



Thinking about Thinking about AI Projects

This is not your typical machine
learning class...

1. Thinking about projects
2. Documentation

1. Thinking about projects

- HW1. Datasets
- HW2. Multimodal Fusion
- HW3. Multimodal LLMs
- HW4. Reasoning & Reinforcement Learning
- HW5. Agents

Final Project

We want to give you complete freedom over what you apply the required techniques to from each week in the assignments.

So here are some examples of
projects

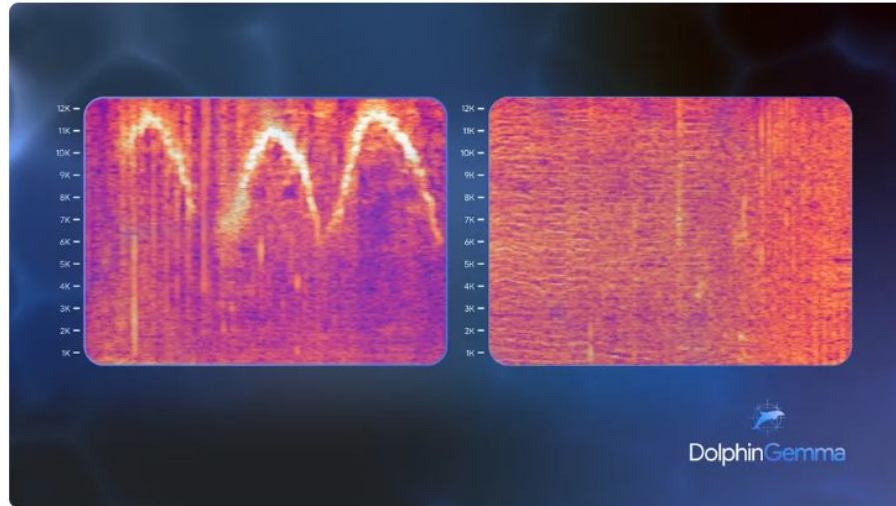


[Credit: Google Teachable Machine](#)

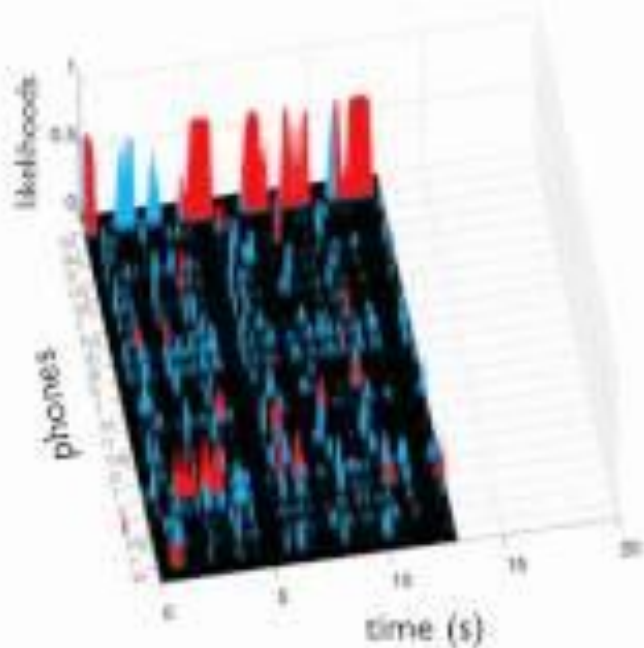
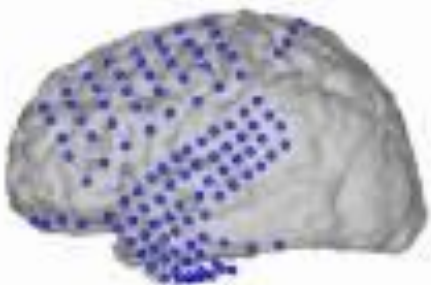
Introducing DolphinGemma

Analyzing dolphins' natural, complex communication is a monumental task, and WDP's vast, labeled dataset provides a unique opportunity for cutting-edge AI.

Enter DolphinGemma. Developed by Google, this AI model makes use of specific Google audio technologies: the SoundStream tokenizer efficiently represents dolphin sounds, which are then processed by a model architecture suited for complex sequences. This ~400M parameter model is optimally-sized to run directly on the Pixel phones WDP uses in the field.



Left: Whistles (left) and burst pulses (right) generated during early testing of DolphinGemma.



Decoded phones:

/w/ /h/ /ch/ /s/ /f/ /h/ /s/ /m/ /i/ /e/ /h/ /ch/ /aa/ /n/ /h/ /aa/ /s/ /ow/ /w/ /e/ /h/ /s/ /b/ /aa/ /n/ /k/ /aa/ /m/ /h/ /t/ /aa/ /r/

This Person Does Not Exist

Using generative adversarial networks (GAN), we can learn how to create realistic-looking fake versions of almost anything, as shown by this collection of sites that have sprung up in the past month. Learn [how it works](#).



This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.

Created by Phillip Wang.



This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.

Created by Ryan Hoover.



This Rental Does Not Exist

Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.

Created by Christopher Schmidt.



This Waifu Does Not Exist

This one is a little stranger than the others, but I bet you'll read more of the plot summary than you expect. It's



This Question Does Not Exist

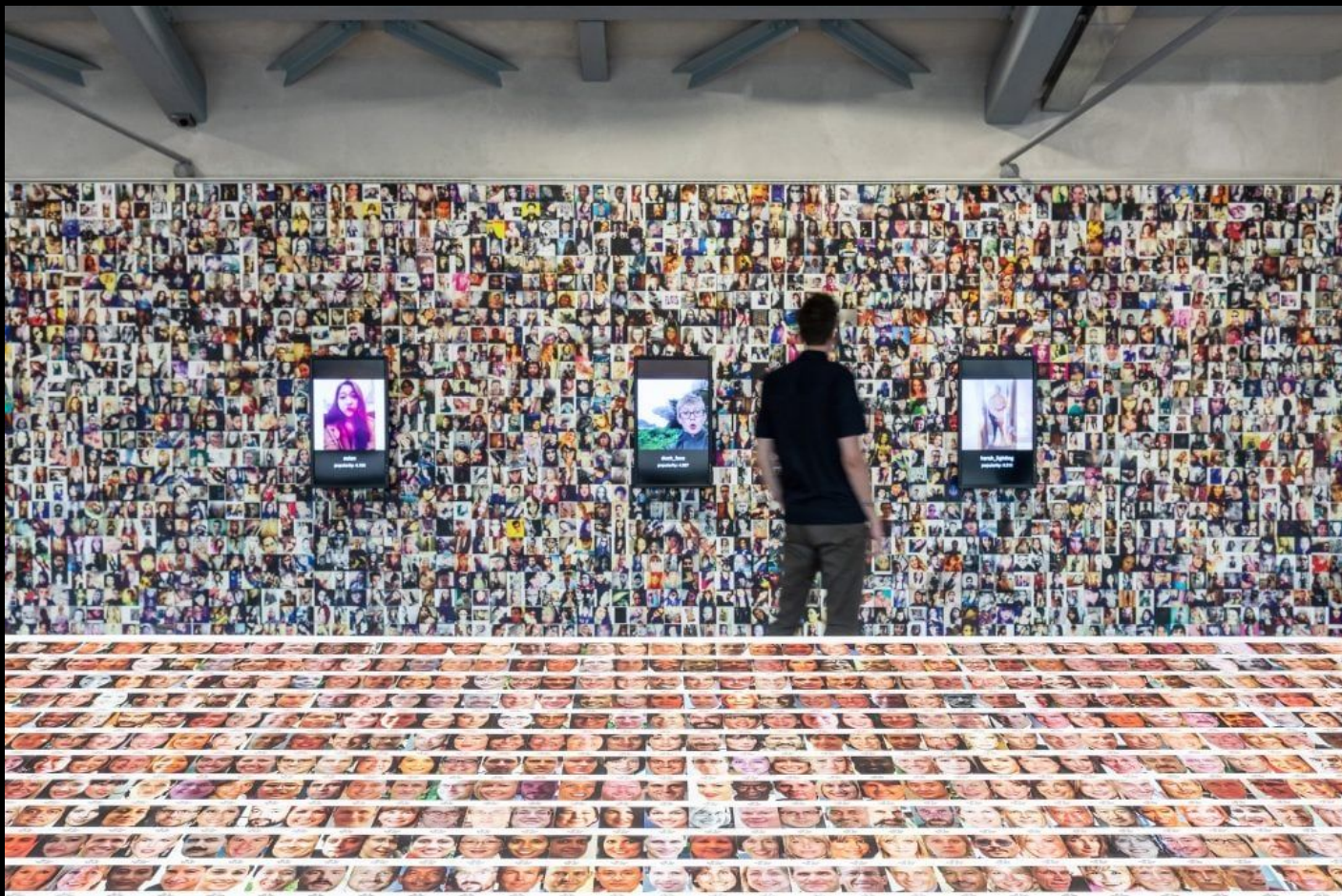
By generating seemingly realistic Stack Overflow questions, you can keep this open on your computer to seem like



