Large World Models: Takeaways & Review

GPUDay 2024, Budapest, Hungary

Natabara Máté Gyöngyössy

Eötvös Loránd University, Department of Artificial Intelligence

2024

How to model the world?

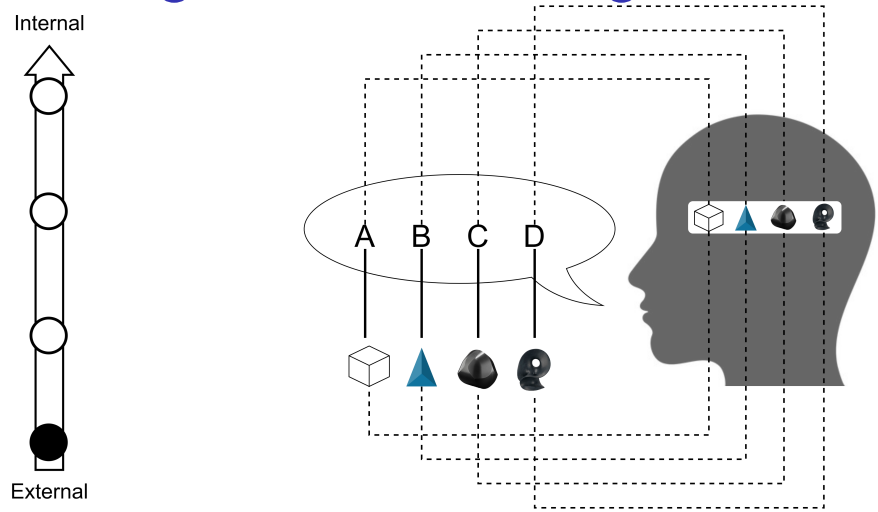
Language

Natural language. . .

- is a symbolic sequence.
- is context-aware.
- could be ambiguous.
- follows a given structure and conveys meaning.

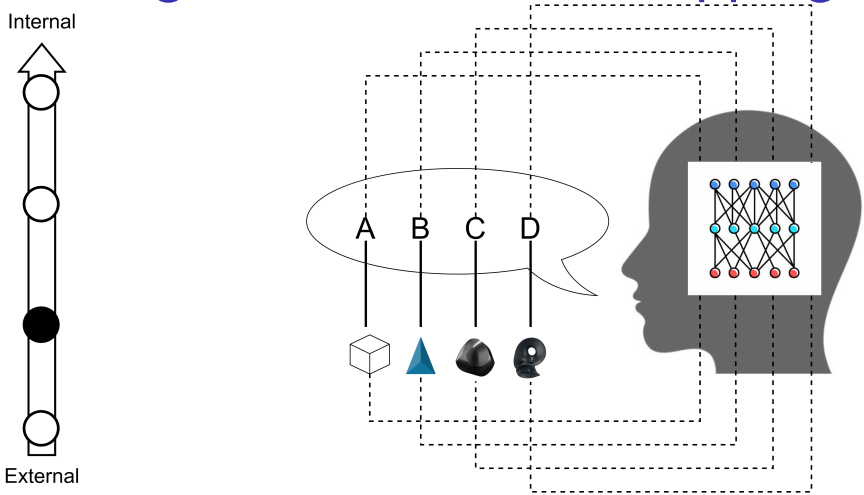
Language, but Grounded

Meanings as Mental Images



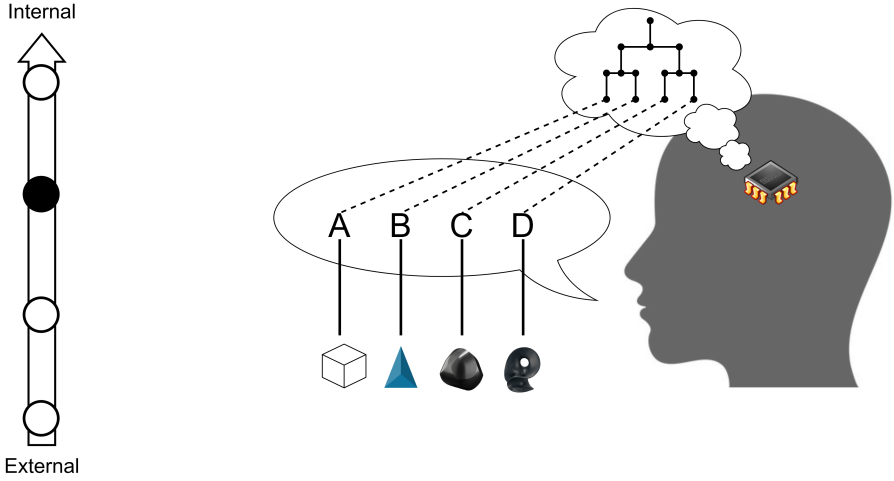
Perception of the language creates a mental image similar to percieving or recalling the object (Deacon 1997).

