The Evolution of NLP: From Specialists to Giants

CS772 · Lecture 2

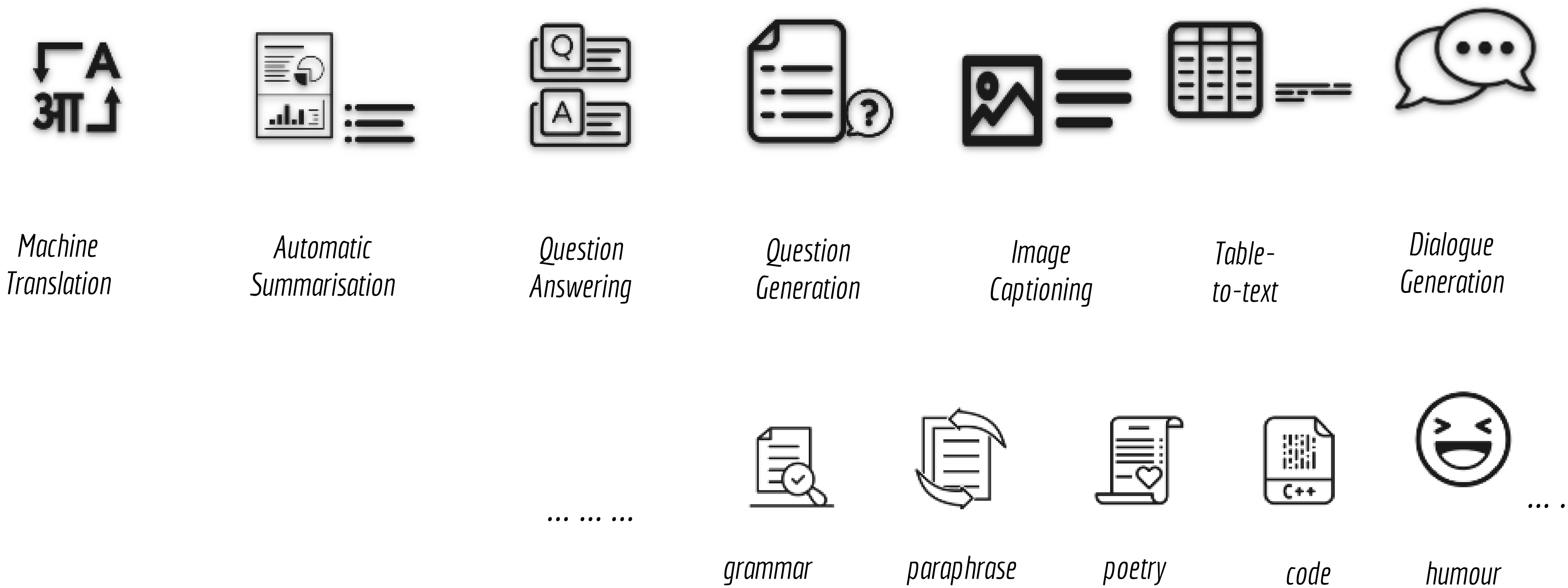
A Journey Through Two Eras of Natural Language Processing

Today's Roadmap

- Era 1 – Task-Based Specialists
- Era 2 – Task-Agnostic Giants
- Frontier directions & current best practices

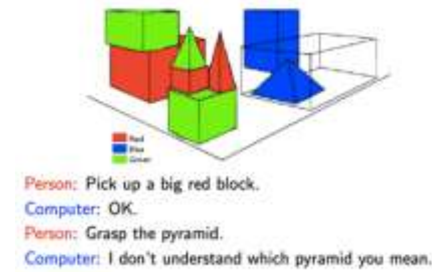
Era 1: The Age of Task-Based Models

Building a different model for every task



1950-1980: Rule-Based Systems

- Hand-crafted linguistic rules
- Precise yet brittle; huge manual effort



6 grammar rules

250 word vocabulary

60 sentences into Russian

IBM 701 mainframe

Georgetown IBM MT Experiment

1990s-2000s: Statistical Revolution

- Naïve Bayes / MaxEnt / CRF for tagging

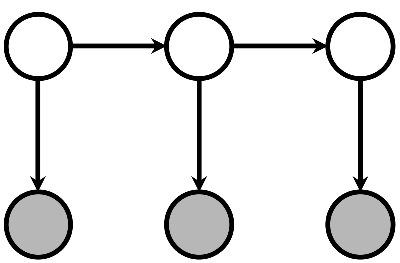
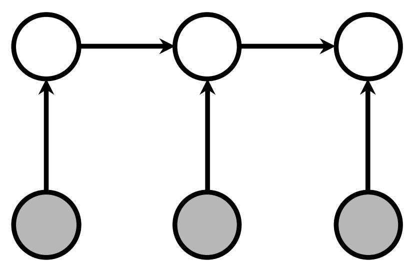
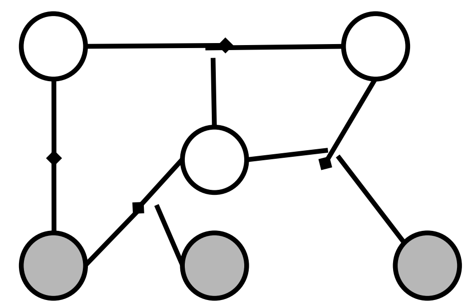
The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

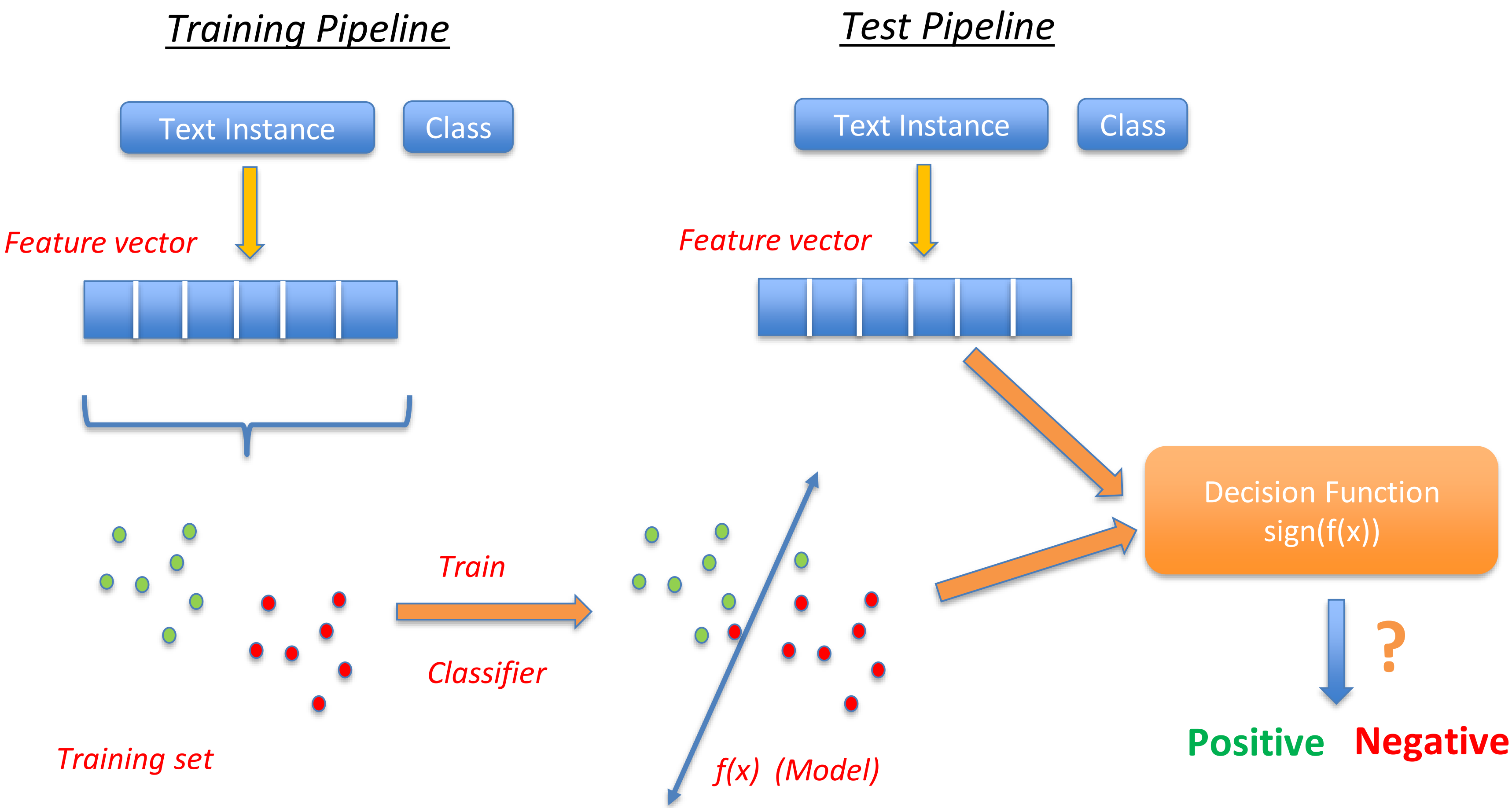
Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