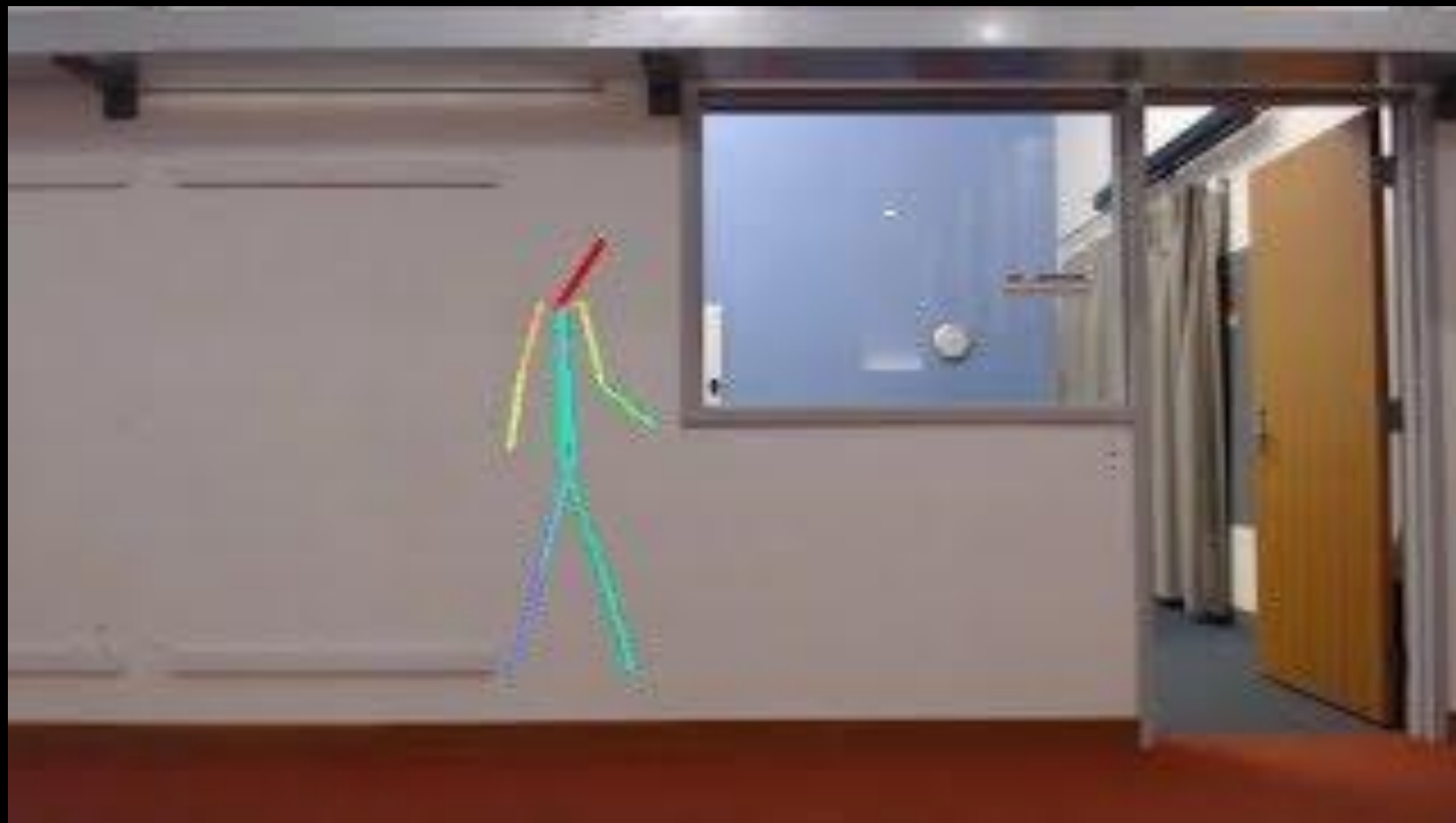
This resume does not exist

This Resume Does Not Exist

To showcase their resume template builder, this company went as far as to incorporate TextgenRNN – not quite a



ImageNet Roulette Exhibition



[Through Wall Post Estimation using RF Signals \(CSAIL\)](#)



SignAloud Gloves (MIT)

Audio Tone Description (C_{atd}):

"The woman in the video speaks with an excited voice."

Lexical Subtitle (C_{ls}):

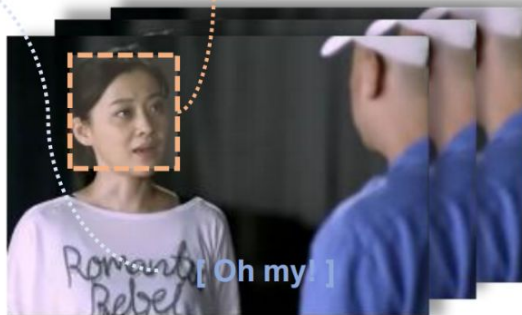
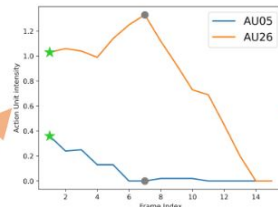
"Oh my!"

Visual Objective Description (C_{vod}):

"The woman in the video is talking to a man, possibly discussing something important or sharing her thoughts and feelings."

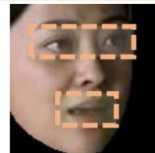
Classification Label (C_{cl}):

"Surprise"



Raw Video & Audio

Visual Expression Description (C_{ved}):



AU-05: 0.36

AU-26: 1.03

"Eyes widened, Wide-mouthed."

Multi-modal Description (C_{md}):

"In the video, a woman is conversing with a man. Her facial expressions, with eyes widened and mouth wide open, clearly show surprise. She amplifies this visual cue with an excited tone as she exclaims, "Oh my." This combination of voice and expression indicates that she is experiencing surprise, likely in response to unexpected news shared during the conversation."



