

# Understanding LLMs and generative AI

and their effect now and in the future on translation and localization  
and on nearly everything else, too

---

Jay Marciano

President, Association for Machine Translation in the Americas

Budapest, 14 June 2024

# SAMSUNG Galaxy S24 Ultra



We speak  
16 languages  
with Live Translate  
and Galaxy AI 🌟



Die Live Übersetzung ist über die Kamera der Samsung Galaxy S24 Ultra verfügbar. Live Übersetzung ist nur in bestimmten Ländern verfügbar. Die Live Übersetzung ist nicht in allen Sprachen verfügbar. Die Live Übersetzung ist nicht in allen Sprachen verfügbar. Die Live Übersetzung ist nicht in allen Sprachen verfügbar. Die Live Übersetzung ist nicht in allen Sprachen verfügbar.



# Introductions

## Jay Marciano

- President of AMTA (Association of Machine Translation in the Americas)
- 26 years' experience in the development & application of MT
- And another decade of experience in publishing and teaching
- Avowed language geek on a mission to increase understanding, cooperation, and collaboration among translators and interpreters, MT researchers and developers, language service companies, and customers





We shape our buildings, and  
afterwards our buildings shape us.  
Winston Churchill

# What is *intelligence*?

The ability to acquire, understand, and use knowledge.

The American Heritage Dictionary

# What does *artificial* mean?

Made by humans, especially in imitation of something natural.

[Middle English, from Old French, from Latin *artificiālis*, belonging to art, from *artificium*, craftsmanship; from *artifex*, *artific-*, craftsman : *ars*, art, + *-fex*, maker.]

The American Heritage Dictionary

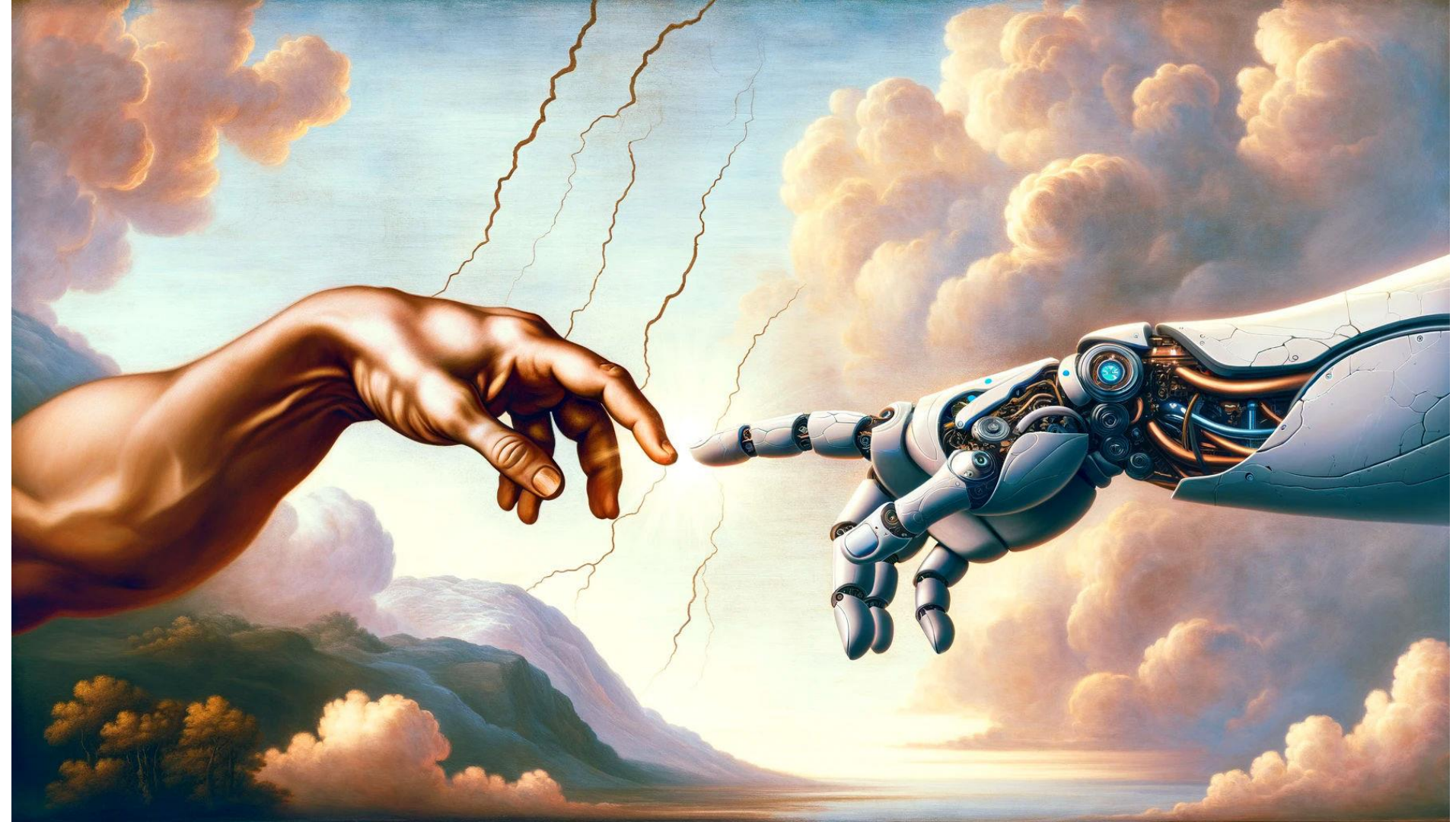
# What is *Artificial Intelligence*?

The ability of a computer or other machine to perform activities that are normally thought to require the capacity to acquire and apply knowledge.

Adapted from The American Heritage Dictionary

**AI may turn out to be humankind's  
greatest achievement.**













# It's really hard to keep up!

arXiv:2405.21060v1 [cs.LG] 31 May 2024

1 Introduction  
Transformers process our app scaling autoregressive language models. Our goal is to make the length of these forms the abstract paper strength

arXiv:2406.02500v1 [cs.LG] 4 Jun 2024

1 Introduction  
While [1, 2, 3] to excise selective these of [7, 8] h

arXiv:2406.00975v2 [cs.CL] 5 Jun 2024

1 Introduction  
Large Language in industry dialogue pressive ability succeed on a variety of 2023). A key challenge is where the model incorrect information (Roller et al., 2023) augmented generalizing know in the LLM prompt conditions in prod

Retrieval-Augmented Models (RAG) are designed to enhance the capabilities of large language models (LLMs) by integrating external information. However, existing RAG systems often suffer from issues such as hallucinations, low latency, and high inference costs. In this paper, we introduce Luna, a novel RAG framework that addresses these challenges through a combination of fine-tuned LLMs, a custom retriever, and a post-processing module. Our method achieves state-of-the-art performance on various benchmarks, demonstrating superior accuracy and reduced hallucinations compared to baseline RAG systems.

arXiv:2406.04303v1 [cs.CV] 6 Jun 2024

1 Introduction  
Language Models (LLMs) to make input sequences to group a sequence can then  
The existing performance goals to Mamba) we introduce as its core non-auto process left. This images. Similar to complex resolutions self-attention

arXiv:2403.06840v2 [cs.CL] 6 Jun 2024

1 Introduction  
Large language models (LLMs) have demonstrated their exceptional knowledge reasoning and output quality across various task domains (Ouyang et al., 2022). However, knowledge stored in LLMs is static and hard to incorporate (Dhingra et al., 2022). In this paper, we address this issue by introducing Retrieval-Augmented Generation (RAG) approaches that leverage external documents, extract relevant knowledge, and incorporate it into the model's context through prompts, thereby embedding the knowledge into the model's reasoning process (Ram et al., 2023).

