Exploiting LLMs for XR Applications

Montse Cuadros

Arantza del Pozo

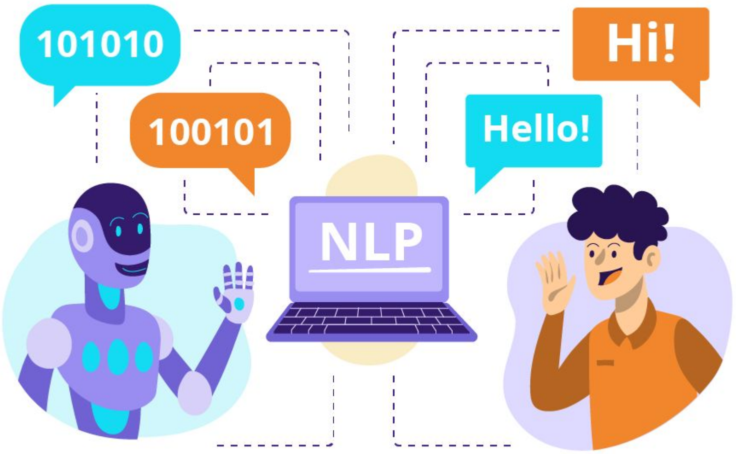
Speech and Natural Language Technologies

Vicomtech

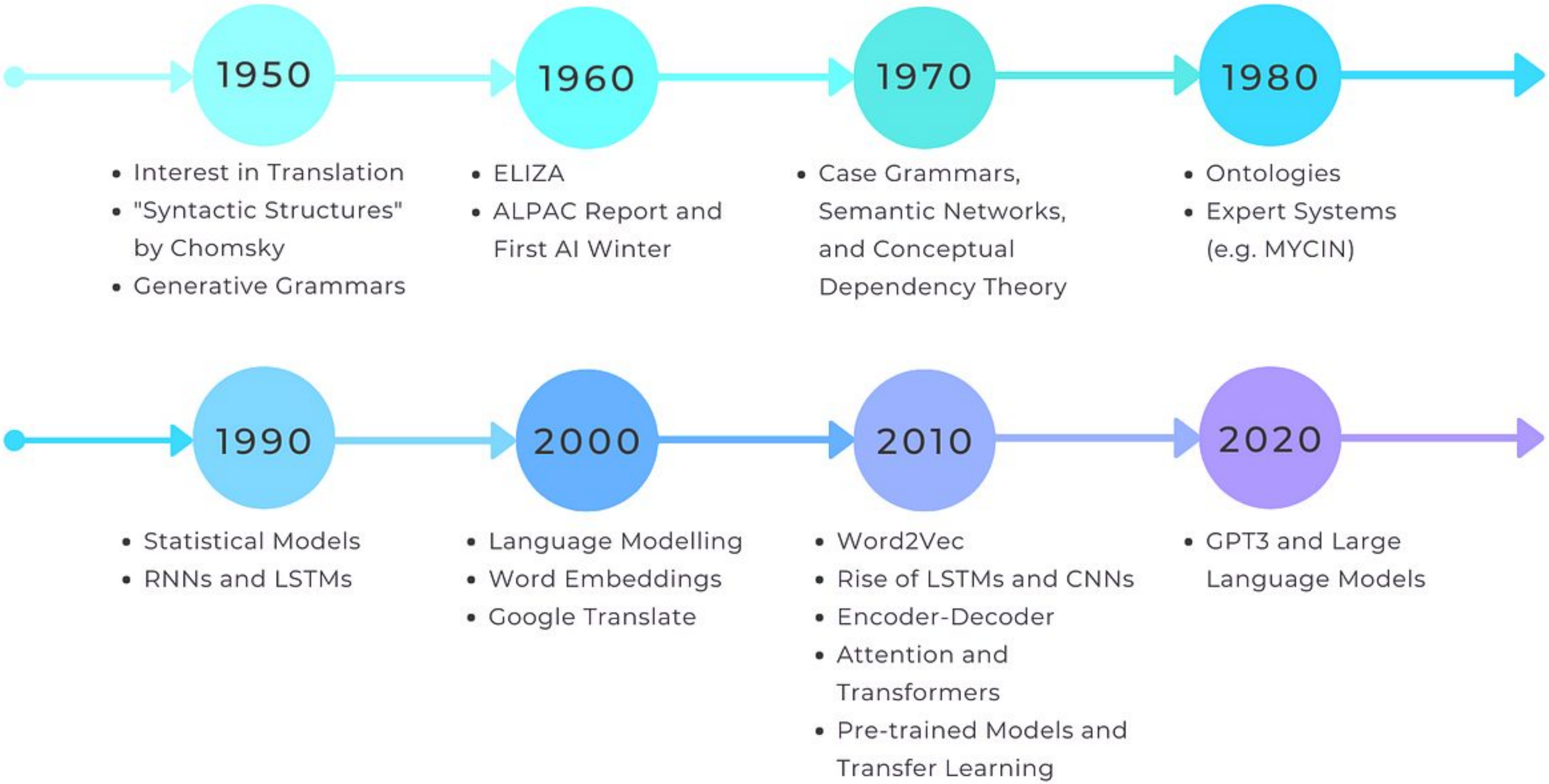
Contents

Introduction

- Natural Language Processing (NLP) is a subfield of IA
- It refers to the analysis of the natural language to understand human language as it is spoken and written.
- NLP can be seen in every application such as virtual assistants like Siri, google assistant, translation and search engines.

