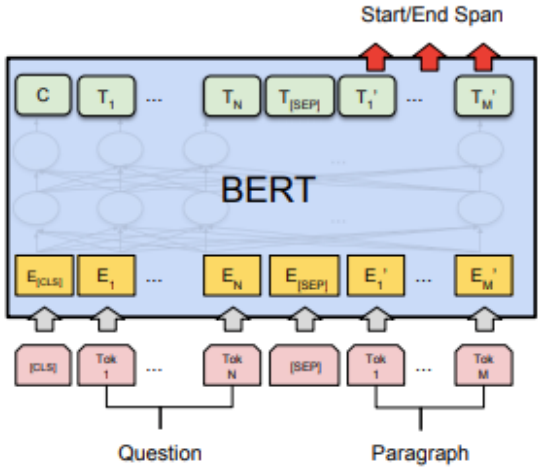
- Download your pretrained BERT of choice (e.g. from HuggingFace)
- Prepare your data with the appropriate tokenizer
- Train the output layer to solve your task, feeding it with the right embedding
  - [CLS] for sentence level tasks
  - Each data token for single element tasks
  - ...



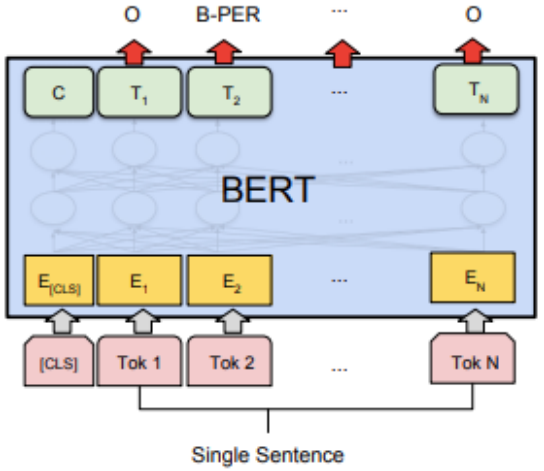
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



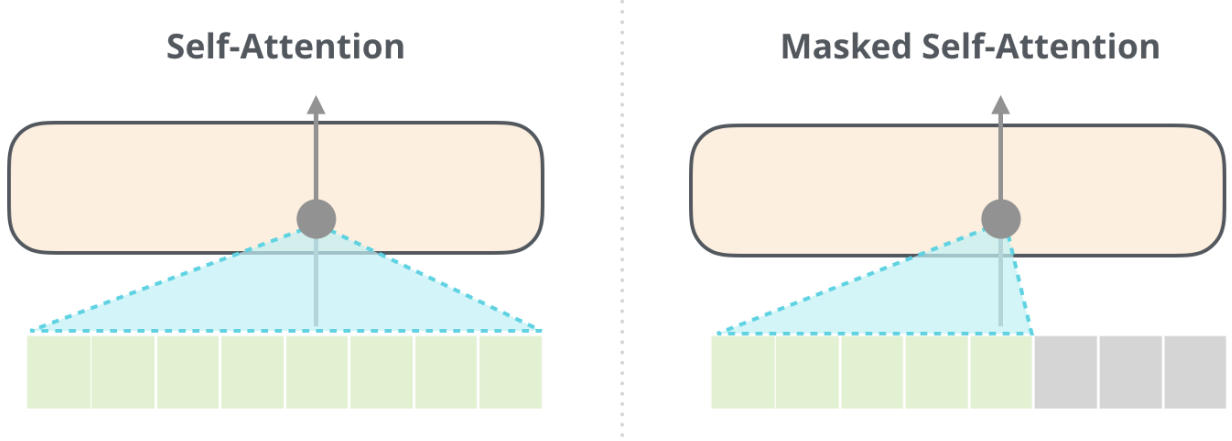
(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

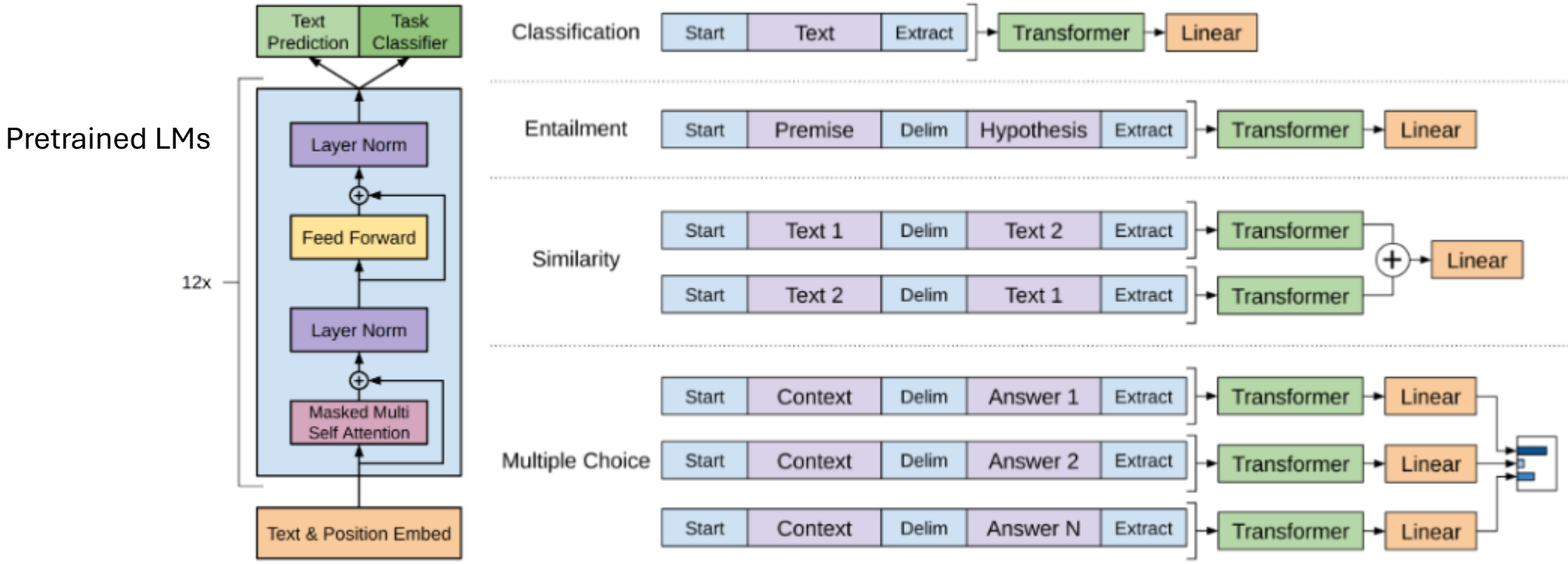
GPT - Generative Pretrained Transformer

- A family of decoder-only models heavily popularized by OpenAI
- From GPT-3 initiated the family of Large Language Models (LLMs)
  - Huge datasets, billions of tokens and parameters
- GPT is trained on a *simpler* fill-in-the-gap task than BERT as it focuses on the next-token prediction
- Uses masked self-attention to avoid cheating by looking into the future

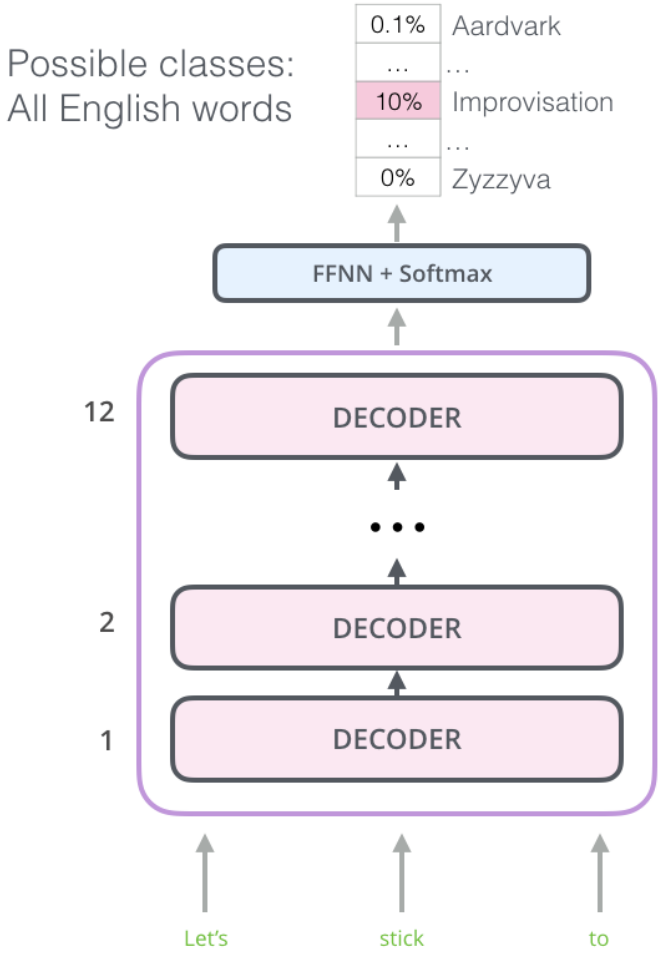


GPT-1 Mixed LM and Supervised Pretraining

Convert all structured inputs into token sequences to be processed by pre-trained LMs, followed by a linear+softmax layer



General GPT Training



- GPT uses self-supervised learning alone to train the language model (again cross-entropy minimization)
- There is no separation of pre-training and fine-tuning like BERT
- It is a naturally multi-task model where tasks are defined through structured token templates

GPT Tasks

- GPT is naturally designed for solving tasks via generation
  - Sequence/sentence completion or generation
  - Sequence classification/regression as generation
  - Summarization (need adequate training with dedicated tokens)
  - Translation (need adequate training with dedicated tokens)
- But its embeddings work surprisingly well also for sentence/token level predictive tasks
- Shows evidence of good zero-shot learning and few-shot learning ability

Processing textual data in healthcare

https://github.com/allenai/scispacy

spaCy Tools

Industry-grade NLP pipelines with customized models for scientific docs (Sci spaCy)

Entity mentions are detected and classified in the Unified Medical Language System (UMLS)

Spinal ENTITY and bulbar muscular atrophy ENTITY (SBMA ENTITY) is an inherited ENTITY motor neuron disease ENTITY caused by the expansion ENTITY of a polyglutamine tract ENTITY within the androgen receptor ENTITY (AR ENTITY). SBMA ENTITY can be caused by this easily.

	text	Canonical Name	Concept ID	TUI(s)	Score	start	end
0	Spinal	spinal	C0521329	T082	1	0	1
1	bulbar muscular atrophy	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	0.909614	2	5
2	SBMA	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	1	6	7
3	inherited	Hereditary	C0439660	T169	1	10	11
4	motor neuron disease	Motor Neuron Disease	C0085084	T047	1	11	14
5	expansion	cell growth	C0007595	T043	0.864297	17	18
6	androgen receptor	AR gene	C1367578	T028	1	24	26
7	AR	AR gene	C1367578	T028	1	27	28
8	SBMA	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	1	30	31

https://github.com/allenai/scispacy

Specialized Entity Recognition

Specialized Entity Recognition models trained on specific Named Entities and used annotate tokens (and spans of tokens) in the text

Spinal and bulbar muscular atrophy **DISEASE** (SBMA **DISEASE**) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor (AR). SBMA **DISEASE** can be caused by this easily.

Recognizes diseases and chemical compounds

Spinal and bulbar muscular atrophy (SBMA) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor **PROTEIN** (AR **PROTEIN**). SBMA can be caused by this easily.

