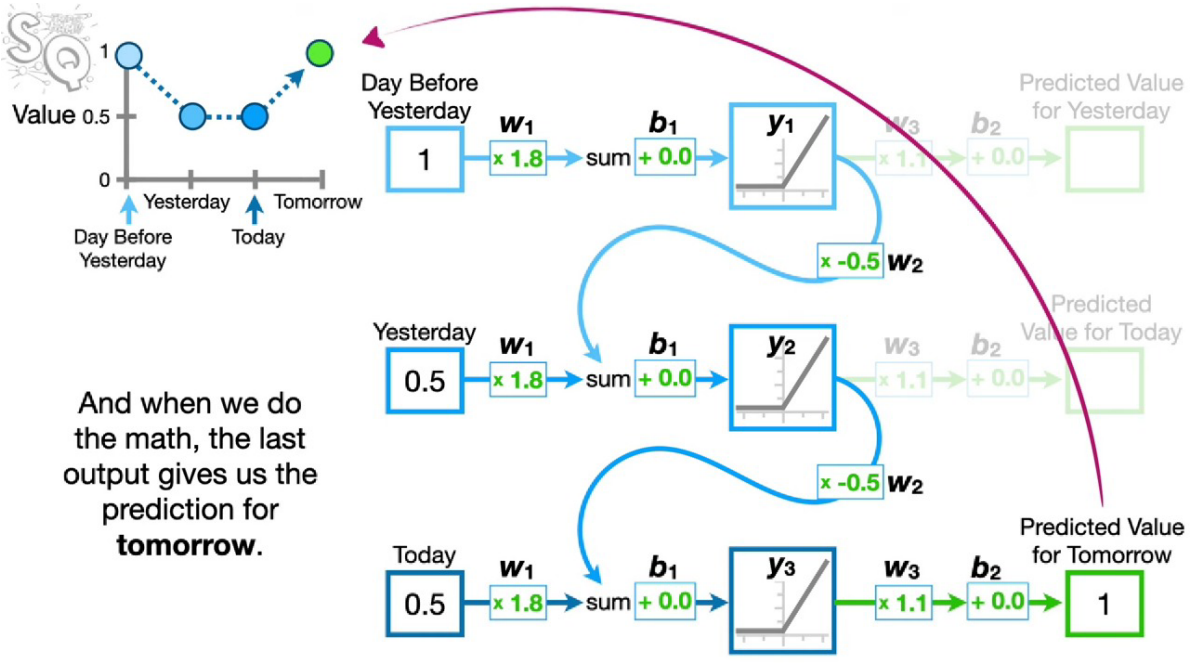
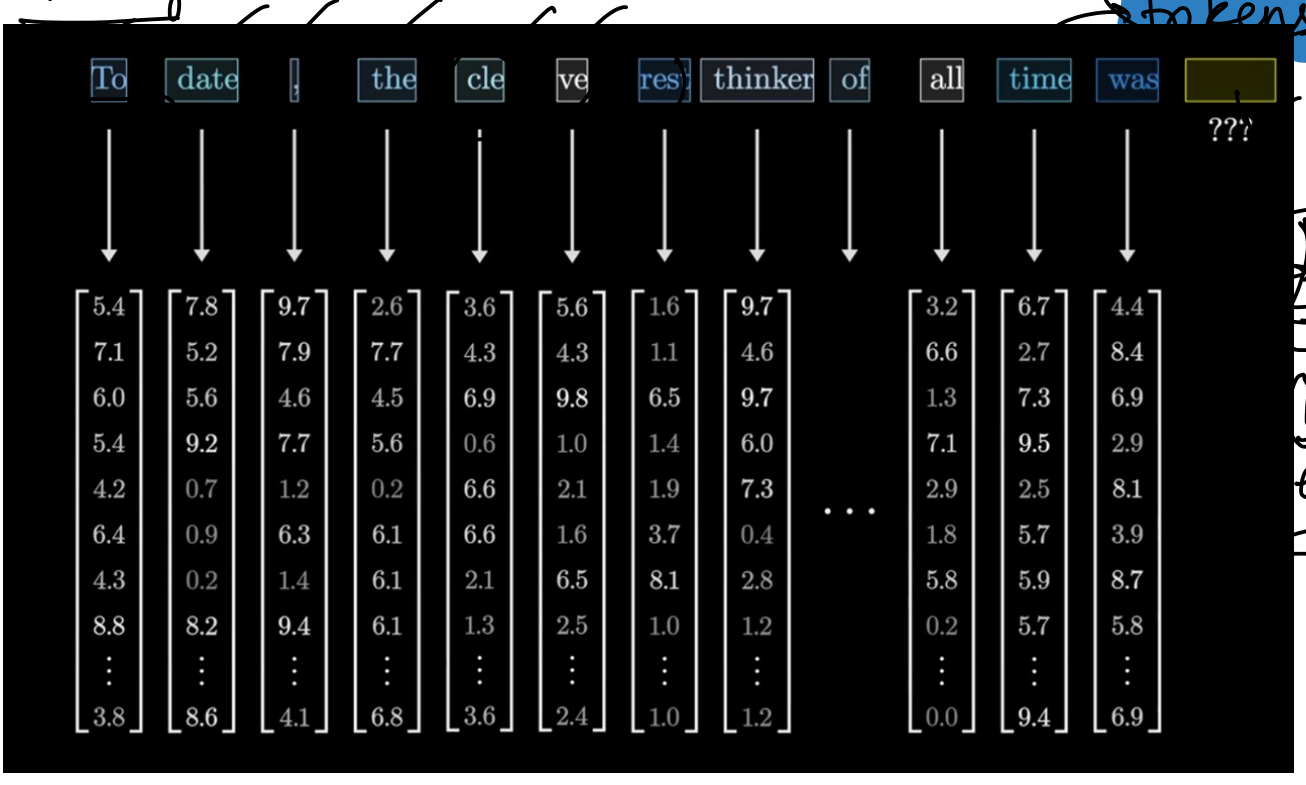
RNNs



