FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

NI-NLM – Lecture 10

Multilinguality & multimodality.

Zdeněk Kasner

28 Apr 2026

Multilinguality

reusing some NPFL140 materials by Tomasz Limisiewicz and Jindřich Libovický

Languages in the world

sources: National Geographic Ethnologue

Question

How many languages exist in the world? How many of them are spoken-only?

There are around **7,170 languages** today.

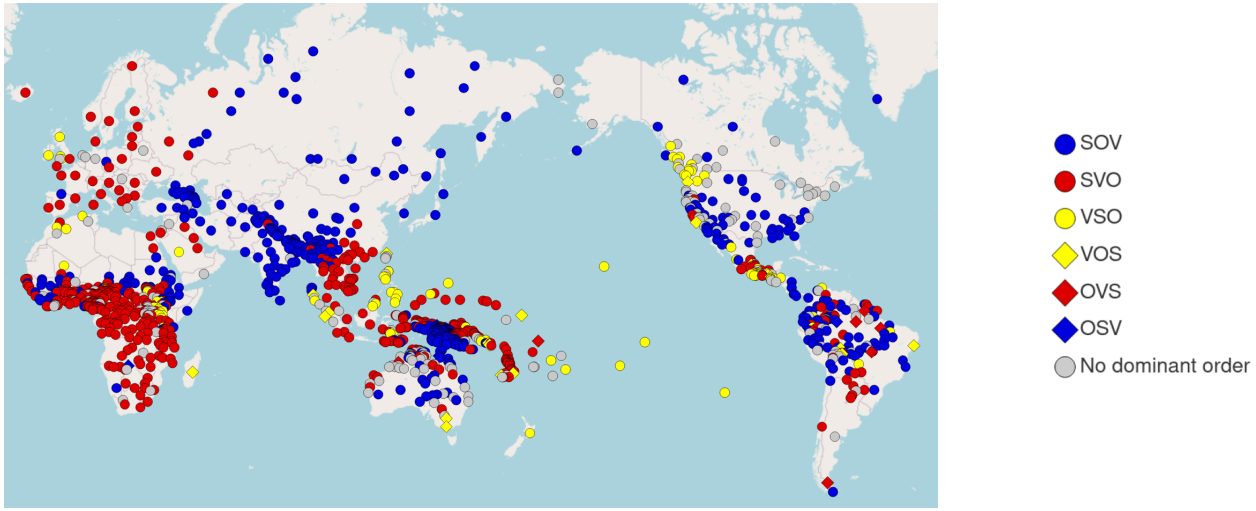
- Out of these, only **4,153 (58%)** have a writing system.
- 94% of people speak 6% of languages.
- Roughly 44% of all languages are now endangered (<1,000 users remaining).

Languages differ

source: WALS

Languages vary widely in their typological properties:

- **Word order**: SVO (English), SOV (Japanese, Hindi), VSO (Irish, Arabic), ...
- **Morphology**: isolating (Chinese) vs. agglutinative (Turkish) vs. fusional (Czech).
- **Writing systems**: Latin, Cyrillic, Arabic, Devanagari, Chinese characters, ...



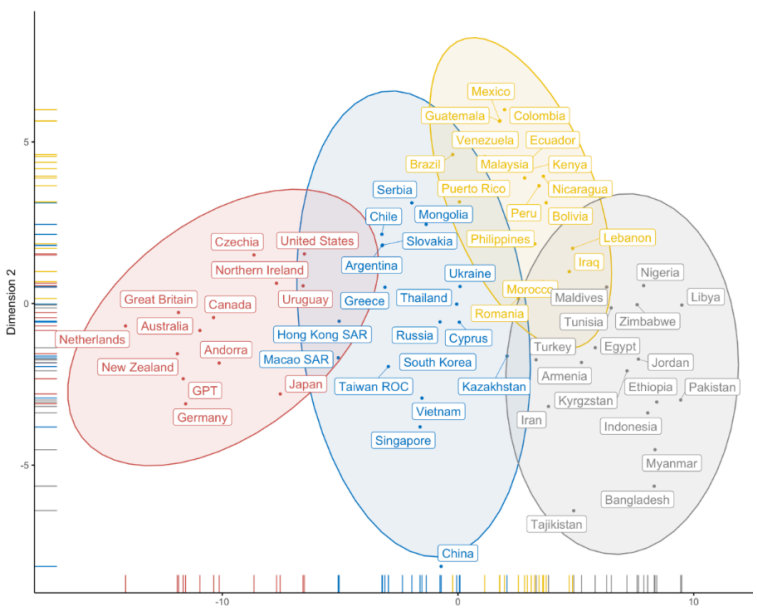
Why multilingual LLMs

source: https://osf.io/preprints/psyarxiv/5b26t_v1

Question

Why do we train multilingual LLMs?

- **Accessibility** to non-English speakers.
- Code-switching & **cross-lingual tasks**.
- **More efficient** than training a specific model for each language.
- Cross-lingual **transfer learning**.
- **Localization** as opposed to translation only.
- Instilling non-English-centric **cultural values**.



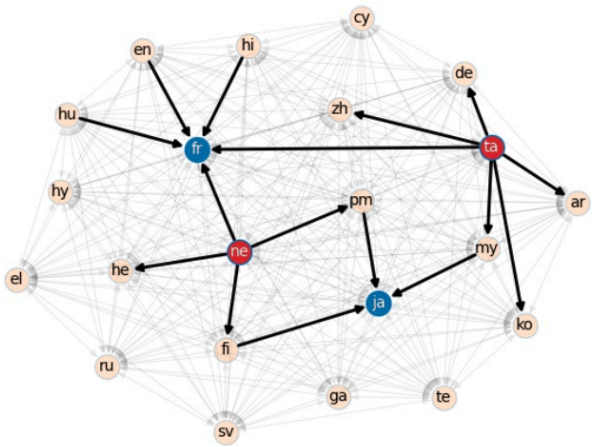
Cross-lingual transfer

source: Conneau et al. (2020)

Idea

Train a model on a **high-resource language** and apply it to a **low-resource** one.

- First explored with multilingual pretrained encoders (mBERT, XLM-R).
- **Zero-shot cross-lingual transfer**: the models can perform tasks such as named-entity recognition **without any target-language training data**.



source: Malkin et al. (2022)

→ Shared multilingual representations must have developed within the models.

Typological similarity helps

source: https://arxiv.org/pdf/1911.03310

Cross-lingual transfer works better between **similar languages**:

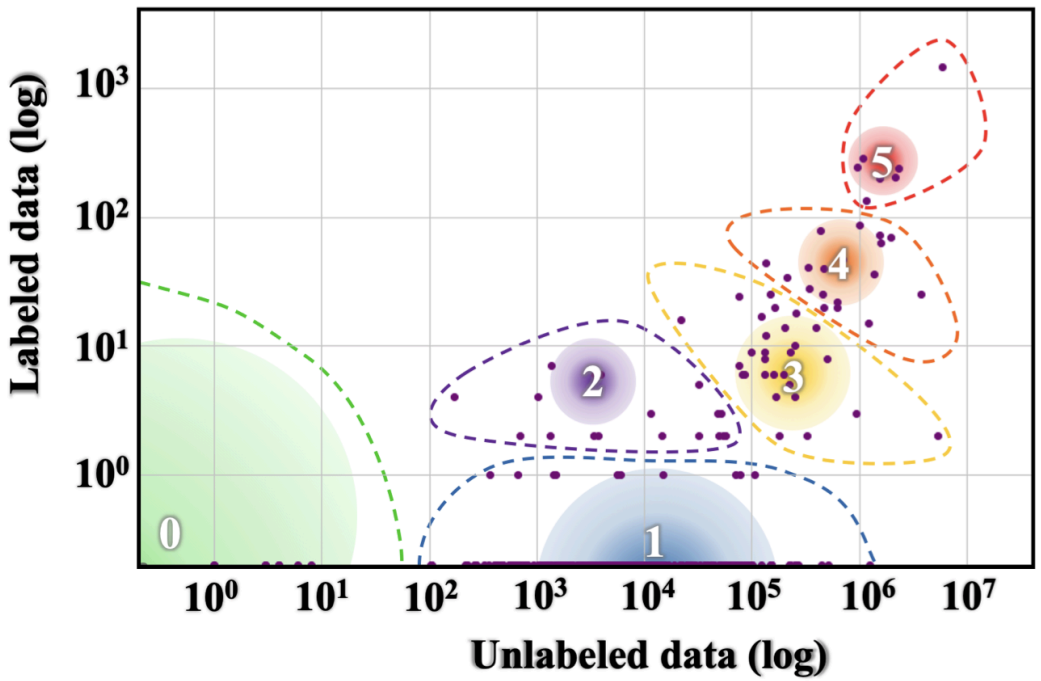
- Languages with **similar word order and morphology** share more representational structure.
- **Shared writing script** also helps the knowledge transfer, especially with respect to tokenization.

