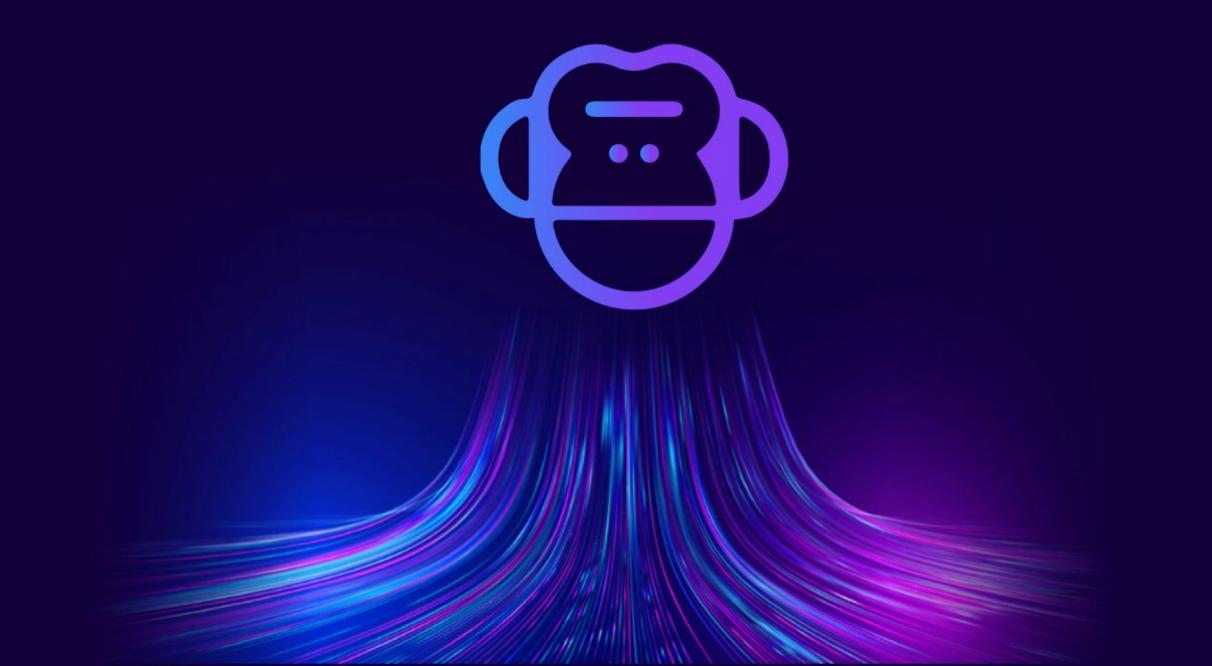
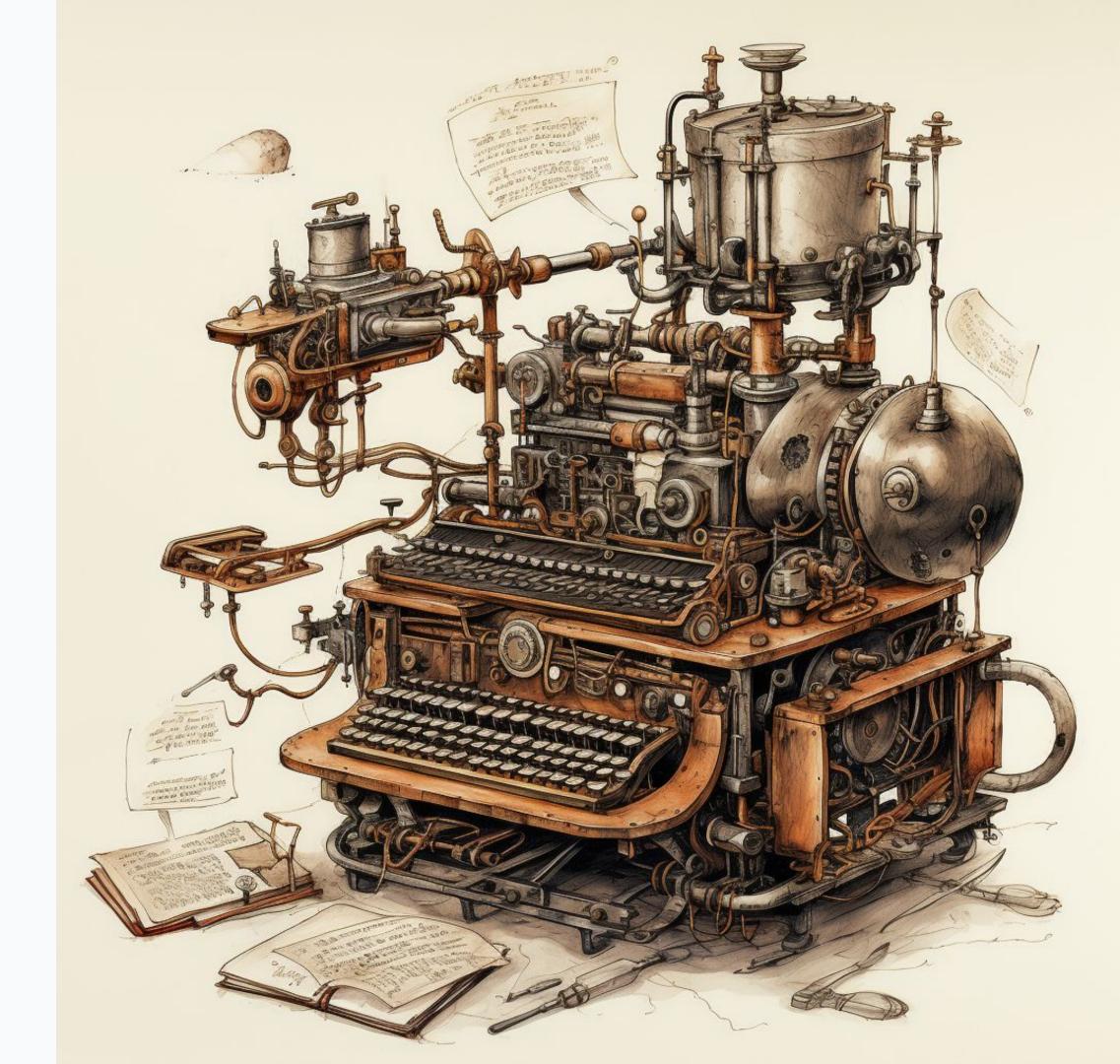


