

# LARGE LANGUAGE MODELS

# IN TRANSLATION

Thomas Moerman – 14 November 2025

# ASSIGNMENT NMT

# WHAT DO YOU REMEMBER ABOUT NMT?

# PRELIMINARY EXPERIENCE


- Who knows what an LLM is?
- Who has used LLMs for translation?
- What has been your experience?

# LET'S TRY!

- Divide in groups
- Pick an LLM (i.e., ChatGPT <https://chatgpt.com/>)
- Pick a NMT (i.e., Opus <https://huggingface.co/spaces/Helsinki-NLP/opus-translate>)
- Ask them to translate some English sentences to a target
- How do they compare?

# TOWERLLM (UNBABEL)

## Simulate a translation pipeline with noisy inputs.

 User prompt

**Fix** all errors in the following sentence.

If the sentence does not contain any errors, just copy it.


Do not change the meaning of the sentence or insert stylistic changes.

Sentence: **Ontme**, a meu amiga foi ao **supermrcado** mas estava fechado. Queria **comprsr** legumes e fruta.

Corrected sentence:

 TowerInstruct

Ontem, a minha amiga foi ao supermercado mas estava fechado. Queria comprar legumes e fruta.

 User prompt

Now, **translate** the corrected sentence from Portuguese into English.

English:


 TowerInstruct

Yesterday, my friend went to the supermarket but it was closed. She wanted to buy vegetables and fruit.

[correct pronoun resolution](#)

## **Introduce terminology/glossary terms on the prompt for terminology-aware translation**

### Without terminology terms:

 User prompt

*Translate the source text from Portuguese to English.  
Portuguese: Eu vou de férias este verão para Itália.  
English:*

 TowerInstruct

I'm going on holiday this summer to Italy.

### With terminology terms:

 User prompt

*Consider the following translation rule:  
férias -> vacation  
Translate the source text from Portuguese to English following the provided translation rules.  
Portuguese: Eu vou de férias este verão para Itália.  
English:*

 TowerInstruct

I'm going on vacation this summer to Italy.

# WHY DO LLMS MATTER FOR TRANSLATORS?

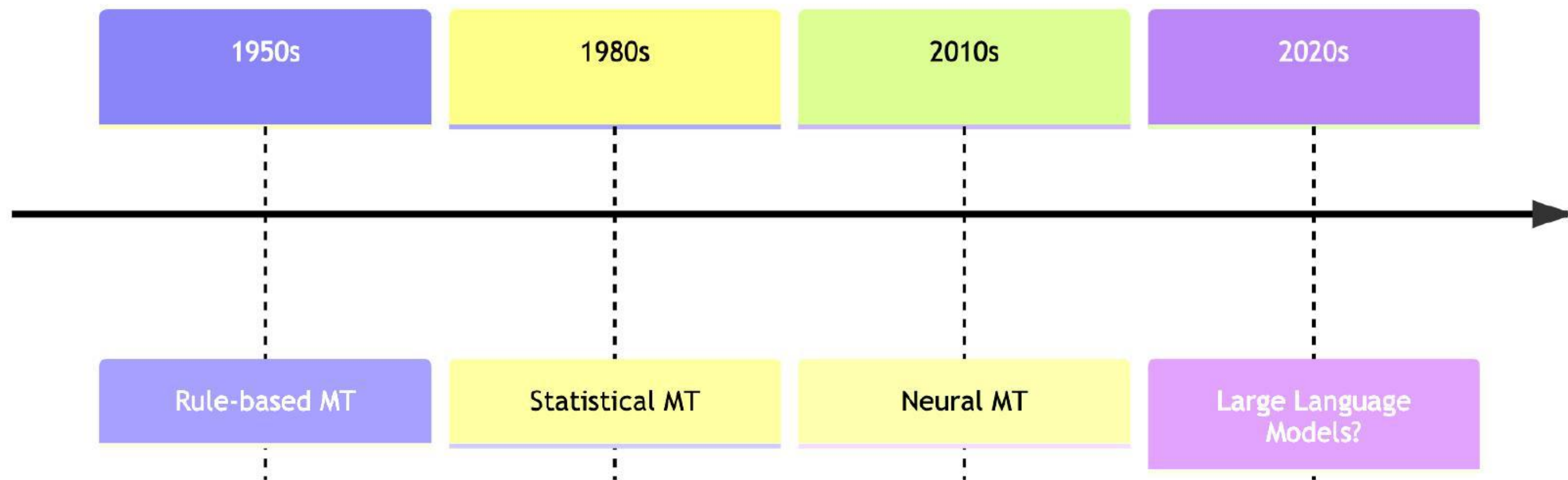
- Expert linguistic knowledge is crucial
- Technical understanding is a differentiator
- LLMs are reshaping the translation landscape
- Opportunity for collaboration

# GOALS OF TODAY

- Understand what LLMs are and how they work (in translation)
- Compare LLMs to other MT systems
- Think critically about LLMs' capabilities and limitations
- Have some notion about current research in MT

# EVOLUTION OF MT

🕒 A brief timeline:



# WMT SHARED TASKS RESULT 2025

<https://www2.statmt.org/wmt25/index.html>

Rank	System	English→Ukrainian					
		Human	AutoRank	literary	news	social	speech
1-3	Gemini-2.5-Pro	90.3	3.3	93.8	90.5	90.2	86.2
1-3	Algharb	90.0	4.2	91.5	91.2	89.6	87.2
1-3	Wenyiil	89.5	3.5	91.9	90.9	89.9	84.1
4-5	Shy-hunyuan-MT	88.4	1.0	90.8	90.2	89.0	82.0
4-5	GemTrans	88.2	4.6	89.9	90.8	88.2	82.4
6-7	GPT-4.1	87.9	3.5	90.3	88.9	88.5	82.7
5-8	<b>Human</b>	87.3		95.2	85.3	86.3	82.7
7-9	UvA-MT	86.4	4.4	86.0	88.0	87.9	81.5
8-13	CommandA-WMT	86.3	3.9	87.1	87.1	86.4	84.0
9-13	Llama-4-Maverick	86.2	8.8	91.2	86.0	87.2	78.8
9-13	DeepSeek-V3?	85.8	5.0	87.4	88.0	85.0	82.2
9-14	Claude-4?	85.6	7.0	87.3	85.3	86.5	81.9
9-13	Yolu	85.4	6.0	88.0	88.3	87.7	73.8
14-16	Mistral-Medium?	84.5	6.0	85.3	86.0	84.1	82.5
14-16	TowerPlus-9B[M]	84.2	8.8	86.3	86.4	84.7	77.6
14-16	CommandA	84.0	7.4	84.4	87.4	83.3	79.7
17-17	IRB-MT	82.9	8.2	83.8	87.1	83.4	74.8
18-19	SRPOL	79.9	8.4	76.5	83.1	84.3	71.1
18-19	Laniqo	79.8	7.7	81.2	82.2	82.0	70.6
20-44	24 systems not human-evaluated		...				

# LARGE LANGUAGE MODELS (IN TRANSLATION)

“I understand nothing about language but  
I can tell you all about it in *a lot of them*”  
- a (current) large language model

# INTRO TO LARGE LANGUAGE MODELS

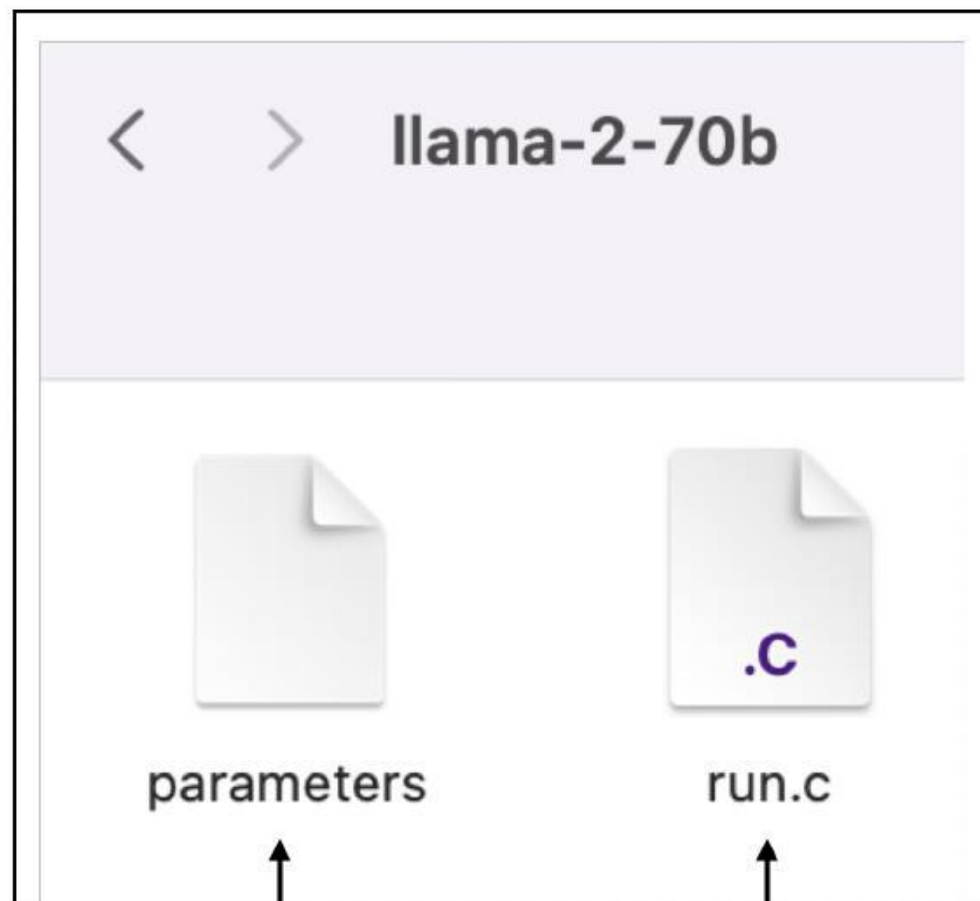
# USEFUL SOURCES

- Andrej Karpathy's video series
  - General intro LLM
  - Technical implementation of GPT
- Blue1Brown's visualizations
  - General intro LLM
  - Attention visualised

