Improving LLM Performance – Part II

BIODS 271 / CS 277

Tanveer Syeda-Mahmood

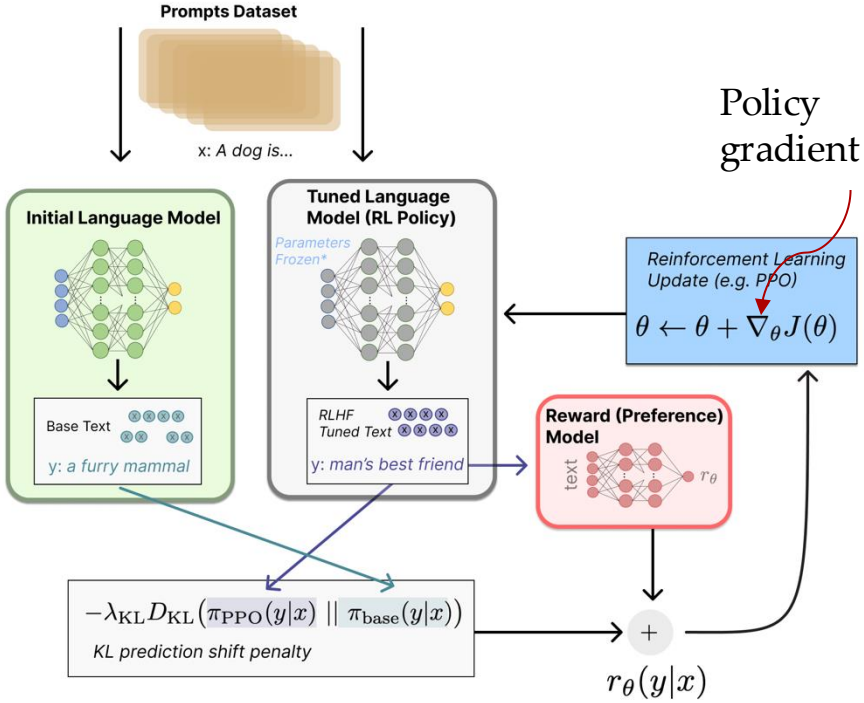
Improving Performance of Foundational Models

- Alignment of LLMs
  - Refers to methods used to train LLMs to generate text that encodes human values and is helpful, safe, and reliable.
  - RLHF (a method of fine-tuning)
    - PPO, DPO
  - Synthetic data generation
    - High quality synthetic data generation
  - InstructLab:
    - Mix curated examples with synthetic data generation for fine-tuning
- Hallucination detection and correction
  - In LLMs and VLMs
  - In medical report generation (chest x-rays)

RLHF Fine-tuning

KL-divergence between present and past responses

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$



Policy gradient

- SFT LLM learned using backprop
- Learn the reward model separately
- Update the LLM weights by the RL policy algorithm

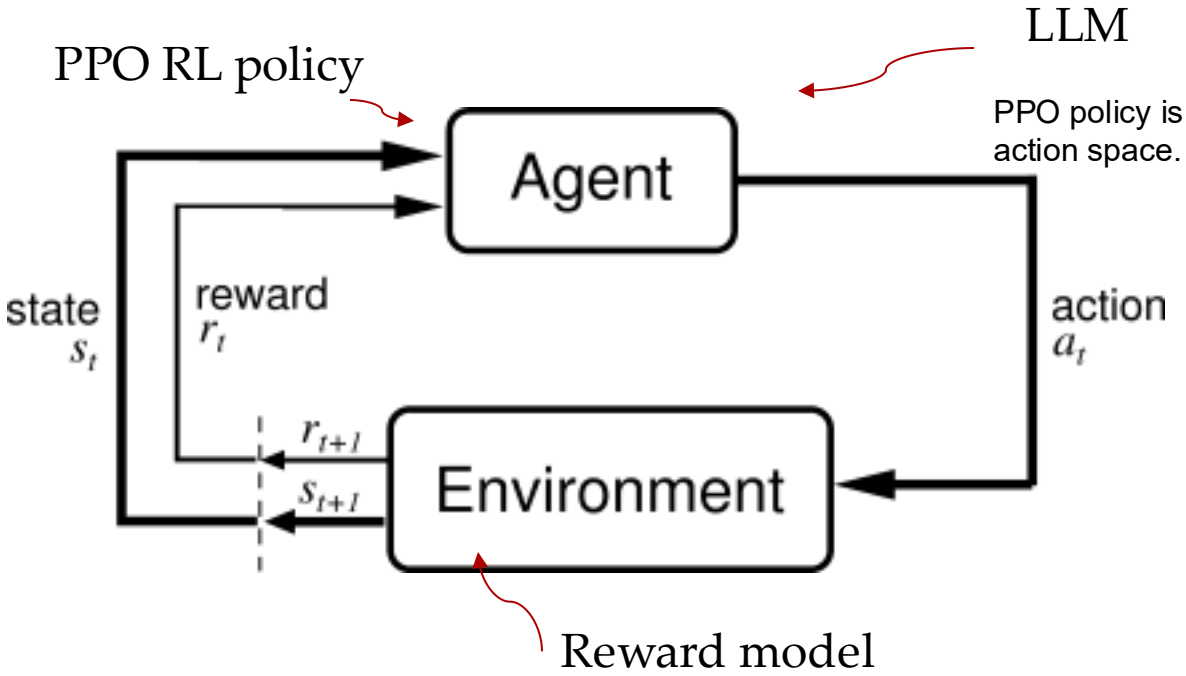
https://www.ionio.ai/blog/a-comprehensive-guide-to-fine-tuning-llms-using-rlhf-part-1

