APRIL 2024 BUSINESS MEETINGS

APRIL 8-17, 2024 | LONDON & ONLINE

OPENING KEYNOTE SPEAKER:

ANTHONY SHEK, PhD

Lead Data Scientist | NLP Engineer |
Guys' and St Thomas' NHS Foundation Trust |
UK

linkedin.com/company/ihtsdo

X.com/SnomedCT

youtube.com/@snomedct

snomed.org

Bridging the Gap: Advancements in NLP and LLM for Clinical Integration with SNOMED CT

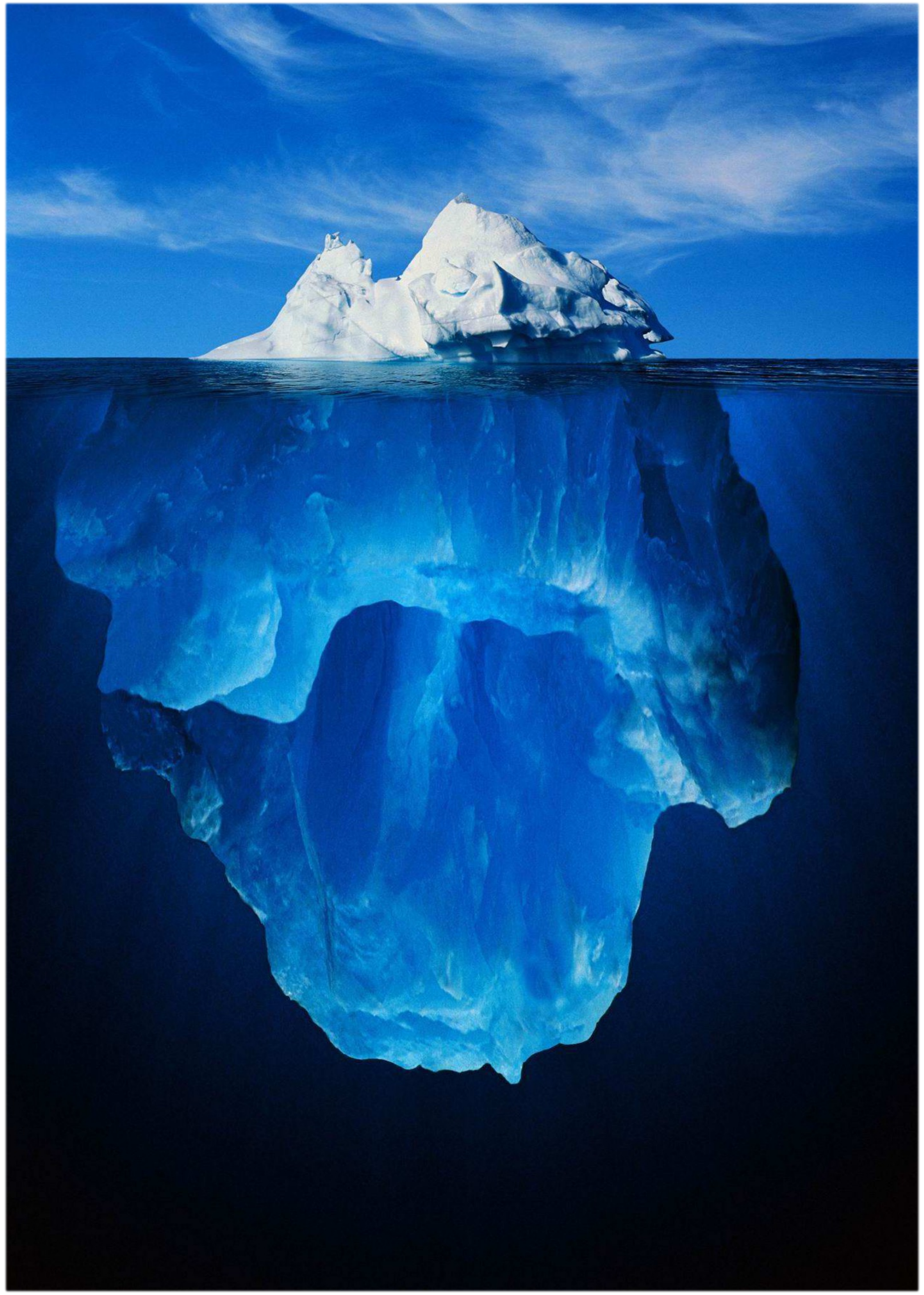
DATE: APRIL 15, 2024
TIME: 09:00-10:00 BST

SNOMED International

Delivering SNOMED CT

Background

- Every activity in a hospital generates data
- Millions electronic patient records documents per hospital
- **80% of information is unstructured** as it is the most natural way to record doctor-patient interactions.



What's the Problem? – Accurate Extraction from Text

Electronic Patient Record (EPR)

NHS
Guy's and St Thomas'
NHS Foundation Trust

Clinic Letter Extract Example

Dear Mr. TGF:

Mr. XY is a client of our Epilepsy clinic.
He suffers from juvenile myoclonic epilepsy and was medically screened by you as a driver.

Your assessment of a diagnosise of jme has resulted in XY being denied the job and we need to point out the following:

DHX:
allergic to chlorquine
oxICarbaZEpine 1200 mg
keppra 1500 mg BID

EPR Characteristics:

- Records clinically valuable information.
- Unstructured – no data standardization requirements
- Difficult to extract information automatically

What's the Problem? – Accurate Extraction from Text

Electronic Patient Record (EPR)

NHS
Guy's and St Thomas'
NHS Foundation Trust

Clinic Letter Extract Example

Dear Mr. TGF:

Mr. XY is a client of our **Epilepsy** clinic.
He suffers from juvenile myoclonic **Epilepsy** and was medically screened by you as a driver.

Your assessment of a diagnosise of jme has resulted in XY being denied the job and we need to point out the following:

DHX:
allergic to chlorquine
oxICarbaZEpine 1200 mg
keppra 1500 mg BID

EPR Characteristics:

- Records clinically valuable information.
- Unstructured – no data standardization requirements
- Difficult to extract information automatically

Q) Extract all mentions of an Epilepsy diagnosis?

"Epilepsy" Keyword search is not ideal

What's the Problem? – Accurate Extraction from Text

Electronic Patient Record (EPR)

NHS
Guy's and St Thomas'
NHS Foundation Trust

Clinic Letter Extract Example

Dear Mr. TGF:

Mr. XY is a client of our Epilepsy clinic.
He suffers from **juvenile myoclonic epilepsy** and was medically screened by you as a driver.

Your assessment of a diagnosise of **jme** has resulted in XY being denied the job and we need to point out the following:

DHX:
allergic to chlorquine
oxICarbaZEpine 1200 mg
keppra 1500 mg BID

EPR Characteristics:

- Records clinically valuable information.
- Unstructured – no data standardization requirements
- Difficult to extract information automatically

Q) Extract all mentions of an Epilepsy diagnosis?

- Ideally you would want a **context dependent extraction** of Epilepsy-related terms.
- Manual data structuring requires **huge** amount of work!!!

Structured Data and Standards

Benefits of Structured Data

- Easier to **analyze** and **create reports** (both clinical and administrative tasks):
  - Audits
  - Quality Reporting and Performance Metrics
  - Clinical Decision Support
  - Patient Safety and Continuity of Care
  - Population Health Management
  - Research
- Interoperability and Information Sharing
  - **Structured** data formats and **standards** enable seamless interoperability between different healthcare systems and platforms, promoting efficient information exchange among healthcare providers, researchers, and other stakeholders.