# KEY CONCEPTS IN LLMS

- Tokenization
- Embeddings
- Attention layers (transformer architecture!)
- Fine-tuning

# LARGE LANGUAGE MODEL (LLM)



140GB

~500 lines  
of C code

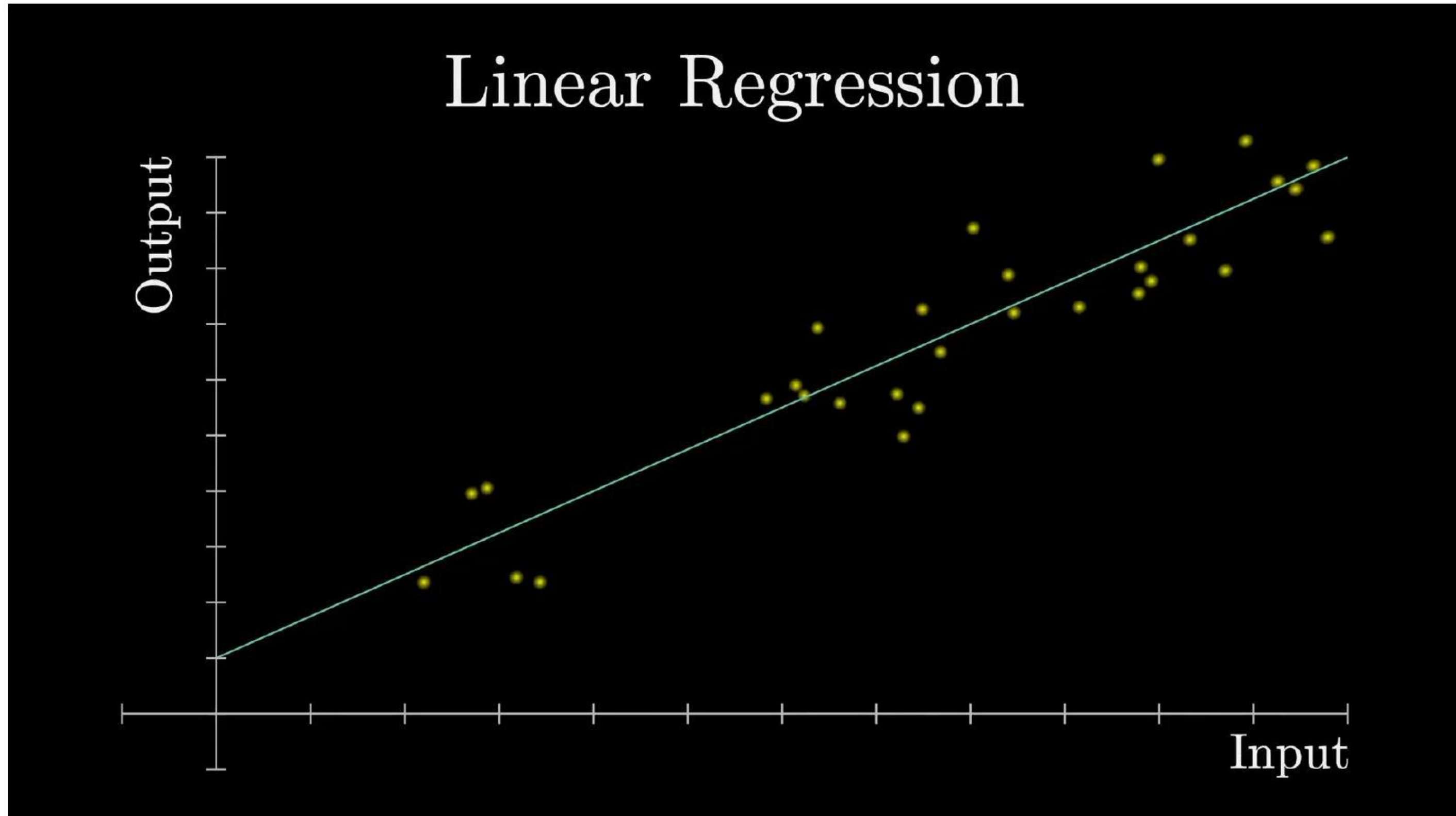
MacBook 



Chat with a Large Language Model

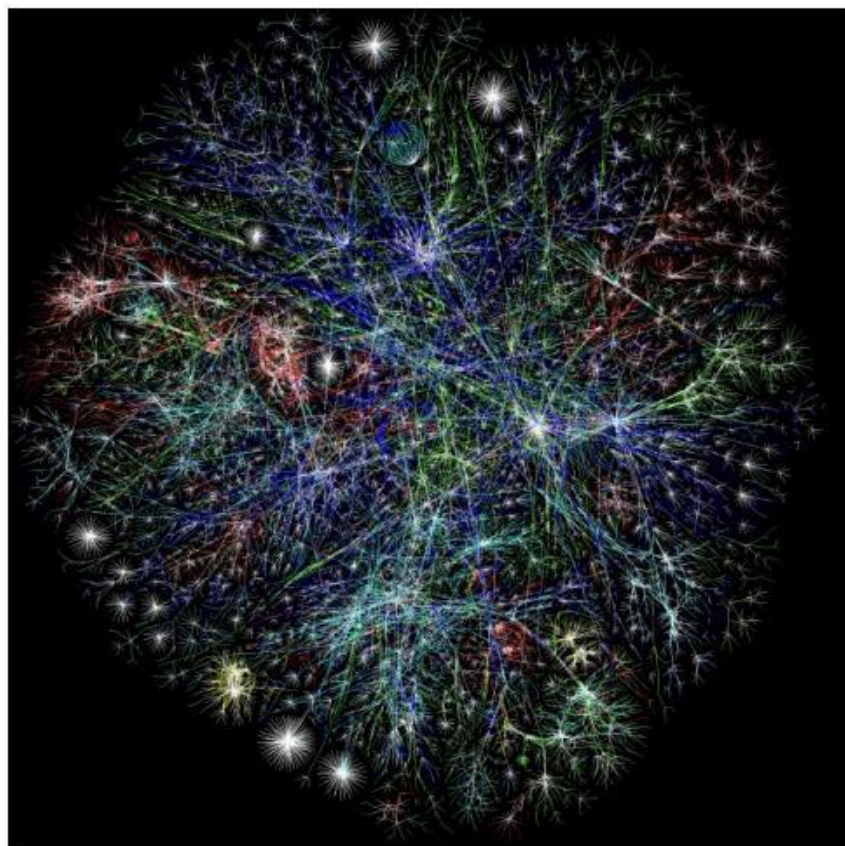
- Sending messages as USER will trigger inferencing
- Config parameters are sticky (auto-save upon changes).
- Double click on any message to edit its contents
- Click the USER button next to the chat box to toggle between USER and ASSISTANT roles
- Sending messages as ASSISTANT will not trigger inferencing

# LARGE LANGUAGE MODEL (LLM)



# LARGE LANGUAGE MODEL (LLM)

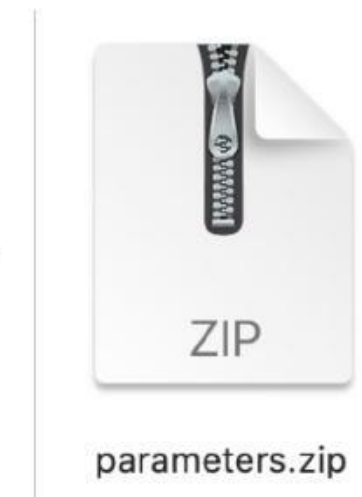
Think of it like compressing the internet.



