Multimodal LLMs

Yu Meng
University of Virginia
yumeng5@virginia.edu

Nov 11, 2024

UNIVERSITY of VIRGINIA

Join at
slido.com
#2705 251

Announcement

- Guest lecture (11/08) grades posted; contact Xu (ftp8nr@virginia.edu) if you have questions
- Assignment 5 has been released (deadline: **12/02 11:59pm**)

Join at
slido.com
#2705 251

Overview of Course Contents

- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling and Neural Language Models
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- Week 9-10: Reasoning, Knowledge, and Retrieval-Augmented Generation (RAG)
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations

(Recap) Overview: Language Model Alignment

Join at
slido.com
#2705 251

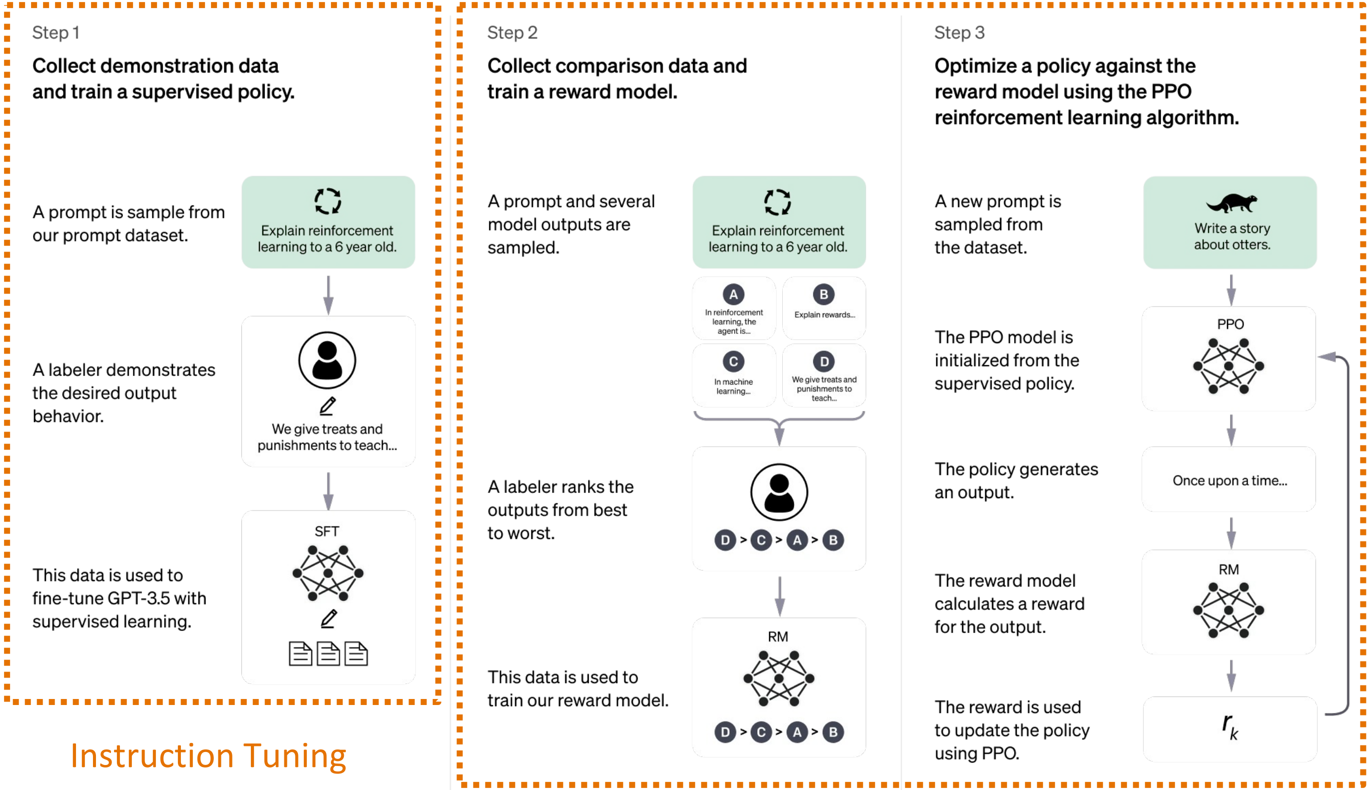
- Ensure language models behaviors are aligned with human values and intent
- "HHH" criteria (Askell et al. 2021):
  - **Helpful**: Efficiently perform the task requested by the user
  - **Honest**: Give accurate information & express uncertainty
  - **Harmless**: Avoid offensive/discriminatory/biased outputs