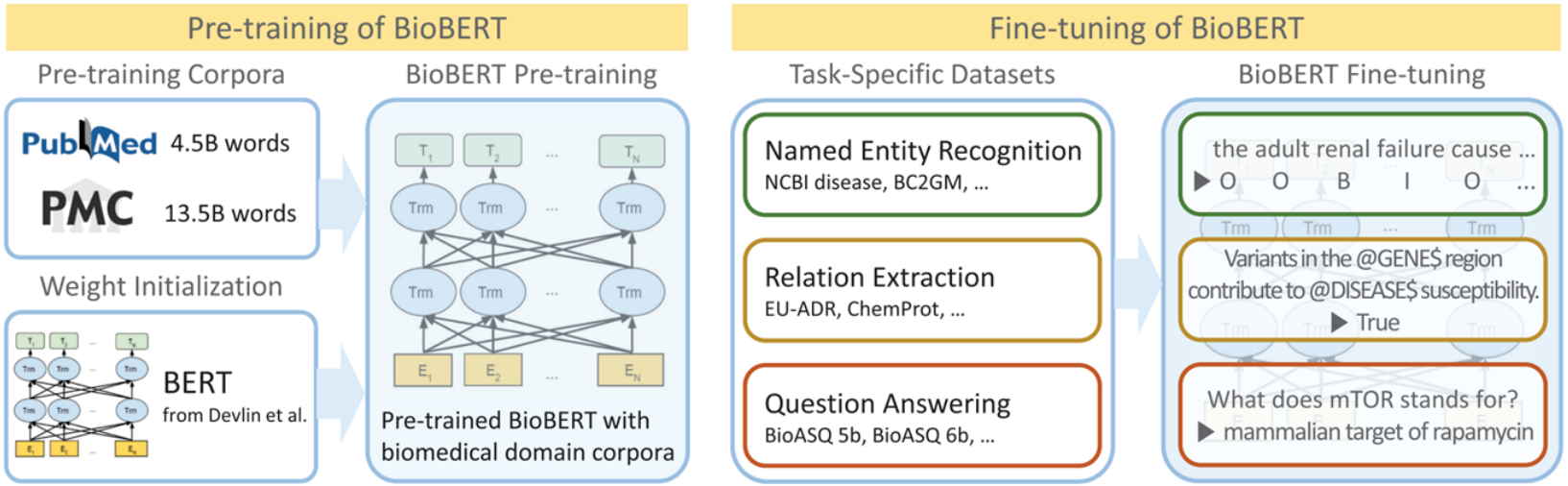
Specialized on proteins, cells and DNA entitiies

Specialized Language Models

- Starting from the architectures introduced so far, specialized language models can be created by
  - Pretraining through self-supervision on healthcare corpora
  - Pretraining through multi-task learning on healthcare corpora
  - Fine-tuning on specialized healthcare tasks
- Often specialized models are not pretrained from scratch, rather we use pretrained generalist models

BioBERT

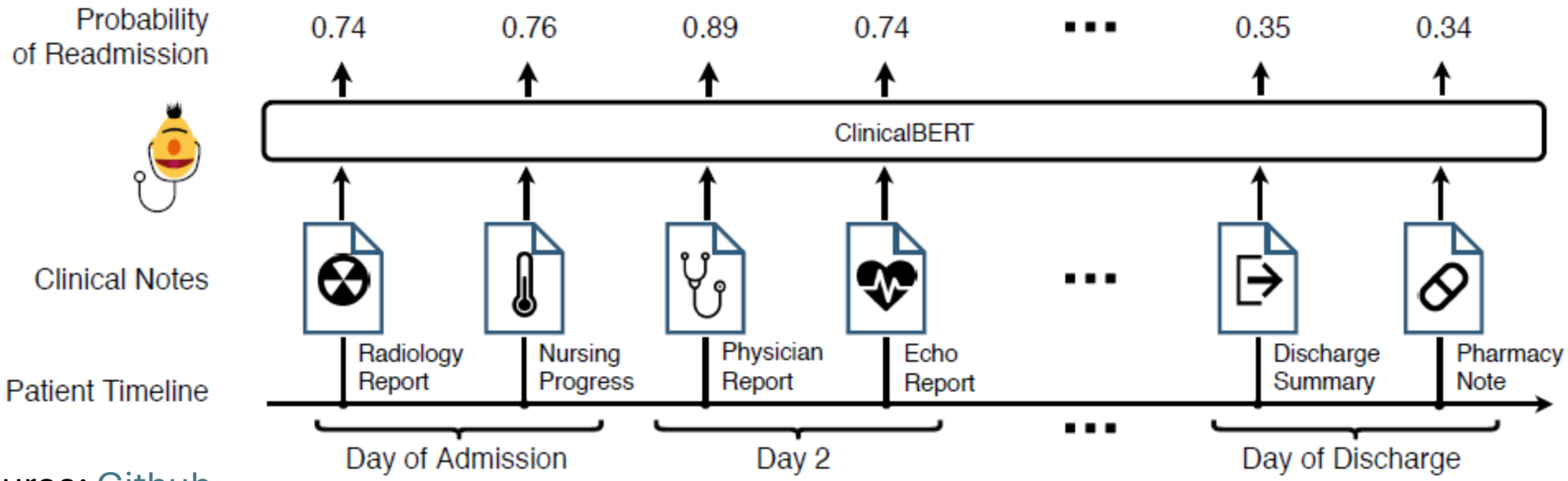
- Initialized with weights from general domain BERT
- Self-supervised pretraining on biomedical domain corpora (PubMed abstracts and PMC fulltext articles)
- Fine-tuned and evaluated on three popular biomedical text mining tasks (Entity Recognition, Relation Extraction, Question Answering)



Lee et al., 2019

ClinicalBERT

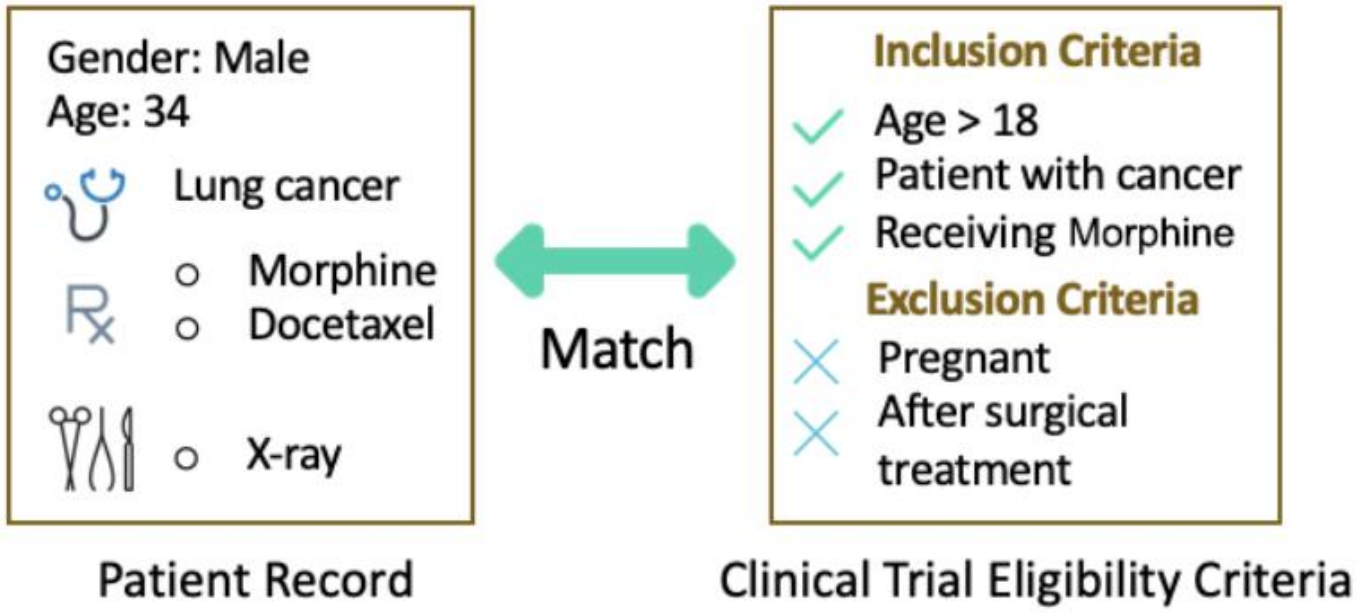
- Pretrained by self-supervision on clinical notes from MIMIC-III EHR dataset
- Goal: improve the prediction of hospital readmission within 30 days based on the information contained in clinical notes



Source: Github

COMPOSE: Find patients for clinical trials from EHR data

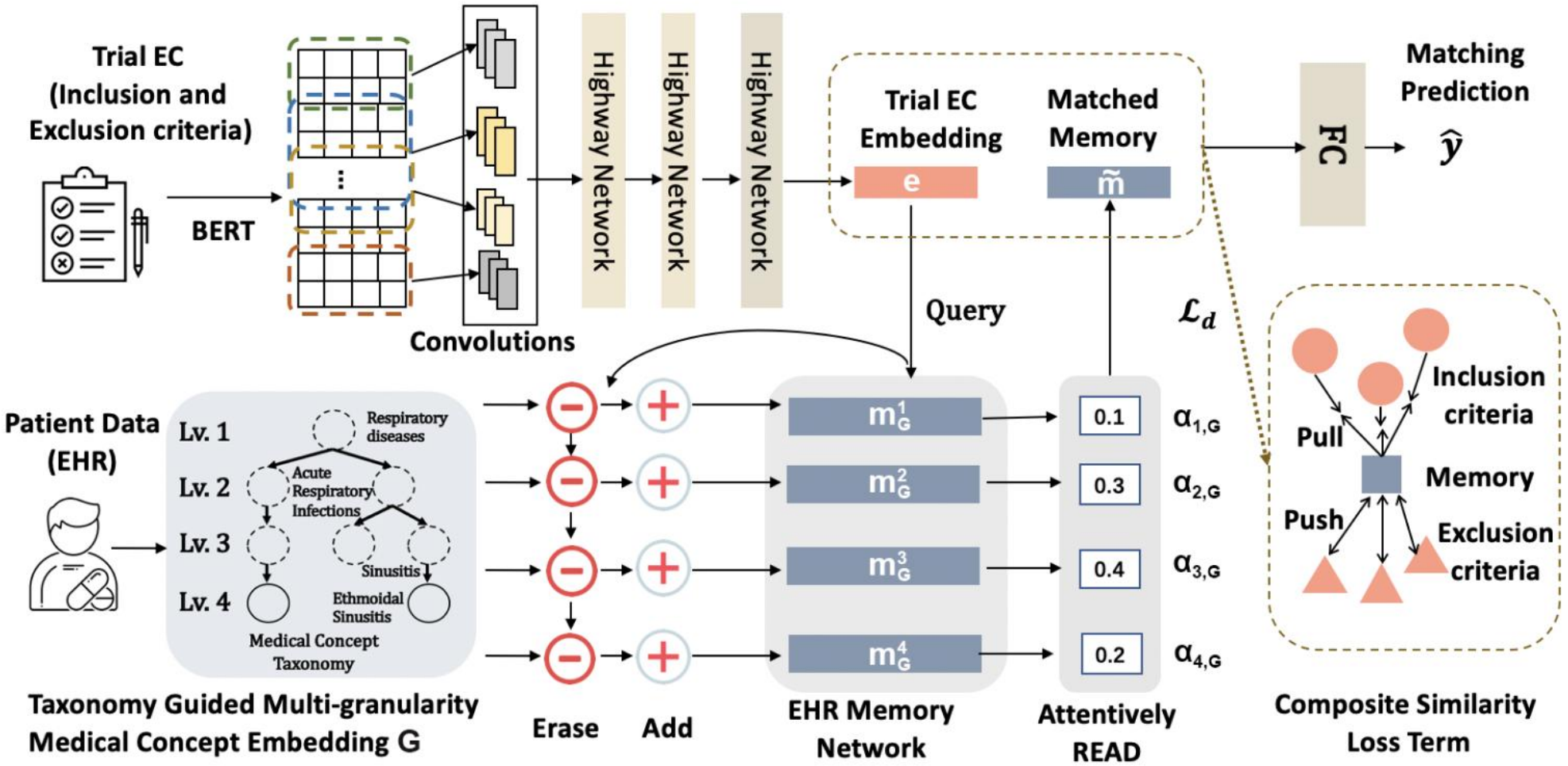
- The problem: patient trial matching
- The approach: system that integrates ClinicalBert embeddings and taxonomy-guided embeddings for patient-criteria matching



