

CSE/DSC 234

Data-Centric AI and AI Engineering

Arun Kumar

Topic 1: Creating Data and Evals

Outline

- ❖ Motivation for Data-Centric AI
- ❖ Creating Data
 - ❖ Human Feedback and Annotation
 - ❖ Data Curation and Structure
-  ❖ Data Synthesis
 - ❖ Multimodal Data
- ❖ Creating Evals

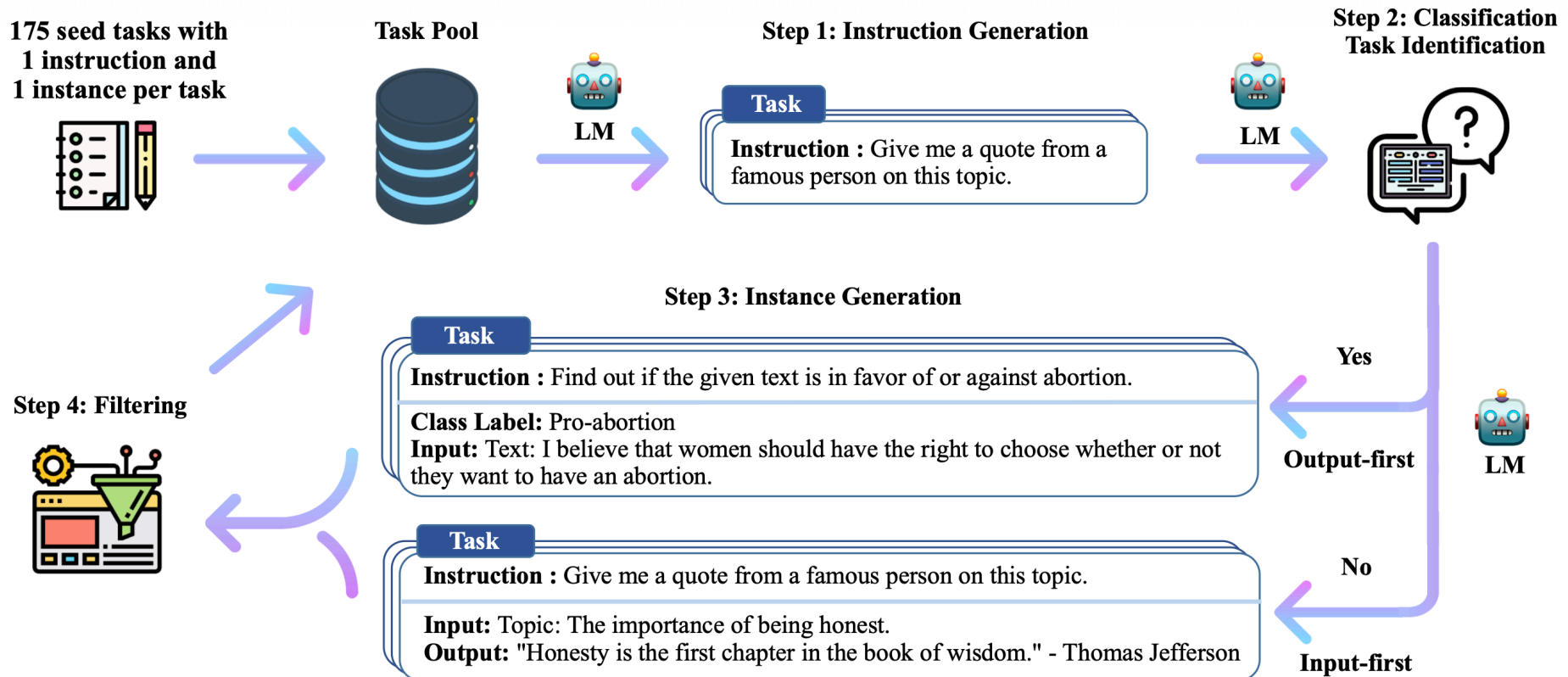
Why Synthetic Data?

- ❖ Human annotation is expensive, slow, may have limited diversity
 - ❖ InstructGPT: ~13K demo examples; does not scale to millions
- ❖ **Synthesis:** LLM-assisted generation/refinement of labeled data for their own use!
- ❖ **Advantages:**
 - ❖ Cheap, fast
 - ❖ Target-customizable; can reach complexity humans find hard
- ❖ **Risks:**
 - ❖ Quality control is harder
 - ❖ “Model collapse”
 - ❖ Bias amplification