Paper: https://arxiv.org/pdf/2112.00861

(Recap) Post-training for Alignment

- Pretrained language models are **not** aligned
- Objective mismatch
  - Pretraining is to predict the next word in a sentence
  - Does not involve understanding human intent/values
- Training data bias
  - Text from the internet can contain biased, harmful, or misleading information
  - LMs don't distinguish between good and bad behavior in training data
- (Over-)generalization issues
  - LMs' generalization can lead to outputs that are inappropriate in specific contexts
  - Might not align with intended ethics/honesty standard

(Recap) Language Model Alignment Techniques



Instruction Tuning

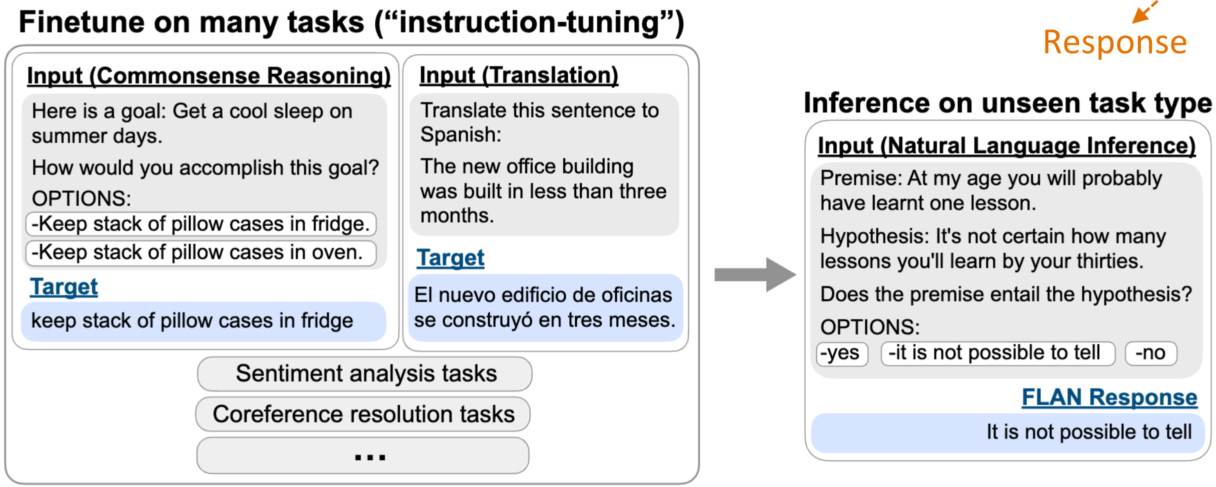
Reinforcement Learning from Human Feedback (RLHF)

Figure source: https://openai.com/index/chatgpt/

(Recap) Instruction Tuning: Method

- **Input**: task description
- **Output**: expected response or solution to the task
- Train LLMs to generate response tokens given prompts $\min_{\boldsymbol{\theta}} - \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$

Response Prompt



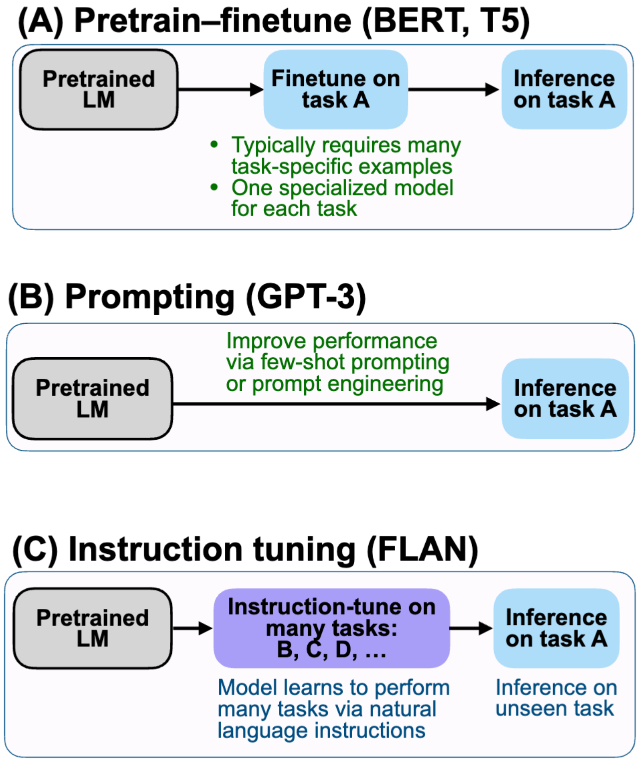
Paper: https://arxiv.org/pdf/2109.01652

UNIVERSITY of VIRGINIA

Join at
slido.com
#2705 251

(Recap) Instruction Tuning vs. Other Paradigms

- Task-specific fine-tuning does not enable generalization across multiple tasks
- In-context learning requires few-shot demonstrations
- Instruction tuning enables zero-shot cross task generalization