Meanings as Associative Mappings



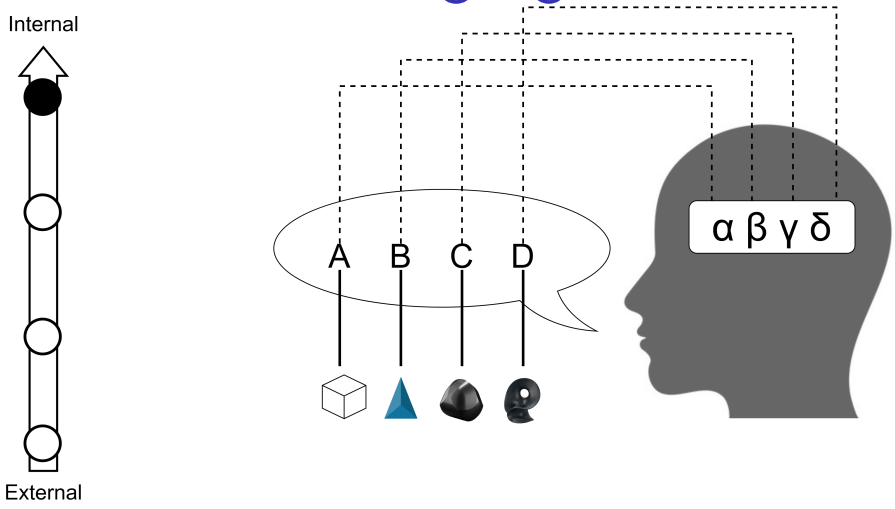
Language is a learned by internalizing distributed probabilistic connections of word-word and word-object structures (Deacon 1997; Skinner 1957).

Innate Universal Grammar



By learning a language we learn the language's connection to the innate Universal Grammar over which we perform inference. (Deacon 1997; Cook and Newson 2014).

Innate Mental Language



Learning a language is a translation task from and to an inner mental language (Deacon 1997; Pinker 2003).

The AI Perspective

- Associative mappings are the closest to how LLMs are trained.
- Transformer circuits try to discover the "mental images" of trained models.
- Ongoing research is eager to incorporate exact "as Universal as possible" grammars into LLMs.
- Translation to and from an LLM's "mental" language is the hottest solution for modality extensions.

LLMs: The Backbone

Characteristics of LLMs

- Context-awareness: Attention mechanism, or similar techniques. Few-shot learning possible.
- Self-supervised learning: Using vast amounts of "unlabeled" data.
- Autoregressive generation: Modeling continuation probabilities over a sequence of symbols (tokens).
- Large-scale: $1 - 2000$B parameters.

The Role of Dynamic Selection

Attention learns a **dynamic** (based on $\mathbf{x}^*$) selection **mechanism** that is used to process each element of the input sequence $\mathbf{x}$. The dynamic selection works by calculating a vector dim. scaled dot-product relevance score between the input and the query after learnable linear projections ($\mathcal{K}$, $\mathcal{Q}$, $\mathcal{V}$) (Vaswani et al. 2017).

$$s(\mathbf{x}_i, \mathbf{x}^*) = \frac{\mathcal{K}(\mathbf{x}_i) \cdot \mathcal{Q}(\mathbf{x}^*)}{\sqrt{d}}$$

$$\text{softmax}(\langle s(\mathbf{x}_1, \mathbf{x}^*), \ldots, s(\mathbf{x}_n, \mathbf{x}^*) \rangle) \cdot \mathcal{V}(\langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle)$$

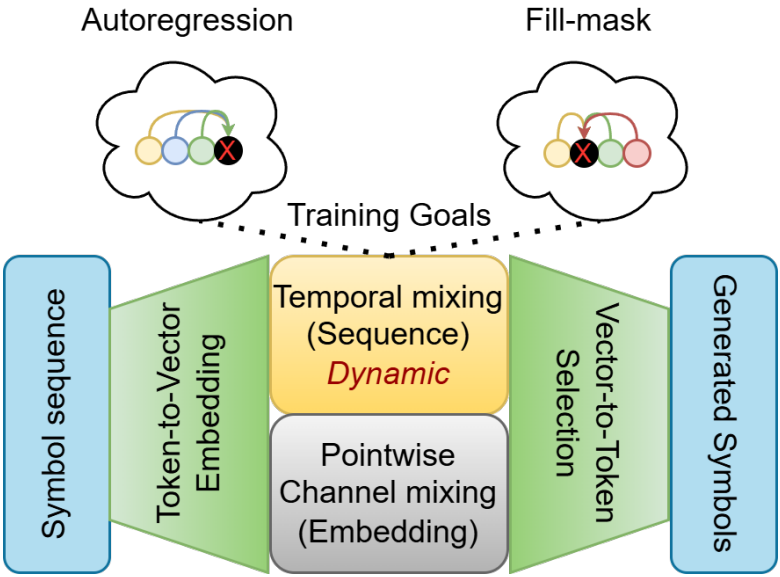
Self-supervision

The distributional semantic approach (Lenci and Sahlgren 2023) assumes:

- Words that occur in similar contexts are semantically similar.
- The meaning of a word could be inferred from the context it appears in.