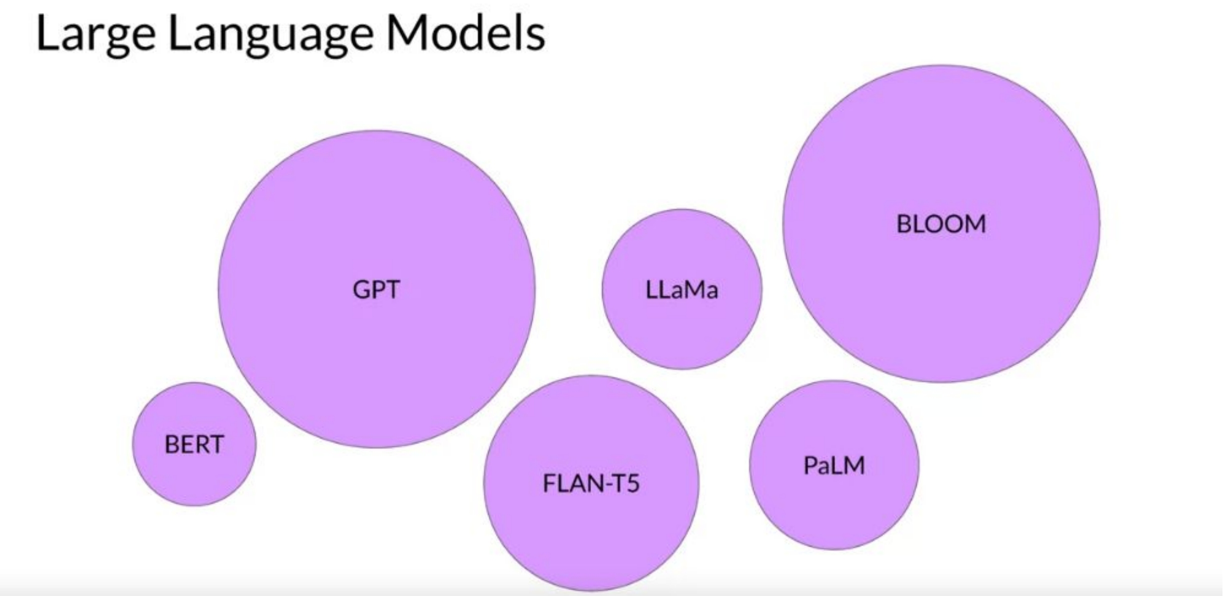


Evolution of NLP



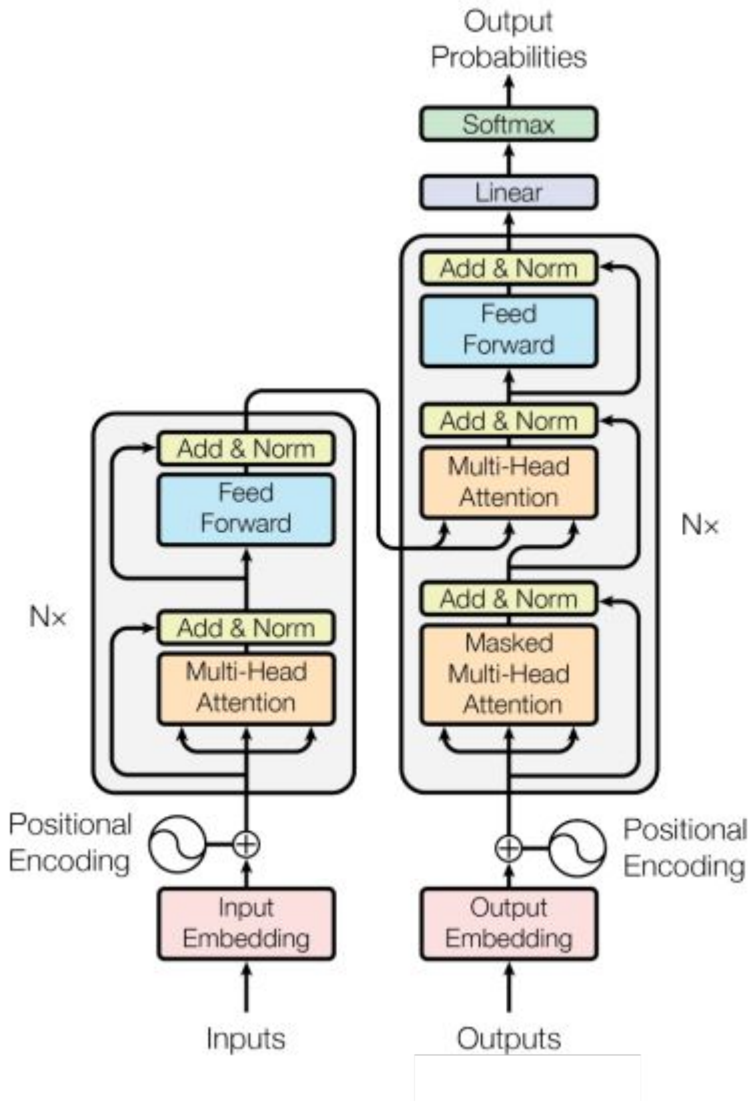
What are LLMs

- AI systems that have been trained to understand and generate natural language text. A well-known example is ChatGPT.
- These models can interpret and produce text in different languages and contexts
- Various applications, some for instance in fields such as text generation and/or summary generation.

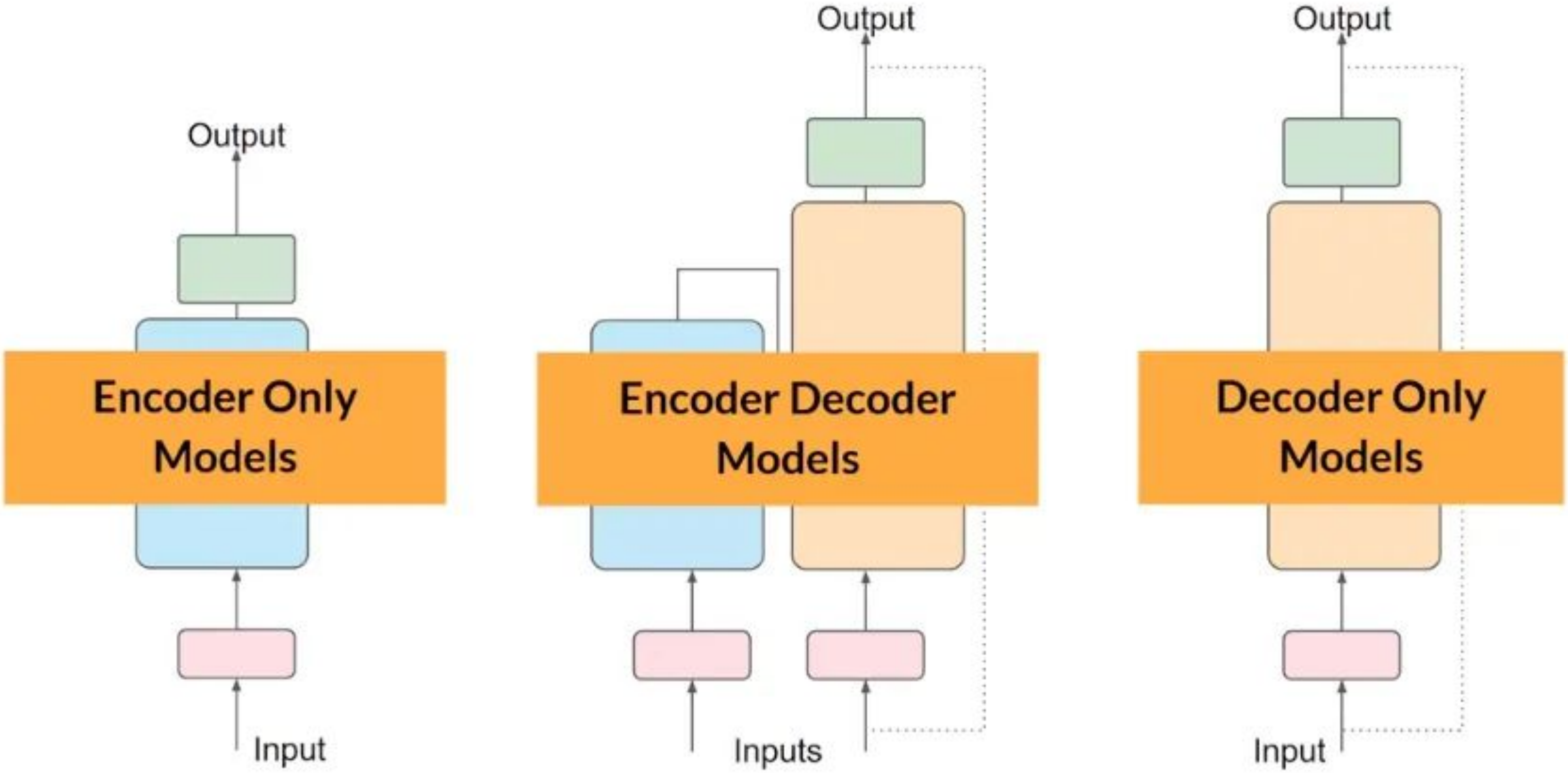


Transformers

- Transformers are neural networks
- The architecture allows models to learn complex relationships between words and sentences paying attention to different parts of text
- Auto-regressive attention, allows the machine to determine which words are relevant for the generation of a coherent response.