PointLLM: Empowering Large Language Models to Understand Point Clouds

Runsen Xu^{1,2}, Xiaolong Wang³, Tai Wang², Yilun Chen², Jiangmiao Pang², Dahua Lin^{1,2}

The Chinese University of Hong Kong¹, Shanghai AI Laboratory², Zhejiang University³

Paper

Code

arXiv

Model

Dataset

Demo

ECCV 2024 Best Paper Candidate

intLLM: Empowering Large Language Models to Understand Point Clouds.

intLLM (Open in Colab)

intLLM Method

File Object ID

Object ID Input

d679a3889c54e4b9a8b89915a77162

3D Model

Chat

What is this?

This is a 3D model of a cartoon-style spaceship that boasts a playful design with a silver body and accented with red details. The model, which is imagined in a fun, child-friendly style, resembles the spacecraft commonly seen in children's animation and video games. It doesn't have specific features identifiable as functional parts of a real spaceship but serves as an aesthetically pleasing object meant to entertain and inspire imagination.

Do you like it?

Yes, as a model designed to evoke fun and imagination, it is a well-crafted piece that can be appreciated.

Enter text and press enter

Send

Clear

0:22 / 0:24

PointLLM is online! Try it at <http://101.230.144.196>. You can chat with PointLLM about the models of the Objaverse dataset or about your own point clouds!

MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing

Siddhant Agarwal^{1*}, Shivam Sharma^{2,3*}, Preslav Nakov⁴, Tanmoy Chakraborty²

¹Indraprastha Institute of Information Technology Delhi, India

²Indian Institute of Technology Delhi, India ³Wipro R&D (Lab45), India

⁴Mohamed bin Zayed University of Artificial Intelligence, UAE

siddhant20247@iitd.ac.in, {shivam.sharma, tanchak}@ee.iitd.ac.in, preslav.nakov@mbzuai.ac.ae

Abstract

Mememes have evolved as a prevalent medium for diverse communication, ranging from humour to propaganda. With the rising popularity of image-focused content, there is a growing need to explore its potential harm from different aspects. Previous studies have analyzed memes in closed settings – detecting harm, applying semantic labels, and offering natural language explanations. To extend this research, we introduce MemeMQA, a multimodal question-answering framework aiming to solicit accurate responses to structured questions while providing coherent explanations. We curate MemeMQACorpus, a new dataset featuring 1,880 questions related to 1,122 memes with corresponding answer-explanation pairs. We further propose ARSENAL, a novel two-stage multimodal framework that leverages the reasoning capabilities of LLMs to address MemeMQA. We benchmark MemeMQA using competitive baselines and demonstrate its superiority – ~18% enhanced answer prediction accuracy and distinct text generation lead across various metrics measuring lexical and semantic alignment over the best baseline. We ana

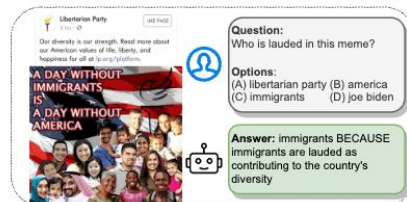


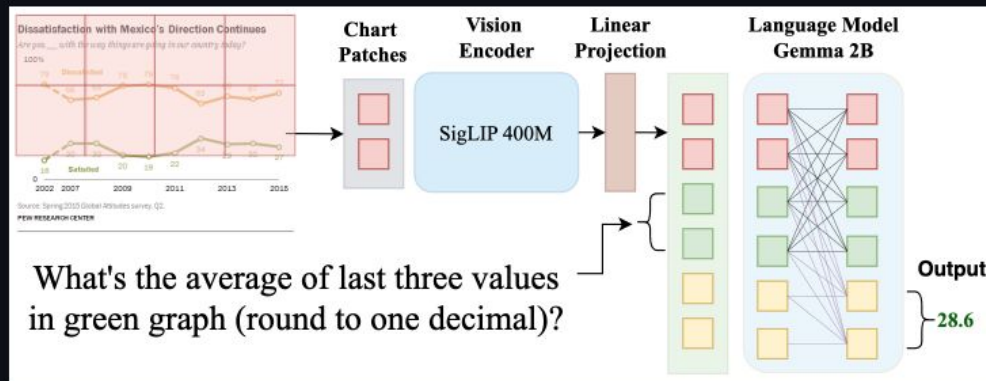
Figure 1: The MemeMQA task: Given an input meme and multiple choices, identify the correct answer and justify.

sometimes be harmful to the general audience, especially in the age of Internet virality. Previous work has explored aspects such as harmfulness in various forms, such as hate speech (Kicla et al., 2020a), cyber-bullying (Sharma et al., 2022b), and offensive languages (Shang et al., 2021), of memes, typically in a black-box setting.

Mememes, with their appealing format and influential nature on social media, necessitate the modeling of complex aspects like harmfulness, targeted social groups, and offensive cues to assess their narrative framing and ensure online content safety.

ChartGemma: Visual Instruction-tuning for Chart Reasoning in the Wild

- Authors: [Ahmed Masry*](#), Megh Thakkar*, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, Shafiq Joty (*equal contribution)
- Paper Link: [ChartGemma](#)



ChartGemma Model Checkpoints

We release the checkpoint for our pretrained model on huggingface.

Protecting Indigenous Languages Using Automatic Speech Recognition

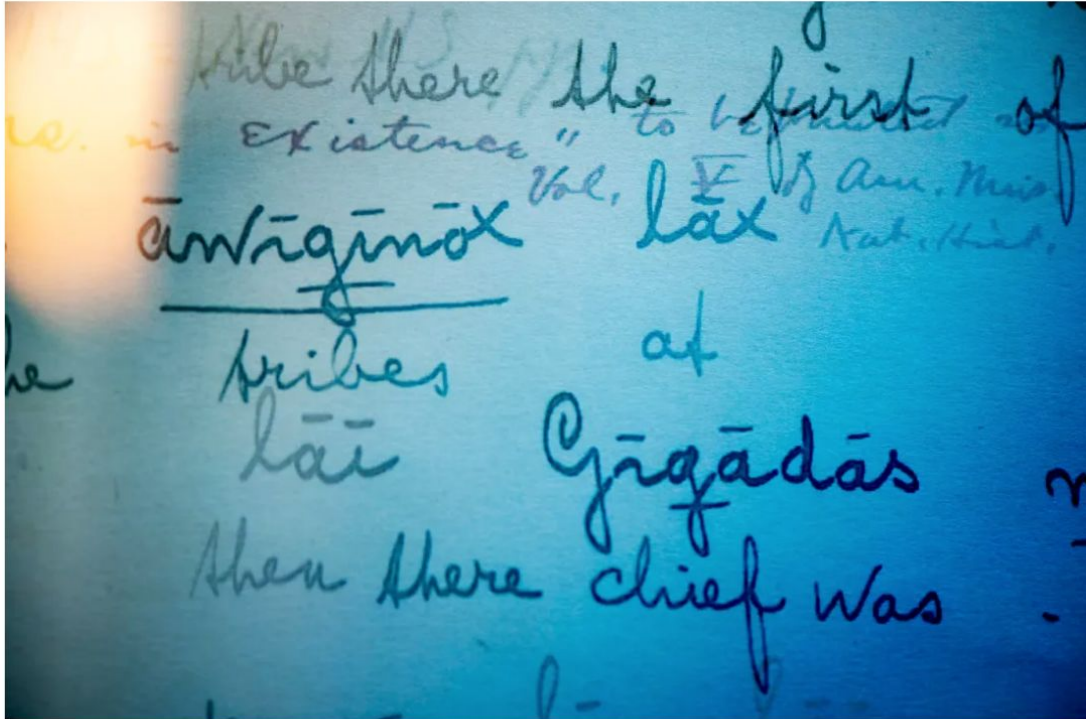





Photo Illustration of Kwak'waka text written by Northeastern clinical instructor Michael Running Wolf. Photo by Alyssa Stone/Northeastern University