Hype – AI now Reads and Generates Text

- Generative AI is a type of **artificial intelligence** that can produce new data, images, text, or music resembling the dataset it was trained on.
- The new language models dazzle us with generation
- But these Large Language Models (LLMs) equally **summarize, simplify, organize, analyze, compare**.

Claude

Gemini

Talk to Pi,
your *personal* AI.

Pi

pi.ai

Solution: Language Models

- The field of **Natural Language Processing** – how computers process and analyze natural language / free text information.

To understand and write text, LLMs must first translate words into a language that they understand.

“They had a seizure in the morning”

- AI has increased the quantity of usable data 100x.

How it Works

- A "Token" is a unit of text that a language model processes. These can be word, subword, or character based on how the text is segmented or tokenized.

First a block of words is broken into "tokens"

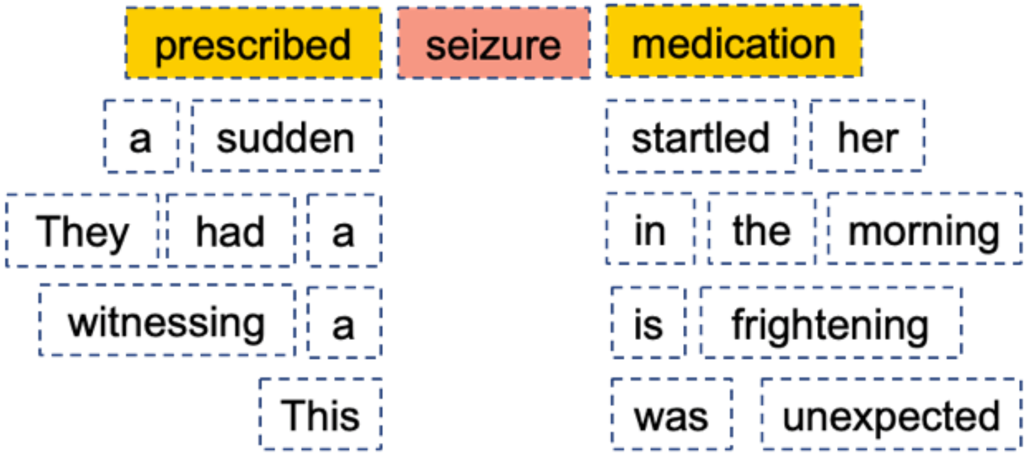
They	had	a	seizure	in	the	morning

- This process is called **<u>Tokenization</u>**
- The component to do it is called a Tokenizer

How it Works

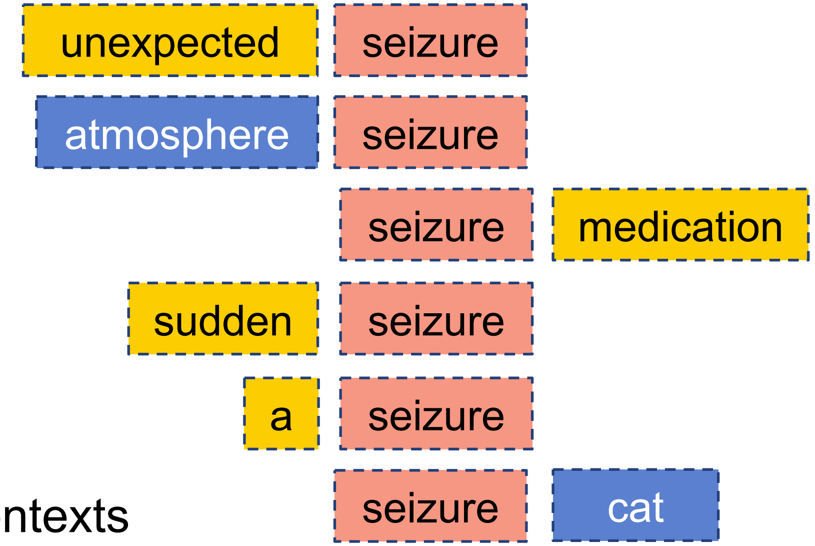
- “Large” language models (LLMs) are called so because they are built on a vast scale with significant number of parameters and data

To grasp a word’s meaning, seizure in our example, LLMs first observe it in context using enormous sets of training data, taking note of nearby words.



How it Works

We end up with a huge set of words found alongside seizure in the training data, as well as those that were distant



- The assumption is that similar words are used in similar contexts

Examples: The patient collapsed because they had a ”seizure”

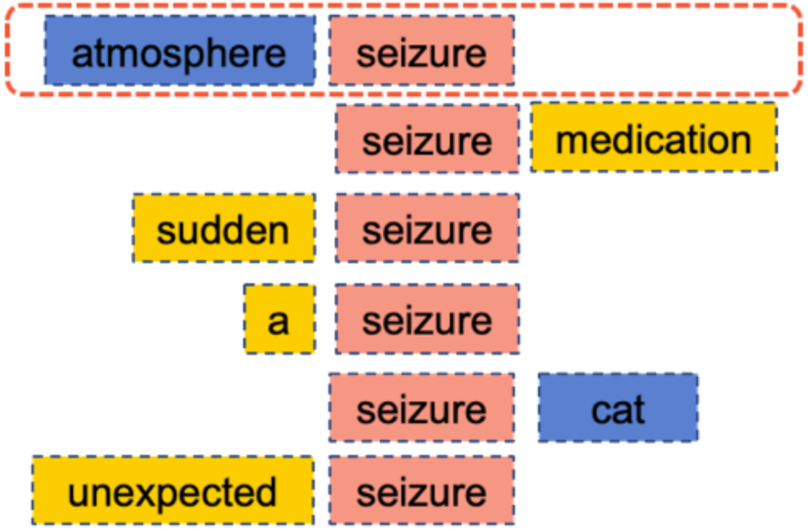
The patient collapsed because they had a ”fit”

A diagnosis of epilepsy does not ”seizure”

How it Works

- A vector representation or embedding is **<u>how machines understand language</u>**

As the model processes this set of words, it produces a "**vector**" and adjusts it based on each word's proximity to seizure in the training data.

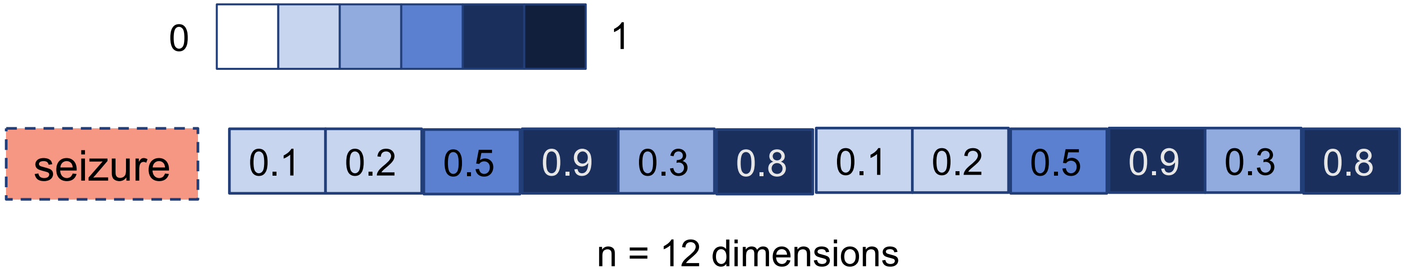


Vector representation (embedding)

How it Works

- Much like you would describe a seizure by its various characteristics—duration, type, triggers, symptoms—the values within the embedding quantify the linguistic features associated with the word seizure

A word embedding can consist of hundreds of values, with each value representing a distinct aspect of the word "seizure's" meaning or context.