**Abstract**  
Large language models (LLMs) have demonstrated exceptional performance in a wide range of tasks, but they still heavily rely on knowledge stored in their parameters. Moreover, updating these models with new information is a challenging task. In this paper, we propose Retrieval-Augmented Generation (RAG) as a solution to this problem. RAG allows LLMs to access external knowledge sources during inference, enabling them to provide more accurate and up-to-date answers. We explore different RAG architectures and evaluate their performance on various benchmarks. Our results show that RAG significantly improves LLM performance on tasks requiring up-to-date information and domain-specific knowledge.

arXiv:2406.03701v1 [cs.MM] 6 Jun 2024

1 Introduction  
In the field of machine learning, the performance of models is often evaluated based on their ability to generalize to new, unseen data. This is a challenging task, especially for deep learning models, which are prone to overfitting. In this paper, we propose a novel method for improving the generalization performance of deep learning models. Our method involves training the model on a diverse set of data, including data from different domains and sources. We also use a variety of regularization techniques to prevent the model from overfitting to the training data. Our results show that this method significantly improves the model's performance on new, unseen data.

arXiv:2406.02543v1 [cs.LG] 4 Jun 2024

res  
ed  
dis  
tin  
so  
2  
g  
n  
a

Artificial Intelligence >

Apple Enters A.I. Fray

The New York Times

Meta's A.I. Scraping

Humane's A.I. Device Flop

OpenAI's 'Reckless' Culture

## Apple Jumps Into A.I. Fray With Apple Intelligence

The iPhone maker, which has been slow to embrace artificial intelligence, will weave it into the technology that runs on billions of devices.



Tim Cook, Apple's chief executive, at the company's developer conference at its headquarters in Cupertino, Calif. Carlos Barria/Reuters



By Tripp Mickle

Tripp Mickle has written about Apple since 2016.

June 10, 2024

Nearly two years after OpenAI ignited a race to add artificial intelligence into

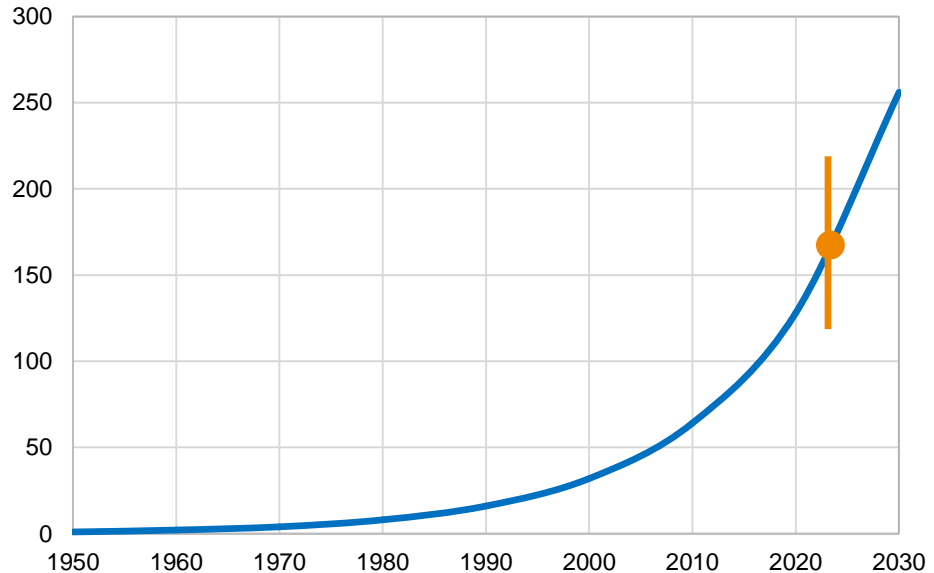


# Technology is advancing faster and faster

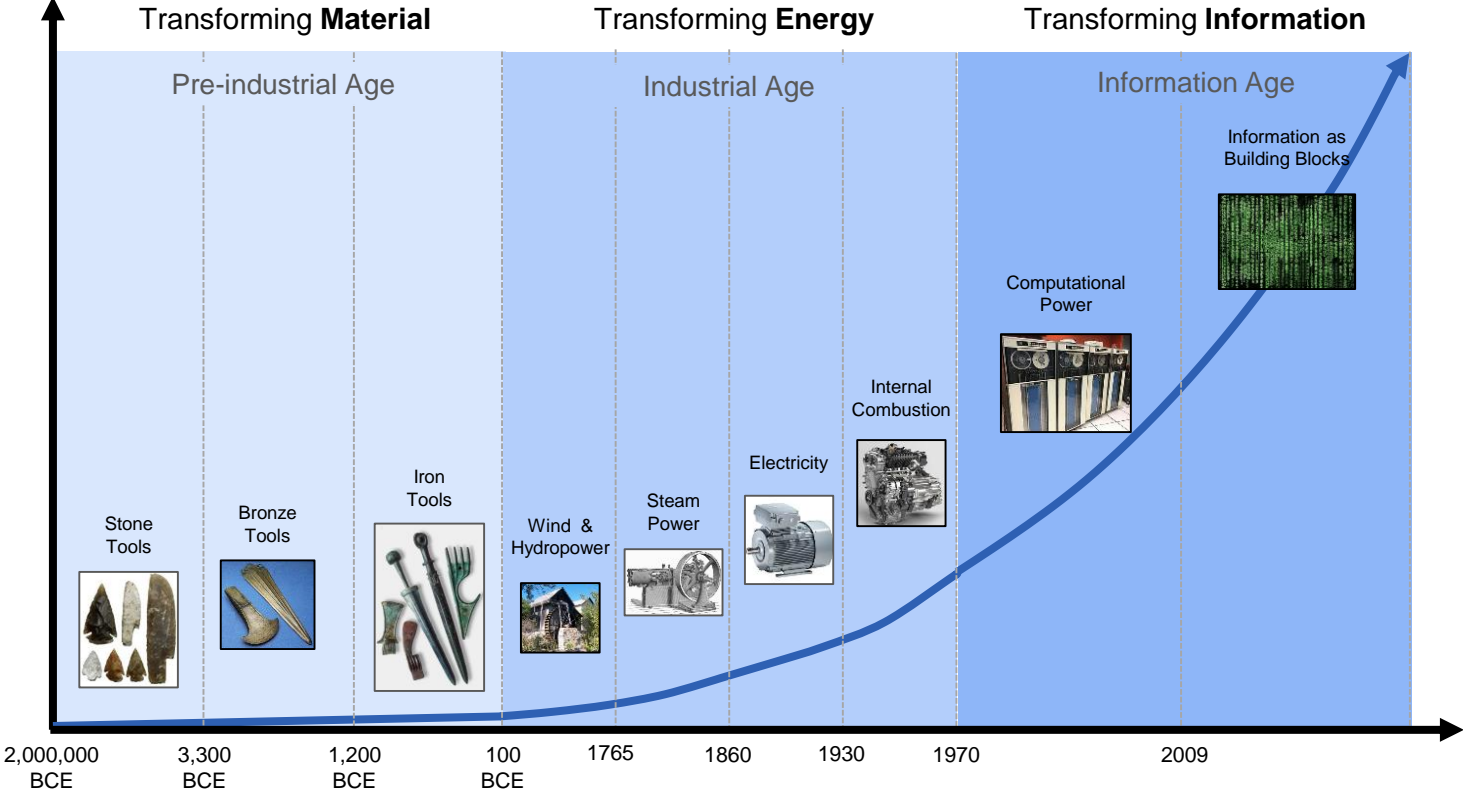
The rate of technological development doubles every ten years

- In 2003, Ray Kurzweil estimated that the rate of technological advance doubles every decade.
- That's exponential growth, with the advances of the past decade helping to increase the rate of development in the next decade.
- In 2024, we're developing technology twice as fast as we were in 2014, and 156 times faster than in 1954, at the dawn of Machine Translation
- And from now until 2034, the rate of technological advance will double again.