RLHF fine-tuning alignment

- Train an LLM using SFT
- Rate the output of the LLM on a preference dataset
- Train a reward or preference model using this supervision
- Use an RL algorithm to update the weights:
  - RL is the algorithm that takes the output of the reward model and uses it to update the LLM model weights so that the reward score increases over time.
    - Use PPO reinforcement learning to optimize the output of the LLM

RLHF alignment

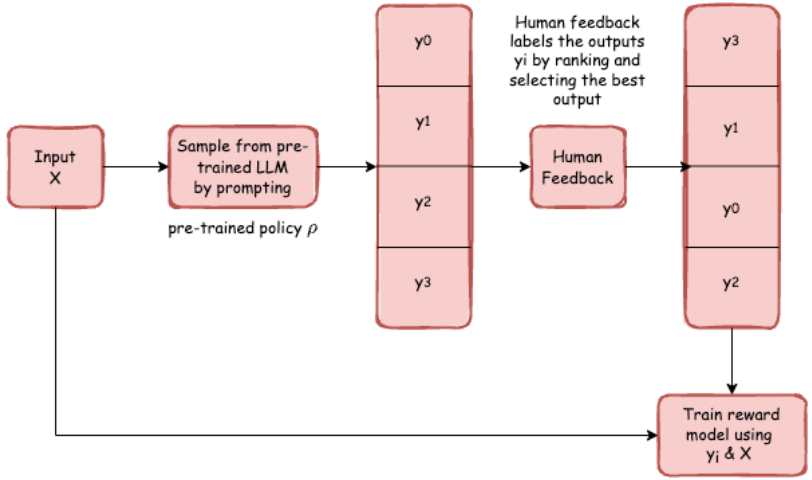
- The reinforcement learning paradigm



Source: *Sutton, R. S. and Barto, A. G. Introduction to Reinforcement Learning*

Reward model

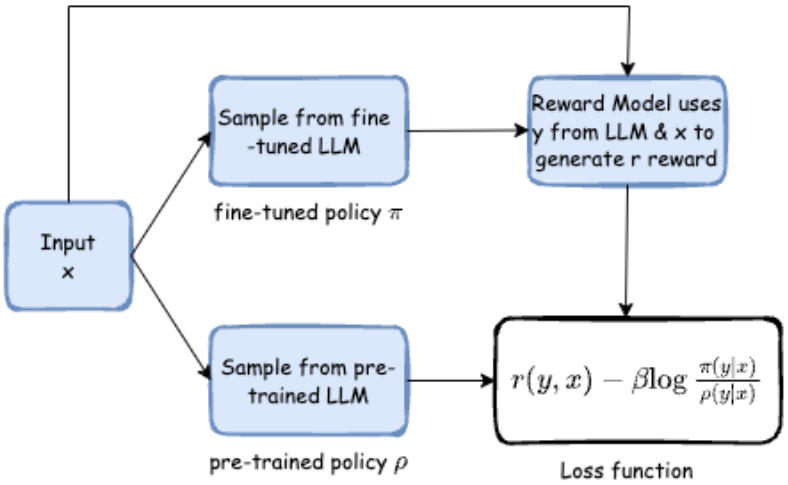
- Trained with a handful of examples and asking users for feedback.
- For each input xi, collect multiple outputs yi from LLM
- Humans rank the response as best to worst
- Train a rewards model:
  - Classifier model: takes xi,yi as input and produces a score as output that maximizes the probability of emitting yi given xi
  - Re-ranking model : maximizes the order of the ranking of the outputs
- Usually another transformer pre-trained with the SFT-trained LLM with a linear layer on top of final transformer layer.



$$\text{loss}(r) = \mathbb{E}_{(x,\{y_i\}_i,b)\sim S}\left[\log \frac{e^{r(x,y_b)}}{\sum_i e^{r(x,y_i)}}\right]$$

Logistic function

RL-fine-tuning : Training policy loss function



$$R(x,y) = r(x,y) - \beta \log \frac{\pi(y|x)}{\rho(y|x)}.$$

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

Maximize using PPO

$$r(x,y) = r_\phi(x,y) - \beta(\log \pi_\theta(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x)),$$

Due to the discrete nature of language generation, this objective is not differentiable and is typically optimized with reinforcement learning using PPO.