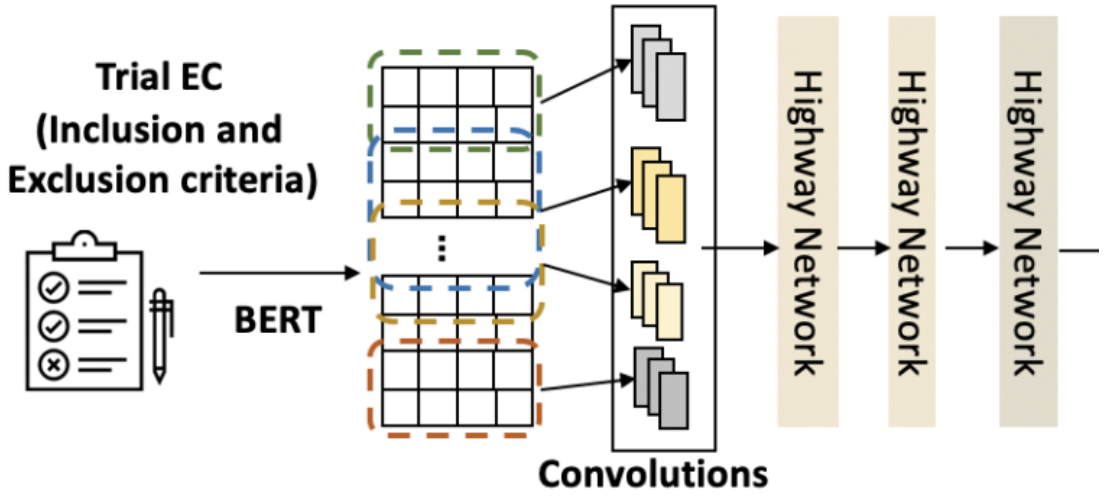
Gao et al., KDD 2020

COMPOSE Model

Gao et al., KDD 2020



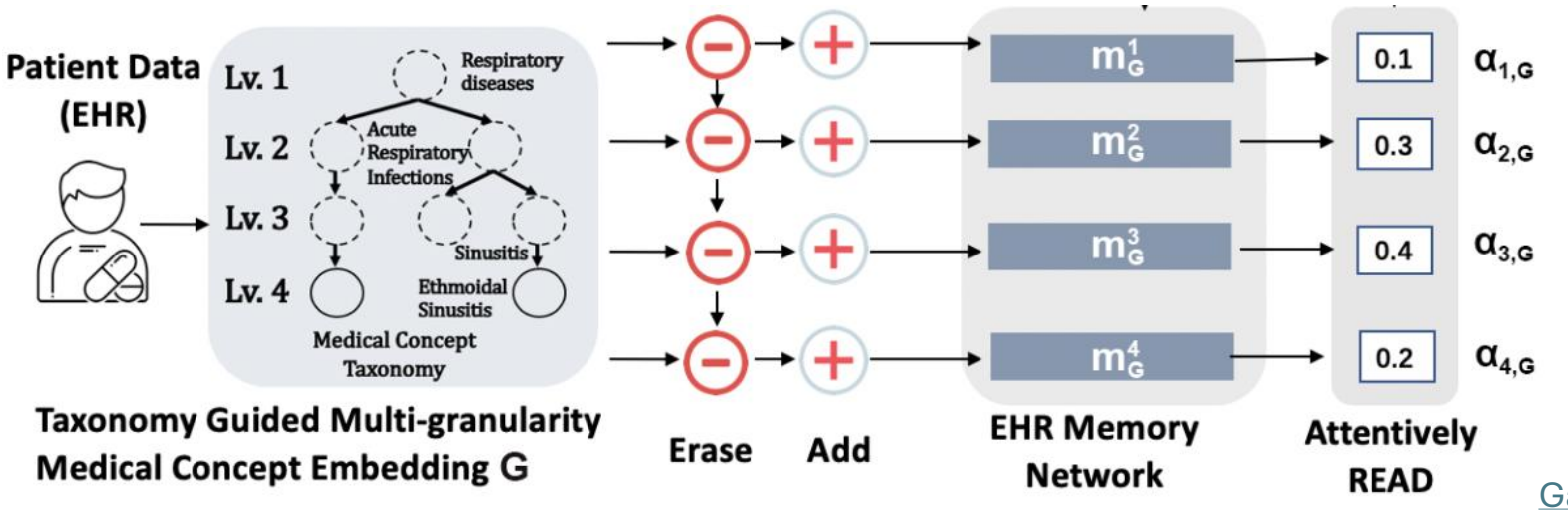
COMPOSE – Criteria Handling



Gao et al., KDD 2020

- ClinicalBert to learn contextual embeddings for the exclusion/inclusion criteria
- Convolutional filters of different size to capture semantics at different levels of granularity
- Highway Networks (structured gated networks, similar to residual layers) and max pooling layers to obtain final embedding

COMPOSE – Patient Embeddings



Gao et al., KDD 2020

- Patient as a sequence of visits, each represented by a diagnosis, procedure and medication
- Use disease, procedure and medication taxonomies (USC) to map EHR concepts onto four levels of abstraction
- Store diagnosis, medications and procedures into a neural memory (attention-based model) organized in the four levels of the taxonomy
- Solves patient-trial matching as a query-matching problem with the items in memory

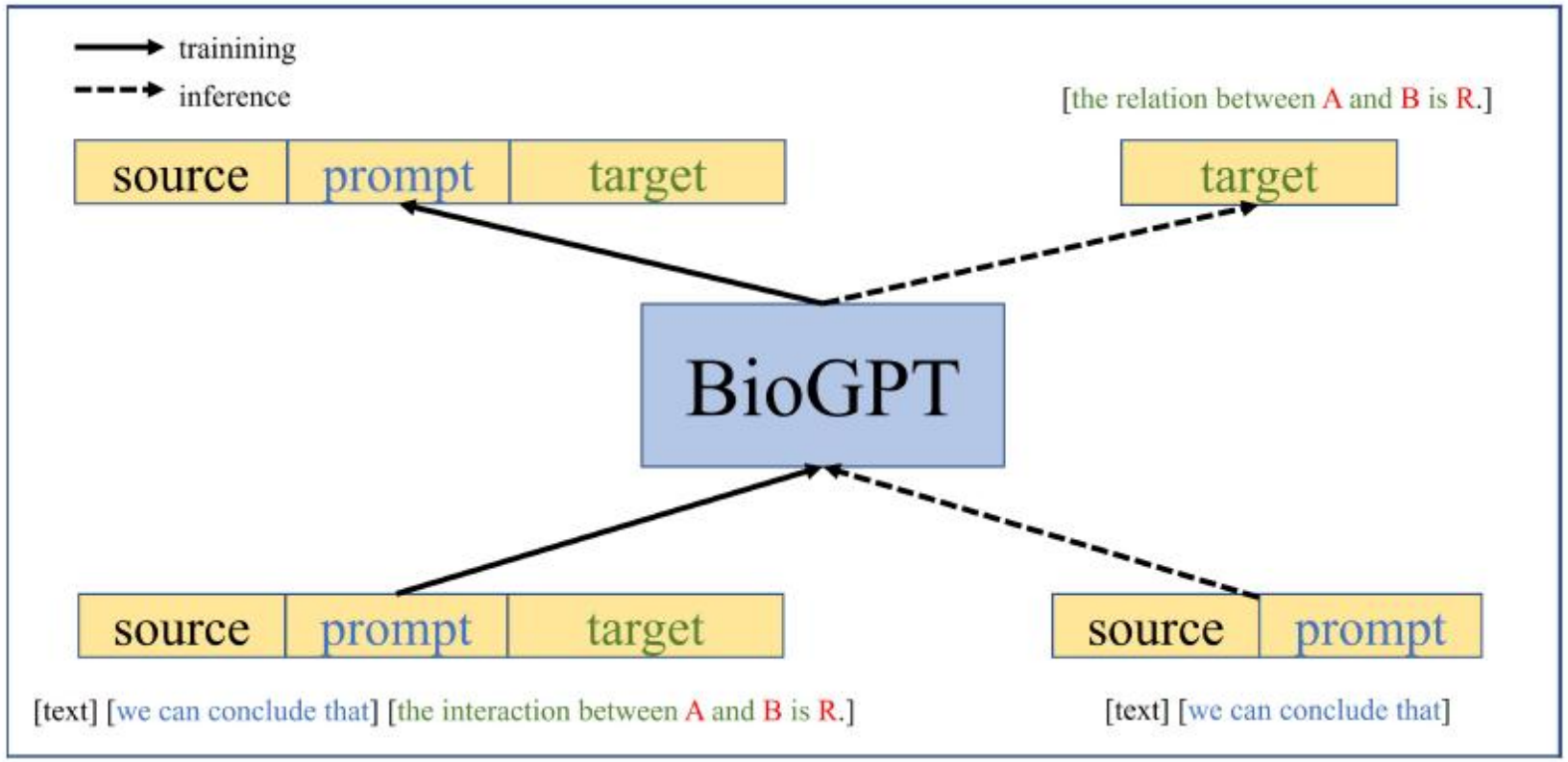
BioGPT

Multi-task fine-tuning on several downstream tasks, including entity recognition, question answering, relation extraction, and document classification

- Hard prompts - Manually designed sentence templates that are prepended to the input text to guide the language model towards a specific task (instruction fine tuning)
- Soft prompts - Continuous embeddings learned during the fine-tuning process