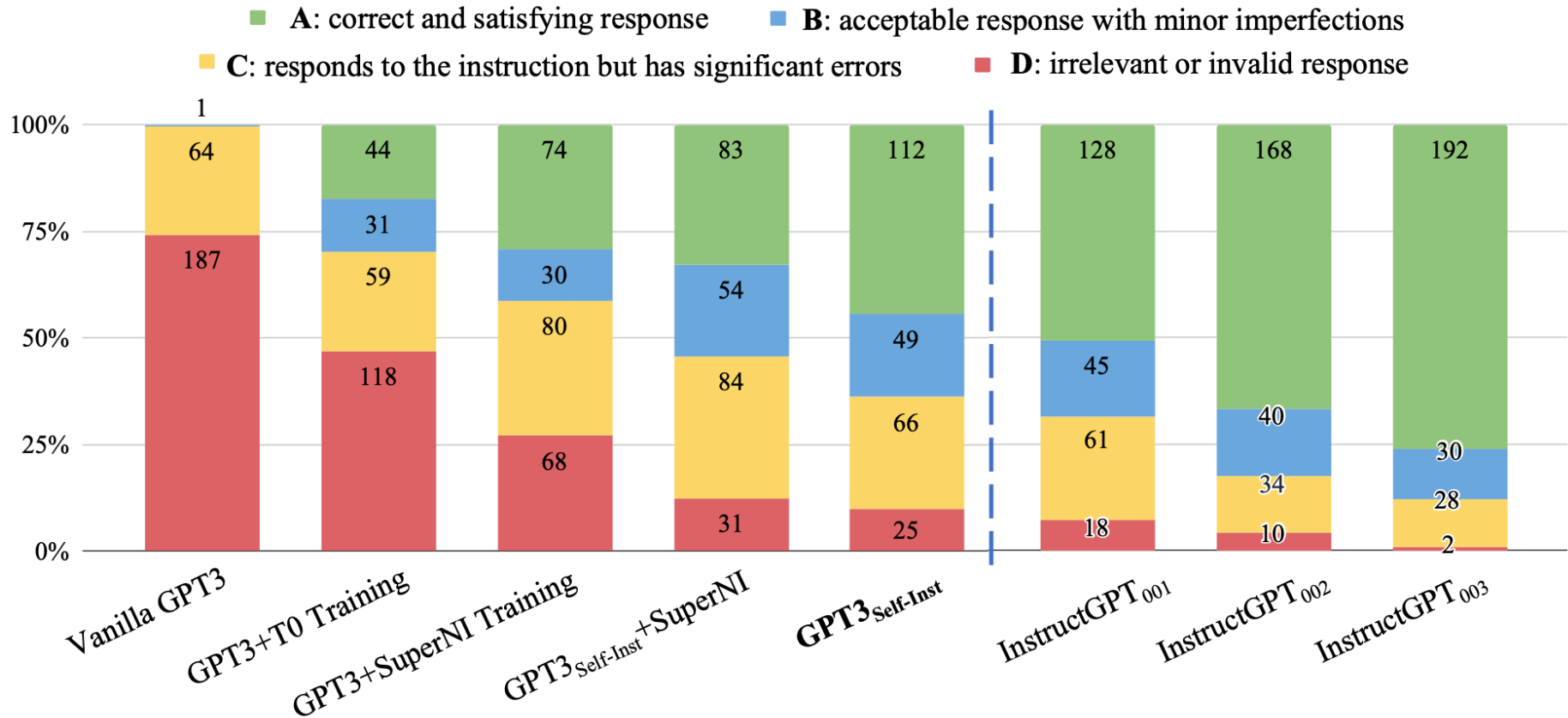
Pivot Point: Self-Instruct

- ❖ First work showing LLMs can bootstrap their own instruction-tuning data
 - ❖ 175 human-written seed tasks -> LLM generates new instructions + input-output pairs -> filter -> fine-tune
 - ❖ 52K instructions, 82K instances suffices for GPT-3 + Self-Instruct to be comparable to InstructGPT-001!
- ❖ Stanford Alpaca: same idea with GPT-3.5 + LLaMA 7B for under \$600; democratized chat models
- ❖ But: generated instructions skew simple and repetitive; heavy on classification, light on reasoning

Pivot Point: Self-Instruct



Pivot Point: Self-Instruct



Self-Instruct: Aligning Language Models with Self-Generated Instructions. ACL 2023

<https://arxiv.org/abs/2212.10560>

Pivot Point: Self-Instruct

```
Input:
Address: 123 Main Street, City: San Francisco ⓘ
Output: 94105
```

