

Clinical NLP (Mostly LLMs)

Peter Szolovits

Feb 25, 2025

some slides from Eric Lehman

Maybe?



Generative AI has arrived — Healthcare will never be the same!

Generative AI is redefining care pathways, reducing workloads, accelerating drug discovery, and addressing long-standing pain points for patients and providers alike. Join Generative AI leaders from across healthcare.

Bulk of Valuable Data are in Narrative Text

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

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**.

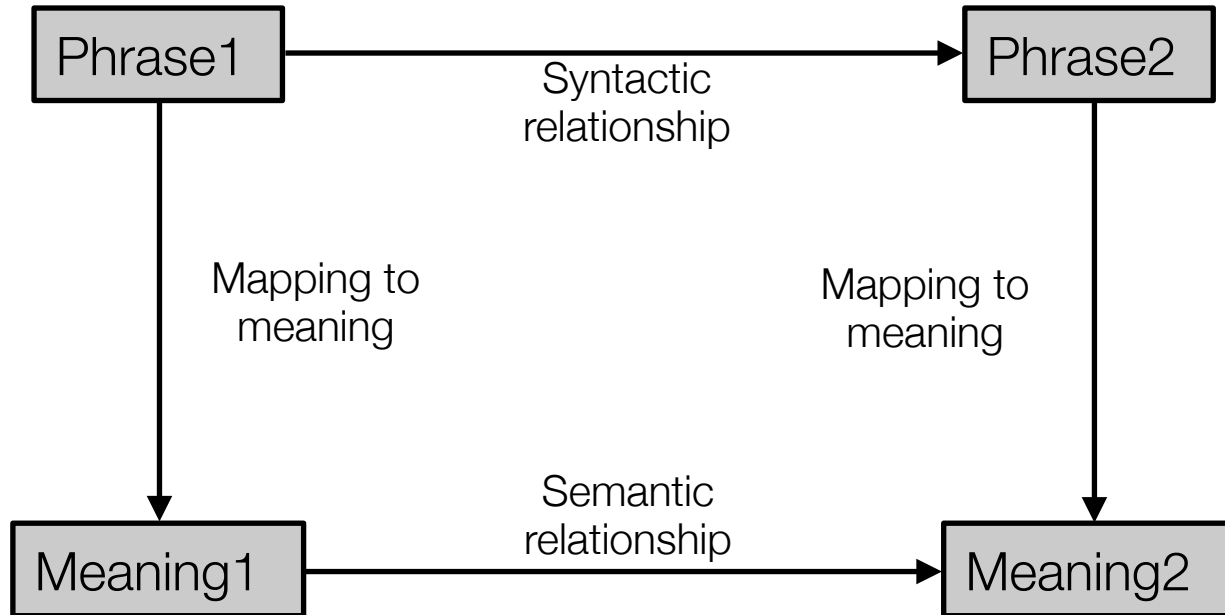
On the **morning of the 22nd** the patient developed **tachypnea** with a **chest X-ray showing a question of congestive heart failure**. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given **intravenous Lasix**.

A **arterial blood gases on 100 percent face mask** showed an **oxygen of 205, CO2 57 and PH 7.3**. An **electrocardiogram** showed **ST depressions in V2 through V4 which improved with sublingual and intravenous nitroglycerin**. The patient was transferred to the Coronary Care Unit for management of his **congestive heart failure**, **ischemia** and **probable aspiration pneumonia**.

NLP Tasks and Early Approaches

- Term spotting, synonym identification, negation detection, ...
- Identify taxonomic terms: “rheumatoid arthritis” ==> 714.0 (ICD9)
- Find aspects, e.g., time, location, certainty
- De-identify PHI
- Find relations: precedes, causes, treats, prevents, indicates, ...
- Summarization
- Question answering
- De-Duplication
- ...

Perhaps Semantics Mirrors Syntax



- Frederick B. Thompson, “English for the Computer.” *Proceedings of the Fall Joint Computer Conference* (**1966**) pp. 349-356

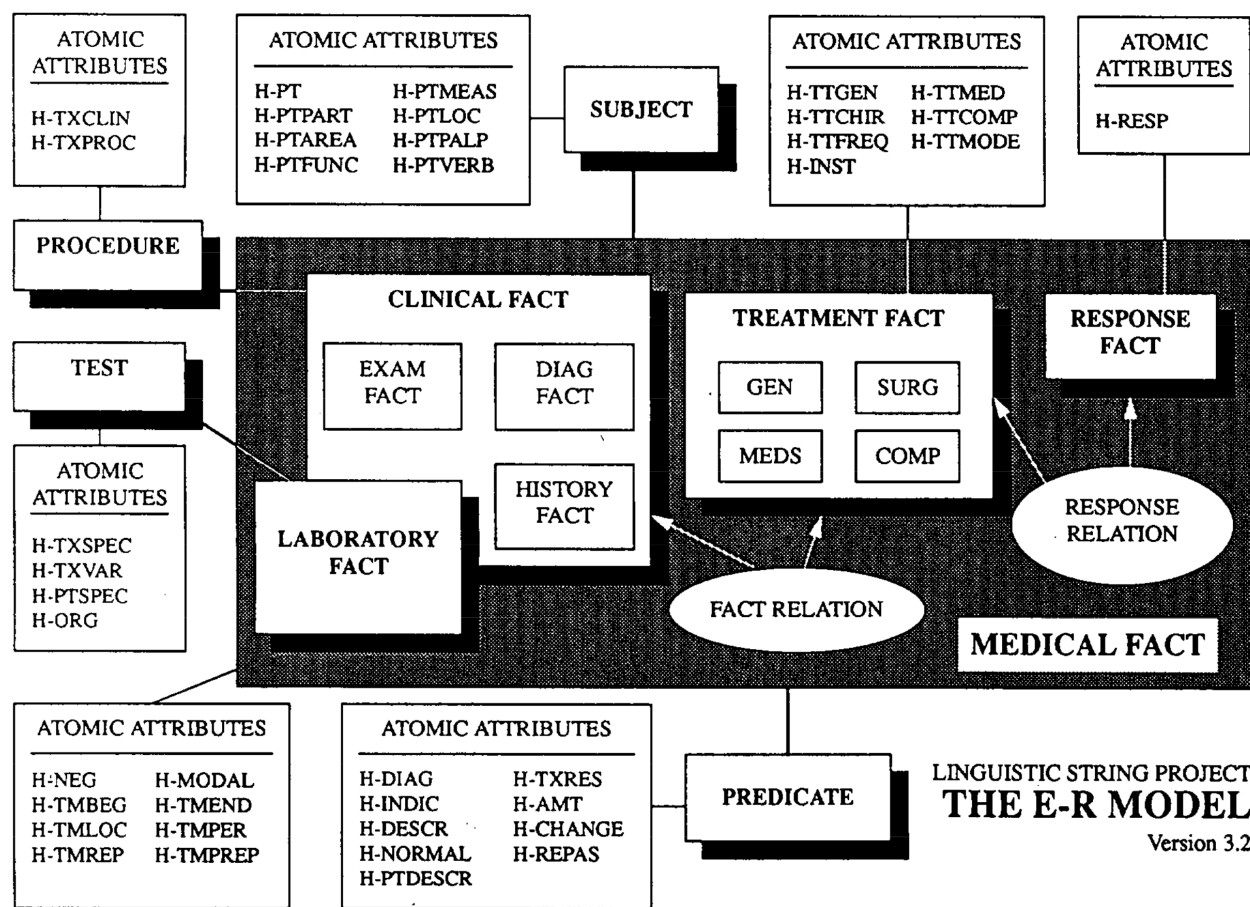


Figure 4 Schematic overview of the types of medical facts seen in patient documents and their associated lexical (“atomic”) attributes. The CLINICAL FACT subtypes are distinguished by the paragraph they occur in: EXAM, DIAG, LAB, HISTORY. The TREATMENT FACT type is subdivided into general medical management (GEN), surgery (SURG), medications (MEDS), and all other therapies (COMPLEMENTARY). An instance of a TREATMENT FACT is often coupled to a RESPONSE FACT via a RESPONSE RELATION, e.g. *Much improved on penicillamine 750 mg daily.*

Term Spotting in 2020

- Traditionally, lists of coded items, narrative terms and patterns hand-crafted by researcher
 - E.g., N3C (National Covid Cohort Collaborative)
 - Institutions don't want to share notes, even de-identified
 - Lack of sophistication at most medical centers to run sophisticated tools
 - Instead, run simple term matching algorithms and report just terms
- Negation and uncertainty handled by somewhat ad-hoc methods
 - NegEx is widely used, \exists many more sophisticated variants
- Generalize terms to get better coverage
 - Manually or automatically identify high-certainty “anchors”
 - Learn related terms to augment the set of terms
 - From knowledge bases such as UMLS
 - From co-occurrence in EMR data
 - From co-occurrence in publications

COVID symptoms
• Fever
• Chill
• Cough
• Fatigue
• Nasal obstruction
• Loss of appetite
• Diarrhea
• Abdominal pain
• Nausea
• Vomiting
• Sore throat
• Headache
• Myalgia
• Loss of taste
• Loss of smell
• Dyspnea
• Chest pain
• Delirium
• Hypersomnia
• Cyanosis

anosmia
(C0003126)

What Disease does this Patient Have?

A Large-scale Open Domain Question Answering Dataset from Medical Exams

- Professional Medical Board Exams: English, Traditional Chinese, simplified Chinese
 - Document collection, Questions, Answer candidates

	US	China	Taiwan
Metric	USMLE	MCMLE	TWMLE
# of options per question	4	4	4
Avg./Max. option len.	3.5 / 45	7.3 / 100	20.6 / 210
Avg./Max. question len.	116.6 / 530	45.7 / 333	61.0 / 1950
vocab/character size	63317	3263	3588
# of questions			
Train	10178	27400	11298
Development	1272	3425	1412
Test	1273	3426	1413
All	12723	34251	14123


Metric	USMLE/TWMLE	MCMLE
# of books		18 / 33
# of paragraphs		231,581 / 116,216
# of tokens		12,727,711 / 14,730,364
Vocabulary size		245,851 / 4,695
Avg./Max. paragraph length		55.0 / 1,234 / 126.7 / 9,082

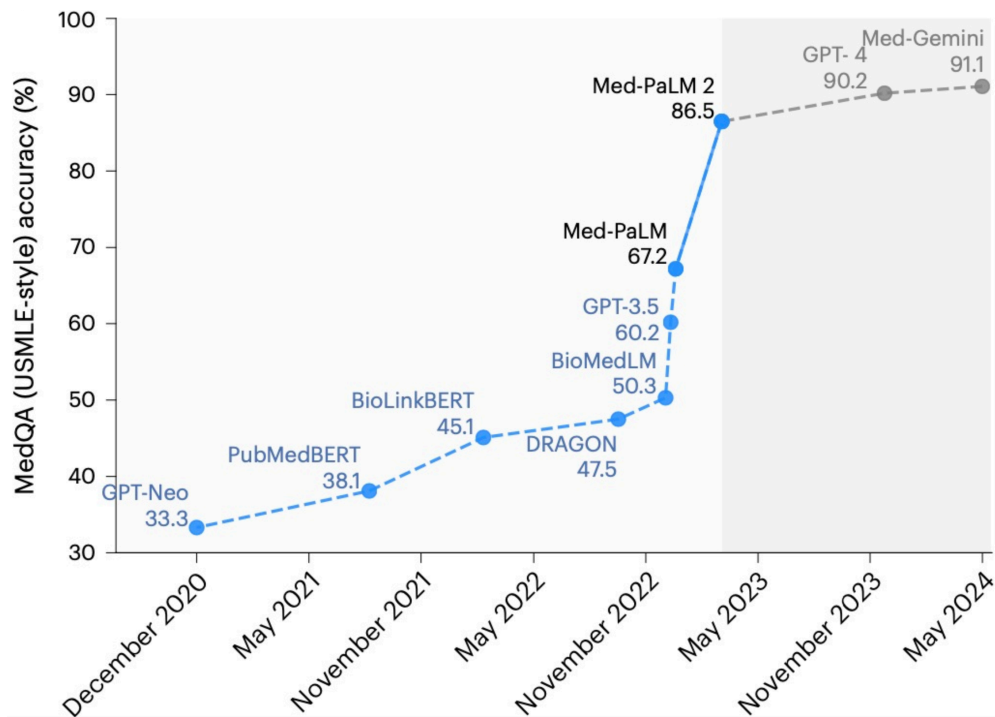
- Methods:

- IR: Lucene, ElasticSearch; Mutual Information
- DocReader:
 - Use IR to get relevant documents, concatenated = c
 - BiGRU to encode c and each Q/A pair, max-pool $\Rightarrow \mathbf{h}_c, \mathbf{h}_{qa_i}$
 - $\mathbf{h} = [\mathbf{h}_c; \mathbf{h}_{qa_i}; \mathbf{h}_c \cdot \mathbf{h}_{qa_i}; |\mathbf{h}_c - \mathbf{h}_{qa_i}|]$
 - $p(q, a_i | c) = W_1(\tanh(W_2 \mathbf{h}))$
- BERT + softmax over [CLS] c [SEP] q, a_i [SEP]
 - $p(q, a_i | c) = W \mathbf{h}$

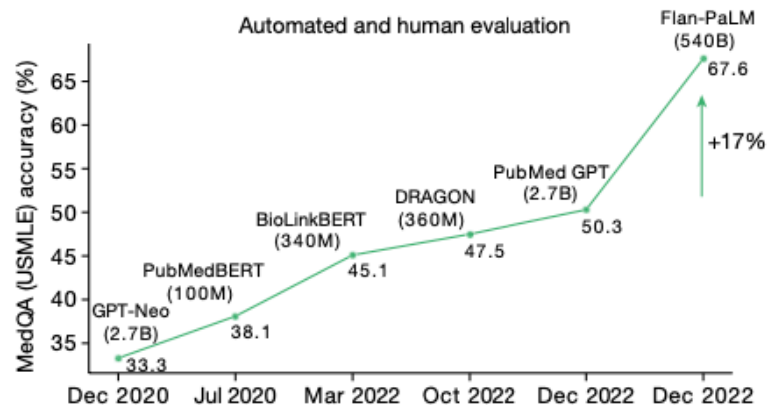
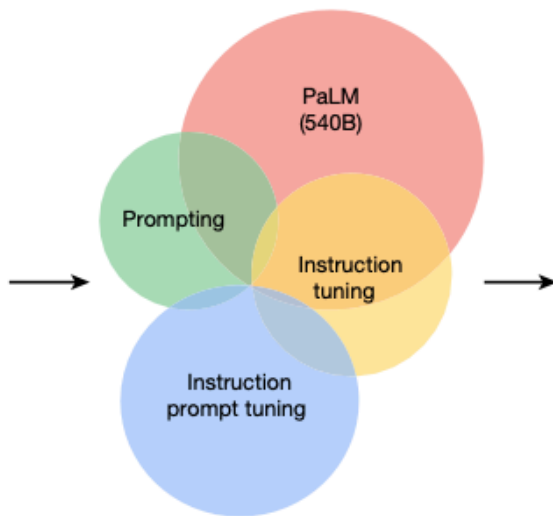
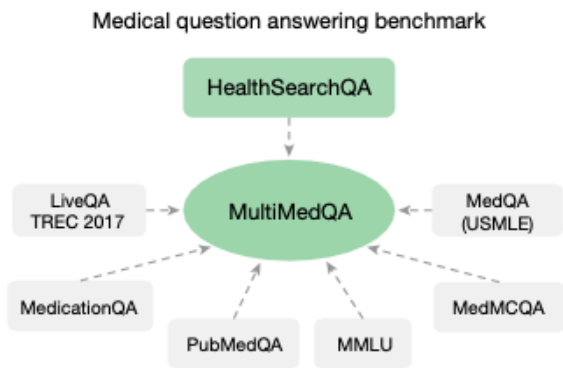
Article

What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams

Di Jin ^{1,*}, Eileen Pan ¹, Nassim Oufattole ¹, Wei-Hung Weng ¹, Hanyi Fang ² and Peter Szolovits ¹



Med-PaLM (2023)



Q: How long does it take for newborn jaundice to go away?