Since GPT-3 tasks are implemented almost uniquely by adjusting prompts

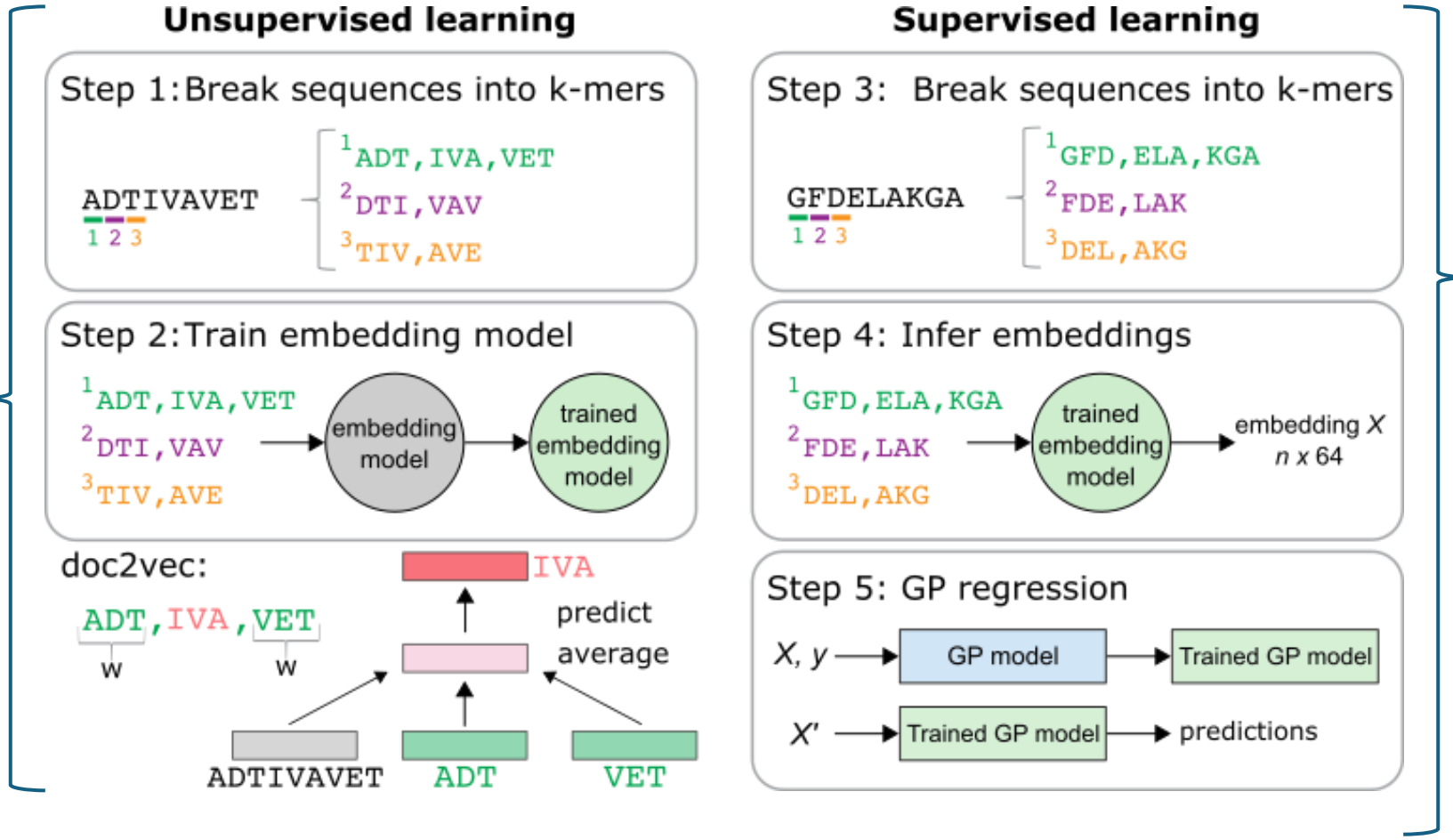
BioGPT



"We have that [head entity] [relation] [tail entity]," "In conclusion, [head entity] [relation] [tail entity]," and "We can conclude that [head entity] [relation] [tail entity]."

Beyond Natural Language - Proteins

Learning embedding for proteins vocabulary (k-mers)

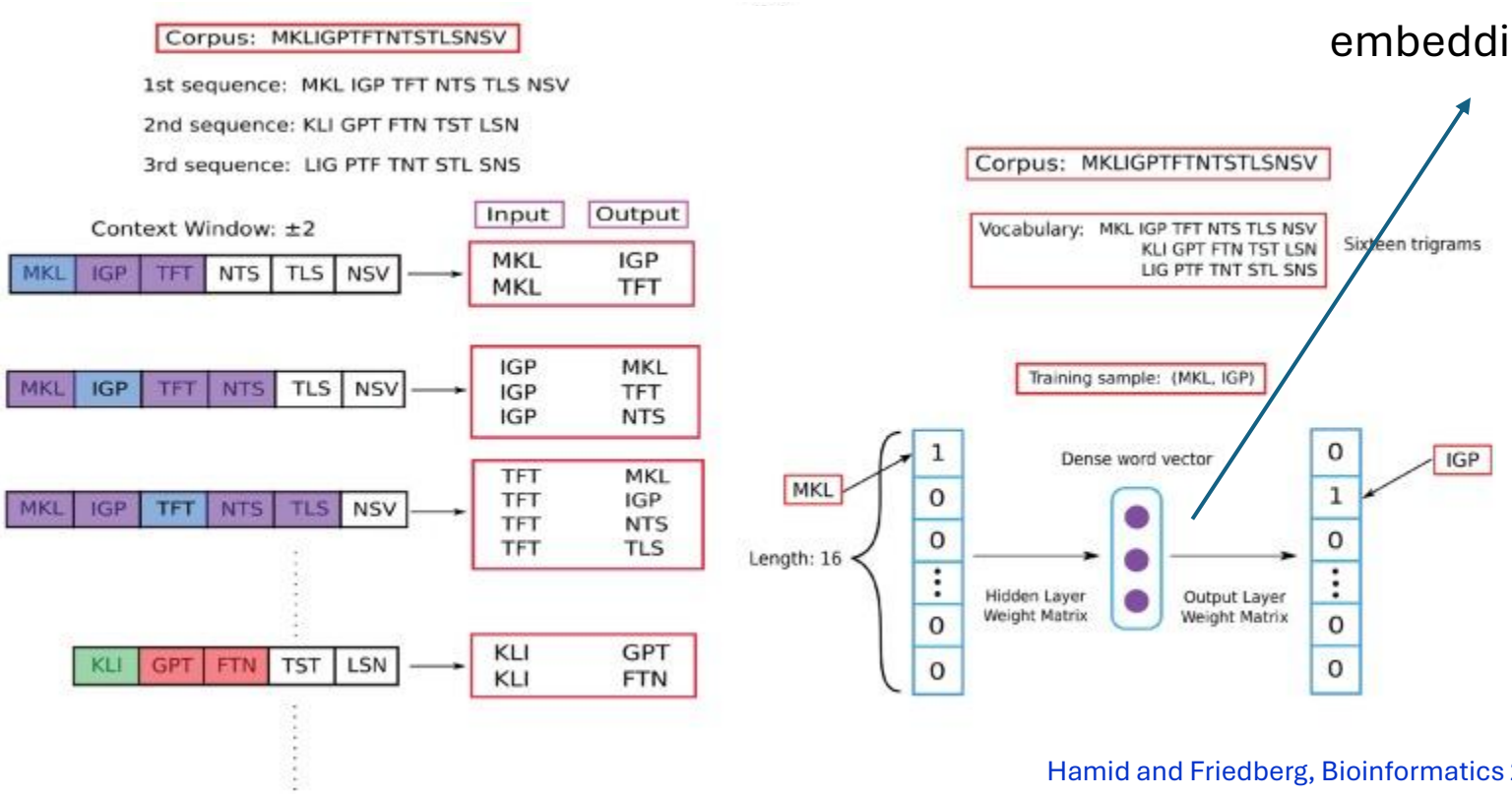


Use k-mers embeddings to predict protein properties

Source: Github

Skipgram protein representation in bacterial toxins identification

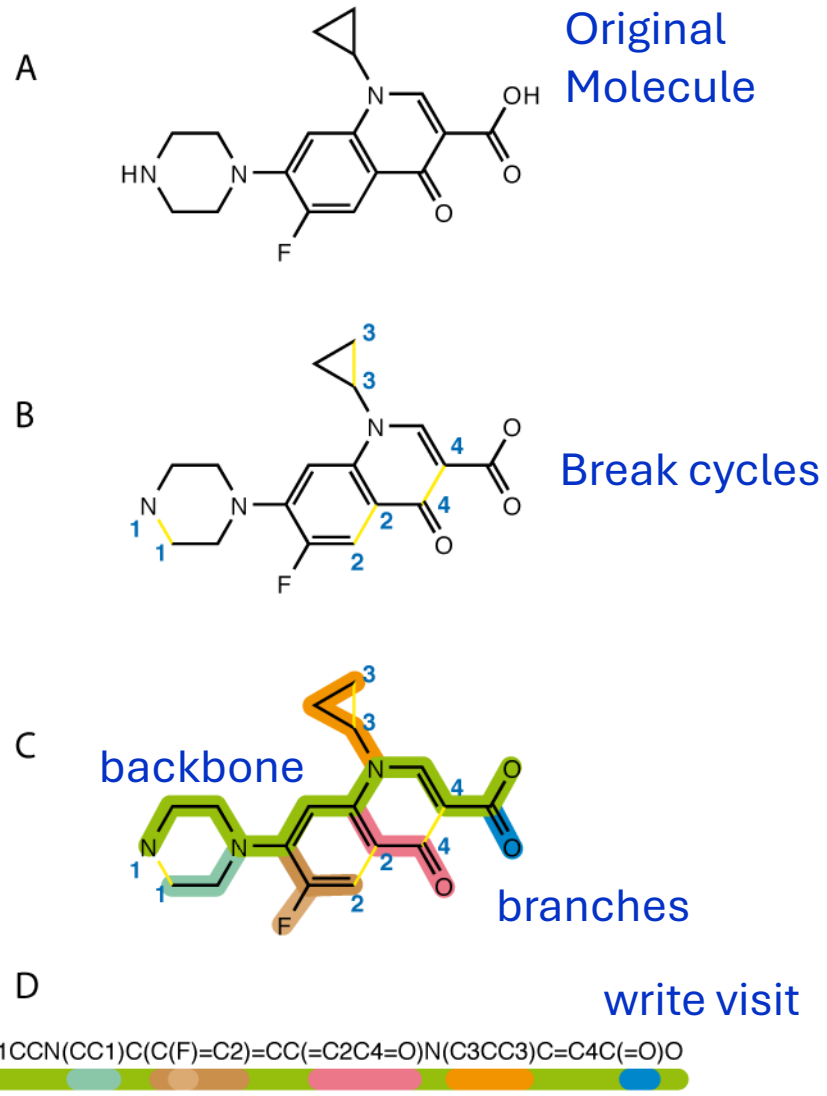
Classification performed by an RNN on the dense skip-gram embeddings of 3-mers



Hamid and Friedberg, Bioinformatics 2018

SMILES-BERT

- Molecules represented as SMILES sequences

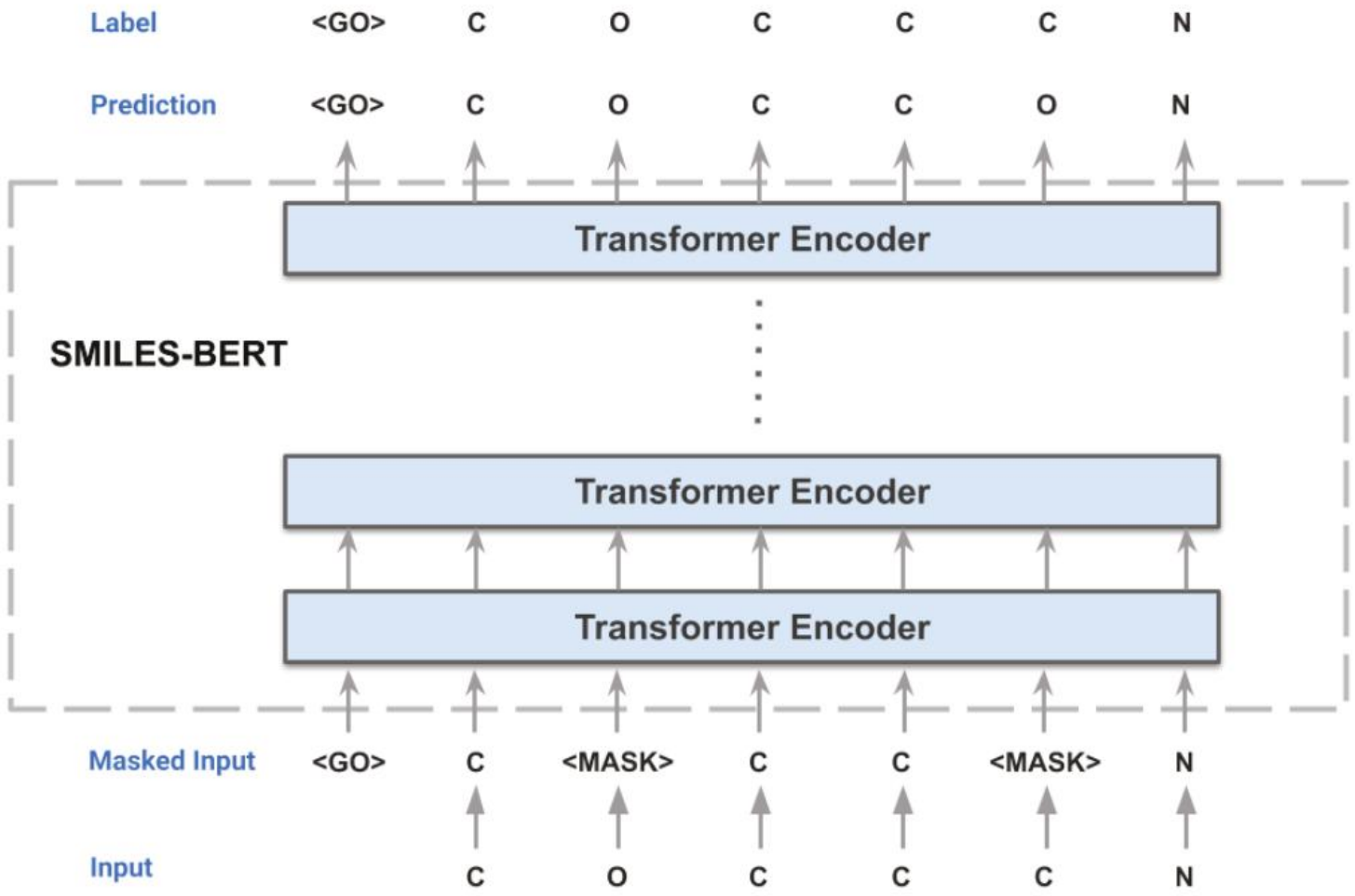


Source: Wikipedia

SMILES-BERT

- Molecules represented as SMILES sequences
- BERT-like trained on masked molecular language modelling
- Fine-tuned on molecular property prediction tasks

