# Scientific Reasoning: Assessment of Multimodal Generative LLMs

Florian Dreyer - Ekaterina Kolos - Daria Matiash

# Structure

- Motivation
- Dataset
- Benchmarking
- Prefix Tuning vs. LoRA
- Knowledge Distillation
- Error Analysis
- Conclusion

# Motivation

Foundation Models are well-known for their ability to extract and generate information on massive data.
These research questions are the main interest in our project:

- RQ1: how can they deal with **scientific reasoning**?
- RQ2: how do **prefix tuning** and **LoRA** affect the reasoning capabilities of smaller pretrained models?
- RQ3: how good is **knowledge distillation** with these methods compared to training on the original dataset?

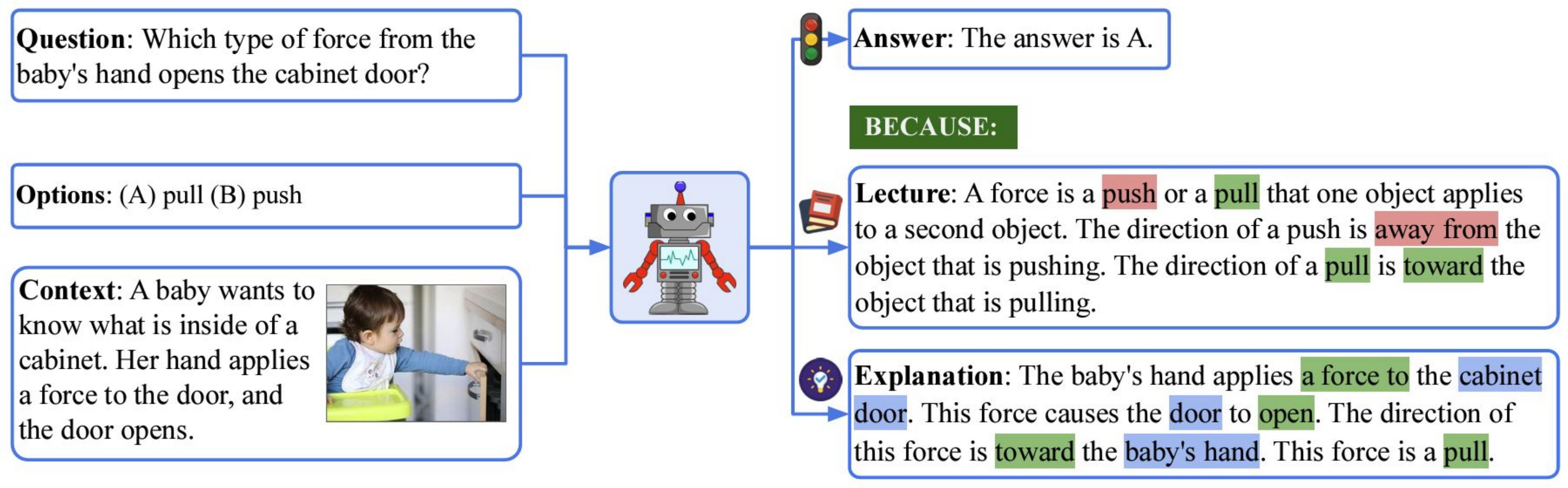
# Dataset: ScienceQA [1]

- Benchmark for multimodal reasoning in science
- Reaches from elementary to high school level questions
- Some irregularities in dataset were filtered and removed

Key Statistics

| | |
|---|---|
| **Total** | 21,208 sets |
| **Subjects** | Natural, social, and language sciences |
| **Topics** | 26 |
| **Categories** | 127 |

# Dataset: ScienceQA




An example question, left side shows the question related data, right side shows the Lecture as further information and the Answer and Explanation, which will be generated by the Foundation Model. [1]