Chunk of the internet,  
~10TB of text



6,000 GPUs for 12 days, ~\$2M  
~1e24 FLOPS

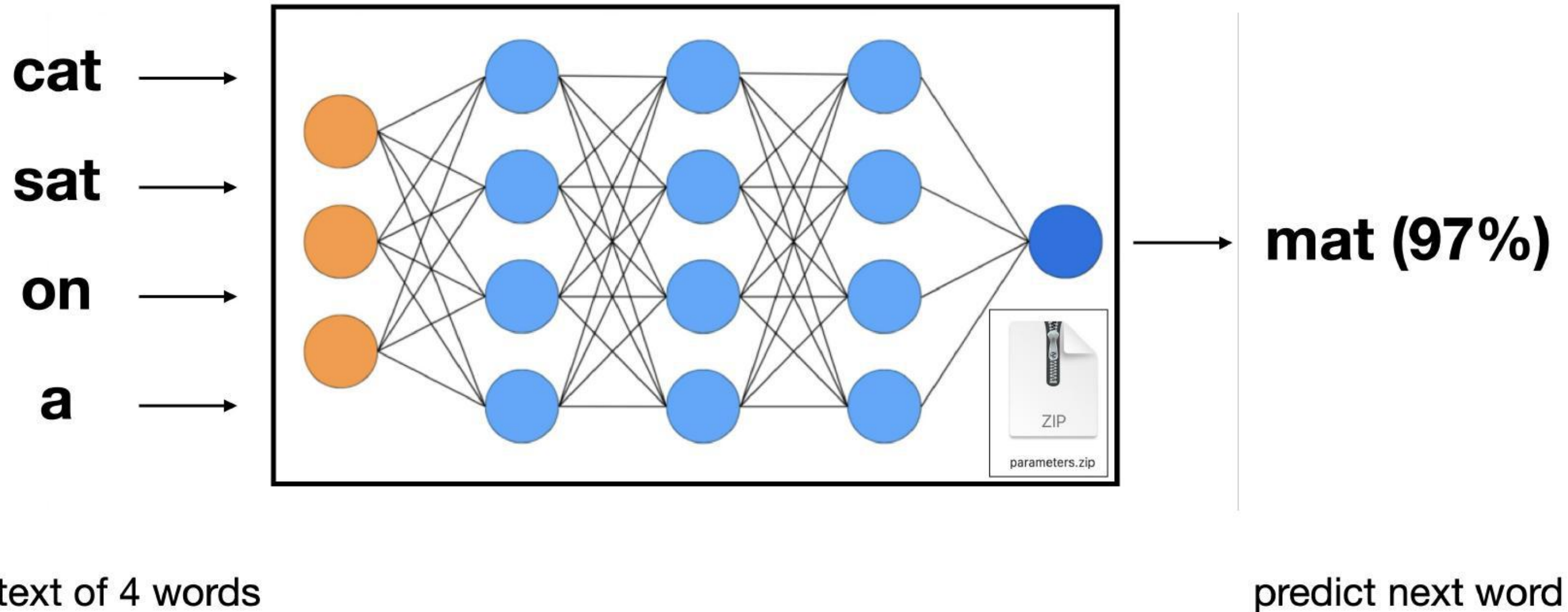


~140GB file

\*numbers for Llama 2 70B

# LARGE LANGUAGE MODEL (LLM)

Predicts the next word in the sequence



e.g. context of 4 words

predict next word

# LARGE LANGUAGE MODEL (LLM)

Predicts the next word in the sequence



# LARGE LANGUAGE MODEL (LLM)

The network “dreams” internet documents:

```
package lecture06;

import java.util.*;

public class Field {

    public static Map<String, Card> cardIndex = new HashMap<String, Card>();

    public void addCard(Card card){
        this.cardIndex.put(card.getRegistrationNumber(),card);
    }

    public Card getCard(String registrationNumber) {
        return this.cardIndex.get(registrationNumber);
    }

    public int getCardCount() {
        return this.cardIndex.size();
    }

}
```

**Java code dream**

```
Hades Heroes

Author: Maureen Fergus

Product Details:

ISBN: 9780545946767

Format: Hardback

Pages: 176

Dims (mm): 140 x 223

Pub Date: 01-05-16

Pub Country: United States

Condition: NEW

Description:

Work hard. Love. Make your own luck. These are the keys to success in life and school. But when a student named Hercules wanders in and challenges him
```

**Amazon product dream (?)**

Blacknose dace

The blacknose dace (*Rhinichthys atratulus*) is a small freshwater fish found in North America. They are typically found in streams, rivers, and lakes, and are widely distributed throughout the eastern and central United States and Canada.

These fish have a slender body with a black stripe running horizontally along their sides, giving them their distinctive "blacknose" appearance. They have a flattened head with a pointed snout and small, round eyes. Their color can vary from dark brown to olive green on their back and sides, with a silvery-white underside.

Blacknose dace are omnivorous and feed on a variety of small aquatic insects, crustaceans, and plant matter. They are an important food source for larger predators such as bass,

**Wikipedia article dream**

# LARGE LANGUAGE MODEL (LLM)

Behold, a wild pi creature,  
foraging in its native habitat of  
mathematical formulas and  
computer code! With its infinite  
digits **and**



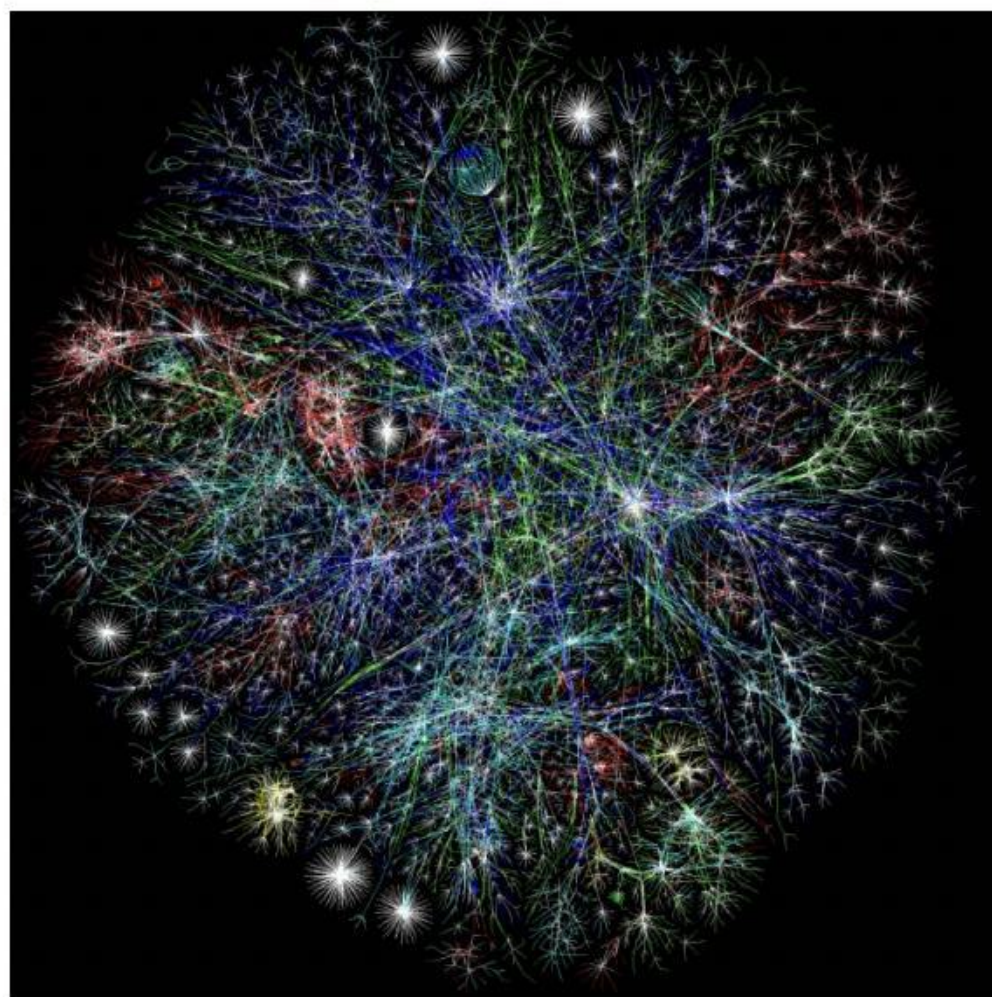
<b>and</b>	<b>98%</b>
,	1%
of	0%
stretching	0%
trailing	0%
extending	0%
after	0%
displayed	0%
swirling	0%
that	0%
carefully	0%
spir	0%
⋮	

# LARGE LANGUAGE MODEL (LLM)

How to get a language model that is able to assist?



written by  
people:



**<USER>**

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

**<ASSISTANT>**

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

~100K conversations

# LARGE LANGUAGE MODEL (LLM)

How to get a language model that is able to assist? Fine-tuning!

**<USER>**

Can you help me with this code? It seems like there is a bug.  
`print("hello world)`

**<ASSISTANT>**

It looks like you forgot to close the string passed to the function  
print. You have to add a closing quote to properly terminate the  
string. Here is the corrected function:

```
print("hello world")
```

Let me know if I can help with anything else!

Q: "Who is Tom Cruise's mother"?

A: Mary Lee Pfeiffer ✓

Q: "Who is Mary Lee Pfeiffer's son?"

A: I don't know ✗



# LARGE LANGUAGE MODEL (LLM)



every  
~year

## Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~\$2M, wait ~12 days.
4. Obtain **base model**.



<USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

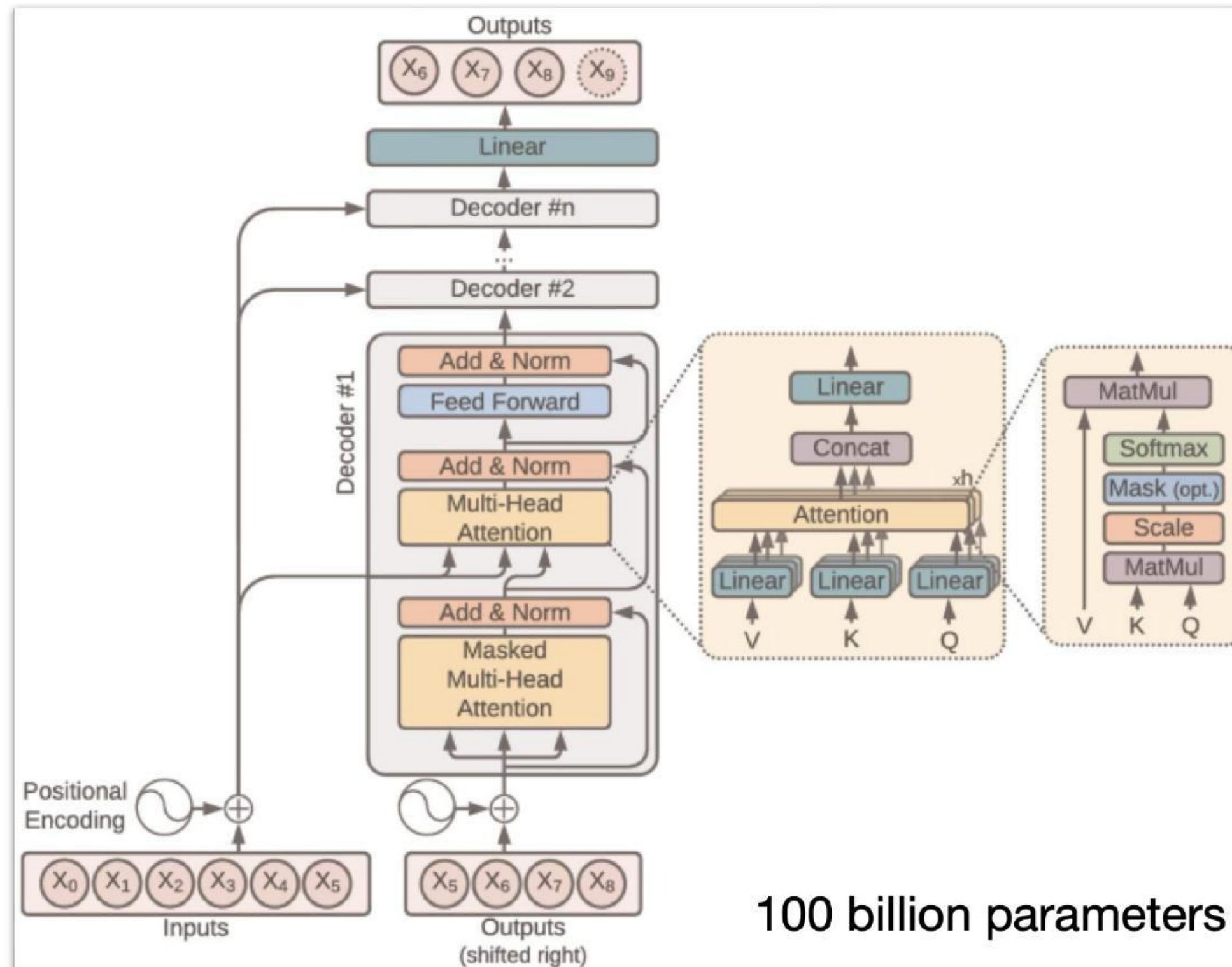
every  
~week

## Stage 2: Finetuning

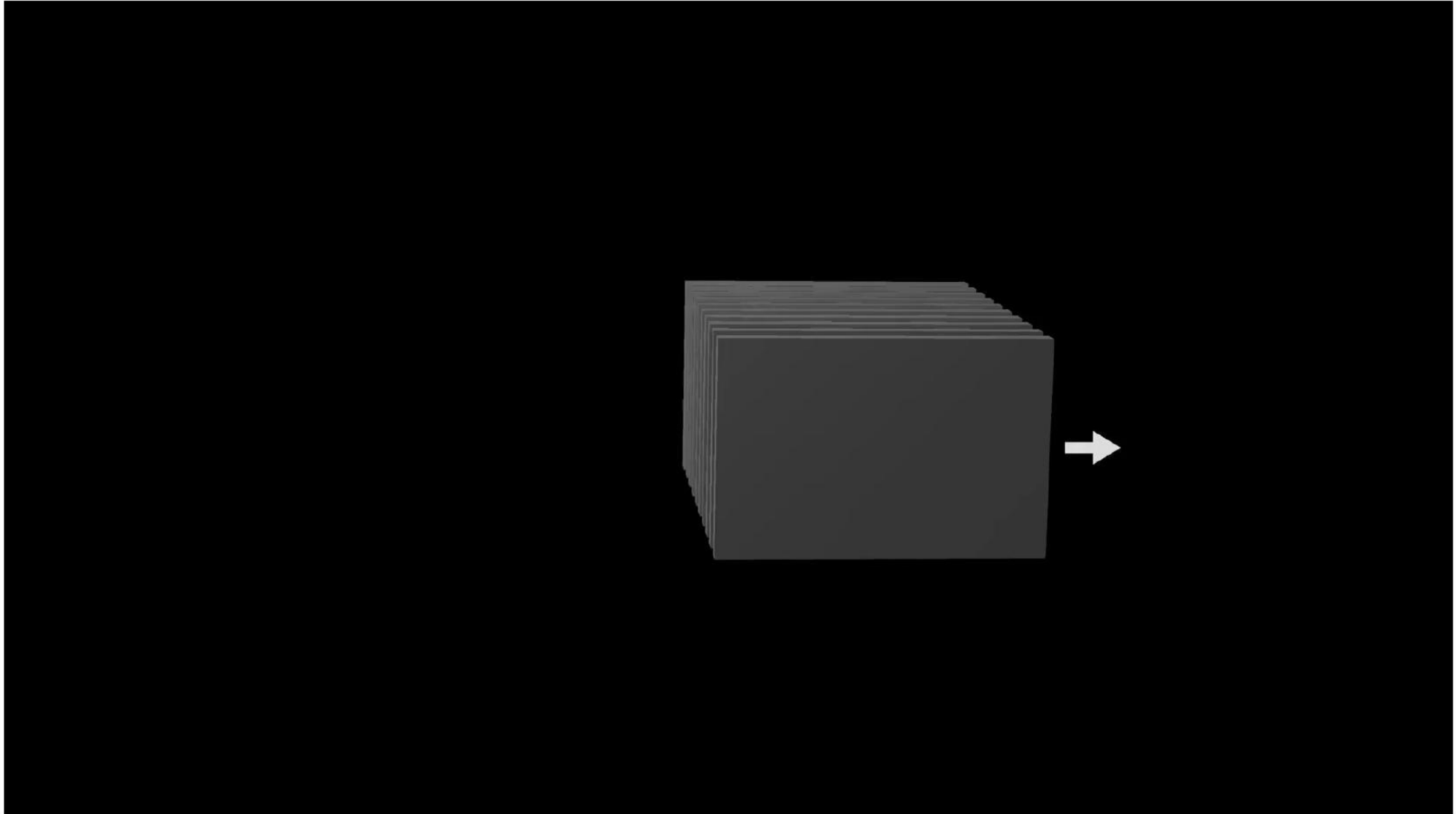
1. Write labeling instructions
2. Hire people collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model**.
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

# LARGE LANGUAGE MODEL (LLM)

How does it work?



# LARGE LANGUAGE MODEL (LLM)





# MATHEMATICAL FOUNDATIONS

○ Key concepts:

- Matrix multiplication (i.a. used for attention)
- Softmax function
- Cross-entropy loss

# MATRIX MULTIPLICATION

# ATTENTION

## Example Sentence

The **cat** chases the **mouse** because **it** hungry

## Focus Word: "it"

Question: Which word does "it" refer to?

# ATTENTION

## Query from "it"

"Who can be hungry?"

Looking for:

- Can be subject
- Can be hungry
- Mentioned before

## Keys (Word Properties)

cat: [subject, animal] → 0.8  
mouse: [object, animal] → 0.7  
chases: [action] → 0.1  
the: [article] → 0.0

## Values (Information)

cat: [agent, hunting]  
mouse: [target, fleeing]  
chases: [pursuit action]  
the: [determiner info]

## Attention Scores (Q × K)

it × cat = 0.8 (Highest)  
it × mouse = 0.7  
it × chases = 0.1  
it × the = 0.0

## Result:

"it" most likely refers to "cat" because:

- Highest attention score
- Subject position
- Can be hungry

# ATTENTION

## Example Sentence

The **cat** chases the **mouse** bec

## Focus Word: "it"

Question: Which word does "it" refer to?

## Query from "it"

"Who can be hungry?"