Input:
Application Form:
Name: _____ Age: _____ Sex: _____
Phone Number: _____ Email Address: _____
Education: _____ ...



```
Input: Null
Output:
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

Input: Null
Output:

Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

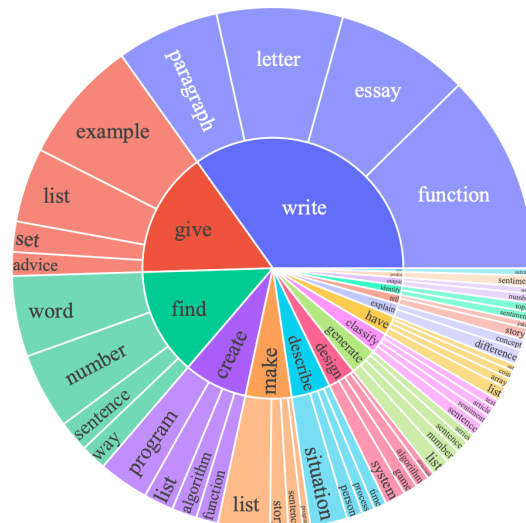


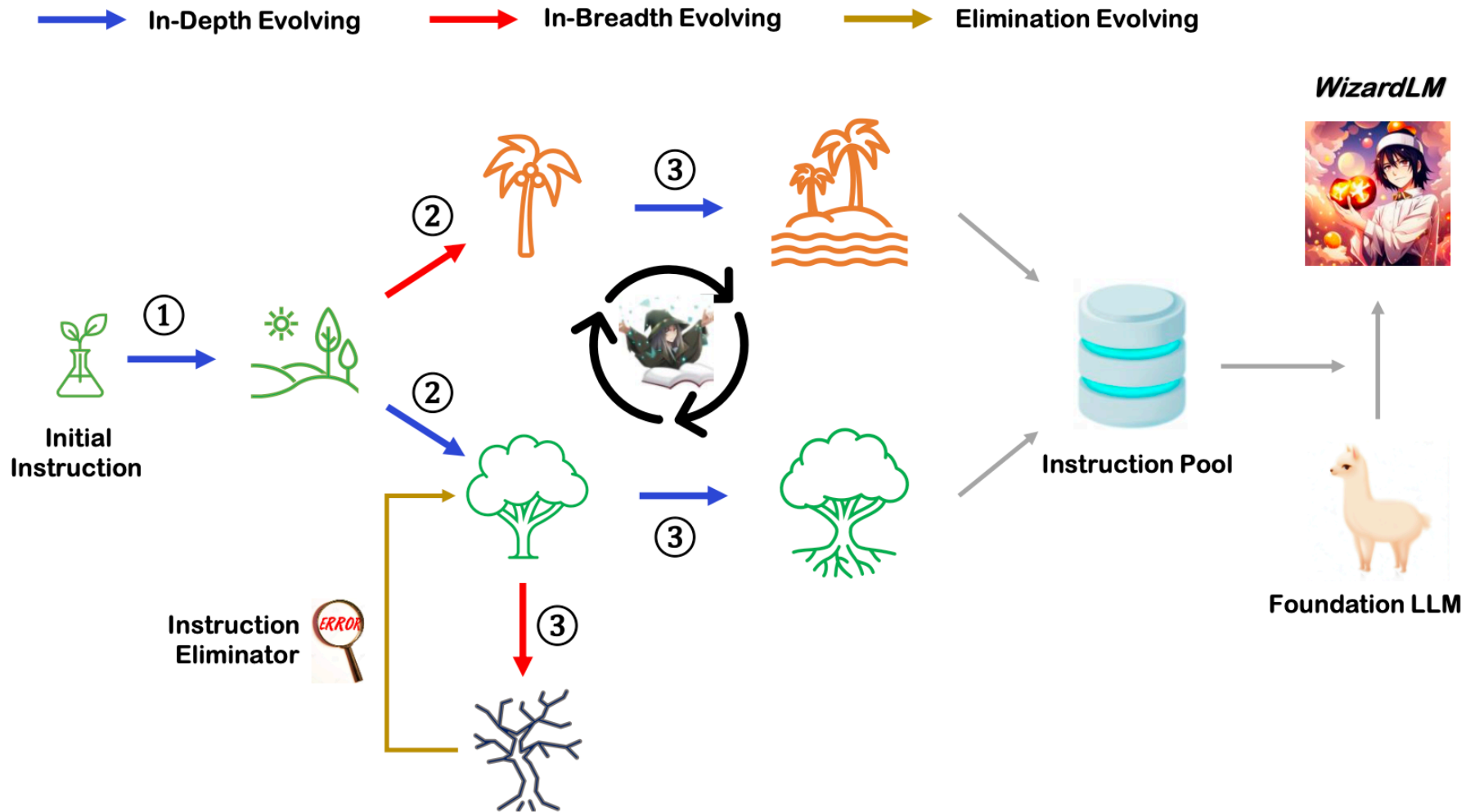
Figure 3: The top 20 most common root verbs (inner circle) and their top 4 direct noun objects (outer circle) in the generated instructions. Despite their diversity, the instructions shown here only account for 14% of all the generated instructions because many instructions (e.g., “Classify whether the user is satisfied with the service.”) do not contain such a verb-noun structure.

Quality Review Question	Yes %
Does the instruction describe a valid task?	92%
Is the input appropriate for the instruction?	79%
Is the output a correct and acceptable response to the instruction and input?	58%
All fields are valid	54%

Case Study: Evol-Instruct/WizardLM

- ❖ Directly tackles the key issue with Self-Instruct, viz., simple, repetitive instructions
- ❖ **Evol-Instruct:** Iteratively rewrite instructions to raise complexity
 - ❖ **In-depth:** Add constraints, deepen reasoning, complicate input
 - ❖ **In-breadth:** Mutate into *new* instructions to raise diversity
 - ❖ Evolution operators are prompt-engineered, not learned
 - ❖ 52K seeds -> 4 epochs -> WizardLM
- ❖ On hard instructions, preferred over ChatGPT!
 - ❖ In-depth was more impactful than in-breadth for quality

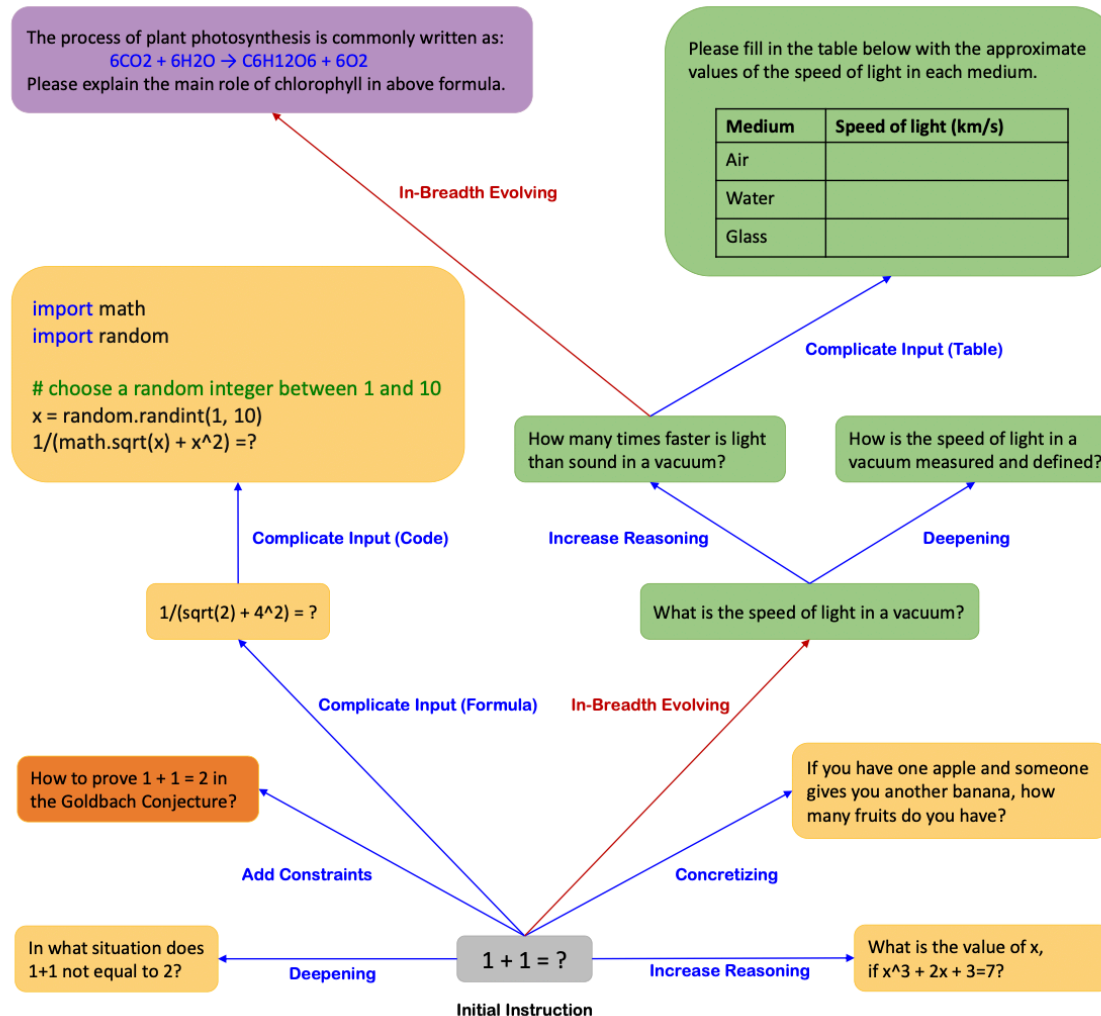
Case Study: Evol-Instruct/WizardLM



WizardLM: Empowering Large Language Models to Follow Complex Instructions. ICLR 2024

<https://arxiv.org/abs/2304.12244>

Case Study: Evol-Instruct/WizardLM



Case Study: Evol-Instruct/WizardLM

Example 3.1: Prompt for Adding Constraints of In-Depth Evolving

I want you act as a Prompt Rewriter.

Your objective is to rewrite a given prompt into a more complex version to make those famous AI systems (e.g., ChatGPT and GPT4) a bit harder to handle. But the rewritten prompt must be reasonable and must be understood and responded by humans.