The transfer does not have to be zero-shot: any amount of data from the target language helps.

Data availability

source: https://arxiv.org/pdf/2004.09095

The amount of available text data varies enormously across languages:



→ Many languages have virtually no usable training data (both unlabeled & labeled).

Data availability

source: https://arxiv.org/pdf/2004.09095

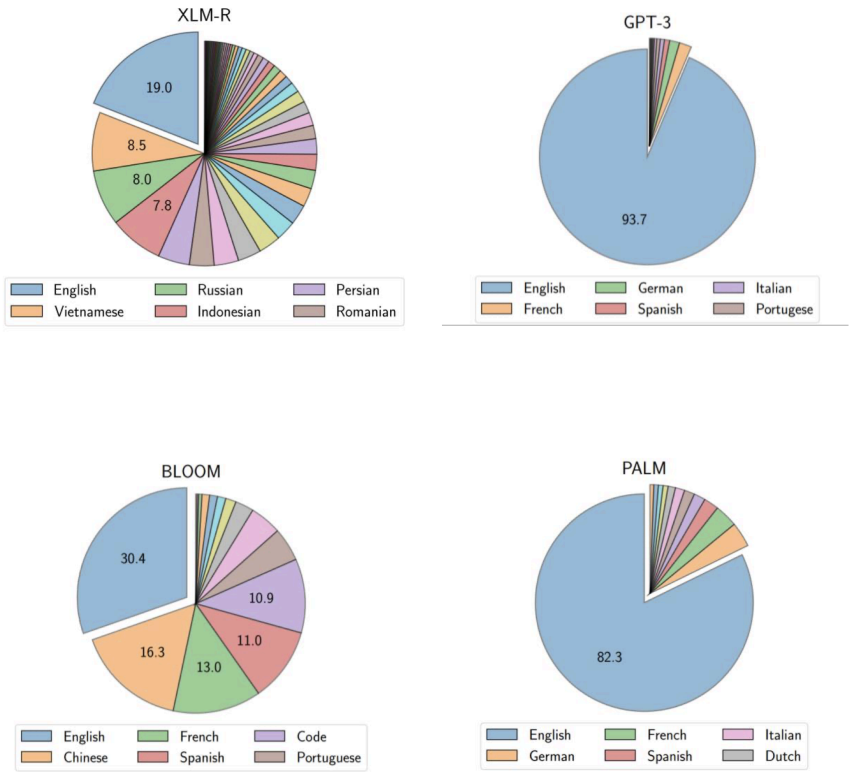
Group	Speakers	Examples of languages
5	2.5 B	English, Spanish, German, French, Arabic, Mandarin
4	1.6 B	Russian, Hungarian, Vietnamese, Czech, Polish, Persian, Hindi, ...
3	1.1 B	Indonesian, Ukrainian, Hebrew, Cebuano, Slovak, ...
2	300 M	Irish, Maltese, Lao, Zulu, Amharic, ...
1	1 B	Cherokee, Fijian, Greenlandic, Navajo, Macedonian, ...
0	1 B	Dhalo, Warlpiri, Popoloca, Wallisian, Bora, ...

Approx. 30% of world population

Training multilingual LLMs

source: NPFL140 - Lecture 8

- **Naive approach**: train on all data proportionally → the model mostly learns English.
- **Sampling**: oversample low-resource, undersample high-resource languages.
  - Still a trade-off: can hurt high-resource performance.
  - Low-resource data are also often of lower quality.



The more balanced the better?

source: BigScience (2023)

An effort to train a **large multilingual LLM** with careful data curation.

- 176B parameters, trained on **46 languages**.
- Special attention to balancing languages.
- One of the first fully open large-scale multilingual models.

→ Performed subpar compared to SotA models.

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

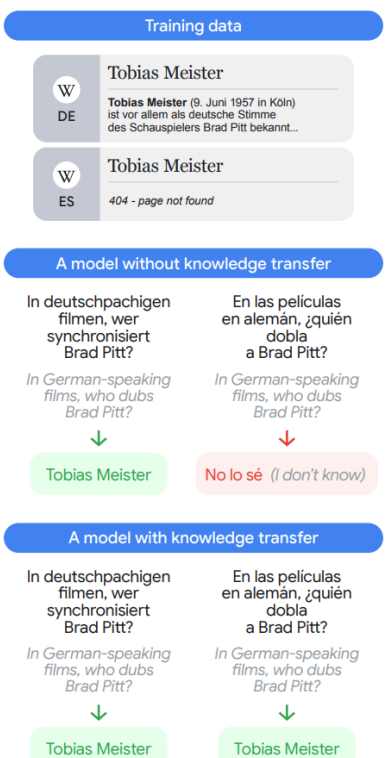
⚠️ **Curse of multilinguality**

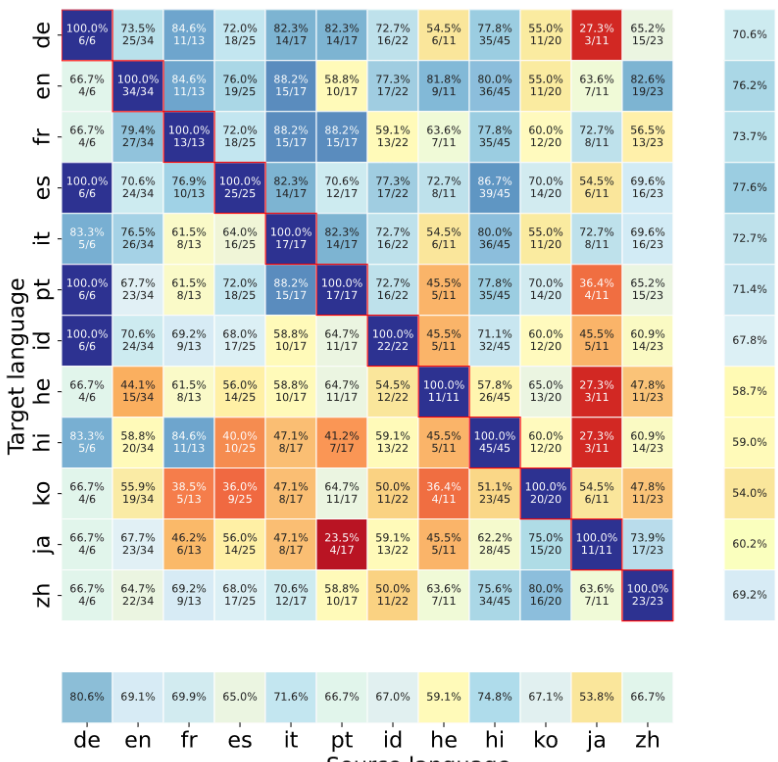
A model has fixed capacity. Adding more languages forces it to share that capacity, hurting per-language performance.

Knowledge might be language-specific

source: https://arxiv.org/abs/2502.21228

Facts from local Wikipedia pages are often **not accessible** in other languages:





Knowledge might also be shared

source: Anthropic (2025)

However, LLMs can develop **language-agnostic concepts**:

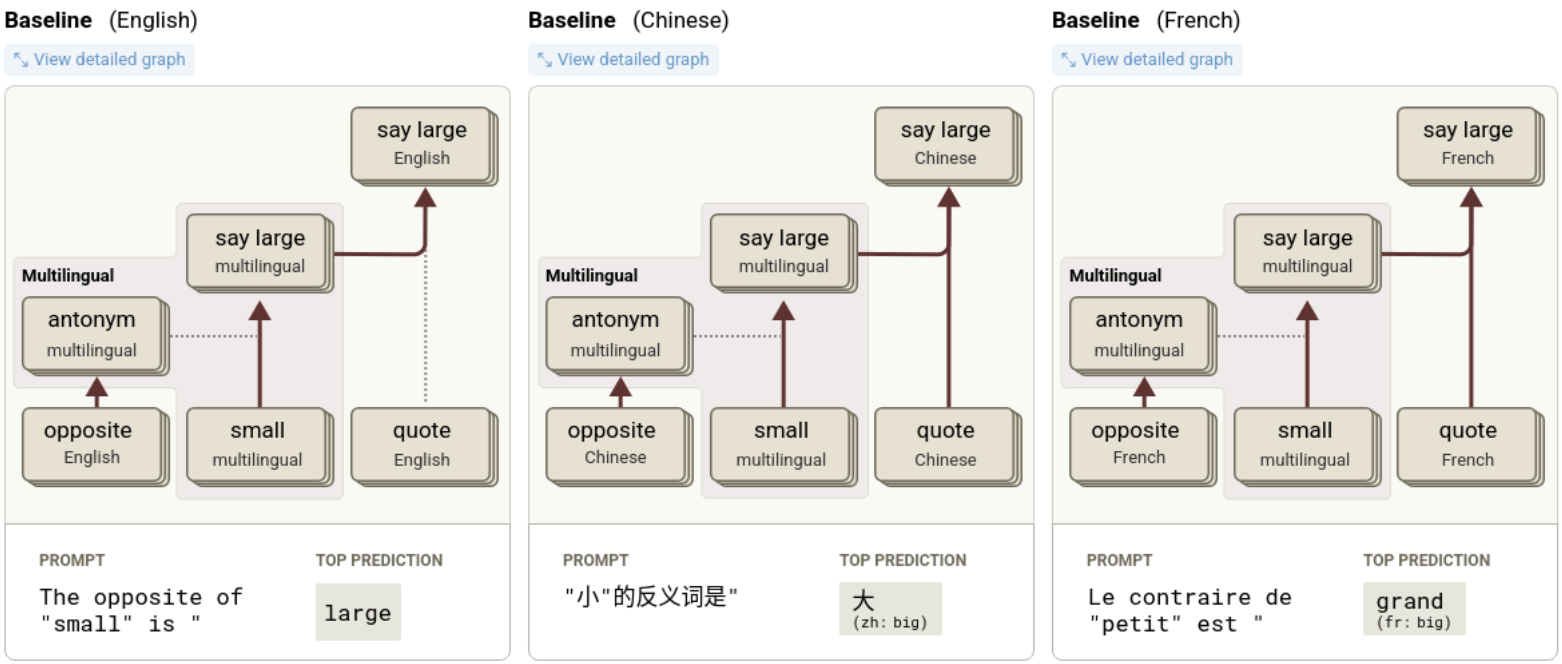


Figure 17: Simplified attribution graphs for translated versions of the same prompt, asking Haiku what the opposite of "small" is in different languages. Significant parts of the computation appear to be overlapping "multilingual" pathways. This is an interactive diagram, and you can hover over supernodes to see visualizations of the constituent features. Note that these are highly simplified, see "View detailed graph" above each to see un-simplified version.

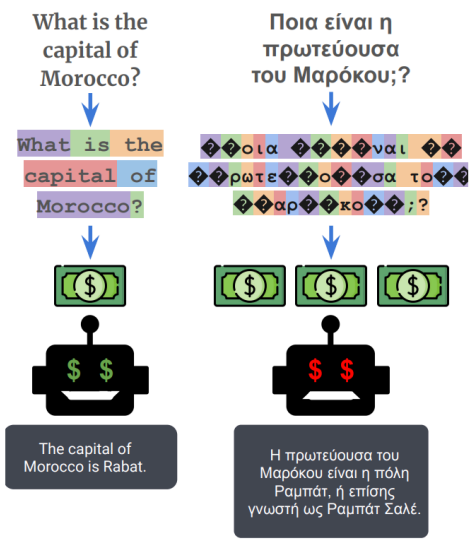
Evaluating multilingual models

Choosing the right model

Question

How would you pick the right model for a specific language?

- Check the model card if the authors **claim to support it** (example)
- Check if the language was included in the **training or fine-tuning data**.
- Check data for similar **languages**.
- Check **tokenization**.
- Check **benchmarks**.



source: Ahia et al. (2023)

UTF-8 encoding

source: NPFL140 - Lecture 7

One **grapheme** (visible character) can span **multiple UTF-8 bytes**:

U+0000 – U+007F (ASCII)
0xxxxxxx
1 byte · 128 code points

U+0800 – U+FFFF (CJK, Devanagari...)
1110xxxx 10xxxxxx 10xxxxxx
3 bytes · 63,488 code points

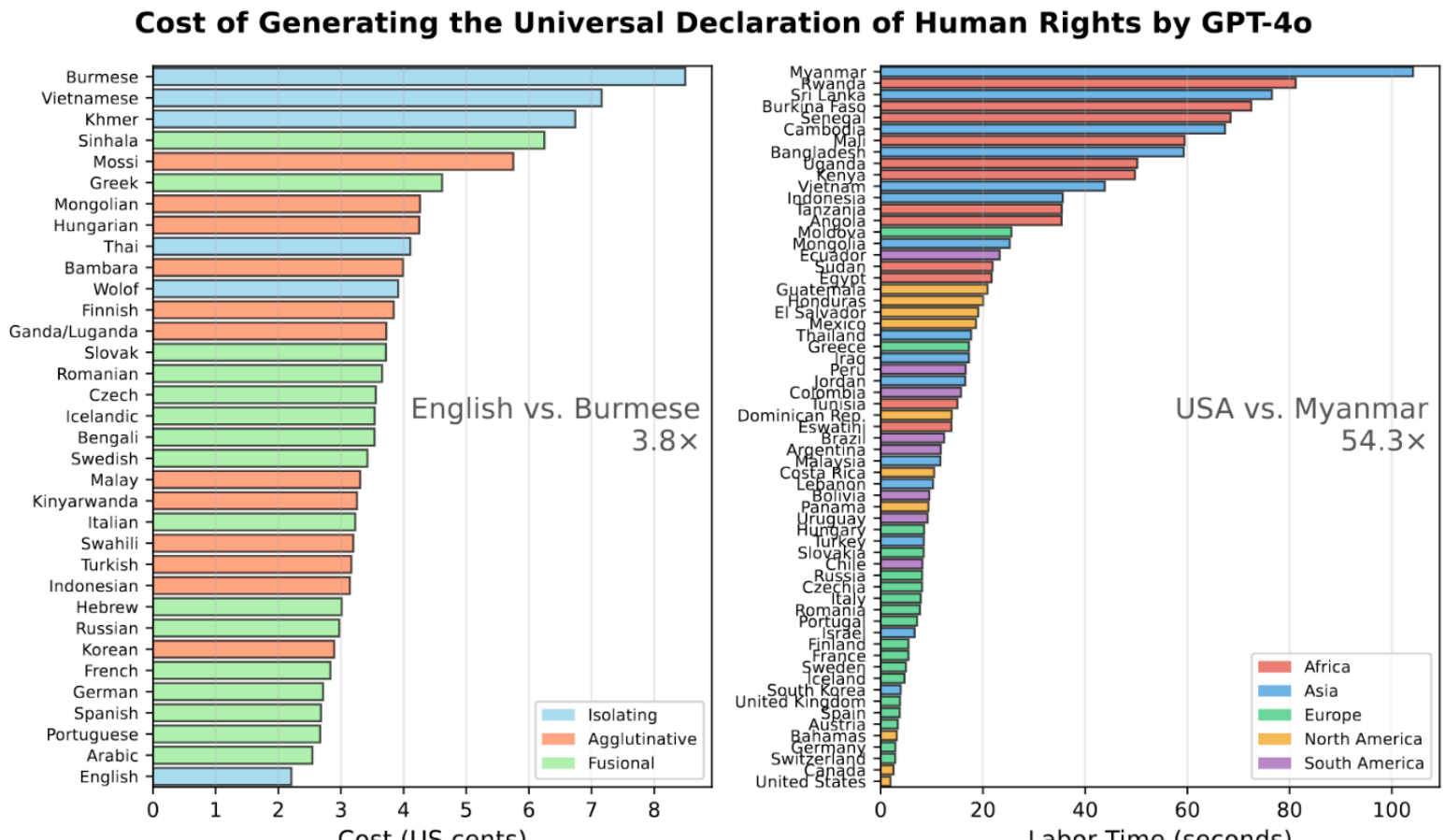
U+0080 – U+07FF (Latin, Arabic, Hebrew...)
110xxxxx 10xxxxxx
2 bytes · 1,920 code points

U+10000 – U+10FFFF (emoji, rare scripts)
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
4 bytes · 1,048,576 code points

	Tamil	Sinhala	Hindi
Grapheme	ன்	ශ්‍ර	वा
Unicode characters	ன + ◌	ශ + ◌	व + ◌ा
Unicode codepoints	U+0BA9, U+0BCD	U+0DC3, U+0DCA	U+0935, U+093E
UTF-8 bytes	e0 ae a9 e0 af 8d	e0 b7 83 e0 b7 8a	e0 a4 b5 e0 a4 be

Models are more expensive for some languages

source: NPFL140 - Lecture 8



လူတိုင်းသည် တူညီ
လွတ်လပ်သော ဂုဏ်သိက္ခာဖြင့်
လည်းကောင်း၊
တူညီလွတ်လပ်သော
အခွင့်အရေးများဖြင့်
လည်းကောင်း၊
မွေးဖွားလာသူများ ဖြစ်သည်။
ထိုသူတို့၌ ပိုင်းခြား
ဝေဖန်တတ်သော ဉာဏ်နှင့်
ကျင့်ဝတ် သိတတ်သော
စိတ်တို့ရှိကြ၍ ထိုသူတို့သည်
အချင်းချင်း မေတ္တာထား၍
ဆက်ဆံကျင့်သုံးသင့်၏။

😱

Where to get multilingual benchmarks?