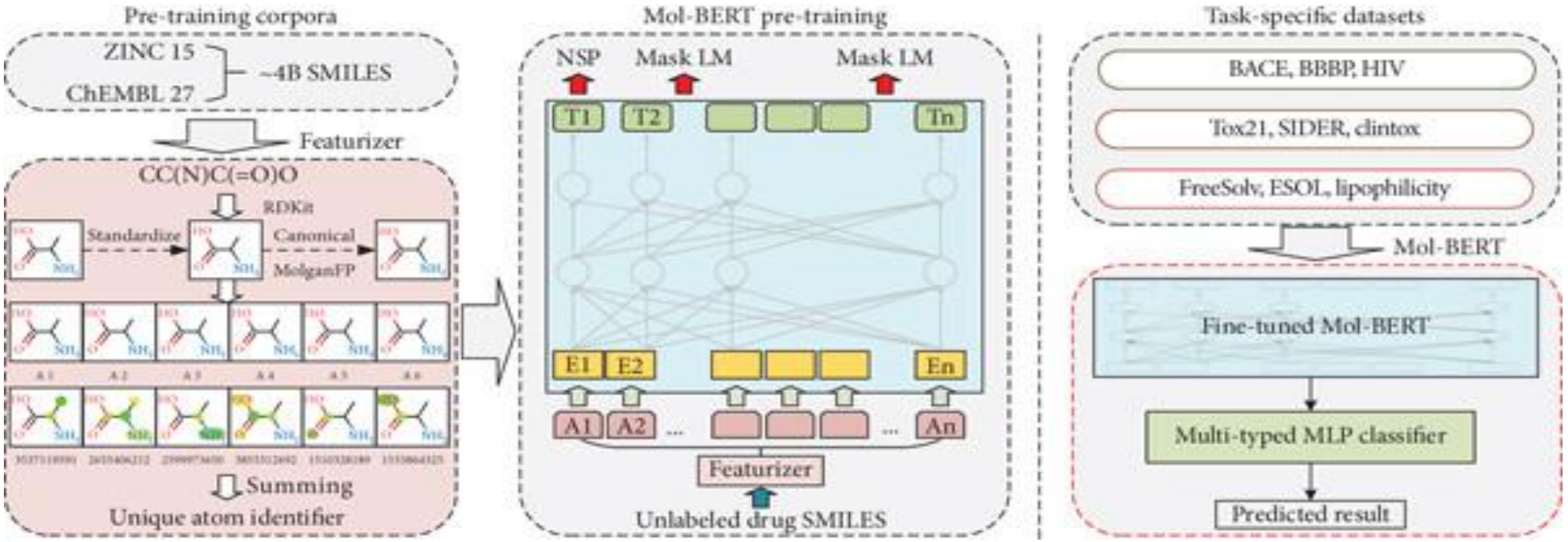
Wang et al, ACM-BCB 2019



Wang et al, ACM-BCB 2019

MOL-BERT

A natural language processing approach to molecular strings



Extracts molecular substructures (fragments) made of multiple atoms

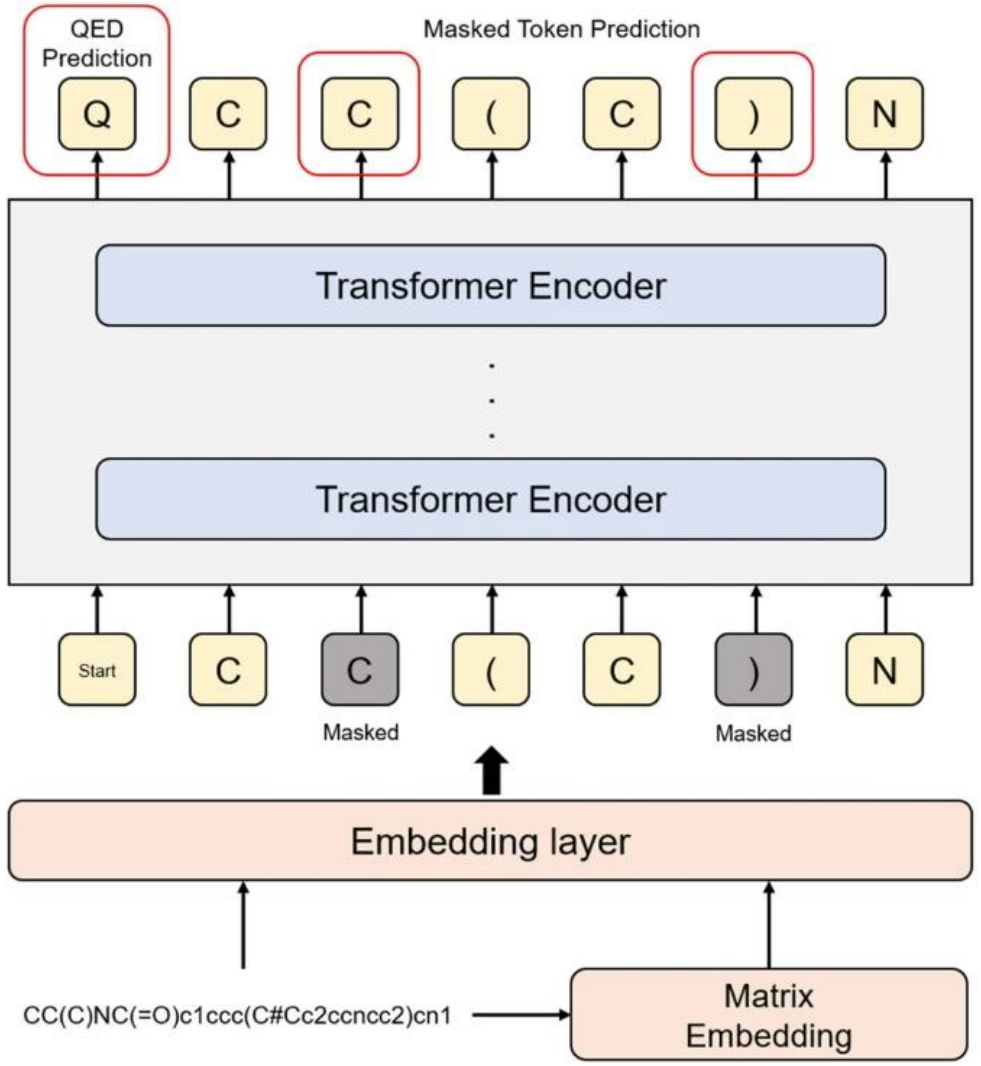
Train a language of molecular fragments (words)

Fine-tuning on classification and regression tasks

Chem-BERT

Extended pretraining that complements language modelling on SMILES with

- Input information on structure-informed embedding of molecule (similar to positional encoding
- Predicting drug-likeness in output to best inform the molecular embeddings