Sequential data, customisable no. of inps.

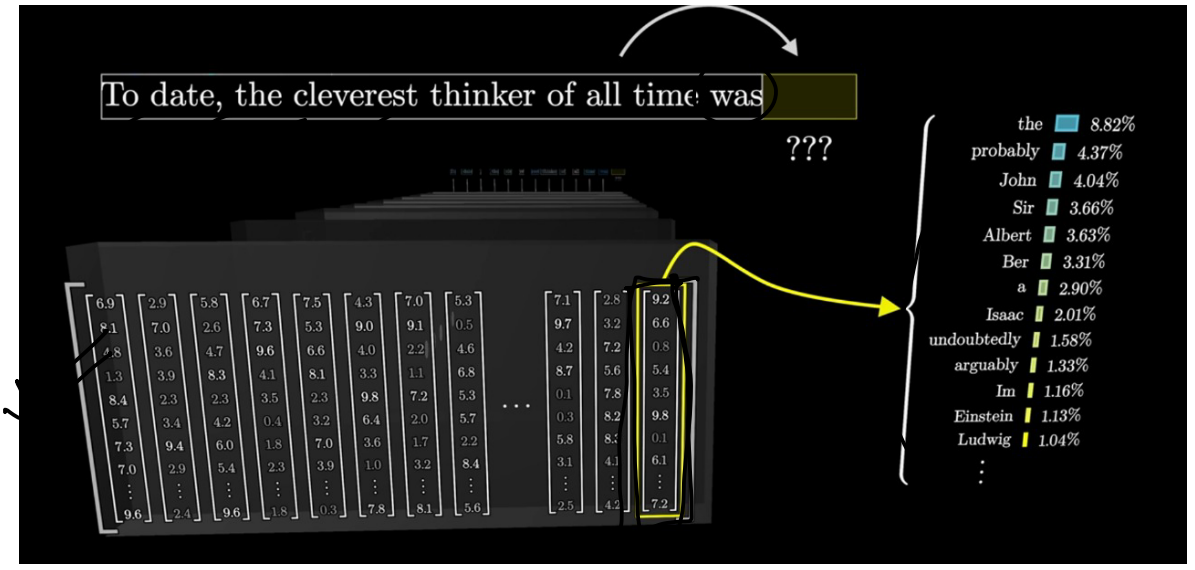
LLMs (Attention mechanism)

→ Predict next word by taking all words in one go.