Rate of Tech Development (1950-2030)



# Drivers of Progress



**In the Information Age,**

**data**

**is the primary driver of  
technological progress.**



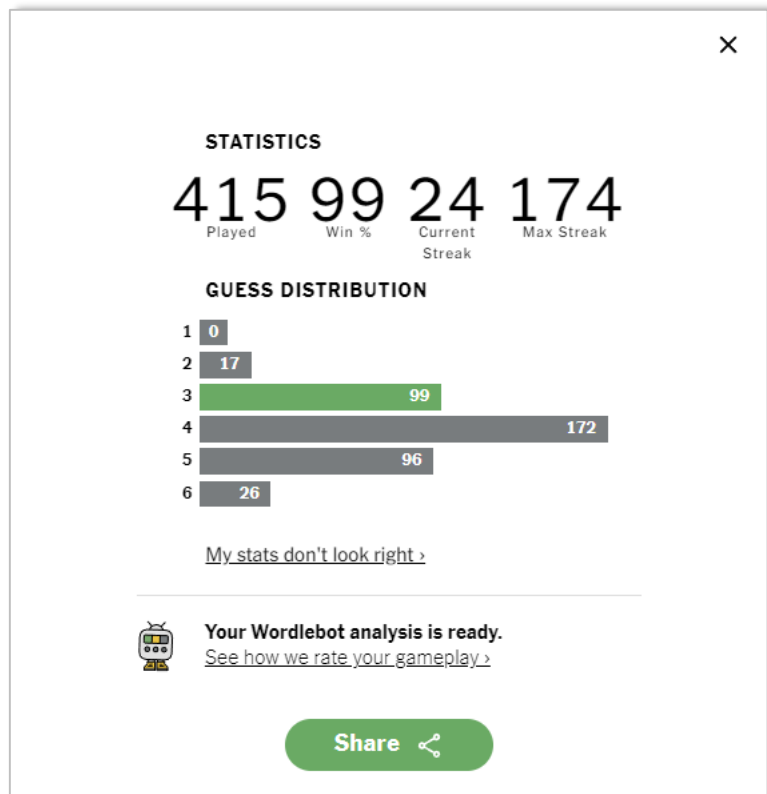
# Truly grasp these novel maybe scary ideas

Something has become very clear to us in the last few years ...

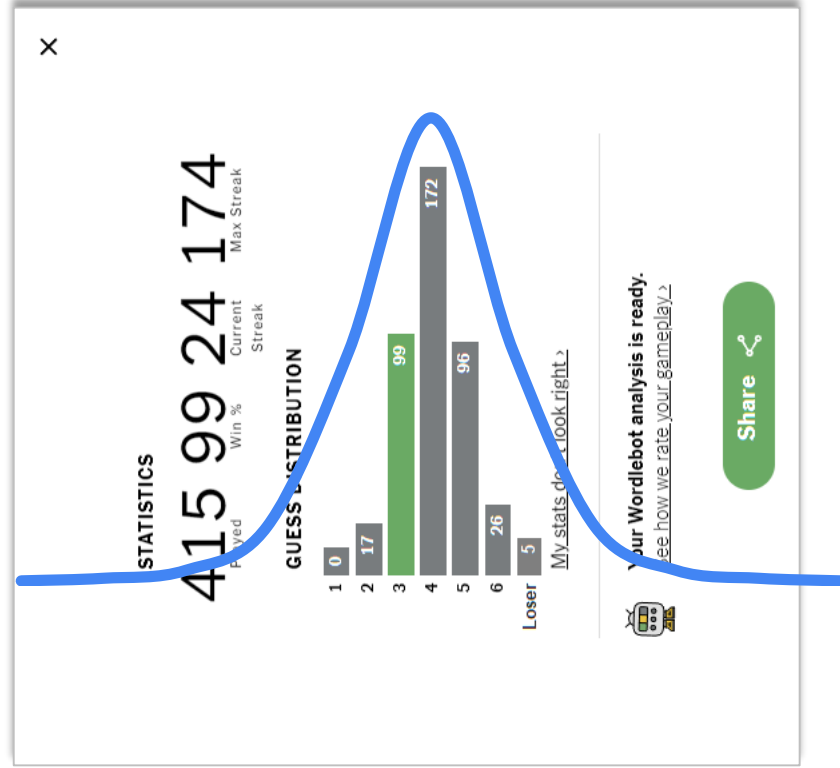
There are a LOT of  
five-letter words  
in English



# Wordle is all about ...



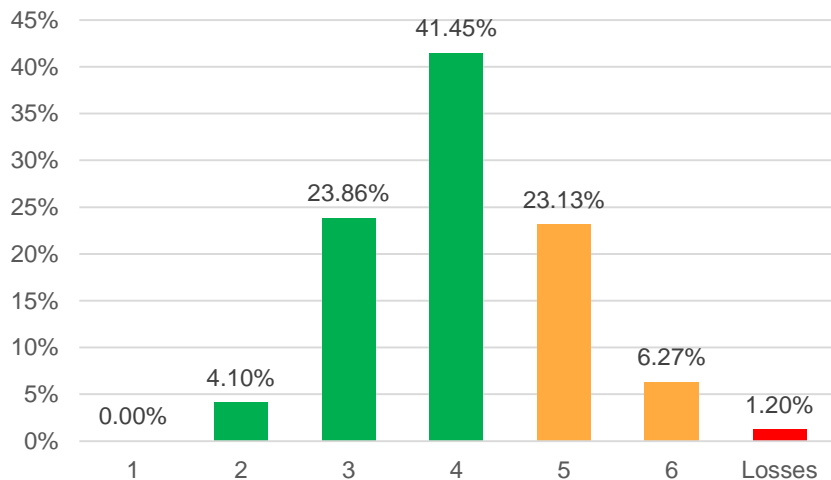
# Wordle is all about ... math?





# Playing Wordle slightly differently

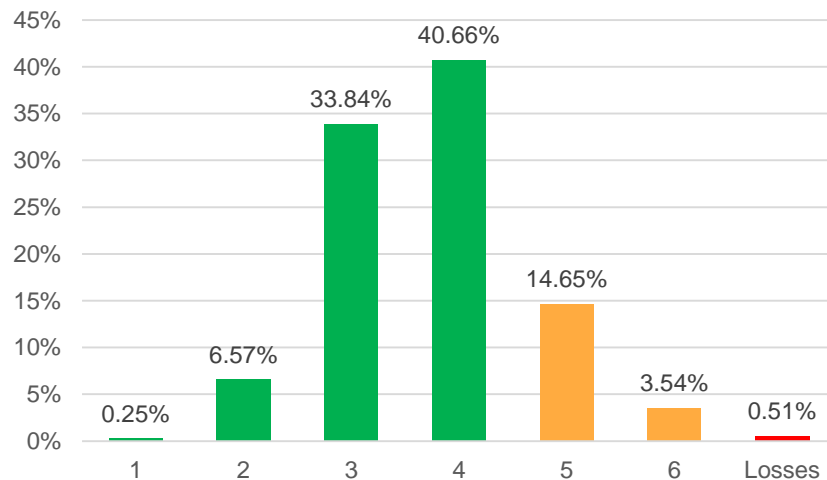
Wordle Results (through May 2023)



Winning percentage: 98.8

Avg guesses needed to win: 4.03

Wordle Results (since June 2023)



Winning percentage: 99.5

Avg guesses needed to win: 3.73

The difference: One additional piece of information about each guess.

# Data use in an AI system

What kind of data does an AI-controlled automobile have about its surroundings?