This context could be bidirectional (fill-mask style) or causal (autoregressive, predict the next style).

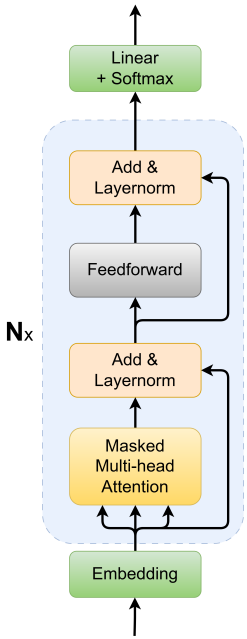
The Language Recipe



Specifics of a GPT-like Model

- Using Causal Multi-Head Attention to mix tokens.
- Feed-forward layers used to mix channels.
- Subword tokenization with Byte-Pair Encoding.
- Autoregressive with k-th order Markov assumption.
- Radford et al. (2019)

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | x_{i-k}, ..., x_{i-1})$$



Alternatives

Selective (input-dependent **B**, **C** and Δ) State-Space Models (Gu and Dao 2023)

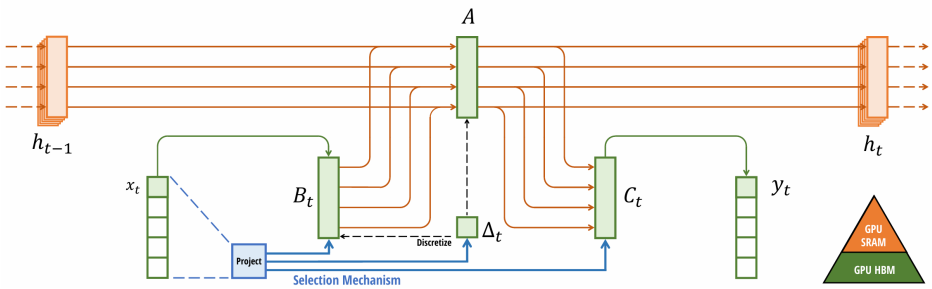
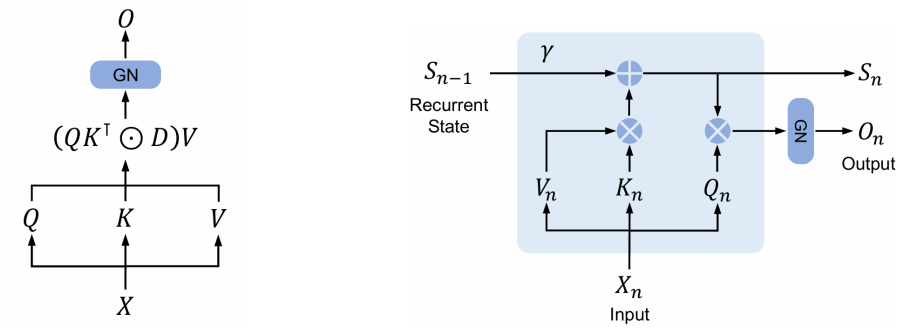


Figure 1: S4 block with SRAM state caching.

Alternatives

Retention with preset decay to construct dual-form (parallel, serial) networks (Sun et al. 2023).



(a) Parallel representation.

(b) Recurrent representation.

Figure 2: Retention for training (left) and inference (right).

How DL Research Benefits from LLMs?

How to Handle a Giant?

Figure 3: From Jones, Goldstone, and Python (1979)

Preference Alignment

Alternatives are not learned due to:

- Data Sparsity (training on all 100K words long sequences is impossible).
- Teacher Forcing (the model is not incentivized to explore alternatives).

But we can do it in a second phase using sequence-level preference training based on a small dataset of human preference data. This instruction fine-tuning produced ChatGPT as well.

Instruction Fine-tuning

PPO-based RL with (reward, reference and policy LLM models) was the first breakthrough in human preference alignment (Ouyang et al. 2022).