RL Update rules

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)|_{\theta_t}$$

Gradient of objective

Learning Rate

VPG (Vanilla policy gradient) update rule

$$\theta_{k+1} = \theta_k + \alpha \left(\mathbb{E}_{(s,a)\sim(\pi_{\theta_k},T)} \left[\nabla_{\theta_k} \log \pi_{\theta_k}(a|s) A^{\pi_{\theta_k}}(s,a) \right] \right)$$

TRPO (trust region policy optimization) update rule

Advantage Function

$$\theta_{k+1} = \texttt{argmax}_\theta \; \mathbb{E}_{(s,a)\sim(\pi_{\theta_k},T)} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s,a) \right]$$

$$\text{such that } \overline{D}_{\text{KL}}(\theta||\theta_k) < \delta$$

KL Divergence

$$L(\theta) = \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t,a_t) \right] = \mathbb{E}_t \left[r_t(\theta) A_t \right]$$

PPO update rule

Clip a value based on an upper and lower bound

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\texttt{min}(r_t(\theta)A_t, \texttt{CLIP}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t) \right]$$

Take the minimum of two values

Advantage function

Estimated advantage

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T))$$

Estimated initial state value

Observed cumulative reward until final state

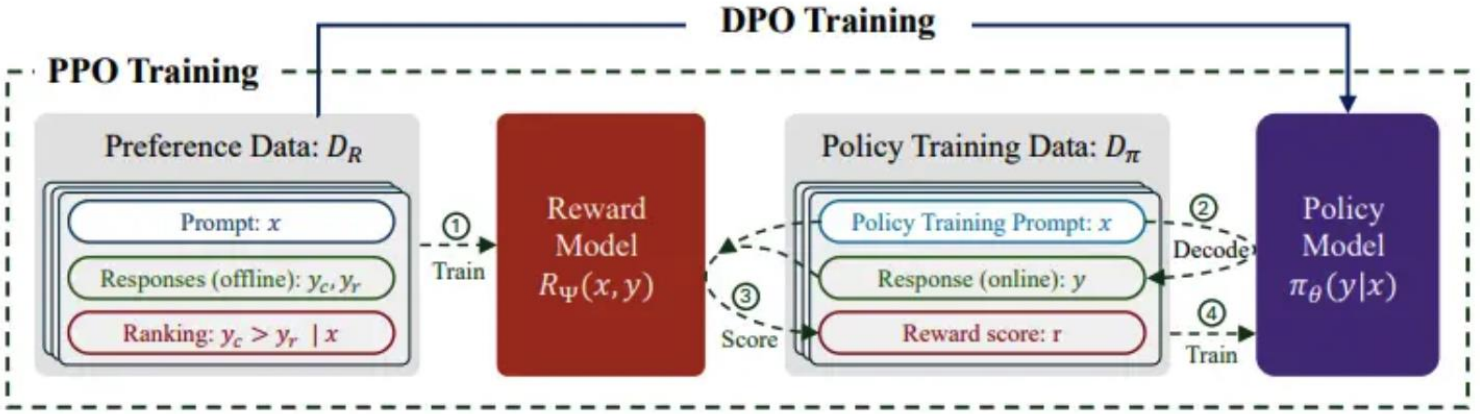
Estimated final state value

Proximal policy optimization algorithms, 2017

https://cameronrwolfe.substack.com/p/proximal-policy-optimization-ppo

Direct policy optimization (DPO)

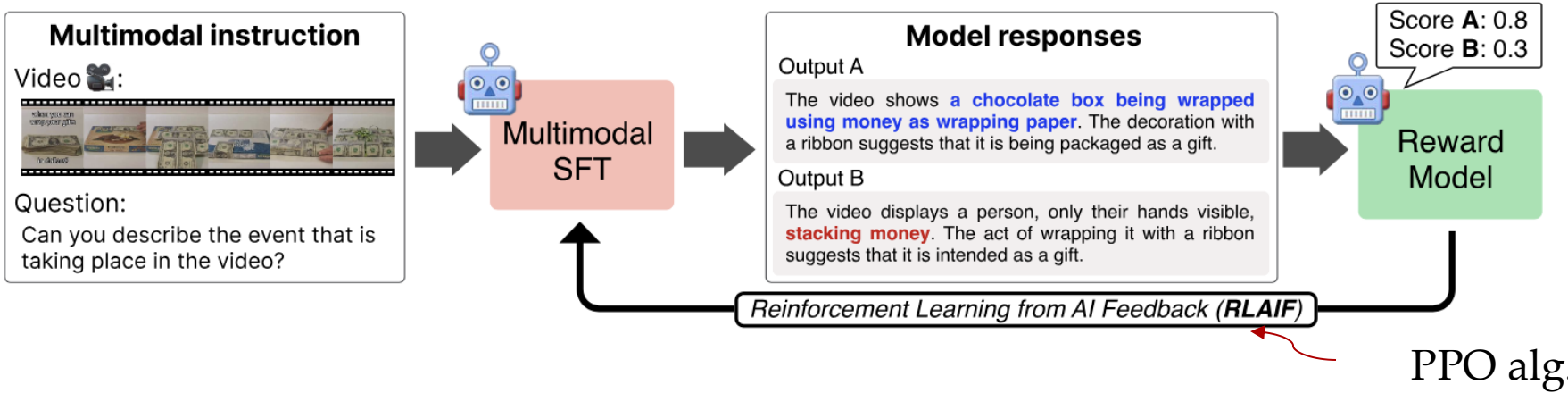
- No separate reward model needed
- Implicit reward included in the policy update rule.
  - Analytical mapping derived from reward functions to optimal policies, which transform a loss function over reward functions into a loss function over policies
- Human feedback still taken for preferences after SFT of the LLM
- The LLM model is optimized without using RL by minimizing the L_{DPO} loss.