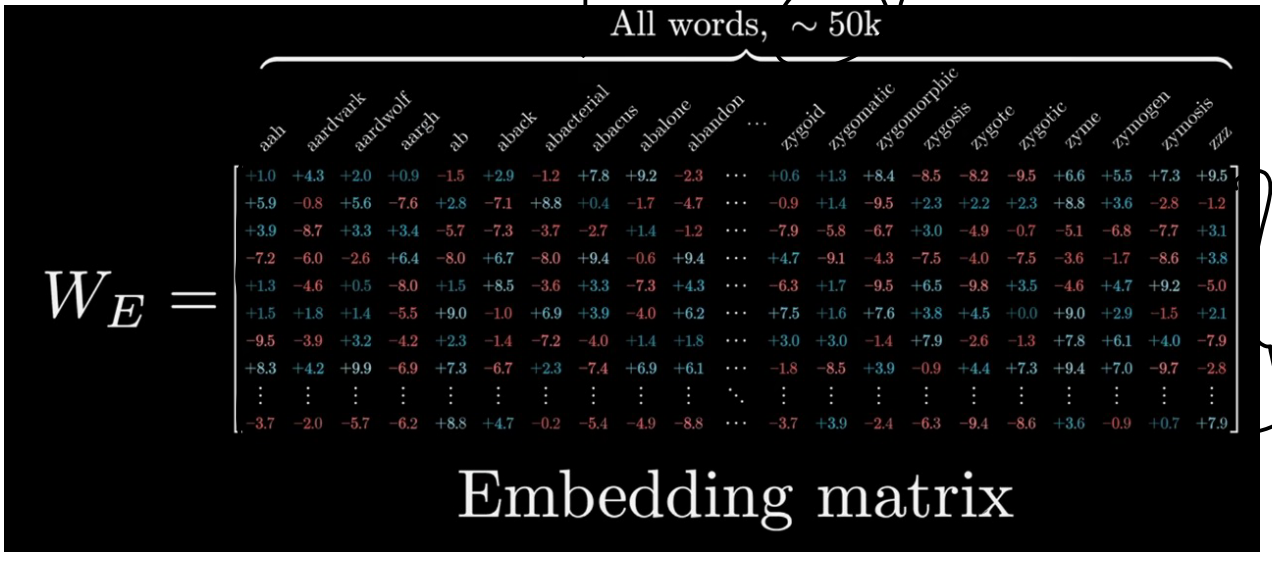
A machine learning model

fashion model

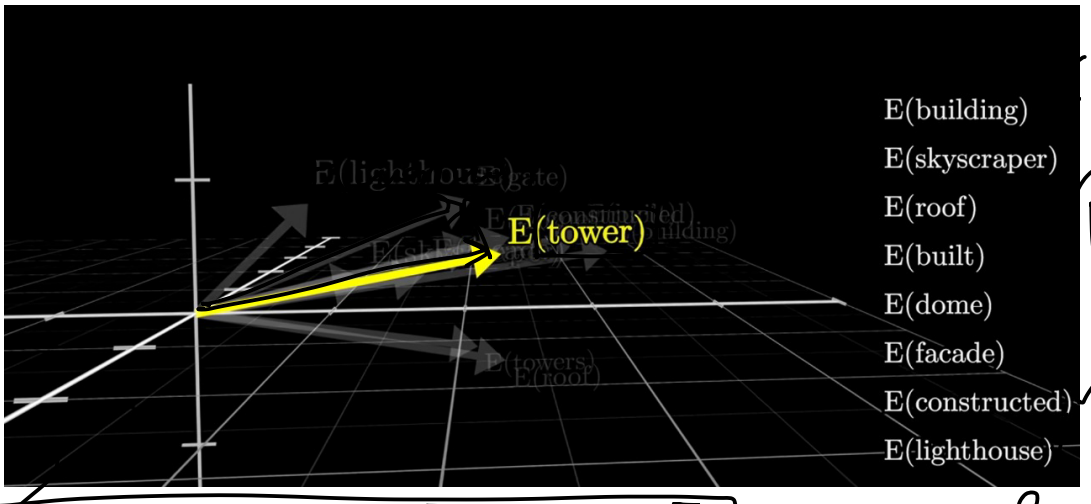


semantic meaning

→ probability of words



100 nos



semantic (positional encoding)

vectors = word$_1$ − word$_2$ → dot & cross product (distance) for Attention mechanism