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[r_\phi(x, y)\right] - \beta\mathbb{D}_{\mathsf{KL}} \left[\pi_\theta(y \mid x) \parallel \pi_{\mathsf{ref}}(y \mid x)\right]$$

Later Direct Preference Optimization (DPO) was introduced that uses maximum likelihood-based training without a reward model (Rafailov et al. 2023).

$$\max_{\pi_\theta} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathsf{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\mathsf{ref}}(y_l|x)}\right)\right]$$

Instruction Fine-tuning

Lately, even the reference model could be omitted by using Odds Ratio Preference Optimization (ORPO) (Hong, Lee, and Thorne 2024).

$$\text{odds}_\theta(y \mid x) = \frac{1-\pi_\theta(y|x)}{\pi_\theta(y|x)}$$

$$\max_{\pi_\theta} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma \left(\log \left(\frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)} \right) \right)$$

Instruction Fine-tuning

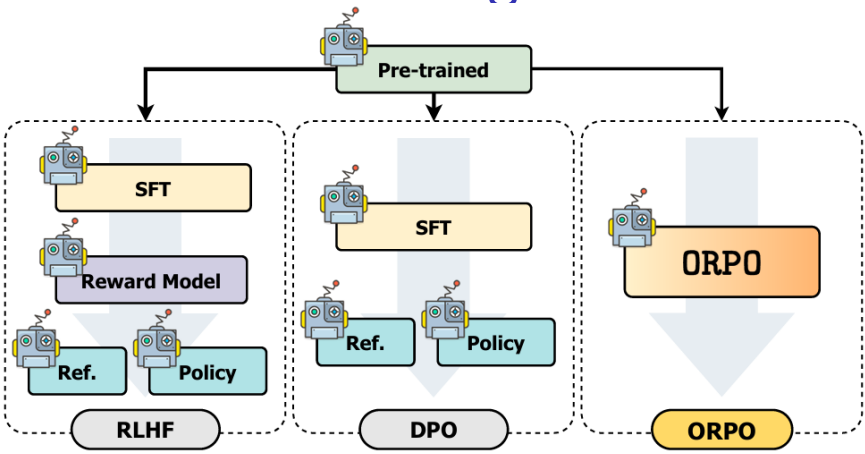


Figure 4: PPO, DPO and ORPO compared in terms of the model versions used during the steps of alignment tuning (Hong, Lee, and Thorne 2024)

Flash Attention

The HBM GPU memory's access is slow, use the SRAM cache instead (Dao et al. 2022)!

- Iterative processing of the QK product
- Parallelized softmax calculation
- Recompute intermediate values during backward pass
- In Flash Attention 2 (Dao 2023) GPU process scheduling is also optimized.

```
torch.backends.cuda.enable_flash_sdp()
```

Flash Attention

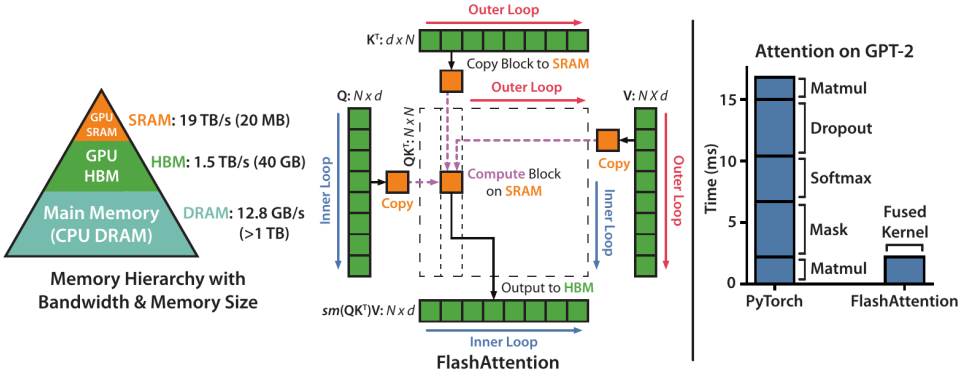


Figure 5: Hierarchy of GPU memory and the benefits of an iterative fused kernel to reduce HBM access. From (Dao et al. 2022)

Adapters

- Full fine-tuning of a GPT-3.5 level model needs 520 GB of memory@fp16.
- Tuning the top layers is inefficient.
- Adapter methods add small trainable parameter sets to all layers of the model.

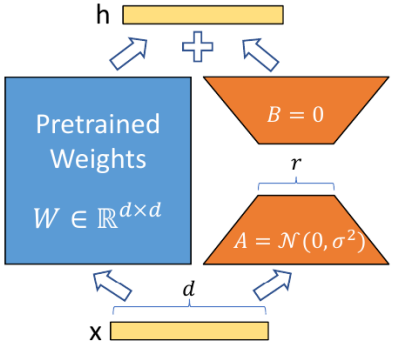
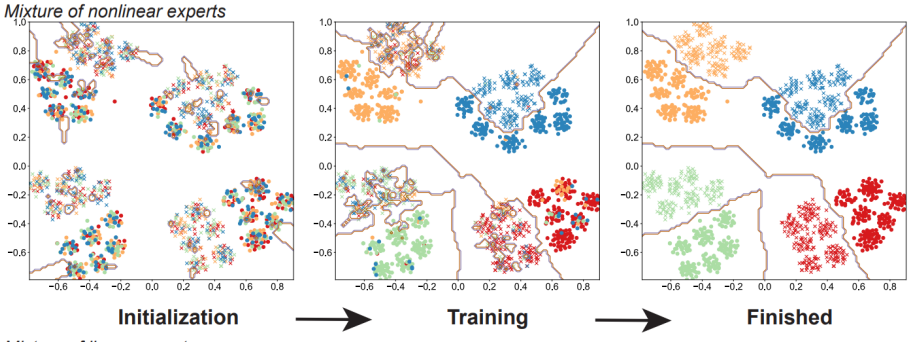