 **ykilcher/gpt-4chan** 

 like 143



Text Generation



Transformers



PyTorch



English

gptj

causal-lm



arxiv:2109.07958



License: apache-2.0



Model card



Files and versions

 xet



Community 14

Access to this model has been disabled

Given its research scope, intentionally using the model for generating harmful content (non-exhaustive examples: hate speech, spam generation, fake news, harassment and abuse, disparagement, and defamation) on all websites where bots are prohibited is considered a misuse of this model. **Head over to the Community page for further discussion and potential next steps.**

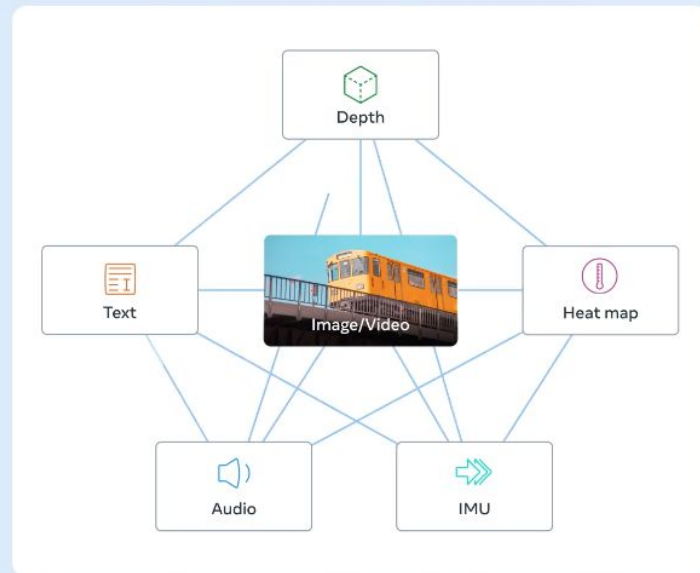
Computer vision

ImageBind: a new way to 'link' AI across the senses

Introducing ImageBind, the first AI model capable of binding data from six modalities at once, without the need for explicit supervision. By recognizing the relationships between these modalities — images and video, audio, text, depth, thermal and inertial measurement units (IMUs) — this breakthrough helps advance AI by enabling machines to better analyze many different forms of information, together.

Explore the demo to see ImageBind's capabilities across image, audio and text modalities.

[See its capabilities](#)





**AI assistants, applications, or other people in a
silent, concealed and seamless manner.**

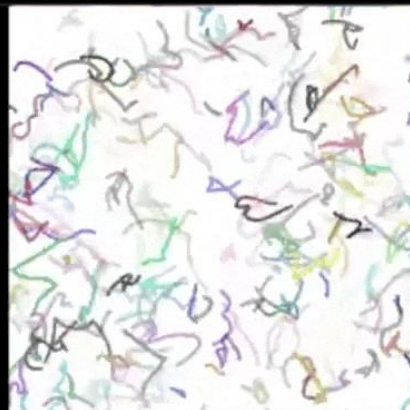




“Underwater
Submarine”



“Forest Temple as
3D Rendered in
Unreal Engine”



“Watercolor painting of
a fire-breathing
dragon”



Valdemar Danry x Cenk Güzelis

PRESENTS

ORGANs WITHOUT BODIES

OPENING: _5PM, JULY 17
MIT MEDIA LAB, E14 LOBBY

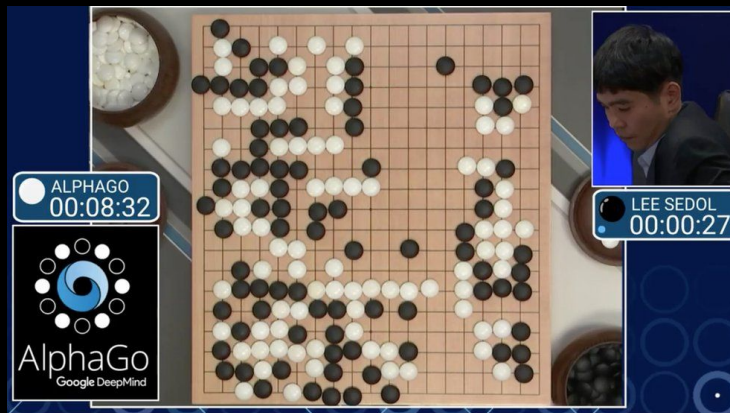




1863 Civil War Union Captain James Cunningham



ReLive: Walking into Virtual Reality Spaces from Video Recordings of One's Past



AlphaGo



AlphaStar



Unity ML Agents

Draw
crocodile
in under 20 seconds

Got It!





“About A Hero”

Nothing Forever



ZOLTAN KAKLER



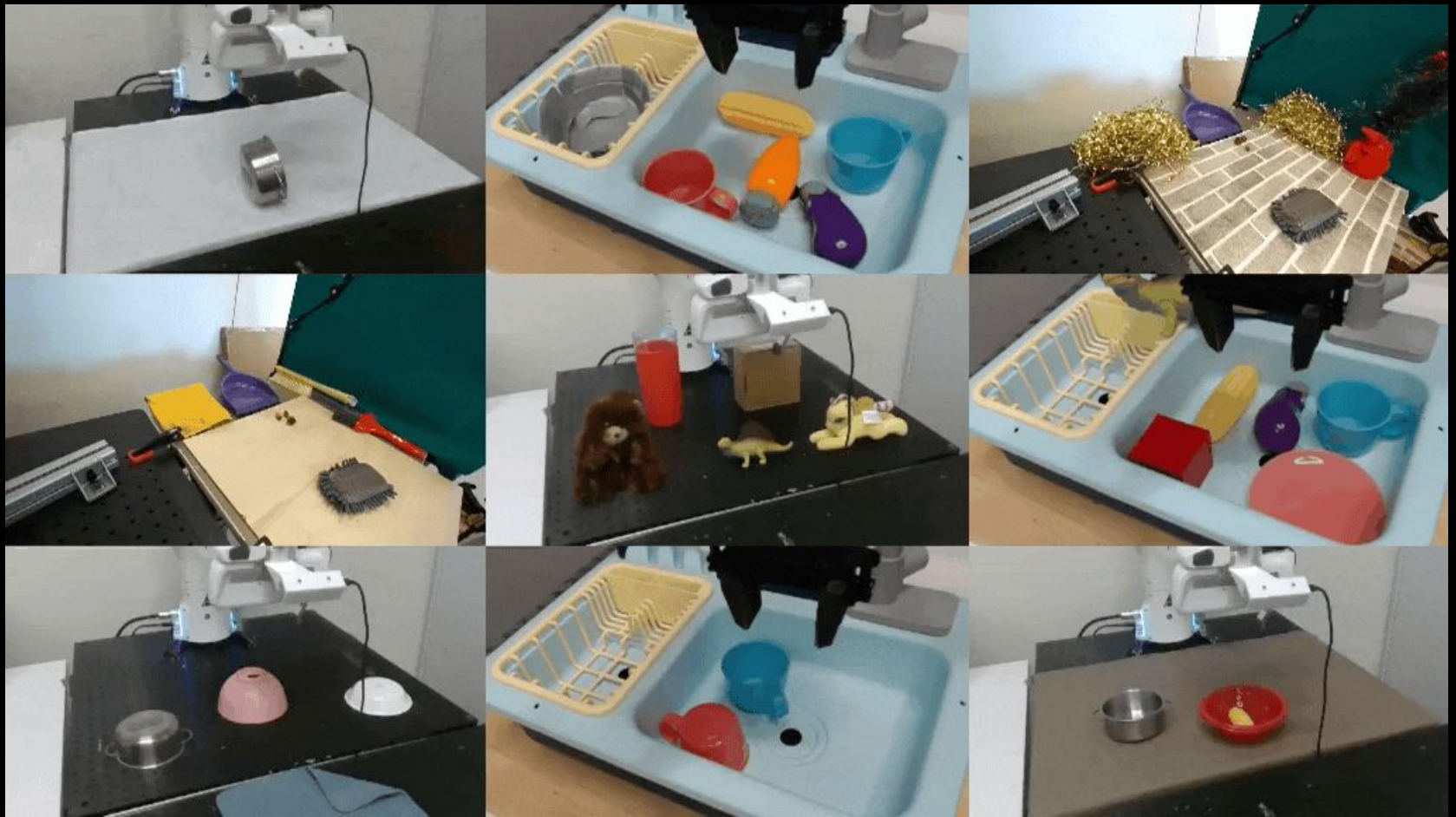


Jordan Rudess and the jam_bot (MIT Media Lab)