Option #1: Translate the English benchmarks.

✅All benchmark items are **equivalent** across languages

→ allows quantifying that language X is better than language Y

⛔ Traces of “translationese”, may not be properly localized.

Option #2: Collect benchmarks directly in the specific languages.

✅ Better reflects actual language use, cultural values etc.

⛔ Hard to compare across languages.

⛔ Getting local expert annotators is difficult.

Localization is *not* translation

source: NPFL140 - Lecture 8

de: Er ist ein typischer SPD-Wähler.

cs: Je to typický volič SPD.

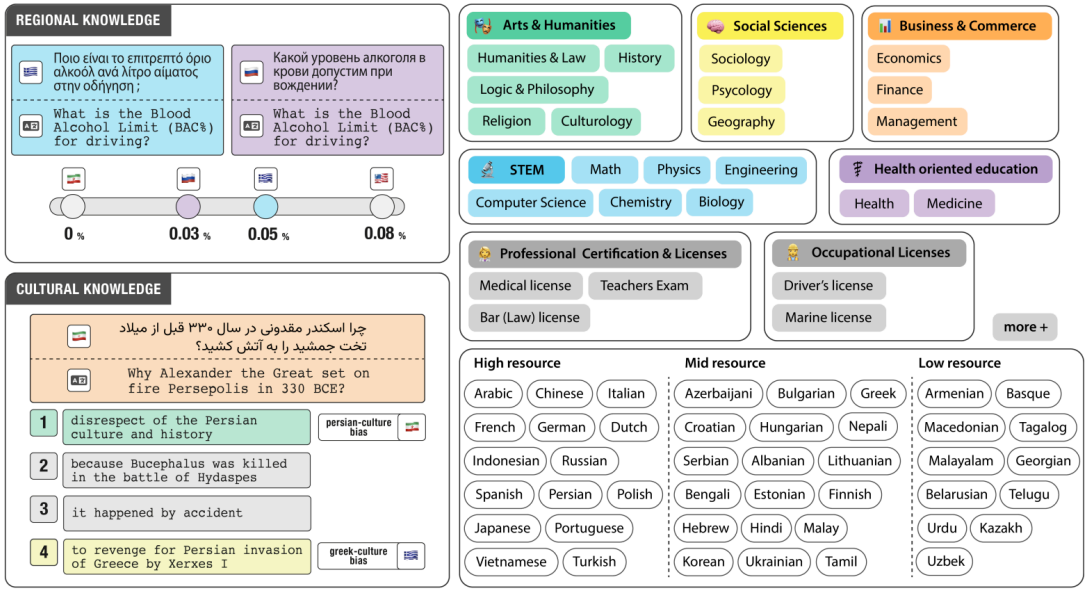
en: He is a typical voter of SPD.

INCLUDE vs. MMLU

source: https://arxiv.org/abs/2411.19799

INCLUDE: Authentic questions from local education systems.

- 55 countries, 44 languages, 15 scripts, 200k QA pairs
- Still, major model releases prefer M-MMLU
  - Heavily US-centric
  - Multilingual versions combine human and machine translation

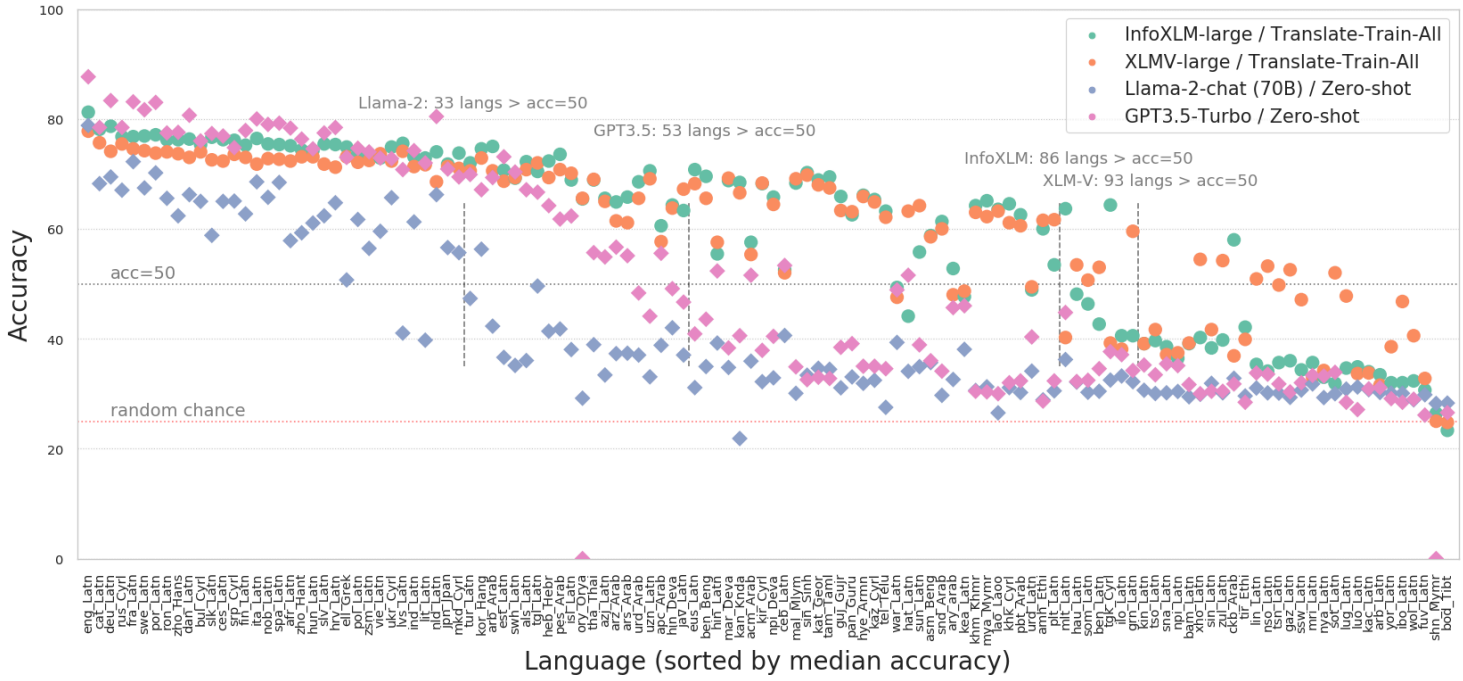


Belebele

source: Bandarkar et al. (2023)

Belebele: Another multiple-choice QA benchmark, spanning **122 languages**.

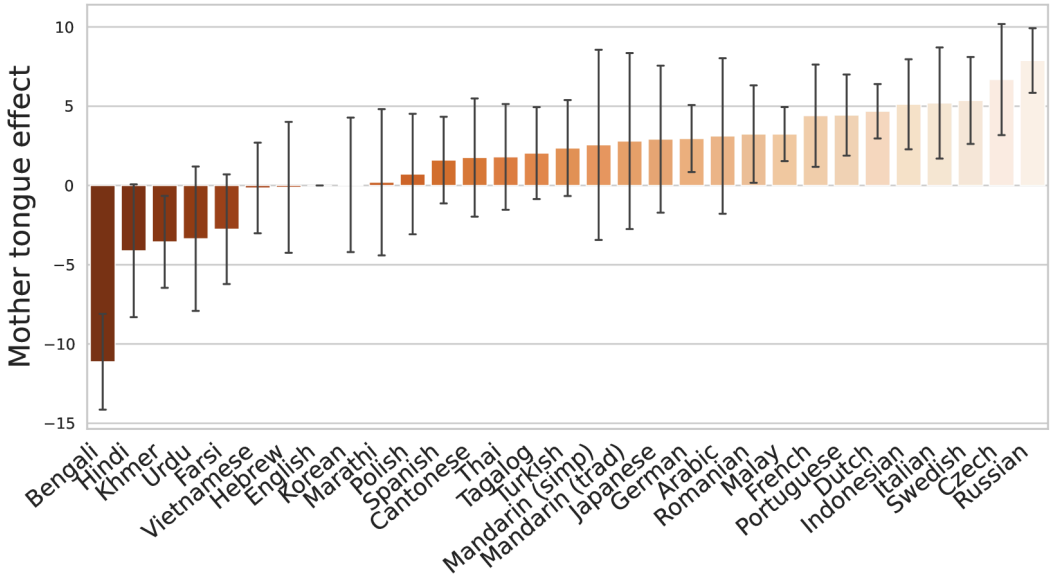
Questions translated from English by native speakers → parallel data.