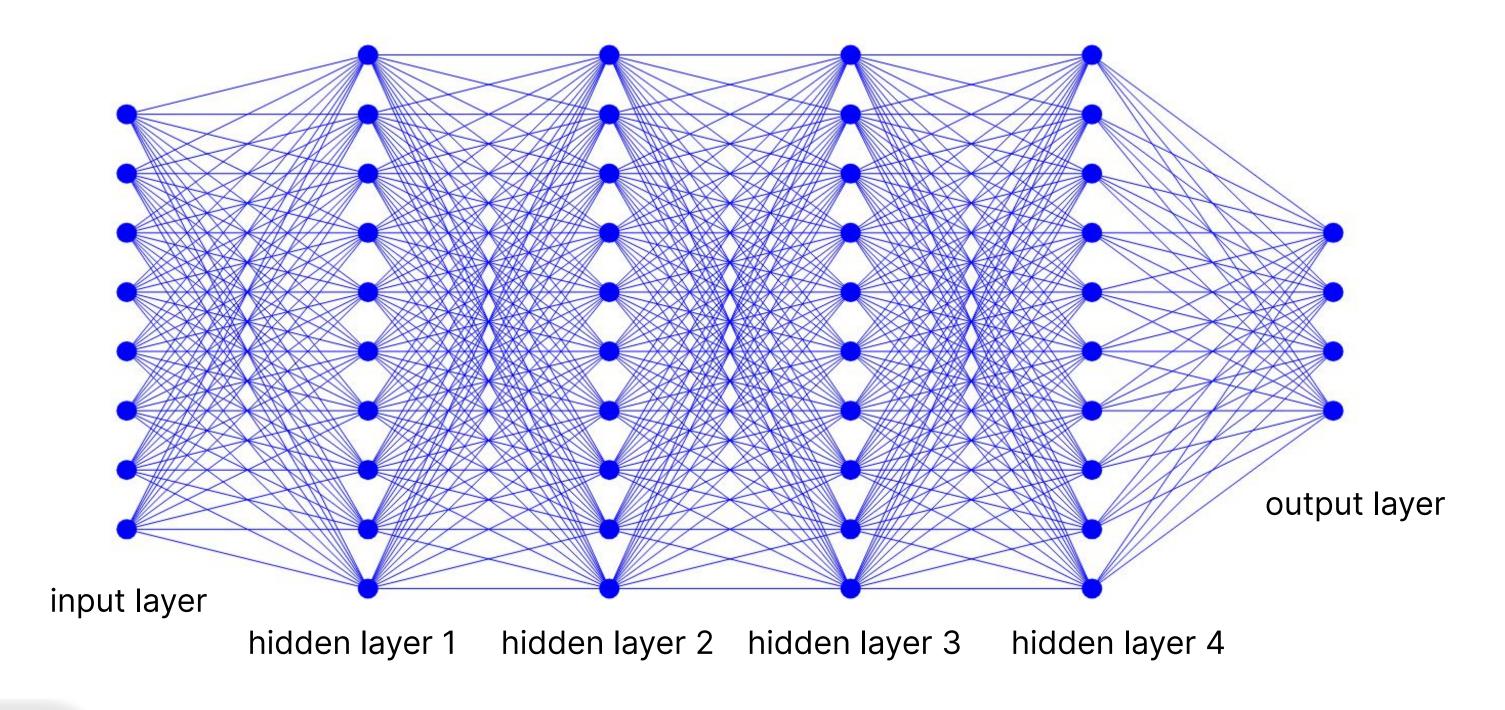
### Product Hour Live - Behind the Scenes of ReMark