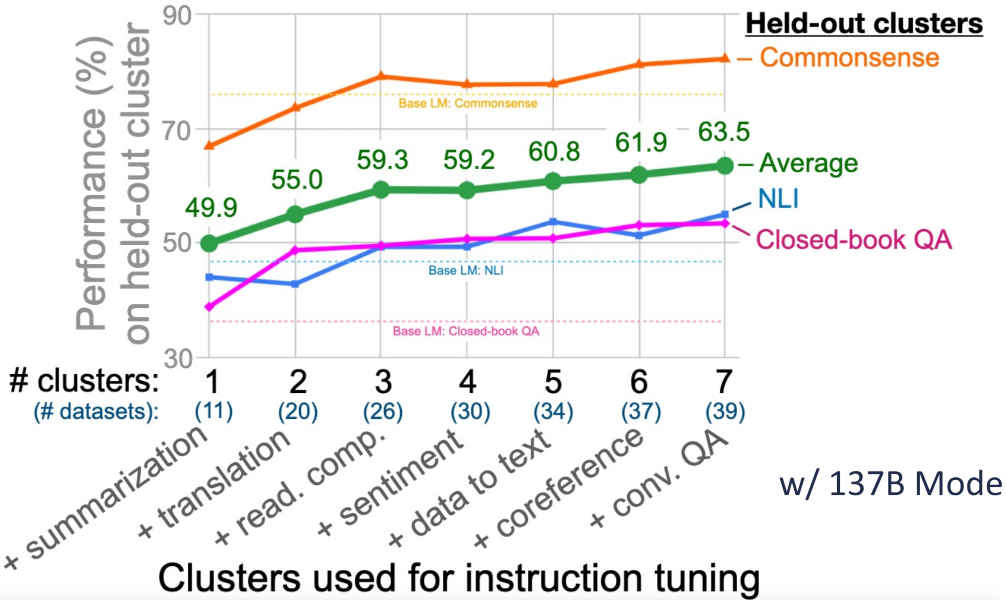


(Recap) Instruction Tuning vs. Pretraining

- Both instruction tuning and pretraining are **multi-task** learning paradigms
- Supervision
  - Pretraining: self-supervised learning (raw data w/o human annotation)
  - Instruction tuning: supervised learning (human annotated responses)
- Task format
  - Pretraining: tasks are implicit (predicting next tokens)
  - Instruction tuning: tasks are explicit (defined using natural language instructions)
- Goal
  - Pretraining: teach LMs a wide range of linguistic patterns & general knowledge
  - Instruction tuning: teach LMs to follow specific instructions and perform a variety of tasks

(Recap) Better Generalization with More Clusters

- Held out three clusters from instruction tuning: Commonsense, NLI, Closed-book QA
- More clusters and tasks used in instruction tuning => better generalization to unseen clusters