→ content size

2 types :- decoder only & encoder-decoder

Pretaining → fine tuning → pref tuning

Response generation

→ One token at a time (small model)

i) probability distribution

$w_1 \quad w_2 \quad \cdots \quad w_t$

$P(w_{t+1} \mid C)$

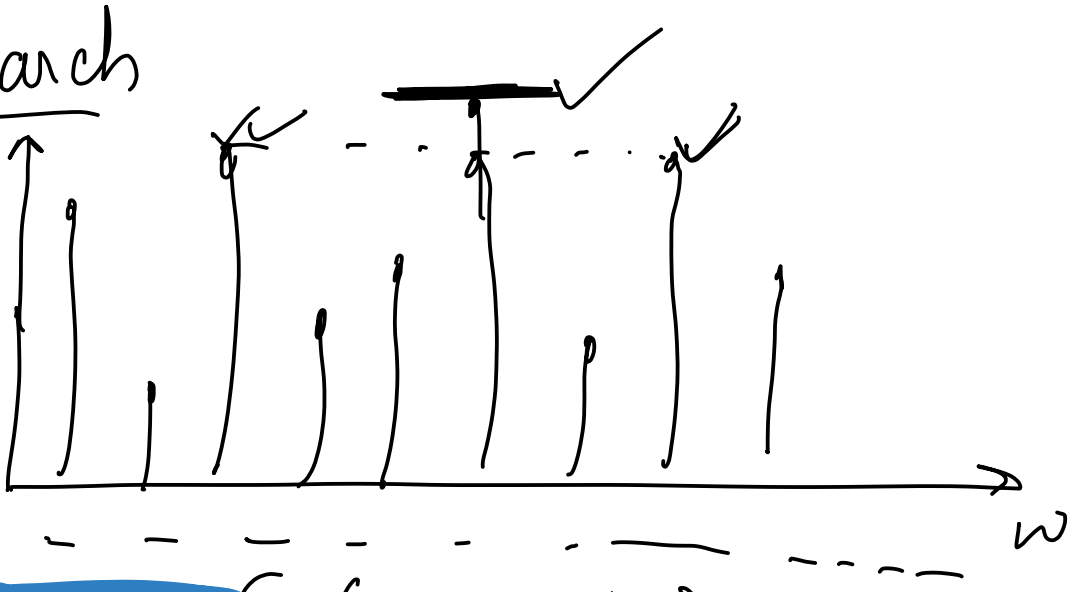
(Vocabulary V)

$P(w_{t+1} = w \mid C)$ → classification

→ via a softmax layer

1) Greedy search

$P(w_{t+1} = w \mid C)$



$\hat{w}_{t+1} = \underset{w \in V}{\operatorname{argmax}} \left(P(w_{t+1} = w \mid C) \right)$