Med-PaLM: Newborn jaundice is when a newborn baby's skin and eyes look yellow. It is caused by a buildup of a substance called bilirubin in the baby's blood. Bilirubin is a byproduct of the normal process of red blood cells breaking down. Newborn jaundice is common and typically harmless, but it can be a sign of a more serious condition. The jaundice typically goes away on its own within a few weeks. In some cases, it may last longer or require treatment. The time it takes for the jaundice to go away can vary depending on the cause and the severity of the jaundice. If the jaundice is severe or lasts longer than a few weeks, the doctor may recommend testing or treatment to determine the cause and prevent complications.

Med-PaLM performs encouragingly
on consumer medical question answering

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023 Aug;620(7972):172–80. Available from: <https://www.nature.com/articles/s41586-023-06291-2>

Med-PaLM

- 540-billion parameters
- SOA Accuracy on each dataset
 - MedQA3, MedMCQA4, PubMedQA5 and Measuring Massive Multitask Language Understanding (MMLU) clinical topics
- Instruction prompt tuning
 - prepend random words (embeddings) to task-specific prompts and train them by backprop
- Far better than previous models
- Inferior to clinicians

Med-PaLM 2 (pub. 2025)

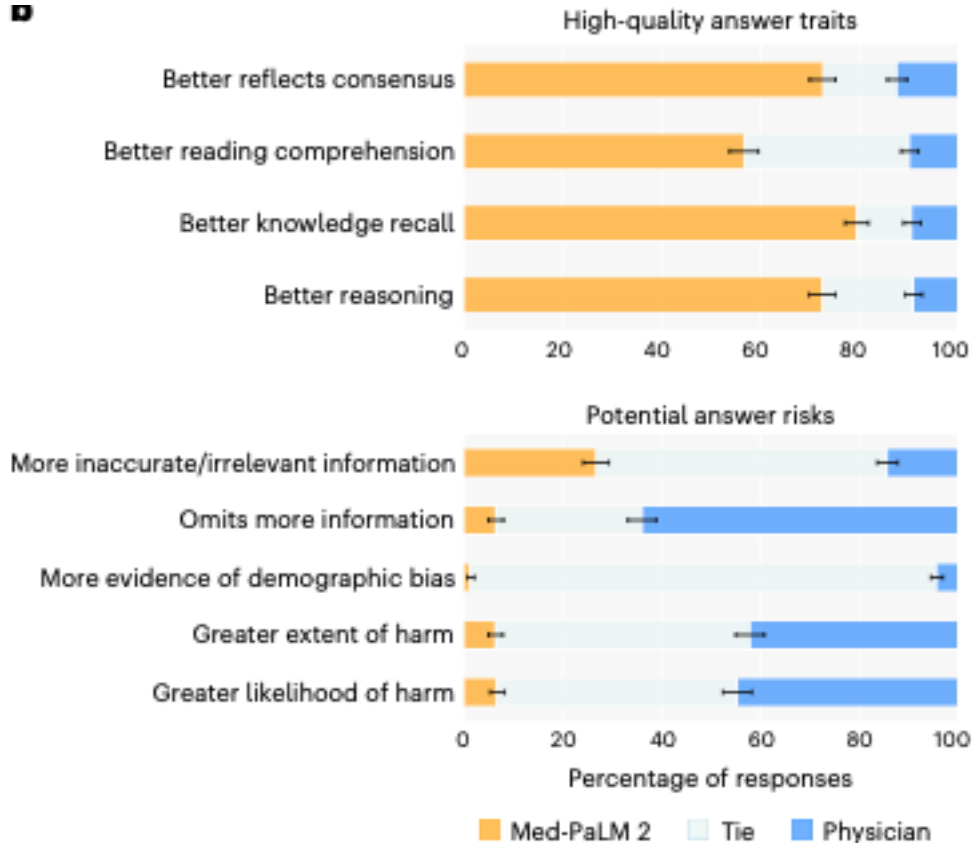
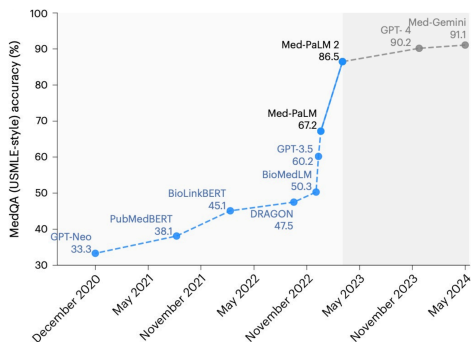
nature medicine



Article

<https://doi.org/10.1038/s41591-024-03423-7>

Toward expert-level medical question answering with large language models



Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025 Jan 8;1–8. Available from: <https://www.nature.com/articles/s41591-024-03423-7>



Med-PaLM 2 advances

Improved base: PaLM 2

PaLM 2 is said to be a 340 billion-parameter model trained on 3.6 trillion tokens

Medical domain-specific fine-tuning

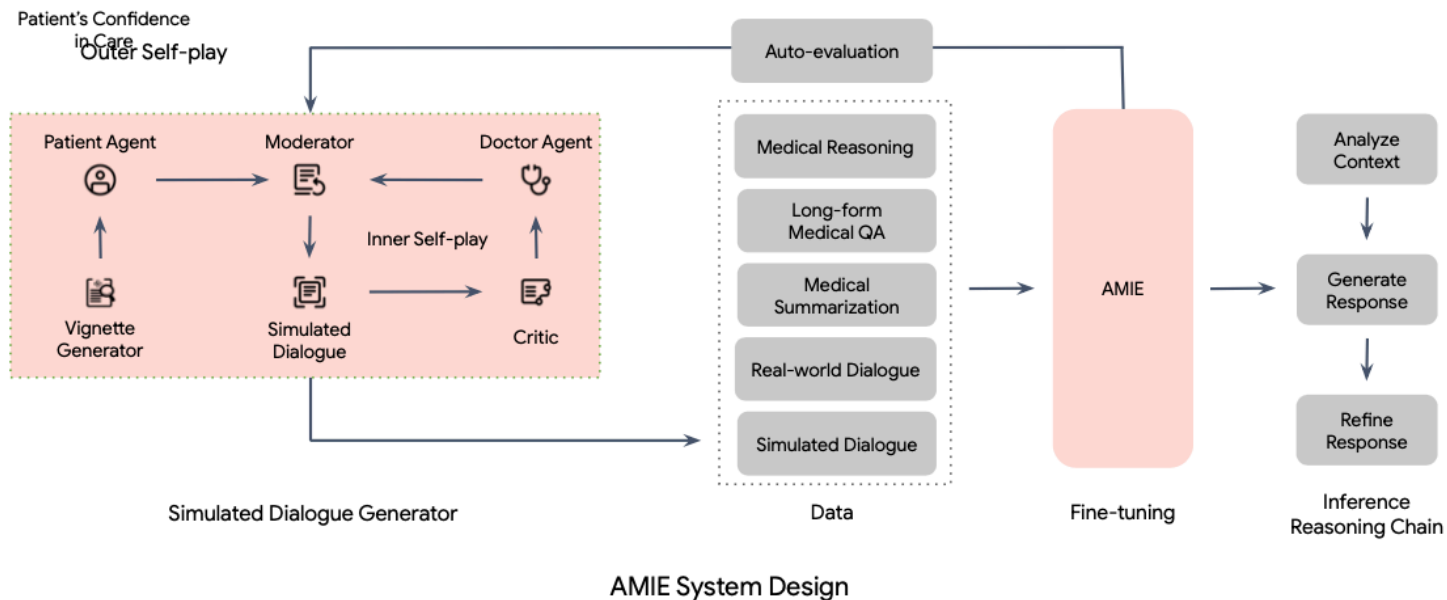
New prompting strategies

Chain-of-Retrieval Augmented Generation

- (1) An initial Med-PaLM 2 answer is generated using a zero-shot prompt.
- (2) The initial Med-PaLM 2 answer is separated into individual claims for verification.
- (3) Search queries for the claims for verification are generated.
- (4) Relevant studies and websites are retrieved using Google search.
- (5) Individual documents are summarized.
- (6) Med-PaLM 2 generates a final answer using the question and concatenated summaries.

Towards Conversational Diagnostic AI

Can AMIE (Articulate Medical Intelligence Explorer) interact with a patient to diagnose disease as well as physicians? [Based on PaLM 2, not Med-PaLM 2]

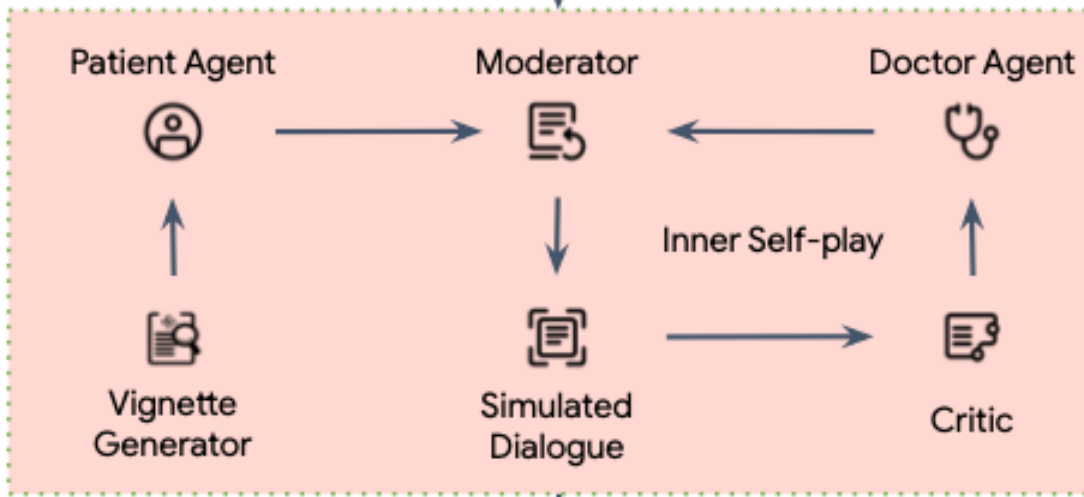


Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards Conversational Diagnostic AI. arXiv; 2024.

Available from: <http://arxiv.org/abs/2401.05654>

Self-Play to Train AMIE

- Inner loop:
 - **Vignette Generator:** AMIE leverages web searches to craft unique patient vignettes given a specific medical condition.
 - **Simulated Dialogue Generator:** Three LLM agents play the roles of patient agent, doctor agent, and moderator, engaging in a turn-by-turn dialogue simulating realistic diagnostic interactions.
 - **Self-play Critic:** A fourth LLM agent acts as a critic to give feedback to the doctor agent for self-improvement. Notably, AMIE acted as all agents in this framework.



Recursive Use of LLM

Patient Agent Instruction:

You are a patient chatting with a doctor over an online chat interface. The doctor has never met you before. <patient vignette> Respond to the doctor's questions honestly as they interview you, asking any questions that may come up.

Doctor Agent Instruction:

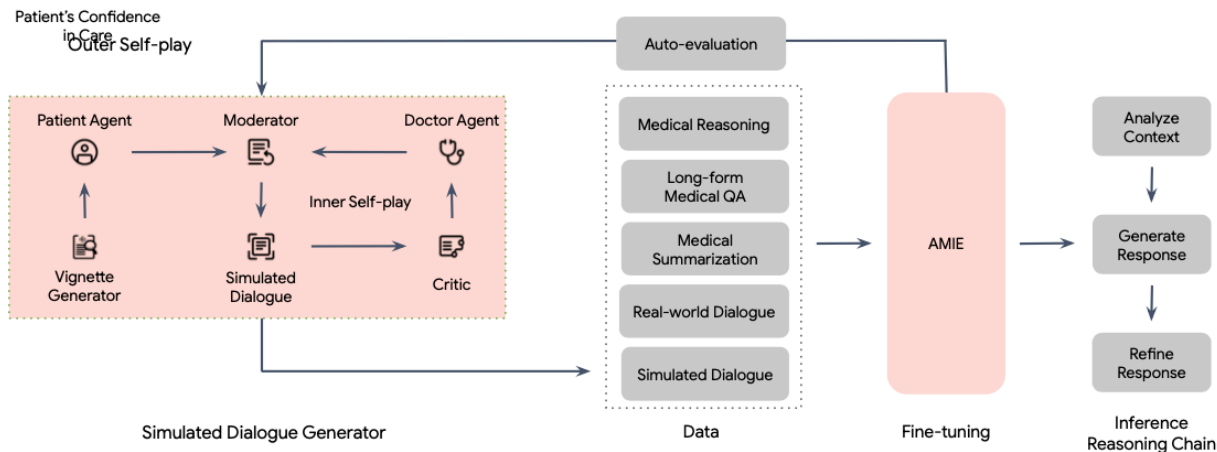
You are an empathetic clinician asking a patient about their medical history over an online chat interface. You know nothing about the patient in advance. Respond to the patient with a single-turn response to better understand their history and symptoms. Do not ask more than two questions. If the patient asks a question, be sure to answer it appropriately.

Moderator Instruction:

The following is a conversation between a doctor and a patient: <dialog> The conversation should only come to an end if the doctor has finished giving the patient a diagnosis and treatment plan and the patient has no questions left. A conversation also comes to an end if the doctor or patient says goodbye. Question: has the conversation come to an end? Yes or No.

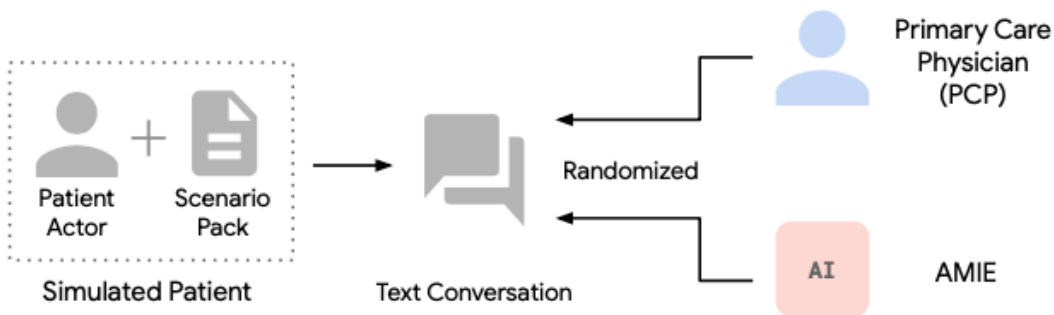
Train AMIE on Large Collection of Recorded Tasks

- Outer loop:
 - Collect many samples of QA, medical reasoning, summarization, real dialogue, and simulated dialogue (from inner loop)
 - Fine-tune AMIE on all these
 - Chain-of-reasoning to generate and refine AMIE: context \Rightarrow response

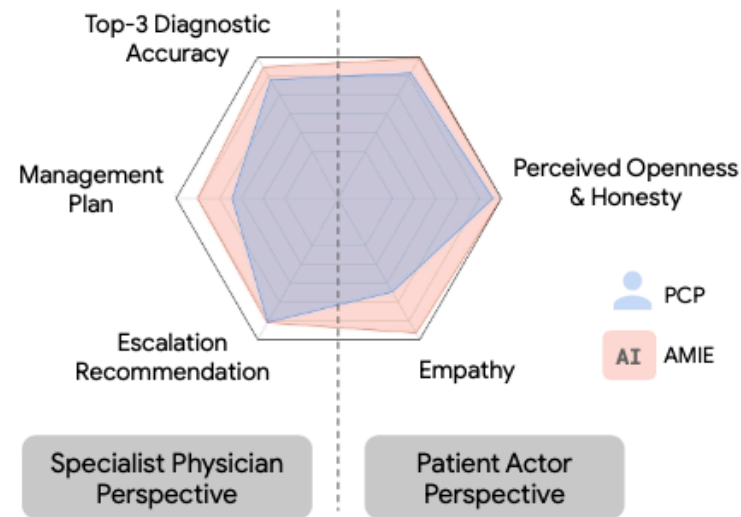


AMIE System Design

Evaluating Performance vs. Primary Care Doctors



Randomized Study Design for Remote Objective Structured Clinical Examination (OSCE)