Your rewriting cannot omit the non-text parts such as the table and code in #Given Prompt#:.. Also, please do not omit the input in #Given Prompt#.

You SHOULD complicate the given prompt using the following method:

Please add one more constraints/requirements into #Given Prompt#

You should try your best not to make the #Rewritten Prompt# become verbose, #Rewritten Prompt# can only add 10 to 20 words into #Given Prompt#. ‘#Given Prompt#’, ‘#Rewritten Prompt#’, ‘given prompt’ and ‘rewritten prompt’ are not allowed to appear in #Rewritten Prompt#

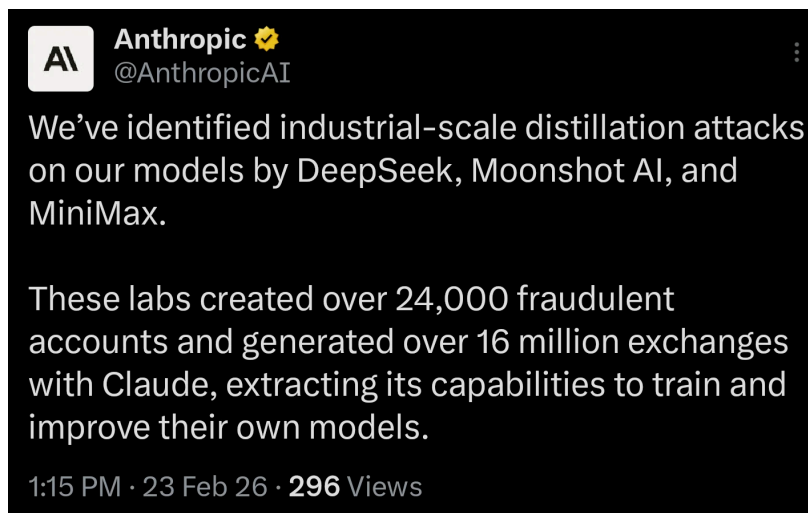
#Given Prompt#:

{Here is instruction.}

#Rewritten Prompt#:

Distillation as a Source of Data

- ❖ Use frontier model as “teacher” to generate data for student
 - ❖ Alpaca: GPT-3.5 -> 52K examples -> LLaMA 7B
 - ❖ Orca: GPT-4’s chain-of-thought -> student learns reasoning
- ❖ **Caveat:** Terms of Service often prohibit training “competitors” on outputs of frontier models!
 - ❖ Open LLMs also increasingly used as teachers



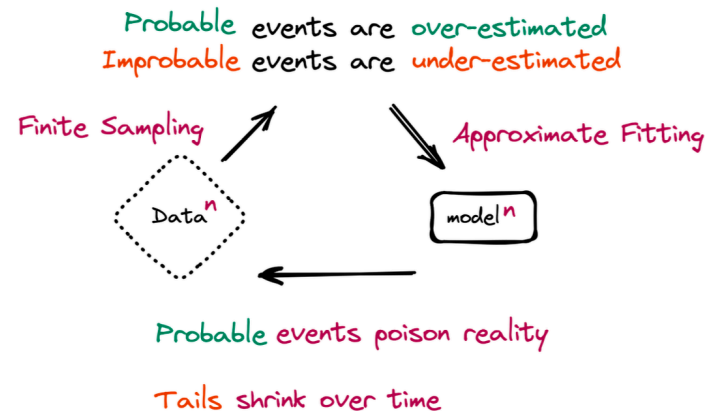
Distillation as a Source of Data



Model Collapse Phenomenon

- ❖ Generative AI models of successive generations being trained on their own outputs -> diversity collapses, tails disappear!
- ❖ **Typical prevention/mitigation:**
 - ❖ Track provenance and mix synthetic with high-quality “real” (human-generated) data
 - ❖ Monitor for diversity metrics
- ❖ Area of active research; still no comprehensive general solution

Model Collapse Phenomenon



(a) Original model



(b) Generation 5



(c) Generation 10



(d) Generation 20

Figure 9: Random latent reconstructions from VAEs. No training data comes from the original distribution. Over the generations, different modes of the original distribution get entangled and generated data starts looking unimodal.

Scaling Diversity in Synthetic Data

- ❖ Single prompt template leads to narrow distribution; need to vary context systematically
- ❖ **Strategies to diversify:**
 - ❖ **Persona-driven:** “You are a biology professor”, “You are a teen”, etc.