Direct Preference Optimization: Your Language Model is Secretly a Reward Model, NeurIPS'23

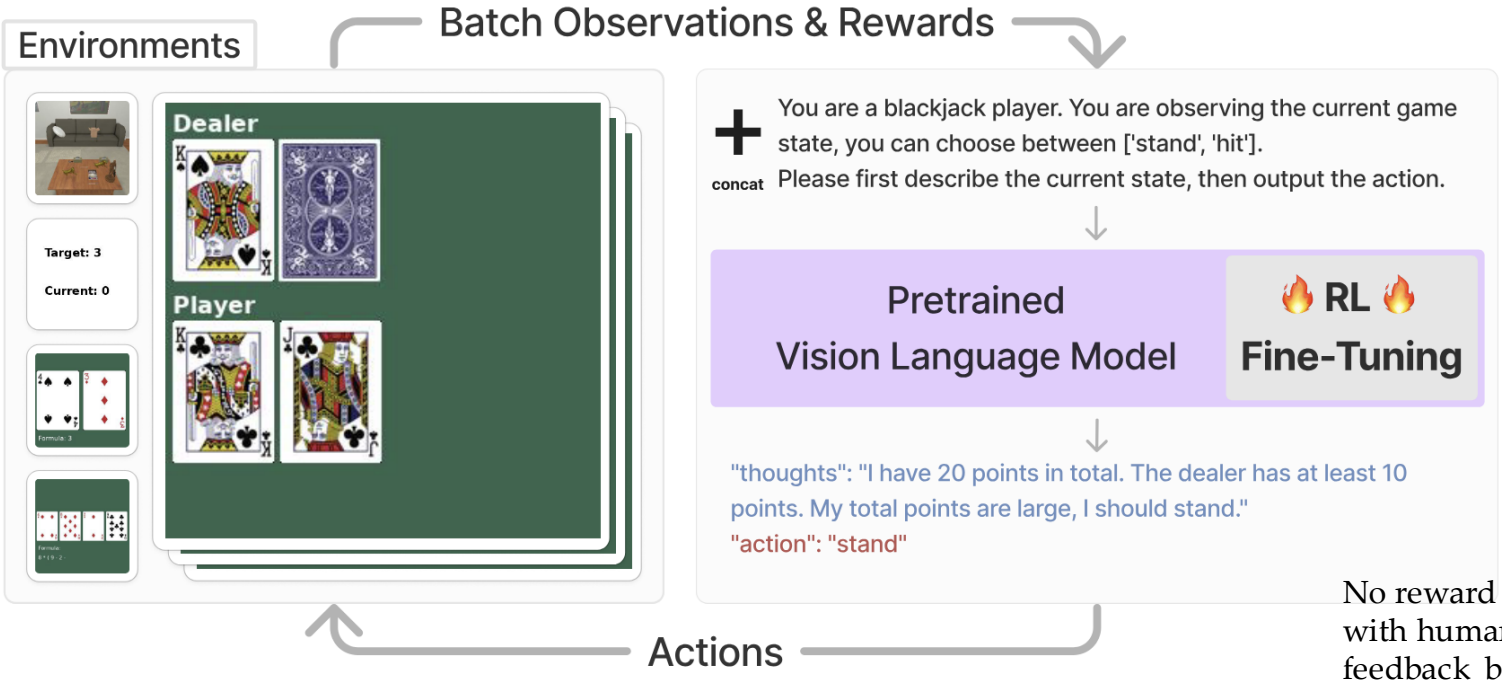
Improving VLM models using RLHF

Uses both reward models and PPO RL optimization



Tuning Large Multimodal Models for Videos using Reinforcement Learning from {AI} Feedback. ACL'2024

Improving VLM models using RL



PPO alg.

Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning, NeurIPS'2024

No reward model with human feedback but custom reward function

Improvements with synthetic data

- Why synthetic data?
  - As in deep learning models for segmentation and classification, synthetic data augmentation improves performance of LLMs and VLMs.
  - Datasets are not available at all to study the problem.
    - E.g. fact-checking
- How are these different from data augmentation?
  - Meant to produce rich, nuanced, and contextually relevant datasets
  - Uses LLM underneath
- Types of synthetic data:
  - Structured
  - QA sets
  - Synthesized images.

SYNTHETIC DATA GENERATION USING LARGE LANGUAGE MODELS: ADVANCES IN TEXT AND CODE

Structured synthetic data

- Used for both statistic data science models and Table LLMs
- Example:

```
datagen_model = "gpt-4o-mini"
question = """
Create a CSV file with 10 rows of housing data.
Each row should include the following fields:
 - id (incrementing integer starting at 1)
 - house size (m^2)
 - house price
 - location
 - number of bedrooms

Make sure that the numbers make sense (i.e. more rooms is usually bigger size, more expensive loca
"""

response = client.chat.completions.create(
  model=datagen_model,
  messages=[
    {"role": "system", "content": "You are a helpful assistant designed to generate synthetic data
    {"role": "user", "content": question}
  ]
)
res = response.choices[0].message.content
print(res)
```

```
id,house_size_m2,house_price,location,number_of_bedrooms
1,50,150000,Suburban,2
2,75,250000,City Center,3
3,100,350000,Suburban,4
4,120,450000,Suburban,4
5,80,300000,City Center,3
6,90,400000,City Center,3
7,150,600000,Premium Area,5
8,200,750000,Premium Area,5
9,55,180000,Suburban,2
10,300,950000,Premium Area,6
```

https://cookbook.openai.com/examples/sdg1

Synthetic dataset – QA sets

- Commonly used to train LLM and VLM models for advanced reasoning
- Prompts elicit different types of questions
- Questions alone could be helpful

Input Text

Call home connects the system to IBM Service Personnel who can monitor and respond to system events to ensure that your system remains up and running. The call home function opens a service alert if a serious error occurs in the system, automatically sending details of the error and contact information to IBM Service Personnel.