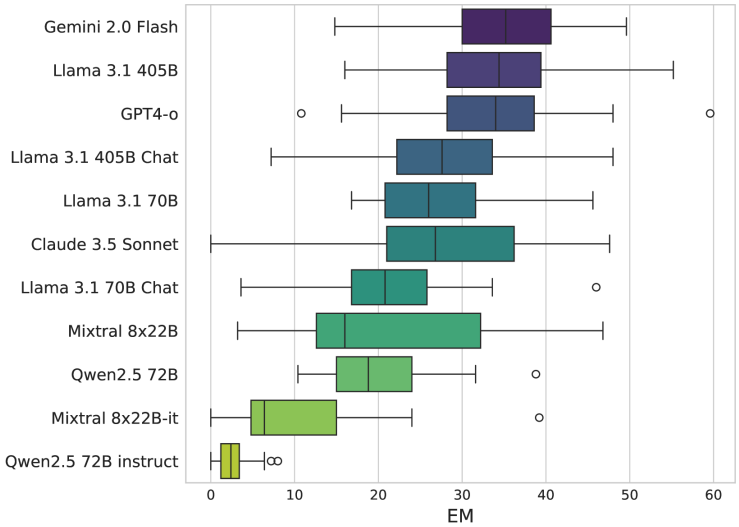


MultiLoKo

source: https://arxiv.org/abs/2504.10356v2

MultiLoKo: A benchmark for multilingual local knowledge: facts that are specific to certain languages and cultures.





Machine translation

LLMs for machine translation

source: https://aclanthology.org/2025.wmt-1.22/

LLMs are increasingly competitive with dedicated machine translation systems.

WMT24:

English→Czech

Rank	System	Human	AutoRank
1-2	HUMAN-A	92.9	
2-2	Unbabel-Tower70B	91.6	1.0
2-3	Claude-3.5 §	91.2	2.1
4-5	ONLINE-W	89.0	2.8
4-6	CUNI-MH	88.4	2.1
6-6	Gemini-1.5-Pro	88.2	2.6
6-8	GPT-4 §	87.7	2.6
8-8	CommandR-plus §	86.9	2.9
8-9	IOL-Research	86.5	2.8
10-11	SCIR-MT	85.4	3.2
10-11	CUNI-DocTransformer	84.3	4.4
12-12	Aya23	84.2	4.3
13-13	CUNI-GA	82.1	2.3
14-14	IKUN	81.7	3.9
15-15	Llama3-70B §	77.4	4.1
16-16	IKUN-C	75.4	4.7

§ means Czech is officially not supported

Claude 3.5 was the overall best system

LLMs Finetuned for MT

Closed general commercial LLMs

Systems based on 2023 winning system

Open-weight general LLMs

source: NPFL140

WMT25:

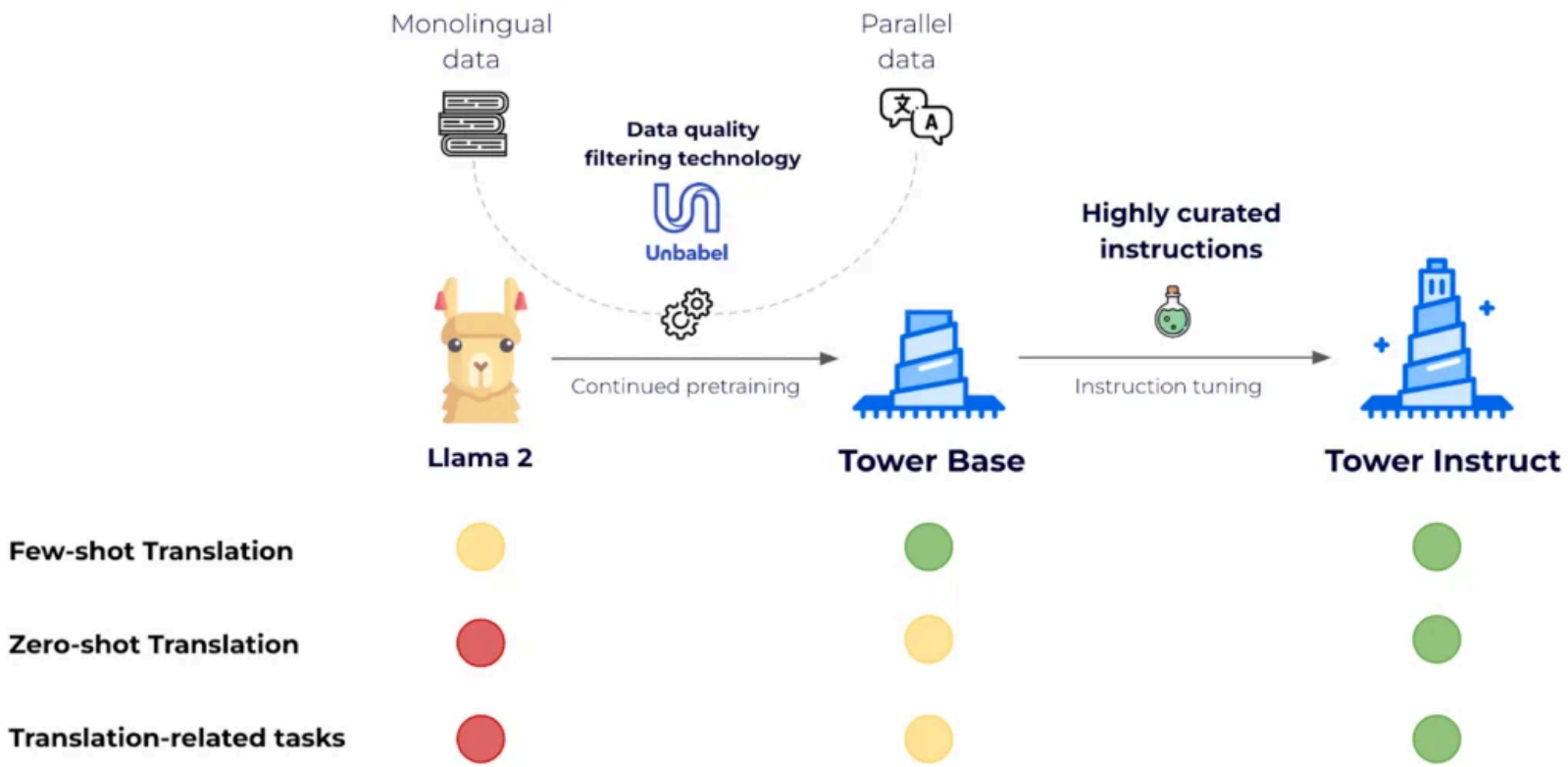
English→Czech

Rank	System	Human
1-1	Gemini-2.5-Pro	88.7
2-2	Shy-hunyuan-MT	87.1
3-4	DeepSeek-V3?	85.1
3-4	**Human**	84.5
5-6	CommandA-WMT	82.6
5-6	Wenyiil	82.4
7-9	GPT-4.1	80.8
7-9	Mistral-Medium?	80.4
7-10	Claude-4?	79.6
9-11	UvA-MT	78.6
10-14	Algharb	76.7
11-14	CommandA	76.4
11-15	Yolu	75.6
11-15	Gemma-3-27B	75.6
13-15	GemTrans	73.2
16-16	CUNI-MH-v2	71.0
17-18	SRPOL	67.5
17-19	Laniqo	66.1
18-19	TowerPlus-9B[M]	65.8
20-20	SalamandraTA	60.3
21-44	23 systems not human-evaluated	

Finetuned LLMs for MT

source: https://arxiv.org/pdf/2402.17733

Smaller LLMs can be competitive with larger models when finetuned for MT.



Multimodal models

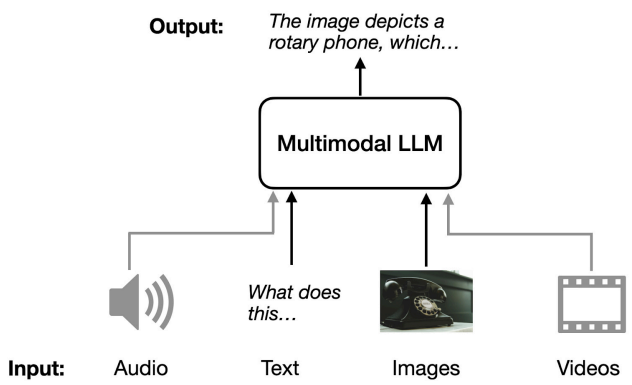
Beyond text

Modern LLMs are being extended to handle **images, audio, and video**.

> **Question**
>
> How would you represent non-text modalities so that a language model can process them?