# Benchmarking: Choosing the Teacher

## 6 LLMs

1. Gemini-1.5-flash
2. Gemini-1.5-flash-8B
3. Llava-1.5-7B
4. Pixtral-12b-2409
5. GPT-4o-mini
6. GPT-4

## **Input**: 4 settings

1. question - choices - hint - image - task
2. question - choices - hint - image - task + lecture
3. question - choices - hint - image - task + lecture + solution
4. question - choices - hint - image - task + solution

## **Output**: answer + solution

## Metrics

1. Answer: Accuracy
2. Solution: Average of BLEU-1, BLEU-4, ROUGE, METEOR, cosine similarity

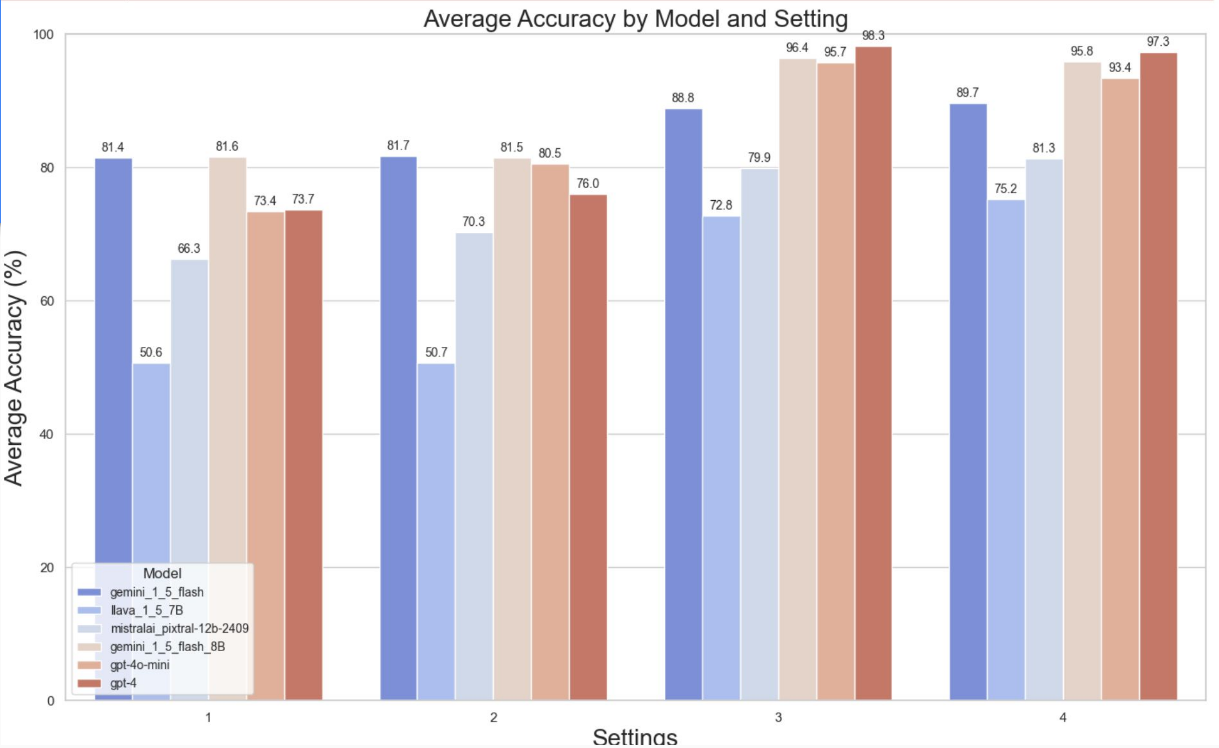
# Benchmarking Take-aways

- Adding *Solution* to input brings consistent significant improvement in accuracy (set. 1&2 vs 3&4)
  → LLMs are capable of extracting information

- Adding *Lecture* to input does not bring any consistent significant change in accuracy (set. 1&4 vs 2&3)

- Best robustness across our 4 experimental settings: Gemini family of models

- Best performance on *lecture*- and *solution*-free tasks: Gemini-1.5-flash-8B




1. question - choices - hint - image - task
2. question - choices - hint - image - task + lecture
3. question - choices - hint - image - task + lecture + solution
4. question - choices - hint - image - task + solution