- Embeddings lengths can vary but are usually 300, 768 or 1024 dimensions
- Trade off between lack expressiveness/representation (small) and computational considerations (large)

How it Works

The way that these characteristics are derived means we don't know exactly what each value represents, but words we expect to be used in comparable ways often have similar embeddings.

Embeddings

seizure	0.1	0.2	0.5	0.9	0.3	0.8	0.1	0.2	0.5	0.9	0.3	0.8
epilepsy	0.2	0.1	0.9	0.9	0.3	0.8	0.2	0.2	0.2	0.8	0.5	0.8
brain	0.1	0.2	0.8	0.4	0.3	0.3	0.5	0.2	0.5	0.7	0.3	0.8
MRI	0.8	0.9	0.8	0.9	0.4	0.3	0.5	0.2	0.5	0.9	0.3	0.9
dialysis	0.8	0.9	0.8	0.9	0.3	0.1	0.1	0.8	0.5	0.9	0.3	0.1
kidney	0.1	0.2	0.5	0.3	0.3	0.3	0.1	0.8	0.2	0.9	0.3	0.1
pain	0.1	0.2	0.5	0.9	0.8	0.3	0.8	0.6	0.5	0.9	0.3	0.4
trauma	0.4	0.2	0.5	0.9	0.9	0.3	0.7	0.5	0.5	0.9	0.3	0.6

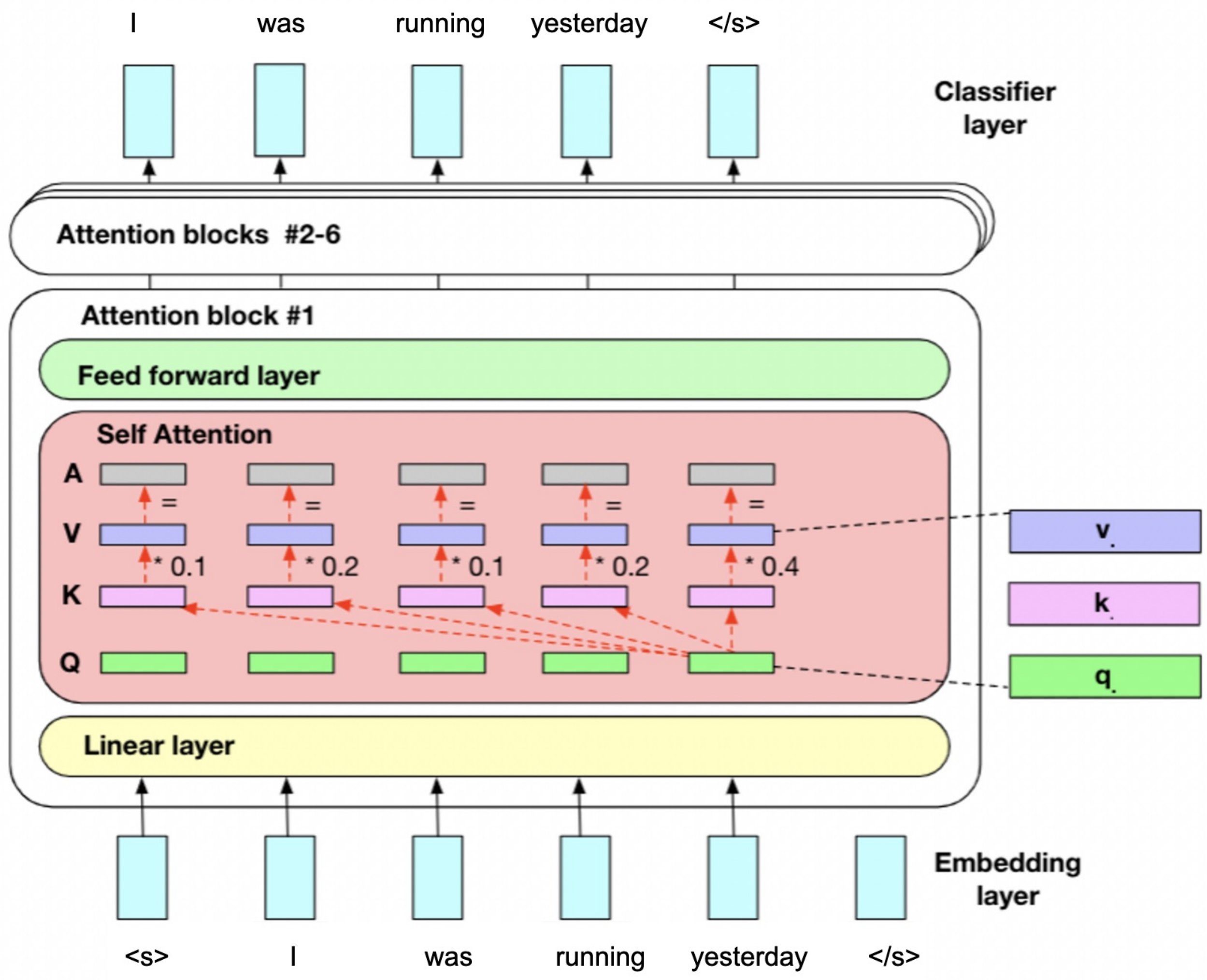
0 1

How it Works – Transformers

Transformer: A deep learning architecture

The crucial building block behind all LLMs

A key concept of the transformer architecture is **Attention**. This is what allows LLMs to understand relationships between words

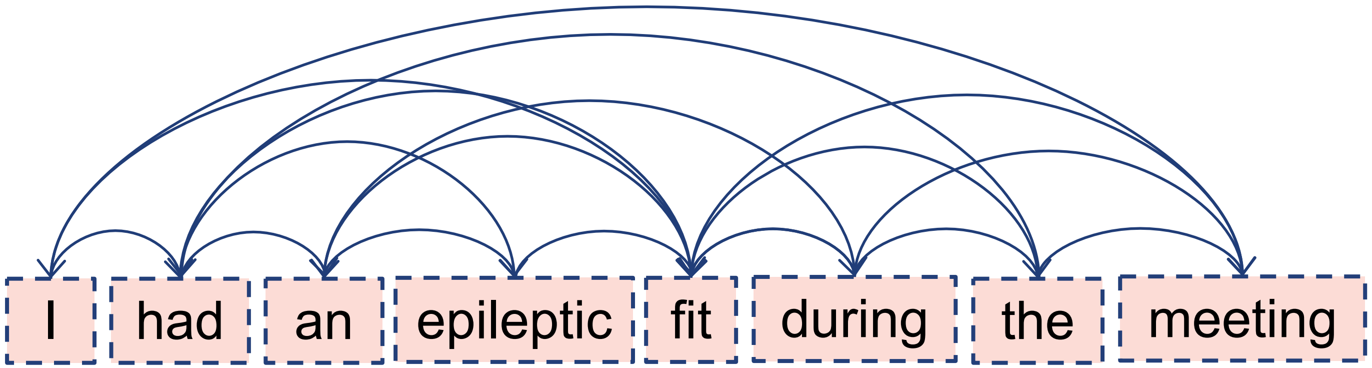


How it Works

Recurrent Neural Networks (RNNs) scanned each word in a sentence and processed it sequentially.