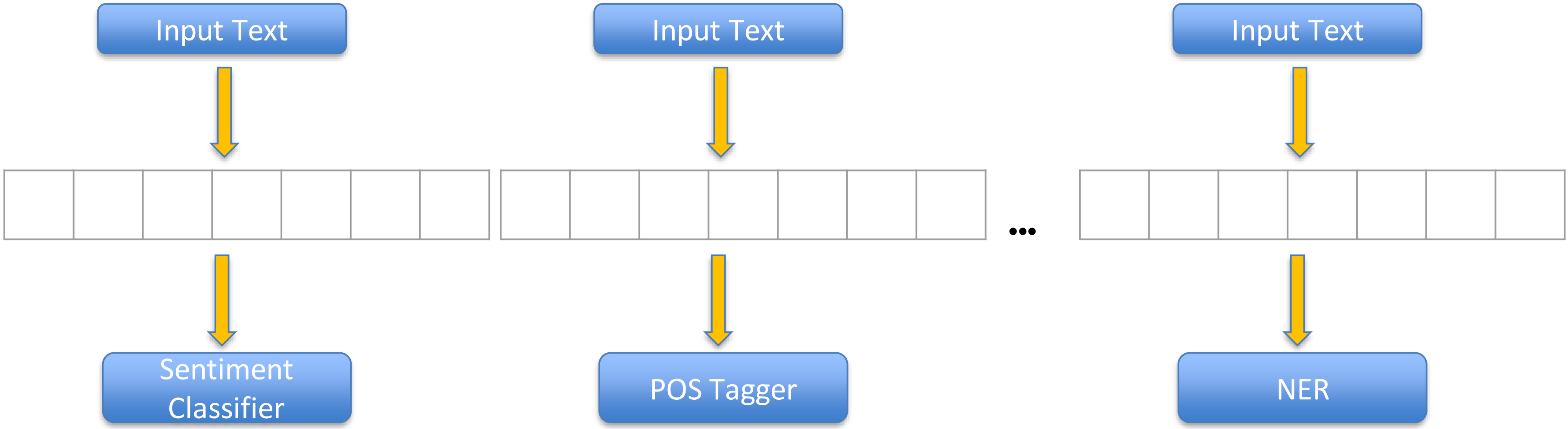
Probabilistic Graphical Models Became Popular

The Data Pipeline in the 90s



Innovate on designing better features and models to capture dependencies between them

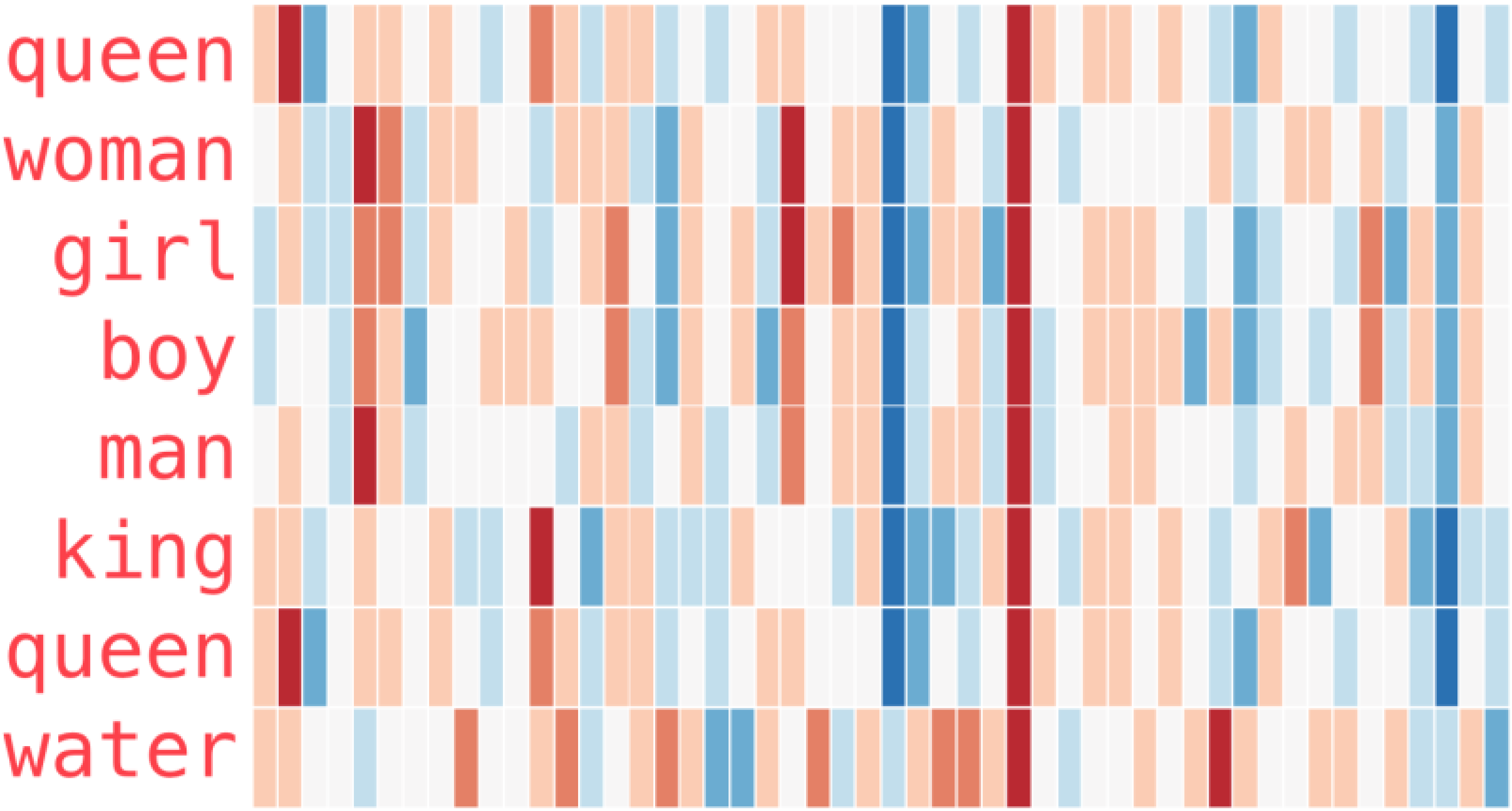
One model per task



<u>Task specific feature extractors</u> (often not reusable across tasks) and <u>task-specific models</u>