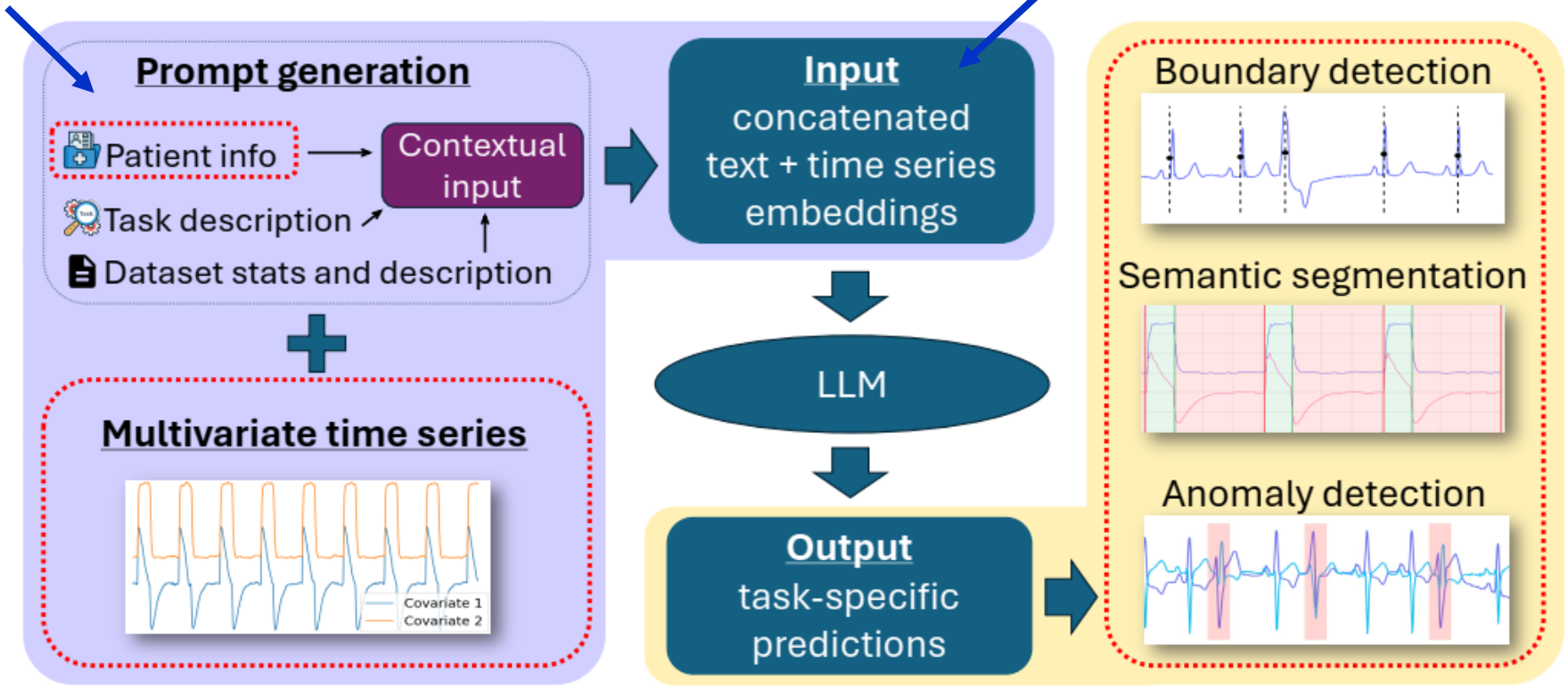


Kim et al, Nat. Sci. Rep 2021

Handling medical timeseries with LMs

Multimodal query packaging multimodal info in LM prompt

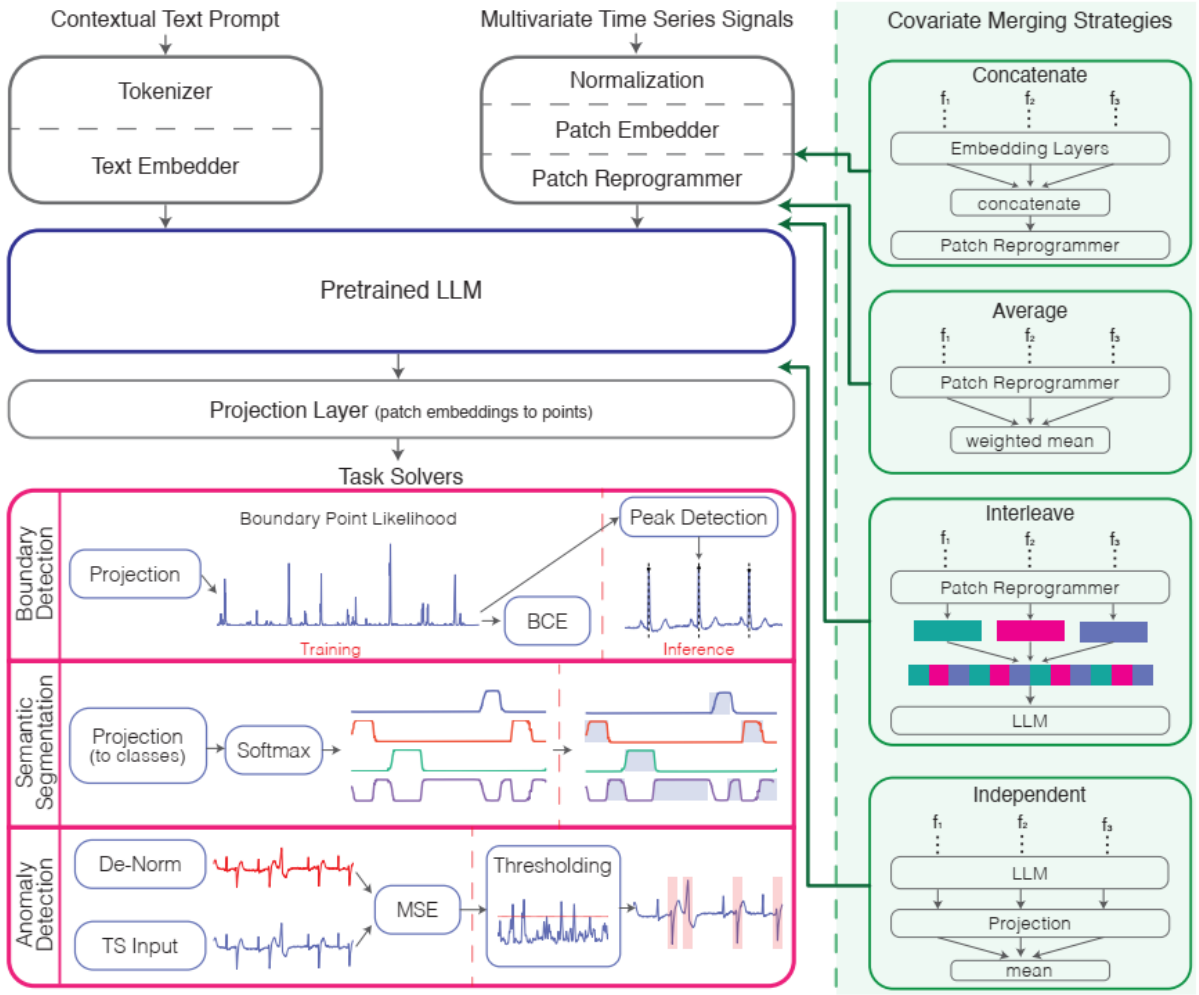
Embedding of context (all treated as text) and time series input



Solving task on respiration and ECG timeseries

Chan et al, MLHC 2024

Proper tokenization of timeseries is key

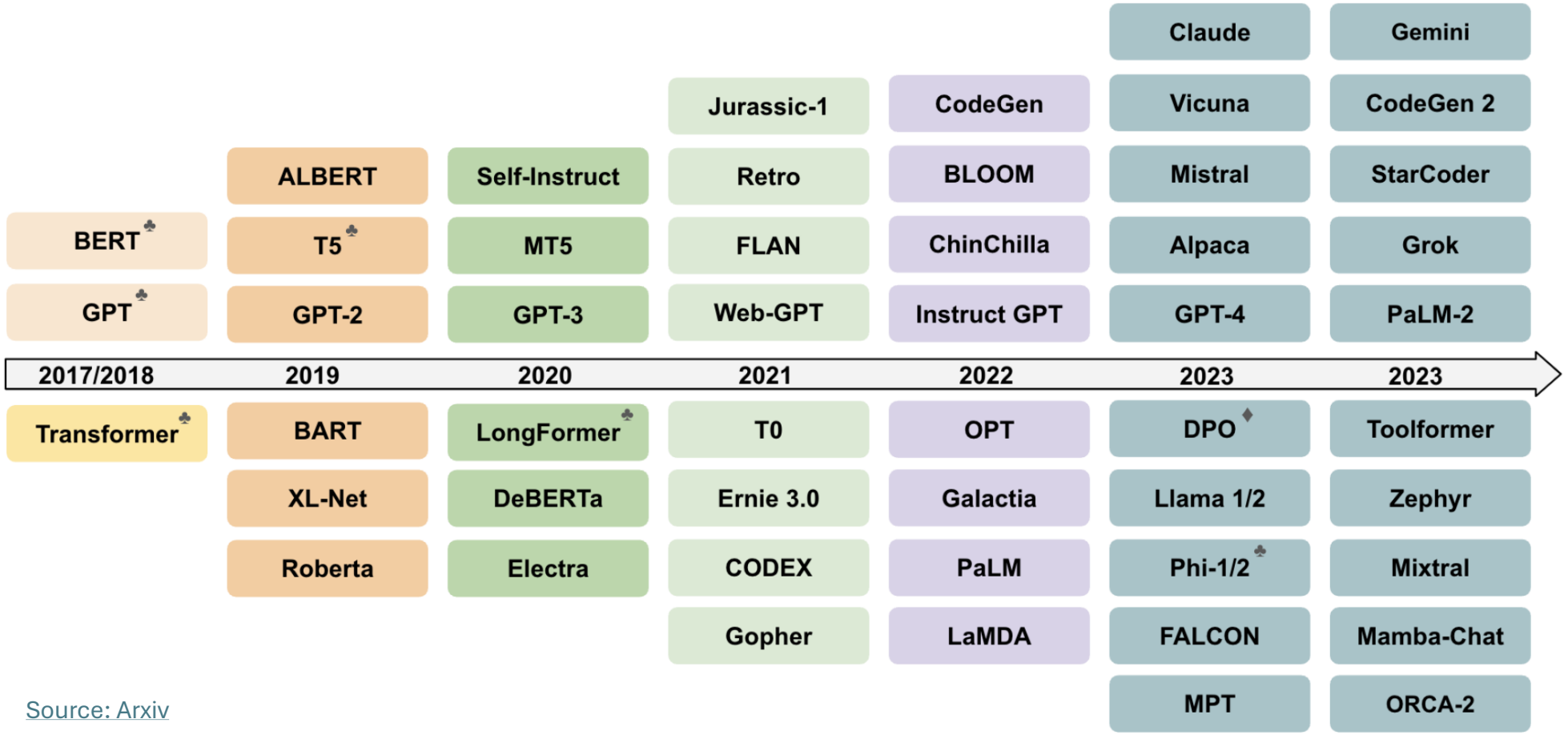


Chan et al, MLHC 2024

Time series tokenized into patches that are aligned with the tokens of the pretrained LLM via a cross-attention mechanism