Two main ingredients:

1. A pretrained **modality encoder** (e.g. vision encoder for images).
2. A **projection mechanism** to align representations with the LLM.



source: Understanding Multimodal LLMs

Vision Transformer

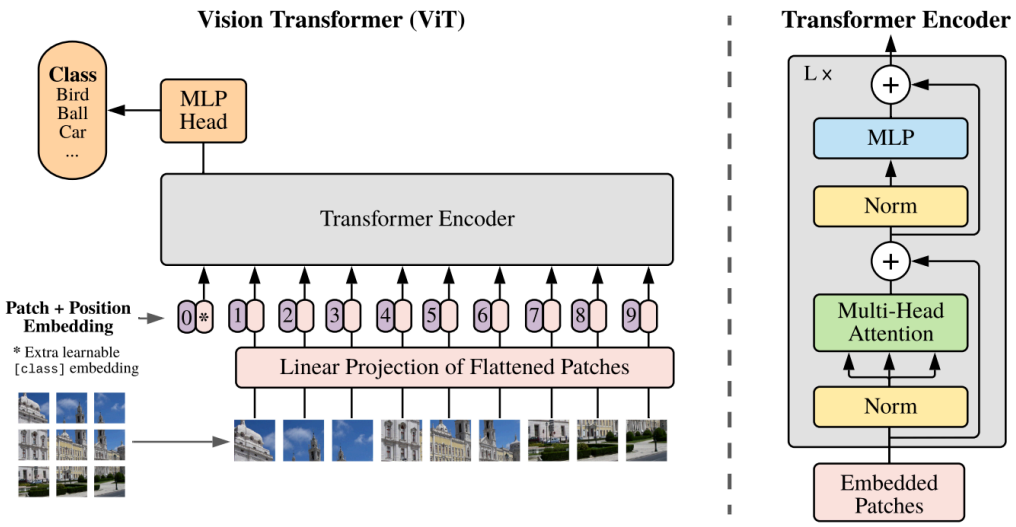
source: Dosovitskiy et al. (2021)

Idea

We can **encode an image similarly to text** if we split it into patches = “tokens”.

Vision Transformer (ViT):

1. Split the image into fixed-size patches (e.g. 16x16 pixels).
2. Project the patch through a linear layer → “embedding”.
3. Process with a Transformer encoder same as we would for the text.



CLIP: contrastive language-image pretraining

source: Radford et al. (2021)

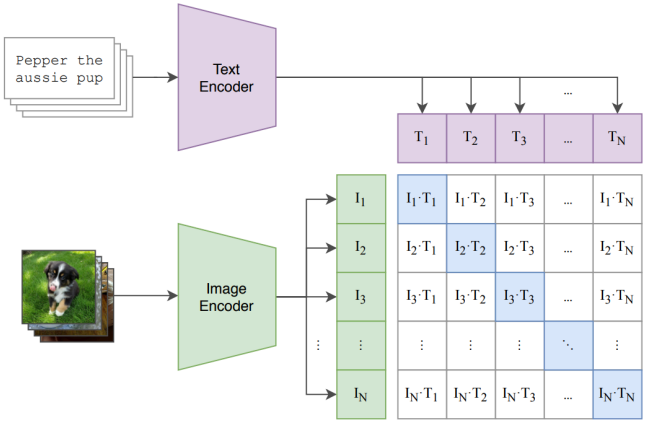
With ViT, the embedding space for the text and for the images can be vastly different.

Idea

Train two encoders (image + text) so that matching pairs end up close in a shared embedding space.

CLIP: trained on 400M image-text pairs.

- Uses **contrastive loss**: maximizing similarity of matching pairs + maximizing dissimilarity of non-matching pairs.
- Enables zero-shot image classification.

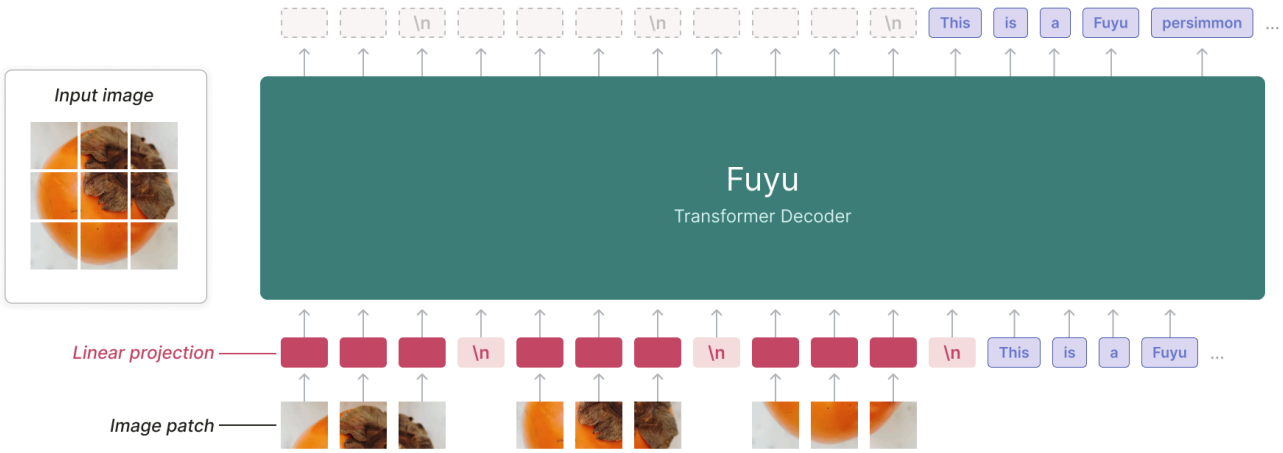


Vision-language models (VLMs)

Option A: unified embedding decoder

The simplest VLM architecture, used by LLaVA, Molmo, Qwen2-VL, Pixtral, …

1. Encode the image into a sequence of **visual tokens** (using ViT + projector).
2. **Concatenate** visual tokens with text tokens.
3. Feed everything through a standard **decoder-only Transformer**.



Option B: cross-modality attention

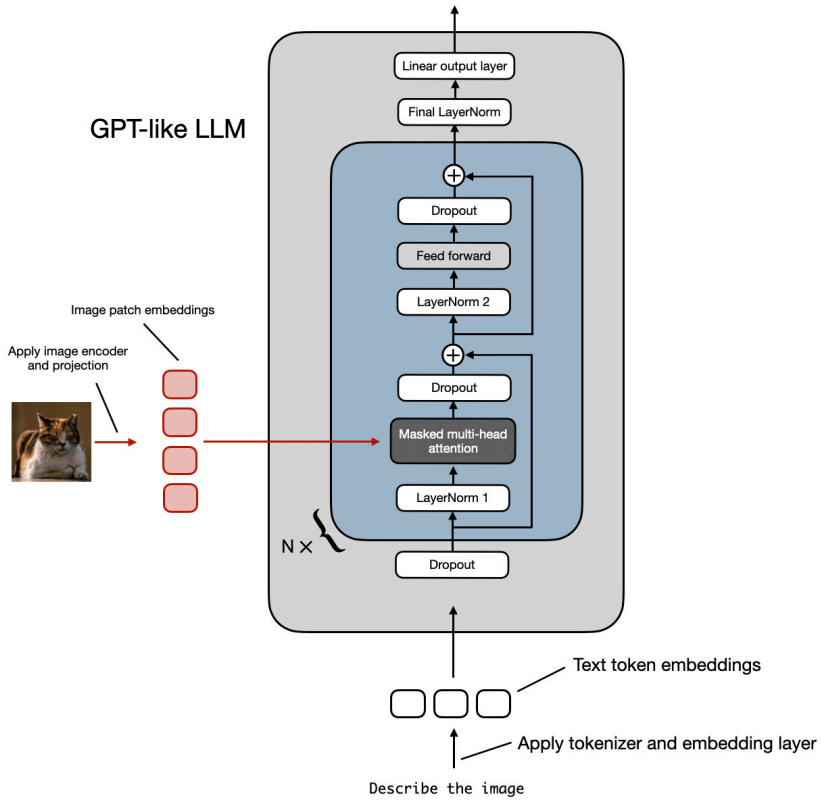
source: Understanding Multimodal LLMs

Visual features are injected via **cross-attention layer**:

1. The LLM has additional cross-attention layers that attend to visual features.
2. The LLM's original self-attention weights can remain **frozen**.
3. Visual features are processed separately and "read" by the LLM on demand.