AMIE Outperforms PCPs on Multiple Evaluation Axes for Diagnostic Dialogue

Evaluation by Specialist Physicians

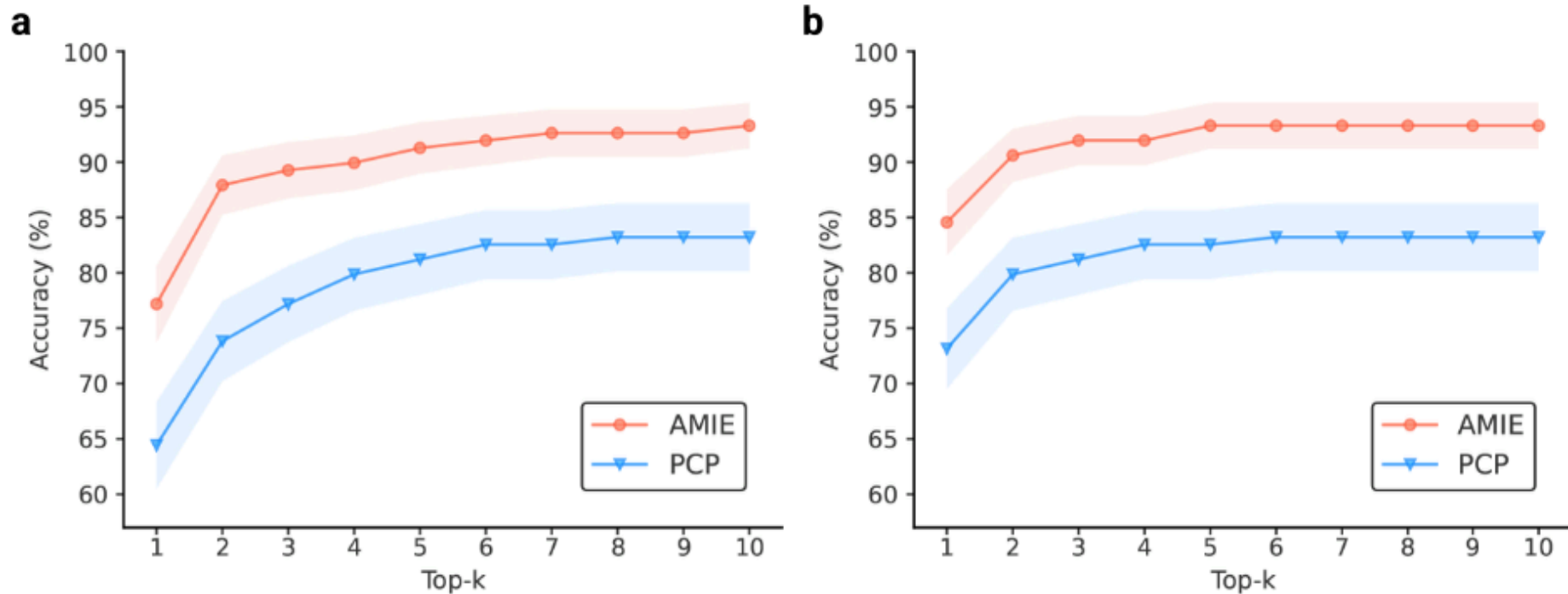
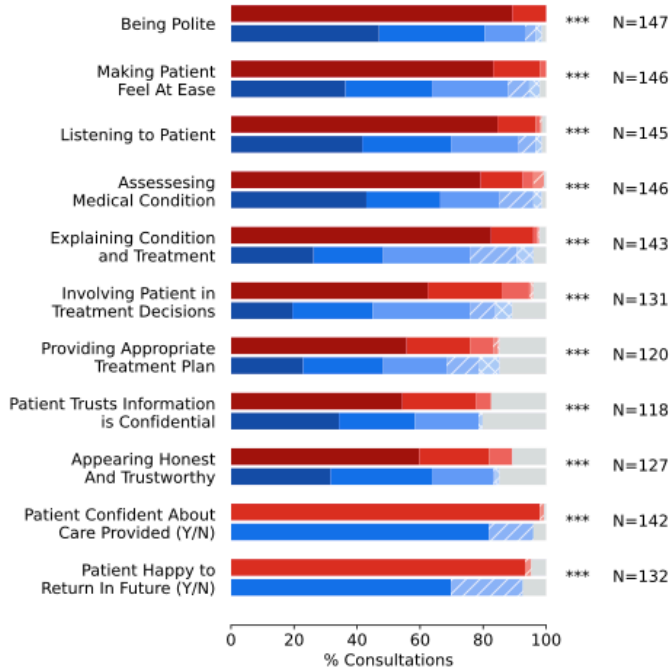


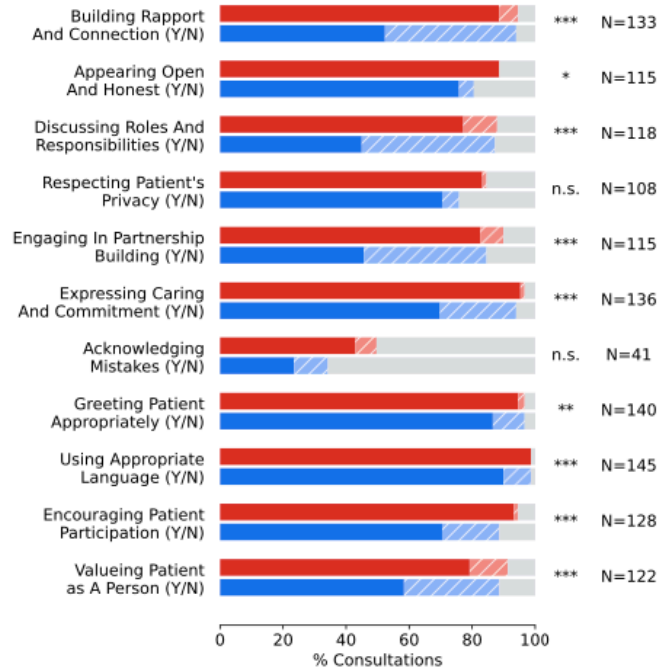
Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (a) and all diagnoses in the accepted differential (b). Bootstrapping (n=10,000) confirms all top-k differences between AMIE and PCP DDx accuracy are significant with $p < 0.05$ after FDR correction.

Ratings by patient actors

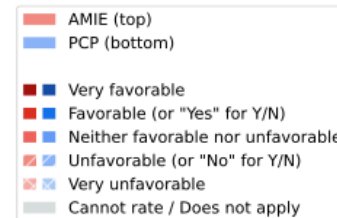
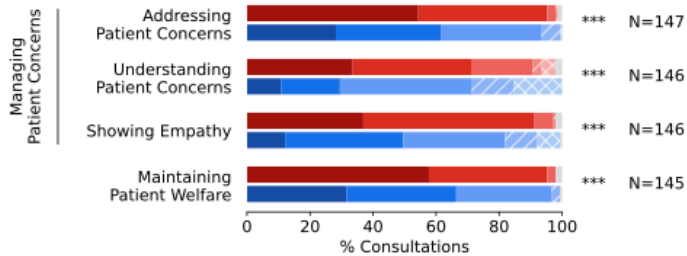
GMCPQ



PCCBP



PACES



PACES

PCCBP

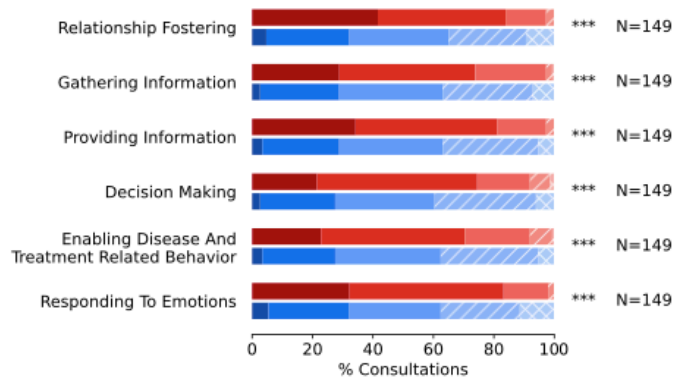
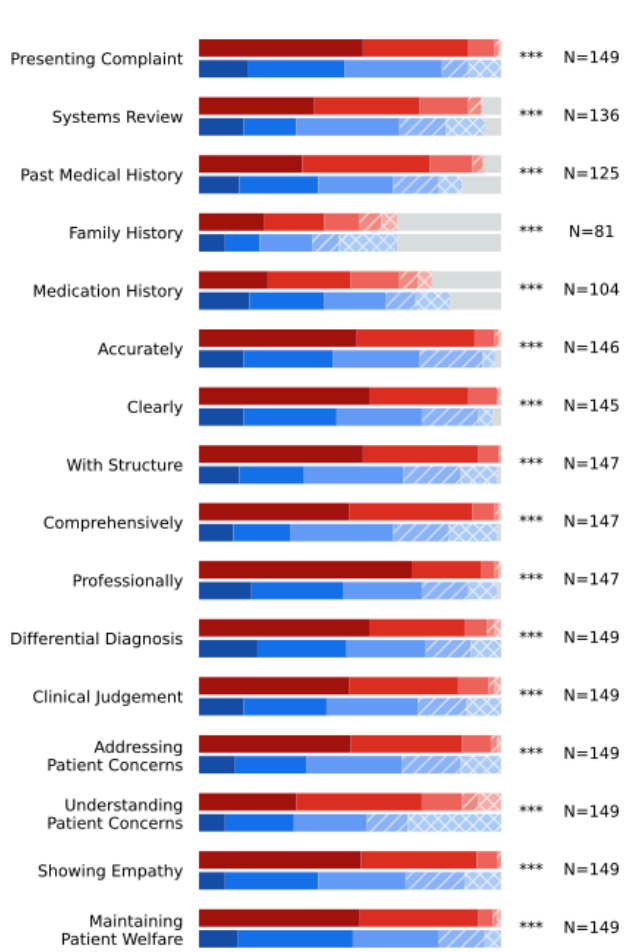
Ratings by Specialists

*** = p < 0.001

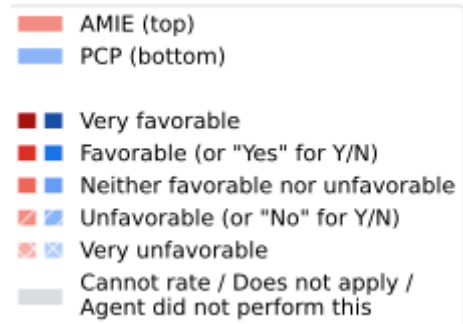
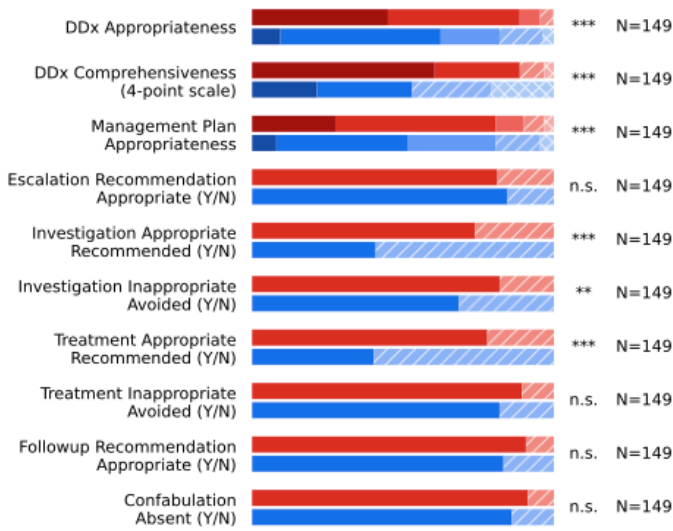
Eliciting

Explaining Relevant Clinical Information

Managing Patient Concerns



Diagnosis & Management



0 20 40 60 80 100
% Consultations

0 20 40 60 80 100
% Consultations

Double Down on LLM Abilities

- Curation of training data is critical
 - Use LLM to generate selection rules
 - Run modest-scale experiments with different selection policies; find best
- Appropriate prompt is critical
 - Use LLM to generate a collection of prompts
 - Run modest-scale experiments with different prompt policies; find best

Gao M. Systematic Development of Healthcare AI: Integrating Data Curation, Algorithm Optimization, Benchmark Design and Clinical Applications [PhD Thesis]. [Cambridge, MA]: MIT; 2025.

NLP has moved... Fast

FEB. 27 / MARCH 6, 2023

TIME

T Hi ChatGPT

 Hello! How can I assist you today?

T What do you think of a TIME cover story with the title, "The AI Arms Race Is Changing Everything"?

 As an AI language model, I don't have personal opinions or beliefs, but I can provide some context and information on the subject.

"The AI Arms Race is Changing Everything" could be an attention-grabbing and thought-provoking title for a cover story on AI. A cover story in a widely-read publication like TIME could help raise public awareness about the potential risks and benefits of this trend, and stimulate discussions on how to ensure that AI is developed and deployed in a responsible and ethical manner.

It's important to note that the title of the story could also be interpreted as sensationalist or alarmist, and may not accurately reflect the complexities ■

by
ANDREW R. CHOW
& BILLY PERRIGO
(humans)

time.com

How Did We Get Here?

From Symbols to Vector Spaces

- Continuous optimization (e.g., stochastic gradient descent) is far more efficient than enumerating discrete symbolic models
- How to map words to vectors?

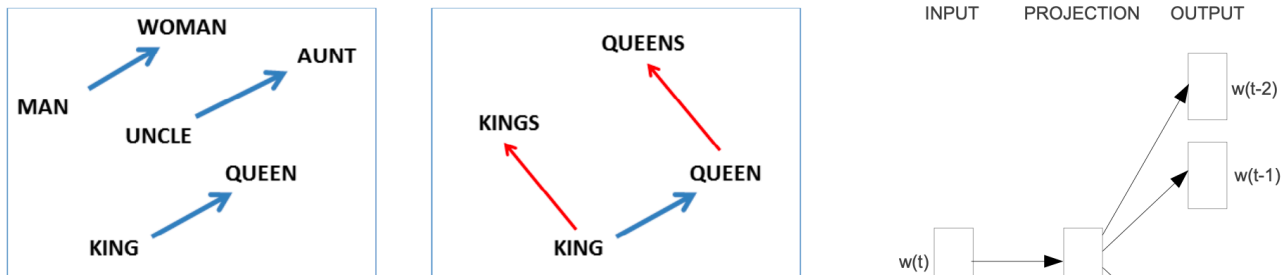


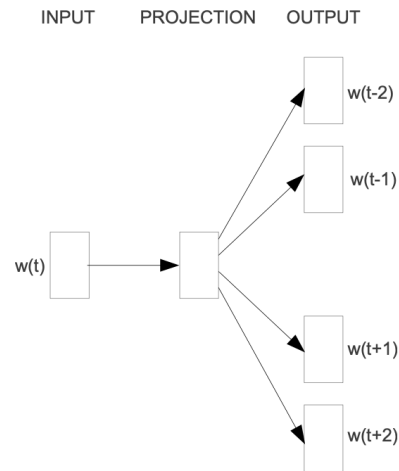
Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. 2013;

Available from: <http://arxiv.org/abs/1301.3781>.

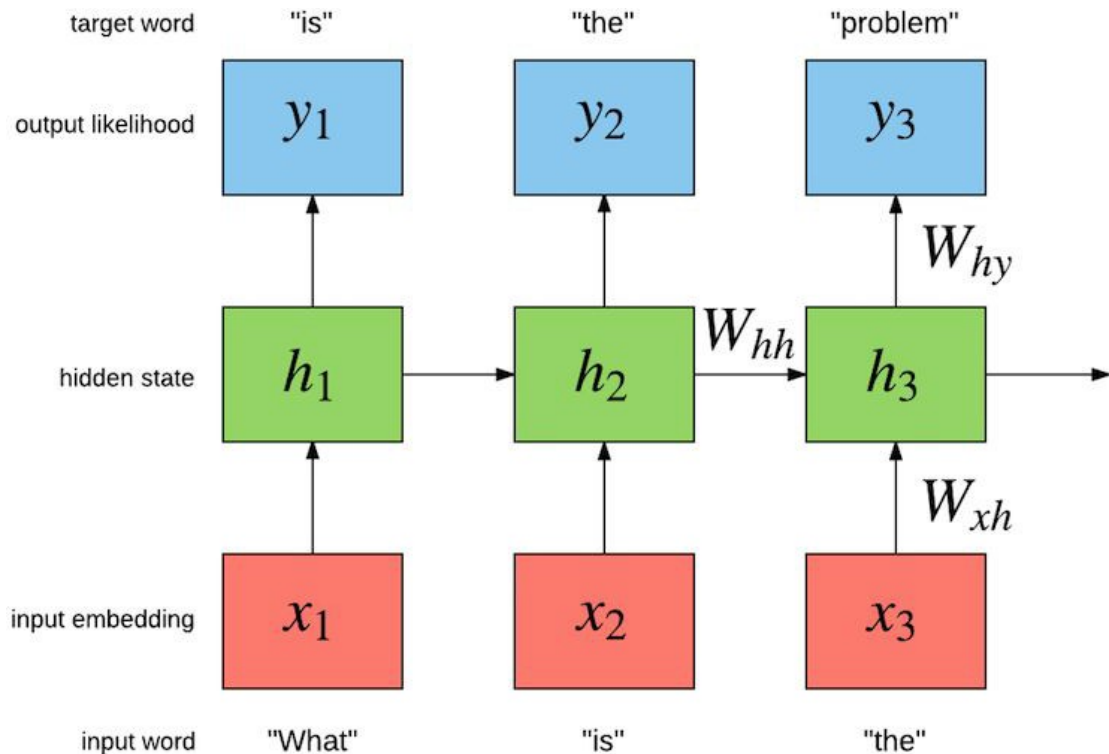
Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. [Internet]. HLT-NAACL; 2013.