 - ❖ **Domain-targeted:** code, math, medical, legal
 - ❖ **Taxonomy-driven:** enumerate knowledge areas, generate per area
 - ❖ **Dimension-based generation:** Define dimensions of variation (feature x persona x scenario), generate tuples, then expand each tuple into realistic queries

Quality Filtering on Synthetic Data

- ❖ Raw synthetic output is usually noisy; must filter aggressively
- ❖ **Typical Approaches:**
 - ❖ **Rule-based:** length, coherence, regex, etc.
 - ❖ **Model-based:** LLM-as-judge scoring; fine-tuned SLMs
 - ❖ **Embedding-based** deduplication
- ❖ **Practical guideline:** Generate ~100K examples -> filter to ~10K; remember: curation > volume
 - ❖ Example: LIMA (Zhou et al., NeurIPS 2023): 1,000 curated examples matched 50K noisy ones

Outline

- ❖ Motivation for Data-Centric AI
- ❖ Creating Data
 - ❖ Human Feedback and Annotation
 - ❖ Data Curation and Structure
 - ❖ Data Synthesis
- ➔ ❖ Multimodal Data
- ❖ Creating Evals

Multimodal Data

- ❖ Many applications now need *joint modalities-aware* annotations: image+text, image+text+tabular, video+text, etc.

Learning Transferable Visual Models From Natural Language Supervision

15

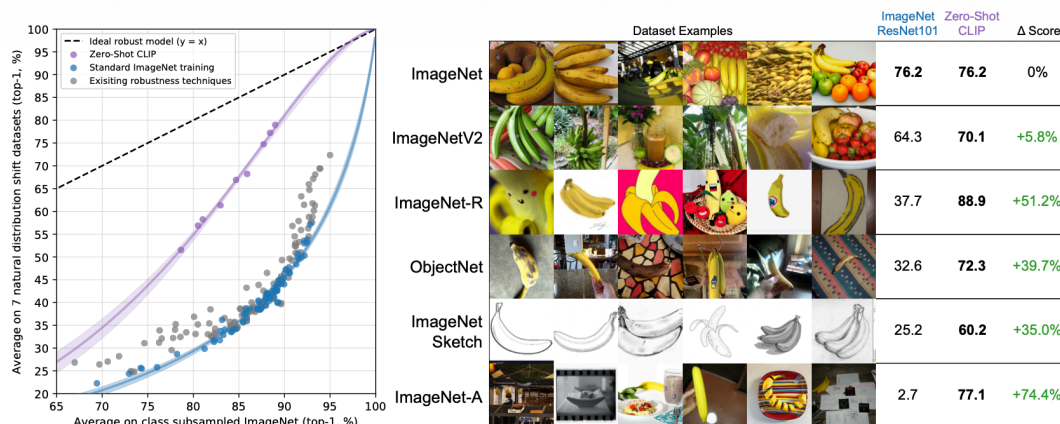


Figure 13. Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models. (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this “robustness gap” by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

ImageNet-R, 3.8% on ObjectNet, 2.8% on ImageNet Sketch, and 1.9% on ImageNet-A. The change in accuracy on the two other datasets, Youtube-BB and ImageNet Vid, is insignificant.

How is it possible to improve accuracy by 9.2% on the Im-

pooling predictions across all sub-classes according to the ImageNet class hierarchy. Sometimes this mapping is much less than perfect. For the person class in Youtube-BB, predictions are made by pooling over the ImageNet classes for a baseball player, a bridegroom, and a scuba diver. With CLIP we can instead generate a custom zero-shot classi-

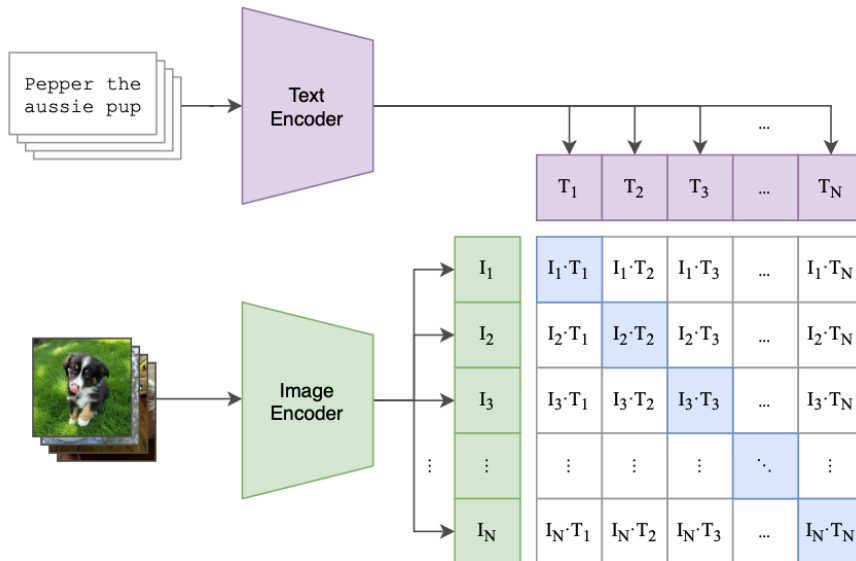
Multimodal Data

- ❖ **Key challenges for multimodal annotations:**