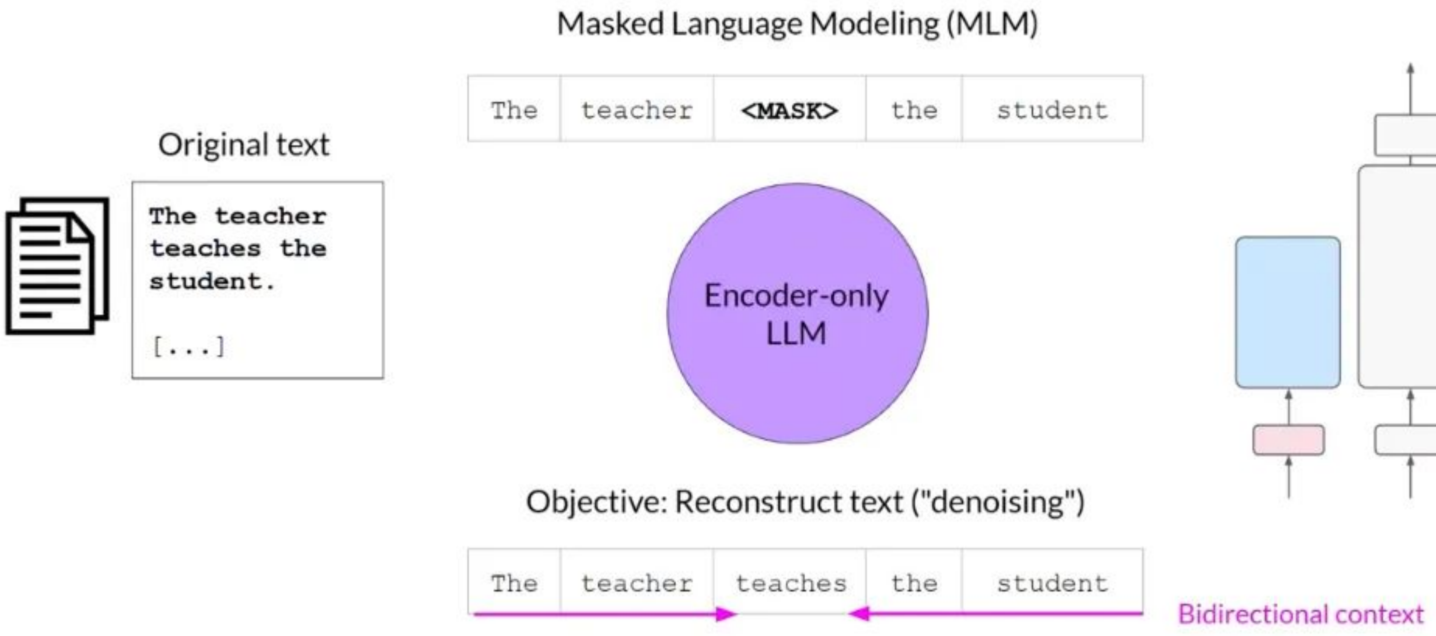
Architecture of LLMs - Transformers



https://medium.com/@yash9439/introduction-to-llms-and-the-generative-ai-part-2-llm-pre-training-and-scaling-laws-71e06c83998b

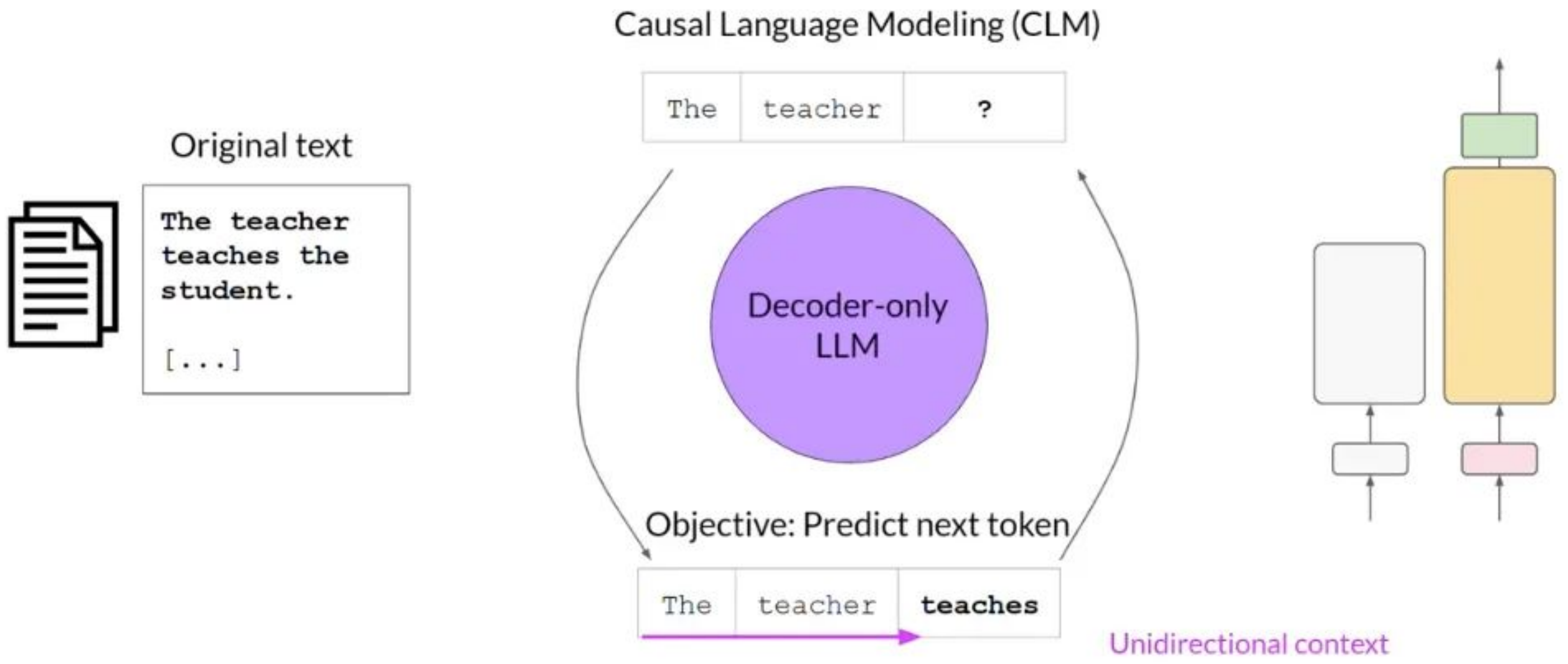
LLMs models: Encoder-only

- Pre-trained using masked language modeling.
- Tokens in the input sequence are randomly masked, and the model's objective is to predict the masked tokens to reconstruct the original sentence.
- Autoencoding models capture bi-directional representations of the input sequence
- Used in tasks such as **sentiment analysis or named entity recognition**.
- Autoencoder examples are BERT and RoBERTa.



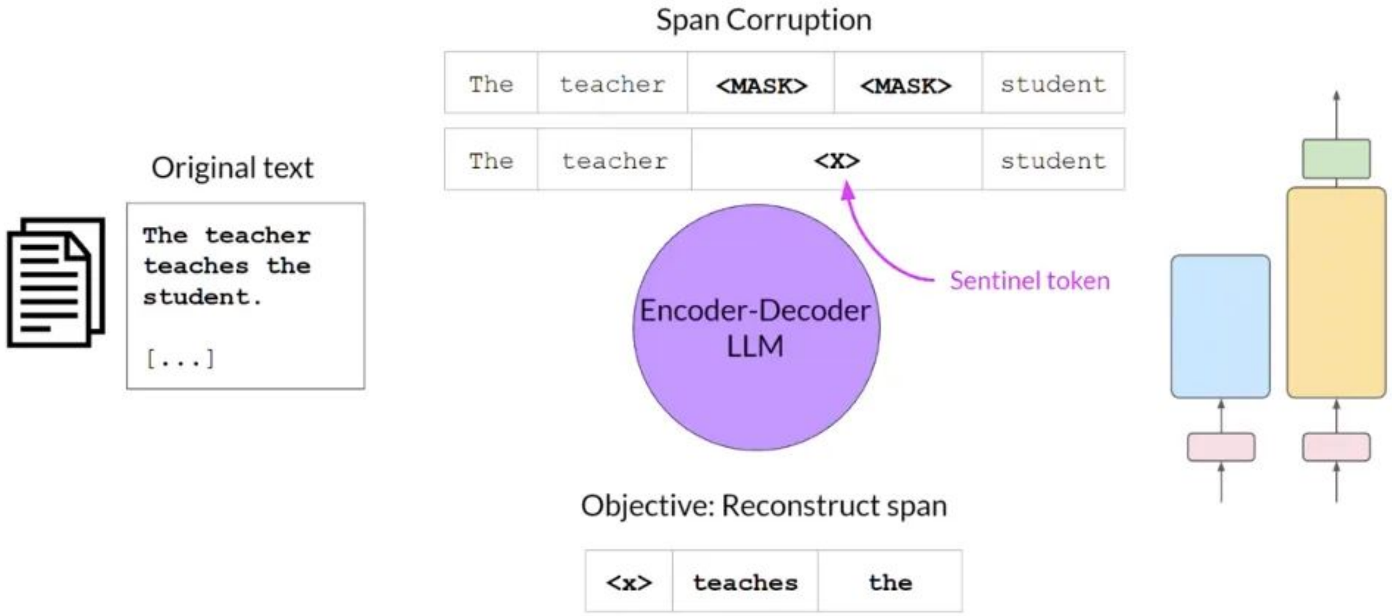
LLMs models: Decoders-only (Generative models)

