Why RNN wasn't enough — The Forgetful Storyteller

Imagine you meet a storyteller who remembers what you told them… but only for a short while.
You tell them a long story about your day, starting from breakfast to bedtime. At first, they recall details just fine — the eggs you had for breakfast, the bus you took to work. But as you go on, they start forgetting the early parts. By the time you reach dinner, they completely forgot about breakfast.

This is exactly what happens with **RNNs (Recurrent Neural Networks)**.

- They process sequences step by step, carrying information forward.
- But as the sequence gets longer, **earlier information fades away** (this is called the **vanishing gradient problem**).
- They're great for short memories but terrible for long stories.

So… we needed a new storyteller — one who could **choose what to remember, what to forget, and what to tell you later**.

Enter LSTM — The Organized Storyteller

LSTM stands for **Long Short-Term Memory**.
Think of LSTM as a storyteller who carries a **magic notebook** with three tools:

1. **Forget Gate** — decides what old details to erase.
2. **Input Gate** — decides which new details are important enough to write down.
3. **Output Gate** — decides what information to share at each step.

Here's how the magic works in plain terms:

- The **Cell State** is like the notebook's "main storyline" — it runs through the entire story and keeps track of key points.
- The **Forget Gate** says, "This part is boring, let's erase it."
- The **Input Gate** says, "Oh, this detail is important, let's add it."
- The **Output Gate** says, "I'll only tell you this part for now."

Because of this, LSTMs can remember breakfast **and** dinner, even if there's a lot of chatter in between.

The Problems with LSTM — The Overloaded Storyteller

But… even the organized storyteller isn't perfect.

- They're **slow** — managing gates and notebooks takes time.
- They need **lots of training data** — otherwise, they can get confused about what to remember.
- They can still forget if the story is extremely long (just slower than RNNs).
- They're **complex** — more gates mean more parameters to train, so they eat up more computation.