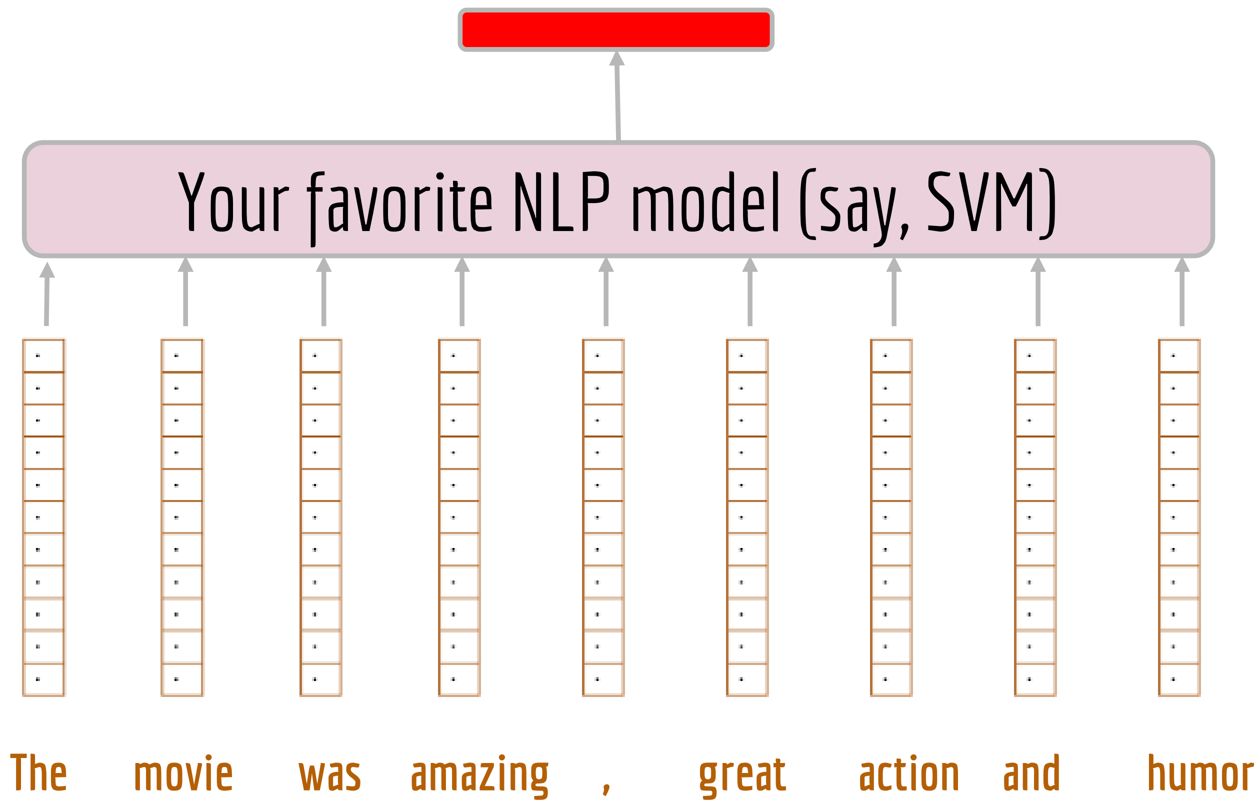
2003-2013: Distributional Semantics

Learn reusable vectorial representations of words and sentences from large scale web corpora



2010-2013: Classical models with distributional features

$$y = f(x; \theta)$$



2014: RNNs & LSTMs

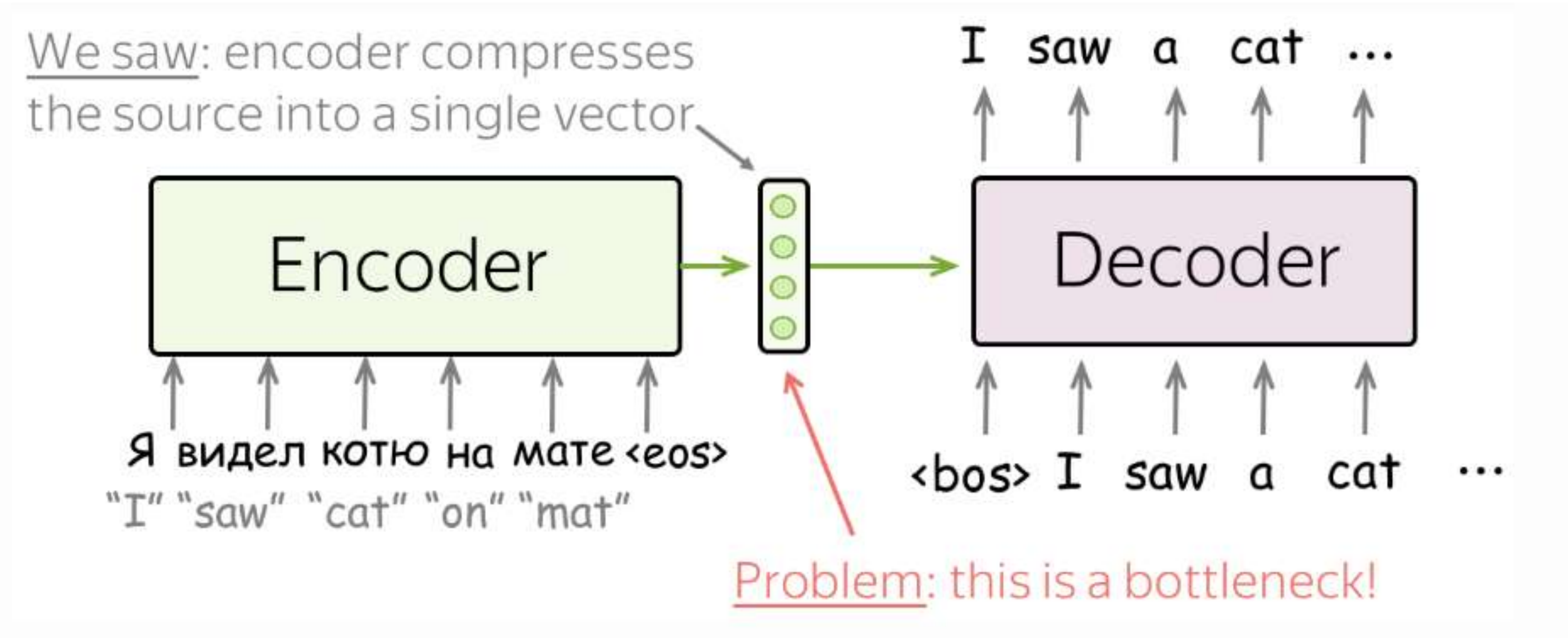


Image source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

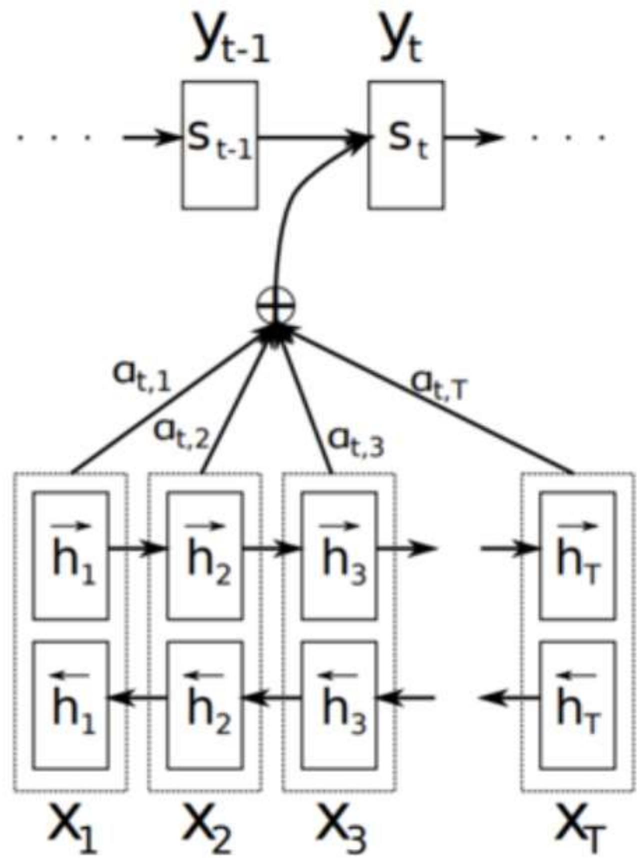
The encoder and decoders are RNNs/LSTMs which are a special type of Deep Neural Networks (bye bye classical models)