# Benchmarking Take-aways

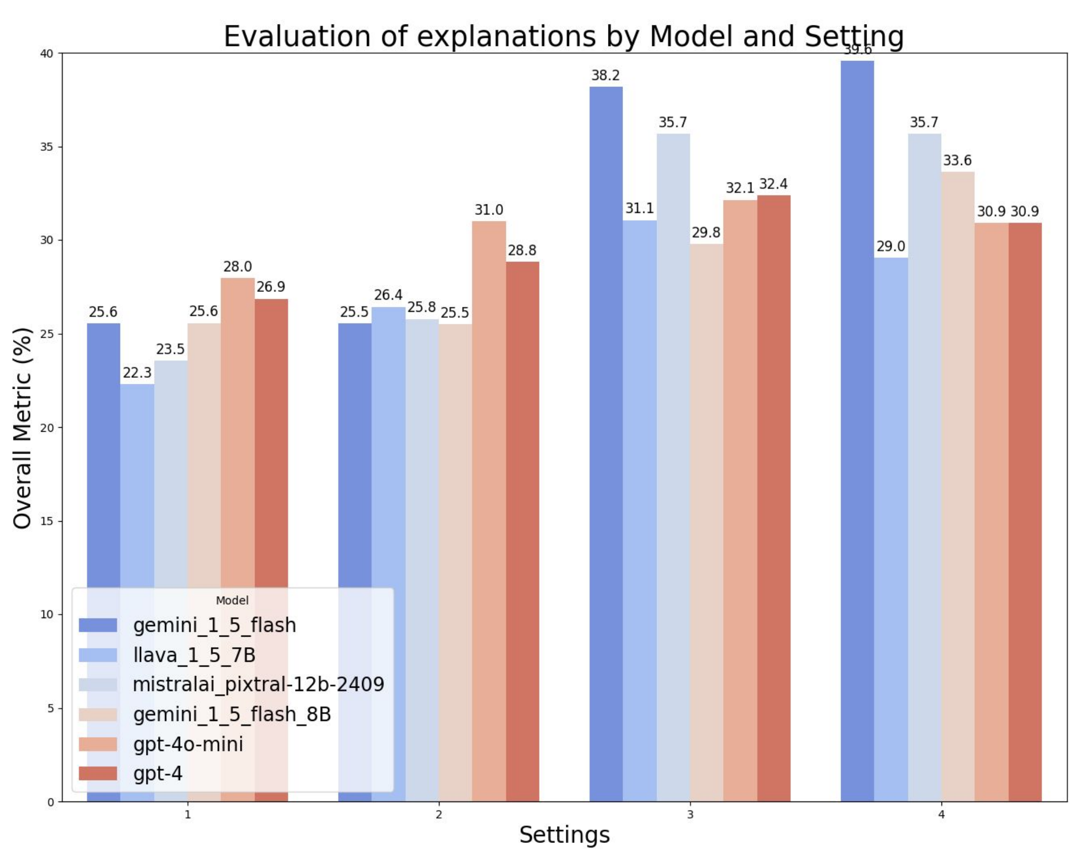
- Overall Metric: avg. of textual similarity metrics

  Adding *Solution* to input brings consistent significant improvement in overall metric (set. 1&2 vs 3&4)
  → LLMs are capable of extracting information

- Adding *Lecture* to input does not bring a consistent significant change in overall metric (set. 1&4 vs 2&3)

- Best textual explanations on *lecture*- and *solution*-free tasks: GPT-4o-mini

- We chose **Gemini-1.5-Flash** as the **teacher** model, because it performs about the same as Flash 8B (set. 1) but has more parameters




1. question - choices - hint - image - task
2. question - choices - hint - image - task + lecture
3. question - choices - hint - image - task + lecture + solution
4. question - choices - hint - image - task + solution

# RQ1: how can LLMs deal with scientific reasoning

Big LLMs can achieve high accuracy in multimodal Scientific Reasoning

Big LLMs can benefit from more context for better reasoning

Extracting information (from solutions) is easier than producing new ideas (from lectures)

# Adapters

**RQ2**: Investigate how Prefix Tuning compares to LoRA for science reasoning tasks
**Method**:

- Take a small LLM, measure zero-shot performance as baseline
- Freeze the LLM and learn Prefix Tuning/LoRA adapter to guide the model

**Why?**

- Practical motivation: Full fine tuning is more expensive
- Research motivation: do these adapter methods work for scientific reasoning in a multimodal setting?