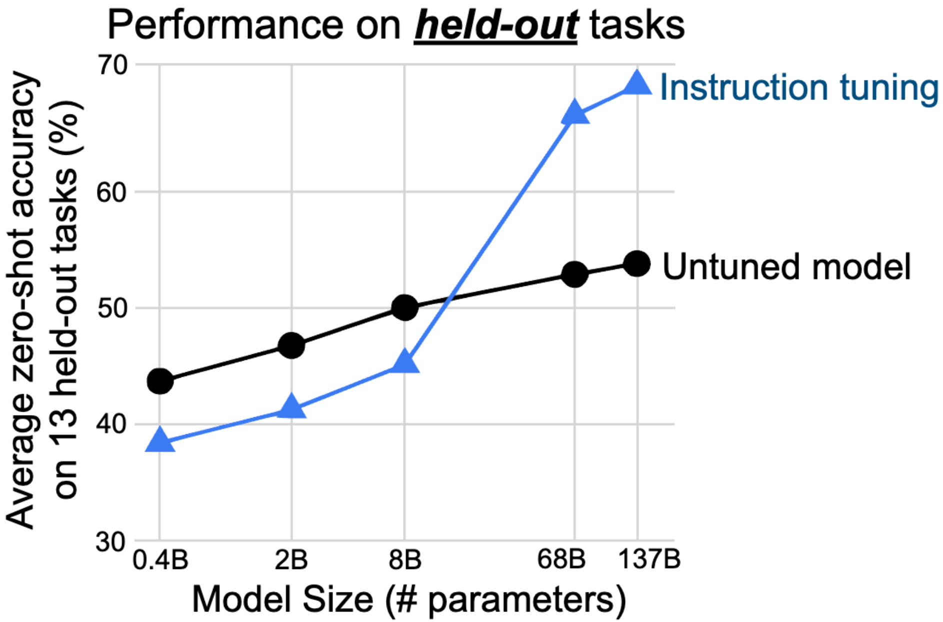


w/ 137B Model

(Recap) Different Model Sizes

Join at
slido.com
#2705 251

- Instruction tuning can hurt small model (< 8B) generalization
- Instruction tuning substantially improves generalization for large models



(Recap) Why RLHF

- **Costly human annotations**
  - Instruction tuning requires human annotators to write down the entire expected responses
  - RLHF only relies on preference labels (which response is better?)
- **Open-ended generation**
  - Open-ended creative generation (e.g., story writing) inherently has no single "right" answer
  - RLHF uses human feedback to determine which response is more creative/appealing
- **Token-level learning**
  - Instruction tuning applies the language modeling loss -> penalizes all token mistakes equally regardless of their impact on the overall quality of the output (e.g., a grammatical error might be less critical than a factual inaccuracy)
  - RLHF uses human feedback to prioritize the error types that are more important to correct
- **Suboptimal human answers**
  - Instruction tuning may learn the suboptimal patterns written by humans
  - Identifying a better answer from a few options is usually easier than writing an optimal answer entirely

