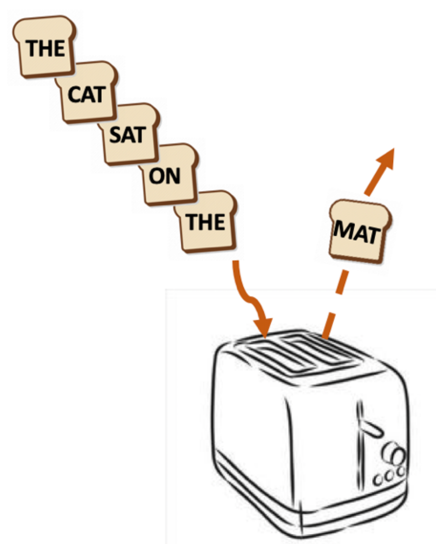
From Words to Wellness: Tapping ChatGPT's "Brain" to Power Longitudinal Health Predictions

Rayan M. Ashish

Imagine you're chatting with ChatGPT. You type, *"I feel like going to the"* and it predicts *"park" "beach"* or *"mall"*. How does it know what you might say next? At its core, ChatGPT is powered by a **Large Language Model (LLM)** that **predicts** the next word in a sentence, by analyzing patterns in the text you've **already written**. Now let's dive into the world of healthcare, where the ability to predict the future can be life-changing. Imagine forecasting the progression of cognitive decline in Alzheimer's patients, offering a glimpse into how the disease might unfold. Picture tracking the subtle but steady motor changes in Parkinson's or predicting the ticking clock of kidney function loss in Chronic Kidney Disease (CKD). Even the gradual weakening of lung capacity in COPD (Chronic Obstructive Pulmonary Disease) can be mapped over time. Longitudinal predictions hold the key to not just understanding these conditions but intervening at the right moment to make a real difference.

Before that however, let us take a moment to understand how does ChatGPT (or any LLM) predict your next word so intelligently.



Large Language Models: *are simply "tell-me-the-next-word" machines*

At their core LLMs, like ChatGPT, are simply “tell-me-the-next-word” machines. Give the machine “The cat sat” and it will tell you “on” or “in”. The machine is smart enough to automatically plug back “on” and then feed itself with “The cat sat on” and then give out “the” and so on. So if you give it “The cat sat” it will give you back “The cat sat on the mat”.

How Does this Help with Predicting Disease Progression ?

LLMs are powered by Deep Learning and by particular models called “Transformers”. But let us now consider the question of how we can take this idea of predicting what word comes next, and apply it to longitudinal health data. Instead of predicting words in a sentence, we predict future measurements in a person’s health data.

Think of a person’s health journey as a story. For example, in a longitudinal study of cognitive decline, researchers collect data from the same person over time. They measure things like:

- Memory test scores
- Brain imaging features
- Lifestyle data like physical activity

Before Transformer model driven LLMs like ChatGPT, the **Recurrent Neural Network or “RNNs”** models were the go-to models for sequence prediction. RNNs process one word at a time, updating their “memory” (hidden state) with each step. While ChatGPT has evolved into something much more complex, the foundational idea of sequence prediction—whether for text or time-based data—remains central.

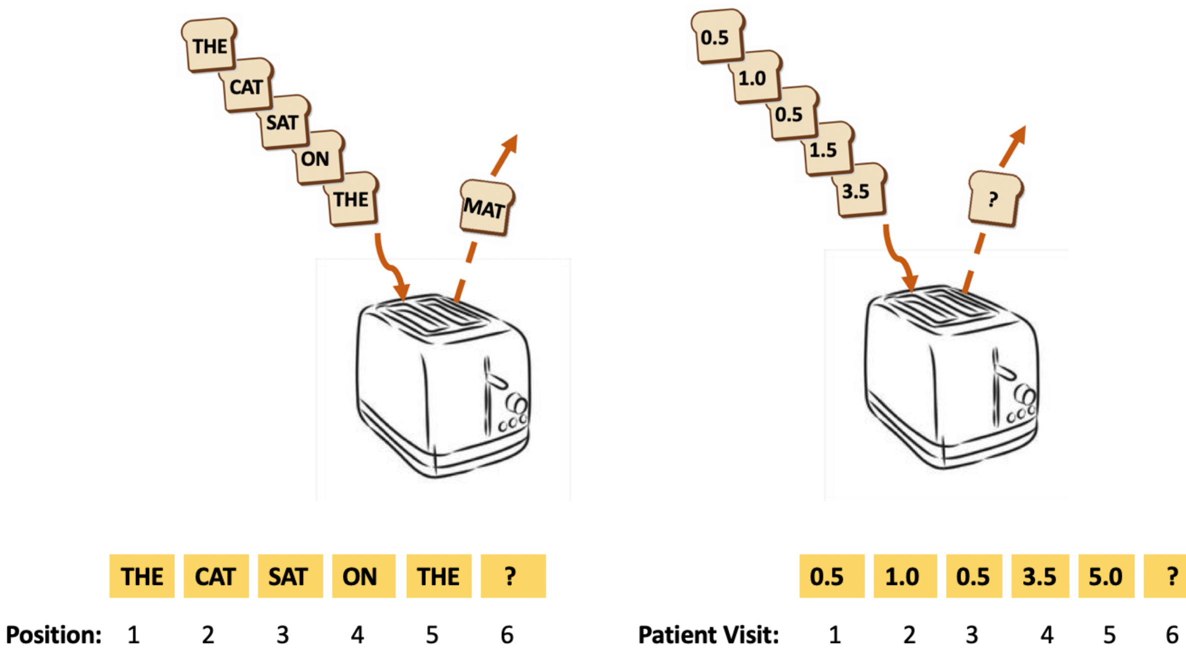
Recurrent Neural Networks (RNNs) for Predicting Disease Progression