# LoRA vs Prefix Tuning Architecture

- LoRA: decomposing weights into low-rank matrices, reducing parameters
- Prefix Tuning: Train a layer that computes a prefix for the input, freeze rest

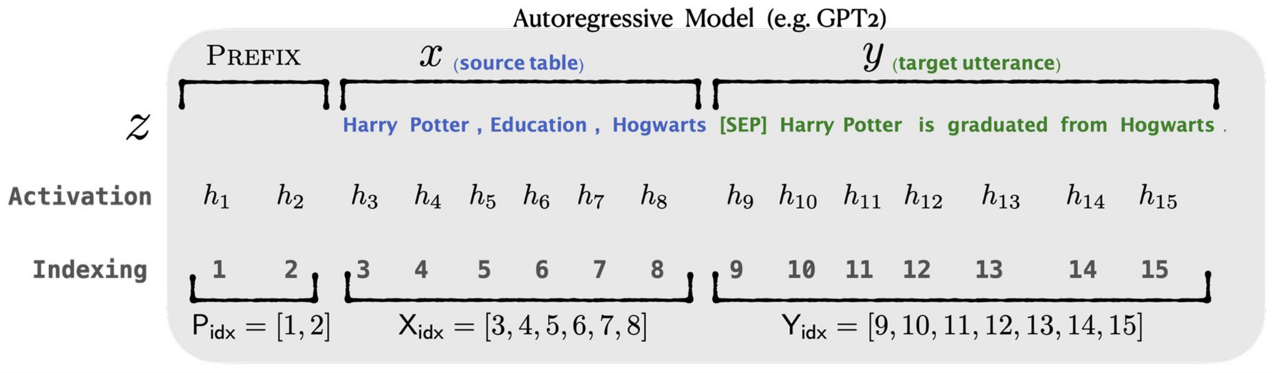



image from [2]