→ statistical method

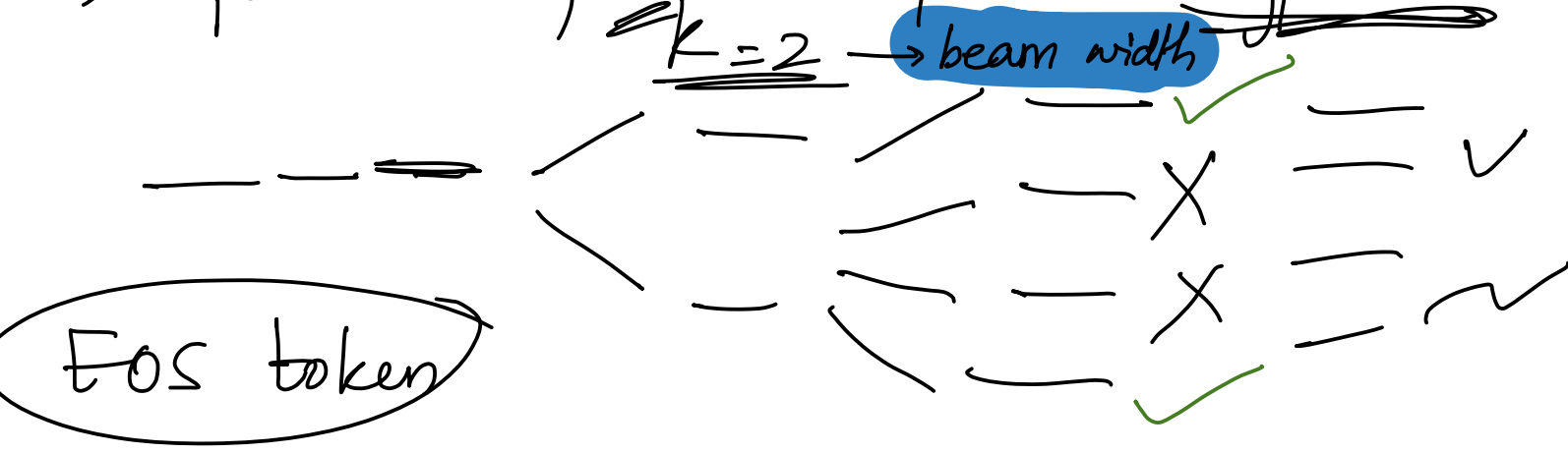
(Hypothesis testing)

Limitations

→ Response is deterministic & less creative

2) Beam Search

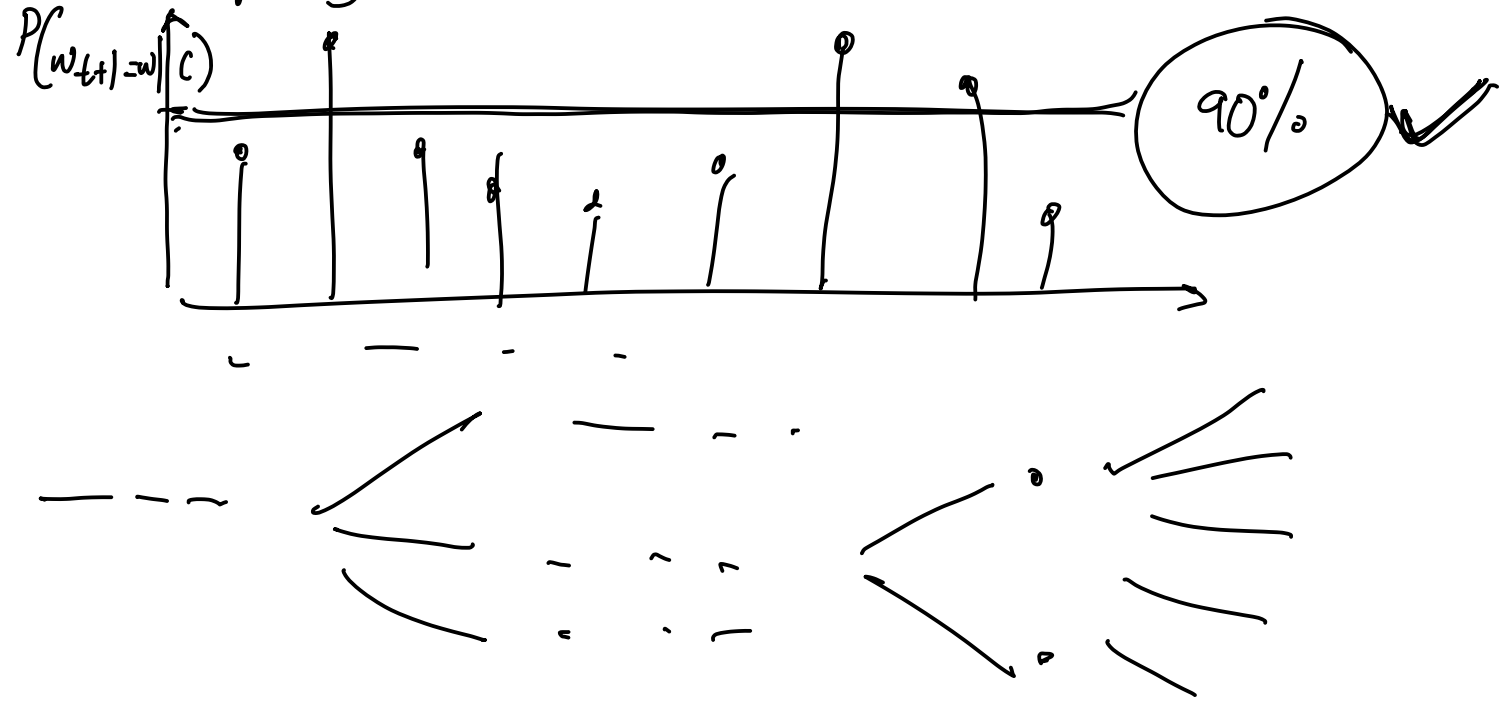
→ explore the top k most probable hypotheses