- This isn't always appropriate

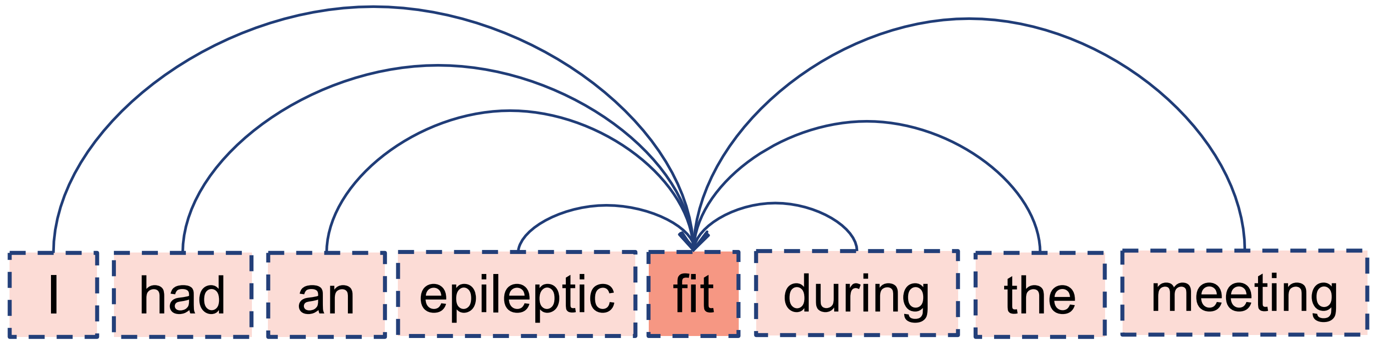
Attention, the transformer computes all words at the same time. Capturing more context giving the LLMs far more sophisticated capabilities to understand language



How it Works

- Attention is all You Need (Vaswani et al., 2017)

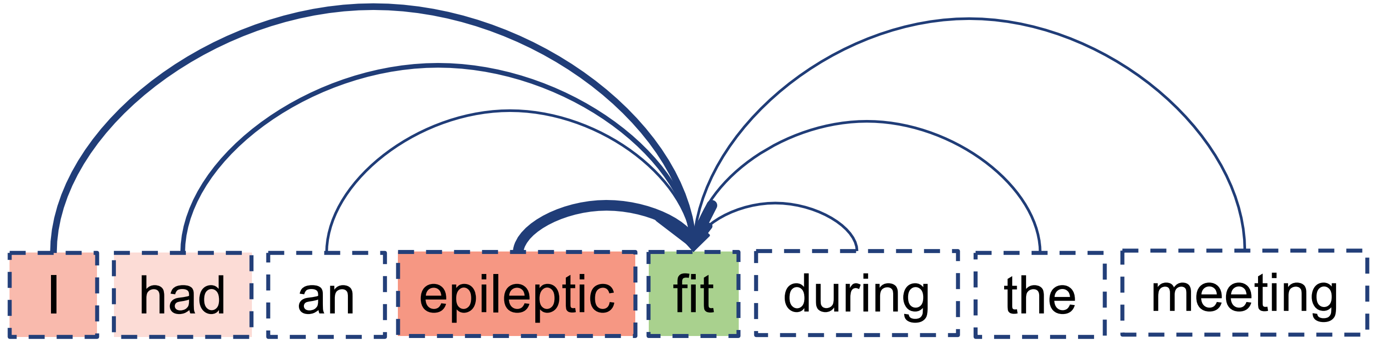
Attention looks at each token in a body of text and decides which others are most important to understanding its meaning



How it Works

- Attention is all You Need (Vaswani et al., 2017)

The **attention** mechanism allows the model to **weigh** the importance of different tokens

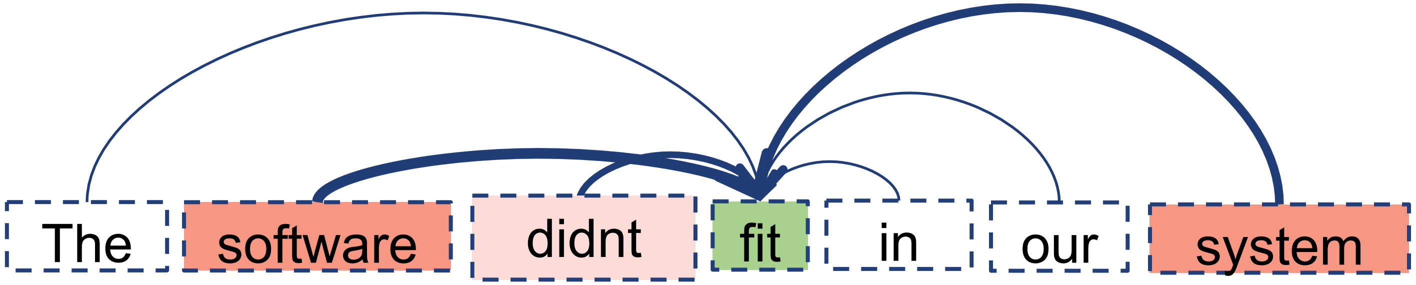


- Attention → focus on specific parts of the sequence when making predictions, enabling it to capture contextual relationships.

How it Works

- Attention is all You Need (Vaswani et al., 2017)

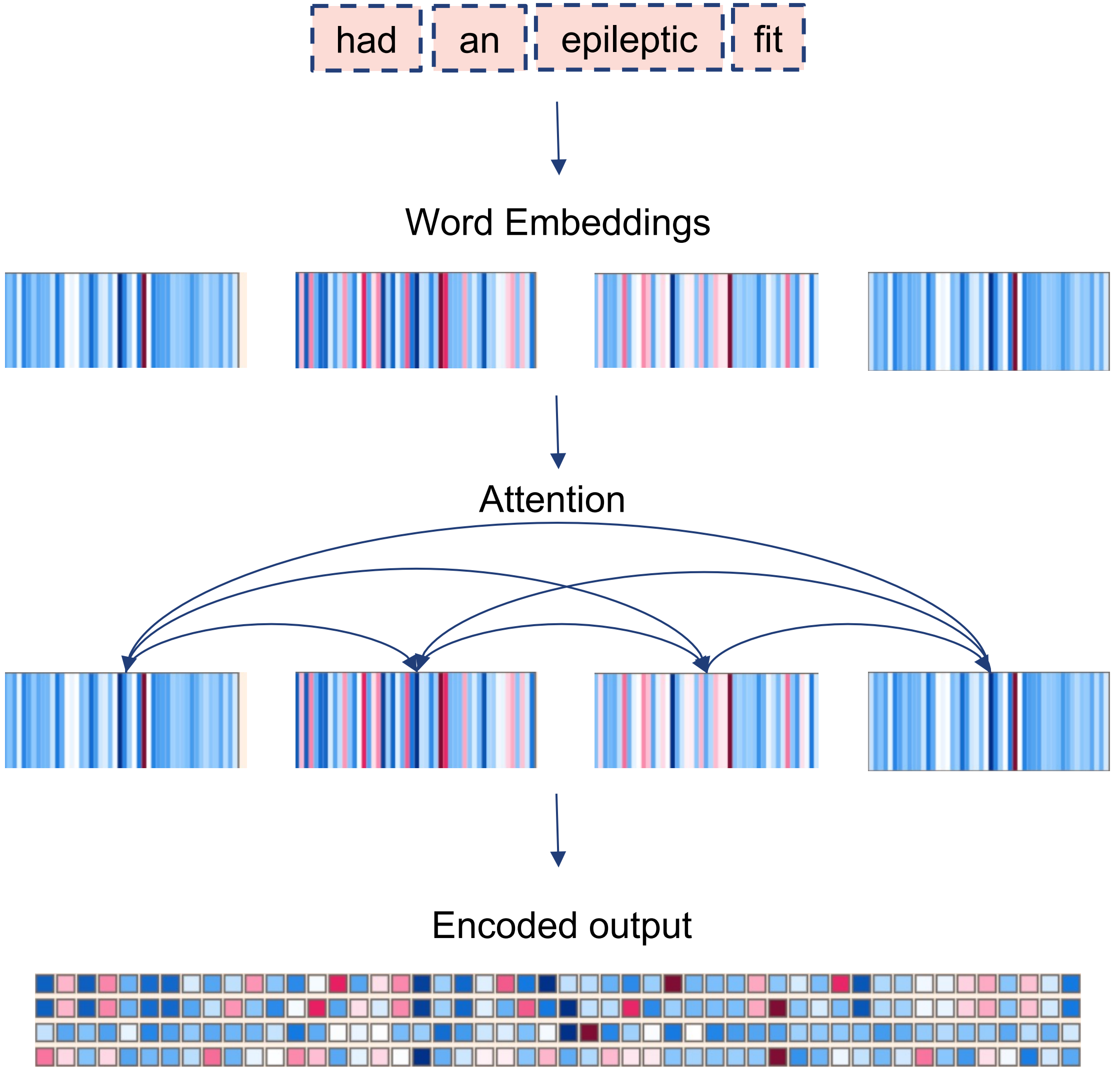
The attention mechanism allows the model to **weigh** the **importance** of different tokens