2014-2017: Seq2Seq + Attention

Encoder-decoder with additive attention

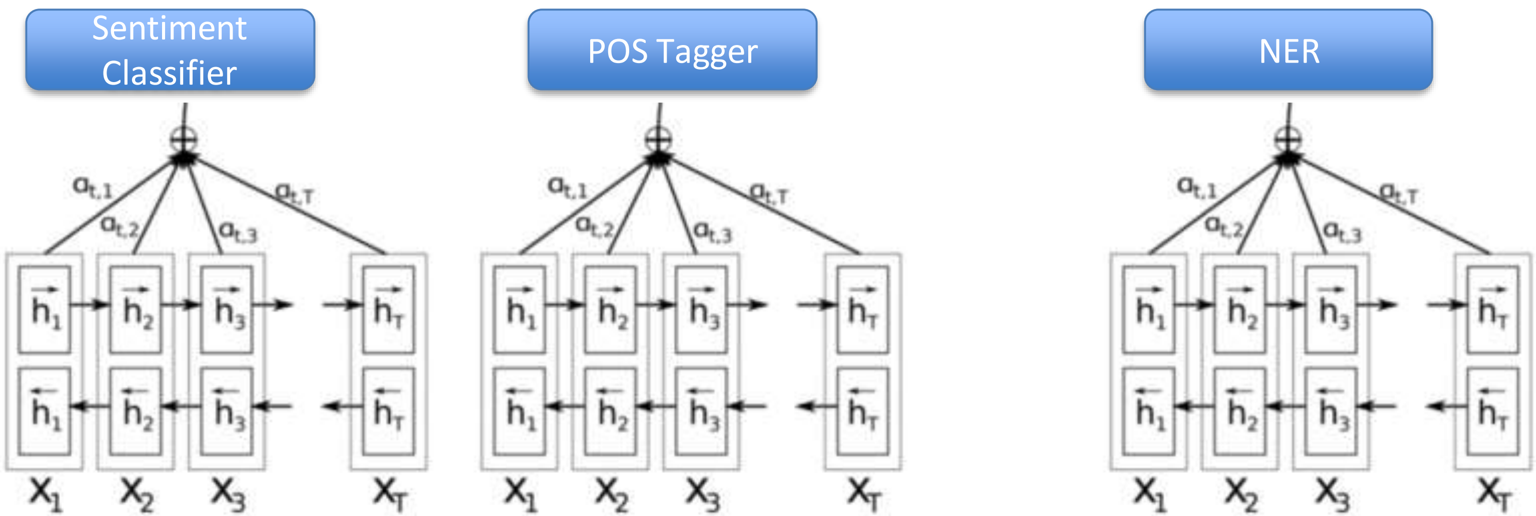
Paved way for Neural MT

Within 2 years toppled Statistical Machine Translation (a technology built over 20+ years)



The idea of attention is perhaps the most important idea of the last decade!

Still one model per task



The idea of attention is perhaps the most important idea of the last decade!

Take-aways from the Specialists Era

Data beats rules

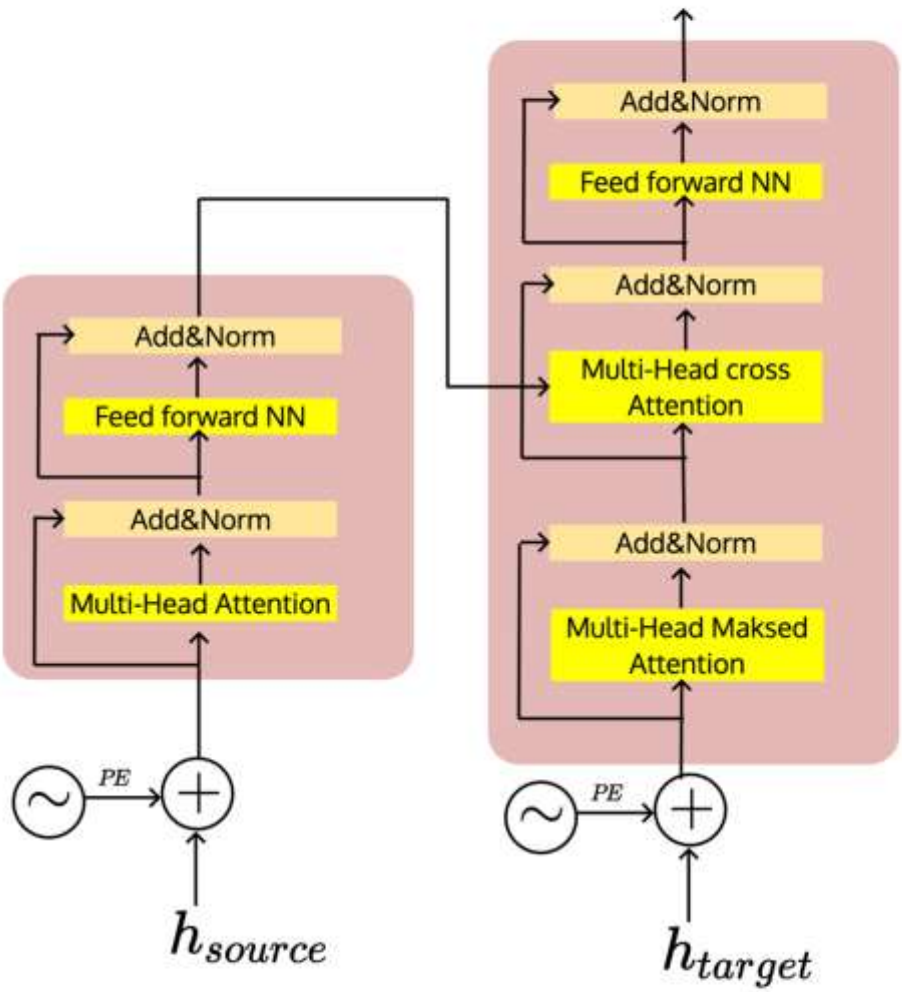
Unified architecture for all NLP tasks (RNN/LSTM with attention)

But scalability & generalisation remain issues (how many models will one train?)

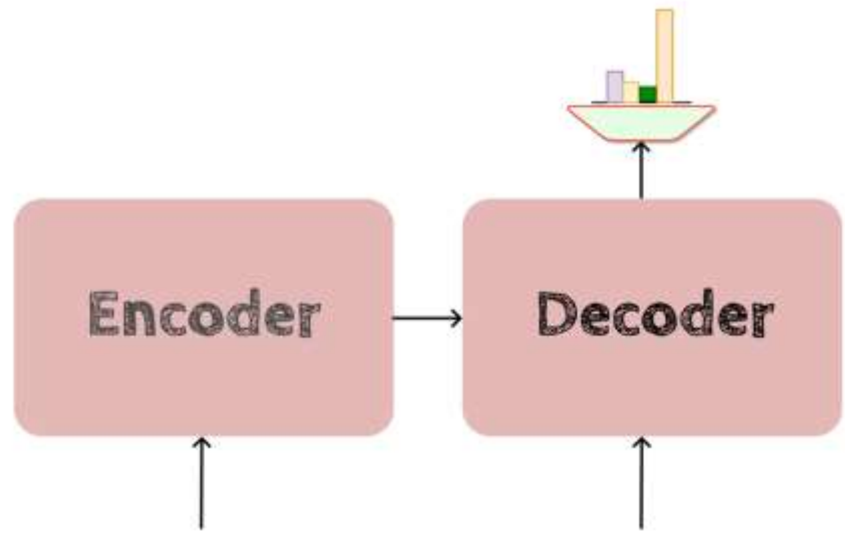
Era 2: The Age of Task-Agnostic Models

One Model to Rule Them All!