Figure 6: Parallel (mergable) low-rank adaptation (LoRA) method. LoRA's are portable. (Hu et al. 2022)

Ensembling

By combining models on the module level, ensembles, such as Mixtures of Experts (MoE) enable large, sparse models with data-specific experts (Z. Chen et al. 2022; Fedus, Zoph, and Shazeer 2022).



Ensembling

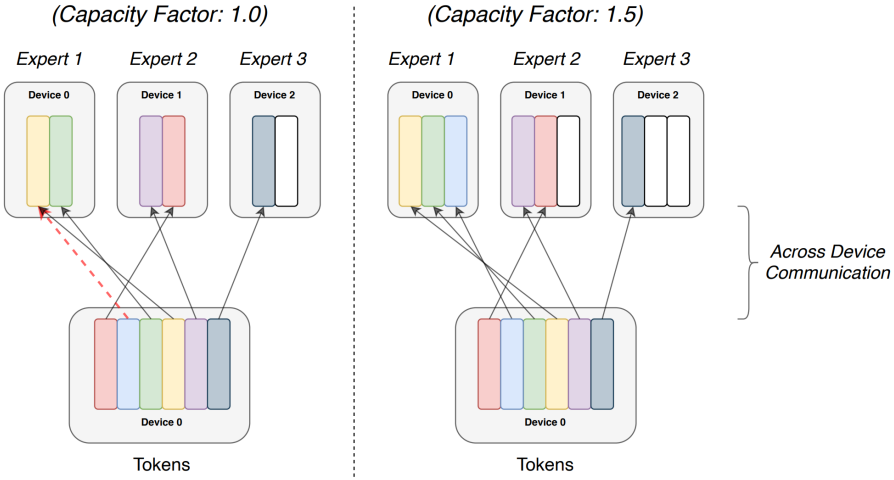


Figure 7: Switch Transformer from Fedus, Zoph, and Shazeer (2022)

Speculative Decoding

- Autoregressive predictions are guided by a smaller model (or medusa heads) (Xia et al. 2023; Leviathan, Kalman, and Matias 2023; C. Chen et al. 2023; Joao Gante 2023; Cai et al. 2024).
- Validation is done by the original model.
- 2-8x speedup with effectively no loss in quality.

Speculative Decoding

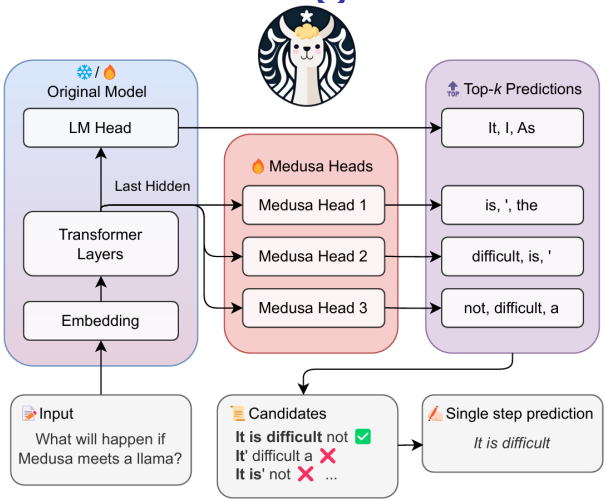


Figure 8: Medusa head k predicts the $1 + k$-th token. Candidates are validated by the main LLM head in the next pass while generating the new candidates as well. (Cai et al. 2024)

In-context Learning

- Context information is used to adapt the model's behavior on the fly enabling zero-shot and few-shot learning.
- This opens up the possibility of input-tuning and answer-engineering (as a ML task).
- The context could be accessed from an external data source as well.
- Reasoning and planning (agents) are also possible.