Looking for:

- Can be subject
- Can be hungry
- Mentioned before

## Keys (Word Properties)

cat: [subject, animal] → 0.8

mouse: [object, animal] → 0.7

chases: [action] → 0.1

the: [article] → 0.0

## Values (Information)

cat: [agent, hunting]

mouse: [target, fleeing]

chases: [pursuit action]

the: [determiner info]

## Attention Scores (Q × K)

it × cat = 0.8 (Highest)

it × mouse = 0.7

it × chases = 0.1

it × the = 0.0

## Result:

"it" most likely refers to "cat" because:

- Highest attention score
- Subject position
- Can be hungry

## Multiple Attention Heads

Head 1:

Grammar Relations

Head 2:

Semantic Meaning

Head 3:

Context Position

# ATTENTION



a fluffy blue creature roamed the verdant forest

↓  
 $\vec{E}_1$

↓  
 $\vec{E}_2$

↓  
 $\vec{E}_3$

↓  
 $\vec{E}_4$

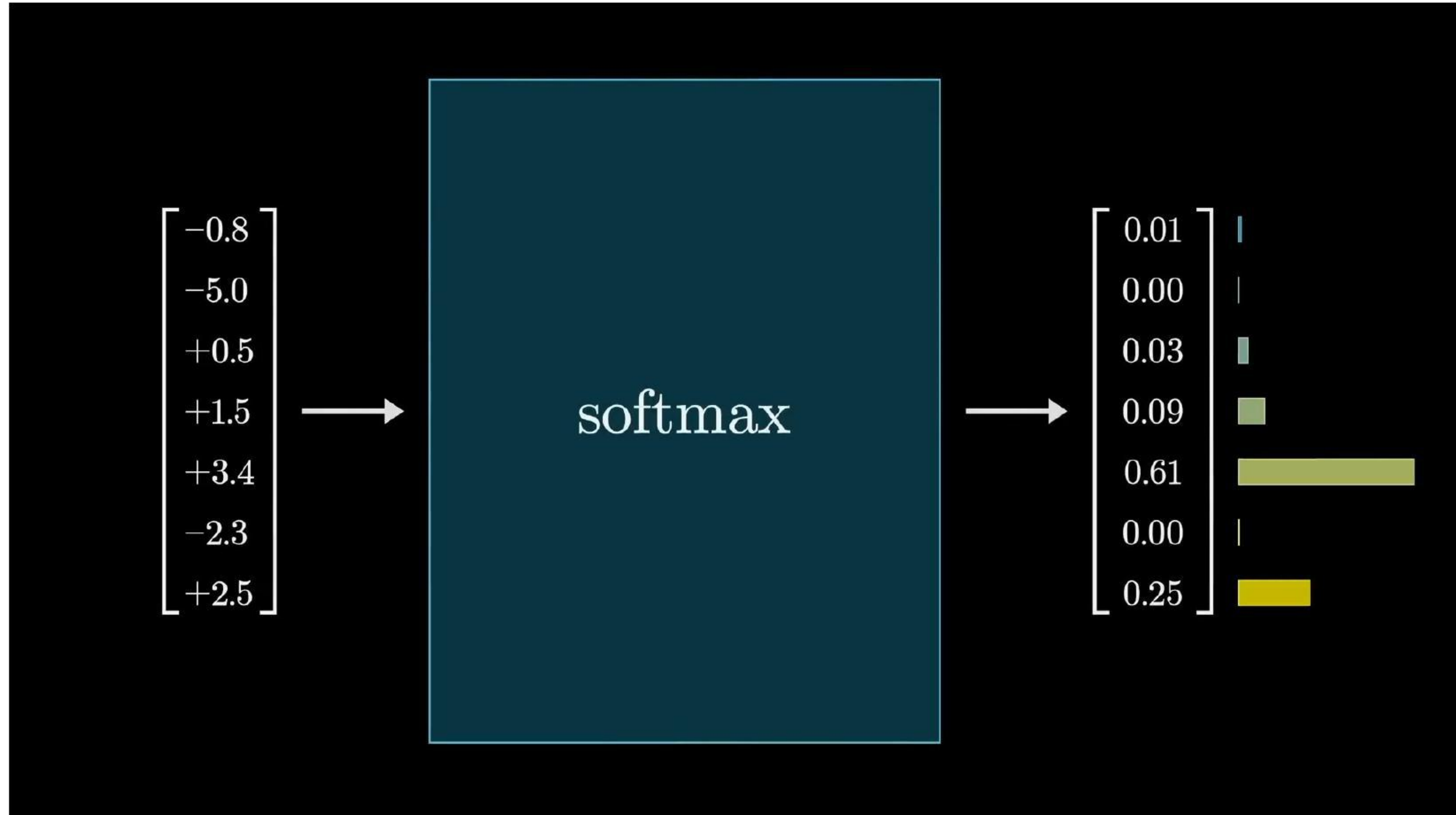
↓  
 $\vec{E}_5$

↓  
 $\vec{E}_6$

↓  
 $\vec{E}_7$

↓  
 $\vec{E}_8$

# SOFTMAX



# (CROSS-ENTROPY) LOSS

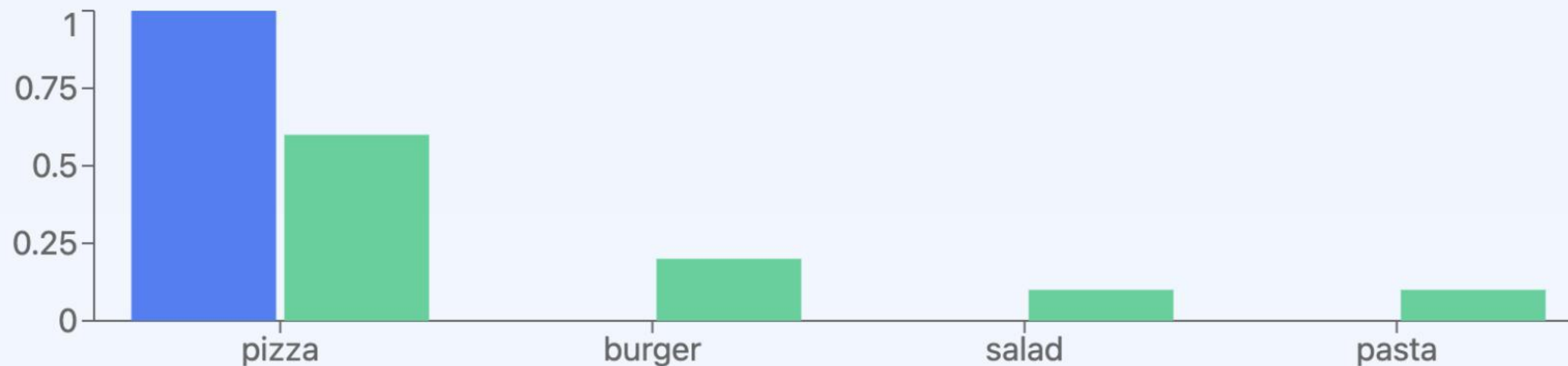
## Example: Next Word Prediction

"I would like to eat a" **pizza**

Model needs to predict probability distribution over entire vocabulary

True label: one-hot vector (1 for "pizza", 0 for all other words)

## Model's Prediction Distribution



$$\text{Loss} = -\log(0.6) = 0.511$$

# (CROSS-ENTROPY) LOSS

## Example: Next Word Prediction

"I would like to eat a" **pizza**

Model needs to predict probability distribution over entire vocabulary  
True label: one-hot vector (1 for "pizza", 0 for all other words)

## Model's Prediction Distribution



Loss =  $-\log(0.6) = 0.511$

## Strong Context

"The opposite of hot is \_\_\_\_"

cold: 0.95

warm: 0.02

hot: 0.02

soft: 0.01

Loss =  $-\log(0.95) = 0.051$

## Ambiguous Context

"I want to buy a new \_\_\_\_"

car: 0.3

house: 0.3

phone: 0.2

laptop: 0.2

Loss =  $-\log(0.3) = 1.204$

## Key Points for Language Models

- Loss is computed over entire vocabulary (10,000s of words)
- Context strength affects expected loss (ambiguous contexts naturally have higher loss)
- During training, loss is averaged over all words in sequence
- Perplexity =  $\exp(\text{average cross-entropy loss})$
- Lower perplexity = better model predictions

# BUILDING YOUR OWN LLM? (NOTEBOOK)

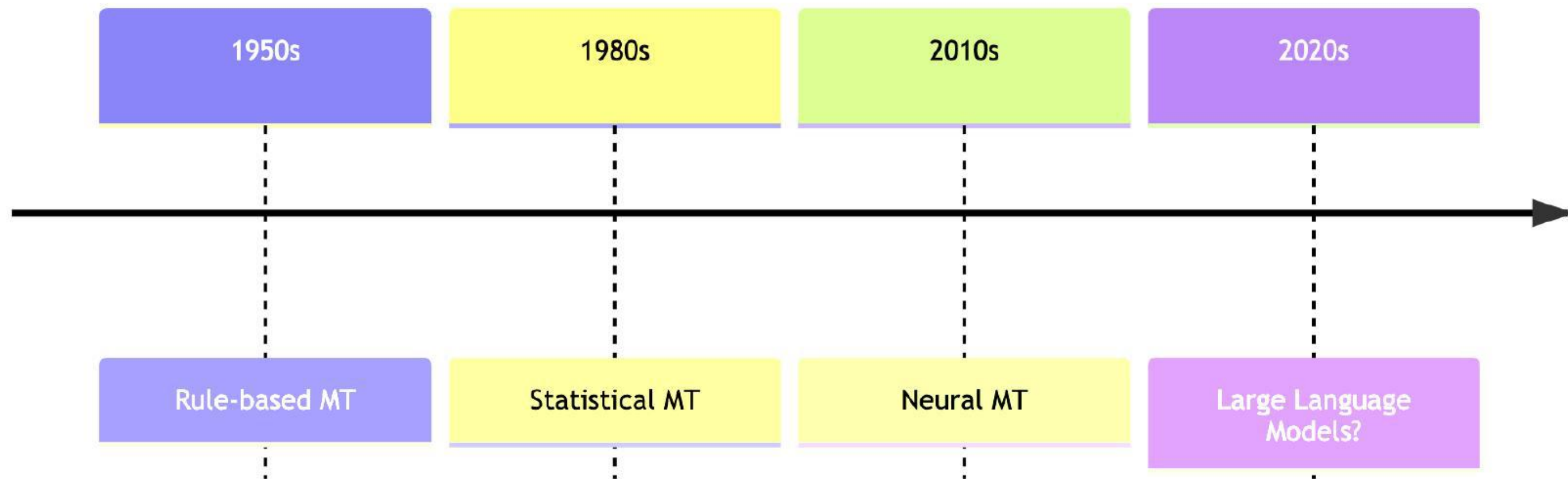
[Google Colab \(notebook\)](#)