Credit: Mateo Larrea Ferro



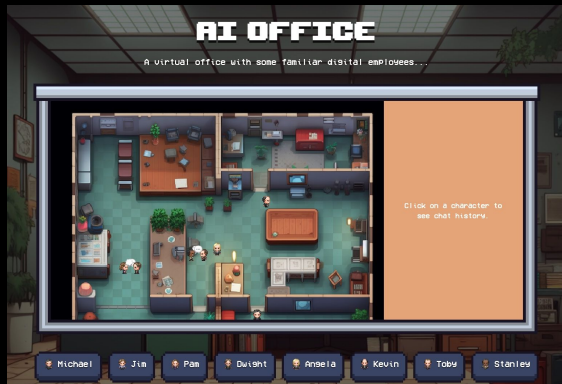
[OpenVLA: An Open-Source Vision-Language-Action Model](#)



Generative Agents



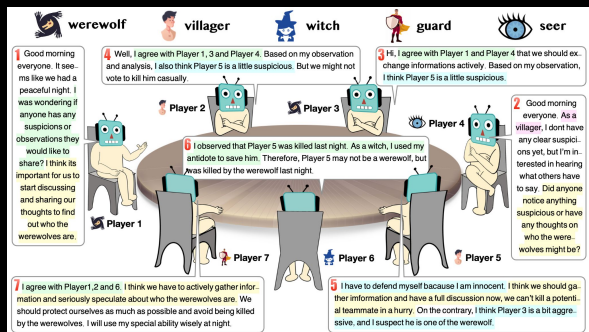
<https://github.com/a16z-infra/ai-town>



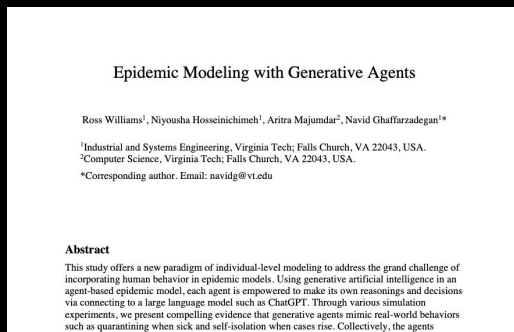
<https://github.com/Parcha-ai/ai-office>



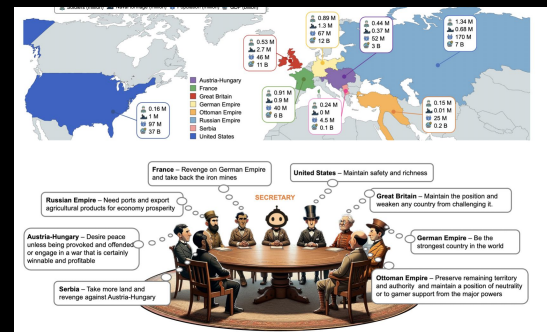
<https://github.com/OpenBMB/ChatDev>



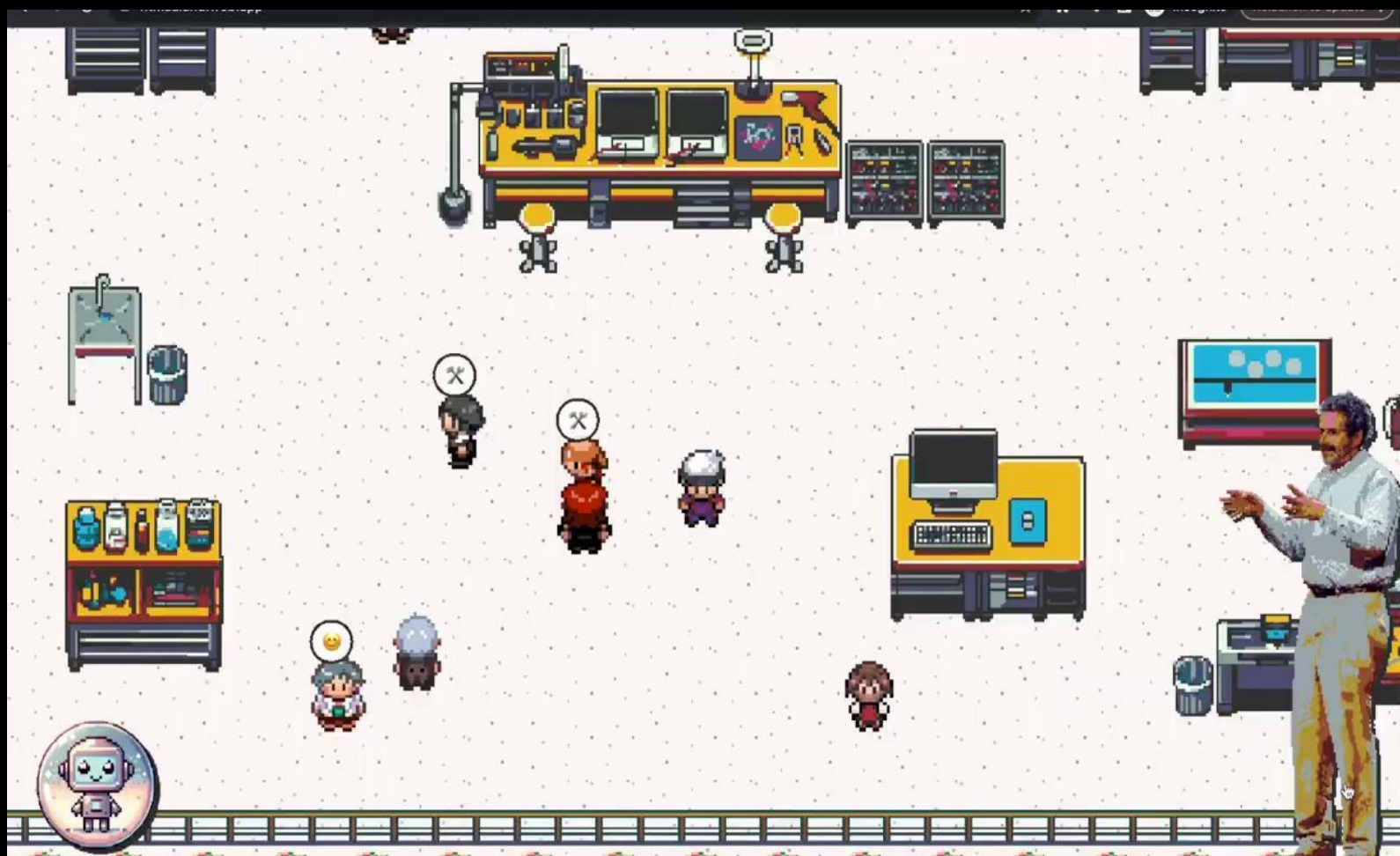
<https://github.com/xuyuzhuang11/Werewolf>



Epidemic Modeling with Generative Agents



<https://github.com/agiresearch/WarAgent>



Credit: Olivia Seow

OPENCLAW
OpenCLAW DASHBOARD
Health OK

Chat

Chat

Control

Overview

Channels

Instances

Sessions

Cron Jobs

Agent

Skills

Nodes

Settings

Config

Debug

Logs

Resources

Docs

Chat

Direct gateway chat session for quick interventions.