# Data use in an AI system

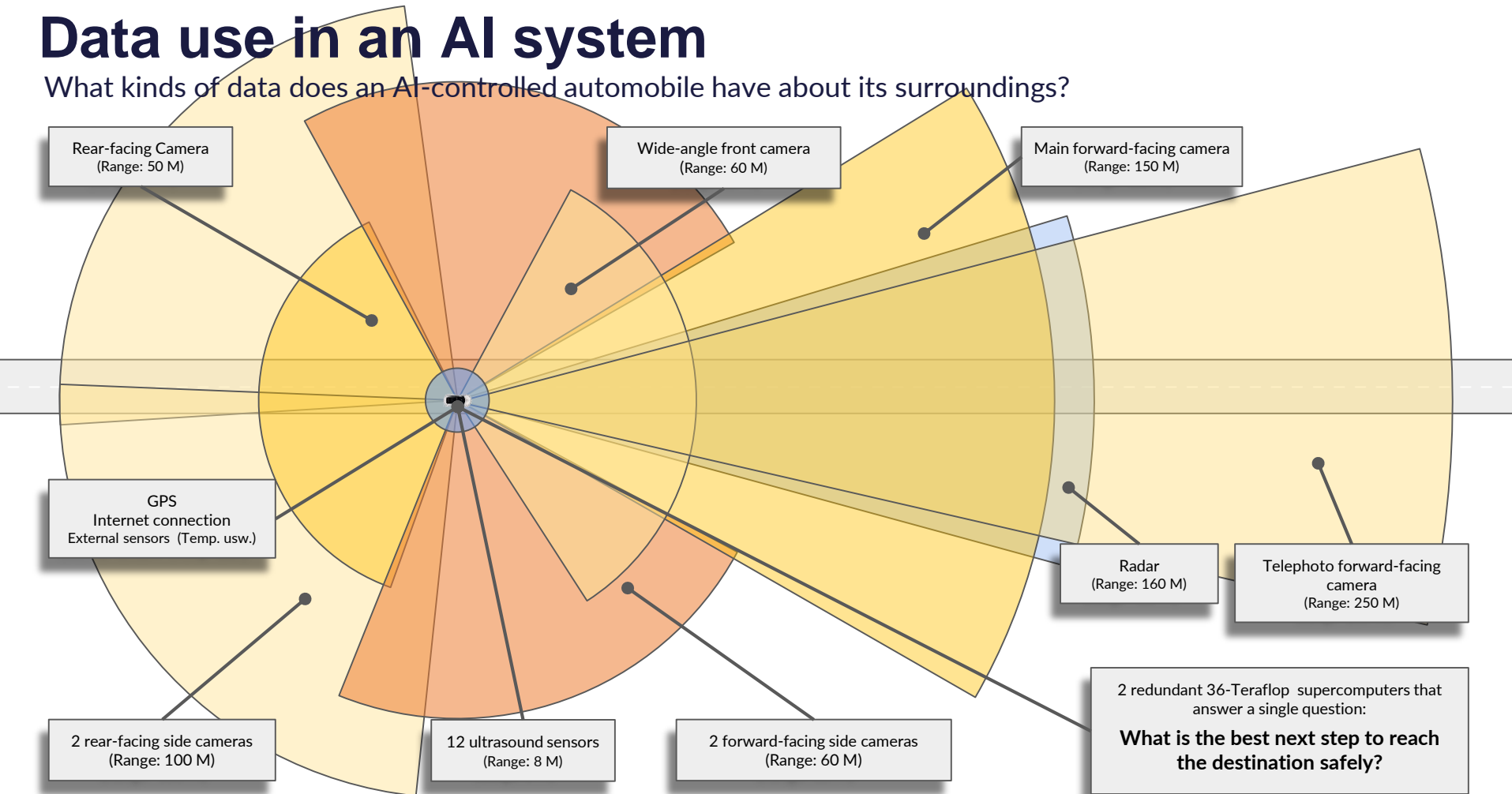
What kinds of data does an AI-controlled automobile have about its surroundings?



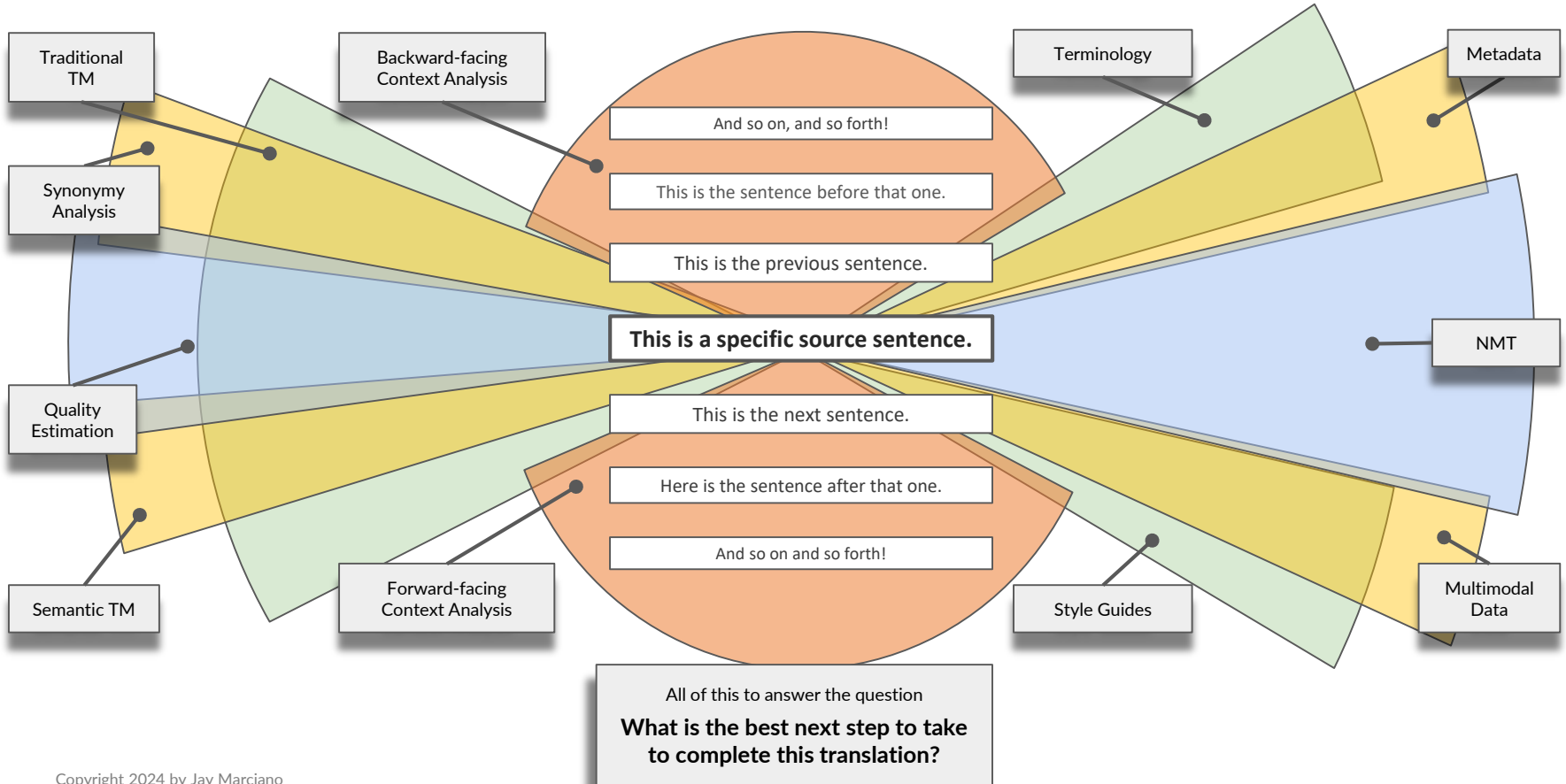


# Data use in an AI system

What kinds of data does an AI-controlled automobile have about its surroundings?



# What data and processing might an AI translation system leverage?



# **Evolving** Machine Translation

# Quickening evolution of Machine Translation



## Rules-based MT

Build a translation with language-specific algorithms and bilingual dictionaries

**State-of-the-art for ~50 years**



## Statistical MT

Find the most probable translation using statistical modeling

**State-of-the-art for ~17 years**



## Neural MT

Predict a translation using a deep neural network

**State-of-the-art for ~6 years**

# Large Language Models



- 
- 
- 
- 
- 
- 

## Large Language Models

Predict translations similarly to NMT, but with a much more flexible and powerful technical foundation

**Will be state-of-the-art for X years**

de  
odels  
d can  
not

n task  
ral

ised

ated

**A rose by any other name ... GPT**



**Generative**

**Pre-trained**

**Transformer**



# Generative



Having the fundamental ability to generate text by predicting the best next word in a response.

The order the words and phrases in its responses are based on the patterns the model has learned from training data.

# Pre-trained

Refers to the machine learning on vast amounts of training data before any explicit task for the resulting neural network is defined.

In other words, the learning algorithms seek out information that will help it to generate fluent text without adapting that information to a particular task.



# Transformer

Refers to the neural network architecture at the heart of LLMs.

Considered state-of-the-art since their proposal by the Google Brain team in late 2017, transformer networks were initially designed to improve performance on sequence-to-sequence problems, such as machine translation.

## Significant advantage:

They process the input all at once (not sequentially), which enables the model to leverage long-range relationships in the input, such as the final word in a long sentence that has a bearing on the meaning of the very first word.