Large Language Model: an Instrument of Language



# Large Language Model is a deep neural network



### Training: The network weights are adjusted to make its predictions better

### Pre-training

Massive unlabeled training data.
Goal to predict next token in the data.

## Instruction fine-tuning

Collections of prompts with expected responses.
Goal to generate expected responses.

### Reinforcement learning from human feedback

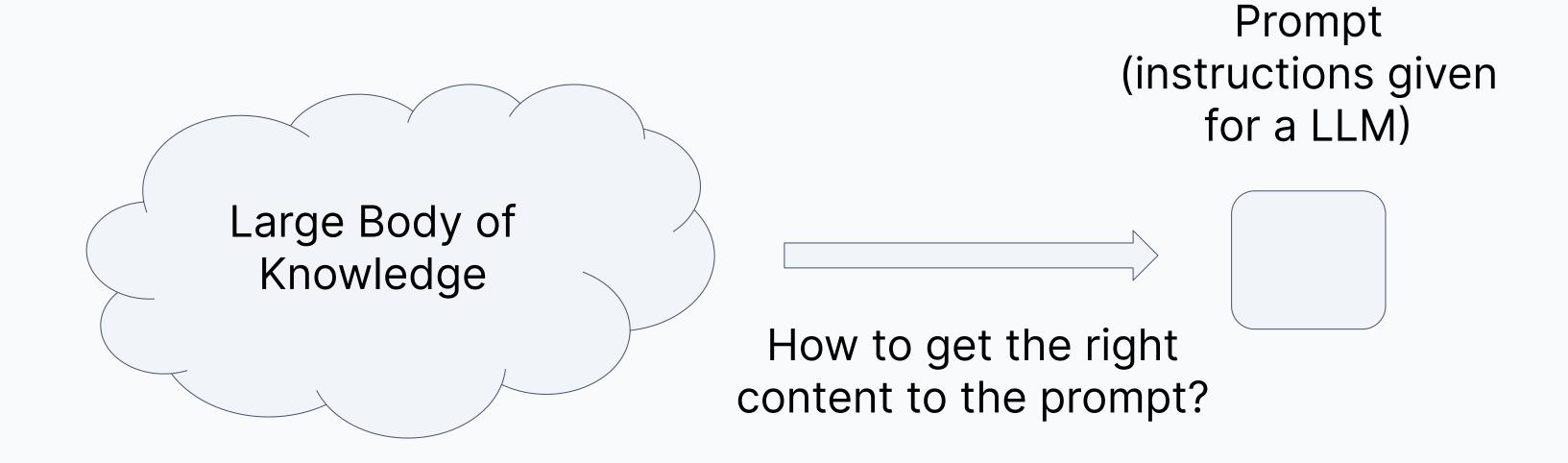
Collection of prompts and human evaluation of responses.



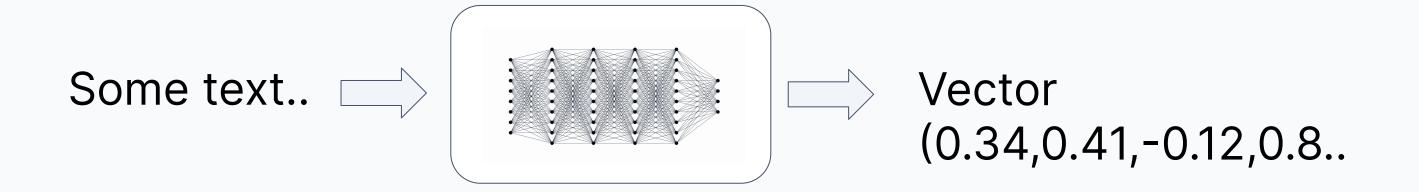
# Context Window size: the number of input and output tokens for a given model.

Model	Context Size (tokens)	Approx pages of text
gpt-4-32k	32768	50
gpt-4	8192	12
gpt-3.5-turbo-16k	16385	25
gpt-3.5-turbo	4097	6

### In-Context Learning



### Embeddings: Turning words into numbers



#### Similarity search





### Retrieval-augmented generation