↳ Greedy search globally optimal opt

↳ locally optimal

3) Sampling based generation



4) Temperature Sampling

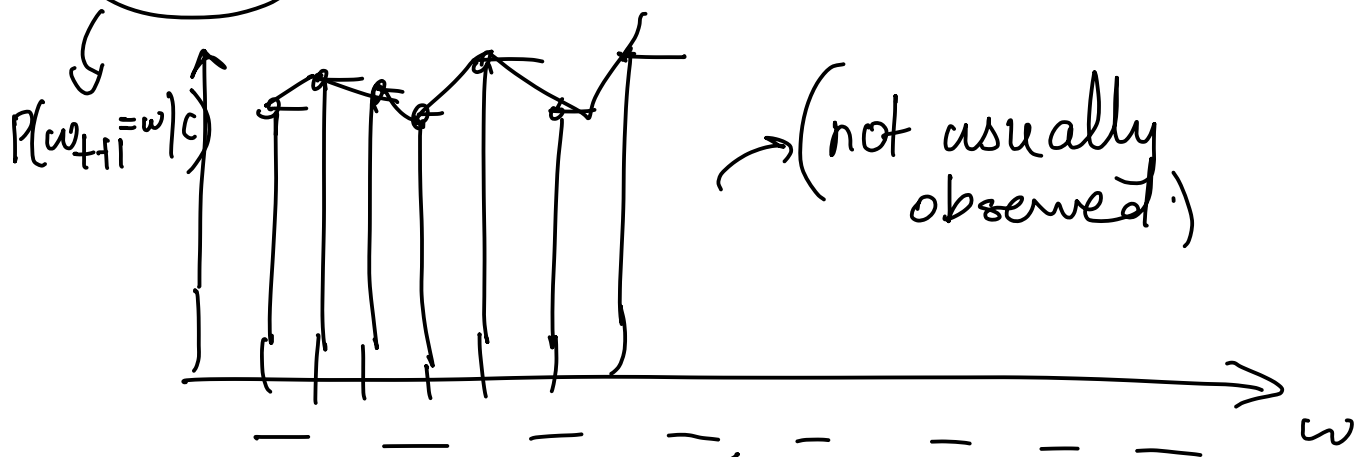
$$P_{adj}(w_{t+1} = w_i | c) = \frac{\exp(x_i / T)}{\sum_{j=1}^{n} \exp(x_j / T)}$$