- Attention → focus on specific parts of the sequence when making predictions, enabling it to capture contextual relationships.

How it Works – Summary

- After tokenising and encoding a prompt, we’re left with a block of data representing our input as the machine understands it, including meanings, positions and relationships between words.

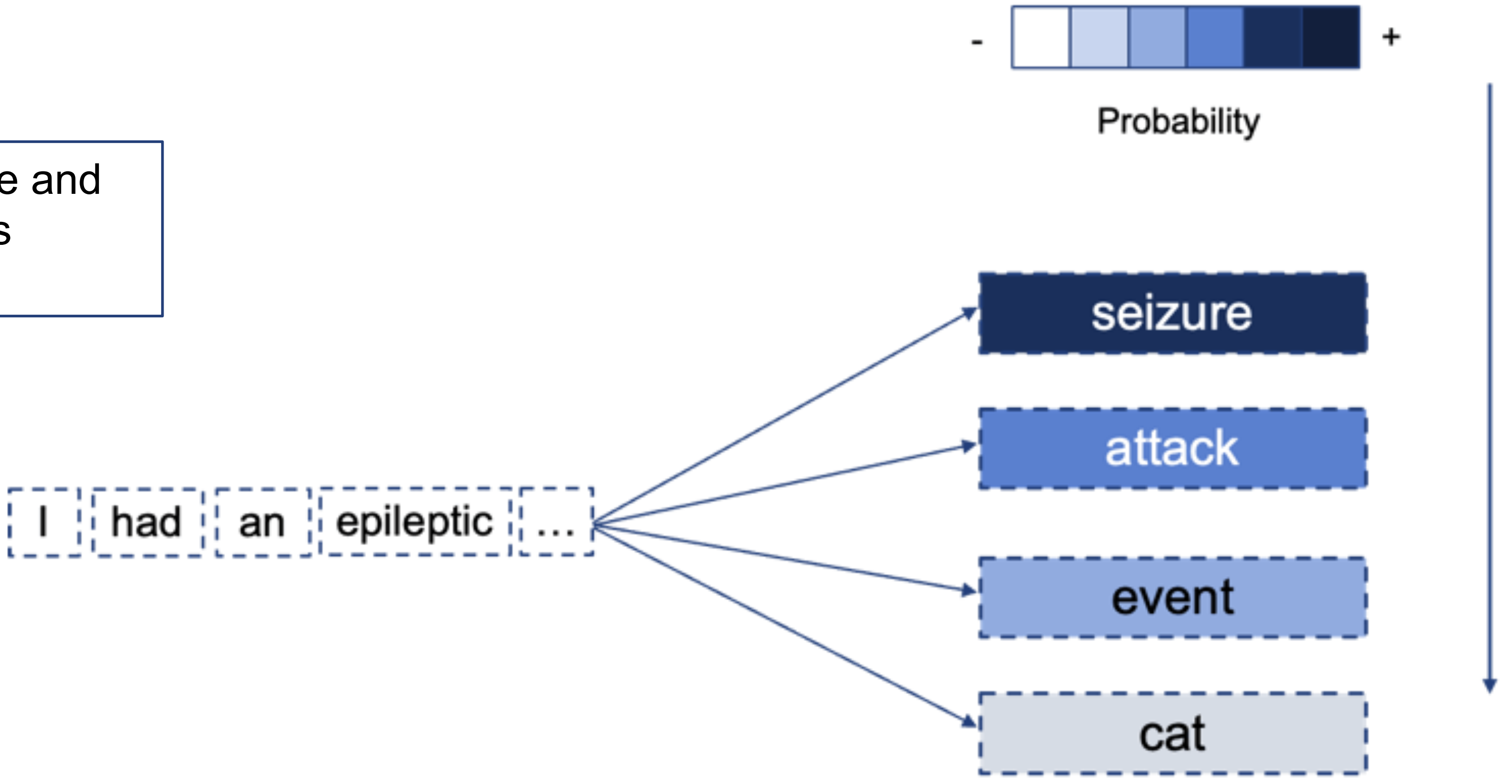


- Machine understandable

Creating a GenAI in Healthcare

TASK

Predict the next word in a sequence and do this repeatedly until the output is complete