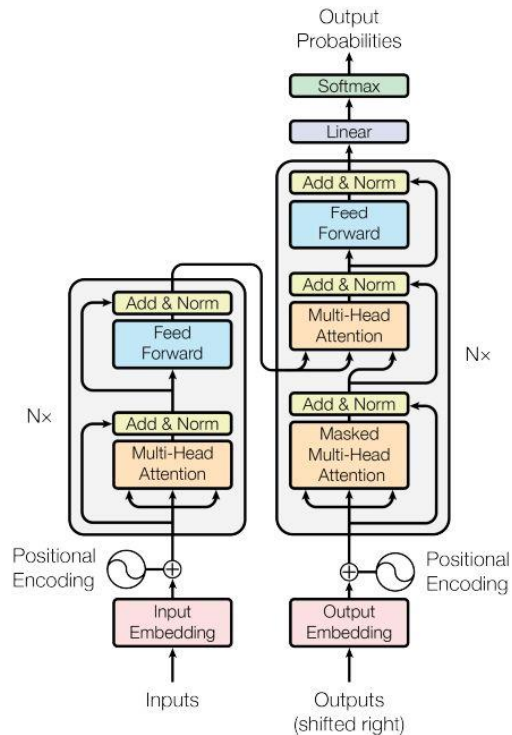
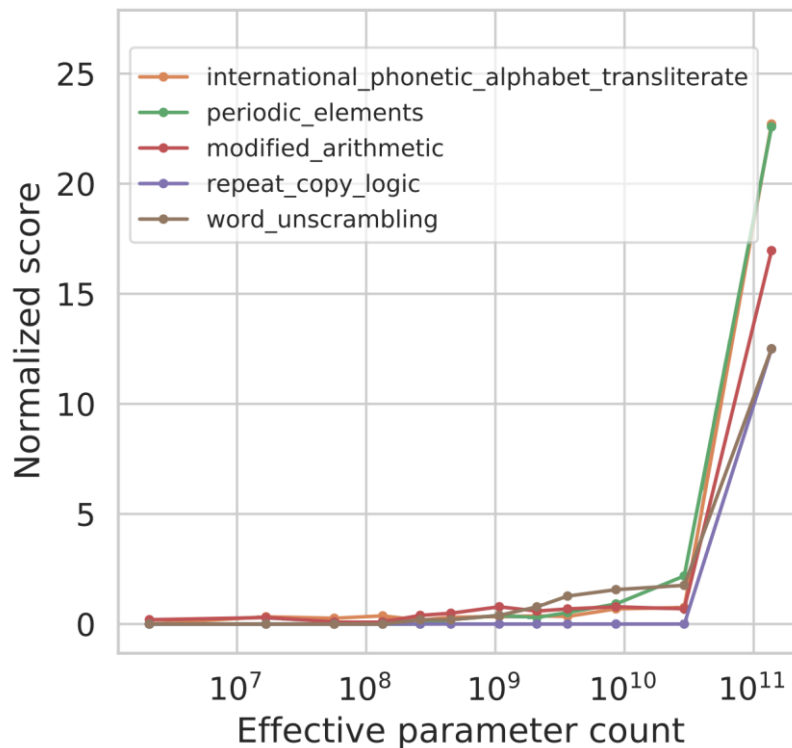


Figure 1: The Transformer - model architecture.

# Emergent

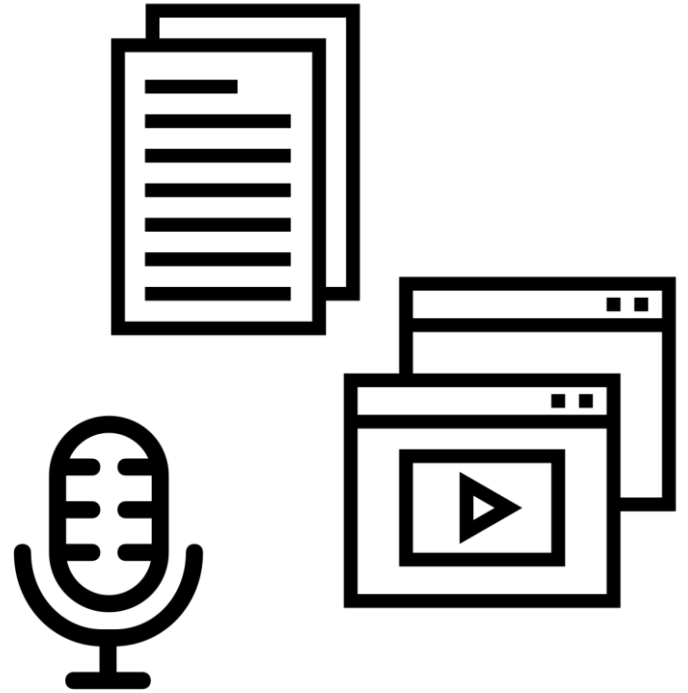
Refers to the sudden improvement in the performance of a task didn't improve in a predictable way in successively larger language models, but then suddenly improves significantly when the model reaches a particular size.



# Multimodal

Refers to a Large Language Model that is trained on material that contains various types of information (or modalities), such as text, images, audio, video, and even computer code.

Multimodal models process different types of information simultaneously, enabling them to perform tasks that require knowledge of more than one type of data.



# Imagining a multimodal LLM



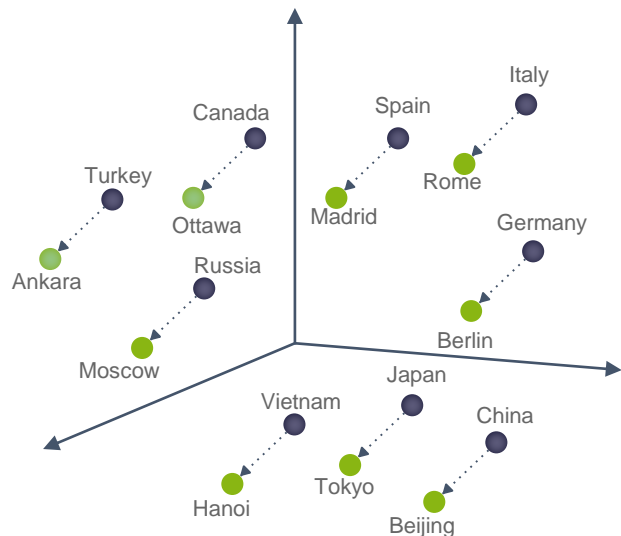
# Start with a pile of Legos



# What is “learned” about those building blocks

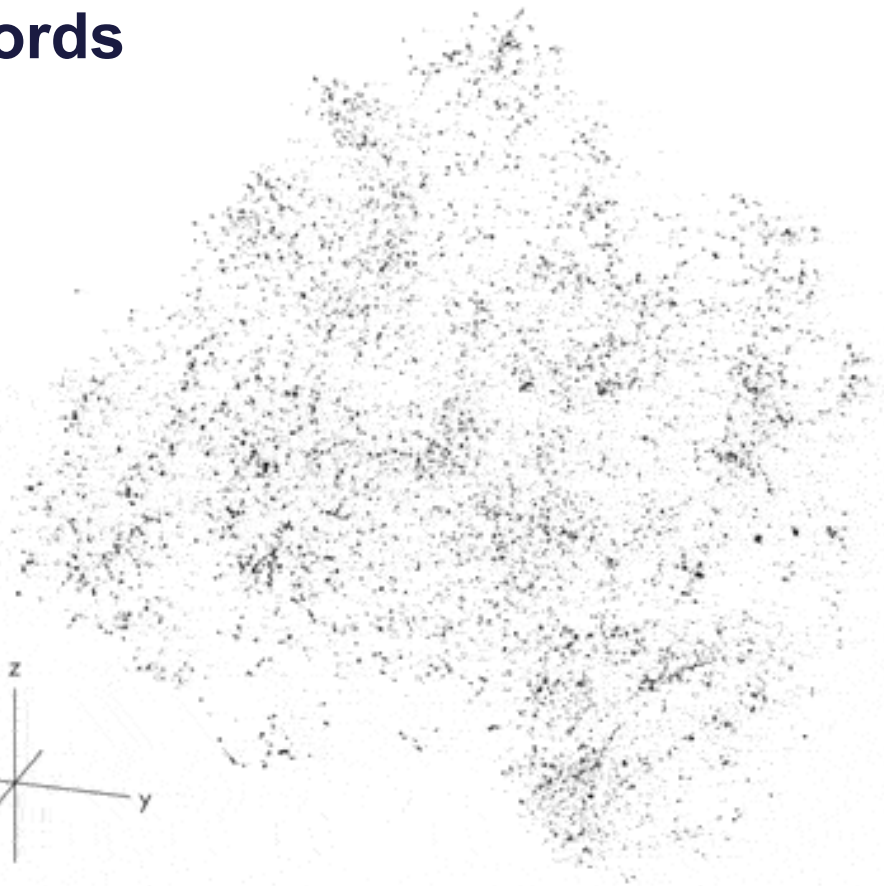
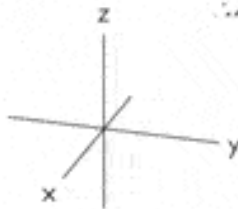
For every building block (word) in the training material, the Deep Learning system calculates a **word embedding**, a vector that contains semantic and grammatical information and indicates relationships among the words.