2017: The Transformer

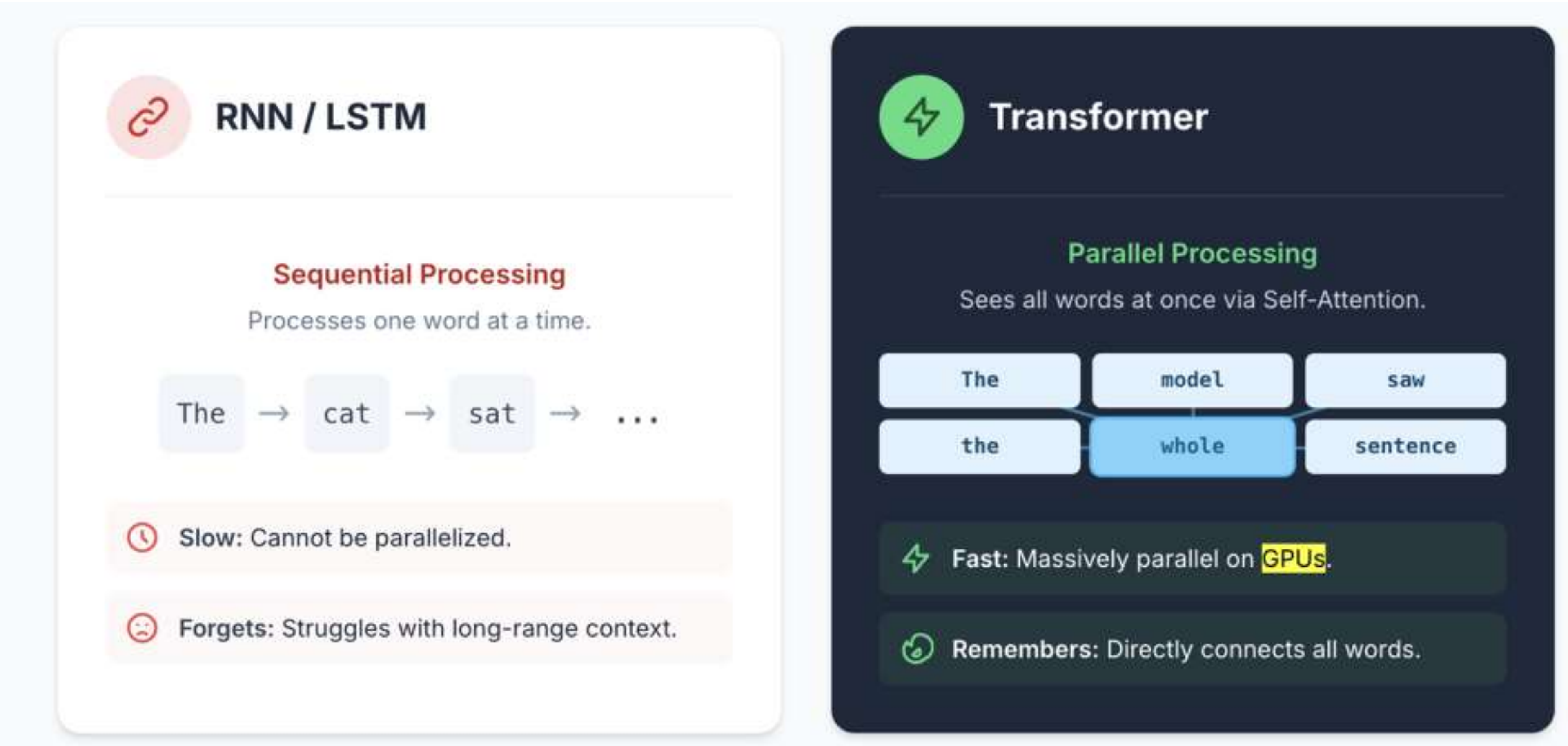


Ignore the complexity for now!

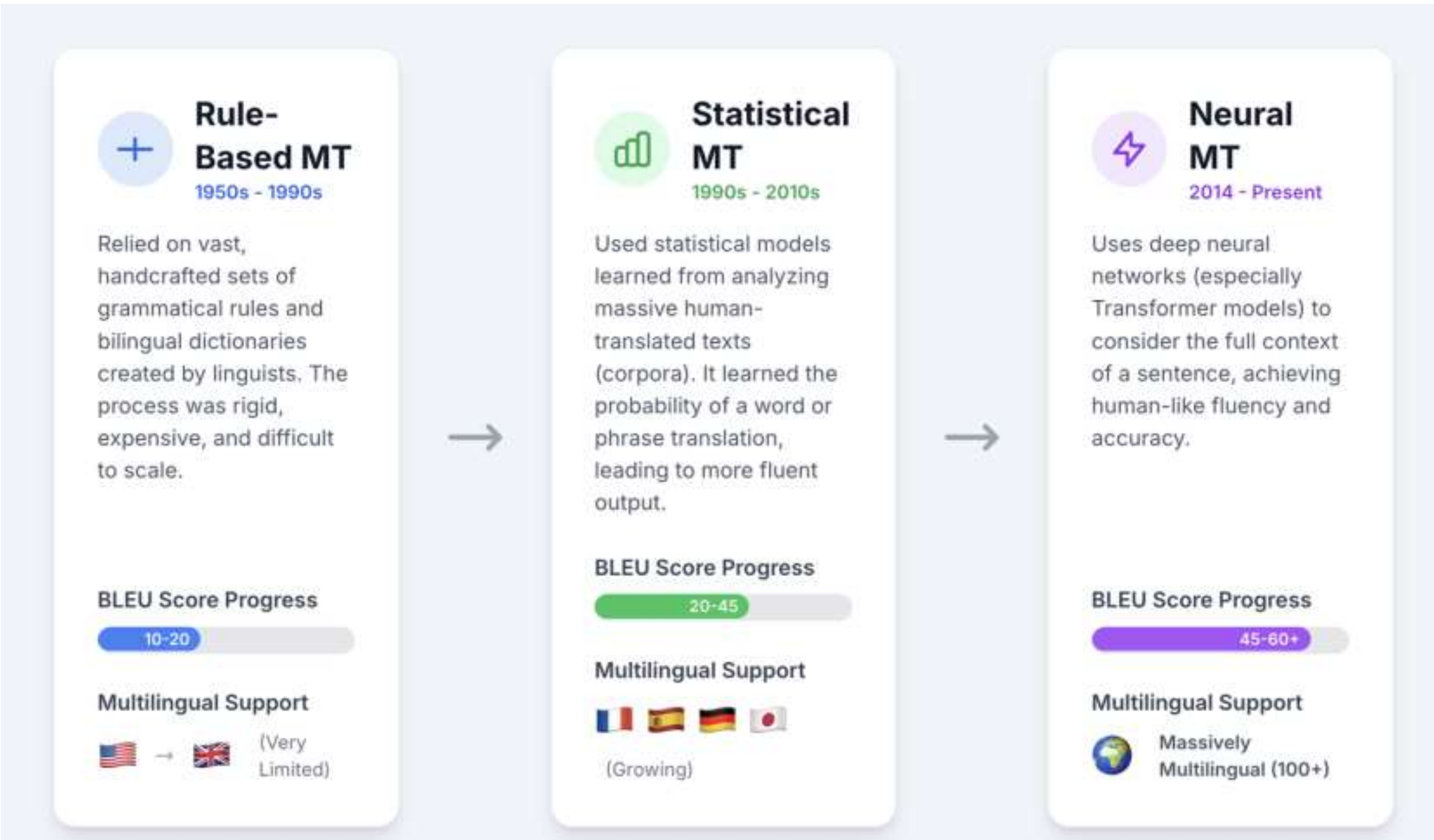


The high level abstraction

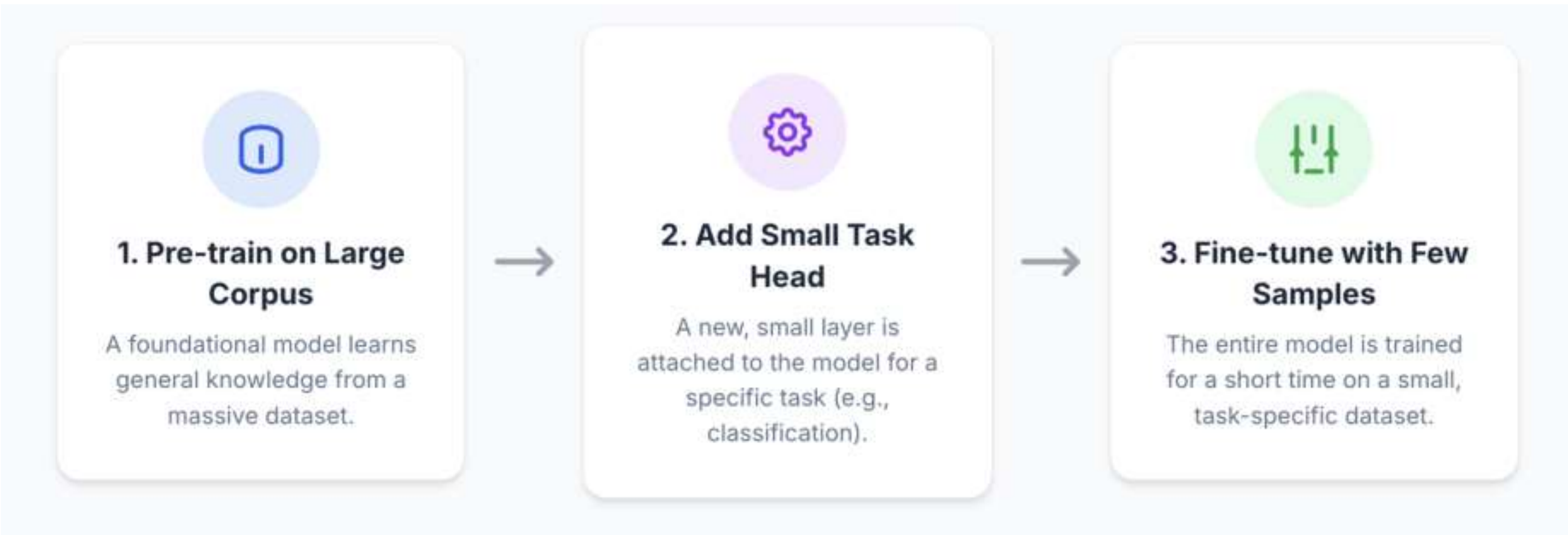
2017-Present: The Transformer Revolution