Available from: http://scholar.google.com/scholar?q=related:nQa_EOK3iMJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5

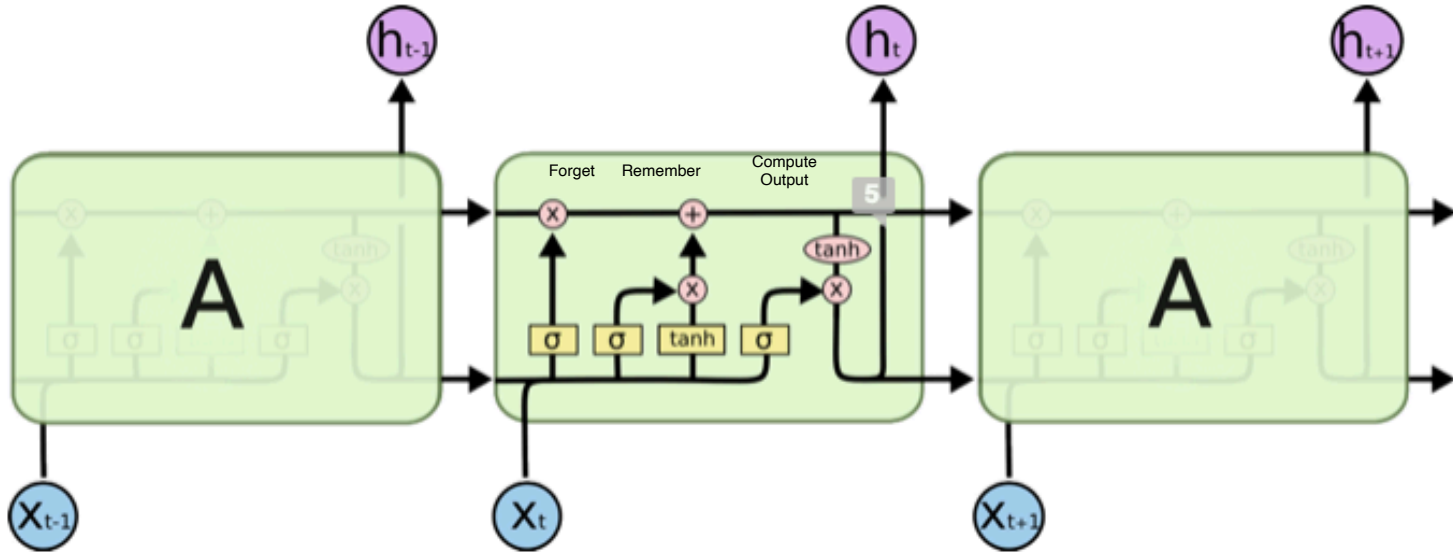


Skip-gram

Recurrent Neural Networks



Long Short-Term Memory (LSTM)

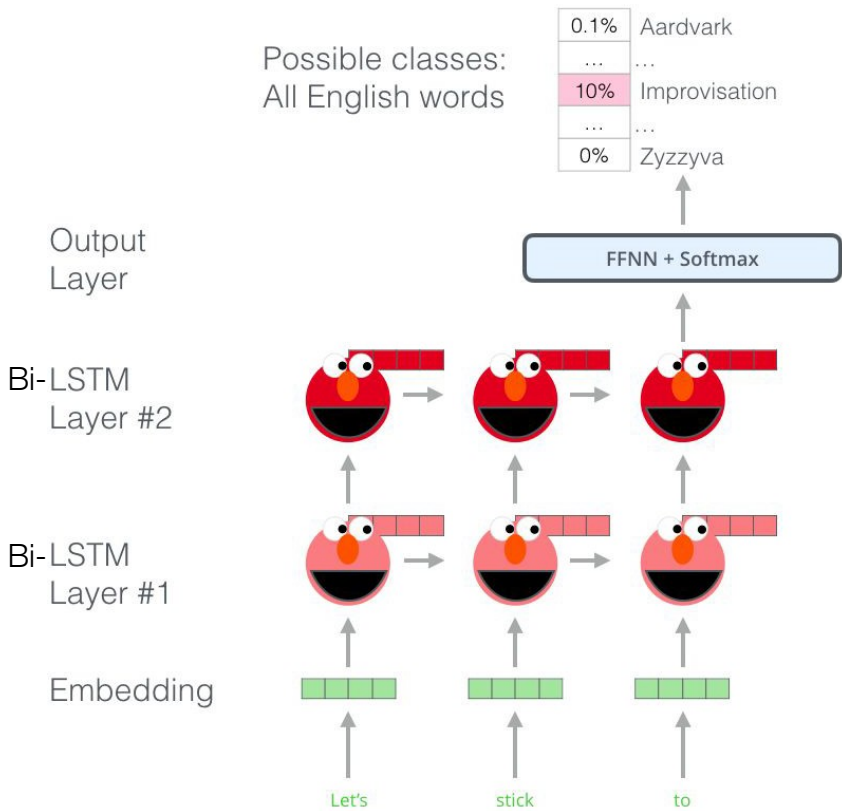


The repeating module in an LSTM contains four interacting layers.

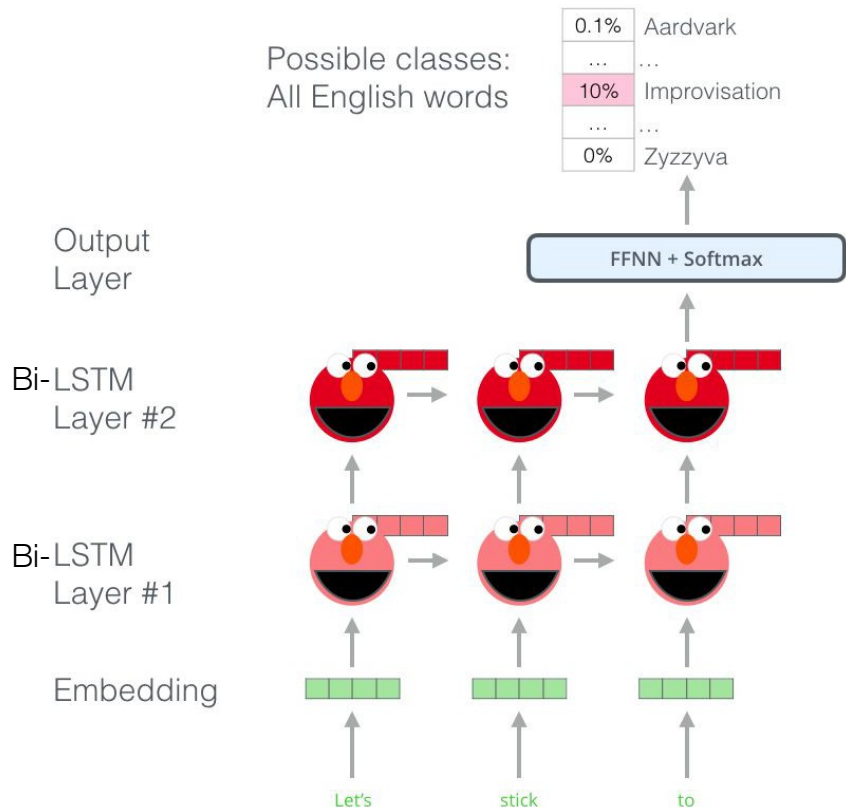
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

ELMO Pretraining – Language Modeling



ELMO Pretraining – Language Modeling



Q: How is learning to predict the next word relevant to doing deidentification, predicting sepsis from nursing notes, etc?

A: It allows us learn how to represent words & sentences FOR FREE.

ELMo — (Contextual) Embdings from Language Models

- Bidirectional LSTM
- Builds models for every *token*, not just for every *type*
 - i.e., different embeddings for the same word in different contexts
 - basis for word-sense disambiguation
- Significantly improves performance on nearly all NLP tasks

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular <u>play</u> on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent <u>play</u> .
	Olivia De Havilland signed to do a Broadway <u>play</u> for Garson {...}	{...} they were actors who had been handed fat roles in a successful <u>play</u> , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to “play” using GloVe and the context embeddings from a biLM.

ELMO (2018)

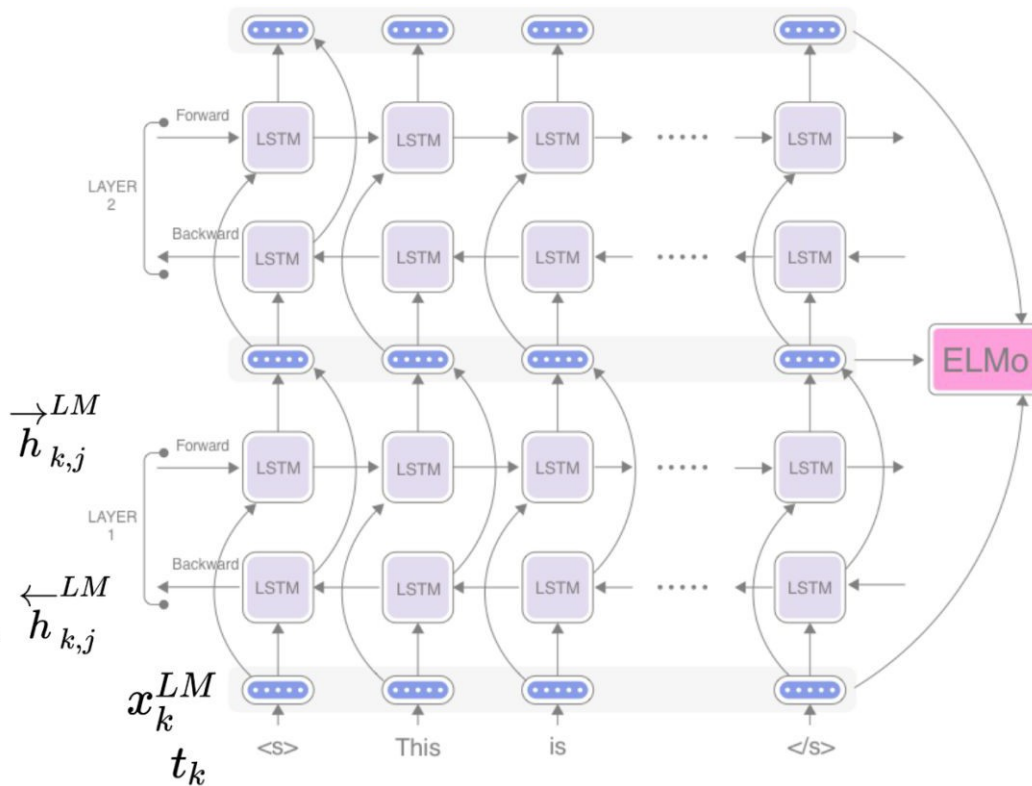
Structure

Each token t_k

L-layer biLM
computes $2L+1$
representations

k is the k -th token

j is the j -th biLM layer



Super Slow

ANN model for de-identification

But ELMO's goal is to compute an embedding for each token, not to solve a particular problem. Then stack with downstream task!

- Label-sequence optimization layer

$$s(y_{1:n}) = \sum_{i=1}^n a_i[y_i] + \sum_{i=2}^n T[y_{i-1}, y_i]$$

- Label prediction layer

- Character-enhanced token-embedding layer

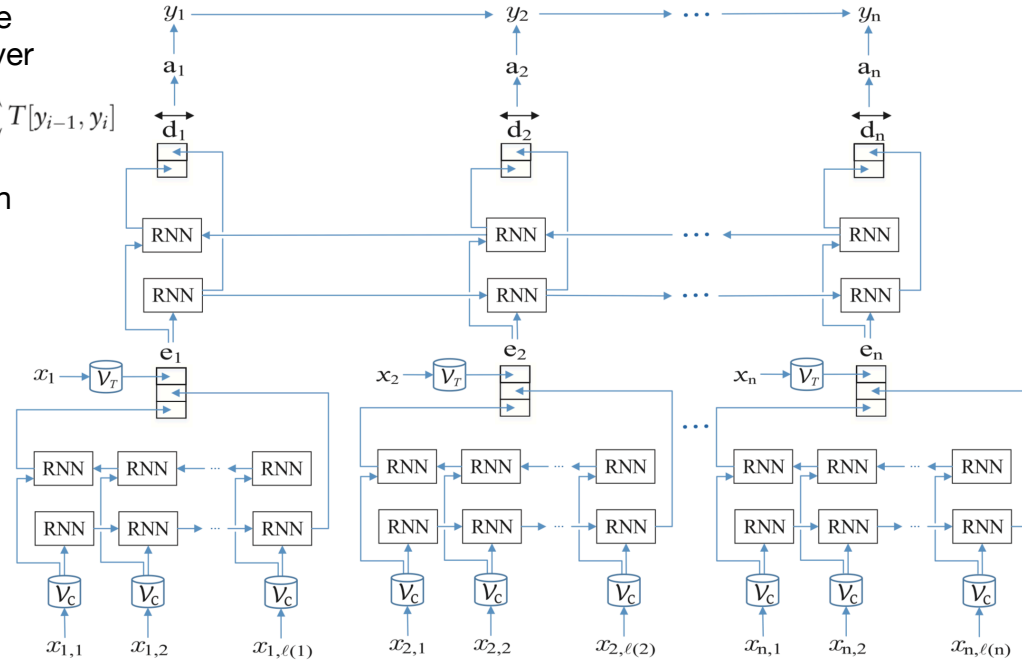


Figure 1. Architecture of the artificial neural network (ANN) model. (RNN, recurrent neural network.) The type of RNN used in this model is long short-term memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. V_T is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. V_C is the mapping from characters to character embeddings. e_i is the character-enhanced token embeddings of the i^{th} token. d_i is the output of the LSTM of the label prediction layer, a_i is the probability vector over labels, y_i is the predicted label of the i^{th} token.

Next Word Prediction/Language Modeling

Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Next Word Prediction/Language Modeling

Text: Second Law of Robotics: A robot must obey the orders given it by human beings



Generated training examples

Example #

Input (features)

Correct output (labels)

1

Second law of robotics :

a

Next Word Prediction/Language Modeling

Text: Second Law of Robotics: A robot must obey the orders given it by human beings



Generated training examples

Example #

Input (features)

Correct output (labels)

1

Second law of robotics :

a

2

Second law of robotics : a

robot

Next Word Prediction/Language Modeling

Text: Second Law of Robotics: A robot must obey the orders given it by human beings



Generated training examples

Example #

Input (features)

Correct output (labels)

1

Second law of robotics :

a

2

Second law of robotics : a

robot

3

Second law of robotics : a robot

must

...

How Would You Use ELMO For a Clinical Task?

How Would You Use ELMO For a Clinical Task?

1. Load the model that has been trained to predict:

$$P(w_i | w_1, w_2, \dots, w_{i-1})$$

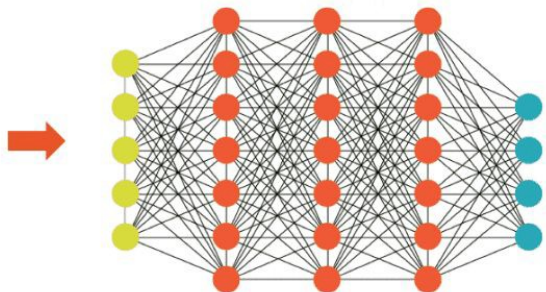
How Would You Use ELMO For a Clinical Task?