- Autoregressive models build a statistical representation of language.
- Pre-trained using causal language modeling.
- The objective is to predict the next token based on the previous sequence of tokens.
- These models mask the input sequence and can only see the input tokens leading up to the token in question. By learning to predict the next token from numerous examples
- Often used for **text generation** tasks, example models for instance are GPT and BLOOM

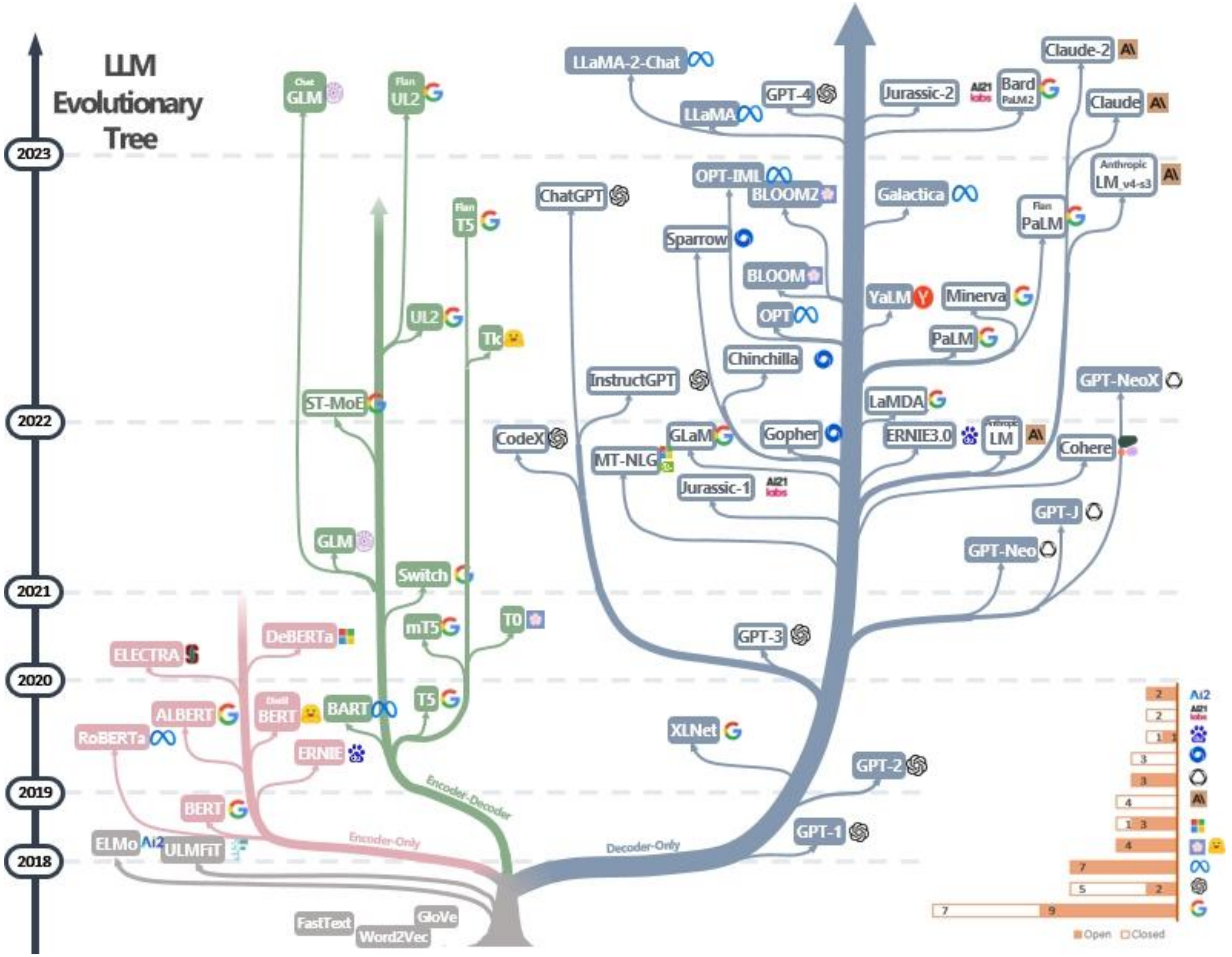


LLMs models: Encoder-decoder

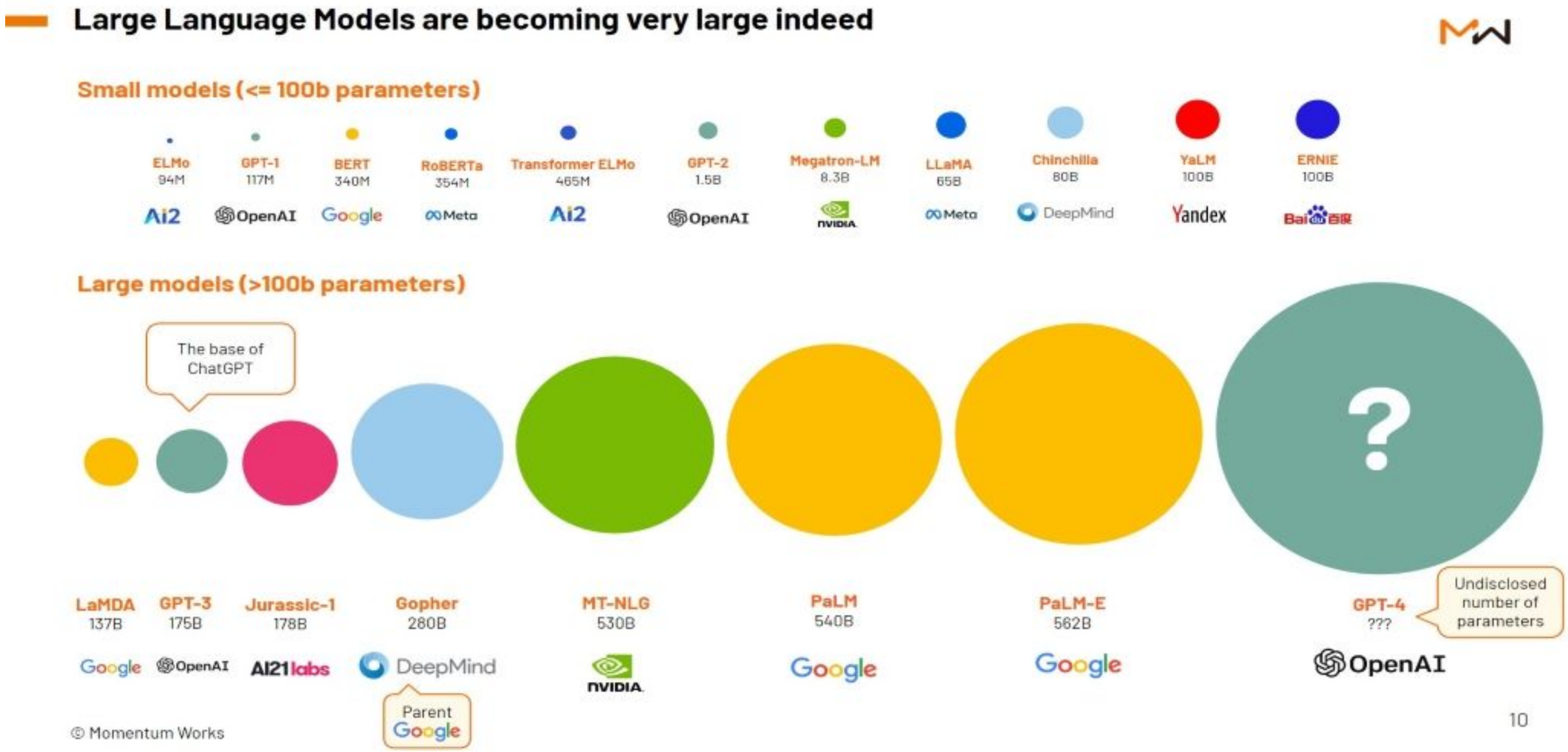
- Sequence-to-sequence models utilize both the encoder and decoder components of the original transformer architecture.
- Pre-training objective for these models varies depending on the specific model. For example, the T5 model is pre-trained using span corruption, where random sequences of input tokens are masked and replaced with a unique Sentinel token.
- Decoder is then tasked with reconstructing the masked token sequences auto-regressively.
- Used for **translation**, **summarization**, and **question-answering** tasks. Example BART