Case Study: Progress in MT



2018: The Pre-training Revolution



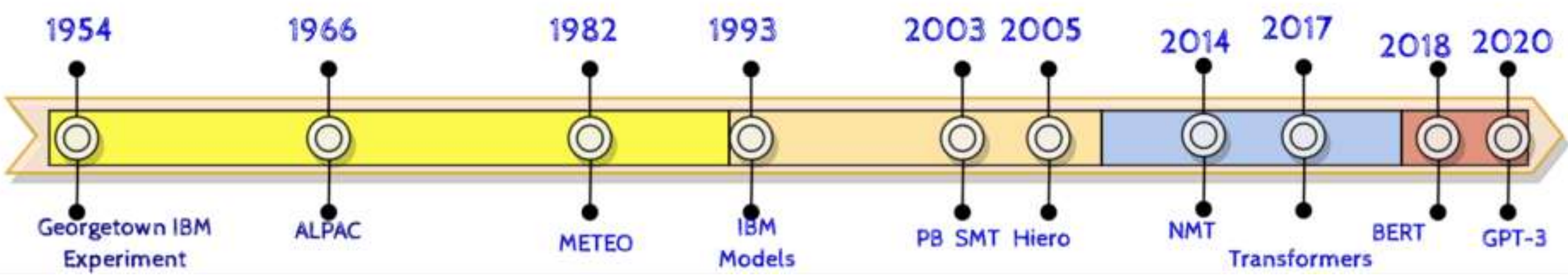
Post 2020: The Era of scale

The Billion Parameter Club

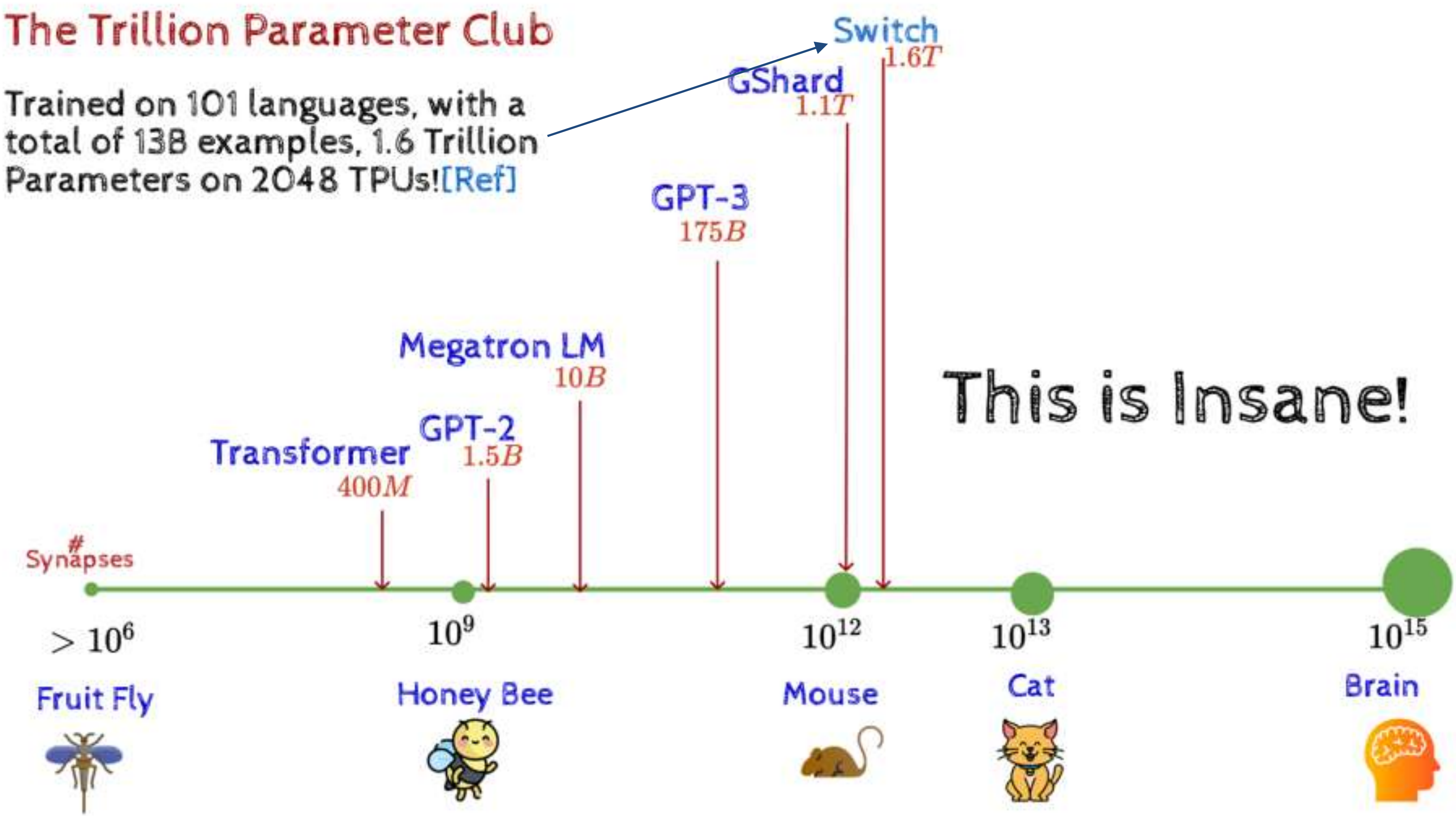
The models are becoming bigger and bigger and bigger!

GPT-3 has 175 billion parameters

Capabilities like in-context learning emerge as size increases.



Post 2020: The Era of Scale



The Data Revolution

STAGE 1: PRE-2018

The Curated Era

Scale: Small (GBs)

Variety: Clean Text Only

Datasets: Wikipedia, BookCorpus

High-quality but limited data, creating models with narrow world knowledge.

STAGE 2: c. 2018-2020

The Web-Scale Era

Scale: Massive (100s of GBs)

Variety: Mostly Web Text

Datasets: Common Crawl, WebText

Sheer volume unlocked general capabilities, but specialized skills were lacking.

STAGE 3: c. 2021-2023

The Diverse Pre-training Era

Scale: Vast (Terabytes)

Variety: Text, Code, Dialogue

Datasets: The Pile, GitHub, ArXiv

Deliberately adding code and scientific text dramatically improved reasoning abilities.