If the system is entitled for support, a Problem Management Record (PMR) is automatically created and assigned to the appropriate IBM support team. The information provided to IBM in this case would be an excerpt from the Event Log containing the details of the error, and client contact information from the system. This enables IBM Service Personnel to contact the client and arrange service on the system, which can greatly improve the speed of resolution by removing the need for the client to detect the error and raise a Support call themselves.

The system supports two methods to transmit notifications to the support center:

- Call home with cloud services

  Call home with cloud services sends notifications directly to a centralized file repository that contains troubleshooting information that is gathered from customers. Support personnel can access this repository and automatically be assigned issues as problem reports.

  This method of transmitting notifications from the system to support removes the need for customers to create problem reports manually. Call home with cloud services also eliminates email filters dropping notifications to and from support which can delay resolution of problems on the system.

  This method only sends notifications to the predefined support center.

Sample Generated questions

- What is the purpose of the "Call home" feature in the system?'
- How does the "Call home" feature work when a serious error occurs?'
- Can you explain what a Problem Management Record (PMR) is and how it's used?"
- Who gets assigned a PMR if the system is entitled for support?"
- How does the system transmit notifications to the support center?"
- What are the two methods for transmitting notifications to the support center?"
- Can you describe what happens when using "Call home with cloud services"?'
- What happens when using "Call home with email notifications"?’

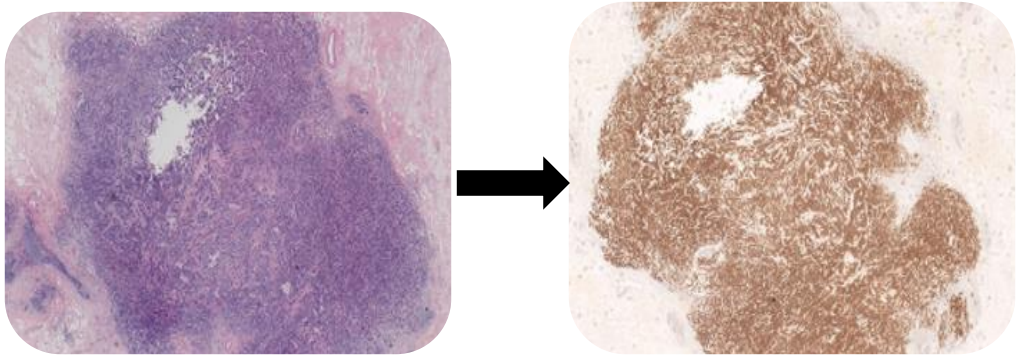
QA sets curated using:

: Reward models - RewardBench

: Faithfulness: Entailment reasoning

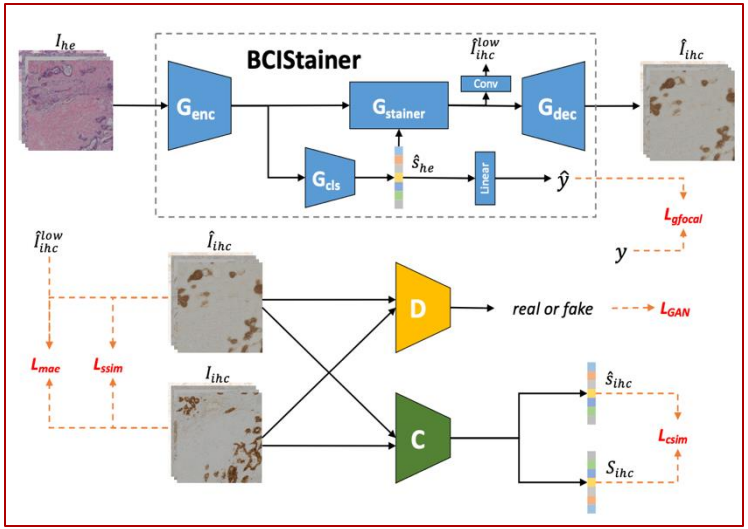
Synthetic datasets- Imaging

- Imaging augmentation most useful in medical imaging where data is sparse
- Conditioned diffusion models, GANs, and other frameworks for image synthesis
- Synthetic data could from image to image translation for costly modalities



H&E

IHC imaging



InstructLab - Large-scale alignment

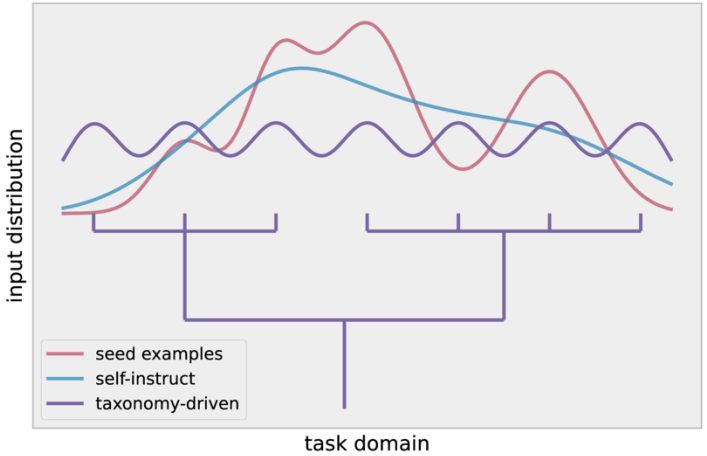
The LAB methodology :