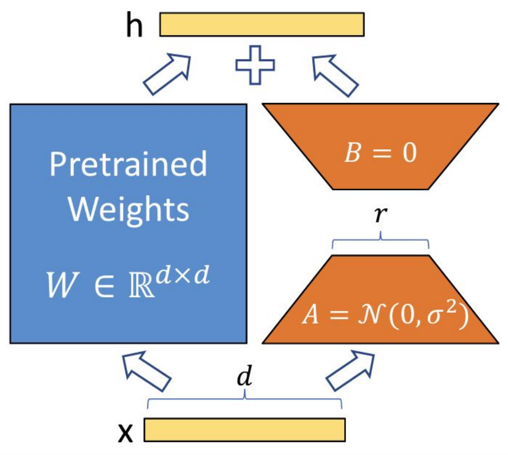
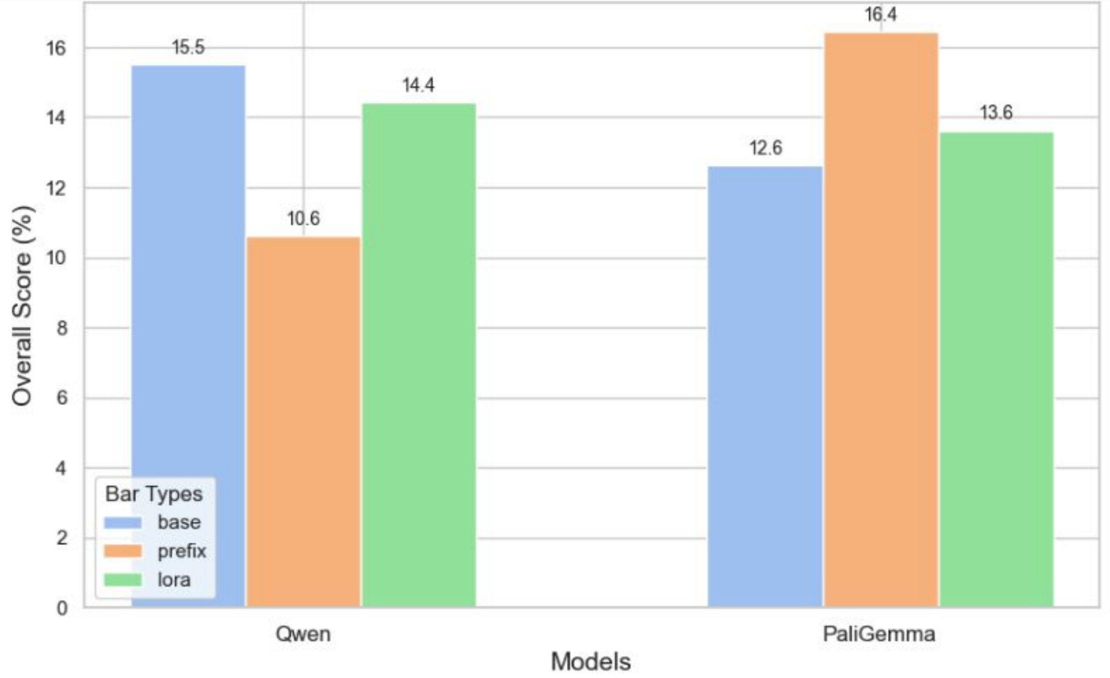



image from [3]

# Performance Comparison

- We chose solution as label
- Metrics compared: avg of textual similarity metrics on test data
- Results not consistent over different models
- We punish the model for any divergence from golden solution
- Gemini 1.5 Flash achieves score of 20



# RQ2: how do prefix tuning and LoRA affect the reasoning capabilities of smaller pretrained models?

Prefix tuning and LoRA do not bring consistent improvement to the reasoning capabilities of smaller pretrained LLMs