STAGE 4: c. 2023 - PRESENT

The Specialized & Synthetic Era

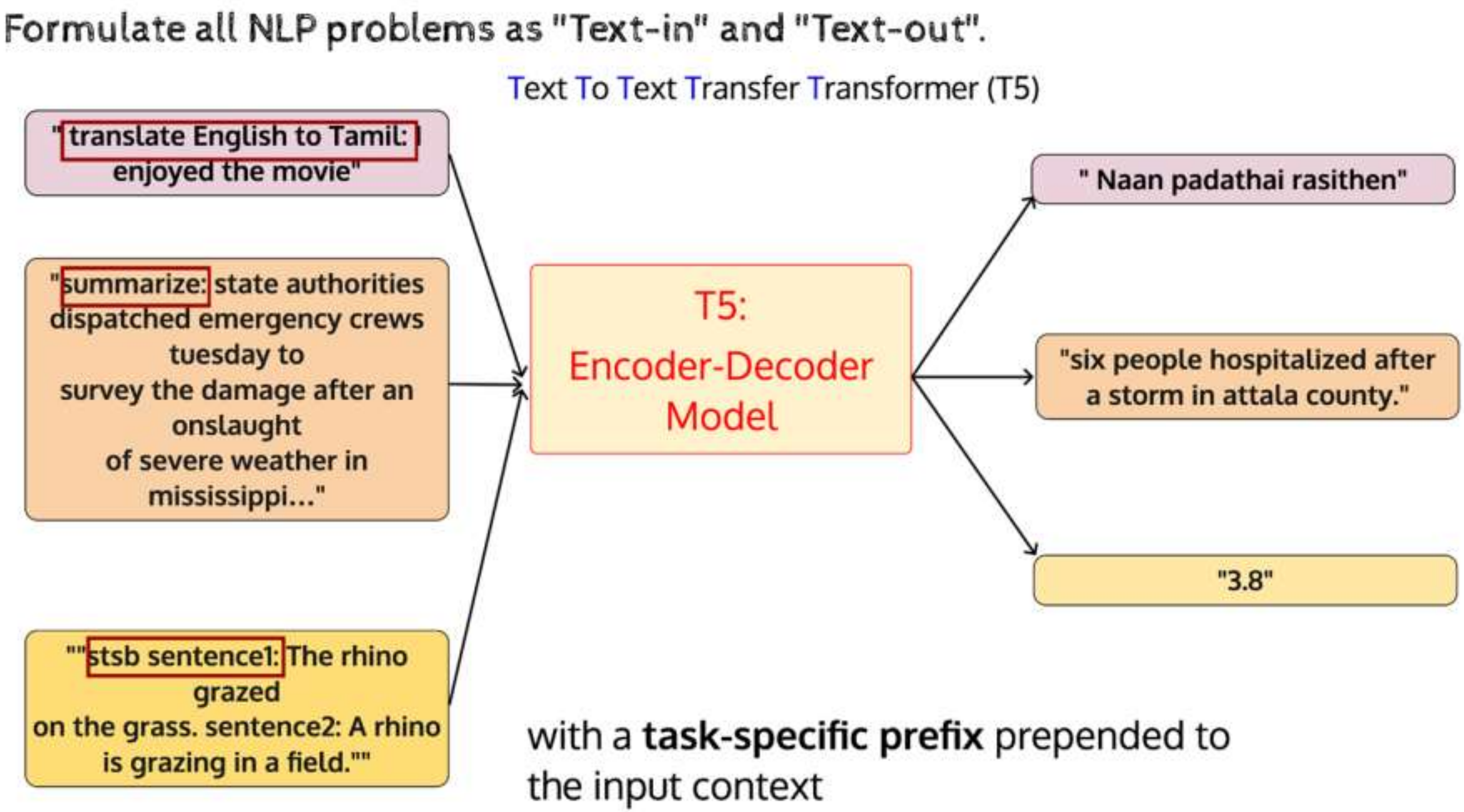
Scale: Vast + Quality Focused

Variety: Reasoning & Synthetic Data

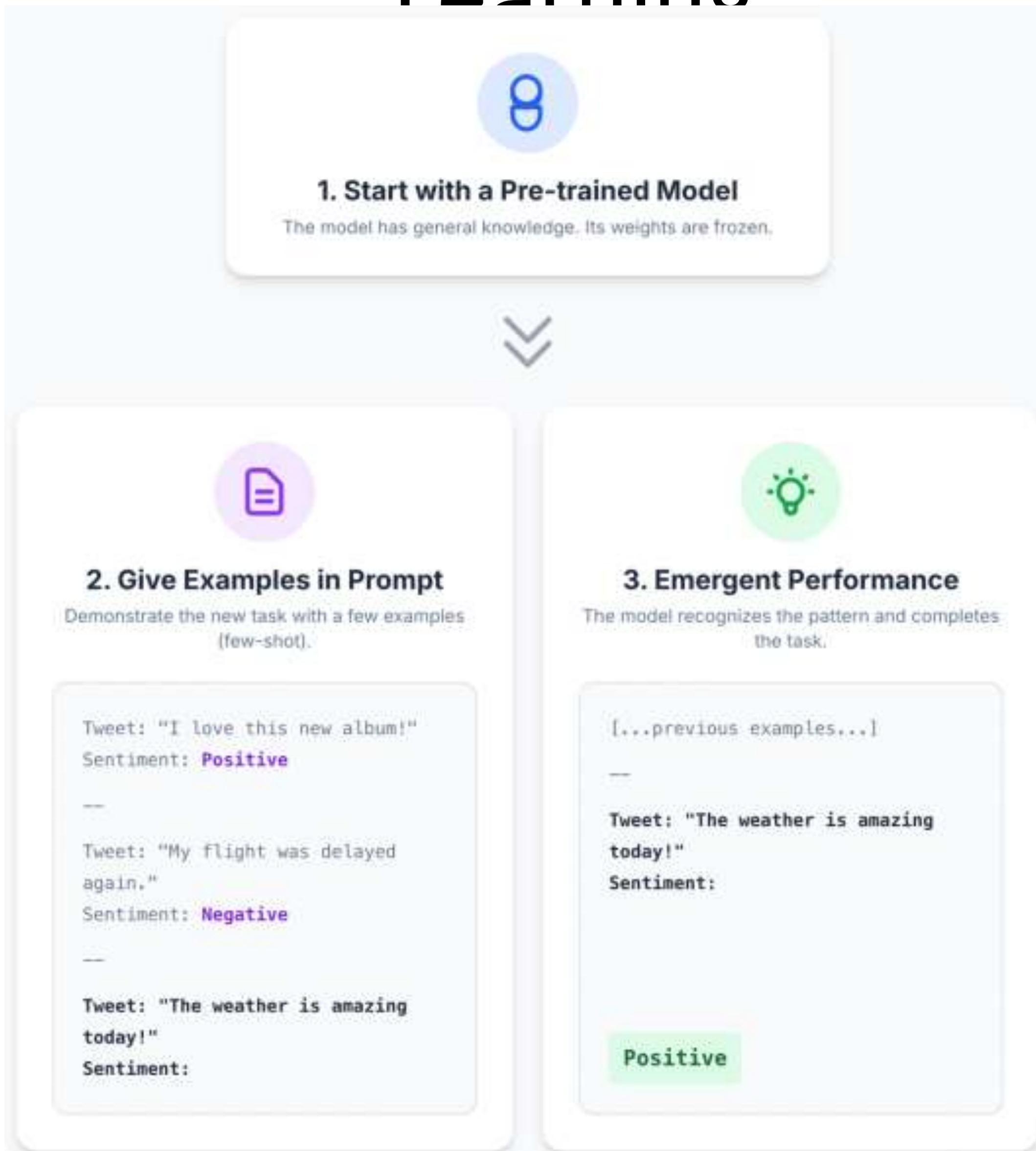
Datasets: Proprietary Mixes, GSM8K

The focus is now on creating better data to teach nuanced skills and complex reasoning.

The idea of a prompt



Emergent Behavior In-Context Learning



2022: Aligning to human needs

To make AI models…

Helpful

Follow instructions accurately and assist users in achieving their objectives.

Honest

Provide truthful information and avoid making things up (hallucinations).

Harmless

Refuse to generate unsafe, unethical, or malicious content.

Without Alignment ChatGPT would not have become popular when it was released

Alignment in action!

User Asks:

"How do I make a strong cleaning solution at home?"

Helpful

"A simple and effective all-purpose cleaner can be made by mixing equal parts white vinegar and water in a spray bottle. It's great for countertops and windows."