How an RNN Predicts the Next Word in a Sentence

1. **The Sequence is Key**: Let’s say you type, "The cat is on the". The model tries to predict the next word—"mat", "roof", or even something unexpected like "internet". It looks at the sequence of words and makes an educated guess.
2. **Memory Matters**: An RNN remembers what you’ve typed so far by storing "hidden states". These hidden states act like a short-term memory. If the sentence

started with "In the dark forest, the cat is on the", the model would take "forest" into account and might guess something like "tree" instead of "mat".

3. **Training on Lots of Text**: The model gets really good at these predictions by being trained on huge amounts of text. It learns the patterns of language—what words usually follow others.
4. **Making the Guess**: Finally, the model gives probabilities for the next word. For example:
   - "mat": 70%
   - "roof": 20%
   - "internet": 10%



Position	*2*
Type	*Living thing*
POS	*Noun*
Capitalized	*Yes*
Common	*Moderately*

Visit	*3*
Education	*Living thing*
Physical Activity	*Noun*
MMSCORE	*Yes*
SPARE_AD	*Moderately*

Adapting RNNs for Longitudinal Health Predictions

How do we make the RNN work for Longitudinal Health Predictions (Disease Progression)

1. **Turning Visits into Data Points**: Let's say a person has six visits to a clinic over several years. For each visit, we record data like their age, education level, memory scores, and brain imaging features. This creates a sequence of data points, just like words in a sentence.
2. **Learning from the Past**: Just like with text, the RNN looks at the earlier visits to predict what will happen at the next one. For example, if someone's memory scores have been steadily declining, the RNN might predict that the decline will continue.
3. **Storing the Sequence**: RNNs use their hidden states to store information from earlier visits. This helps them understand long-term patterns, like whether a decline is accelerating or leveling off.
4. **Making Predictions**: Instead of predicting the next word, the model predicts the patient's cognitive status—whether they'll remain cognitively normal (CN), develop MCI, or progress to Alzheimer's disease (AD).

Why This Is a Big Deal

Longitudinal data is everywhere in science and healthcare. Whether it's tracking a patient's health, a student's learning progress, or even the behavior of animals in a study, these kinds of predictions can have a massive impact. Here's why this approach is so powerful:

- **Early Warnings**: For diseases like Alzheimer's, predicting the trajectory of cognitive decline early can help doctors intervene sooner.
- **Personalized Predictions**: By training the RNN on individual patients' data, the model can make predictions tailored to each person.
- **Wider Applications**: This doesn't just apply to neuroscience. You can use this method for other conditions such as pulmonary or cardiovascular domains.

Challenges and Future Directions

Of course, applying RNNs to longitudinal data isn't as straightforward as working with text. Here are some of the challenges and how we're tackling them:

1. **Irregular Time Gaps**: In text, the time between words is basically zero. In longitudinal data, there might be months or years between visits. To address this, we can add time as an input feature to the RNN.
2. **Small Datasets**: Unlike the billions of sentences used to train LLMs, longitudinal studies often have much smaller datasets. Techniques like oversampling or data augmentation can help.
3. **Interpreting the Results**: It's important to understand why the model makes certain predictions, especially in healthcare. Adding attention mechanisms to the RNN can highlight which parts of the sequence are most important for the prediction.

NeuroLAMA: Deep Learning for Longitudinal Health Predictions

NeuroLAMA, www.neurolama.org is an open machine learning resource designed for analyzing longitudinal health data. Currently, NeuroLAMA focuses on neurodegenerative conditions such as Alzheimer's disease. Its mission includes:

1. Providing comprehensive resources to the community, leveraging standard datasets like ADNI and BIDS.
2. Encouraging contributions from developers to expand its capabilities.

While the initial focus is on neurodegenerative diseases, the long-term vision is to extend its application to longitudinal health predictions for a wide range of conditions.

Conclusion: From Words to Wellness

Next-sequence prediction started with language models like RNNs and LLMs, revolutionizing how we interact with technology. But the same principles can be applied to longitudinal data, offering new insights into health, education, and beyond. By treating each person's journey as a sequence, we're able to predict what's coming next—giving researchers and clinicians a powerful tool to improve lives.

Whether you're working with sentences or health data, understanding the past is the key to predicting the future.