Used by: Flamingo, Llama 3.2 Vision, …

Method B: Cross-Modality Attention Architecture



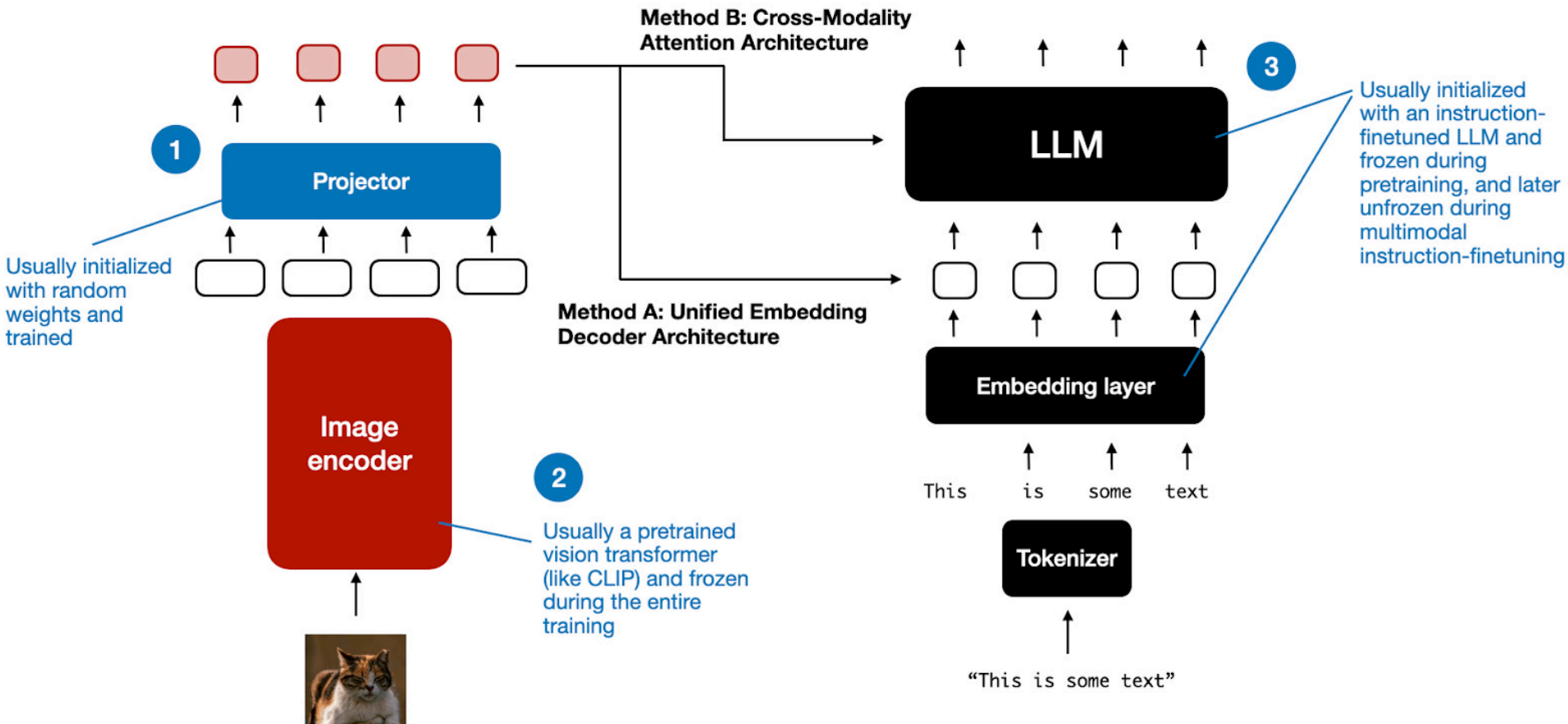
Training a vision-language model

VLM training typically has **two stages**:

1. **Alignment pretraining**
   - Freeze both image encoder and LLM.
   - Only train the **projector** (linear layer or MLP).
   - Goal: align visual representations with the LLM's embedding space.
2. **Visual instruction tuning**
   - Unfreeze the LLM (and optionally the image encoder).
   - Train on visual instruction-following data.
   - The model learns to answer questions about images, describe scenes, etc.

Training a vision-language model

source: Understanding Multimodal LLMs



Vision-language tasks

Visual QA

Answer questions about images

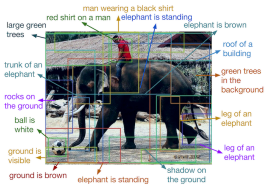
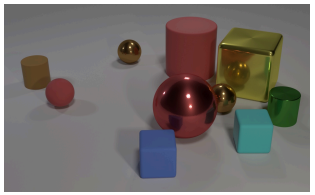


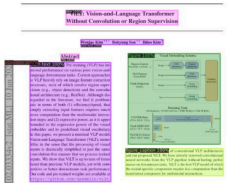
Image captioning

Generate text descriptions



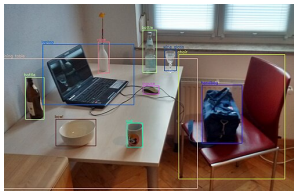
Visual reasoning

Multi-step inference over images



OCR / document parsing

Read text from images



Grounding

Localize objects by description

Image generation

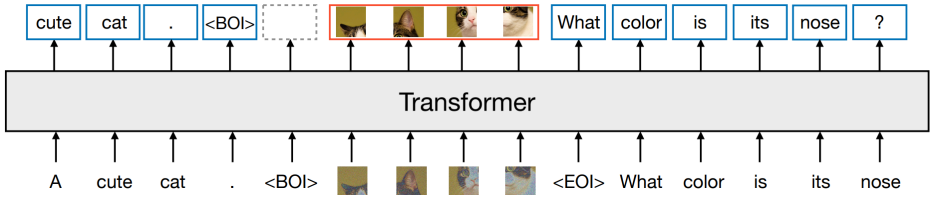
Generate images from text

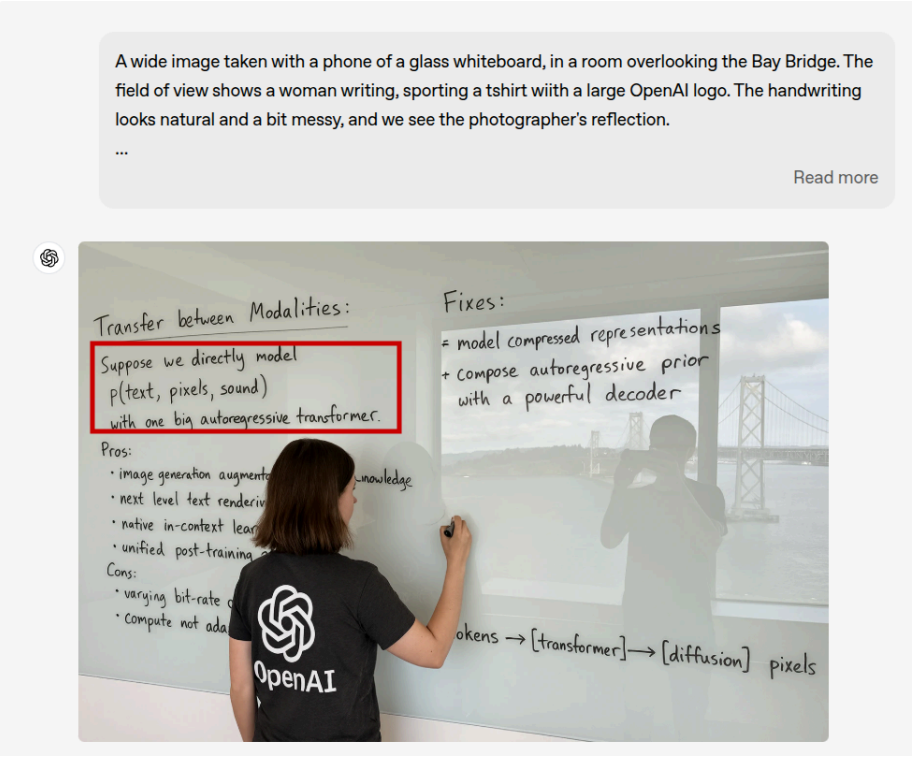
Image generation

source: https://arxiv.org/abs/2408.11039

Early commercial LLMs called an **external image generation model** (e.g., DALL·E) when they recognized an image generation request.

Some current LLMs are **trained jointly** for text generation & diffusion → better interplay between modalities.





source: OpenAI blog

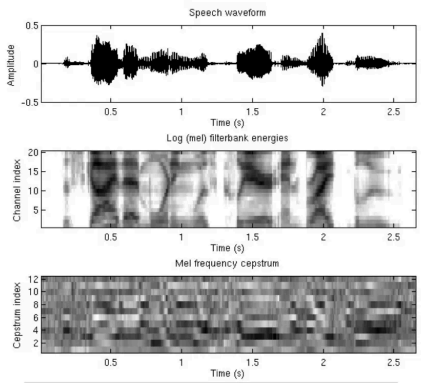
Speech & video

Speech representations

How do we represent audio for neural models?