UNIVERSITY of VIRGINIA

Join at
slido.com
#2705 251

(Recap) RLHF Illustration

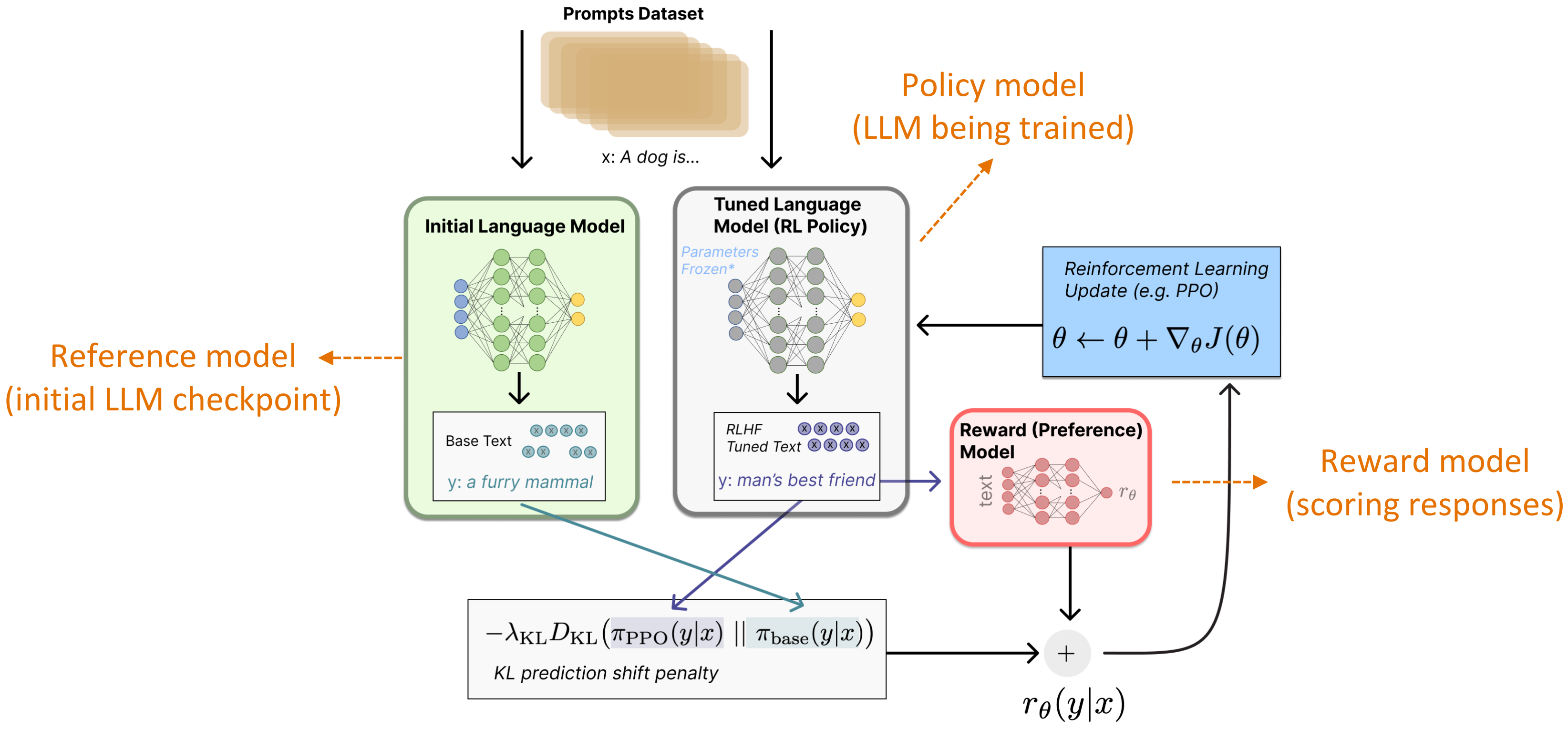


Figure source: https://huggingface.co/blog/rlhf

(Recap) Preference Data

Humans are better at relative judgments than absolute ones

What are the steps for making a simple cake?

1. *Preheat oven to 350°F (175°C).*
2. *Grease and flour a cake pan.*
3. *In a bowl, combine 2 cups flour, 1.5 cups sugar, 3.5 tsp baking powder, and a pinch of salt.*
4. *Add 1/2 cup butter, 1 cup milk, and 2 tsp vanilla; mix well.*
5. *Beat in 3 eggs, one at a time.*
6. *Pour batter into the pan.*
7. *Bake for 30-35 minutes or until a toothpick comes out clean.*
8. *Let cool, then frost or serve as desired.*

What are the steps for making a simple cake?

1. *Warm up the oven.*
2. *Grease a cake pan.*
3. *Blend dry ingredients in a bowl.*
4. *Incorporate butter, milk, and vanilla.*
5. *Mix in the eggs.*
6. *Pour into the prepared pan.*
7. *Bake until golden brown.*
8. *Add frosting if desired.*

Preference data: $(\boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{y}_l)$

prompt

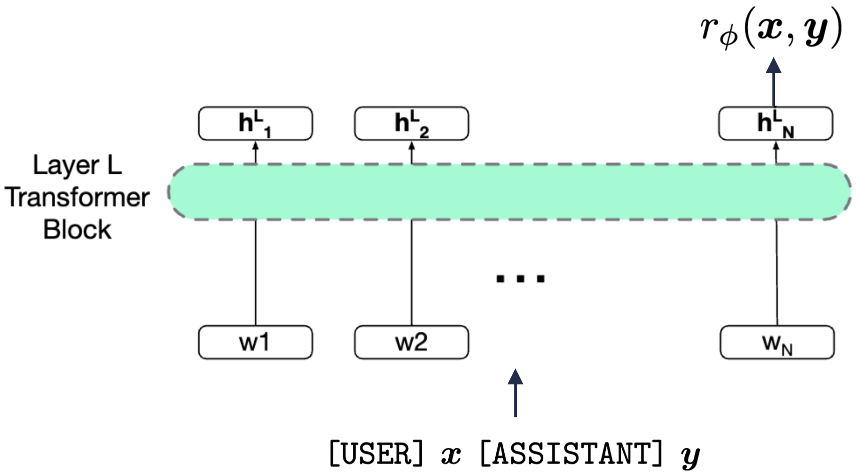
preferred (winning) response

dispreferred (losing) response

Figure source: https://lm-class.org/lectures/14%20-%20aligning%20llms.pdf

(Recap) Reward Model Setup

Goal: train a reward model to assign a higher reward to $\boldsymbol{y}_w$ than $\boldsymbol{y}_l$