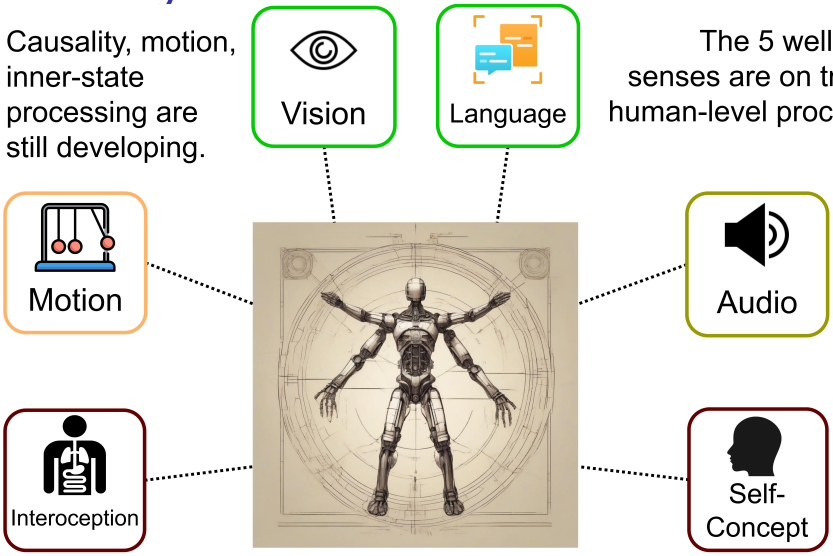
Agent-loops

Figure 9: A ReAct-style agent observes the current state, reasons about it, generates a candidate action and reflectively improves it before execution (Yao et al. 2023).

World Models and the Future

Modality Extension

Causality, motion, inner-state processing are still developing.



The 5 well known senses are on track for human-level processing.

Sense of Self (narrow case) is a controversial and relatively unexplored modality.

Emerging Modality Connections

Aligning modality pairs $\mathcal{M}_i$ and $\mathcal{M}_j$ along a spanning tree of all modalities we get weakly aligned modalities for each $\mathcal{M}_i$ and $\mathcal{M}_{k \neq j}$ as well. Language is a good candidate for a modality that can form pairs with most other modalities.

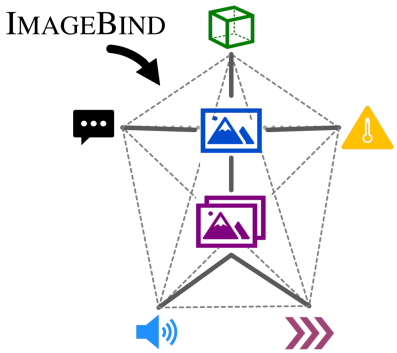


Figure 10: Modality pairs with training data (solid) and without training data (dotted) from (Girdhar et al. 2023)

Language as a Transporter of Meaning

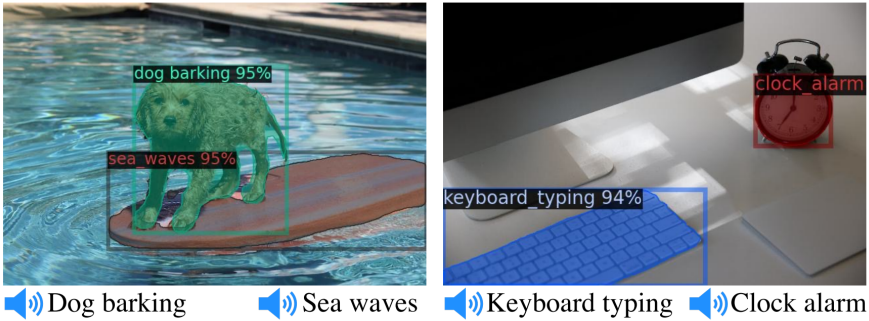
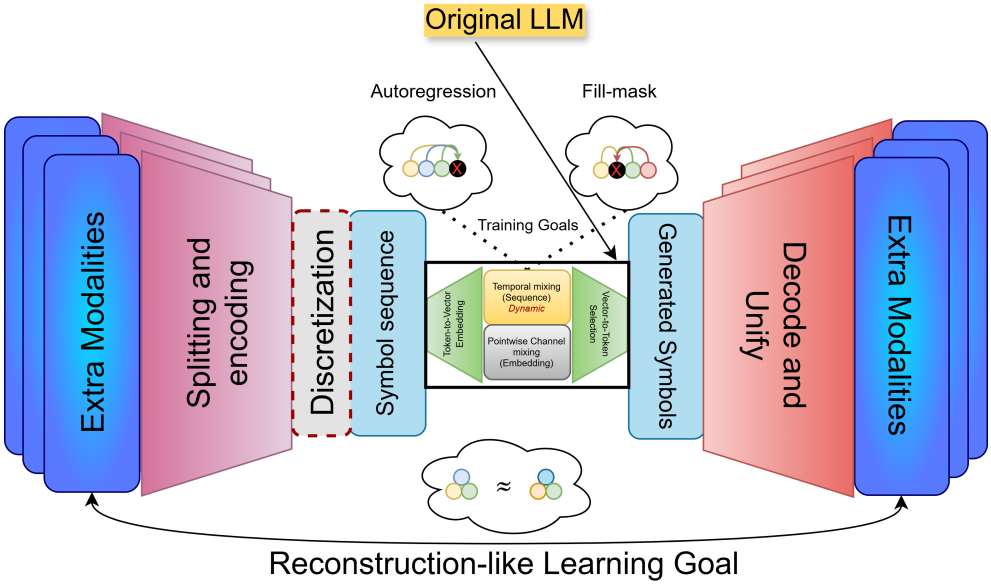