I need an automatic meeting preparation agent that researches attendees and sends me briefings before external meetings. **Calendar Integration** - Connect to my Google Calendar (or Outlook/Office 365) - Monitor for upcoming external meetings (exclude internal team meetings) - Identify external meetings by: attendees with external email domains, meeting titles containing keywords like "call", "demo", "pitch", "discussion" **Meeting Detection Criteria:** External meetings include: - Sales calls/demos - Investor/fundraising meetings - Partnership discussions - Customer success check-ins - Vendor/procurement calls - Industry networking calls - Board meetings with external board members Exclude: - Internal team meetings (all attendees @mycompany.com) - 1:1s with direct reports - All-hands or company meetings - Personal calendar events **Research Requirements:** For each attendee (non-internal), research: 1. **LinkedIn Profile:** - Current role and company - How long in current position - Career history (previous companies/roles) - Education background - Mutual connections (people we both know) - Recent posts or activity (if public) 2. **Company Intelligence:** - Company overview (what they do, size, location) - Recent funding rounds or financial news - Recent product launches or announcements - Recent press coverage or media mentions - Growth trajectory (hiring, expansion) - Tech stack (if relevant for sales calls) 3. **Our Relationship History:** - Pull all email threads from Gmail involving this person - Summarize past conversations and context - Identify any open action items or commitments - Note last interaction date and topic - Check CRM (Salesforce/HubSpot) for deal history or notes 4. **Social Media & News:** - Recent Twitter/X posts (if public profile) - Recent company blog posts - TechCrunch, news mentions in last 30 days - LinkedIn company page updates 5. **Context & Insights:** - Why are we meeting? (infer from calendar title/description) - What stage is this relationship? (first meeting, ongoing, renewal, etc.) - Any red flags or concerns from past interactions? - Opportunities to mention (mutual interests, shared connections) **Briefing Format (WhatsApp Message):** Send 30 minutes before meeting: 📅 MEETING PREP: [Meeting Title] ⏰ In 30 minutes [Time] 👥 ATTENDEES: [Name 1] - [Title] at [Company] - Role: [Current position], [tenure] at company - Background: Previously [relevant experience] - 🔄 Mutual connections: [Names or "None found"] - 📄 Recent: [Company raised \$10M Series A / They just launched new product / etc.] [Name 2] - [Title] at [Company] - (Similar format) 🏢 COMPANY CONTEXT: - [Company Name]: [Brief description] - Size: [employees/revenue if available] - Recent news: [Funding/launches/press in last 30 days] - Tech stack: [If sales call and relevant] 📅 OUR HISTORY: - Last contact: [Date] - [Topic summary] - Email summary: [2-3 key points from past conversations] - Open items: [Any pending commitments or questions] - CRM

Message (↵ to send, Shift+↵ for line breaks, paste images)

New session
Send



The goal of this class is to teach you
how to use multimodal AI to make
(almost) anything

After each assignment there will be a
show and tell of everyone's projects

For final your project this can
include new techniques, datasets,
applications or empirical findings

Examples of Previous Project Reports

Project reports from student teams who participated in the MMAI course

We list here only project reports that were publicly released by students. It should be noted that some of these links are for the follow-up submissions to conferences, after some revisions of the original project reports.

Phoebe Chua, Cathy Mengying Fang, Takehiko Ohkawa, Raja Kushalnagar, Suranga Nanayakkara, Pattie Maes. [EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language](#). arXiv 2025

Chenyu Zhang, Minsol Kim, Shohreh Ghorbani, Jingyao Wu, Rosalind Picard, Patricia Maes, Paul Pu Liang. [When One Modality Sabotages the Others: A Diagnostic Lens on Multimodal Reasoning](#). NeurIPS 2025 Workshop

Shohreh Ghorbani, Chenyu Zhang, Minsol Kim, Jingyao Wu. [Beyond Accuracy: A Diagnostic Protocol for Fairly Evaluating Multimodal Reasoning](#). NeurIPS 2025 Workshop

Haofei Yu, Zhengyang Qi, Lawrence Jang, Russ Salakhutdinov, Louis-Philippe Morency, Paul Pu Liang. [MMoE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts](#). EMNLP 2024

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssef Kebe, Paul Pu Liang, Louis-Philippe Morency. [Social-iq 2.0 Challenge: Benchmarking Multimodal Social Understanding](#). ICCV 2023 Challenge

Vedant Palit, Rohan Pandey, Aryaman Arora, Paul Pu Liang. [Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP](#). ICCV 2023

Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, Louis-Philippe Morency. [Multimodal Lecture Presentations Dataset: Understanding Multimodality in Educational Slides](#). ICCV 2023

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, Louis-Philippe Morency. [Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions](#). ACL 2023

1. Thinking about projects
2. Documentation

Setting up your class page



Thinking about Thinking about AI Projects



Massachusetts Institute of Technology

MAS.S60/6.S985

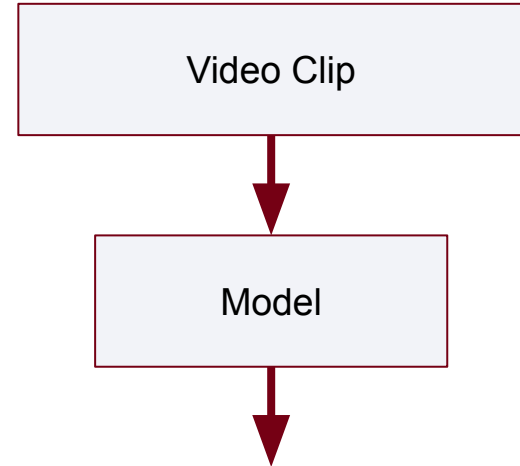
Dataset Collection and Preprocessing

Part One

Data Collection

First Steps

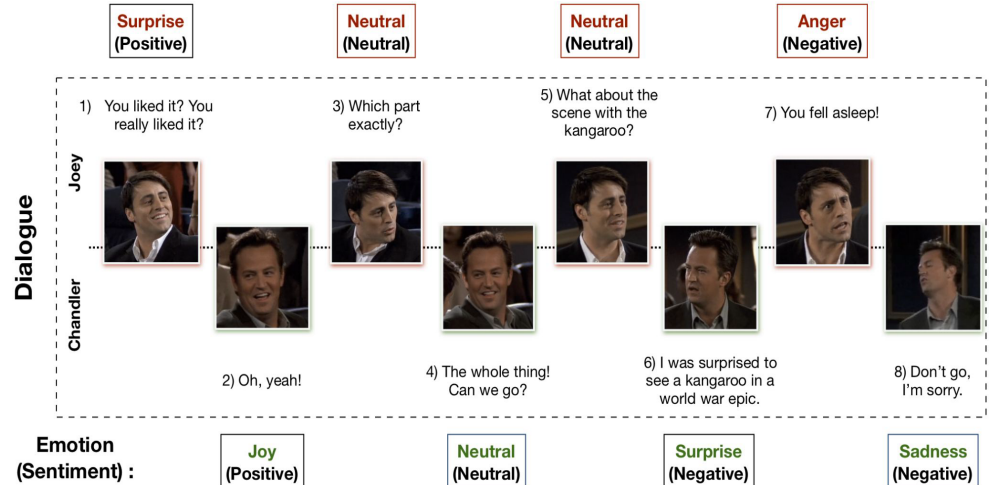
- Obtaining a good dataset is imperative for any ML/AI project
 - Bad data -> Bad results (Bad!)
- Before beginning collection, we should first identify the following:
 - What is the overall goal?
 - For multimodal projects: What specific modalities are we interested in utilizing?
- These questions help shorten the search or creation of datasets.