 - ❖ Curation across disparate modalities harder to commoditize
 - ❖ Quality (e.g., image alt-text usually too coarse)
 - ❖ Diversity across concepts/styles/languages tougher to ensure
 - ❖ Noise and inconsistencies
 - ❖ Scaling to large corpora

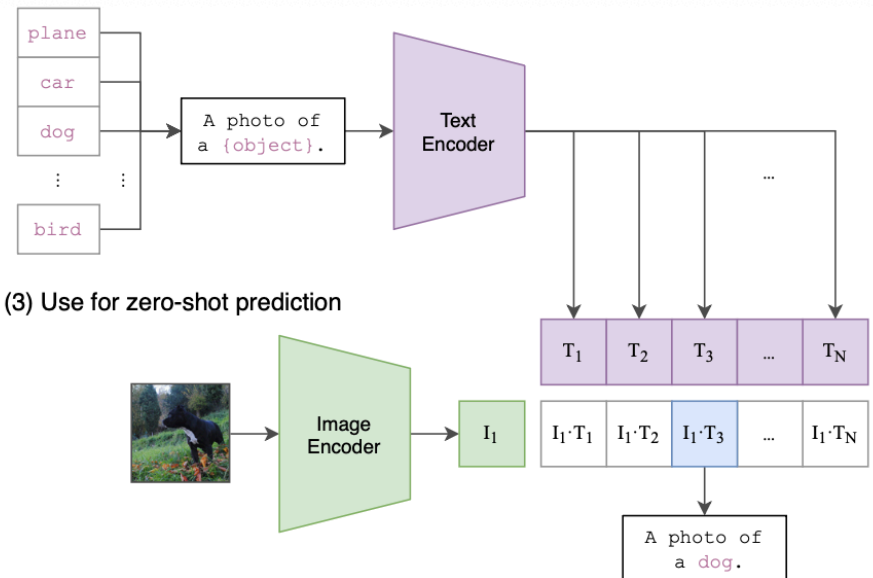
Pivot Point: CLIP for Filtering

- ❖ CLIP: Contrastive Language-Image Pretraining
 - ❖ OpenAI sourced 400mil pairs from Web data
 - ❖ Simple dual encoder DL architecture
 - ❖ Contrastive loss to learn shared image-text embedding space

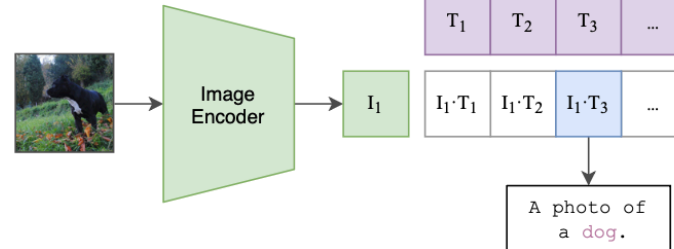
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Pivot Point: CLIP for Filtering

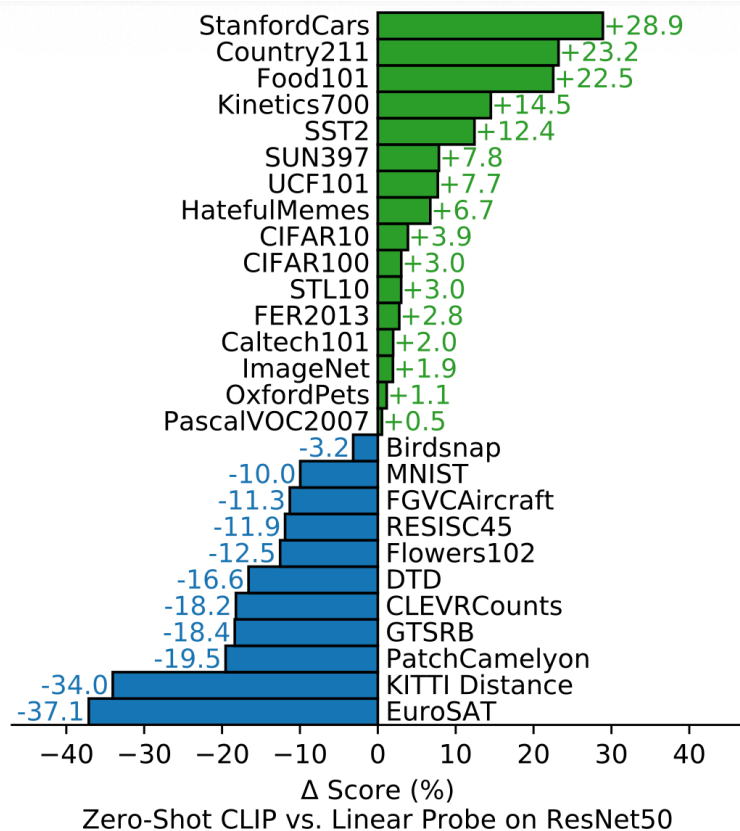


Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

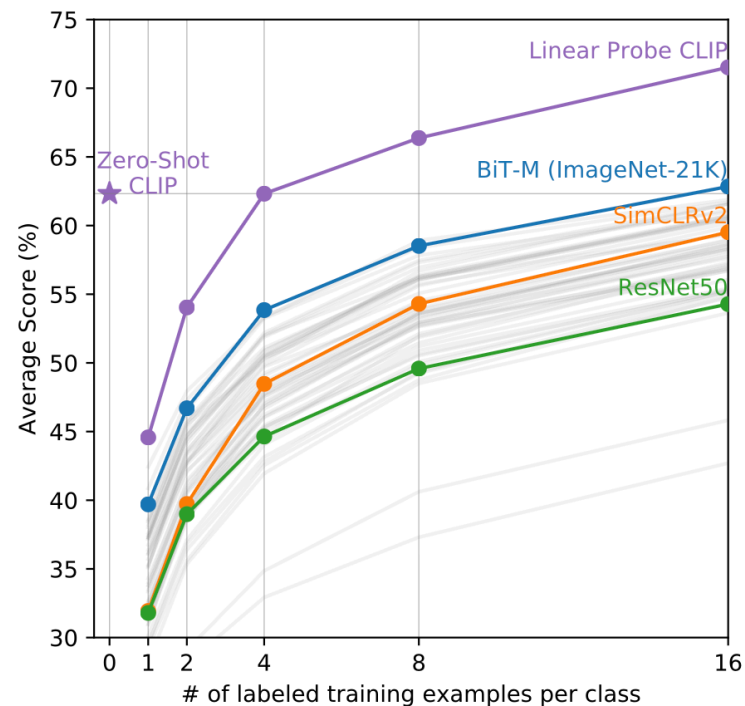


Figure 6. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

Pivot Point: CLIP for Filtering

- ❖ **CLIP score** = cosine similarity between image and text
 - ❖ Valuable quality signal for multimodal data curation
 - ❖ Keep pairs above threshold => huge quality improvement
 - ❖ Threshold is model-dependent and calibrated per dataset
 - ❖ For instance, LAION-5B used ~ 0.28 with ViT-B/32 on web crawled image-text pairs
 - ❖ Now standard for image->text and text->image tasks

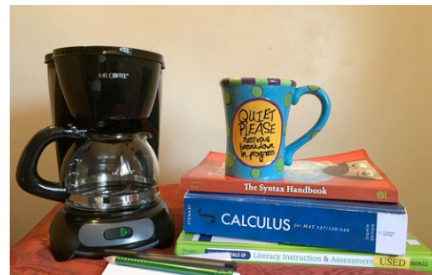
Pivot Point: CLIP for Filtering



- **Accurate:** The darker-haired cat is behind the white cat.
CLIP Score: 22.28
- **Category Error:** The darker-haired **bird** is behind the white cat.
CLIP Score: 20.43 (-1.85)
- **Attribute Error:** The **light-haired** cat is behind the white cat.
CLIP Score: 21.22 (-1.06)
- **Spatial Error:** The darker-haired cat is **to the right of** the white cat.
CLIP Score: 23.66 (+1.38)



- **Accurate:** The white refrigerator is closer to the camera than brown door.
CLIP Score: 29.74
- **Category Error:** The white **cabinet** is closer to the camera brown door.
CLIP Score: 25.66 (-4.08)
- **Attribute Error:** The **dark** refrigerator is closer to the camera than **opened** door.
CLIP Score: 29.39 (-0.35)