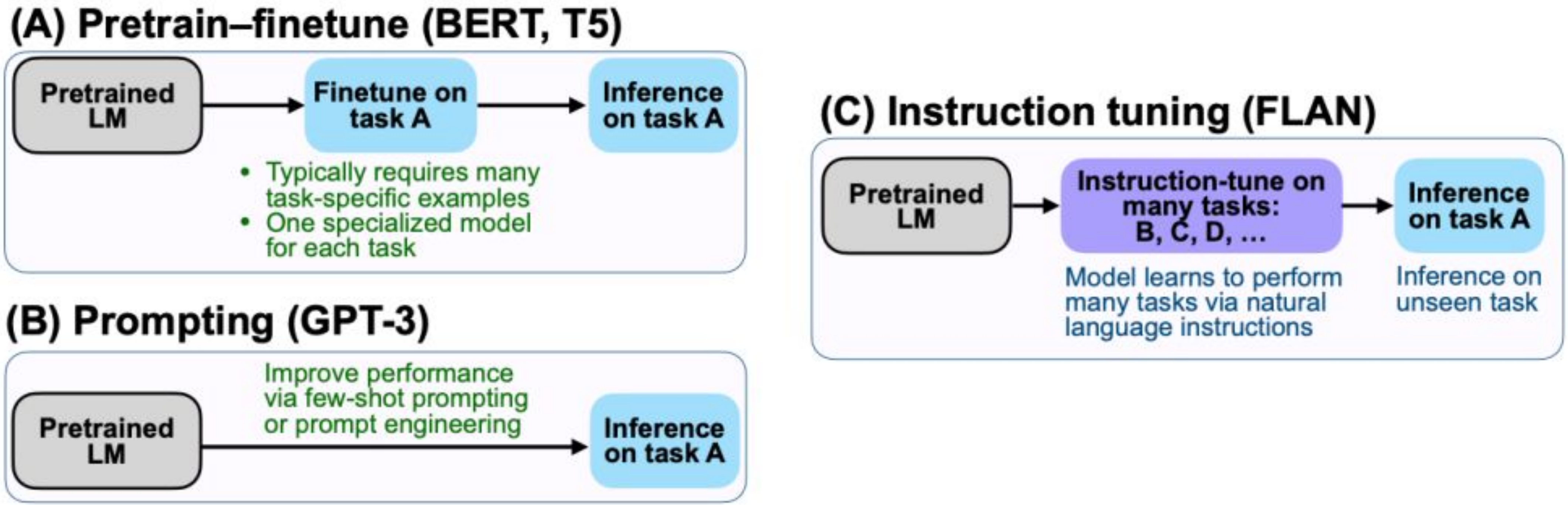
Example of evolution of LLMs



Size matters



Ways of using LLMs



Prompting

- Method of providing a specific input or instruction to an LLM to generate desired output
- Text or a set of instructions that guide the model on what kind of response is expected.
- The choice of words and the structure of the prompt can significantly influence the output generated by the LLM.
- Crafting effective prompts is an important skill when working with these models.
- Generative models, can solve tasks with different prompting approximations, depending on the problem and on the data available

Prompting in Generative models

Zero-shot prediction

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  cheese =>                          ← prompt
```

One-shot prediction

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  sea otter => loutre de mer        ← example
3  cheese =>                          ← prompt
```

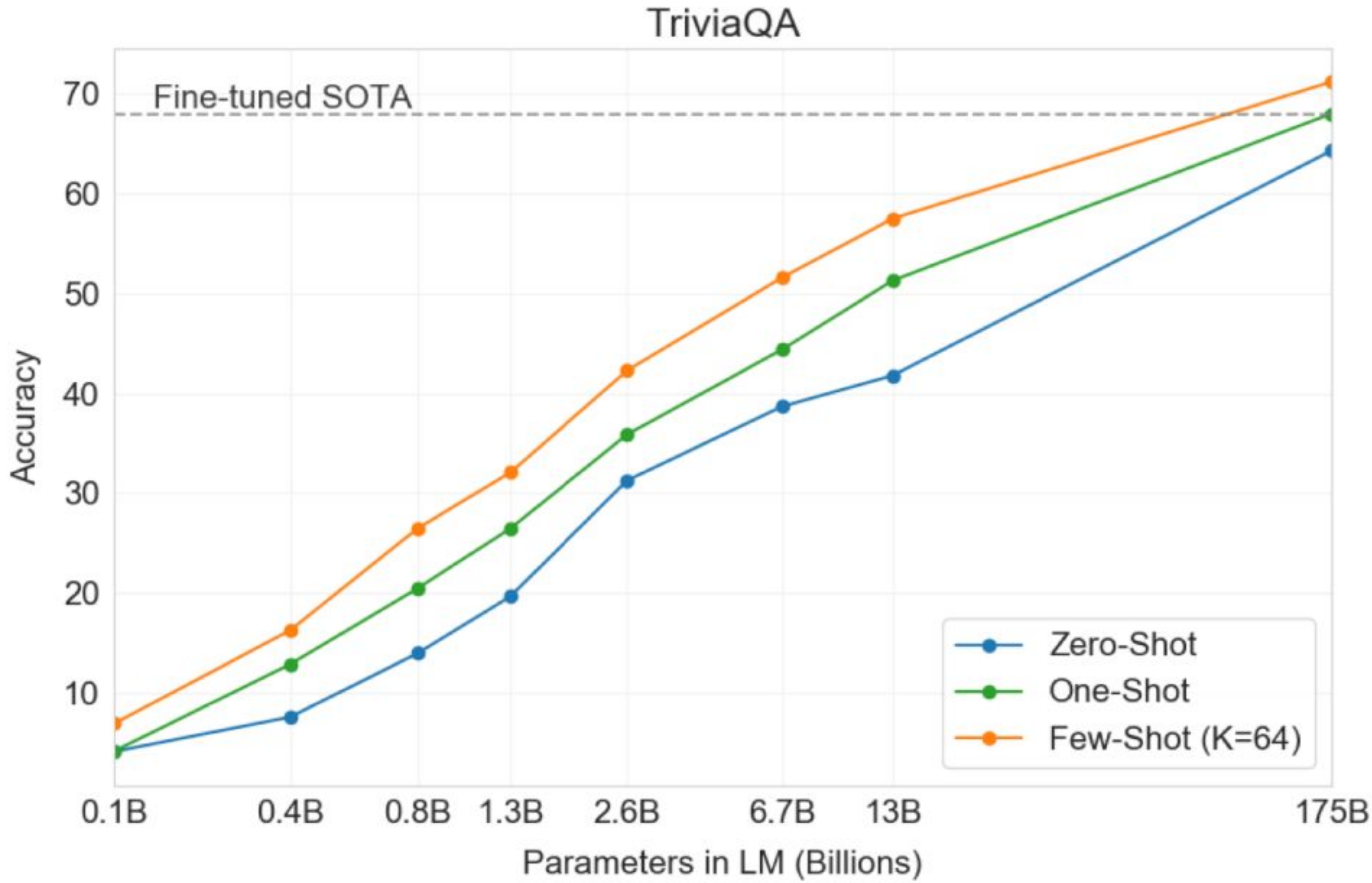
https://datascience.stackexchange.com/questions/115554/how-exactly-does-in-context-few-shot-learning-actually-work-in-theory-under-the

Prompting in Generative models

Few-shot prediction

Fine-tuning

https://datascience.stackexchange.com/questions/115554/how-exactly-does-in-context-few-shot-learning-actually-work-in-theory-under-the



TriviaQA
Fine-tuned SOTA
Accuracy
70
60
50
40
30
20
10
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)
Zero-Shot
One-Shot
Few-Shot (K=64)

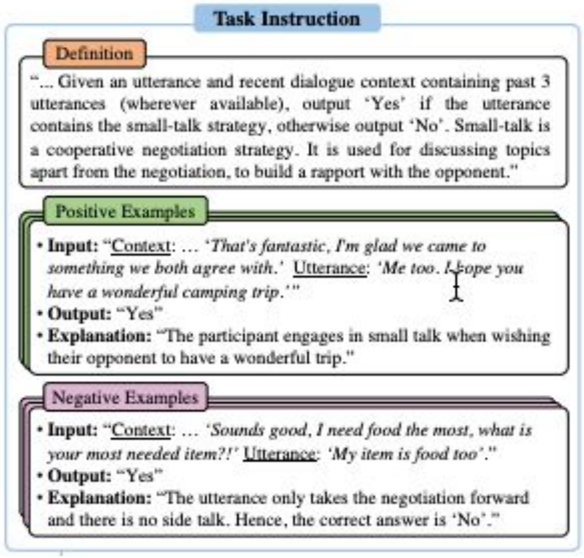
Prompting in Generative models

- Instruct-based prompting help LLM follow instructions.