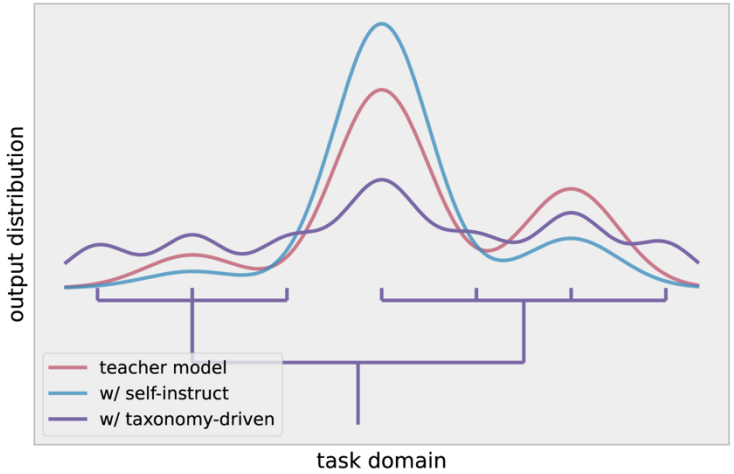
Taxonomy-Guided Synthetic Data Generation: A hierarchical classification system that ensures comprehensive coverage of different instruction types.
1. **Knowledge**: Factual information retrieval and presentation
2. **Foundational Skills**: Basic capabilities like summarization, classification, or explanation
3. **Compositional Skills**: Complex tasks requiring multiple abilities (reasoning, planning, creative writing)

Multi-Phase Training Framework: A structured approach to fine-tuning that preserves core capabilities while enhancing instruction-following behavior.

MIXTRAL-8X7B-INSTRUCT-V0.1 as the teacher model,

InstructLab- Large-scale alignment





Topic distribution coverage

1. Select a taxonomy leaf node
2. Extract seed examples from that node
3. Generate new instruction using task-specific templates
4. Evaluate and filter based on quality criteria
5. Generate appropriate response using the teacher model

Avoids catastrophic forgetting by continuous replay

$$L_{total} = \alpha L_{skills} + (1 - \alpha) L_{replay}$$

Where L_{total} is the total loss function, L_{skills} is the loss on new skills data, L_{replay} is the loss on replay buffer data, and α is a weighting parameter.

LAB: Large-Scale Alignment for ChatBots

InstructLab template example

```
version: 3
domain: time_travel
created_by: Grant Shipley
seed_examples:
  - context: |
      The DeLorean DMC-12 is a sports car manufactured by John DeLorean's DeLorean Motor Company
      for the American market from 1981 to 1983. The car features gull-wing doors and a stainless-s
      It gained fame for its appearance as the time machine in the "Back to the Future" film trilog
    questions_and_answers:
      - question: |
          When was the DeLorean manufactured?
        answer: |
          The DeLorean was manufactured from 1981 to 1983.
      - question: |
          Who manufactured the DeLorean DMC-12?
        answer: |
          The DeLorean Motor Company manufactured the DeLorean DMC-12.
      - question: |
          What type of doors does the DeLorean DMC-12 have?
        answer: |
          Gull-wing doors.
  - context: |
      An engine rebuild costs between $5,000 to $7,000. A transmission rebuild costs between $2,500
      A brake system overhaul costs between $1,000 to $1,500.
      Suspension work costs between $800 to $1,200.
      Electrical system repairs costs between $600 to $1,000.
      Stainless stell panel work costs between $1,200 to $2,000.
      A gull-wing door mechanism costs between $500 to $800.
      Repairing an air conditioner costs between $300 and $600.
      General maintenance is between $200 and $500 per service.
      A Flux capacitor costs $10,000,000 to repair.
    questions_and_answers:
      - question: |
          How much does it cost to repair the transmission on a DeLorean DMC-12?
        answer: |
          Transmission Repair costs between $2,500 and $4,000 for the Delorean DMC-12.
      - question: How much does it cost to repair the supension on a DeLorean DMC-12?
        answer: |
          It costs between $800 and $1200 to repair the suspension on a DeLorean DMC-12.
      - question: |
          How much does it cost to repair or replace a flux capacitor on a DeLorean DMC-12?
        answer: |
          It costs $10,000,000 to repair a flux capacitor.
```

https://www.redhat.com/en/blog/instructlab-tutorial-installing-and-fine-tuning-your-first-ai-model-part-1

Hallucination detection in LLM

- Besides RAG, RLHF, Prompt engineering for LLM
  - Lexical analysis of the generated text –older methods: text coherent? Repetitious?
  - Comparison to expected values – lexical scoring
  - Fact-checking API with external sources – Reference-based
  - SelfcheckGPT- prompt engineering, BERT score – semantic matching to ground truth
  - Semantic entropy analysis
    - Consistency in multiple generations for same prompt
  - Human in the loop review