Apply a linear layer at the last token representation to learn a scalar output

Join at
slido.com
#2705 251

(Recap) Reward Model Training

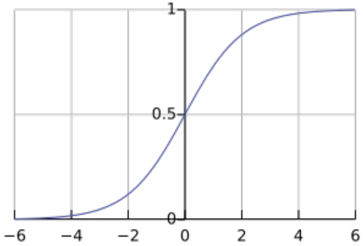
Bradley-Terry pairwise comparison objective

$$\mathcal{L}_{\mathrm{RM}}(r_\phi) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}}\left[\log \sigma\left(r_\phi(\boldsymbol{x},\boldsymbol{y}_w) - r_\phi(\boldsymbol{x},\boldsymbol{y}_l)\right)\right]$$

reward of winning response

reward of losing response

$y = \sigma(x)$



(Recap) Regularized Reward Optimization

- Add a penalty for drifting too far from the initial SFT checkpoint

$$\max_{\theta} \mathbb{E}_{\boldsymbol{y} \sim p_\theta(\cdot|\boldsymbol{x})} \left[\underbrace{r_\phi(\boldsymbol{x}, \boldsymbol{y})}_{\text{Maximize reward}} - \underbrace{\beta}_{\text{hyperparameter}} \underbrace{\log \left(\frac{p_\theta(\boldsymbol{y}|\boldsymbol{x})}{p_{\text{SFT}}(\boldsymbol{y}|\boldsymbol{x})} \right)}_{\text{Prevent deviation from the initial (SFT) model}} \right]$$

- Penalize cases where $p_\theta(\boldsymbol{y}|\boldsymbol{x}) > p_{\text{SFT}}(\boldsymbol{y}|\boldsymbol{x})$
- In expectation, it is known as the Kullback-Leibler (KL) divergence $\mathrm{KL}(p_\theta(\boldsymbol{y}|\boldsymbol{x}) \| p_{\text{SFT}}(\boldsymbol{y}|\boldsymbol{x}))$

UNIVERSITY of VIRGINIA

Join at slido.com #2705 251

(Recap) Optimization with RL

- Why reinforcement learning:
  - No supervised data available (only a reward model)
  - Encourage the model to explore new possibilities (generations) guided by the reward model
- Optimization: policy gradient methods
  - Optimize the policy (LLM) by adjusting the parameters in the direction that increases expected rewards
- REINFORCE (simplest policy gradient method):

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a|s) R$$

cumulative reward

step size

policy model (LLM)

action (generating the response)

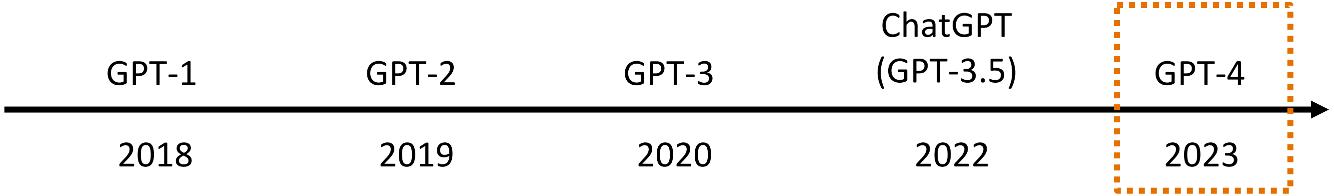
state (user prompt + conversation history)

Further Reading on RLHF

- RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment [Dong et al., 2023]
- Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint [Xiong et al., 2023]
- SLiC-HF: Sequence Likelihood Calibration with Human Feedback [Zhao et al., 2023]
- SimPO: Simple Preference Optimization with a Reference-Free Reward [Meng et al., 2024]

The Evolution of GPT Models: GPT-4

- GPT-1: decoder-only Transformer pretraining
- GPT-2: language model pretraining is multi-task learning
- GPT-3: scaling up & in-context learning
- ChatGPT: language model alignment
- **GPT-4: multimodality**