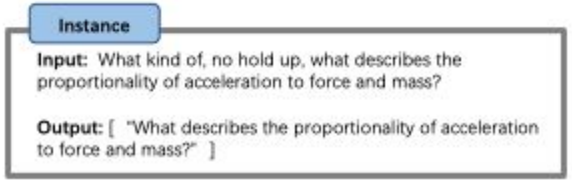
- Get the model to more reliably perform complex tasks like planning, intent & entity recognition, and even avoid hallucinations.

- Exist multiple Instruct Datasets even Multimodal: https://github.com/yaodongC/awesome-instruction-dataset#the-multi-modal-instruction-datasets

- Models such as InstructGPT, ChatGPT



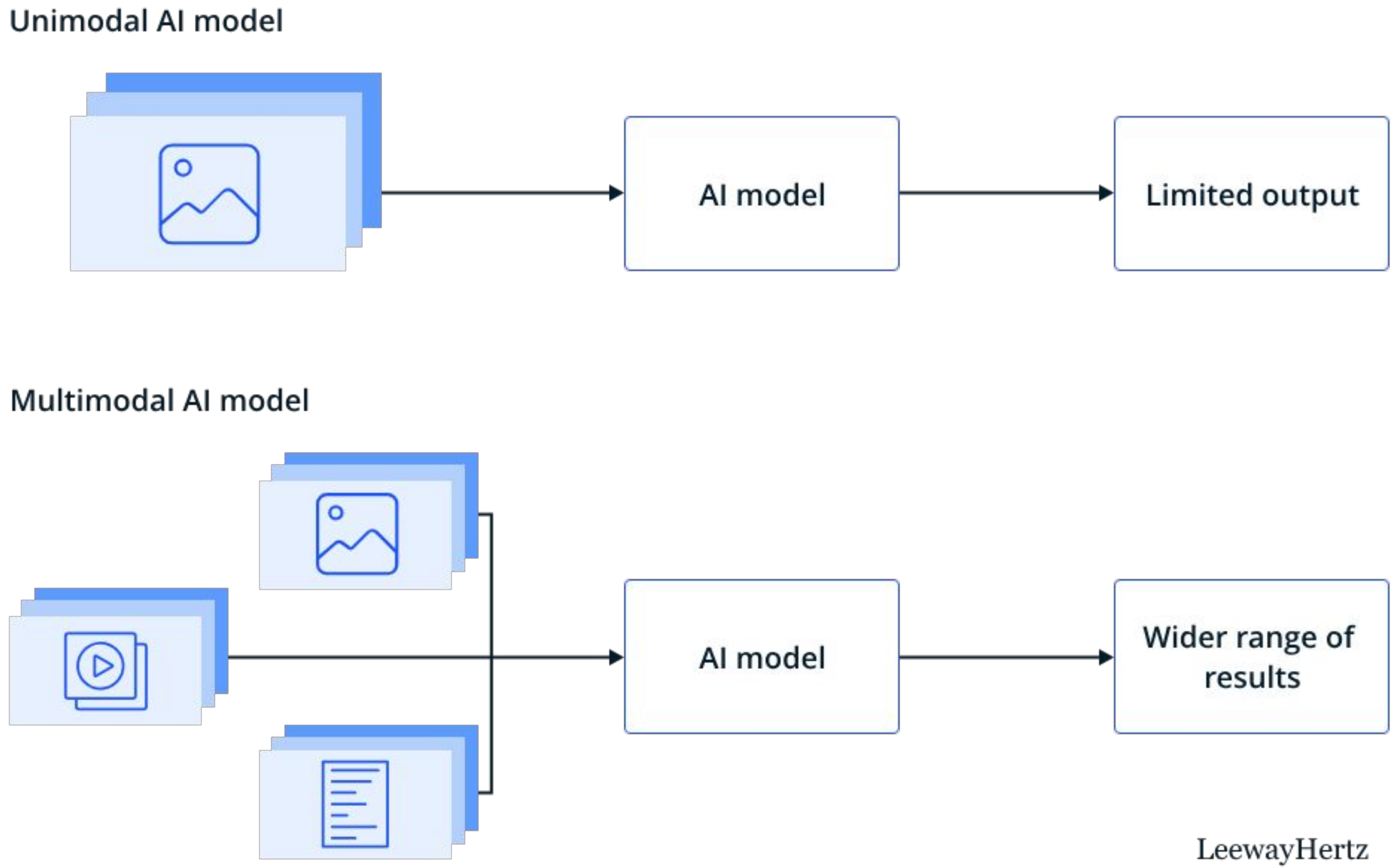
(a) An example of INSTRUCTIONS in Super-Natural Instruction dataset.



Multimodal LLMs (MLLMs)

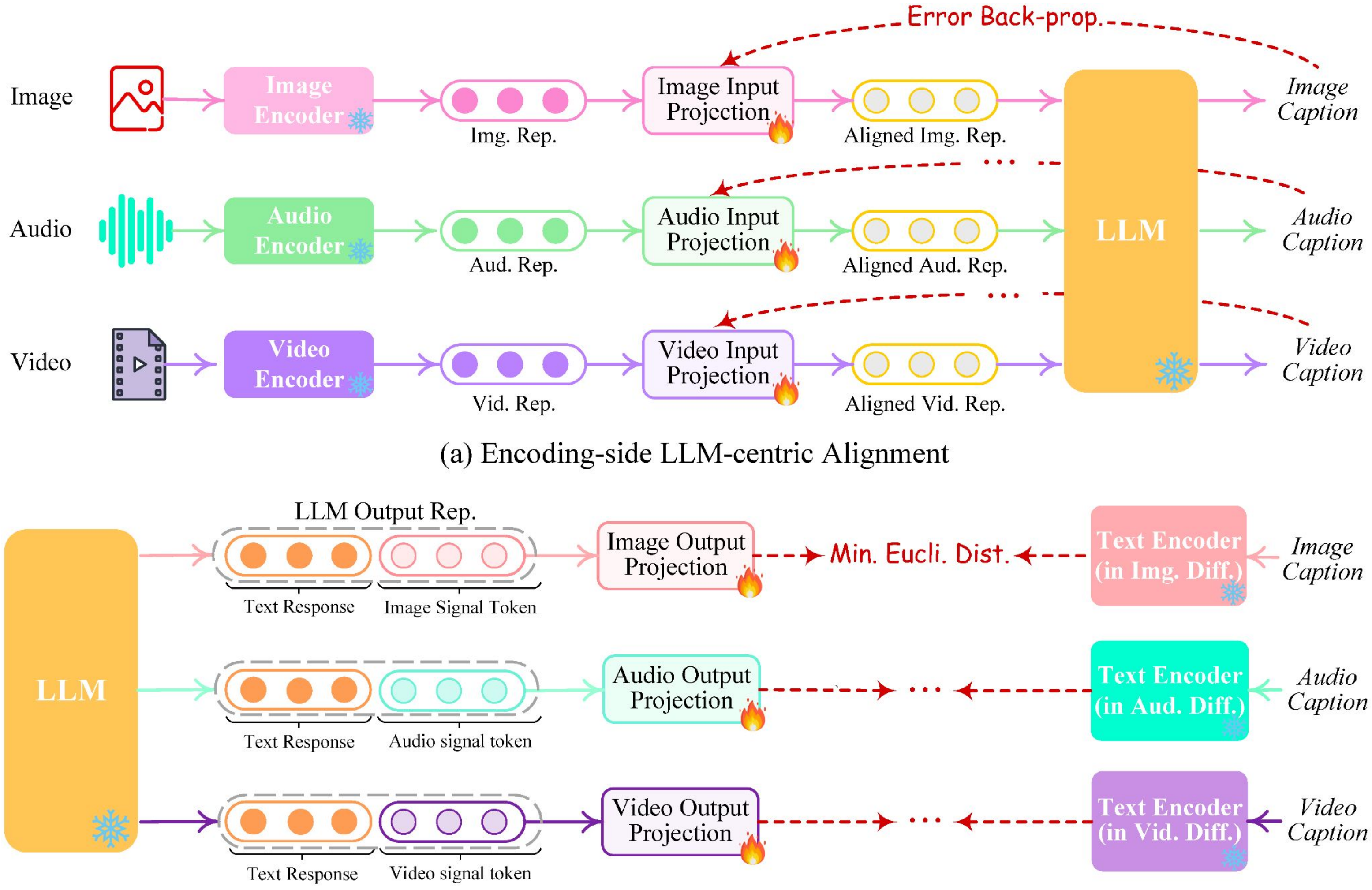
- Able to process text, interpret and generate information from a variety of sources and in different formats, such as images, text and audio.
- Combination different modalities allows them to provide more accurate and contextualised results.
- Based on neural network architectures that can process and relate data from different modalities, allowing them to learn to relate and contextualise information from different sources.