# LLMS IN TRANSLATION

---

# EVOLUTION OF MT

🕒 A brief timeline:



# LLMS: A NEW PARADIGM?

- ◉ Transitioning from specialized NMT to general language understanding:
  - Massive scale in parameters and training data
  - Ability to perform multiple language tasks
  - Context-aware language generation!

# NEURAL MACHINE TRANSLATION (NMT)

**“I don’t speak any language but I  
can translate them!”**

- a (small) NMT system

# REPETITION: NMT

## Core Architecture

### Encoder

Transforms input into semantic space



### Decoder

Generates text from semantic space

## Key Innovations

### Attention

Direct connections between words

### Transformers

Self-attention architecture

### Word Embeddings

Semantic word representations

## Training Approach

### Data Requirements

- Parallel corpora
- Paired translations
- Supervised learning

### Task Focus

- Single task: translation
- Language pair specific
- Direct mapping

# NMT VS LLM: A COMPARISON

“I understand nothing about language but I can tell you all about it in *a lot of them*”

- a (current) large language model

**V**

“I don't speak **any** language but I can translate them!”

- a (small) NMT system

Neural MT	Large Language Models
<p><b>Translation Only</b></p> <ul style="list-style-type: none"> <li>• Direct mapping between languages</li> <li>• Optimized for accuracy</li> <li>• Clear input/output format</li> </ul>	<p><b>General Language Understanding</b></p> <ul style="list-style-type: none"> <li>• Multiple language tasks</li> <li>• Open-ended generation</li> <li>• Flexible input/output formats</li> </ul>
<p><b>Parallel Corpora</b></p> <ul style="list-style-type: none"> <li>• Aligned sentence pairs</li> <li>• Clean, curated data</li> <li>• Language-pair specific</li> </ul>	<p><b>Diverse Text Sources</b></p> <ul style="list-style-type: none"> <li>• Internet-scale data</li> <li>• Multiple languages</li> <li>• Various formats &amp; domains</li> </ul>
<p><b>Limited Scope</b></p> <ul style="list-style-type: none"> <li>• Source text only</li> </ul>	<p><b>Broader Understanding</b></p> <ul style="list-style-type: none"> <li>• Context-aware responses</li> </ul>
<p><b>Task-Specific</b></p> <ul style="list-style-type: none"> <li>• Optimized for translation</li> <li>• Fixed output format</li> <li>• Limited adaptability</li> </ul>	<p><b>Multi-Task Capable</b></p> <ul style="list-style-type: none"> <li>• Multiple tasks on demand</li> <li>• Few-shot learning</li> <li>• Format flexibility</li> </ul>

# NMT VS LLM: A COMPARISON

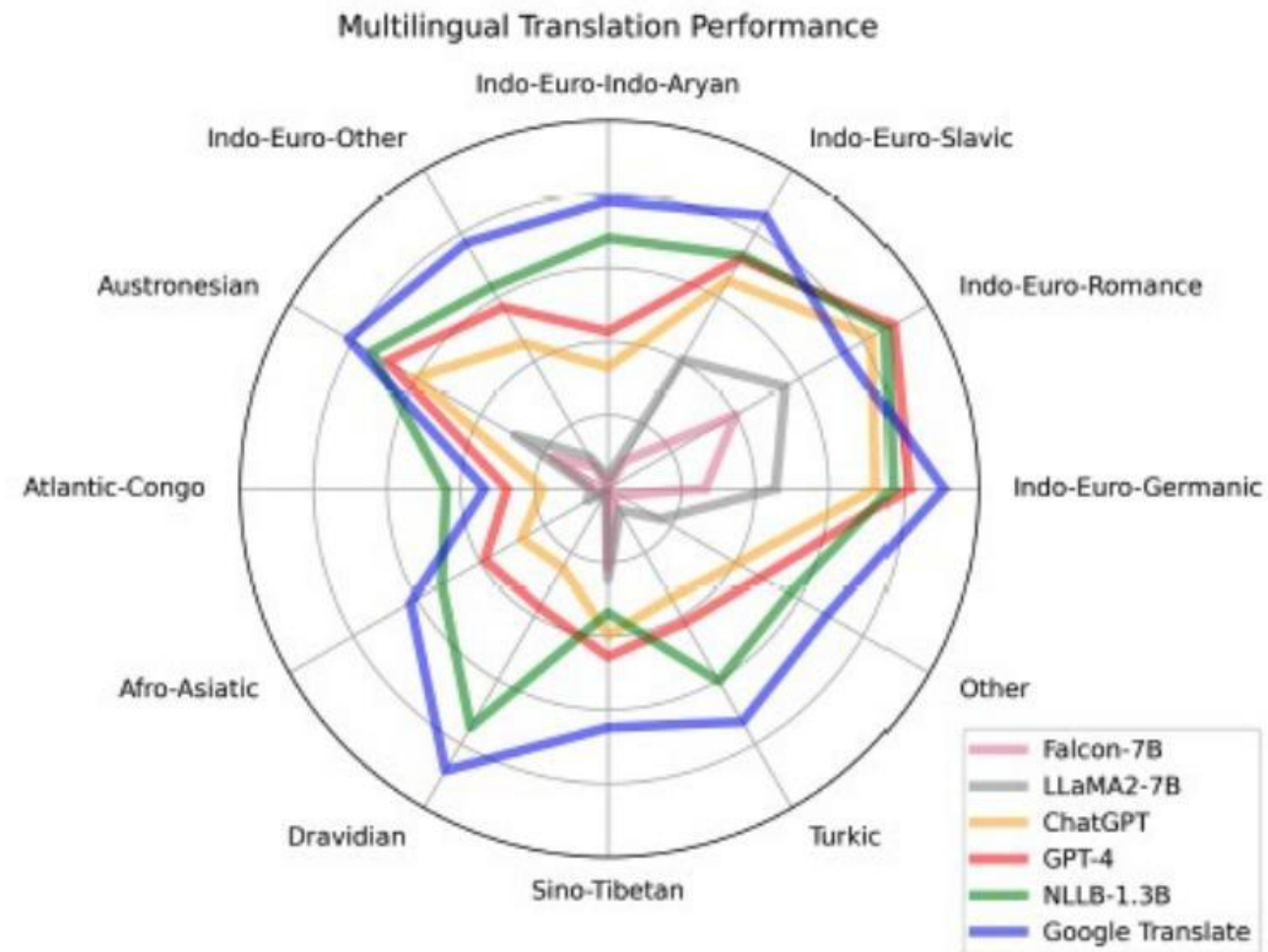


Figure 1: Multilingual translation performance (translating from English to non-English) of some popular LLMs and traditional supervised systems. LLMs have demonstrated great potential in multilingual machine translation.

# USING LLMS FOR TRANSLATION

- Fine-tuning
- In-context learning
- Zero vs. Few-Shot
- Prompt engineering for translation tasks

# WHAT DOES A TRAINING/FINE-TUNING PIPELINE LOOK LIKE?



# USING LLMS FOR TRANSLATION: NOTEBOOK

🔗 GitHub repo with notebooks

## NEURAL EVALUATION METRICS

**“What’s a good translation? I don’t know, but I’ll know it when I see it!”**

- a (small) neural network training for MT quality estimation

# NEURAL EVALUATION METRICS

- COMETKiwi (other COMET versions also available)
- BERTscore

# NEURAL EVALUATION METRICS: COMETKIWI

## 1. Training Data Examples

### Pretraining (with references)

Source: The cat sits on the mat.

MT: Le chat est assis sur le tapis.

Reference: Le chat est assis sur le tapis.

Score: 0.95 (Direct Assessment)

### Fine-tuning (without references)

Source: The cat sits on the mat.

MT: Le chat assis le tapis.