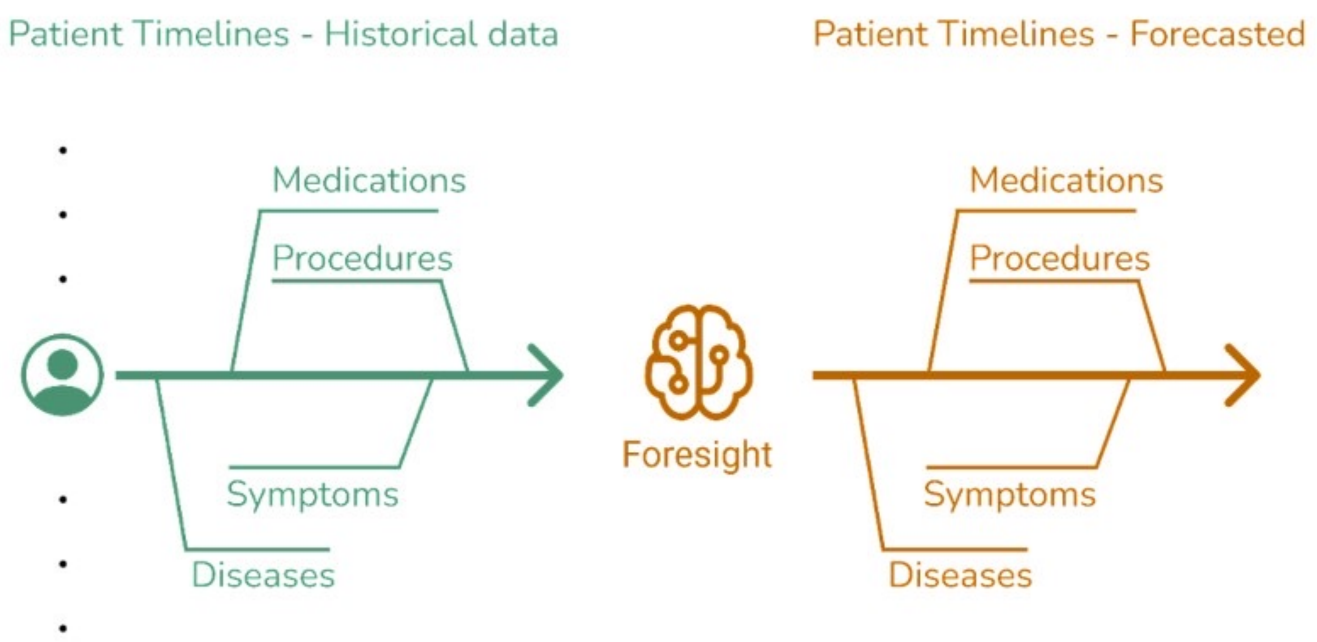


Real-world Example of Clinical LLMs using SNOMED-CT

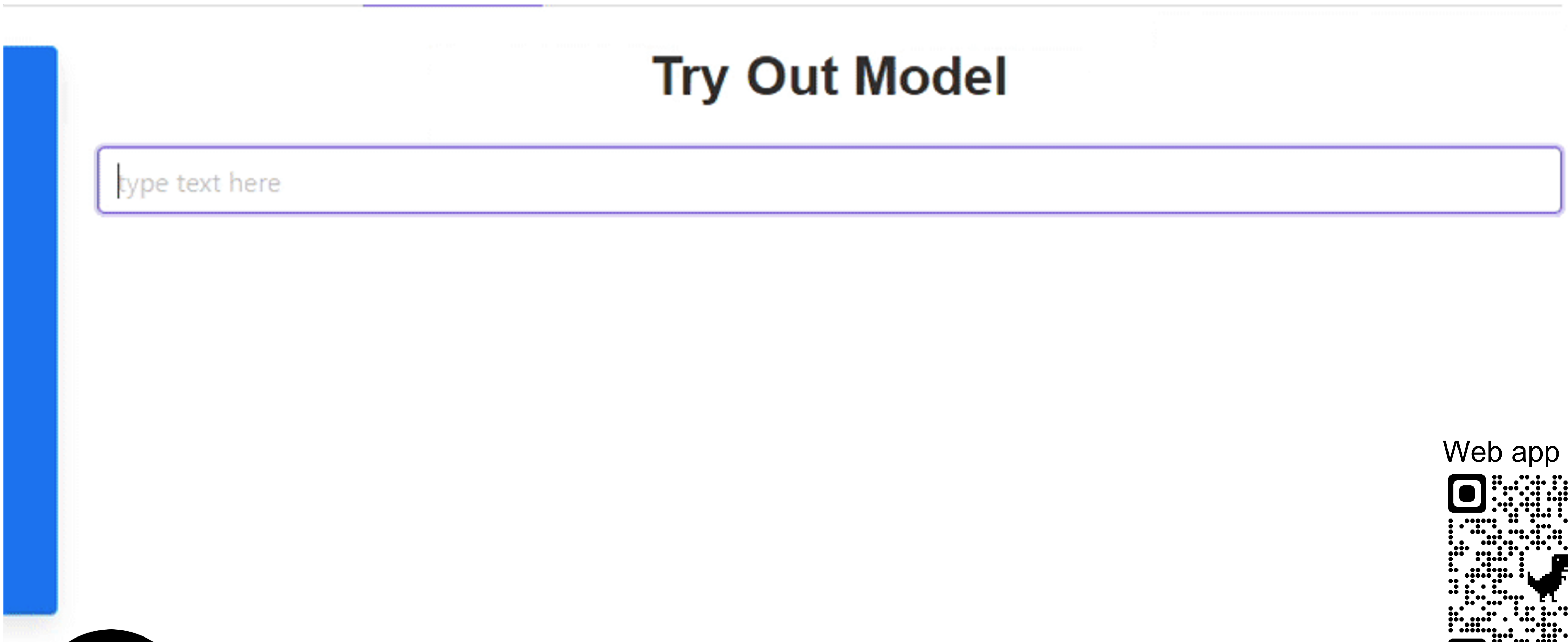
https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00025-6/fulltext

Foresight

- Uses the existing knowledge from electronic health records
- A patient’s medical history can be seen as a sequence of SNOMED-CT concepts



Extracting SNOMED-CT from Free-Text



Web app Demo

Try it out yourself!

Free-Text to Timeline

MedCAT annotations

Clinic letter Extract example

Dear Mr. TGF:

Mr. XY is a client S-325511000000105 - CLIENT (PERSON) - T-15300 - PERSON - 1.0 of our Epilepsy S-84757009 - EPILEPSY (DISORDER) - T-02100 - DISORDER - 0.69 clinic S-257585005 - CLINIC (ENVIRONMENT) - T-03100 - ENVIRONMENT - 0.5 . He suffers from juvenile myoclonic epilepsy S-6204001 - JUVENILE MYOCLONIC EPILEPSY (DISORDER) - T-02100 - DISORDER - 1.0 and was medically screened S-20135006 - SCREENING PROCEDURE (PROCEDURE) - T-10000 - PROCEDURE - 0.33 by you as a driver S-236320001 - VEHICLE DRIVER (OCCUPATION) - T-15200 - OCCUPATION - 1.0 . Your assessment of a diagnosise S-439401001 - DIAGNOSIS (OBSERVABLE ENTITY) - T-05000 - OBSERVABLE ENTITY - 0.5 of jme S-6204001 - JUVENILE MYOCLONIC EPILEPSY (DISORDER) - T-02100 - DISORDER - 1.0 has resulted in S-79409006 - RESULTING IN (ATTRIBUTE) - T-14310 - ATTRIBUTE - 1.0 XY being denied S-441889009 - DENIED (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 1.0 the job S-14679004 - OCCUPATION (OCCUPATION) - T-15200 - OCCUPATION - 1.0 and