Multimodal LLMs (MLLMs)



MLLMs - architecture

NExT-GPT



(a) Encoding-side LLM-centric Alignment

(b) Decoding-side Instruction-following Alignment

Shengqiong Wu Hao Fei* Leigang Qu Wei Ji Tat-Seng Chua, NExT-GPT: Any-to-Any Multimodal Large Language Model, arXiv 2309.05519, 2023

MLLMs - data requirements

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
▶ *Existing data*							
MiniGPT-4 [70]	CC [7], CC3M [45]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [32]	SD [43]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [65]	COCO [33]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [67]	MS-COCO [33], VG [26]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [65]	COCO [33], CC3M [45], LAION [44]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [29]	WebVid [4]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [36]	ActivityNet [17]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [64]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [11]	Multiple	T+I/V→T	Auto	✗	-	-	~ 1.6M
MIMIC-IT [27]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [49]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [68]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^3IT [30]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [61]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [69]	Clotho [13], VGGSS [8]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [60]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
▶ *In this work*							
T2M	Webvid [4], CC3M [45], AudioCap [24]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

MLLMs - applications

- Example-1: Text → Text+Image+Audio

I've recently started practicing mindfulness and meditation as part of my daily routine, but sometimes it's challenging. Can you help me with some techniques?

Of course! Mindfulness and meditation can be a wonderful way to cultivate a sense of calm and focus. One technique you might find helpful is called "mindful breathing." It involves taking slow, deep breaths and focusing your attention on the sensation of the breath entering and leaving your body. Another technique is "body scan meditation," where you consciously scan your entire body from head to toe, paying attention to any sensations or discomfort. You can also try guided meditations that provide a structured practice. Would you like me to find a video that demonstrates mindful breathing or body scan meditation?

I wish to listen to a music for the meditation first.

I guess you may want some musics that sound soft.

0:00 / 0:09

Base on the audio content, visualize it in an image.

Can you infer what time is in the picture?

Based on the image, it is likely that the time is either early morning or late afternoon. The sunlight is either low or setting, creating a serene atmosphere.

- Example-5: Text+Audio → Text+Image+Video

Describe what can be heard from the sound:

0:00 / 0:01

The sound of a growling lion is a deep, rumbling noise that starts low in the chest and gradually builds in intensity. It is often associated with a sense of power and strength, as well as a sense of danger or warning. The growl of a lion can be heard from a distance and is often used by the animal to communicate with other members of its pride.

Displaying an image to help visualize it would be fantastic.

Sure, here is an image of a lion growling:

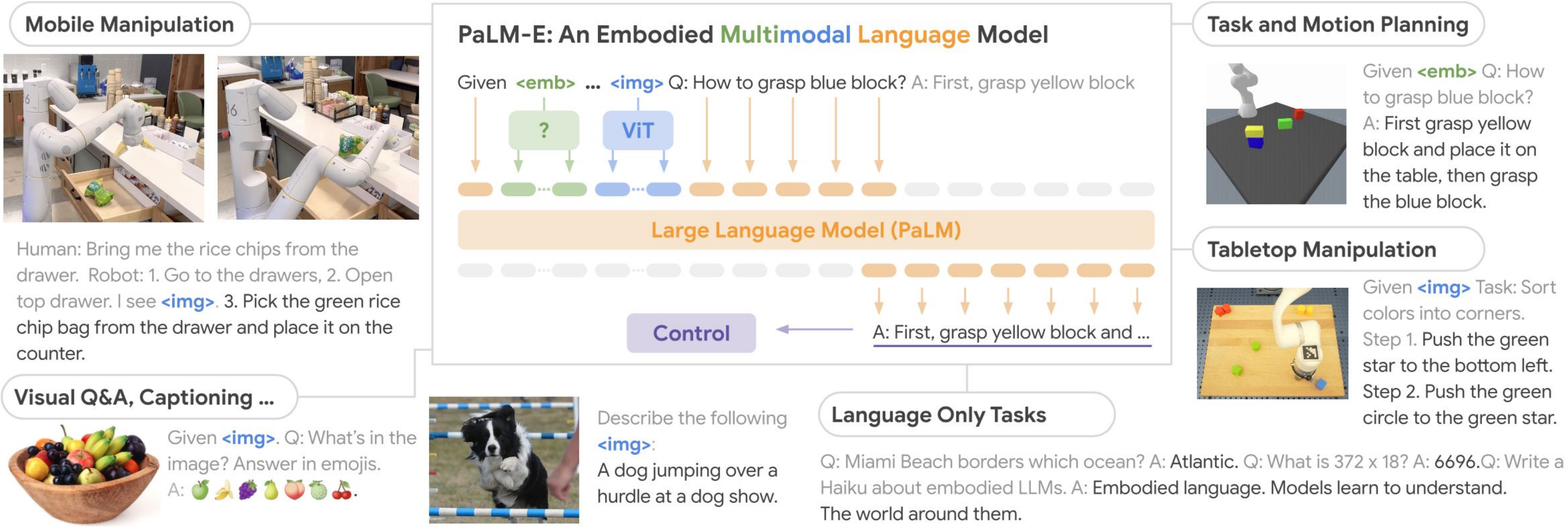
That looks great! How about showing a video that explains how the lions catch its prey?