Figure 11: ImageBind retrievals of non-trivial modality pairs (with object detection in the visual modality) (Girdhar et al. 2023)

The Large World Model Template



LLava = LLama + Vision

- LLaVa uses an LLM + a CLIP-like vision encoder.
- It prepends a single image prefix to the text input and generates text.
- GPT-4V used a similar approach early 2023.

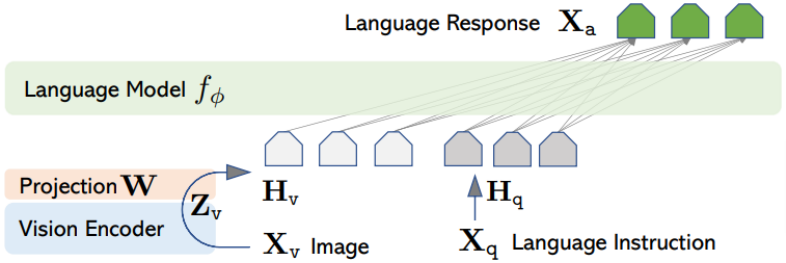


Figure 12: LLaVA architecture from Haotian Liu et al. (2023).

Interleaved Input & Proper Decoding

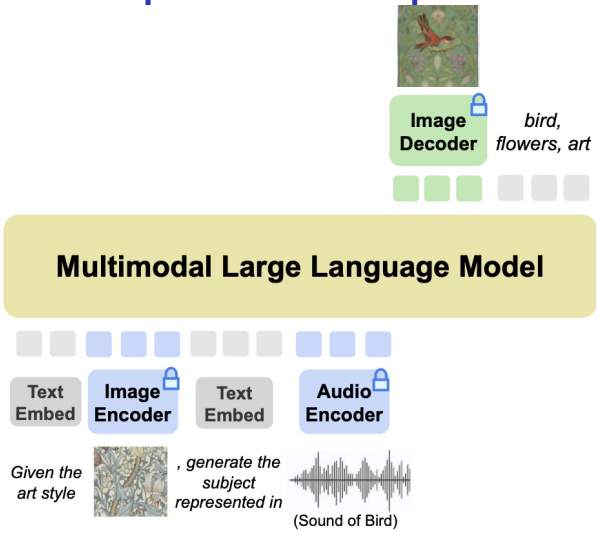


Figure 13: By applying the corresponding encoders and decoders Tang et al. (2023) train an any-to-any model.

LWMs in Action

LWMs are capable of **summarizing** lectures, **generating** toned audio responses, performing **speech recognition** at SOTA levels.

OpenAI (OpenAI 2024) and Google (Team et al. 2023) each provide LWM services for development **beating single-modality models** in many tasks. Input and output streaming is also possible to reduce latency (taking timing information into account).

LWMs in Action

Figure 14: Video-based Q&A by Hao Liu et al. (2024)

LWMs in Action

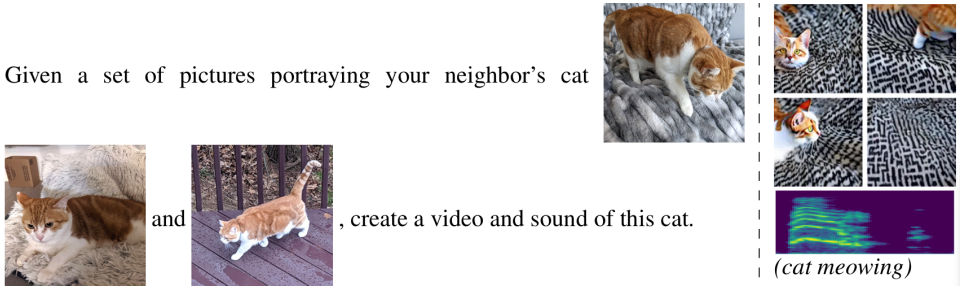


Figure 15: Multimodal generation based on interleaved input sequences by Tang et al. (2023)

And many more...

- Robot control (Collaboration et al. 2024)
- Action spaces & environment modeling (Bruce et al. 2024)
- Modelling priors for image generation (Ramesh et al. 2022)
- Time Series (Das et al. 2024)
- Motion (Jiang et al. 2023)
- 2D-to-3D object generation (Xu et al. 2024)

What we lack

- Stronger Reasoning (avoiding hallucinations)
- Continual Learning (personalization, adaptation)
- Symbolic Logical Inference (e.g. for theorem proving)
- Massively Multimodal Models (for dozens of modalities)

Strong AI?