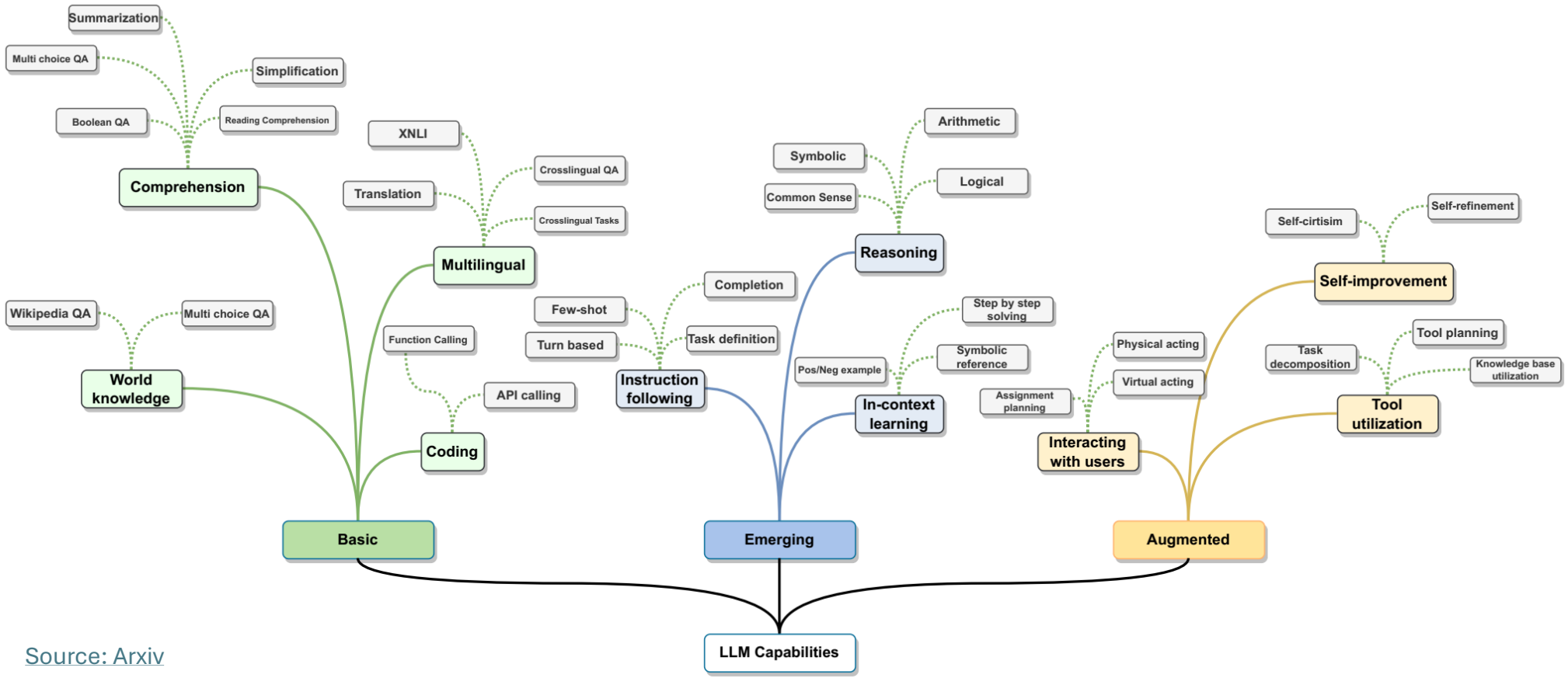
Towards General and Reusable Models

The (Large) LM Landscape



Source: Arxiv

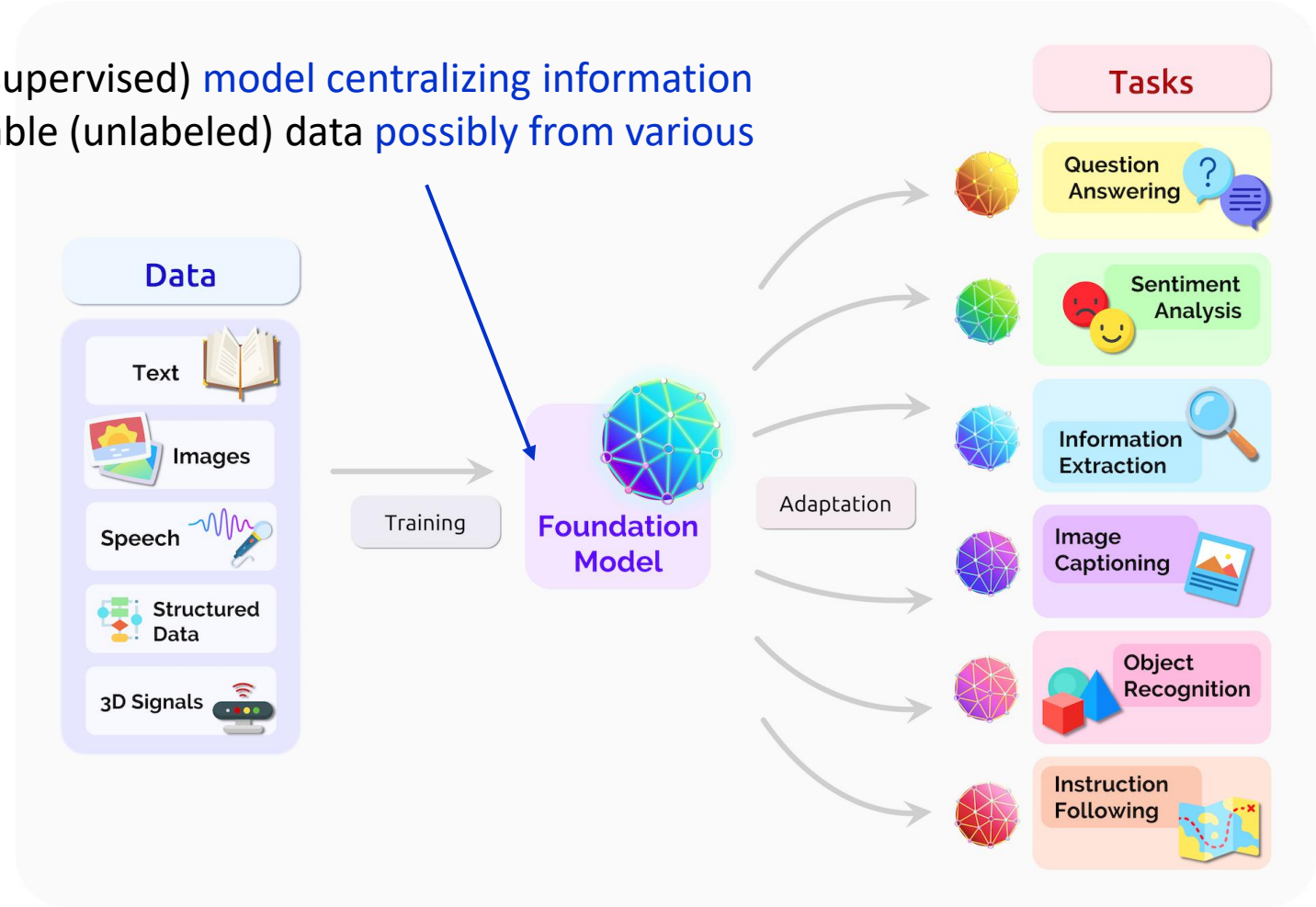
Useful LLM Capabilities



Source: Arxiv

Foundation Models

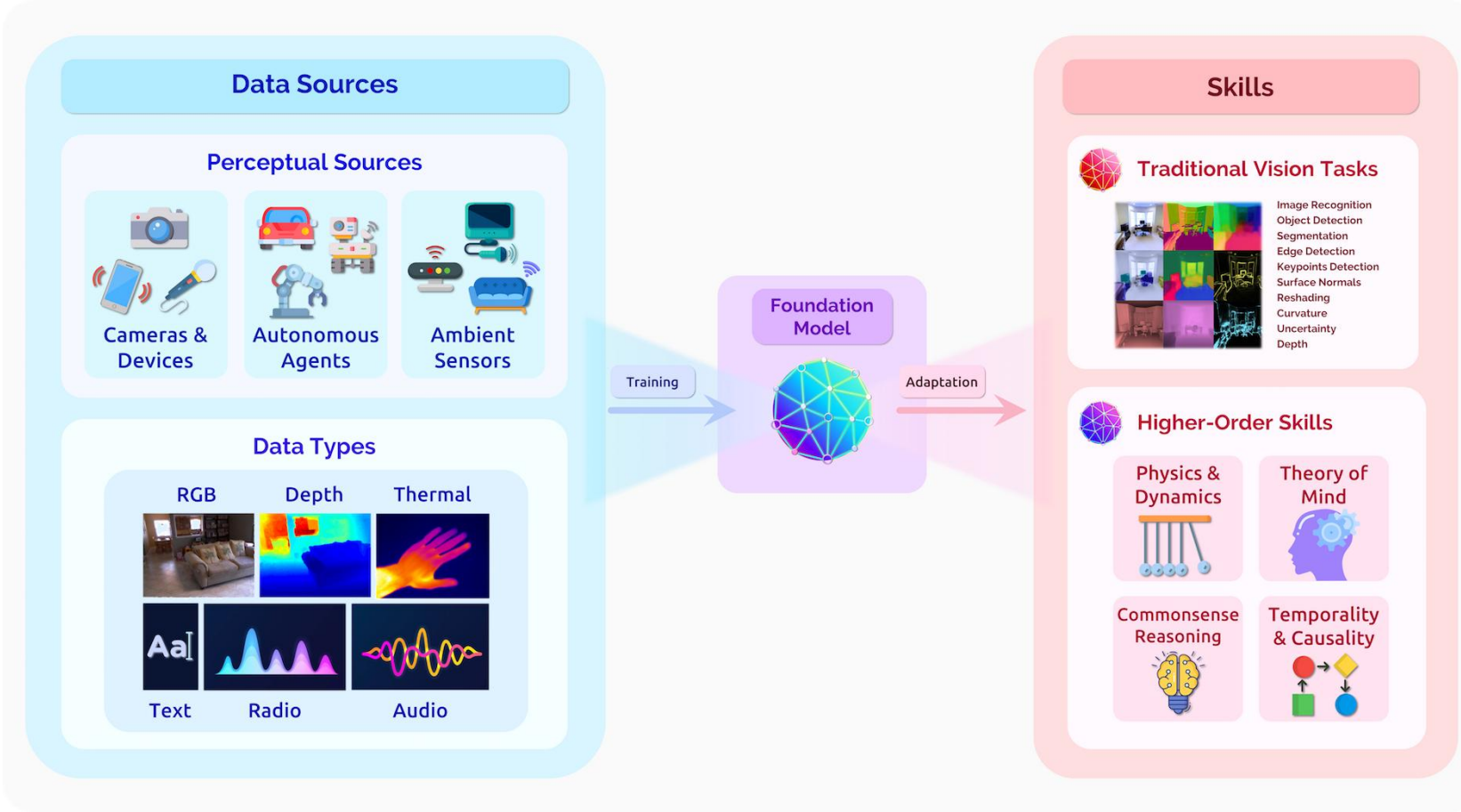
Learn a (self-supervised) model centralizing information from all available (unlabeled) data possibly from various modalities



Adapt the general model to a wide range of downstream tasks (using fewer labelled data) in specific domains