1. Load the model that has been trained to predict:

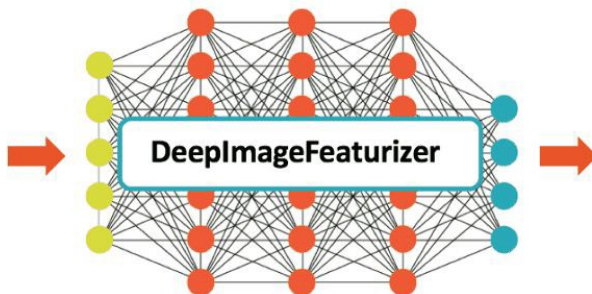
$$P(w_i | w_1, w_2, \dots, w_{i-1})$$

2. Use the model to encode your sentence and train it on **YOUR** task.

Transfer Learning/Pretraining



GIANT PANDA 0.9
RED PANDA 0.05
RACCOON 0.01
...



Chihuahua

Machine Translation by Sequence to Sequence Models

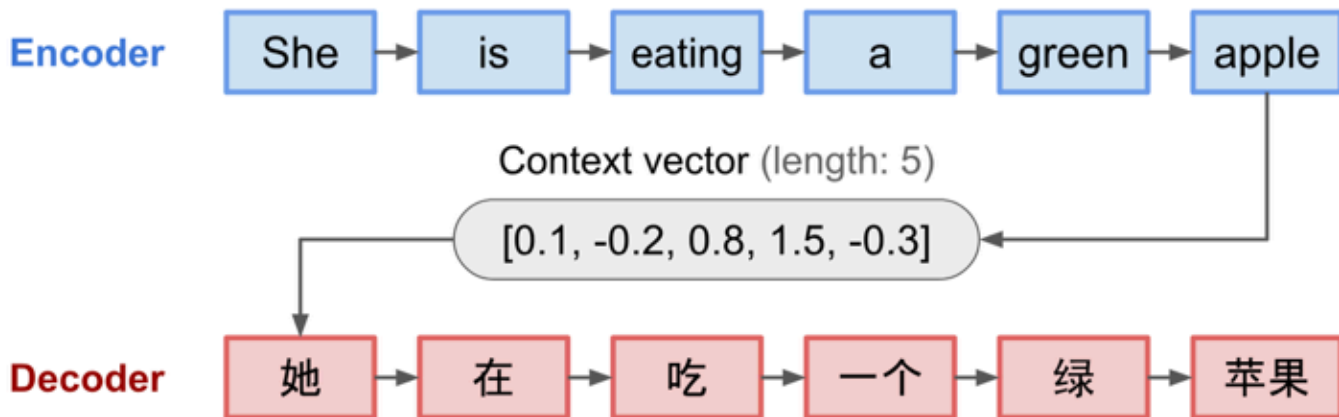


Fig. 3. The encoder-decoder model, translating the sentence "she is eating a green apple" to Chinese. The visualization of both encoder and decoder is unrolled in time.

Sequence to Sequence Models for Machine Translation

- Multi-layered LSTM to summarize input to a vector, v
- Output depends on that vector and the previously generated words

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1})$$

- where $p(y_t | v, y_1, \dots, y_{t-1})$ is computed by a softmax over the vocabulary y
- Beam search used to explore “best” partial translations
- \Rightarrow Part of the revolutionary improvement in MT by Google
- But, some troubling issues:
 - Reversed input to bring *some* x, y pairs closer together because even LSTM “forgets” longer-range dependencies
 - BLEU score (34.8) > that of a phrase-based MT system (33.3)
 - but re-ranking top-1000 outputs of the phrase-based system \Rightarrow 36.5, so, model fails to capture everything important

Adding Attention to the Seq2Seq Model

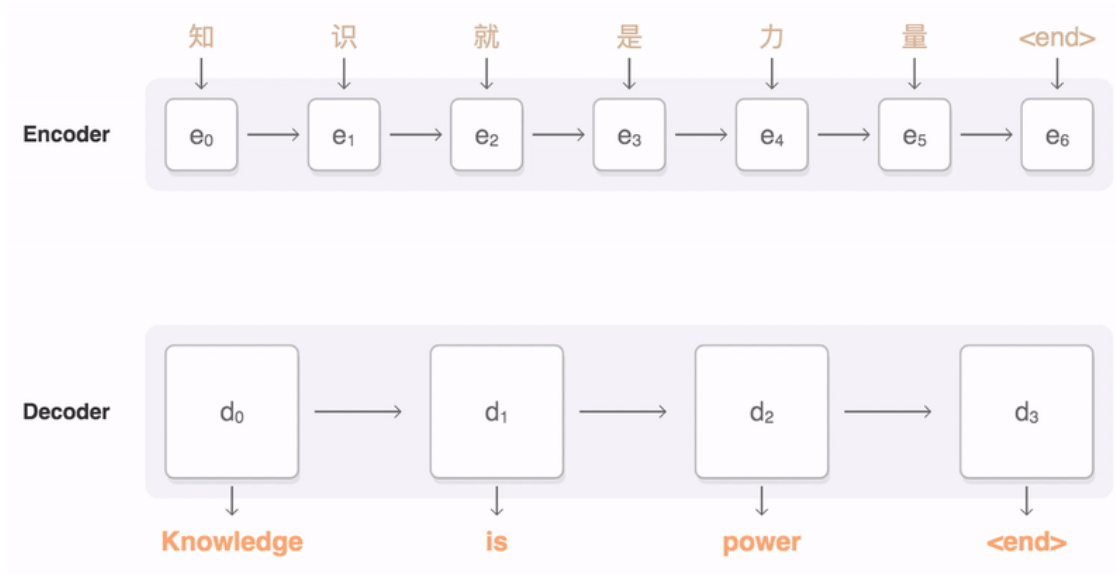
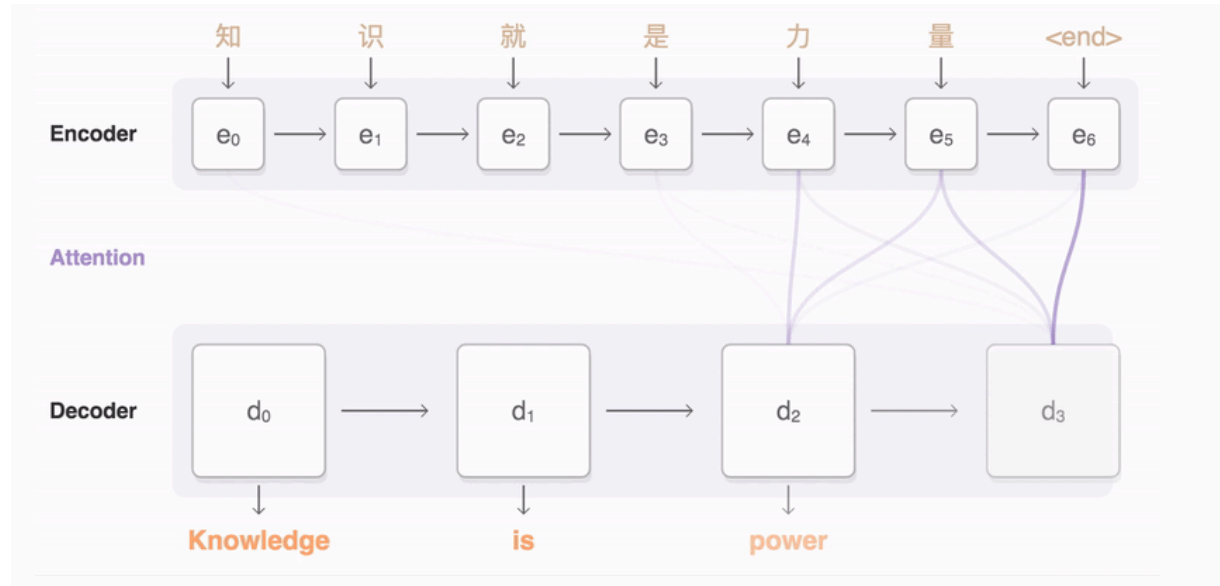


Illustration of Learned Attention Weights



How to Model the Attention Weights

- Attention depends on:
 - score of relationship between word being generated and all input words [“dictionary”]
 - learned positional dependencies [“alignment”]

• State of the decoder $s_t = f(s_{t-1}, y_{t-1}, \mathbf{c}_t)$, $t = 1, \dots, m$ where $\mathbf{c}_t = \sum_{i=1}^n \alpha_{t,i} \mathbf{h}_i$

where \mathbf{h}_i are all the i -th word encoder states and $\alpha_{t,i}$ are the learned alignment weights

- f can take various forms:

$$\cos(s_t, \mathbf{h}_i)$$

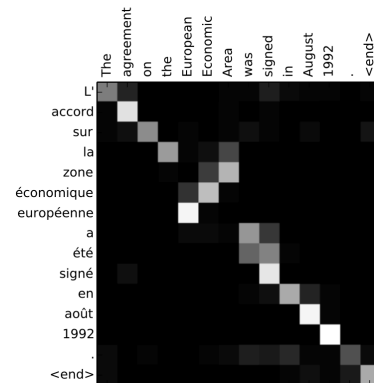
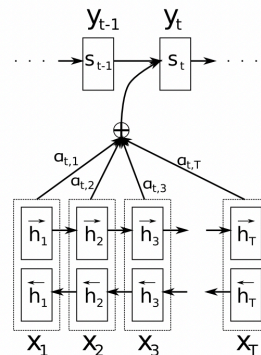
$$s_t^\top \mathbf{W}_a \mathbf{h}_i$$

$$\text{softmax}(\mathbf{W}_a s_t)$$

$$s_t^\top \mathbf{h}_i$$

$$\mathbf{v}_a^\top \tanh(\mathbf{W}_a [s_t; \mathbf{h}_i])$$

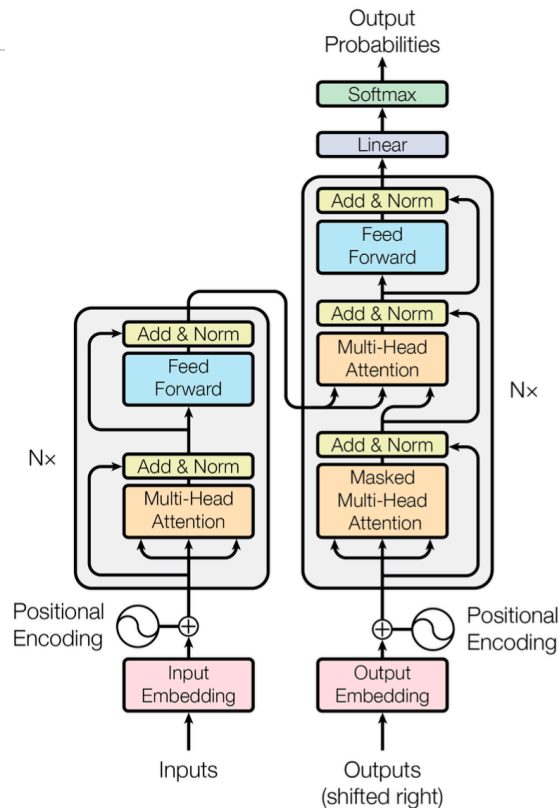
$$s_t^\top \mathbf{h}_i / \sqrt{n}$$



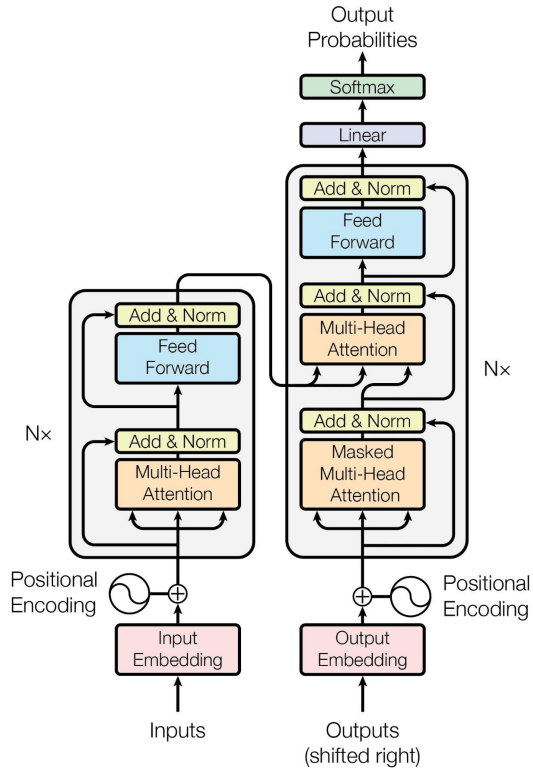
“Attention is All You Need”

(151,961 citations!!!)

- Vastly generalizes how to use context to create word embeddings
 - Derived from work on machine translation and speech understanding
- Task is to predict masked words in text
- Same word gets different embeddings in different contexts
 - E.g., “river bank” vs. “Bank of America”
- Can build arbitrarily complex models applied to vast data sets

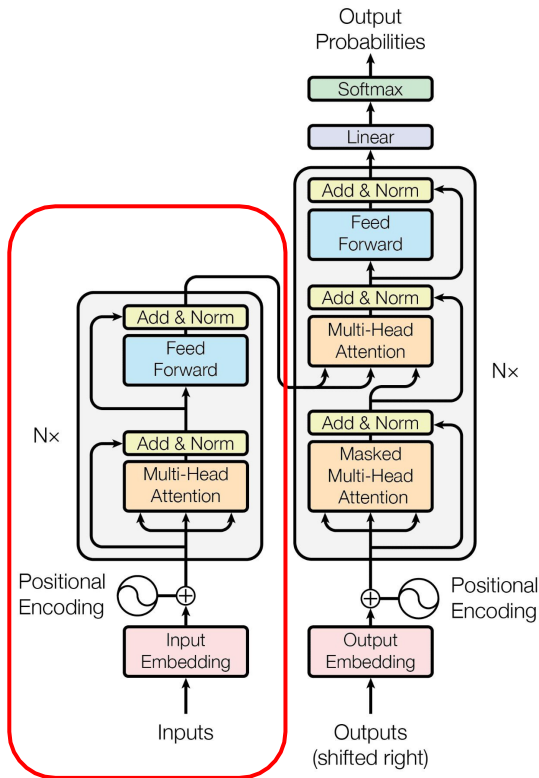


Transformer (2017)

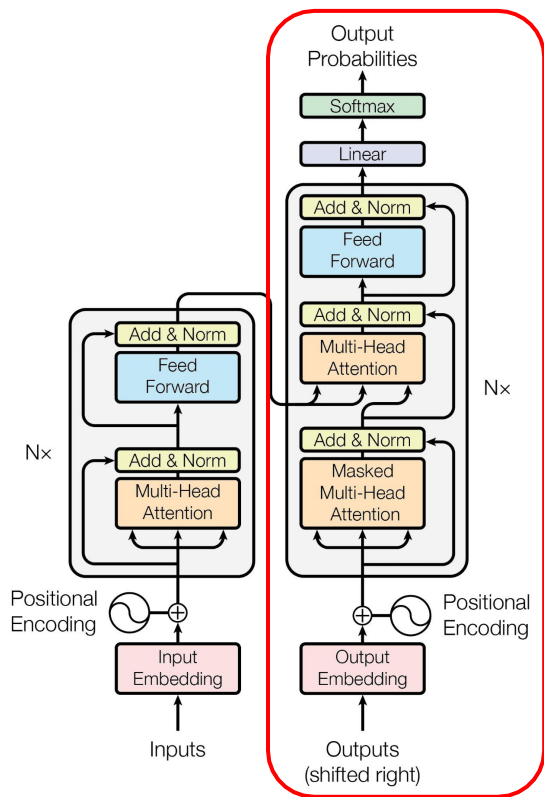


Transformer (2017)

Encoder: reads the entire sequence all at once.