GPT-4: Multimodal Input Processing

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

accept both images and texts as input

Screenshot source: https://openai.com/index/gpt-4-research/

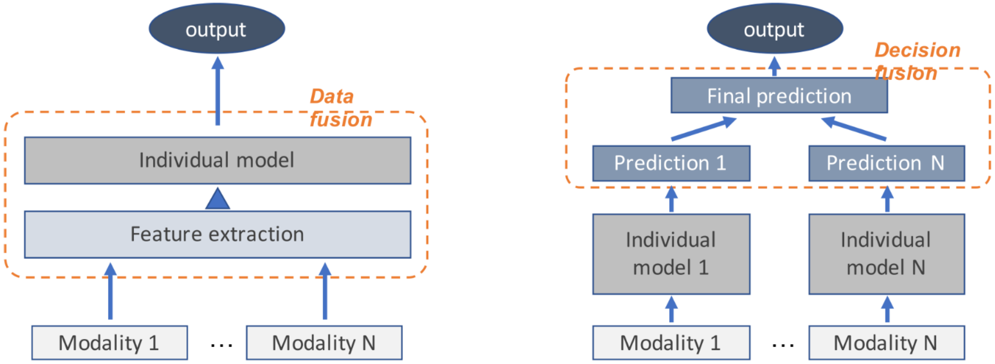
Overview: Multimodal LLMs

- Process and understand multiple types of data (e.g., text, images, audio, and video)
- More comprehensive and contextually rich understanding & generation
- Multimodal input processing (common):
  - Accept and process different types of input data
  - Examples: understanding the content of an image, transcribing and interpreting speech, analyzing video content, or integrating information from sensor data
- Multimodal output generation (less common):
  - Generate output in various modalities
  - Examples: creating realistic images from text descriptions, translating speech to text, or generating music according to user descriptions

Overview: Multimodal Architecture

- Architecture:
  - Require modality-specific architectures (e.g., vision/audio/video encoders)
  - Usually LLMs serve as the strong base
- Multimodal fusion: fuse information from different modalities
  - Early fusion: Combine raw input data from different modalities before processing
  - Late fusion: Process each modality separately and then combine the representations later

Early fusion



Late fusion

Figure source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9548039

Overview: Multimodal Datasets

Join at
slido.com
#2705 251

Training datasets need to contain paired examples of different modalities => teach the model the relationships between different types of data

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board.

Figure source: https://arxiv.org/pdf/2304.08485

Learning Aligned Visual Representations

- Goal: learn a joint embedding space where images and their matching text descriptions are close together
- **CLIP** (Contrastive Language-Image Pretraining): predict the correct pairings of a batch of (image, text) training examples

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Paper: https://arxiv.org/pdf/2103.00020

CLIP: Contrastive Pretraining

Maximize similarity between correct image-text pairs and minimize for incorrect ones

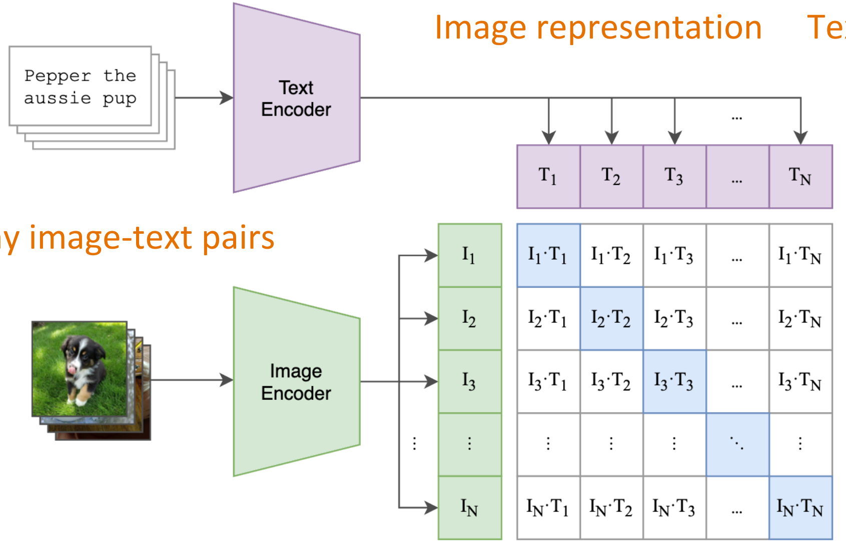
$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\cos(\boldsymbol{I}_i,\boldsymbol{T}_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(\boldsymbol{I}_i,\boldsymbol{T}_j)/\tau)} + \log\frac{\exp(\cos(\boldsymbol{T}_i,\boldsymbol{I}_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(\boldsymbol{T}_i,\boldsymbol{I}_j)/\tau)}\right)$$