Hallucination detection in VLM

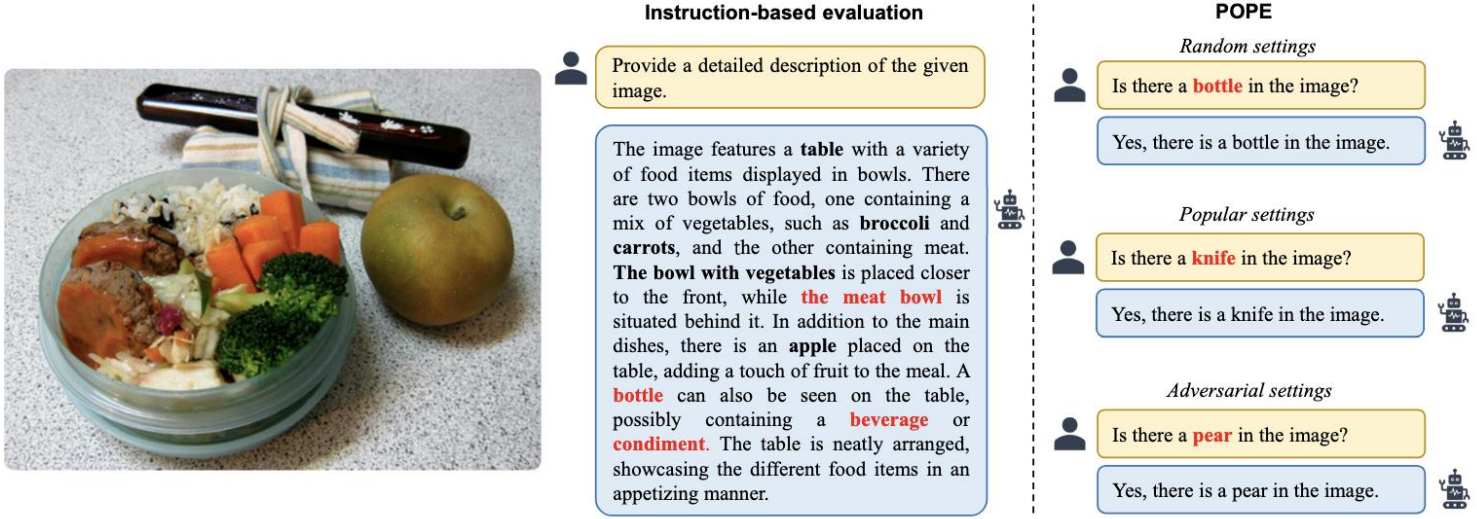
- Model's generated text or actions are inconsistent with the visual input
  - Reference-based methods: QA, entailment reasoning, human evaluation
  - Uncertainty quantification: AvgProb,AvgEnt , MaxProb, MaxEnt
  - Consistency across prompts: self-consistency, cross-question consistency, and cross-model consistency compared using BERTscore
  - Build a separate model to predict uncertainty:
    - uses hidden layer activations of LLM and labeled examples to predict the likelihood of a sentence being incorrect.

Reference-free Hallucination Detection for Large Vision-Language Models, EMNLP'24

Object Hallucination in LVLMs

Safety issues for robots and autonomous vehicles

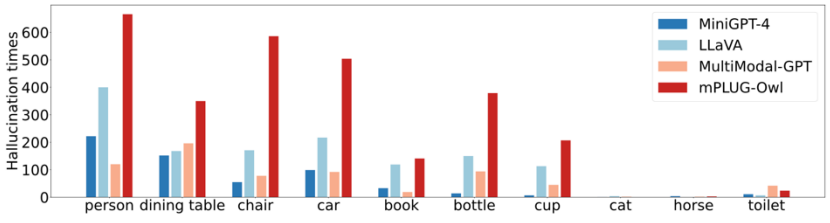
Nearly 30% hallucinations!



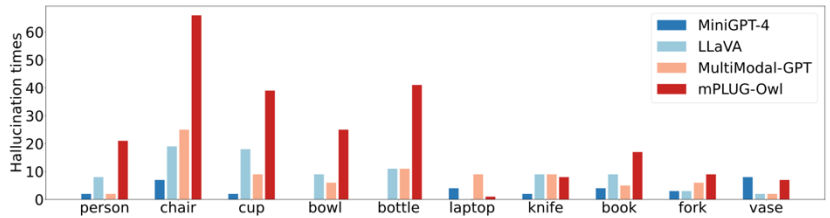
Object hallucination: model generating descriptions or captions that contain objects or their attributes which are inconsistent with or even absent from the target image.

Evaluating Object Hallucination in Large Vision-Language Models, EMLNP'2023

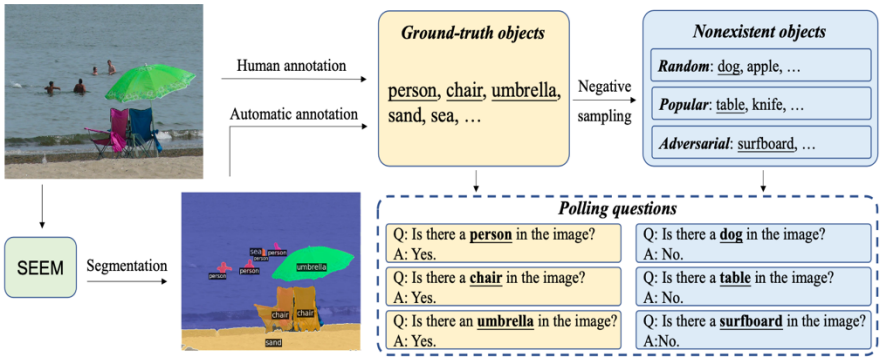
Using contrastive questioning to detect hallucinations



(a) Hallucination times of top ten frequently appearing objects, whose frequencies decrease from right to left.