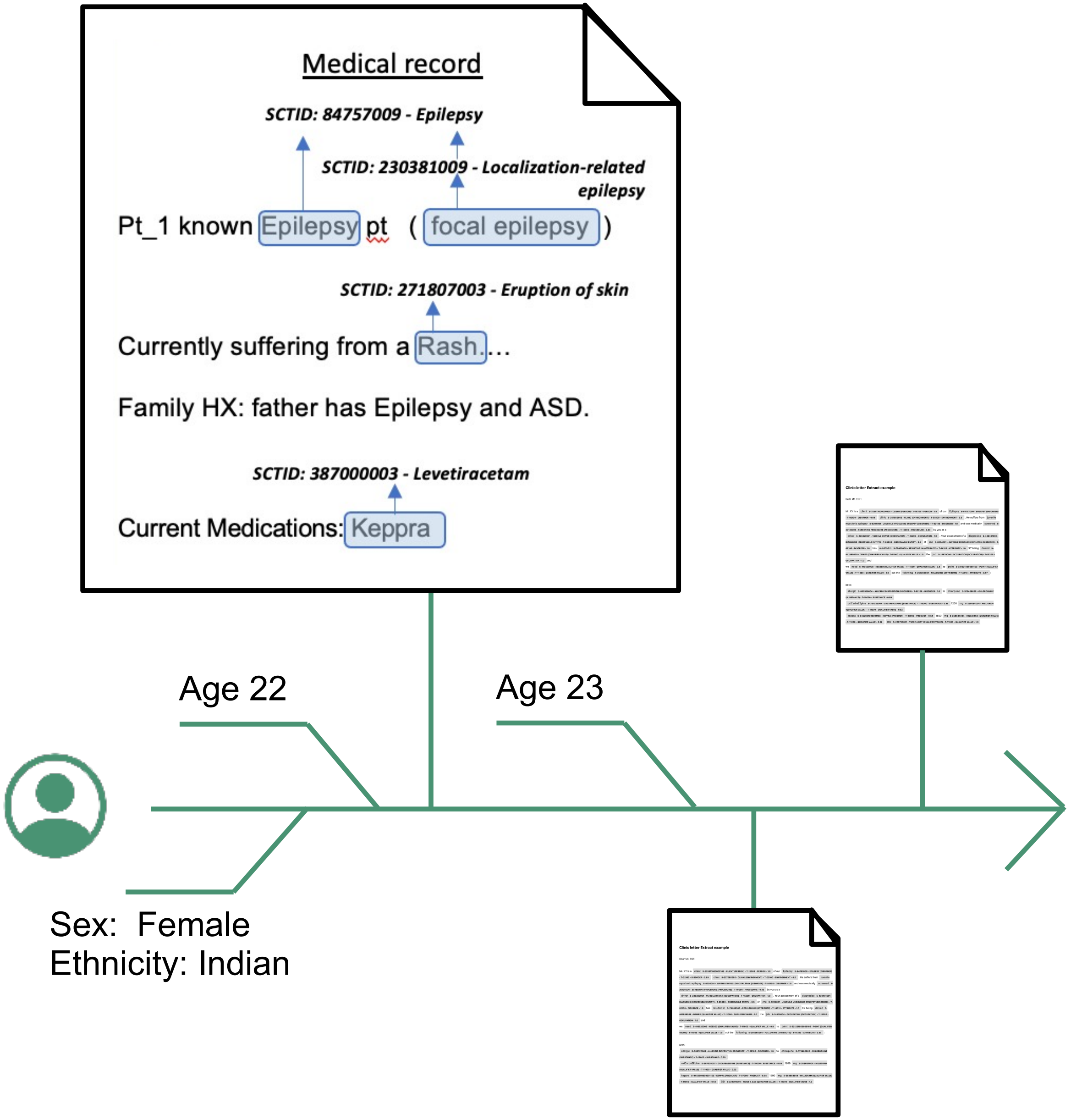
we need S-410525008 - NEEDED (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 0.8 to point S-321221000000103 - POINT (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 1.0 out the following S-255260001 - FOLLOWING (ATTRIBUTE) - T-14310 - ATTRIBUTE - 0.57 :

DHX:

allergic S-609328004 - ALLERGIC DISPOSITION (DISORDER) - T-02100 - DISORDER - 1.0 to chlorquine S-373468005 - CHLOROQUINE (SUBSTANCE) - T-19000 - SUBSTANCE - 0.69

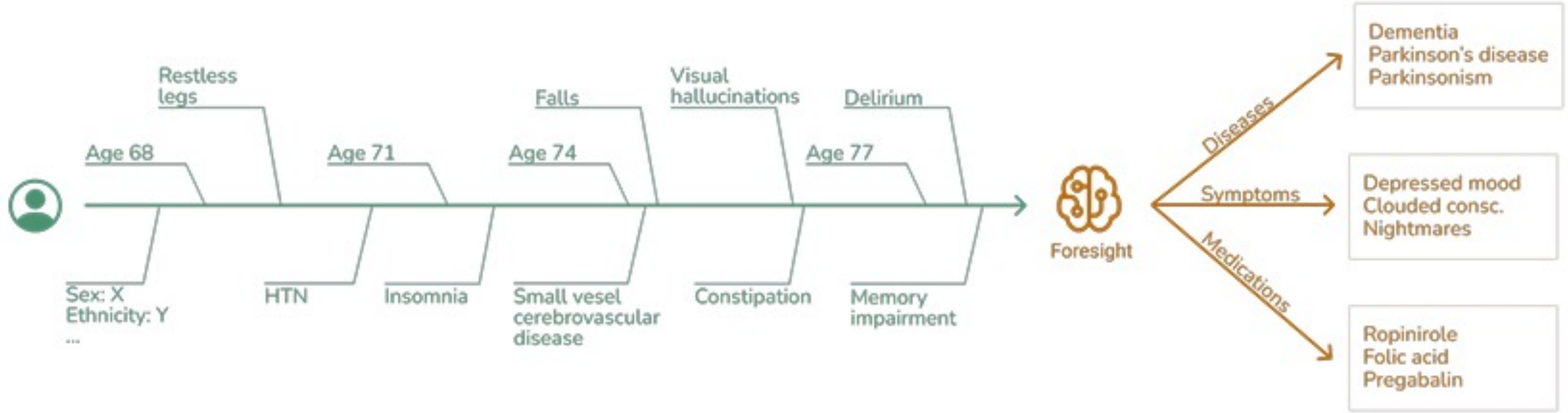
oxlCarbaZEpine S-387025007 - OXCARBAZEPINE (SUBSTANCE) - T-19000 - SUBSTANCE - 0.85 1200 mg S-258684004 - MILLIGRAM (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 0.52

keppra S-9452601000001103 - KEPPRA (PRODUCT) - T-07000 - PRODUCT - 0.54 1500 mg S-258684004 - MILLIGRAM (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 0.52 BID S-229799001 - TWICE A DAY (QUALIFIER VALUE) - T-11000 - QUALIFIER VALUE - 1.0



Foresight

- Train on existing patient timelines found in your dataset. (Patient as a sequence of concepts)
- Input a timeline of a new patient's health trajectory (timeline)
- Forecast Then you can forecast the rest of their timeline.



Input

Forecast

Foresight

CogStack

Foresight

Foresight

A generative transformer model trained on ~1M patients from King's College Hospital and ~20k patients from South London and the Maudsley Mental Health NHS Foundation Trust. Please do not use this to diagnose yourself or someone else. The main use of this webapp is to test the capabilities of the underlying models. Mistakes or biases are possible and reflect problems in the dataset or simply the inability of the model to generalize well enough.

The model is geared towards common high-level SNOMED-CT concepts, and performance is better with long timelines.

Citation

Preprint on arXiv

Examples Timeline (Physical Health)

- CoVid
- Intracranial Hypertension
- Kidney Disease
- Sleep Apnoea
- Wernicke Encephalopathy
- Parkinson's Disease
- Kidney Disease [Masked]

Examples Text (Physical Health)

First two examples are taken from BMJ

Timeline | Text

Search for concepts and press ENTER to add them to the timeline.