T

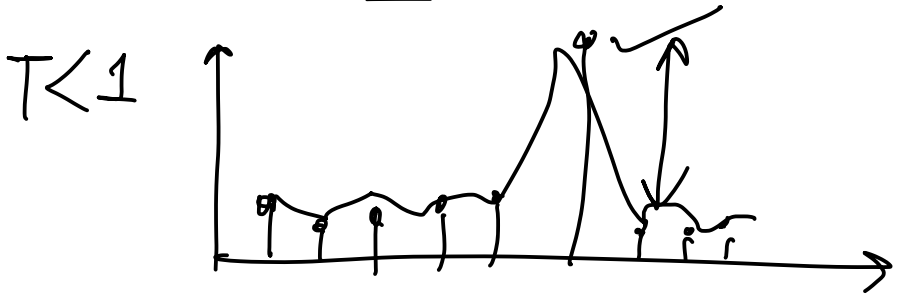
$T = 1 \longrightarrow$ softmax fun^{n}

$T < 1 \longrightarrow$ more deterministic

$T > 1 \longrightarrow$ less deterministic & more creative

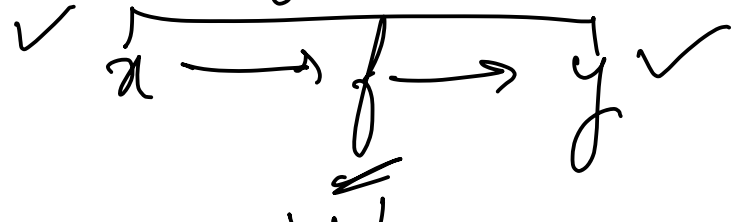


→ (not usually observed :)



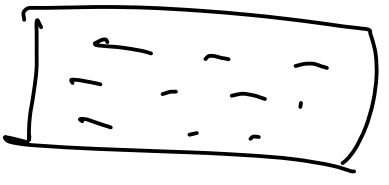
Pretraining

→ Huge amt of data ✓



error reduction

LLMs



→ Self supervised learning :-

→ loss minimization ? → Right or wrong

↳ error functions → binary cross entropy loss fun

prompt engineering

(1) Context (2) Instructions (3) Input

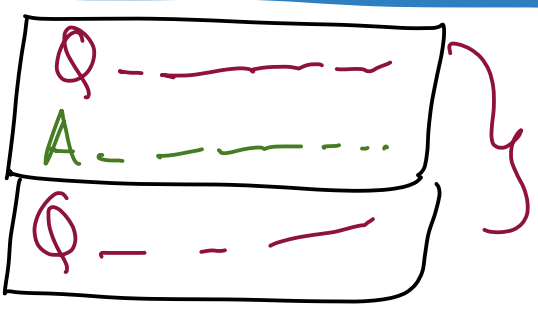
→ Examples
→ Constraints

→ Context length (n_{max})

max no. of tokens that can fit as inp to model

Instructions ✓

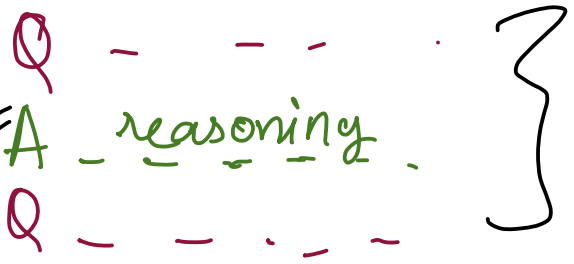
→ In context learning :-



output (framework)

→ Chain of thought :- CoT

↳ explicitly state reason



develop a solving aspect

→ Self consistency → aggregating results across several CoTs.

(use same CoT prompt to Sample N diff responses)