Traditional: MFCCs

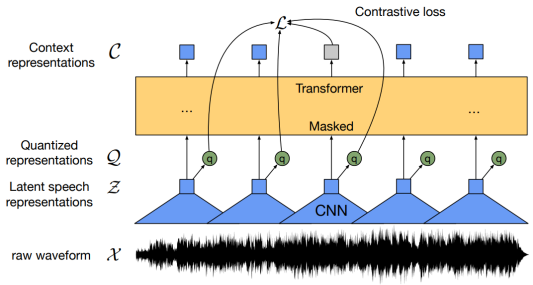
- Extract spectral features from short audio frames.
- Hand-engineered, lossy.



source: Medium.com

Current: Raw waveforms

- Self-supervised training on unlabeled audio to build representations directly from the raw waveform.



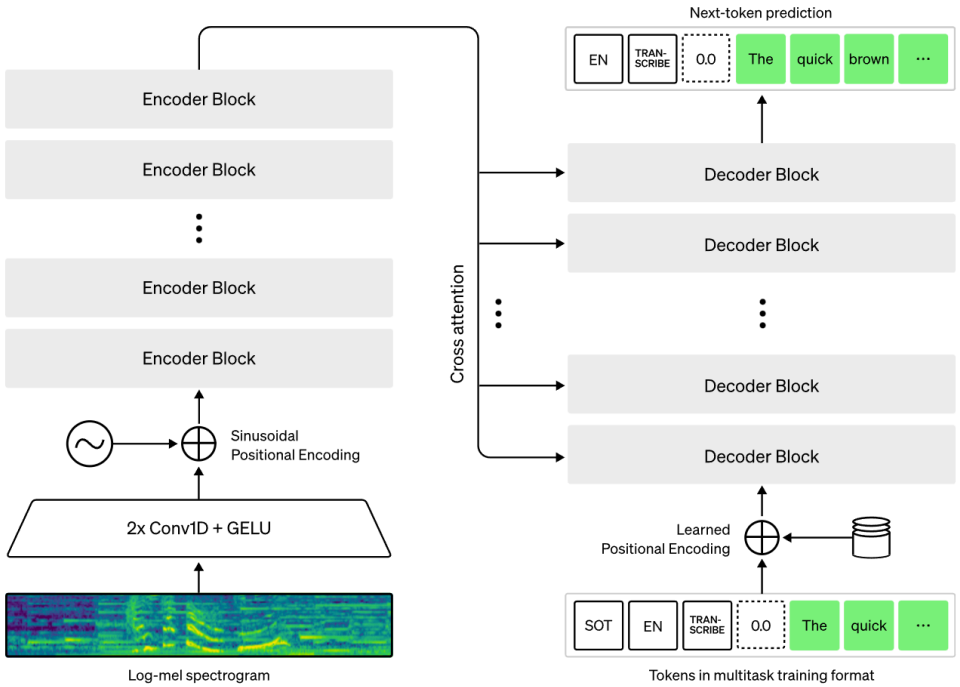
source: Baevski et al. (2020)

Whisper

source: OpenAI blog

Whisper: An encoder-decoder Transformer for automatic speech recognition.

- Trained on 680,000 hours of labeled audio from the web.
- Multi-task: ASR, translation, language identification, timestamps.
- Works well across many languages out of the box.
- **ÚFAL extensions**: WhisperStreaming and SimulStreaming → real-time processing, simultaneous translation.



Video understanding and generation models

source: OpenAI blog

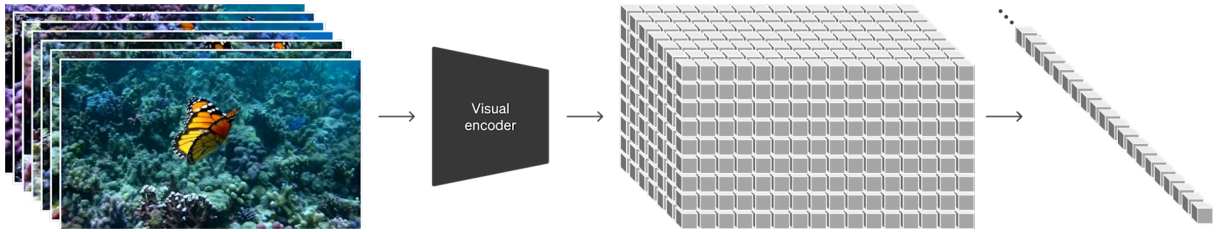
Video adds the **temporal dimension**: a sequence of image frames.

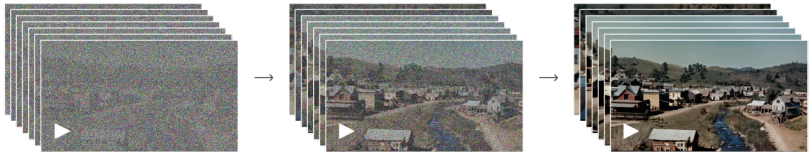
Input: spacetime latent patches

1. We embed the video into a lower-dimensional latent space (→ for efficiency).
2. We flatten the representation into a series of patches (→ similar to images, but in 3D).

Output: video diffusion

Given noisy video patches + text prompt, predict the original clean patches.

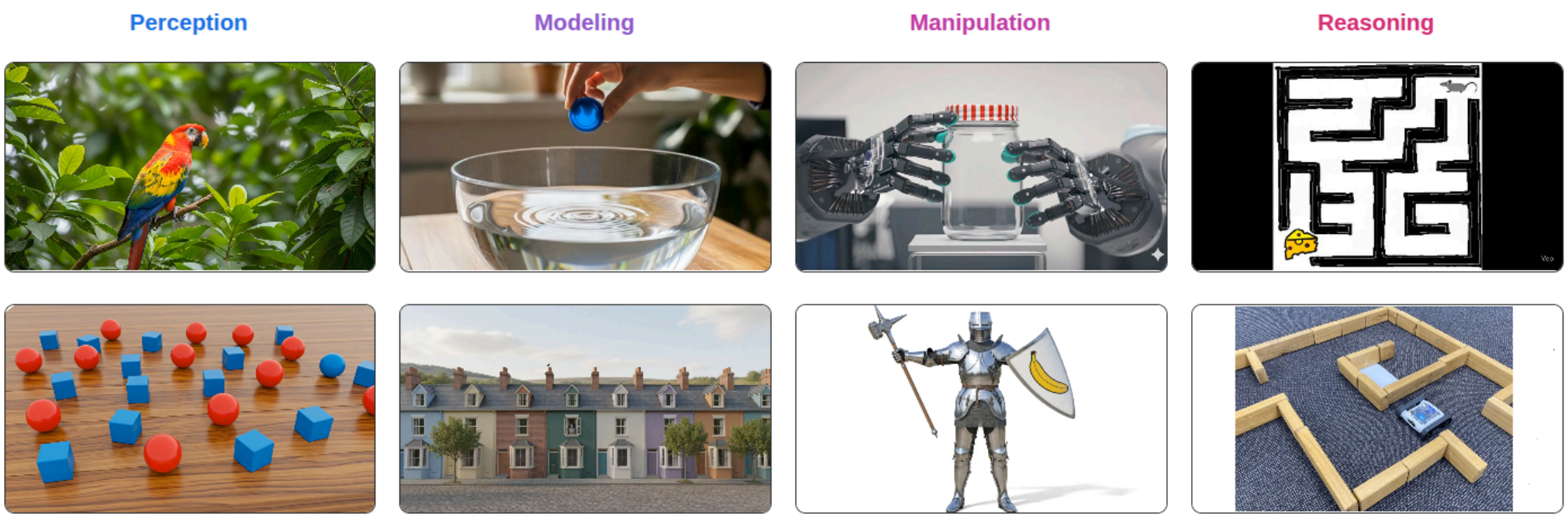




Emerging video capabilities

source: Google DeepMind (2025)

Recent video generation models (e.g. Veo 3) show surprising **emergent zero-shot capabilities** for image-related tasks (edge detection, style transfer, simulations, ...)



Summary

Multilinguality

- 7,000 languages, but most NLP research focuses on a handful.
- Multilingual models enable cross-lingual transfer but face the curse of multilinguality.
- LLMs are increasingly competitive for machine translation.

Multimodality

- Vision Transformers and CLIP bridge the gap between images and text.
- VLMs combine a vision encoder with an LLM via a projector (or cross-attention).
- The same pattern extends to speech and video understanding.

Links and resources

- Conneau et al. (2020): Unsupervised cross-lingual representation learning at scale
- Malkin et al. (2022): Studying multilingual language models through transfer
- Ahia et al. (2023): Do all languages cost the same? Tokenization in multilingual models
- Dosovitskiy et al. (2021): Vision Transformer
- Radford et al. (2021): CLIP
- Raschka: Understanding multimodal LLMs
- Baevski et al. (2020): wav2vec 2.0
- Whisper overview