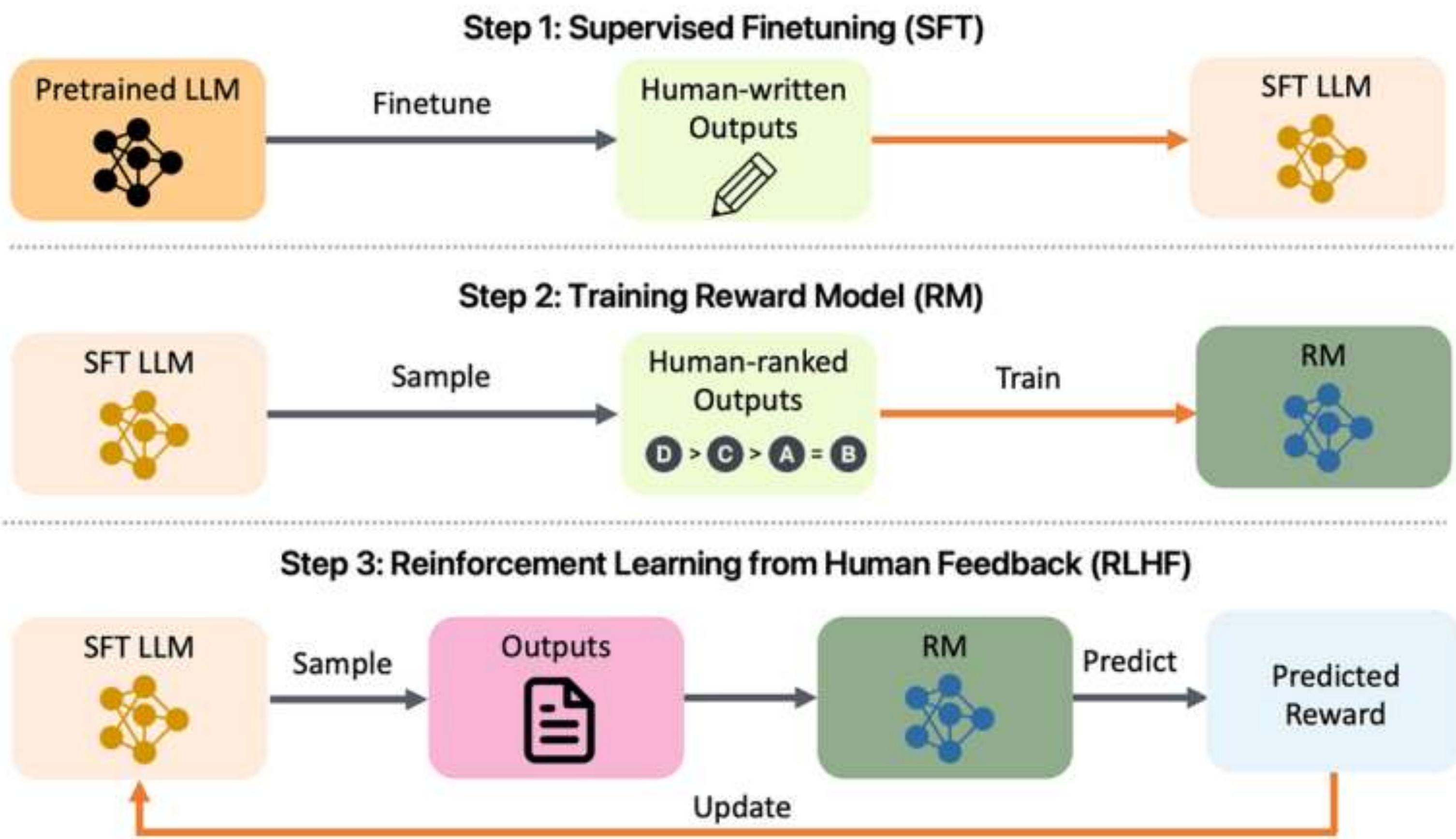
Honest

"While homemade solutions are useful, they may not disinfect as effectively as commercial products registered with the EPA. Always check if a surface is safe for acidic cleaners like vinegar."

Harmless

"I cannot provide instructions on how to mix chemicals like bleach and ammonia. Combining them creates toxic chloramine gas, which is extremely dangerous and can cause serious respiratory damage."

Alignment: A complex DL-RL pipeline



Source: https://cobusgreyling.medium.com/llm-alignment-hallucination-misinformation-a1673d96629f

The Power

(What LLMs Can Do)

- Generate Code
- Translate Languages
- Summarize Text
- Power Chatbots

...Comes Great Responsibility

(The Challenges We Face)

- Hallucinations
- Safety & Misuse
- Bias & Fairness
- Cost & Impact

The new frontier of NLP Challenges

Bias, Fairness & Safety
c. 2016 - Present

Ensuring models don't amplify stereotypes or generate harmful, toxic, or unsafe content.

Explainable AI (XAI)
c. 2017 - Present

Understanding and interpreting the "why" behind a model's decisions, moving beyond black boxes.

Green AI & Efficiency
c. 2019 - Present

Reducing the immense computational and environmental cost of training and running large models.

Data Cleaning at Scale
c. 2020 - Present

Developing methods to automatically curate and filter petabytes of web data for high-quality training.

Better Alignment
c. 2020 - Present

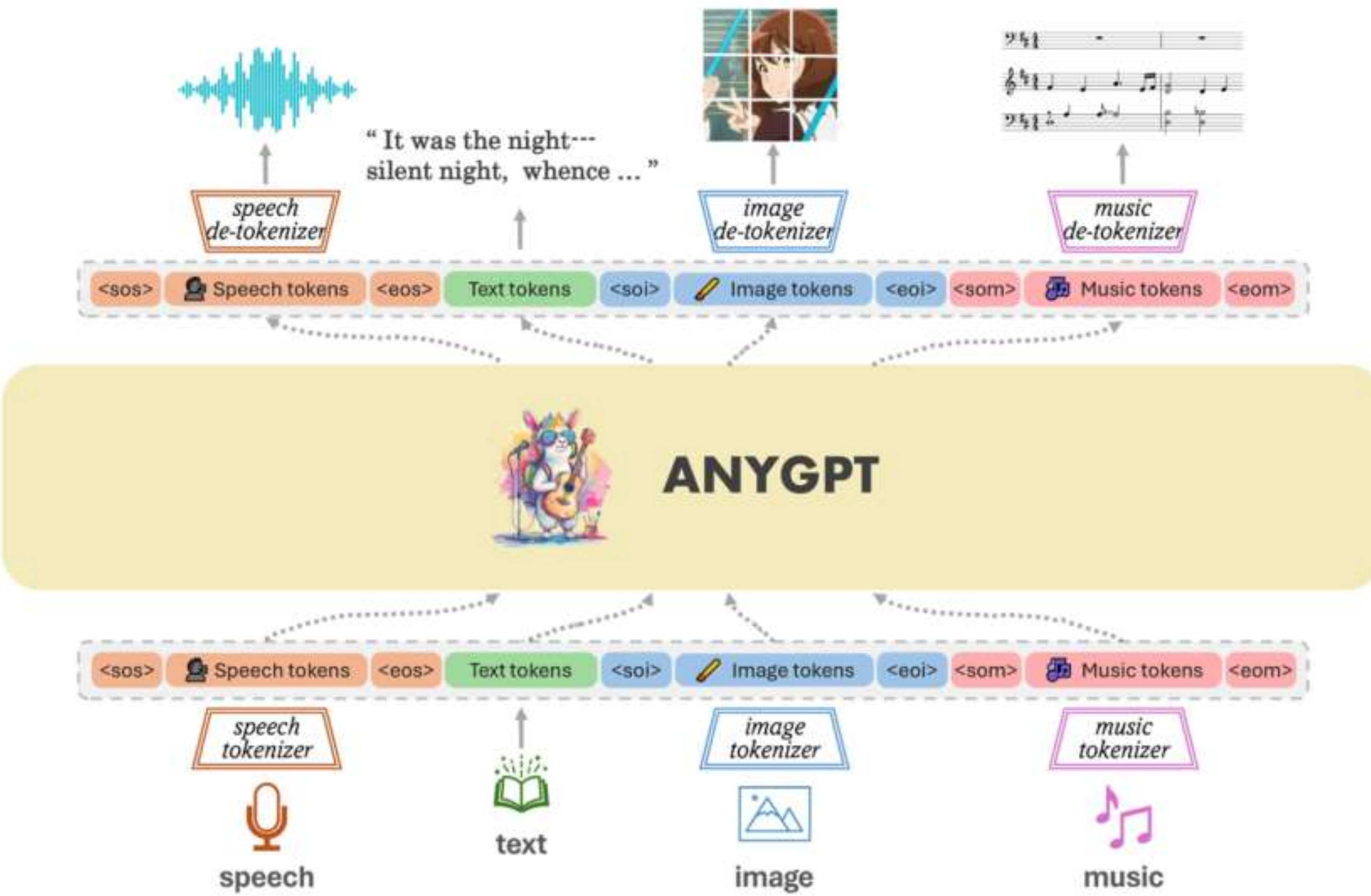
Ensuring models follow complex instructions and adhere to human values and preferences.

Deployment & Scalability
c. 2022 - Present

Making it practical to serve massive models to millions of users efficiently and affordably.

From language to vision

All-in-one models



Deep Learning: the great unifier

1

PRE-2000s

Rule-Based Systems

Unified: Nothing. Rules were bespoke per problem.

Systems: ELIZA (1966)

Isolated, brittle systems that could not scale or generalize.

2

2000s - EARLY 2010s

Task-Specific Models

Unified: Nothing. Each task had unique features & models.

Systems: SVMs, CRFs, SMT

A fragmented ecosystem of specialized statistical solutions.

3

c. 2013

Unified Features

Unified: The feature space, via universal word representations.

Innovations: word2vec (2013)

The first major step toward generalization, creating a common language.

4

c. 2014-2017

Unified Architecture

Unified: The core model architecture for sequence tasks.

Innovations: Seq2Seq (2014)

RNNs/LSTMs became the standard blueprint, but each task still needed a separate model.

5

c. 2018

Unified Model & Language

Unified: The model itself; one base model for many tasks & languages.

Innovations: BERT (2018), GPT

The Transformer ushered in the era of pre-training and fine-tuning.

6

c. 2021 - PRESENT

Unified Modalities

Unified: The data itself; one model for text, images, audio & video.

Innovations: CLIP (2021), Gemini

The final stage: a single AI reasoning across different types of information.