Forming a Dataset

- Now we need to find a dataset that best suits our needs
- Ideally try to find one that has all the modalities that you plan on using
 - If not → find one which you can then extract any missing modalities from
- Don't restrain yourself to the limitations of current datasets → sometimes you have to be creative if trying to solve issue!

MELD Dataset:



Part Two

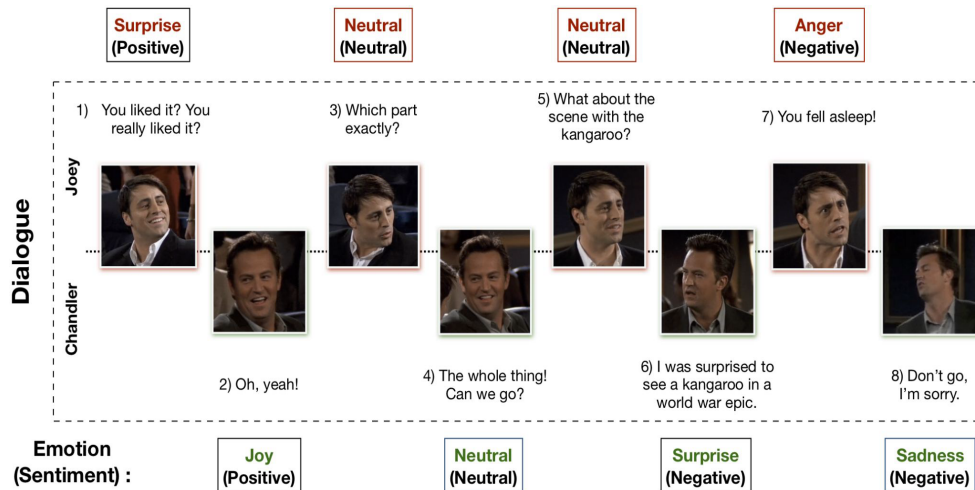
Data Preprocessing/Feature Extraction

When doing preprocessing or feature extraction, here's how to code up all the algorithms for a modality of interest.

Do Not.

Data Preprocessing/Feature Extraction

- Of course there can be exceptions to what was said, like when a project focuses on that or if there does not exist a better way to do it, but...
- These days there are
 - Multimodal datasets with the modalities of interest already extracted
 - Open source packages that can do this extraction for you pretty easily (you may have to code up some logic to do it in a loop)
- But generally, you should not need to completely code up from scratch logic to extract a modality. Just use what is already out there if possible.



This already has video audio and text modalities → just download and use what I need.

Example Packages for Extracting Modalities

- Video
 - FFMpeg
 - OpenCV
 - MoviePy
- Audio
 - Low-Level Features
 - OpenSMLE
 - PyAudioAnalysis
 - FFMpeg
- Facial Features
 - MediaPipe
 - OpenFace

Example Packages for Extracting Modalities

- Text
 - OpenAI Whisper
 - MoviePy
 - OpenCV
 - SpeechRecognition
- Pose
 - OpenPose
 - MediaPipe
- For other more niche modalities, either there does exist a package to extract it, the dataset has already done it or you may need to do it yourself. Look deep and be creative if needed!
- Most importantly, be aware of package dependencies! Always make sure to set up these packages in a virtual environment!

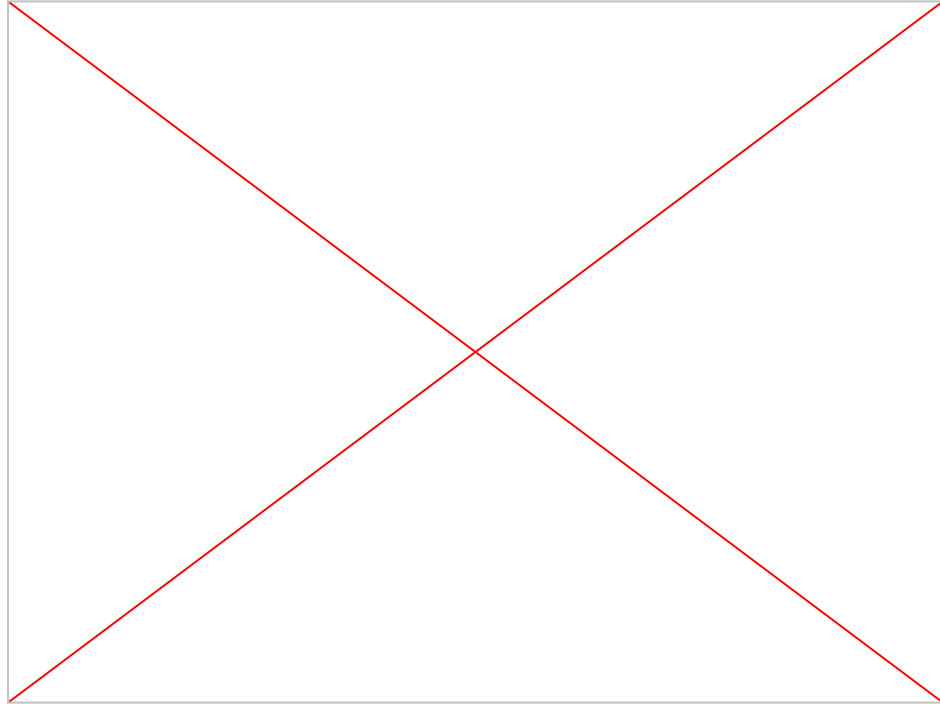
Part Three

A More Real Example of This

A More Real Example (Background)

ADOS (Autism Detection)

- Official diagnostic for placing children within the spectrum
- Children are present in a very specific room with examiner and parents
 - Presented with various stimuli and responses are recorded and noted by examiner, with video recording to later review
- Responses are used to score items in the exam which are group by specific type of response
 - Scores are aggregated to create final score



A More Real Example (Background)

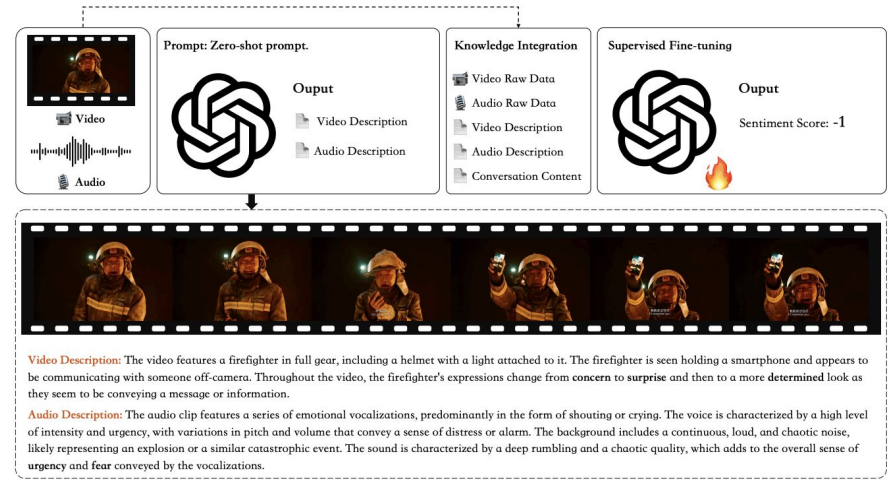
- Raw data is in the form of these video sessions
 - These are very long videos that current multimodal models cannot handle (this is a open question/problem currently being researched!)
- Just training on this would not likely yield beneficial results/models
 - Subtleties in body movement, pitch, facial expressions could be lost
- Extracting modalities could allow models to capture these subtleties and better score ADOS items