Image representation

Text representation

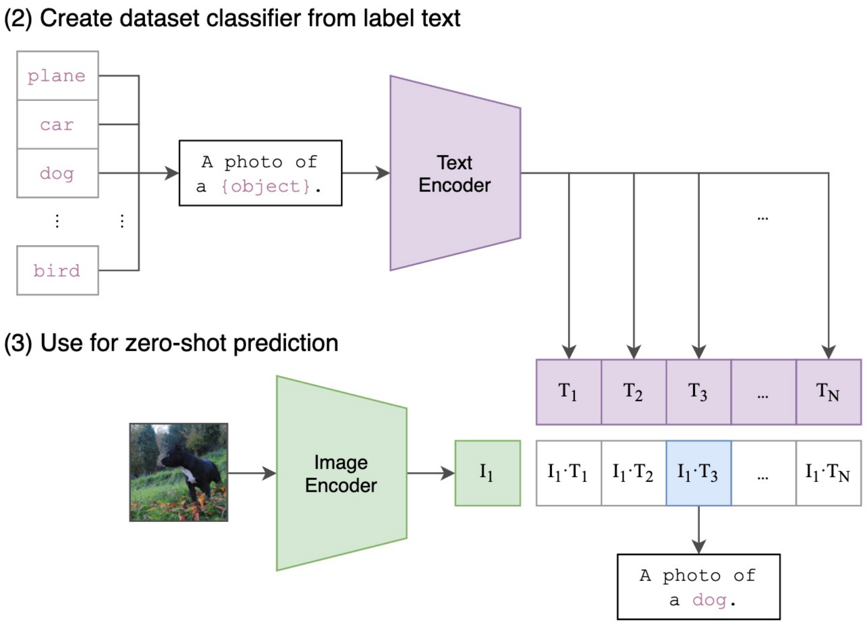
Learnable temperature

Many image-text pairs



CLIP: Zero-shot Generalization

After training, the text encoder/image encoder can embed the target class names/test images for zero-shot image classification



UNIVERSITY of VIRGINIA

Join at
slido.com
#2705 251

Visual Instruction Tuning

- Goal: fine-tune a multimodal LLM to learn to follow instructions for tasks that involve both visual and textual information
- **LLaVA** (Large Language and Vision Assistant): combine a pretrained vision encoder (e.g., CLIP) with a large language model (e.g., Llama) for visual instruction tuning

Visual Instruction Tuning

Haotian Liu[1*], Chunyuan Li[2*], Qingyang Wu[3], Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

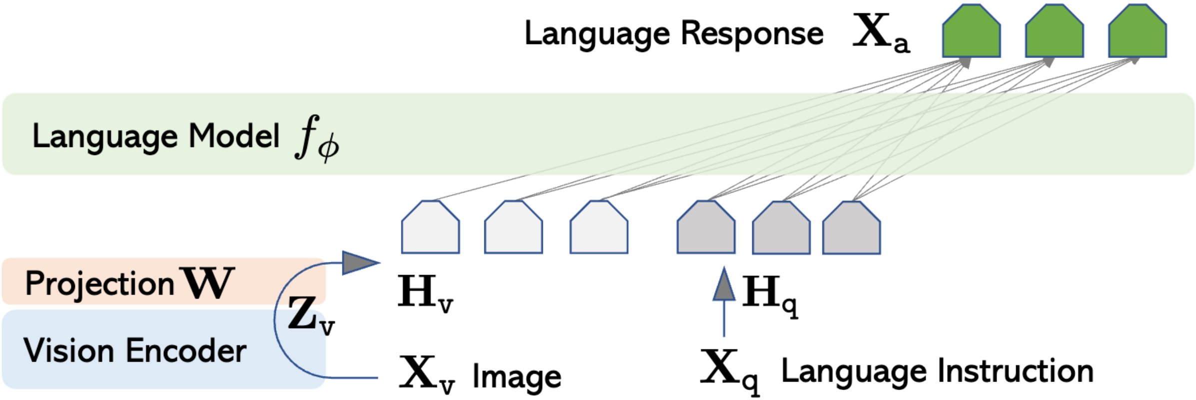
Paper: https://arxiv.org/pdf/2304.08485

UNIVERSITY of VIRGINIA

Join at
slido.com
#2705 251

LLaVA: Architecture

- Learn a projection matrix ($\boldsymbol{W}$) to convert imagine representations ($\boldsymbol{Z}_v$) to text embeddings ($\boldsymbol{H}_v$)
- Concatenate visual tokens ($\boldsymbol{H}_v$) with text tokens ($\boldsymbol{H}_q$) as input to the model



Adopted in latest multimodal Llama models

meta-llama/**Llama-3.2-90B-Vision**

meta-llama/**Llama-3.2-11B-Vision**

LLaVA: Results

User Can you explain this meme in detail?

LLaVA The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are

Further Reading on Multimodal LLMs

- Zero-Shot Text-to-Image Generation [Ramesh et al., 2021]
- Flamingo: a Visual Language Model for Few-Shot Learning [Alayrac et al., 2022]
- AudioLM: a Language Modeling Approach to Audio Generation [Borsos et al., 2022]
- Movie Gen: A Cast of Media Foundation Models [Polyak et al., 2024]

Thank You!

Yu Meng
University of Virginia
yumeng5@virginia.edu