Thank you for your attention!

References

References I

Bruce, Jake, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, et al. 2024. "Genie: Generative Interactive Environments." https://arxiv.org/abs/2402.15391.

Cai, Tianle, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. "Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads." https://arxiv.org/abs/2401.10774.

Chen, Charlie, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. "Accelerating Large Language Model Decoding with Speculative Sampling." https://arxiv.org/abs/2302.01318.

Chen, Zixiang, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. 2022. "Towards Understanding Mixture of Experts in Deep Learning." https://arxiv.org/abs/2208.02813.

References II

Collaboration, Embodiment, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, et al. 2024. "Open x-Embodiment: Robotic Learning Datasets and RT-x Models." https://arxiv.org/abs/2310.08864.

Cook, Vivian, and Mark Newson. 2014. *Chomsky's Universal Grammar: An Introduction*. John Wiley & Sons.

Dao, Tri. 2023. "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning." https://arxiv.org/abs/2307.08691.

Dao, Tri, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. "Flashattention: Fast and Memory-Efficient Exact Attention with Io-Awareness." In *Advances in Neural Information Processing Systems*, 35:16344–59.

Das, Abhimanyu, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. "A Decoder-Only Foundation Model for Time-Series Forecasting." https://arxiv.org/abs/2310.10688.

References III

Deacon, Terrence William. 1997. *The Symbolic Species: The Co-Evolution of Language and the Brain*. 202. WW Norton & Company.

Fedus, William, Barret Zoph, and Noam Shazeer. 2022. "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity." https://arxiv.org/abs/2101.03961.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. "Imagebind: One Embedding Space to Bind Them All." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–90.

Gu, Albert, and Tri Dao. 2023. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." https://arxiv.org/abs/2312.00752.

Hong, Jiwoo, Noah Lee, and James Thorne. 2024. "ORPO: Monolithic Preference Optimization Without Reference Model." https://arxiv.org/abs/2403.07691.

References IV

Hu, Edward J, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9.

Jiang, Biao, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. "MotionGPT: Human Motion as a Foreign Language." https://arxiv.org/abs/2306.14795.

Joao Gante. 2023. "Assisted Generation: A New Direction Toward Low-Latency Text Generation." Hugging Face Blog. https://doi.org/10.57967/hf/0638 .

Jones, Terry, John Goldstone, and Monty Python. 1979. "Monty Python's Life of Brian." In *Proceedings of the Monty Python Comedy Collection*, edited by Monty Python. United Kingdom: HandMade Films.

Lenci, Alessandro, and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge University Press.

References V

Leviathan, Yaniv, Matan Kalman, and Yossi Matias. 2023. "Fast Inference from Transformers via Speculative Decoding." https://arxiv.org/abs/2211.17192.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. "Visual Instruction Tuning." *arXiv Preprint arXiv:2304.08485*. https://arxiv.org/pdf/2304.08485.pdf.

Liu, Hao, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. "World Model on Million-Length Video and Language with Blockwise RingAttention." https://arxiv.org/abs/2402.08268.

OpenAI. 2024. "GPT-4o Introduction Page." https://openai.com/index/hello-gpt-4o/.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." https://arxiv.org/abs/2203.02155.

References VI

Pinker, Steven. 2003. *The Language Instinct: How the Mind Creates Language*. Penguin uK.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. "Language Models Are Unsupervised Multitask Learners."

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." https://arxiv.org/abs/2305.18290.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. "Hierarchical Text-Conditional Image Generation with CLIP Latents." https://arxiv.org/abs/2204.06125.

Skinner, Burrhus Frederic. 1957. *Verbal Behavior*. New York: Appleton-Century-Crofts.

References VII

Sun, Yutao, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. "Retentive Network: A Successor to Transformer for Large Language Models." https://arxiv.org/abs/2307.08621.

Tang, Zineng, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023. "CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation." https://arxiv.org/abs/2311.18775.

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. 2023. "Gemini: A Family of Highly Capable Multimodal Models." *arXiv Preprint arXiv:2312.11805*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008. https://arxiv.org/pdf/1706.03762.pdf.

References VIII

Xia, Heming, Tao Ge, Si-Qing Chen, Furu Wei, and Zhifang Sui. 2023. "Speculative Decoding: Lossless Speedup of Autoregressive Translation." https://openreview.net/forum?id=H-VlwsYvVi.

Xu, Jiale, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. 2024. "InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-View Large Reconstruction Models." https://arxiv.org/abs/2404.07191.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. "ReAct: Synergizing Reasoning and Acting in Language Models." https://arxiv.org/abs/2210.03629.