Transformer (2017)

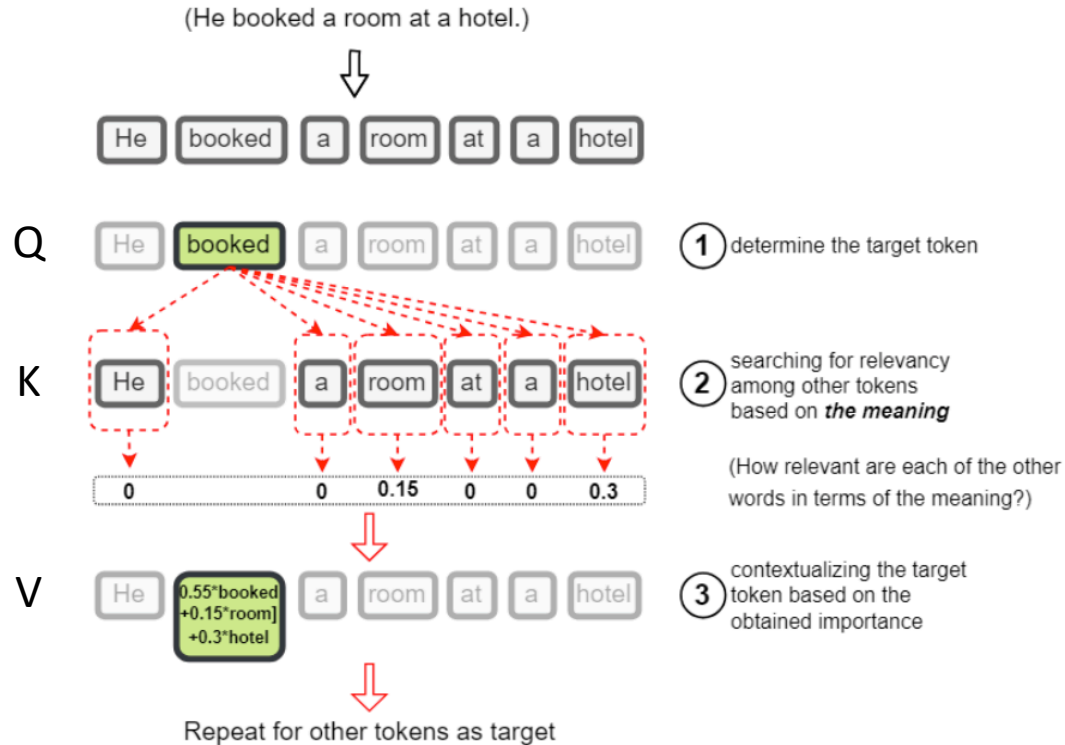


Encoder: reads the entire sequence all at once.

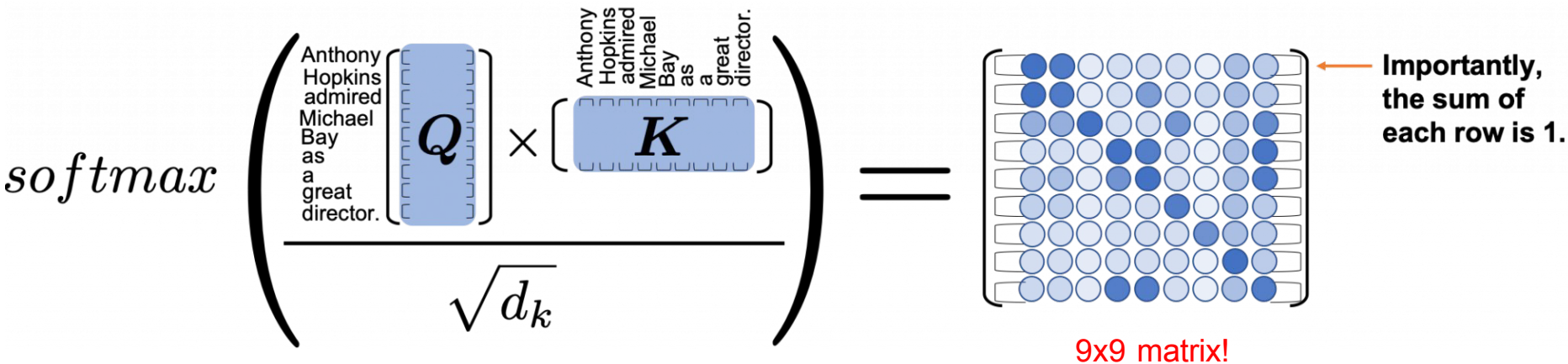
Decoder: reads left to right (but parallelized)

Self-Attention

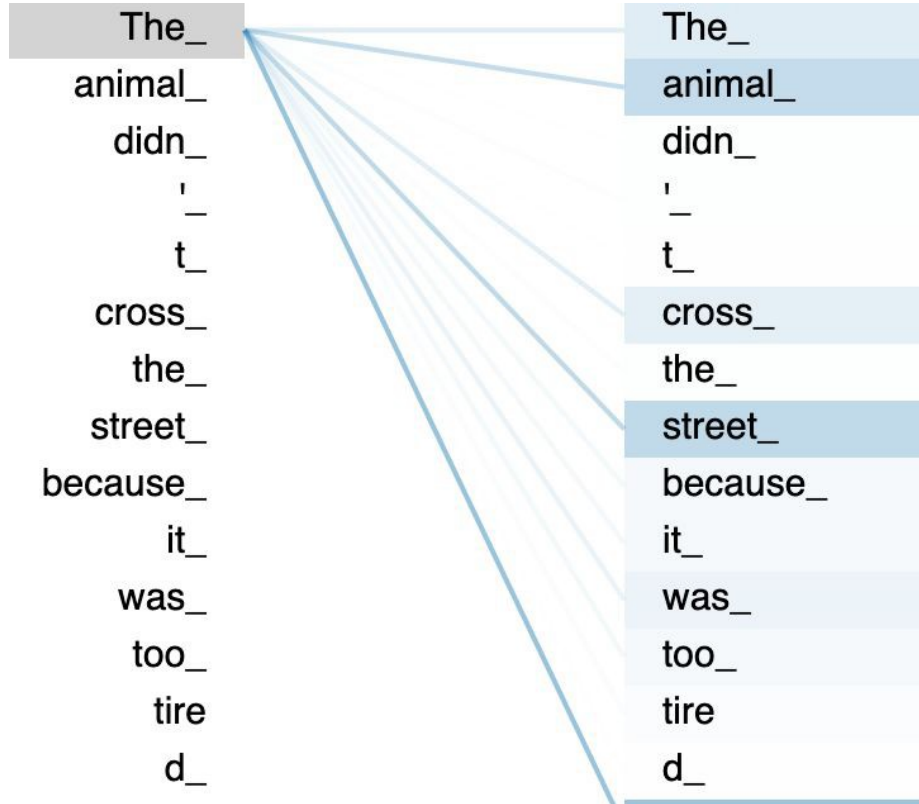
- Q = Query
 - Embedding of text
- K = Keys
 - Embeddings of each other word
- V = Value
 - Component of contribution from word K to contextual embedding of Q



Self-Attention

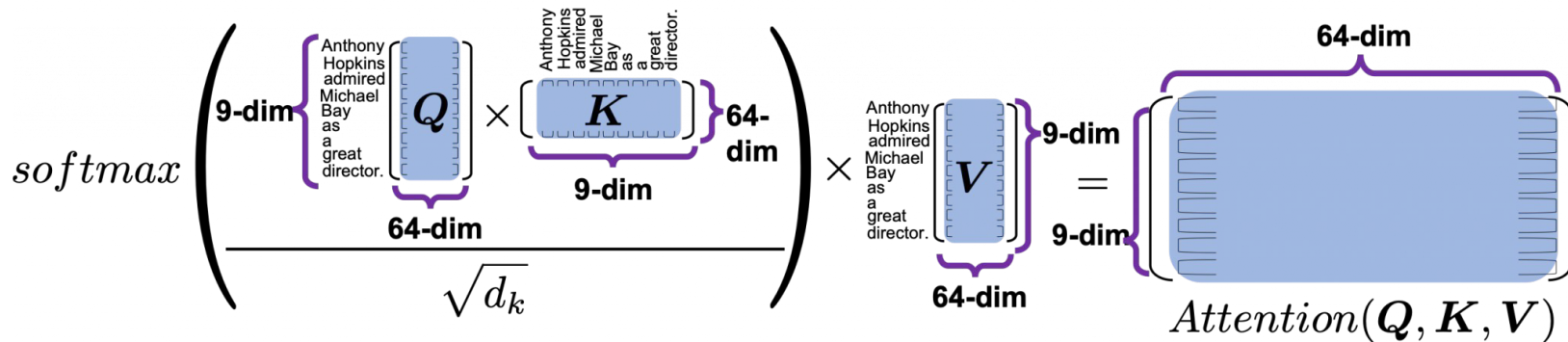


Transformer (2017)

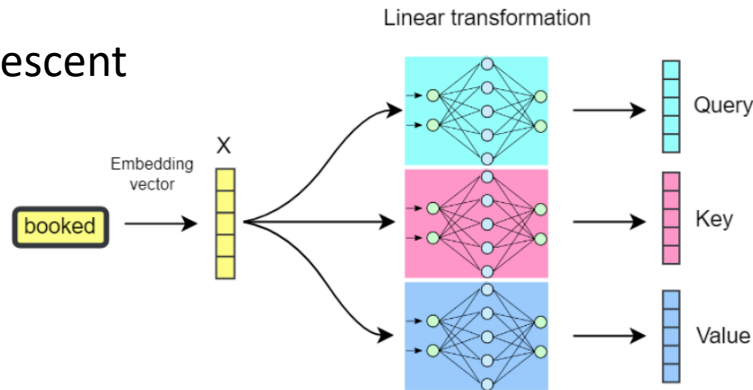


Idea: How important is *this word*, with respect to ALL other words?

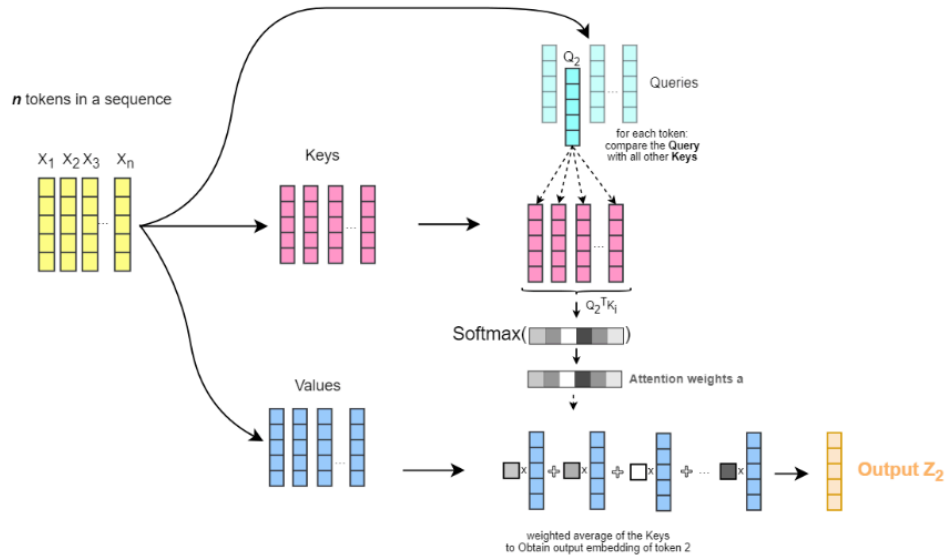
Self-Attention



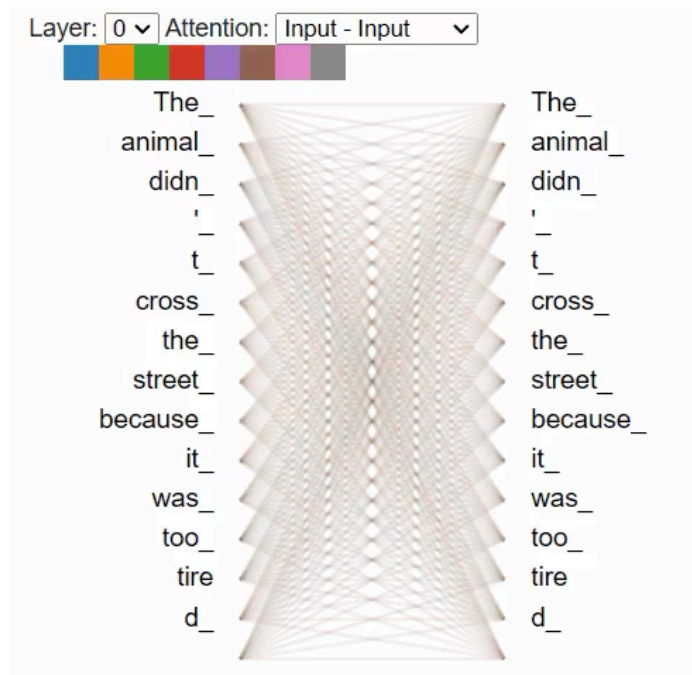
The Miracle of Gradient Descent



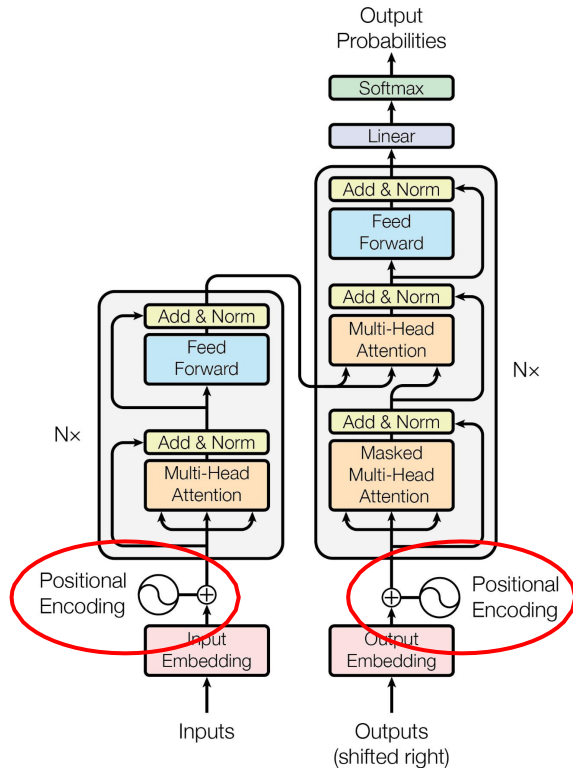
Self-Attention



Example of Attention Among Words in a Sentence

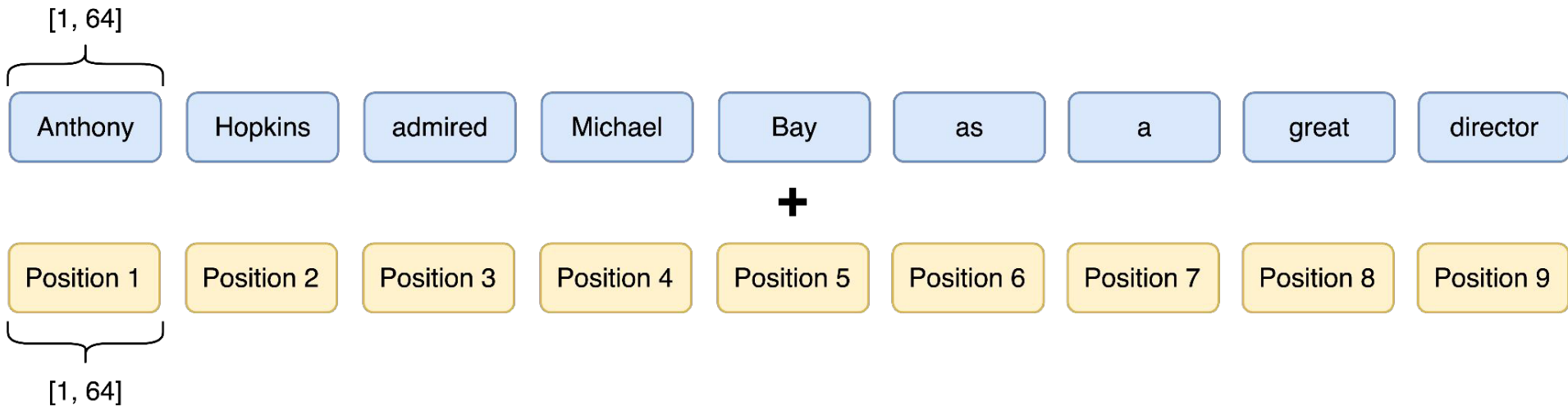


Transformer (2017) – Position Embeddings

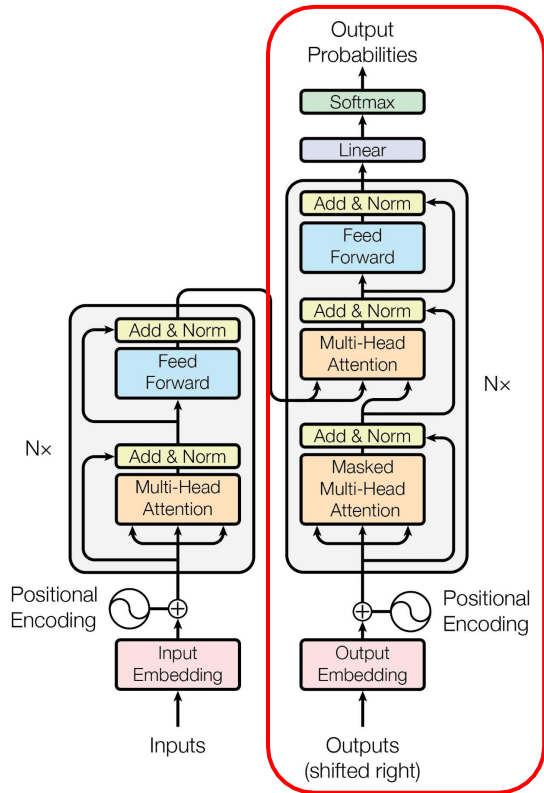


- Every position (i.e., the first, second, third... word) has a unique vector that represents its position in the sentence/paragraph
- These are randomly initialized and learned by the model!