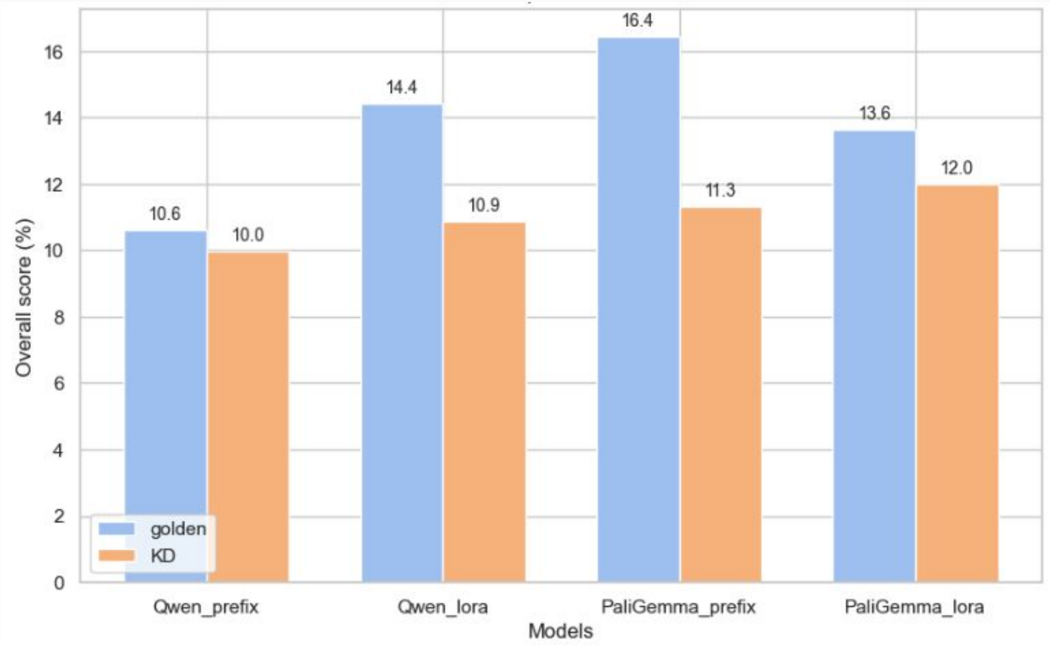
# Knowledge Distillation

- **RQ3**: Investigate how good **knowledge distillation** with adapter methods is compared to training on the original dataset?

- Instead of taking the solution from the dataset as the label to train on, take the output of the teacher model

# Knowledge Distillation Performance

- The overall score (avg. similarity) is low for all models
- Learning from Teacher's outputs consistently drops the performance on test data for both adapters and both models



# RQ3: how good is knowledge distillation with these methods compared to training on the original dataset?

Learning on Teacher’s outputs drops performance as measured by textual similarity metrics

# Errors of Large Foundation Models

- The solutions can contain irrelevant information, or be too abstract

  **Question**: What's the difference between weather and climate?
  **Solution**: *Climate is the pattern of weather in a certain place. It got down to 3°C in Athens, Greece, last night!*

  **Question**: Which property do these three objects have in common? **Choices:** ['stretchy', 'transparent', 'rough']
  **Solution**: *An object has different properties. A property of an object can tell you how it looks, feels, tastes, or smells. Properties can also tell you how an object will behave when something happens to it. Different objects can have properties in common. You can use these properties to put objects into groups. Grouping objects by their properties is called classification.*


- Less mistakes were made in answers to tasks on Natural Sciences
- More mistakes were made in answers to tasks on Social Sciences

# Errors in Format

- Only biggest MLLMs (*GPT4*, *Gemini*) could consistently produce valid JSON of expected format
- Controlled decoding for smaller local models did not help produce JSON
- *Paligemma-FT* only outputs answers without providing a solution
- *Qwen* can sometimes output in Chinese instead of English
- Answer is inconsistently given as copy of choice string, number of choice or letter of choice. The number/letter is not guaranteed to be in range [0;5]

# Errors of Adapters

- Adapters made the LLMs “afraid” to generate sequences that they produced in zero-shot settings
- The new outputs are short (often empty) and often only contain functional words, numbers, or rare irrelevant or foreign tokens: “名”, “loon”, “6”, “ⓧ”

- Possible explanation: the model was discouraged every time it’s solution was not 100% as expected

# Conclusion

- Very **large foundation models**:
  - are good in multimodal scientific QA;
  - extracting information from available *solutions* improves performance;
  - reasoning from *lectures* does not improve, and sometimes drops performance.
- **Small foundation models** such as *Qwen* and *Paligemma* (pretrained) show low performance on the free-text science reasoning task both in zero-shot inference and with adapter tuning
- The metrics (zero-shot/Prefix/LoRA) are **not consistent** in *Qwen* and *Paligemma*
- Bad performance of adapter tuning is probably affiliated with our loss function
- Knowledge Distillation is outperformed by training on golden dataset

# Future Work

- Custom error function using BLEU, ROUGE, …
- Multihead architecture (solution vs answer)
- Knowledge distillation with different prompting techniques on teacher model (e.g. learn from mistakes) [5] [6]

# References


- [1] Pan Lu et al., Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. NeurIPS 2022
- [2] Li, X. L., & Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint (2021)
- [3] Hu, E. J. et. al., Lora: Low-rank adaptation of large language models. arXiv preprint (2021)
- [4] Jianping Gou et. al., Knowledge Distillation: A Survey. IJCV (2021)
- [5] Lucie Charlotte Magister et. al., Teaching Small Language Models to Reason. NeurIPS 2022
- [6] Kumar Shridhar et. al., Distilling Reasoning Capabilities into Smaller Language Models. ACL 2023