- Word embeddings are mapped in multidimensional space
- Similar words have similar values (or locations)
- The mathematical relationship between words that have a related meaning resembles the relationship between two other words that share that semantic relationship
  - Example: Countries and their capital cities



# What is “learned” about words

This information provides a multidimensional map of each supported language, showing the relationships among all of the words in those languages.





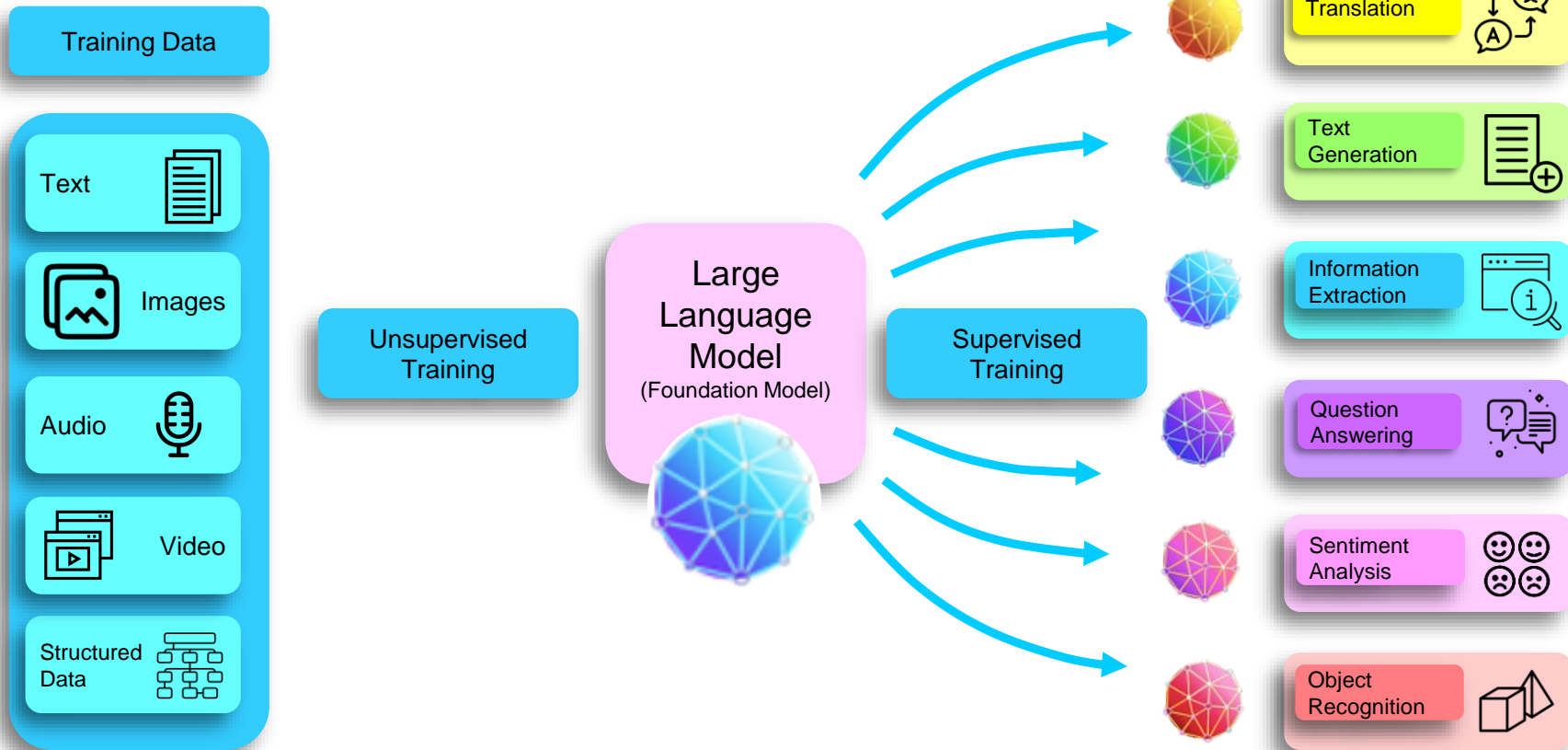
**Can we even imagine multidimensional space?**





# Multimodal Large Language Models

The new platform for MT ... and much more



**Let's see a multimodal LLM work**



# ChatGPT

Text generation in response to a text prompt

- Developed by OpenAI
- GPT 3.5 (released November 2022)
  - Released November 2022
  - 175 billion parameters
  - Context: 3,000 tokens
- GPT-4 (released March 2023)
  - 1.76 trillion (10x over 3.0)
  - Context: 25,000 tokens
- GPT-4o (released May 2024)
  - The “o” stands for “omni”
  - Context: 128k tokens
  - Supposedly a single LLM that provides all of the available services



Write a sonnet about post-editing machine translation



Amidst the whirring gears and hum of steel,  
A marvel of technology resides,  
A tool to make our language barriers yield,  
And bring our tongues and cultures side by side.

Yet though its speed and reach we all admire,  
And welcome the relief it brings us all,  
There still remains a need for human fire,  
To give the text that certain human drawl.

For only in the mind of man, we find,  
The depths of nuance, humor, and emotion,  
A richness that machines can't quite unbind,  
A certain je ne sais quoi of expression.

So let us hail this wondrous, modern aid,  
And wield it deftly with our human blade.

# ChatGPT

Graphic generation in response to a prompt

Prompt:

Create a photorealistic image of Michelangelo's David but instead of David's head use Mark Twain's



# ChatGPT

Graphic generation in response to a multimodal prompt

## Prompt:

Create a photorealistic image of the man in this photograph giving a presentation about generative AI to an audience of 200 people

## Additional Response:

Here is a photorealistic image of a middle-aged man with a grey beard and short hair, wearing a green sweater and a plaid shirt, giving a presentation about generative AI to an audience of 200 people.



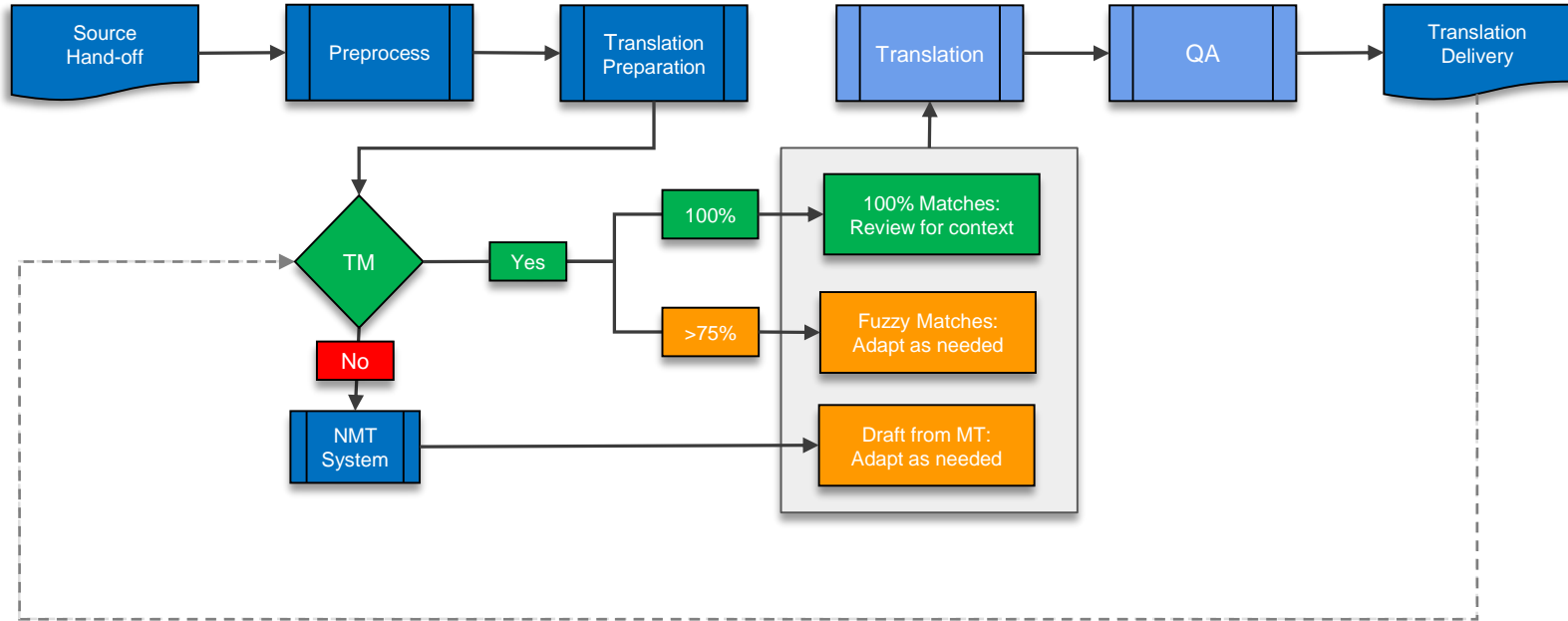
# That's impressive, but let's not freak out