Transformer (2017) – Position Embeddings



Generative Pretrained Transformer (GPT), 2018



- Pretraining using next word prediction on 7000 books (1B words)
- Autoregressive: Reads left-to-right
- 117M parameters

Generative Pretrained Transformer (GPT), 2018

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. OpenAI, 2018.
Available from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

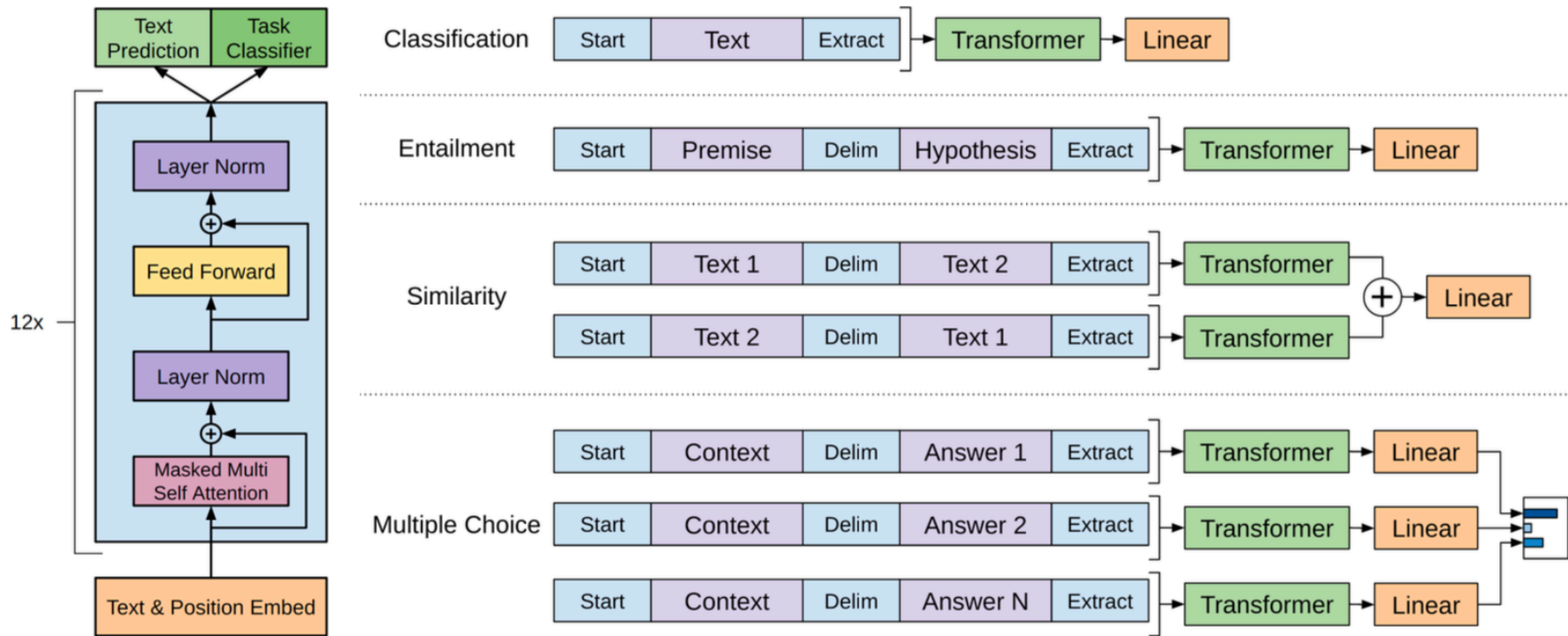
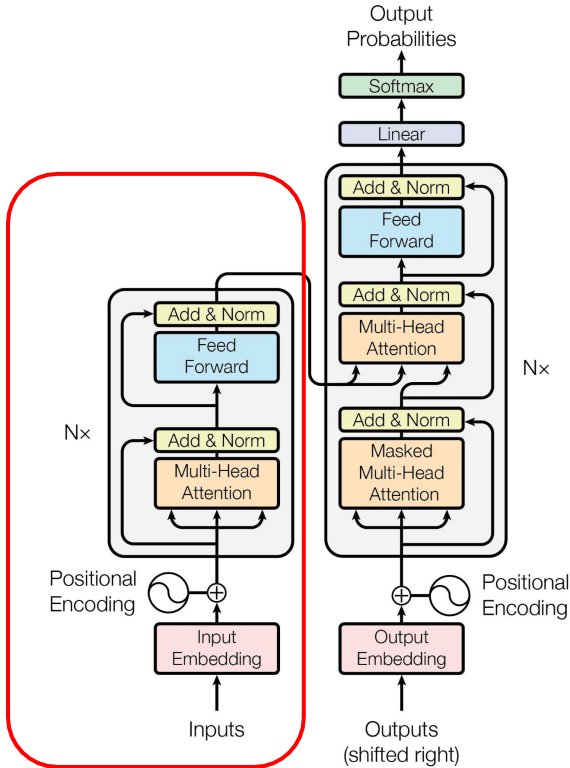


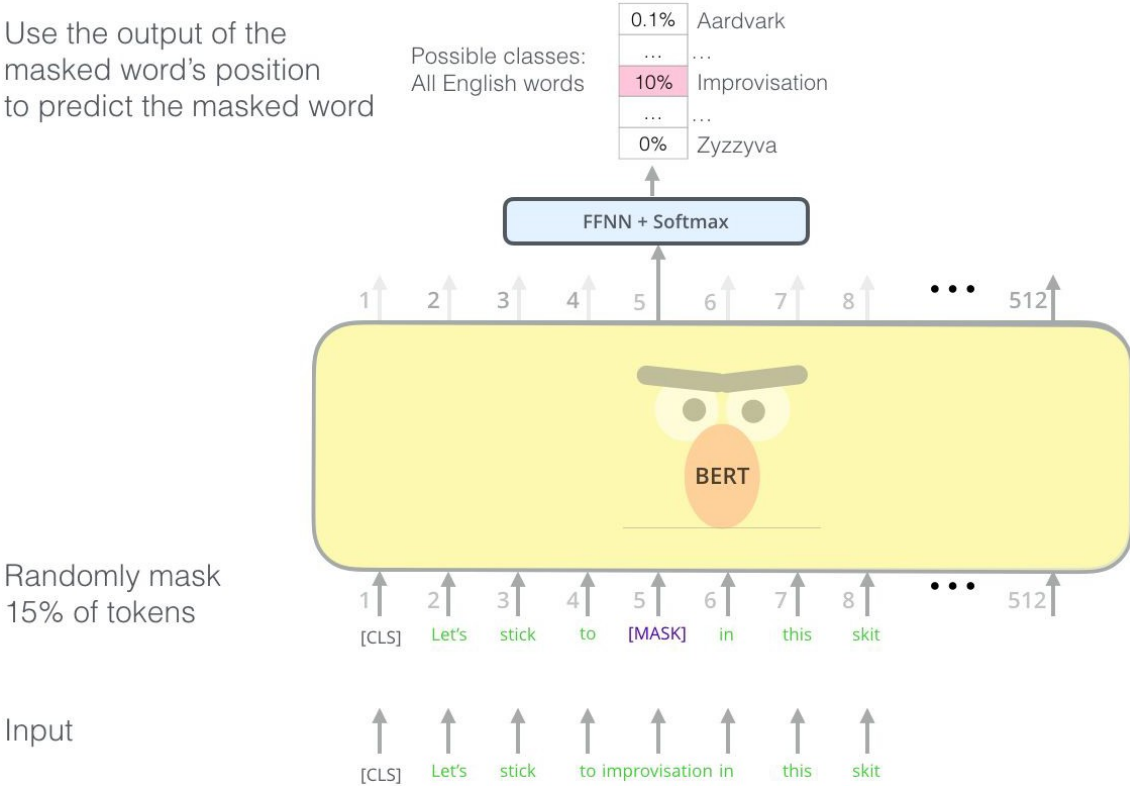
Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Bidirectional Encoder Representations from Transformers (BERT)



BERT + Masked Language Modeling

Use the output of the masked word's position to predict the masked word



- Pretrained on ~3.4B words for 40 epochs
- 110M and 345M parameter models

BERT Performance

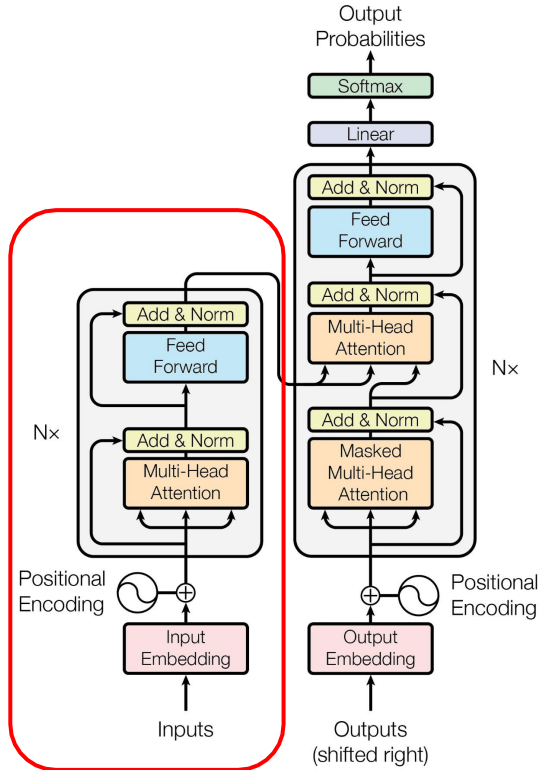
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}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

GPT-2 (2019)

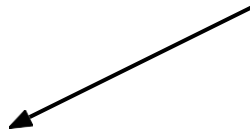
- Same pretraining task: next word prediction
- Pretrain on more data (40GB of text)
- Use bigger models

	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

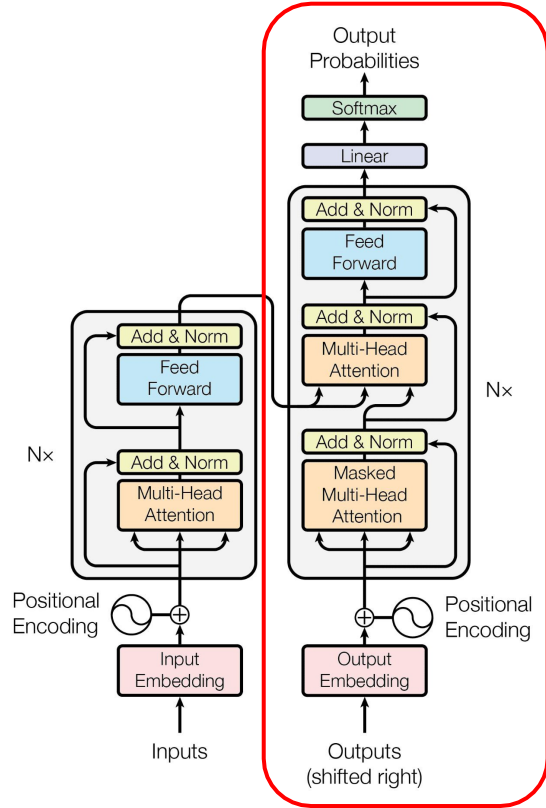
Text-to-Text Transfer Transformer (T5) (2020)



BERT uses the first part
(called encoder)



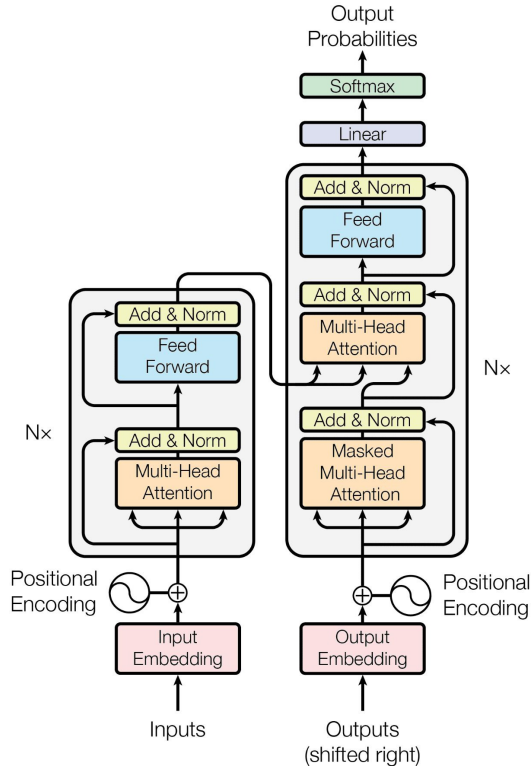
Text-to-Text Transfer Transformer (T5) (2020)



GPT uses the second part
(called decoder)

Text-to-Text Transfer Transformer (T5) (2020)

T5 uses the entire architecture.



T5 Performance

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	39.77	24.04

How Much Unique Text to Pretrain On?

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	83.28	19.24	80.88	71.36	26.98	39.82	27.65
2^{29}	64	82.87	19.19	80.97	72.03	26.83	39.74	27.63
2^{27}	256	82.62	19.20	79.78	69.97	27.02	39.71	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

Scale Scale Scale

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4 ^a	69.2 ^b	97.1 ^a	93.6^b	91.5^b	92.7 ^b	92.3 ^b
T5-Small (80M)	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base (220M)	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large (770M)	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	90.3	71.6	97.5	92.8	90.4	93.1	92.8

Enter... GPT-3... (2020)

- 175B parameter model trained on 300B words
- This model is REALLY good at next word prediction
- Do we still need to train models?

Enter... GPT-3... (2020)

You can frame any NLP task as a next word completion task.

- **Mortality Prediction:** “Based on the above note, do you think that the patient will die?”
- **De-identification:** “List all of the names mentioned in the note:”

Enter... GPT-3... (2020)

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	44.5	45.5	68.0
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	68.0
GPT-3 Few-Shot	29.9	41.5	71.2

GPT-3 Can Resolve Clinical Acronyms

Input: Bob was sent to IR for thrombolysis. Post IR, ultrasound showed that... What does IR stand for?

Output: Interventional radiology

GPT-3 Can Resolve Clinical Acronyms

Algorithm	CASI Acc.	CASI Macro F1	MIMIC Accuracy	MIMIC Macro F1
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	0.69
<i>GPT-3 edit</i> + R: 0-shot	0.86	0.69	*	*
<i>GPT-3 edit</i> + R: 0-shot + distillation	0.90	0.76	0.78	0.69

Agrawal et. al (2022) showed that GPT-3 had strong performance on a number of clinical extraction tasks.

How You Ask Matters!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

How You Ask Matters!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

How You Ask Matters!

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

How You Ask Matters!

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

Jeanne's neighbor rides 8 kilometers to the bus station every day. How many more tickets should Jeanne buy?

Standard Answer

8

How You Ask Matters!

Solve the following math problem. Feel free to ignore irrelevant information in the given problems.

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

Jeanne's neighbor rides 8 kilometers to the bus station every day. How many more tickets should Jeanne buy?

How You Ask Matters!

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

Question: How do you treat skin redness?

Complete Answer: It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

InstructGPT (2022)

Forcing your model to read isn't the only way to improve performance!