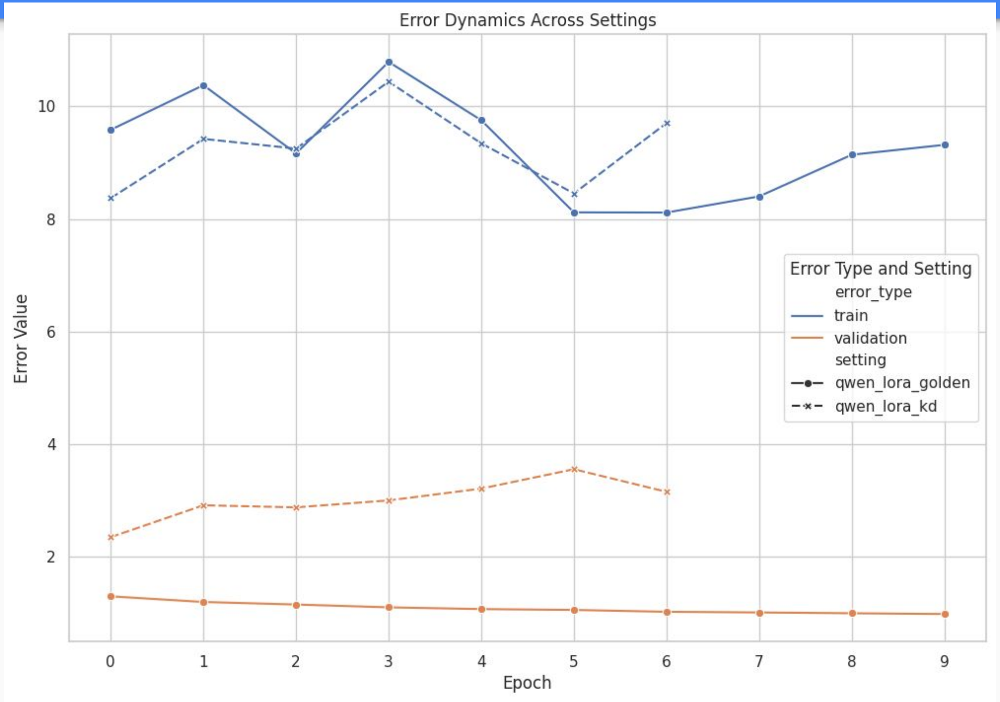

Computations were partly done on BWCluster

# Thank you for your attention!

## Questions?

We have more details in the appendix :)