- **Spatial Error:** The white refrigerator is **further from** the camera than brown door.
CLIP Score: 29.62 (-0.12)



- **Accurate:** The dark coffee machine is to the left of the mug.
CLIP Score: 22.10
- **Category Error:** The **ceramic cup** is to the left of the mug.
CLIP Score: 19.48 (-2.62)
- **Attribute Error:** The **old broken** coffee machine is to the left of the mug.
CLIP Score: 21.27 (-0.83)
- **Spatial Error:** The dark coffee machine is **in front of** the mug.
CLIP Score: 22.53 (+0.43)



- **Accurate:** Two people are standing behind a pile of huge-sized pipes.
CLIP Score: 26.62
- **Category Error:** Two people are standing behind a pile of huge-sized **papers**.
CLIP Score: 16.40 (-10.22)
- **Attribute Error:** Two people are standing behind a pile of **small** pipes.
CLIP Score: 25.78 (-0.84)
- **Spatial Error:** Two people are standing **inside** a pile of huge-sized pipes.
CLIP Score: 27.18 (+0.56)

LLM-based Multimodal Annotations

Q: Can text-only LLMs help “see” images in multimodal data?

- ❖ **LLaVA:** Large Language and Vision Assistant
 - ❖ Key insight: Structured image descriptions (captions + bounding boxes from COCO) enable a text-only LLM to produce *visual instruction* data without new human annotation!
 - ❖ GPT-4 generates visual instruction data from just image captions + bounding boxes; never sees the images!
 - ❖ 150K visual instruction examples across 3 types: conversations, detailed descriptions, complex reasoning QA
- ❖ Bridges the previous sections we discussed: data synthesis with LLM help + distillation + multimodal annotation

LLM-based Multimodal Annotations

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

LLM-based Multimodal Annotations

Visual input example, Extreme Ironing:



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

User	What is unusual about this image?
LLaVA	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. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
User	[Start a new conversation, and clear the history]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car . The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user's instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