A More Real Example

- Modalities Captured
 - Video (Broken into shorter-length segments)
 - Audio
 - Text (Video Captions/dialogue)
 - Facial Features
 - Video Summaries
 - Mid Level Features
 - Emotion
 - Sentiment
 - Anxiety
 - Depression

A More Real Example

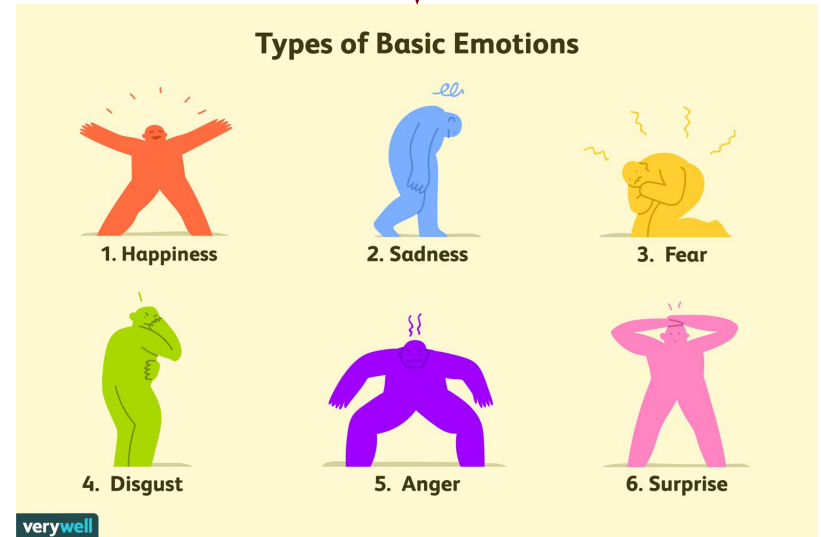
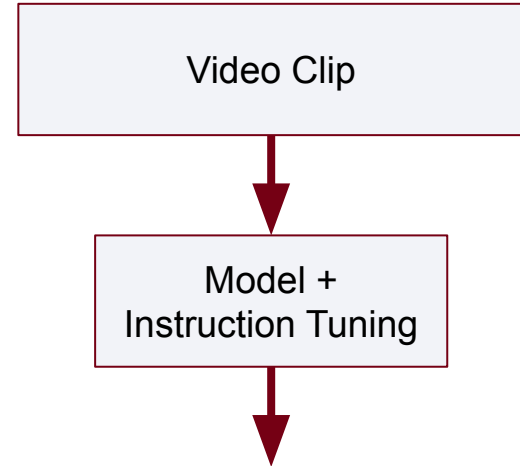
- How to get video summaries?
- Idea from another paper:
 - Use available models to do this for us
- Resulting summaries may not be perfect, but with a little bit of nudging and proper wording we can get decent results
- **Rationale:** Current Multimodal LLMs are still very text-hungry/text-dependent. Providing additional text in the form of generated video summaries as extra context, even if not perfect → improved performance!



Idea + Image credits: Luo, M., Long, J., Li, Z., Yang, Y., Jiang, Y., & Mai, S. (2025). Multimodal Large Language Models for End-to-End Affective Computing: Benchmarking and Boosting with Generative Knowledge Prompting. arXiv preprint arXiv:2508.02429.

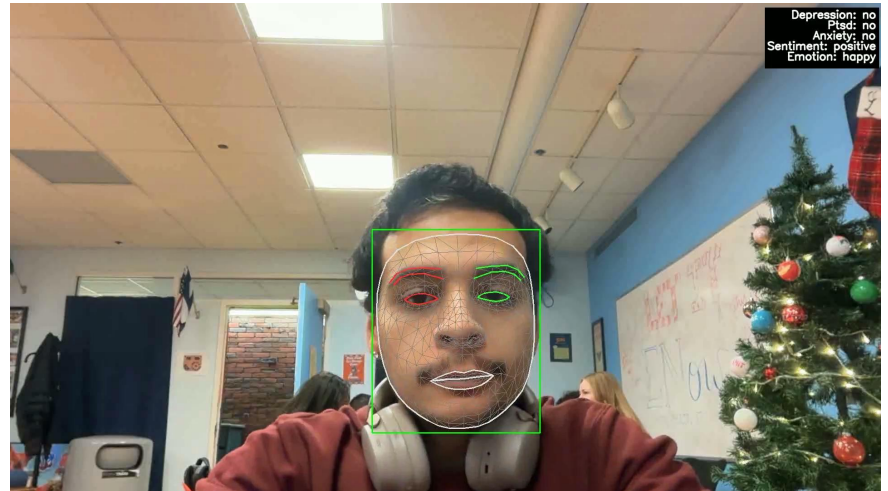
A More Real Example

- If we only have the video, how am I also getting mid level features?
 - We can't have humans providing this info as nice as it would be (subjective + painstaking)...
- So we'll use models to do this with instruction tuning!
 - Essentially tell a baseline model or one more trained for our purposes: "Here is video. Tell me what X is using only the following options: {1, 2, 3, 4...}"



A More Real Example

- Constructed a feature extraction pipeline for ados sessions to test if including these features as part of the dataset to train on → improved performance.
- Showcase of some extracted features from current results!
 - Top: ADOS video
 - Bottom: Result of doing feature extraction live





MAS S60: PyTorch & Huggingface Tutorial

Before class:

Register for an Huggingface & Wandb Account

Open Colabs below:

Pytorch: torch.dd.works

Huggingface: hf.dd.works

View Slides at s60.dd.works

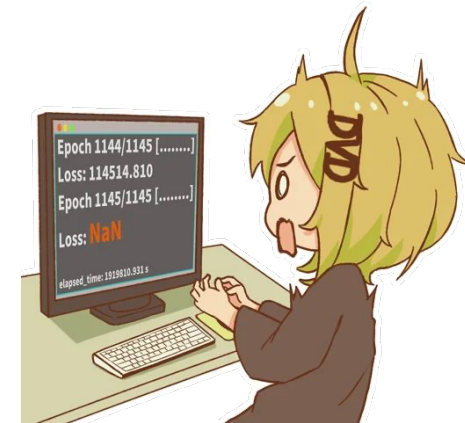
How to Design ML Models for New Data

- Look at the data first
- For simple, low dimensional data, start with simple models (SVM, Random Forest, Shallow MLP/CNN)
- For vision/language data, try pretrained model
- Start simple, then add complexity. Simple ones can be used as baselines.

How to Debug Your Model

- Look at the data first. Is the input data & label correct?
 - Ensure no data leakage;
- Look at the outputs. Is model only predicting one label?
 - Label imbalance: Data Augmentation; loss scaling
- Look at the training loss
 - Loss is nan: Inspect weights and inputs for NaN values. Make sure weights are initialized. LLM: Use bfloat16 instead of float16.
 - Loss not changing: Model underfitting. Increase learning rate; decrease weight decay; Add more complexity; Use better optimizer*.
- Understand every line of your code.

* Personal tip: I recommend trying second order optimizers from packages like [Heavyball](#)



How to Debug Your Model (Continued)

- Look at Loss (Continued)
 - Loss highly varied/increasing: Decrease learning rate; Gradient Clipping; Use better Optimizers
- Look at train vs val accuracy (or any other metrics)
 - Train \gg Val: Model overfitting. More weight decay, reduce model complexity, data augmentation, get more data
 - Train \approx Val \approx 100%: Check for data leakage
 -