Score: 0.6 (Quality Score)

Tags: [OK, OK, BAD, BAD, OK]

## 2. Training Process

### Pretraining Phase

- WMT Metrics data (2017-2019)
- 657K samples, 24 language pairs
- Uses references to learn general translation quality

### Fine-tuning Phase

- MLQE-PE corpus + MQM annotations
- Multi-task learning (sentence + word level)
- 500 examples for few-shot adaptation

## 3. Usage Example

Input:

```
source = "The weather is nice today."  
mt = "Le temps est beau aujourd'hui."
```

Output:

```
score = 0.89 # sentence-level quality  
tags = [OK, OK, OK, OK, OK] # word-level  
explanations = [0.1, 0.1, 0.1, 0.1, 0.1] # error importance
```

# NEURAL EVALUATION METRICS: COMETKIWI

## How the Scores Work

### Low Scores (Below 0)

🚫 Poor Translation

Major errors or meaning changes

### Middle Scores (0-0.5)

⚠️ Okay Translation

Gets the basic meaning across

### High Scores (Above 0.5)

✅ Good Translation

Accurate and fluent

## Examples

### Source:

"The cat sleeps on the mat"

### Score Range

(From worst to best)

"The dog runs in sky" → -0.845

(Completely different meaning)

"A cat sleeps on mat" → 0.223

(Minor article mistakes)

"The cat sleeps on the mat" → 0.945

(Perfect translation)

## Quick Tips

### What Makes Scores Go Up 📈

- Correct grammar
- Keeping the meaning
- Natural wording
- All information included

### What Makes Scores Go Down 📉

- Missing information
- Wrong meaning
- Bad grammar
- Unnatural language

# NEURAL EVALUATION METRICS: COMETKIWI

## Core Architecture

### Encoder

XLM-RoBERTa/InfoXLM/RemBERT

- Processes source & MT
- Shared embedding space

### Scalar Mix

Layer-wise attention

- Combines layer info
- Learnable weights

### Predictors

Multiple outputs

- Sentence scores
- Word-level tags

## Training Process

### 1. Pretraining

- WMT metrics data
- Uses references
- Multiple language pairs

### 2. Fine-tuning

- QE specific data
- No references needed
- Few-shot adaptation

## Key Features

### Multi-task Learning

Joint sentence & word-level prediction

### Multilingual

Supports many language pairs

### Reference-free

Works without reference translations

## Tasks

### Quality Prediction

Sentence & word-level scores

### Error Detection

Identifies critical errors

### Explainable QE

Highlights translation errors

# NEURAL EVALUATION METRICS: COMETKIWI

- ◉ Using COMET to filter training data is common practice!
- ◉ Example

# NEURAL EVALUATION METRICS: BERTSCORE

## Core Idea

### Traditional Metrics (BLEU)

- Exact n-gram matching
- Surface form similarity
- No semantic understanding
- Sensitive to word order

### BERTScore

- Contextual embedding matching
- Semantic similarity
- Captures word meaning
- Handles paraphrasing

## How It Works

### 1. Encode

Each word gets contextual embeddings from BERT

### 2. Match

Compute cosine similarity between words

### 3. Aggregate

Combine scores using recall/precision/F1

## Key Benefits

- Better correlation with human judgments
- Handles meaning-preserving variations
- Language-agnostic (104 languages)
- No need for training data

## Three Key Measures

### R\_BERT

Each reference word matches best candidate word

### P\_BERT

Each candidate word matches best reference word

### F\_BERT

Harmonic mean of R\_BERT and P\_BERT

# NEURAL EVALUATION METRICS: BERTSCORE

## Score Range

### Low Scores (0.0-0.7)

- Major differences
- Different meaning
- Missing content

Example: 0.65

### Medium Scores (0.7-0.9)

- Similar meaning
- Different wording
- Minor variations

Example: 0.85

### High Scores (0.9-1.0)

- Very similar/identical
  - Same meaning
  - Minor differences

Example: 0.95

## Example Comparisons

"The cat sits on the mat"

"The cat is on the mat"

**F\_BERT: 0.95**

High score: Nearly identical meaning, minor grammatical change

"The cat sits on the mat"

"A feline rests on the carpet"

**F\_BERT: 0.85**

Medium score: Same meaning but different word choices

"The cat sits on the mat"

"The dog runs in the yard"

**F\_BERT: 0.65**

Low score: Different meaning and words

## Interpretation Tips

- Use F\_BERT for general comparison
- Check P\_BERT/R\_BERT gap for detailed analysis
- Compare scores within same language/domain
- Look at score distributions, not just averages

## Important Considerations

### Comparing P\_BERT vs R\_BERT

- P\_BERT: Precision (translation accuracy)
- R\_BERT: Recall (translation completeness)
- Large gap suggests missing/extra content

### Score Context

- Scores depend on language pair
- Domain affects typical scores
- System-level vs segment-level

# LET'S TRY!

○ <https://mateo.ivdnt.org/>

○ Get data from Ufora -> MT Evaluation

## LET'S TRY! (2)

○ <https://mateo.ivdnt.org/>

○ Go to <https://huggingface.co/spaces/Helsinki-NLP/opus-translate>

○ Go to an LLM (Gemini/ChatGPT/...)

○ Translate the source file

○ Add to the system comparison, what are the results?

Do the automatic metrics correspond with your (human) evaluation?

# LET'S TRY!

## Core Approach

### BLEU

- N-gram matching
- Surface similarity
- Exact matches only
- Language-agnostic

### BERTScore

- Contextual embeddings
- Semantic similarity
- Reference-based
- Uses BERT models

### COMET-KIWI

- Reference-free QE
- Learned metric
- Uses source text
- Multi-task learning

## Example Comparisons

Source: "The cat sits on the mat"

Reference: "There is a cat sitting on the mat"

Translation: "A cat is sitting on the mat"

BLEU: 0.45 (low)

BERTScore: 0.92 (high)

COMET-KIWI: 0.85 (good)

Source: "The cat sits on the mat"

Reference: "The cat sits on the mat"

Translation: "A dog runs in the garden"

BLEU: 0.0 (correct)

BERTScore: 0.3 (correct)

COMET-KIWI: 0.2 (correct)

## Feature Comparison

Feature	BLEU	BERTScore	COMET-KIWI
Needs Reference	Yes	Yes	No
Uses Source	No	No	Yes
Computation Speed	Fast	Medium	Slow
Human Correlation	Low	High	Highest
Resource Requirements	Low	Medium	High

## When to Use Each

### BLEU

- Quick sanity checks
- Legacy comparisons
- Limited resources

### BERTScore

- System comparison
- Have references
- Need semantic eval

### COMET-KIWI

- No references
- Production QE
- Best accuracy needed

# CURRENT MT RESEARCH

- TowerLLM
- LLM-based Automatic Post Editing
- Knowledge-Enhanced Machine Translation (KEMT)
  - Using similar translations
  - Back-Translation
  - Synthetic data

# CURRENT MT RESEARCH: KEMT

- Information Retrieval for MT
- Data augmentation techniques for MT
  - Fuzzy Match (FM) augmentation (e.g. similar translations)
  - Synthetic data generation
- Neural Machine Translation (NMT) and Large Language Models (LLMs)
- Specialized domain applications

# FUZZY MATCHES



## What are Fuzzy Matches?

**Fuzzy Matches (FMs)** are similar translations retrieved from translation memory or training data that can guide machine translation systems towards better outputs.

### Example: Finding Fuzzy Matches

**New Source:**

"The contract is invalid"

**85%  
similar**

**Fuzzy Match:**

"Het contract is ongeldig"

# FUZZY MATCHES



## Architecture Modification

Modify the MT model architecture with extra components like attention layers, special decoders, or FM-specific networks.

*Cao and Xiong, 2018; He et al., 2020*



## Shared Goal

**Steer MT output towards translations of fuzzy matches**

All three approaches aim to leverage similar existing translations to improve the quality and consistency of machine translation systems.

*Moslem et al., 2023*

S

## Data Augmentation (FR)

Repair: Augment training data by repairing source sentences with fuzzy matches. No architecture modification needed.

```
<sep> fuzzy_target
```

*Rezcan, 2019*

### Step 1: Original Sentence Pair

#### Source

**(S):** We found three studies for inclusion in the review.

#### Target

**(T):** We vonden drie studies voor opname in de review.

### Step 2: Retrieved Fuzzy Match (Similarity: 0.9309)

**FMS:** We identified **nine** eligible studies for inclusion in the review.

**FMT:** We identificeerden **negen geschikte** studies voor opname in de review.

*Differences highlighted in bold*

### Step 3: Creating Augmented Input (S\*)

**Formula:**  $S^* = \text{Source} \langle \text{sep} \rangle \text{FMT}$

#### Actual S\*:

We found three studies for inclusion in the review.

$\langle \text{sep} \rangle$  We identificeerden negen geschikte studies voor opname in de review.