LLM-based Multimodal Annotations

- ❖ Simple DL model architecture!
 - ❖ Vicuna LLM + pre-trained CLIP visual encoder ViT-L/14

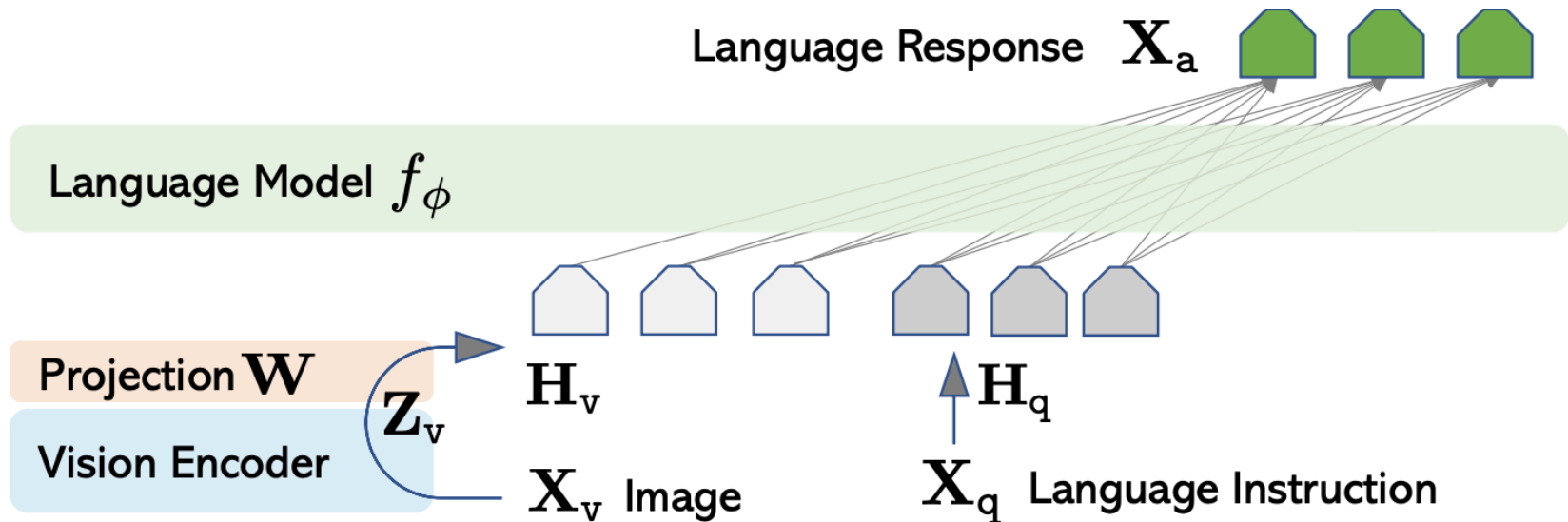


Figure 1: LLaVA network architecture.

Multimodal Data: Bigger Picture

- ❖ DCAI principles apply across all modalities, e.g., filtering (CLIP score), deduplication (perceptual hashing), mixing modalities carefully, documenting choices, etc.
 - ❖ Tool specifics differ, similar data-centric engineering mindset
- ❖ Video and audio face more challenges, e.g., temporal alignment and modality-specific noise
- ❖ Training strategy matters for mixed modality data: *staged training* (align modalities first -> instruction-tune) is now standard vs. mix and train in one go
 - ❖ Active area of research: modality ratios, training ordering, quality variation across modalities

Review Questions

1. Why is the new data synthesis mechanism introduced in Self-Instruct so pivotal in data-centric AI? Explain with an example.
2. Evol-Instruct proposes two evolution strategies: in-depth and in-breadth. Explain the difference with a concrete example. Which strategy was found to be more impactful for data annotation quality?
3. What is distillation in the context of synthetic data? Briefly explain a legal issue that can arise when using frontier models as teachers.
4. Briefly describe a concrete mitigation for model collapse. Explain why it works.
5. Why is annotation diversity important in the context of LLM-based data synthesis? Briefly explain one strategy to achieve diversity.
6. What is CLIP score? Why is it so important in multimodal AI?
7. In LLaVA, GPT-4 never saw any images. So, how did it lead to models that can perform complex reasoning QA over images?
8. Suppose you are part of the AI team at a leading hospital. Your team generates 100K synthetic instruction examples for a multimodal medical Q&A chatbot. Your AI algorithmics colleague proposes to use all examples without filtering to fine-tune a model. List 3 reasons to them why that is likely not a good idea.