MLLMs - applications

ChatGPT-4 deals with speech, text, video and images! :O

Aplications in XR

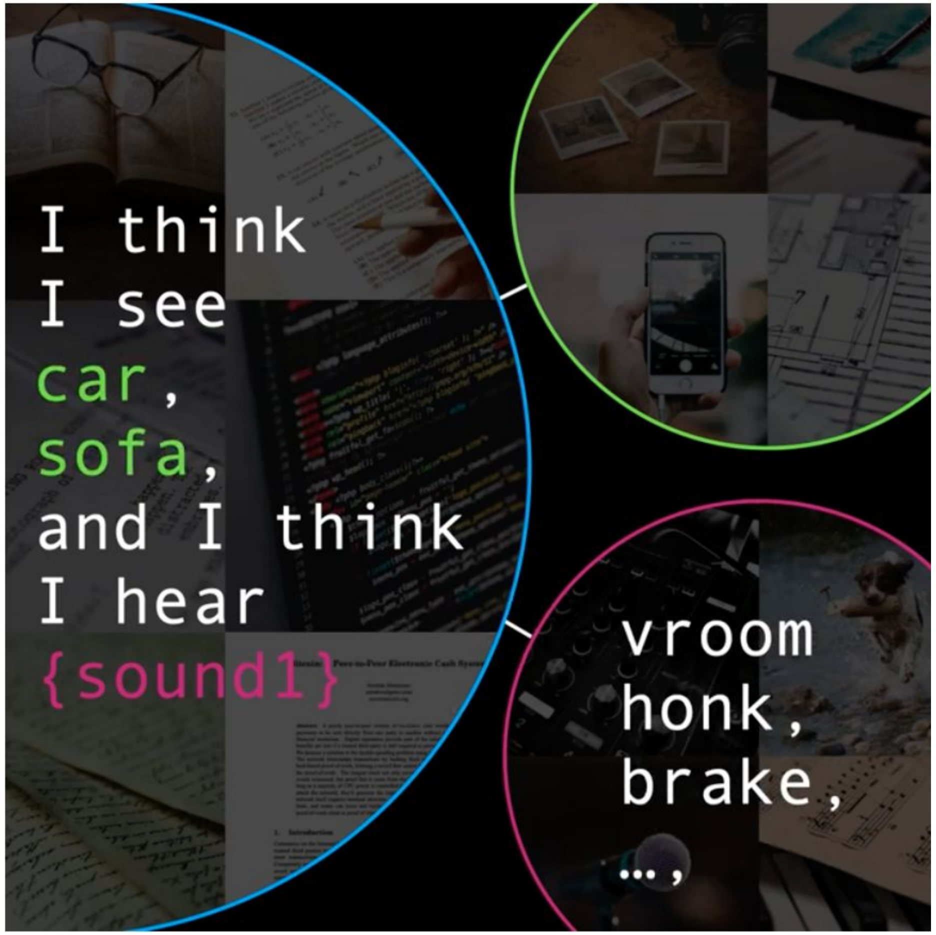
Embodied reasoning



Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Applications in XR

Zero-shot multimodal reasoning with language



Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., ... & Florence, P. (2022). Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Applications in XR

Simplifying user interactions in immersive environments

Project Mellon, NVIDIA

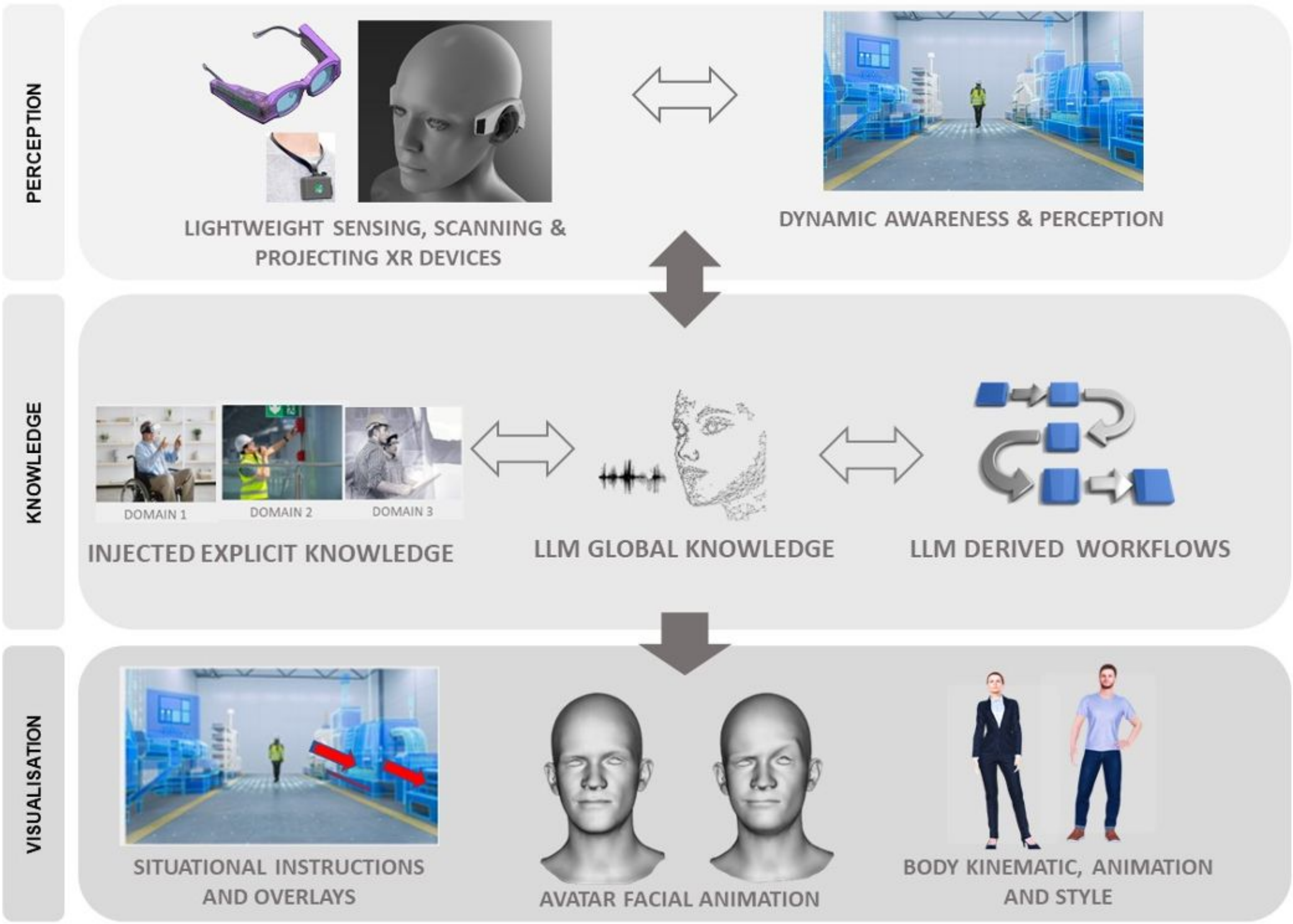
Grounding LLMs

- Using LLMs with information that is use-case specific, relevant and not available as part of the LLM's trained knowledge
- It is crucial for ensuring the quality, accuracy and relevance of the generated output
- While LLMs come with a vast amount of knowledge already, this knowledge is limited and not tailored to specific use-cases
- Grounding involves tasks such as defining the application, preparing relevant data, fine-tuning the model, adapting responses to the context

LUMINOUS Project - aim

Contribute towards the creation of the **next generation of Language Augmented XR systems and applications**, where natural language-based communication and **Large Language Models** redefine the future interaction with novel extended reality (XR) technology and enhances understanding of the users’ situation and environment **even in situations that are encountered for the first time**.

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE


PERCEPTION
LIGHTWEIGHT SENSING, SCANNING & PROJECTING XR DEVICES
DYNAMIC AWARENESS & PERCEPTION
KNOWLEDGE
DOMAIN 1
DOMAIN 2
DOMAIN 3
INJECTED EXPLICIT KNOWLEDGE
LLM GLOBAL KNOWLEDGE
LLM DERIVED WORKFLOWS
VISUALISATION
SITUATIONAL INSTRUCTIONS AND OVERLAYS
AVATAR FACIAL ANIMATION
BODY KINEMATIC, ANIMATION AND STYLE

LUMINOUS Project - MLLM goals

- Creation of innovative state-of-the-art multilingual **MLLMs able to generate situation-specific instructions to guide users through XR workflows**

- **Natural language-based bidirectional conversation via high-resolution avatars** including realistic face & mouth movements to help patients mimic speech sound production. Training material based on **LLM adapted to the environment without previous specific knowledge**.

LUMINOUS
LANGUAGE AUGMENTATION
FOR HUMANVERSE

LUMINOUS - MLLM tasks

- **Grounding** of MLLMs for environment **specific text generation**
- **Adaptation** of MLLMs models **to specific domains and environments**
- **Improving** MLLMs i**nstruction generation** by **including common sense reasoning**
- **Adapting** MLLMs via **situation-specific knowledge injection**

LUMINOUS
LANGUAGE AUGMENTATION
FOR HUMANVERSE

Thank you!