**Same process applied at inference time - source is augmented with FMT before translation**

#### Fuzzy Matches Retrieved From



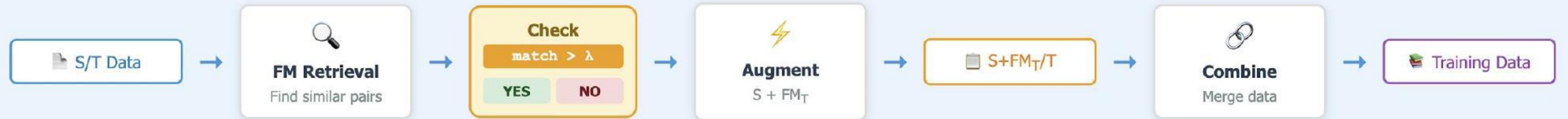
Training Data

OR



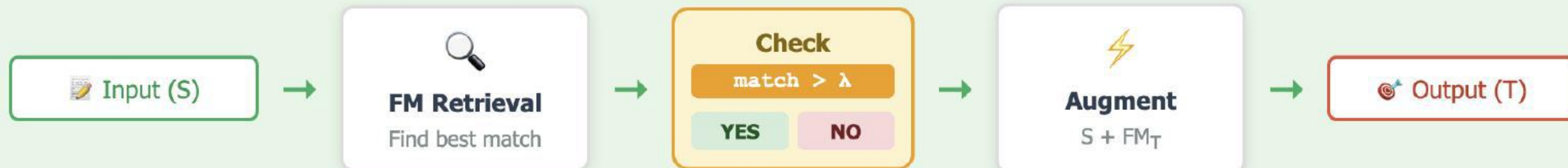
Translation Memory


## 🏆 Training Phase



## 🤖 NMT MODEL

## 🎯 Inference Phase



  
**Limited Parallel Data**  
**300K pairs**  
High-quality human translations  
In-domain (Legal/Technical)  
**HIGH QUALITY**

  
**Abundant Monolingual**  
**1.5M sentences**  
Target language only  
In-domain (Legal/Technical)  
**HIGH QUALITY**

## **Back-Translation Process**



## B BASE

BASELINE

Standard NMT training using only original bilingual data.

### Training Data:

Original Bilingual Data ~300K pairs

## BT BT

BASELINE

Back-translation approach: Original data plus synthetic bilingual data for training.

### Training Data:

Original Bilingual Data ~300K pairs

Back-translated Data ~1.5M pairs

### Total Training:

~1.8M sentence pairs

## BT+NFI BT + NFR

PROPOSED

**Our Approach:** Original + additional synthetic bilingual data used for both training AND fuzzy match augmentation.

### Training Data:

Original Bilingual Data ~300K pairs

Back-translated Data ~1.5M pairs

+ FM Augmentation

### Key Innovation:

Synthetic data for Training + FM Pool

Total: ~1.8M pairs with FM augmentation

## NFR NFR

BASELINE

Neural Fuzzy Repair: Original data with fuzzy match augmentation.

### Training Data:

Original Bilingual Data ~300K pairs

+ FM Augmentation

### Augmentation:

Training + Inference

## NFR NFRmono

BASELINE

NFR with additional monolingual data for FM retrieval during inference only.

### Training Data:

Original Bilingual Data ~300K pairs

### Inference Only:

+ Monolingual Target Data

+ FM Augmentation

### Augmentation:

Inference Only

# 1. Using all available data (bilingual & monolingual)

System	EN→UK			EN→FR			FR→EN		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
<i>BASE</i>	54.90	75.68	90.79	45.76	64.97	79.75	47.76	65.19	80.69
<i>BT (Sennrich et al., 2016)</i>	56.18	76.63	91.61	49.12	67.19	81.42	50.42	67.68	82.32
<i>NFR (Tezcan et al., 2021)</i>	57.71	77.23	91.34	45.61	64.91	79.90	48.05	65.44	80.82
<i>NFR<sub>mono</sub> (Tamura et al., 2023)</i>	60.26	78.58	91.55	46.73	65.48	79.88	49.02	66.05	81.05
<i>Proposed system</i>	<b>66.79</b>	<b>81.99</b>	<b>92.23</b>	<b>57.07</b>	<b>71.83</b>	<b>82.82</b>	<b>60.04</b>	<b>72.96</b>	<b>83.55</b>



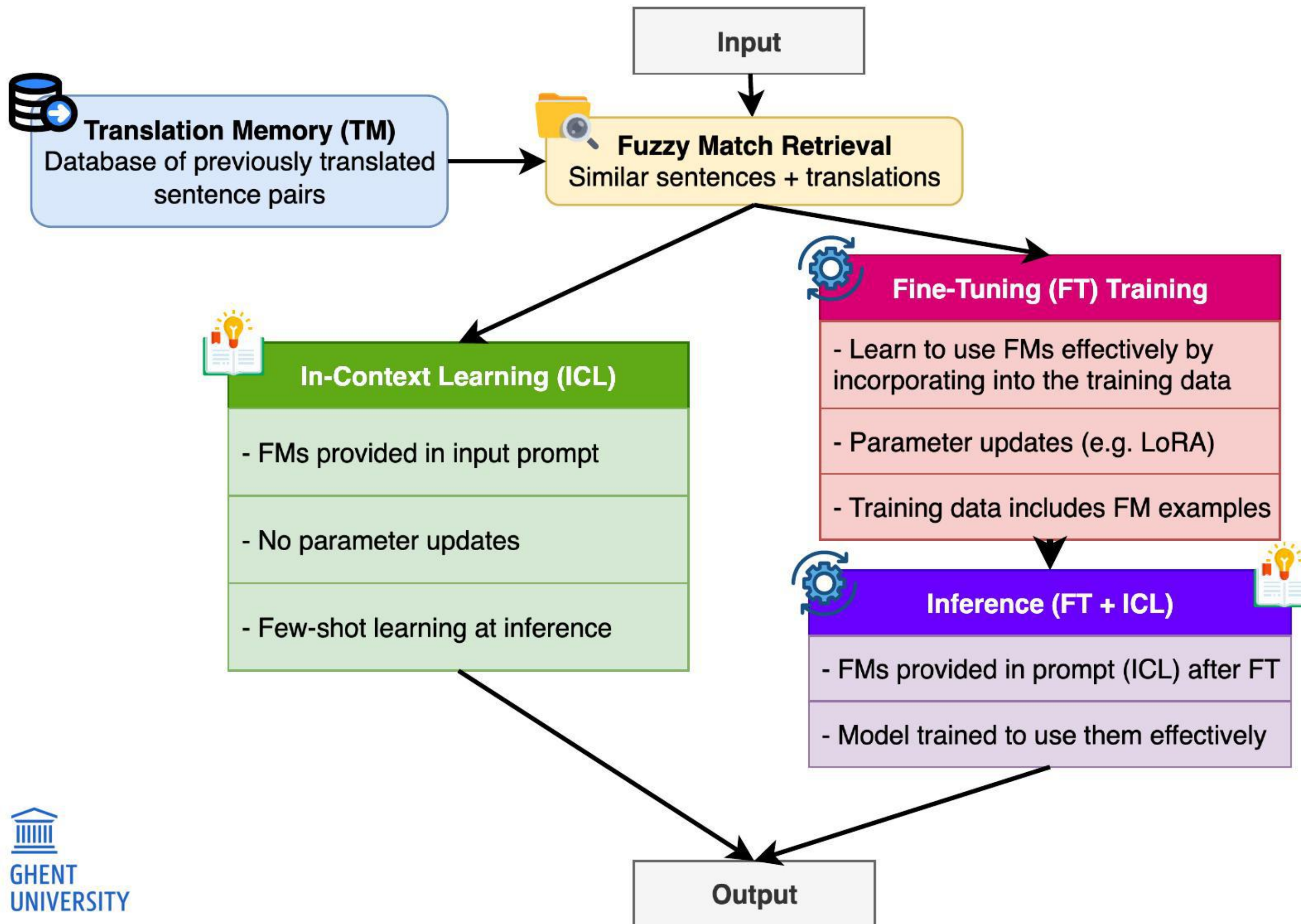
## Baseline Analysis

- ✓ **BT** is the best baseline for EN ↔ FR
- ✓ **NFRmono** is the best baseline for EN → UK
- ✓ **NFR** not useful for EN ↔ FR (confirmed)
- ✓ **NFR** useful for EN → UK



## Our System Performance

- ✓ Outperforms **all baseline systems**
- ✓ Up to **+9.62 BLEU** over best baseline
- ✓ Up to **+12.28 BLEU** over BASE
- ✓ Consistent improvements across all language pairs



## LLM with Fuzzy Matches

Uses FMs as context examples

**0 shot**      No fuzzy match examples

```
Translate the following from SOURCE to TARGET:  
Source: [source sentence]  
Target:
```

**1- $n$  shot**    Up to  $n$  fuzzy match examples

```
Translate the following from SOURCE to TARGET:  
FM_Source1: [FM_SRC1]  
FM_Target1: [FM_TGT1]  
...  
FM_Source $n$ : [FM_SRC $n$ ]  
FM_Target $n$ : [FM_TGT $n$ ]  
Source: [source sentence]  
Target:
```

**Inference:** prompt as above and the model generates the target.

**Training:** prompt as above + expected target translation

# KEY TAKEAWAYS

- LLMs represent a paradigm shift in machine translation
- They will not be the last
- They offer broader language “understanding” but face challenges
- (Human) evaluation and quality assurance remain critical
- The future for translation likely involves human-AI collaboration
- Continuous learning and adaptation are essential for professionals

# Thomas Moerman

PhD Researcher

LT<sup>3</sup> - Language and Translation Technology Team

E Thomas.Moerman@ugent.be

<https://research.flw.ugent.be/nl/thomas.moerman>