TIME

Age: 59

Alcohol dependence (disorder)

Age: 61

Ischemic stroke (disorder)

Atrial fibrillation (disorder)

<SEP>

Evaluation of oral and pharyngeal swallowing function (procedure)

Dysphagia (disorder)

<SEP>

Acute confusion (finding)

Diplopia (disorder)

Nystagmus (disorder)

Impaired cognition (finding)

Predict

Relative Probability	SNOMED ID	Name	Show saliency
0.02275	21007002	Wernicke's disease (disorder)	Saliency
0.01693	286933003	Confusional state (disorder)	Saliency
0.01343	191480000	Alcohol withdrawal syndrome (disorder)	Saliency

What should be predicted by the model

- [x] Disorders
- [ ] Medications and Substances
- [ ] Procedures
- [ ] Symptoms and Findings

Model (First pick a model and then add concepts to the timeline)

- (•) KCH Physical Health
- ( ) SLaM Mental Health
- ( ) MIMIC-III Physical Health

Concept Status

- [x] New Concepts
- [ ] Recurring Concepts

Ignore concepts that are parents/children/siblings of current concepts

- [ ] Ignore Siblings
- [x] Ignore Children
- [x] Ignore Parents

Filter to SNOMED codes and all its children (can be a comma separated list)

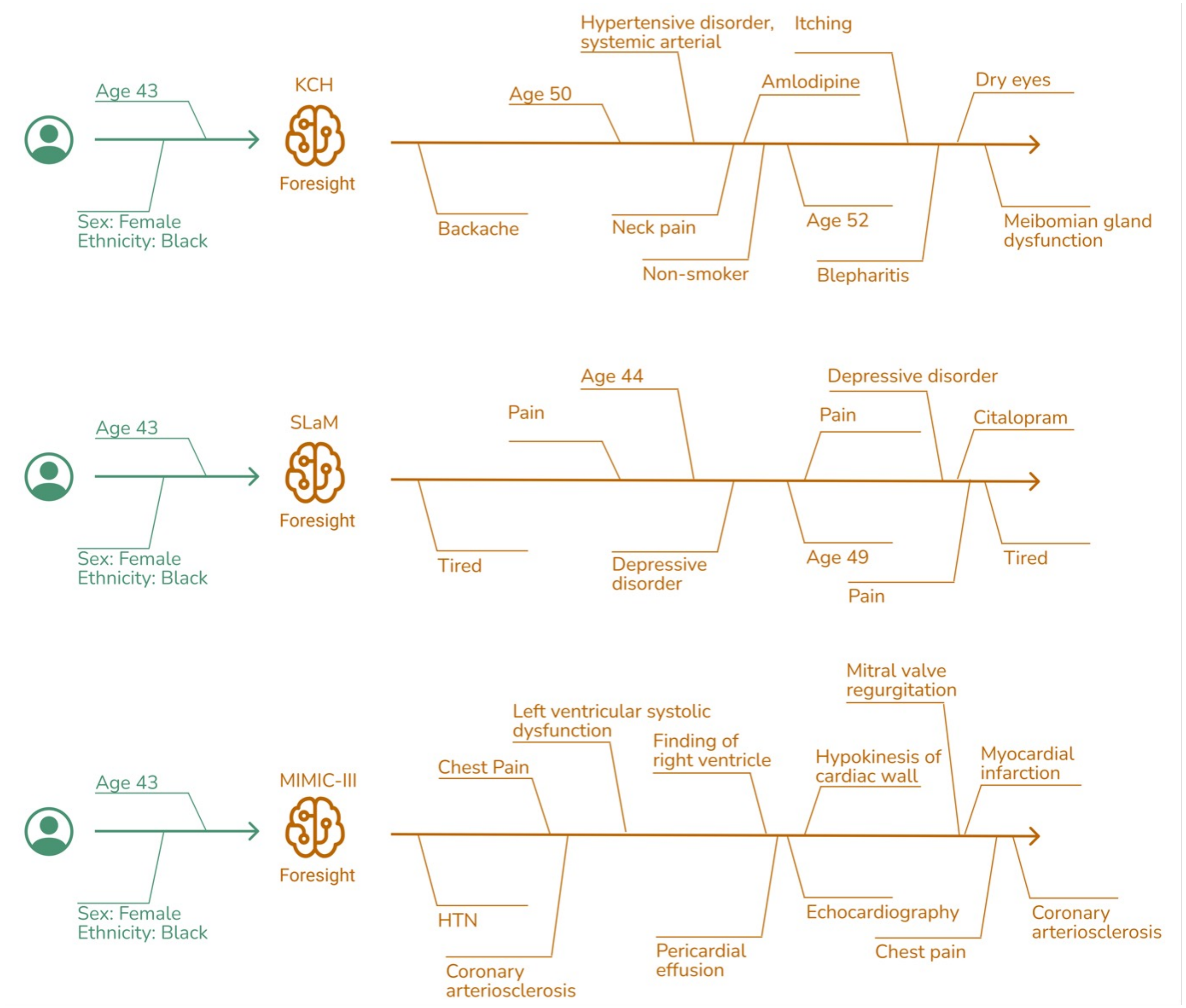
Enter SNOMED codes

Detected concepts from text

Save/Load your timeline

Foresight: Timeline Generation

Examples of generated synthetic timelines



Simple Prompt:

Age: 43-year-old

Sex: Female

Ethnicity: Black

Performance

Task: Next Concept Prediction

Semantic Tags	KCH (Precision)			SLaM (Precision)		
	TOP-1	**TOP-5**	**TOP-10**	**TOP-1**	**TOP-5**	**TOP-10**
Precision Overall	0.667	0.875	0.917	0.658	0.890	0.938
Precision Disorders	0.605	0.825	0.874	0.637	0.872	0.917
Precision Findings	0.604	0.855	0.908	0.624	0.879	0.935
Precision Substances	0.716	0.912	0.950	0.731	0.921	0.958

Try it for yourself

https://foresight.sites.er.kcl.ac.uk/

Applications and Limitations

Applications:

- Clinical Risk prediction
- Diagnosis suggestion
- Digital twins / virtual trial emulation
- Forecasting cost

Base Patient Timeline Prompt	Scenario	Time+1 5 events	Time+2 5 events	Time+3 5 events	Time+4 5 events	Time+5 5 events
"This 45 year old male has 1 month of intermittent confusion. He presented with confusion and motor seizures. He was drowsy. He reports olfactory hallucinations. Sometimes he feels a sensation of deja vu. His EEG shows slowing in the left temporal leads."	A: Levetiracetam	Seizure Depression Fall Pneumonia Cataracts	UTI STEMI Confusion Anxiety Chest pain	AKI Dehydration Pneumonia Sepsis Anorexia	Depression Anxiety Alopecia Seizure Rash	Anxiety Elation Urosepsis UTI Depression
	B: Lamotrigine	Eczema Dysarthria Glaucoma Diplopia Rash	Rash Pneumonia UTI Seizure Tinnitus	Seizure Rash Fall Pneumonia Cataracts	Depression Fall Pneumonia Seizure Covid	SUDEP Arryhthmia NSTEMI Anxiety Chest pain

Limitations:

- Can only predict concepts seen in the training dataset
- Cannot accurately estimate time
- Learns the biases in the data

How to Introduce LLMs into Healthcare Environments

Application Frameworks	Are the **cornerstone of the GenAI stack** and provide the foundation for building and running generative AI applications.
Application Models	Models are the **brain of GenAI systems**, generating new data or content based on learned patterns.
Data	Essential for **training and feeding information to the AI models**. To make the models more effective and precise, developers need to operationalize their data. Systems need to be able to ingest structured and unstructured data.
Evaluation Platform	Provide tools and metrics for **assessing the performance of generative AI models**. This includes metrics such as accuracy, loss, and convergence rates, as well as visualization tools. They also track model performance in real-time to detect anomalies or drifts over time.
Deployment	Transitioning GenAI applications from development environments to production environments to be **used by end-users**. Includes packaging the application, configuring deployment infrastructure, and ensuring scalability, reliability, and security in the production environment.

Key Consideration: Experience and Resources

Things to consider for choosing an AI tech stack:

- **Team Expertise**: Development team's skills impact technology choice.
- **Resource Availability**: Access to hardware and software influences tech stack selection.
- **Training and Support**: Availability of resources affects technology choice.
- **Budget Constraints**: Project budget influences tech stack decisions.
- **Maintenance Needs**: Maintenance and ongoing validation requirements impact technology selection.

Questions / End

Contact

anthony.shek@gstt.nhs.uk