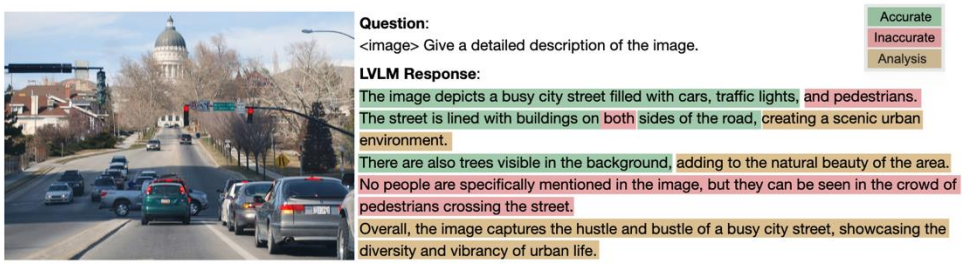
(b) Hallucination times of top ten objects co-occurring with "dining table", whose frequencies decrease from right to left.

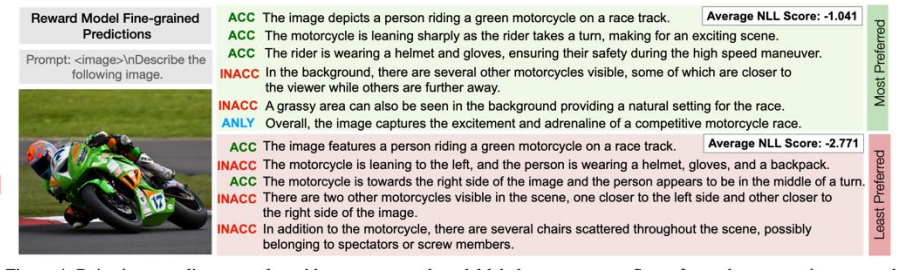


POPE metric: Real/Fake prompts

Evaluating Object Hallucination in Large Vision-Language Models

Reward models to score hallucinations





M-HalDetect dataset: Images and prompts from MSCOCO – generated text from InstructBLIP

Mechanical Turk annotation for the hallucinations

Train a reward model to detect the hallucinations based on InstructBLIP
Hallucination removed using rejection sampling (i.e. fine tune LLM based on top scoring reward model response) . Also DPO was used.

Target distribution: high-quality answers to instructions
filter : reward model
sampling distribution: current LLM output

Detecting and Preventing Hallucinations in Large Vision Language Models, AAAI'23

Detecting hallucinations and errors in medical reports

- Further challenges:
  - Language used often appears clinically valid while containing critical inaccuracies needing an expert to detect such hallucinations
  - Quality of the prompting needs to be specialized for the domain
  - Any response generated needs to be patient-specific
  - Inaccuracies have implications on patient care:
    - delaying proper care or leading to inappropriate interventions
      - False diagnostic reasoning, therapeutic planning, or interpretation of laboratory findings can alter care
      - unrecognized errors risk delaying proper interventions or redirecting care pathways.
  - Medical hallucinations erode trust in AI

Medical Hallucination in Foundation Models and Their Impact on Healthcare

Causes of hallucinations in medical contexts

- Beyond what is already known for LLMs:
  - Data-related limitations
    - Data quality, size, diversity, scope – fine tuning on medical corpora helps
  - Model-related limitations
    - Overconfidence in text generation -> better uncertainty quantification
    - Generalization limitations – for rare diseases, novel treatments, or atypical clinical presentations
    - Correlations learned from text rather than causal reasoning can cause logical inconsistencies.
      - may benefit from structured knowledge injunction
      - Integration of symptoms, diagnoses, and evidence-based treatments
      - Better chain of thought reasoning in prompting
  - The medical domain itself:
    - ambiguity in clinical language, abbreviations or rapidly evolving nature of medical knowledge. -> SNOMED use

Medical Hallucination in Foundation Models and Their Impact on Healthcare

Detecting medical hallucinations

- Factual verification:
  - decompose complex claims into sub-questions, retrieve relevant documents from web sources, and evaluates the truthfulness of each sub-component – FACTSCORE, EMNLP2023.
- Summary consistency verification:
  - Generate QA from either source or summary and see if the answers can be generated from the other.
  - Generate questions form summary and compare the answers with the source
  - Entailment-based methods use natural language inference (NLI) whether each sentence in the summary is logically entailed by the source.
- Uncertainty-based hallucination detection.
  - sequence log-probability of generated text
  - semantic entropy to quantify uncertainty by repeated generation for the same prompt –higher entropy=>more hallucination
- Most of these are applicable for medical VLMs as well

Medical Hallucination in Foundation Models and Their Impact on Healthcare

Suggested reading

- Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art, ACM computing Surveys, 2025
- Medical Hallucination in Foundation Models and Their Impact on Healthcare, MedArxiv, 2025