- LLMs are tools that organize and leverage massive collections of data.
- This leveraging is *not* done because the LLM wants to but because a person has prompted the LLM to do something.
- In other words, an LLM does not have “agency”
- *An LLM does not have* the capacity to act independently and to make free choices, based on its will

# What about Gen AI for Translation?

# 1. Old-school Post-Editing

Translation workflow with MT and post-editing



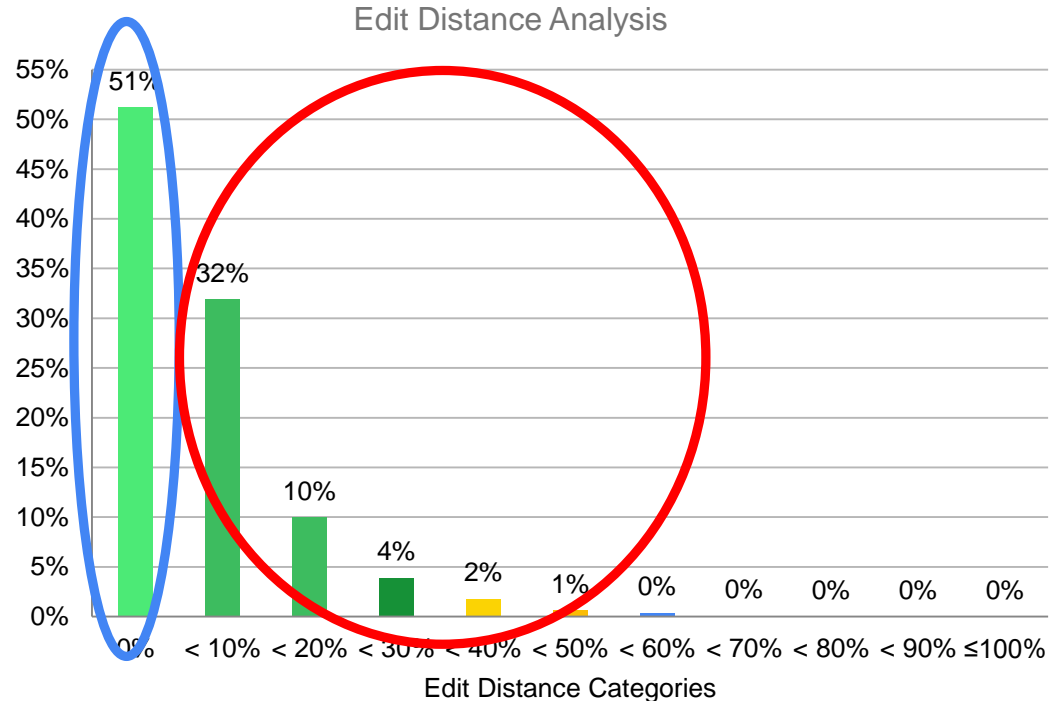


## 2. Automatic quality estimation

- Edit-distance analysis from a Berlin-based LSP
- Trained NMT
- Single customer
- All languages
- All content
- 12 months' work with customer

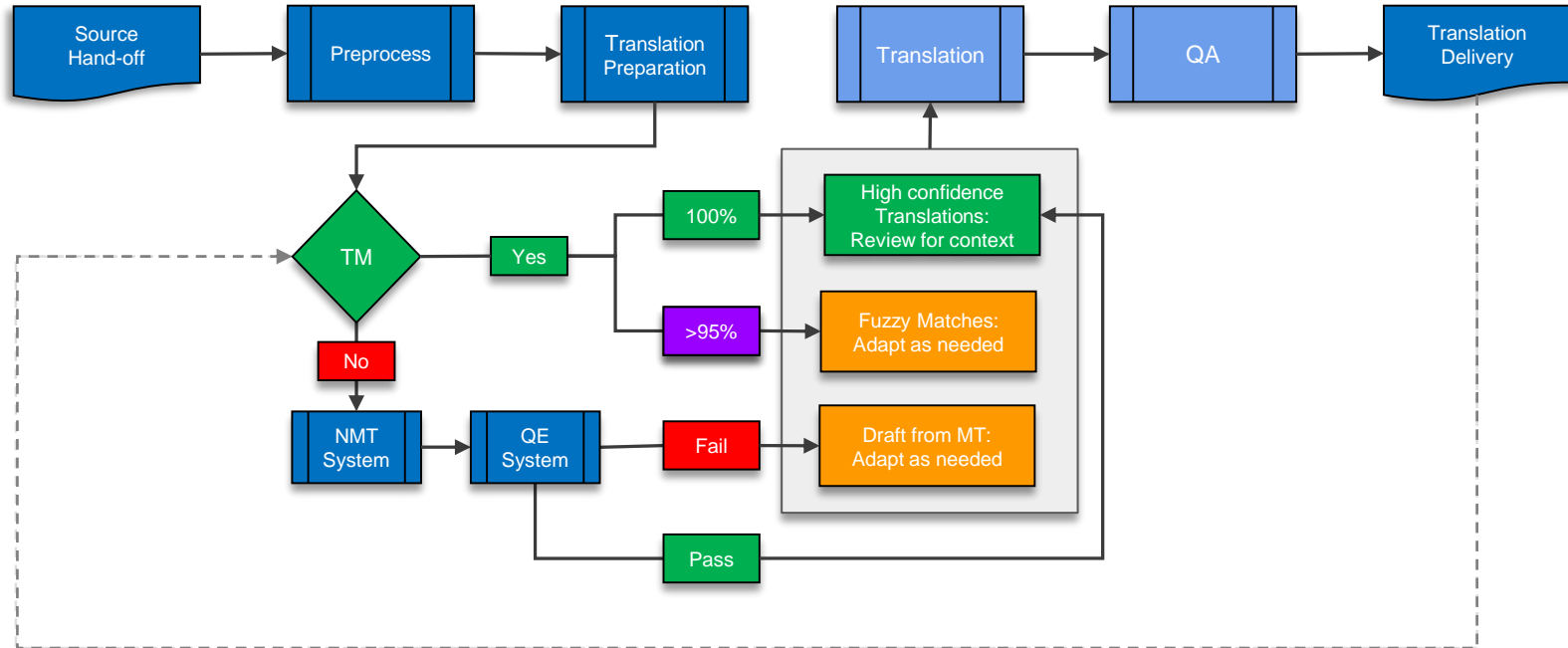
QE systems can be used to:

- Predict which segments will fall into the 0% Edit Distance Range
- Provide annotation for segments that are predicted to need edits