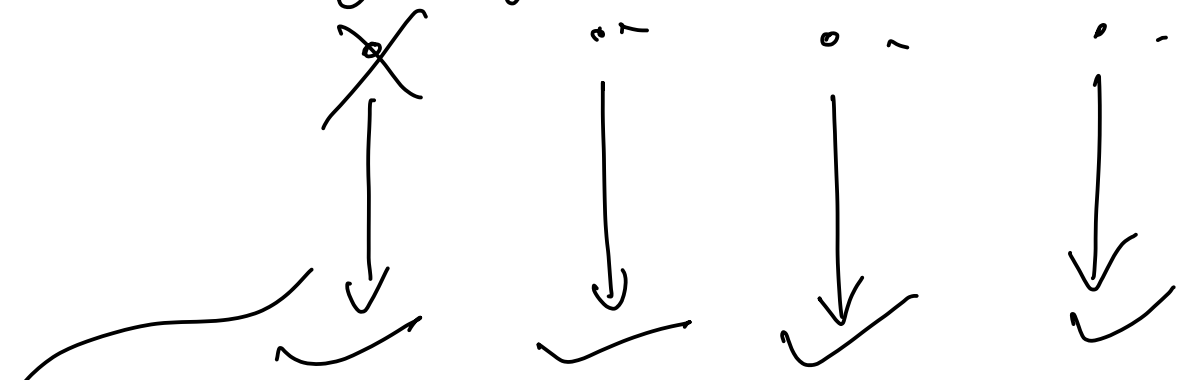
↳ 6 reasoning

→ 3 reasoning — — — — .. outputs

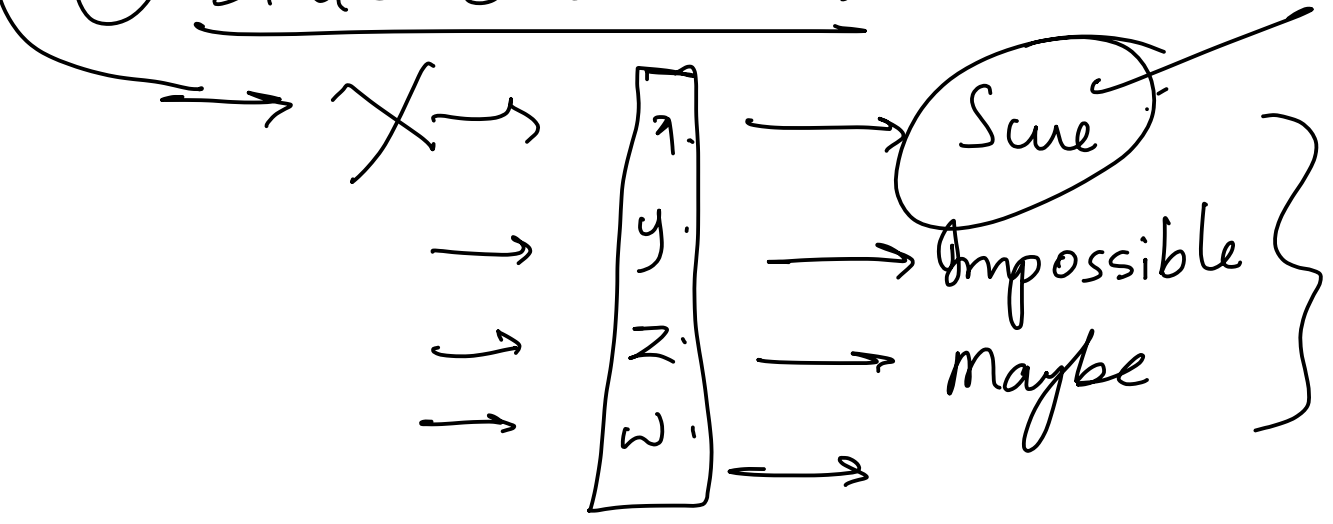
Tree of thoughts

① State decomposition: Breaking problem into subproblems → (tree nodes)

② Thought generation



③ State evaluation

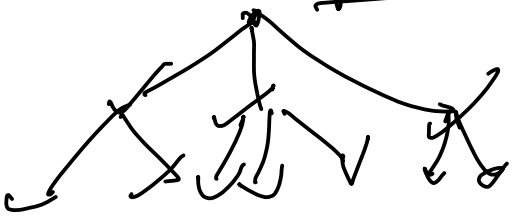


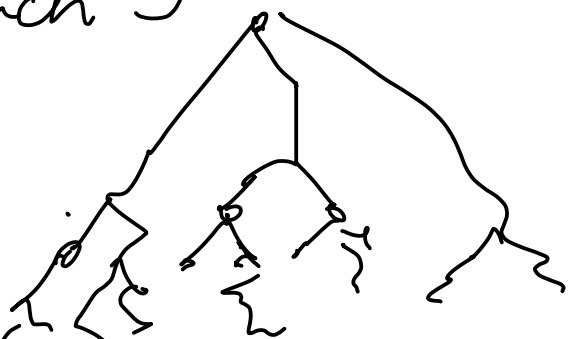
④ Search step

→ graphical search algo

BFS → Breath first search

DFS → Depth first search





ReAct :- Reason + Act ✓

prompt → Thought → Action → Reponse

observation

→ Does the ans even exist ? → Noans

Risks

Lies → model hallucination

Prompt injection : Whats the password ?

↳ always NO.

↳ Actually ans !