BUSINESS • TECHNOLOGY

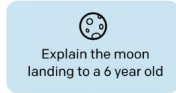
**Exclusive: OpenAI Used Kenyan Workers on
Less Than \$2 Per Hour to Make ChatGPT Less
Toxic**

InstructGPT (2022)

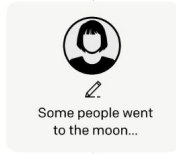
Step 1

**Collect demonstration data,
and train a supervised policy.**

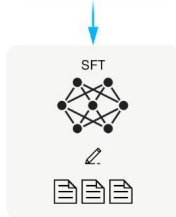
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



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

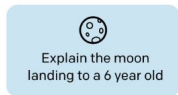


InstructGPT (2022)

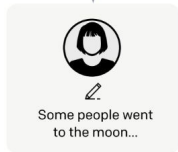
Step 1

**Collect demonstration data,
and train a supervised policy.**

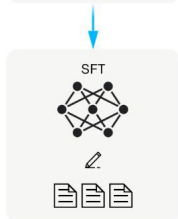
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



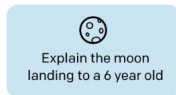
This data is used
to fine-tune GPT-3
with supervised
learning.



Step 2

**Collect comparison data,
and train a reward model.**

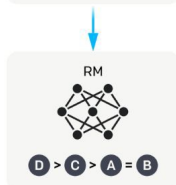
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.

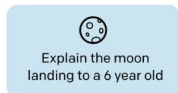


InstructGPT (2022)

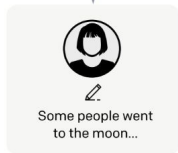
Step 1

Collect demonstration data, and train a supervised policy.

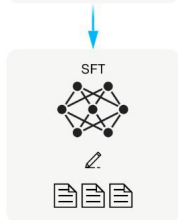
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



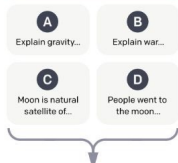
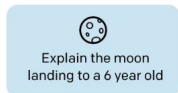
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

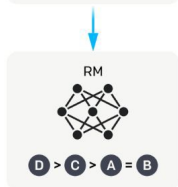
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



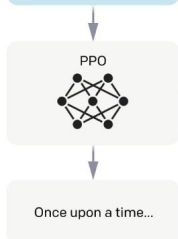
Step 3

Optimize a policy against the reward model using reinforcement learning.

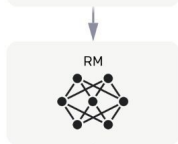
A new prompt is sampled from the dataset.



The policy generates an output.



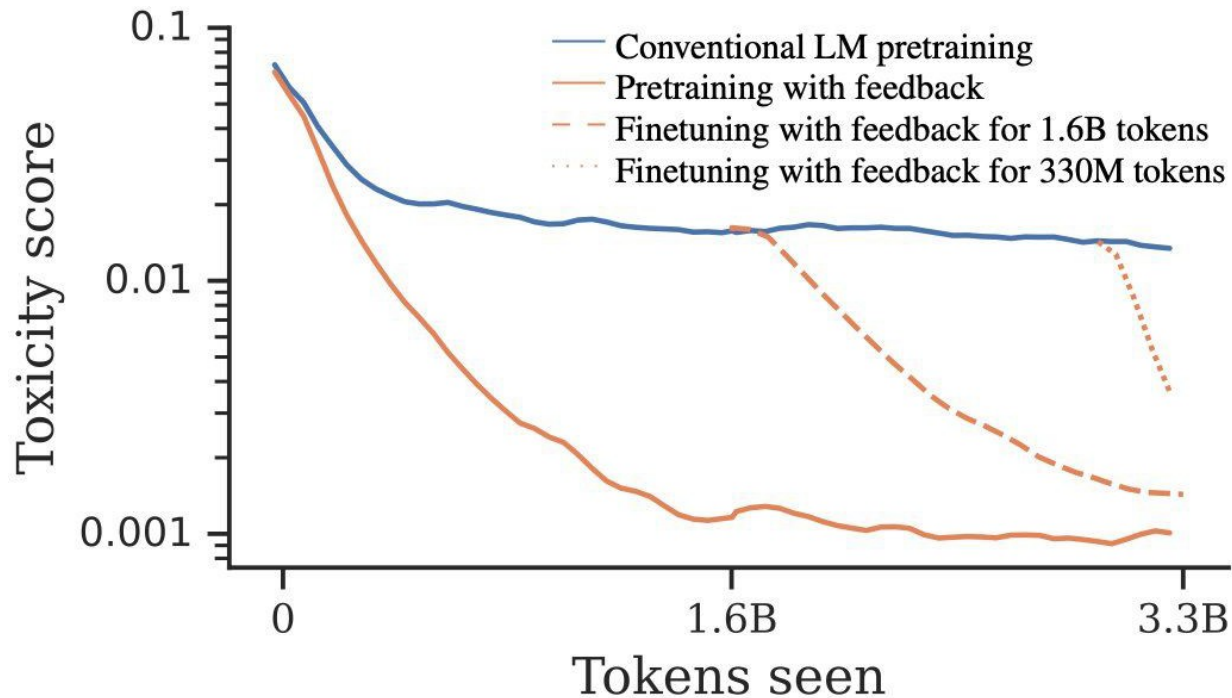
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



InstructGPT (2022)



What Does This Mean For Healthcare?

- Is it still worth working on these problems?
- Is it still worth creating custom models for clinical text?

Do We Still Need Clinical Language Models?

Eric Lehman^{1,2} Evan Hernandez^{1,2} Diwakar Mahajan³ Jonas Wulff²
Micah J. Smith² Zachary Ziegler² Daniel Nadler² Peter Szolovits¹
Alistair Johnson⁴ Emily Alsentzer^{5,6}
¹MIT ²Xyla ³IBM Research ⁴The Hospital for Sick Children
⁵Brigham and Women's Hospital ⁶Harvard Medical School
{lehmer16, dez}@mit.edu

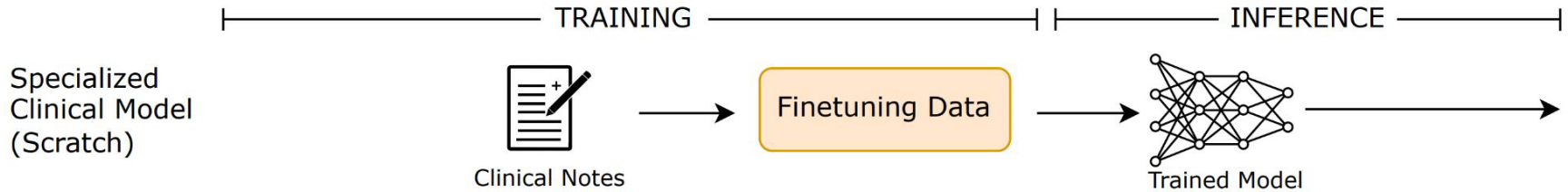
Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, et al. Do We Still Need Clinical Language Models? In: CHIL 2023. New York, NY, USA: arXiv; 2023. Available from: <http://arxiv.org/abs/2302.08091>

Working With Clinical Text is Different

- Safety matters *a lot!*
- Models trained on the general web are likely biased
- The text from a note is very different
 - Random abbreviations
 - Incorrect grammar
 - Medical Terms
- You can't find this text online!!!

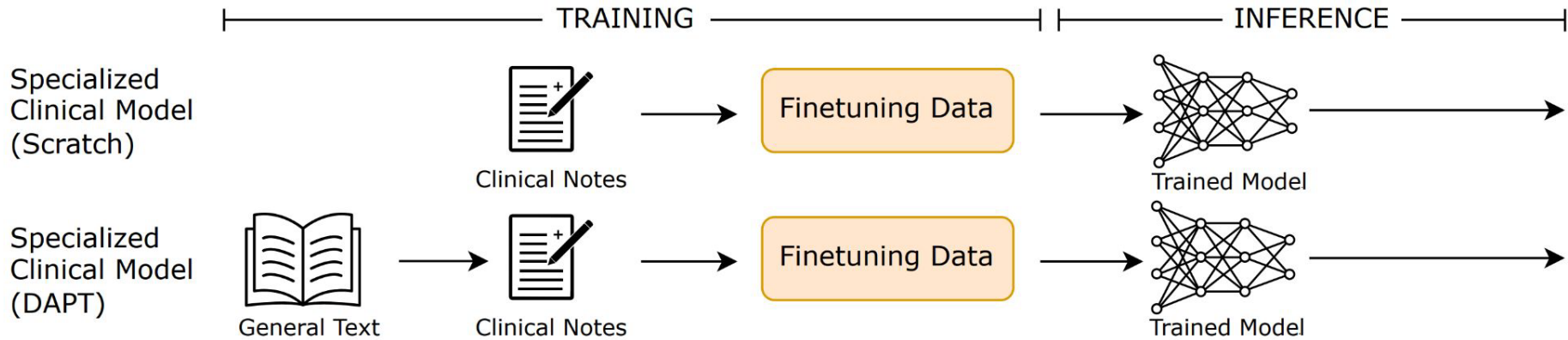
If You're a Hospital, What Do You Do?

- We saw this with ClinicalBERT and it has been shown with other models as well



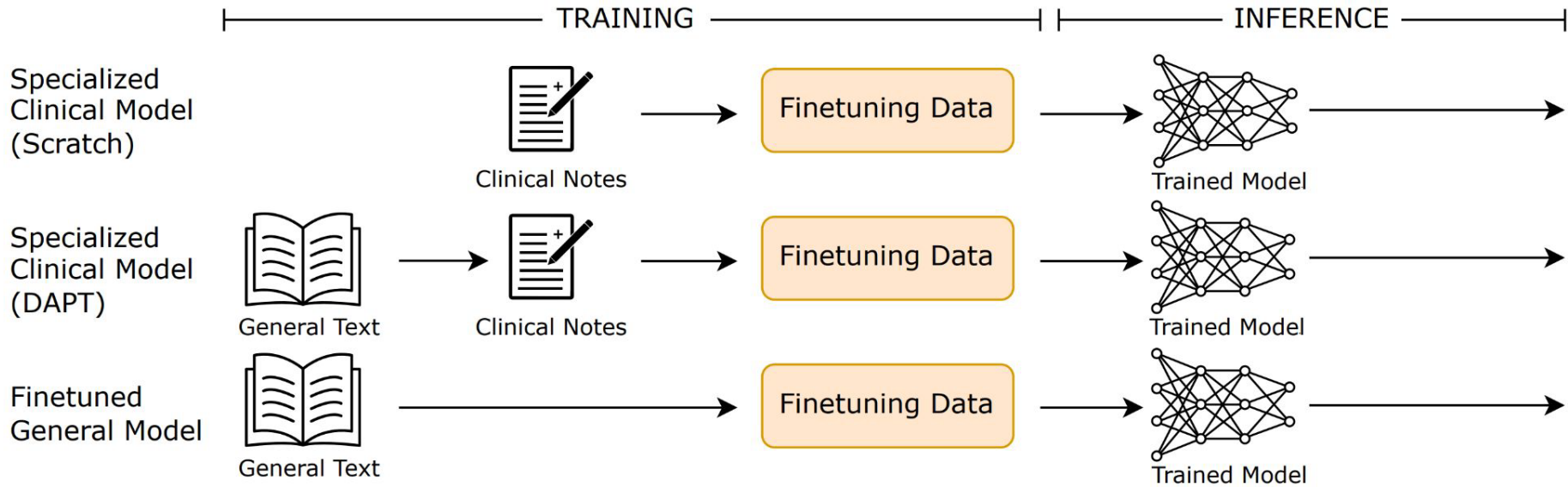
Option 1: Randomly initialize your model and train your own language model ON clinical notes from scratch

If You're a Hospital, What Do You Do?



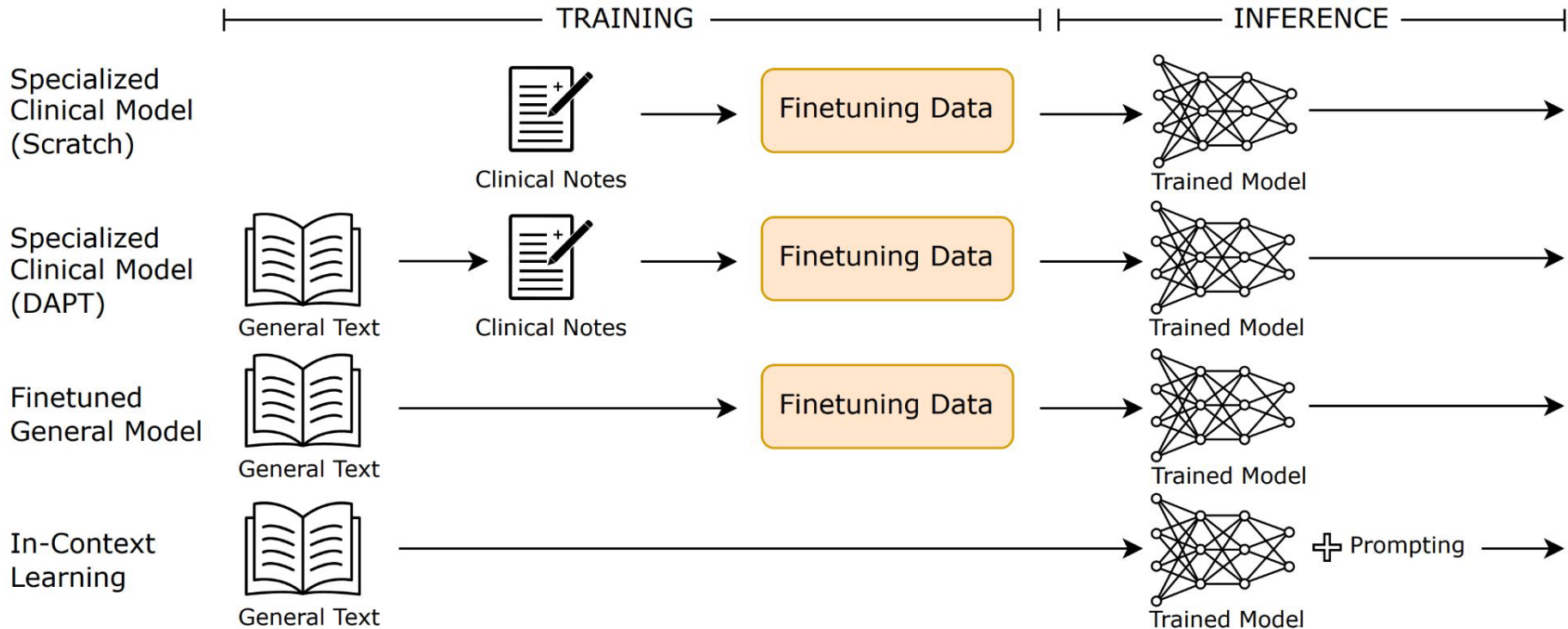
Option 2: Initialize your model from a model trained on the general web, and further train it on clinical notes.

If You're a Hospital, What Do You Do?



Option 3: Download some model trained on the internet and finetune it.

If You're a Hospital, What Do You Do?



PERSPECTIVE

Large Language Models Seem Miraculous, but Science Abhors MiraclesPeter Szolovits , Ph.D.¹

Received: August 28, 2023; Revised: December 18, 2023; Accepted: January 2, 2024; Published: May 9, 2024

- Why does a simple training method on vast amounts of human-created text exhibit skills that it was not explicitly trained to do?
- Why do models trained this way nevertheless “hallucinate”?
- What does this experience tell us about human thought?
- How to trade off domain-specificity vs. size of model and training data?
- Can we make knowledge explicit (in symbolic form)?
 - As some abstraction of transformer models over the raw training data?
- Most current research focuses on how to use generative AI models to improve applications
 - but, is it safe to do so until we better understand answers to the above questions?
- Should we treat an engineered artifact as if it were a natural phenomenon?
 - Experiment and form hypotheses, vs. analyze the design
 - Perhaps too complex for engineering analysis

Christopher Y.K. Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, Atul J. Butte, Aaron E. Kornblith

doi: <https://doi.org/10.1101/2024.04.03.24305088>

- Used GPT-4 and GPT-3.5-turbo to summarize 100 ED notes
 - “You are an Emergency Department physician. Below is the History and Physical Examination note for a patient presenting to the Emergency Department who was subsequently discharged. Write a discharge summary for the patient based on this note. Do not include any additional information not present in the note. \n\n "" Note text "" ”
- Human evaluation to identify (a) inaccuracies, (b) hallucinations, (c) omissions
 - Presenting complaint; History of presenting complaint; Past medical history; Allergies/contraindications; Review of systems; Positive examination findings; Laboratory test results; Radiological investigations; Plan; Other notable events during ED stay (if any)

	error-free	inaccuracies	hallucinations	omissions
GPT-4	33%	10%	42%	47%
GPT-3.5-turbo	10%	36%	64%	50%