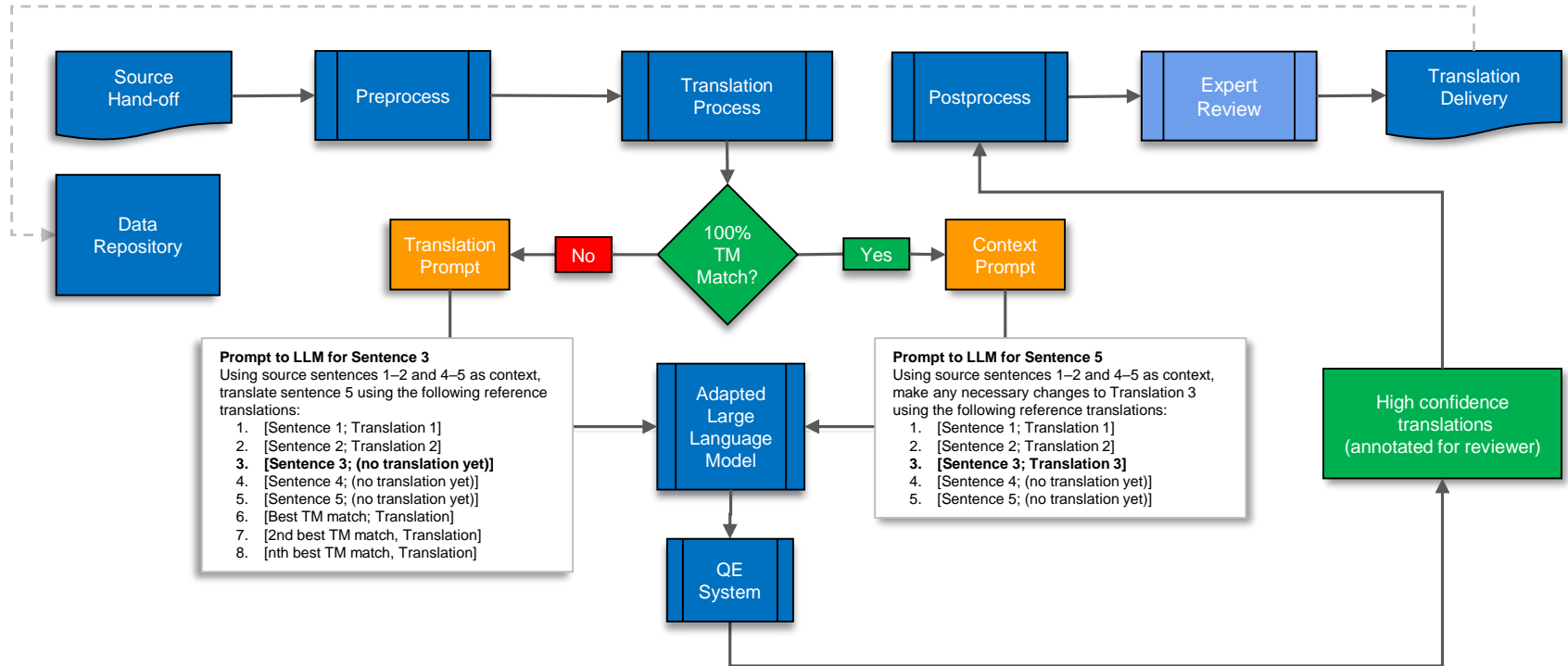
## 2. Automatic quality estimation with post-editing

Translation workflow with MT, QE, and post-editing



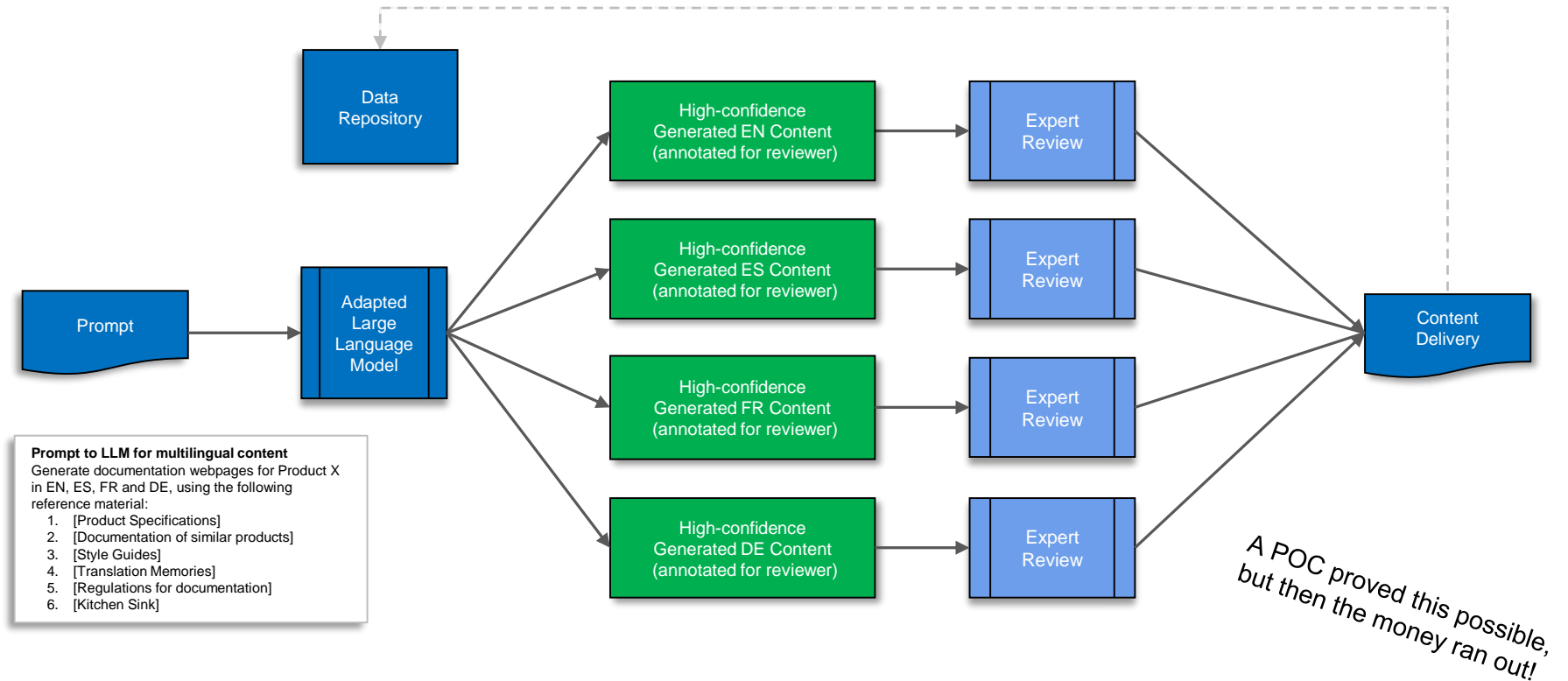
# 3. Translation workflow with generative AI

Translation workflow contextual and TM prompts to a Large Language Model



# 4. Simultaneous generation of multilingual content

Multilingual content generation workflow with AI and expert review



**Welcome to the**  
**Post-post-editing World**

# What tasks are being created by AI in our industry?

- Data Curation
- Data Science
- Terminology Management
- Corpus Linguistics
- Computational Linguistics
- Prompt Engineering
- Language Technology Assessment
- Language Process Analysis
- Communication Analysis
- AI Ethics Review
- Subject Matter Expert Review
- Machine Learning Supervision
- “Gatekeeper” Review
- AI Evaluation



# Data Curator

## Tasks:

- Ensure data quality and consistency through the implementation of data governance practices
- Define and enforce data styling and terminology standards
- Develop and maintain a data catalog to ensure efficient data discovery and accessibility
- Work with stakeholders to understand data requirements and ensure data availability and accuracy
- Collaborate with data scientists and machine learning teams to support the development of new models and features
- Manage data retention and archival policies to ensure compliance with data privacy regulations

## Requirements:

- Experience in data management and data governance
- Familiarity with machine learning and data science methodologies
- Strong understanding of Knowledge Graphs
- Experience with SQL and NoSQL databases
- Ability to work with cross-functional teams and manage stakeholder relationships
- Strong problem-solving and analytical skills
- Bachelor's or Master's degree in Computer Science, Data Science, or a related field

## What strengths will be required to thrive in these jobs?

- All the great skills language professionals already have
  - Subtle and sophisticated knowledge of language
  - Excellent proficiency in two or more languages
  - No fear of other languages
- Deep curiosity
- An appreciation for “algorithmic thinking”
- Comfort level (or better) with data and databases
- The audacity to work on skills that will make the old you redundant

**Sure, but what else?**

# **Imagination and Intention**



**We shape our technology, and afterwards that technology shapes us.**

# **Thank you for your attention!**

[www.linkedin.com/in/jaymarciano/](https://www.linkedin.com/in/jaymarciano/)

**Join us in Chicago for AMTA 2024  
30 September – 2 October  
In Chicago**

**Tomorrow (14 June 2024) is the last day to submit a proposal!**

**Information at [amtaweb.org](http://amtaweb.org)**