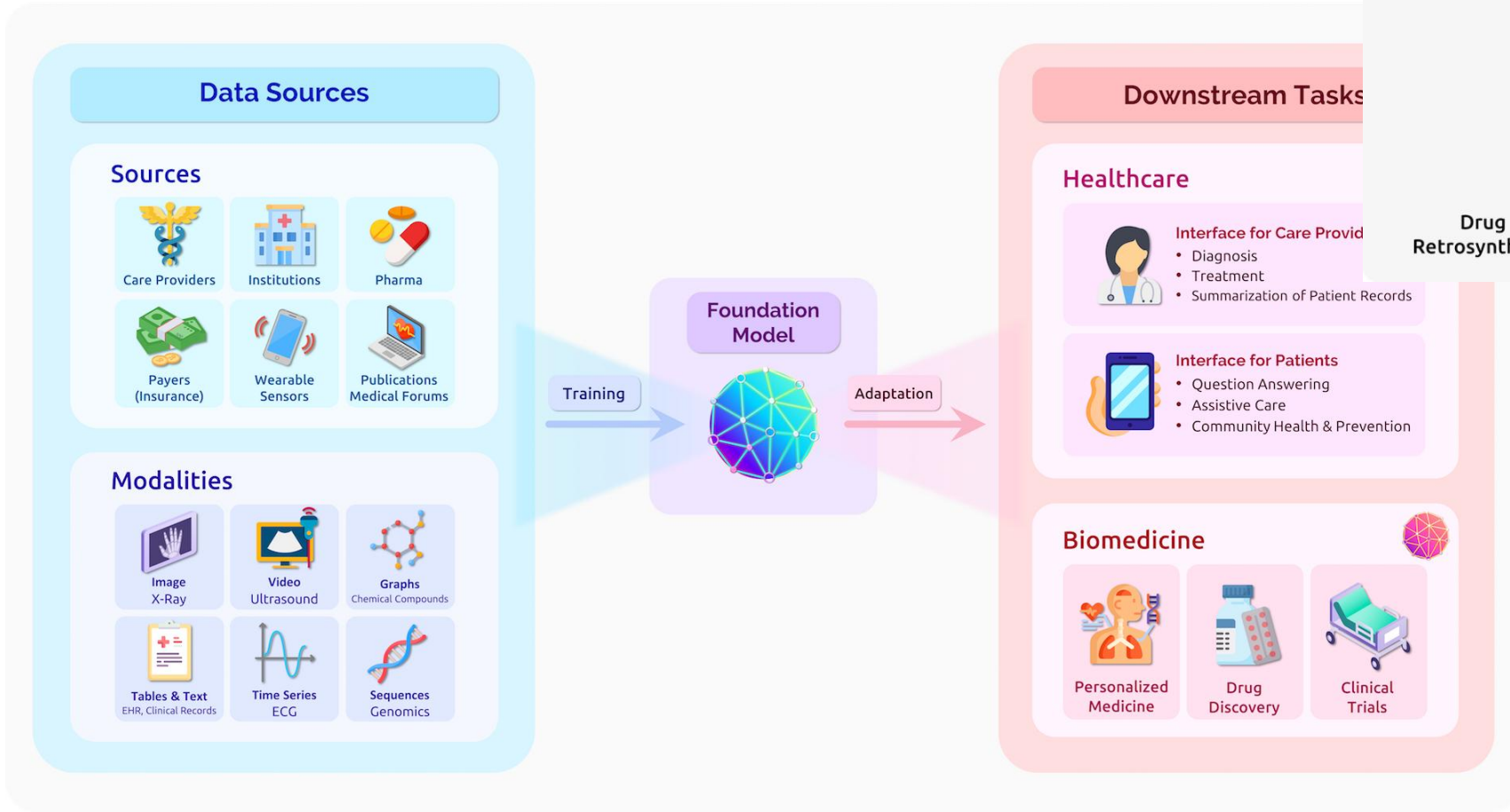
Source: Arxiv:2108.07258

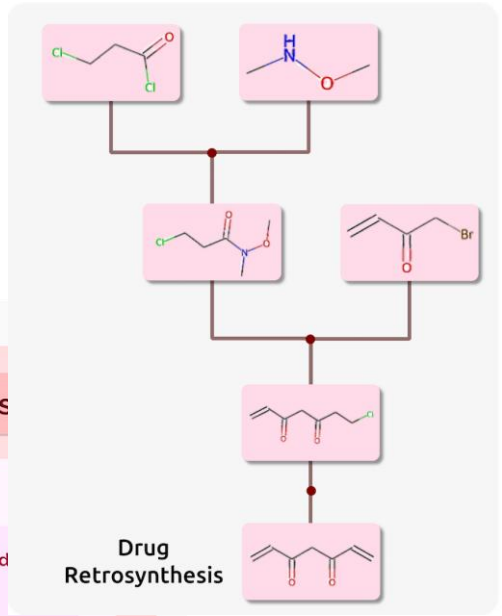
Beyond Only Language



Source: Arxiv:2108.07258

Foundation Models in Healthcare





..and beyond

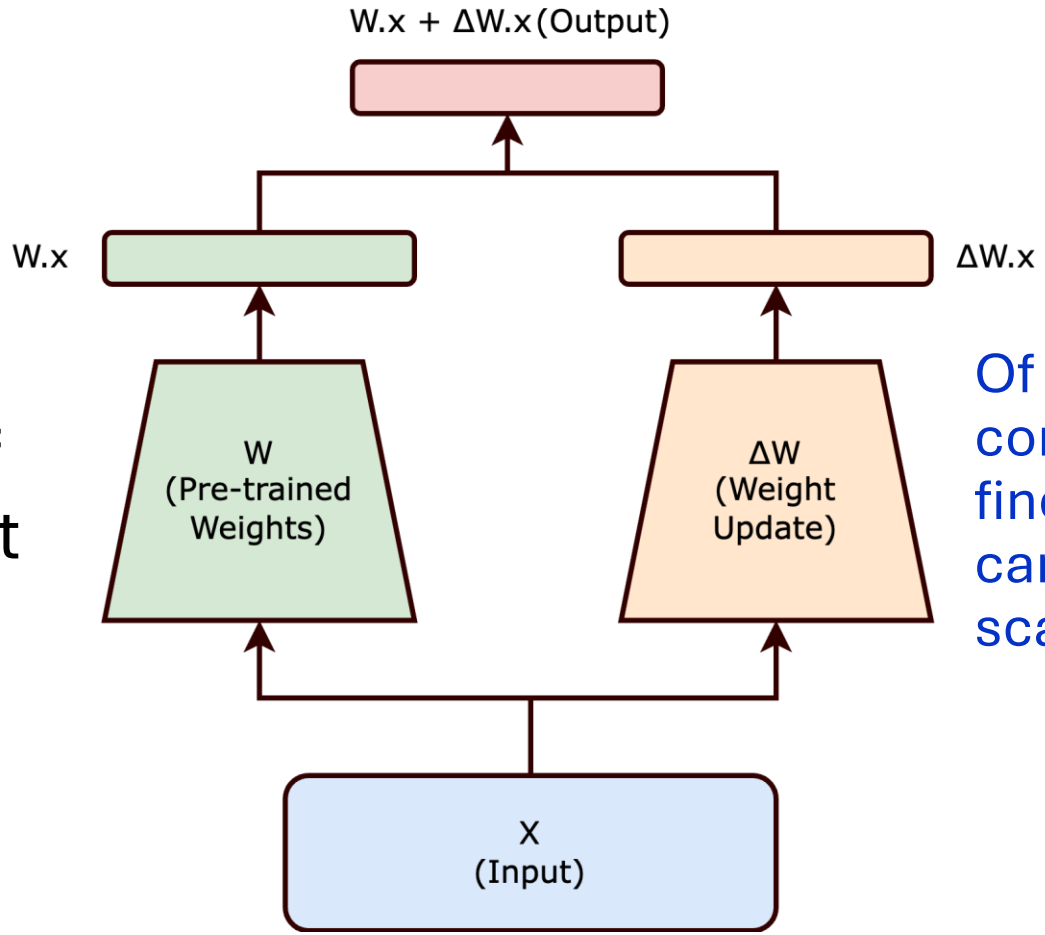
Source: Arxiv:2108.07258

Pretrained weights adaptation revised

The update of a generic set of weights W can be seen as

$$W' = W + \Delta W$$

If W are pretrained weights (e.g. of a foundation model), we may want to leave them unchanged and instead learn the weight update ΔW



Of this can be compressed, fine tuning can be made scalable

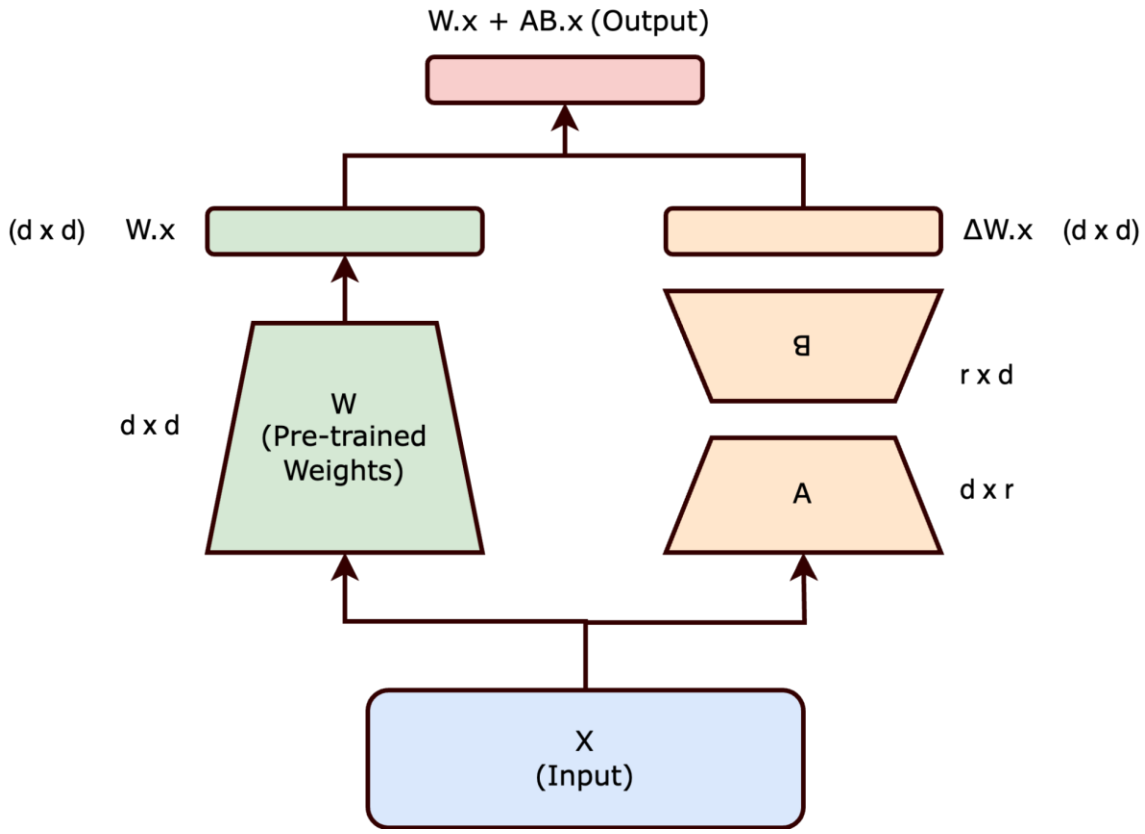
Img source: link

Low Rank Adaptation (LoRA)

Intrinsic rank hypothesis: significant weight changes can be captured using a lower dimensional representation

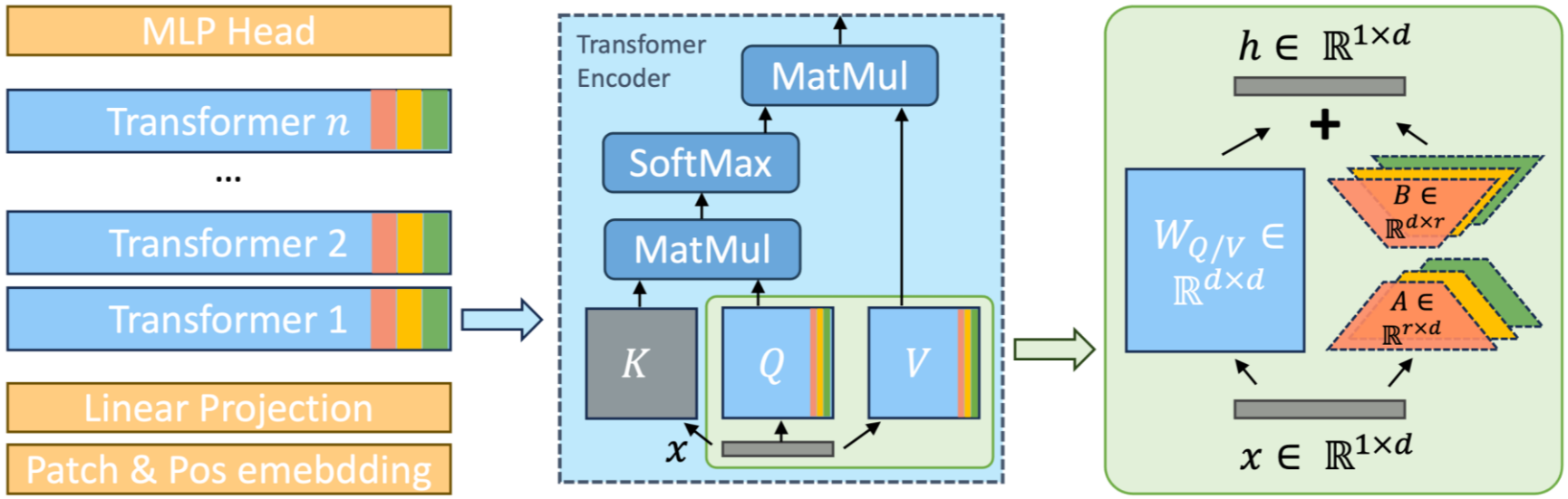
So, ΔW can be represented effectively by its low rank factorization in two simple matrices A and B

Often used together with other efficient designs (weight quantization, sparsification, …)



Img source: link

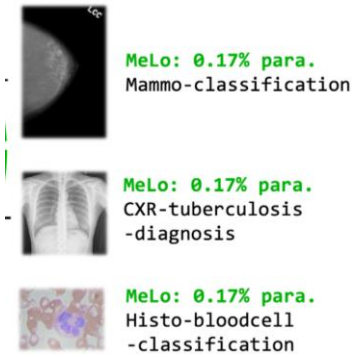
MeLO – Low Rank for Medical Imaging



Vision Transformer (ViT) pretrained on general purpose images (ImageNet)

Low-rank adapters (one per task type) to adjust Query and Value embedding matrices

Comparable performance to fine tuning using 0.1% adaptive parameters



Source & code: GitHub

Wrap-up

Take Home Lessons

- Language processing techniques are key to enable intelligent healthcare applications
  - Allow to squeeze information from otherwise “far-from-ideal” textual data sources
  - Textual data in healthcare and bioinformatics is much more than natural language data
- Dense embeddings are key to enable effective processing of symbolic textual data in highly metric machines such as neural networks
  - One key intuition above all: semantics can be provided by the context (distributional hypothesis)
- Language modelling builds on the distributional hypothesis to create a label-less training mechanism (self-supervision)
  - Scale is key, to achieve emergence of general and generalizable abilities
  - Three key scheme: encoder only; decoder only; encoder-decoder
- Foundation models to generalize language modelling results outside text
  - Low rank adaptation to make fine-tuning sustainable

Next Lectures

- Language processing laboratory

Deep learning for graphs

- Graph formalism and learning tasks on structured data
- Neural message passing paradigms
- Components of a graph neural network
- A preview on advanced neural models for graphs
- Applications to healthcare and bioinformatics