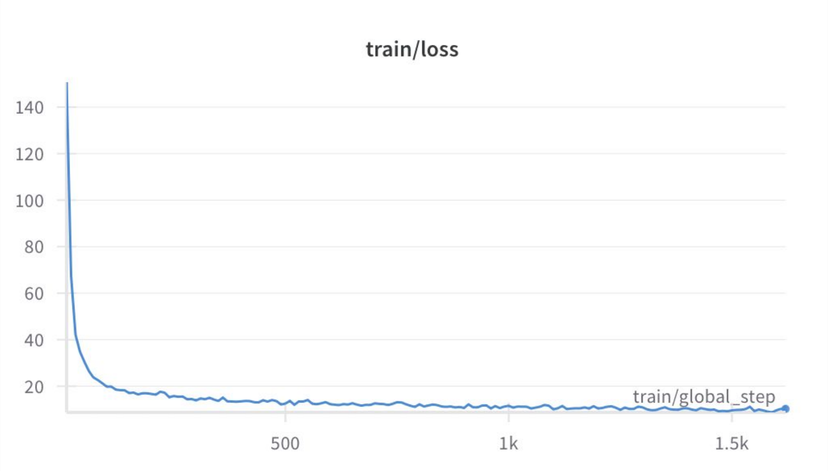
# Appendix
# Training Plots: LoRA Qwen



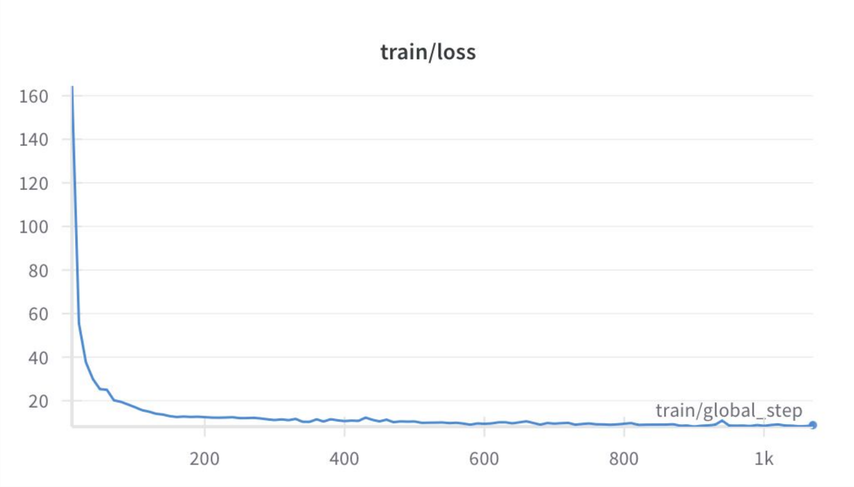

Due to technical issues and limited resources not all errors were saved

# Appendix
# Training Plots: LoRA Paligemma

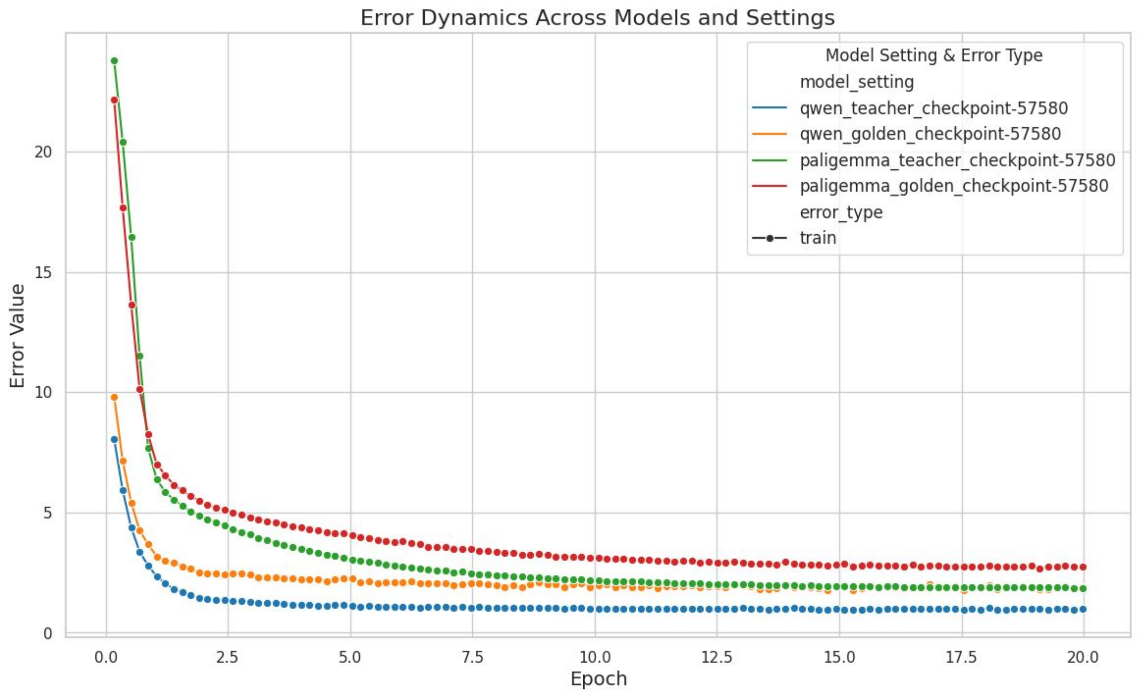
## Paligemma LoRA Golden



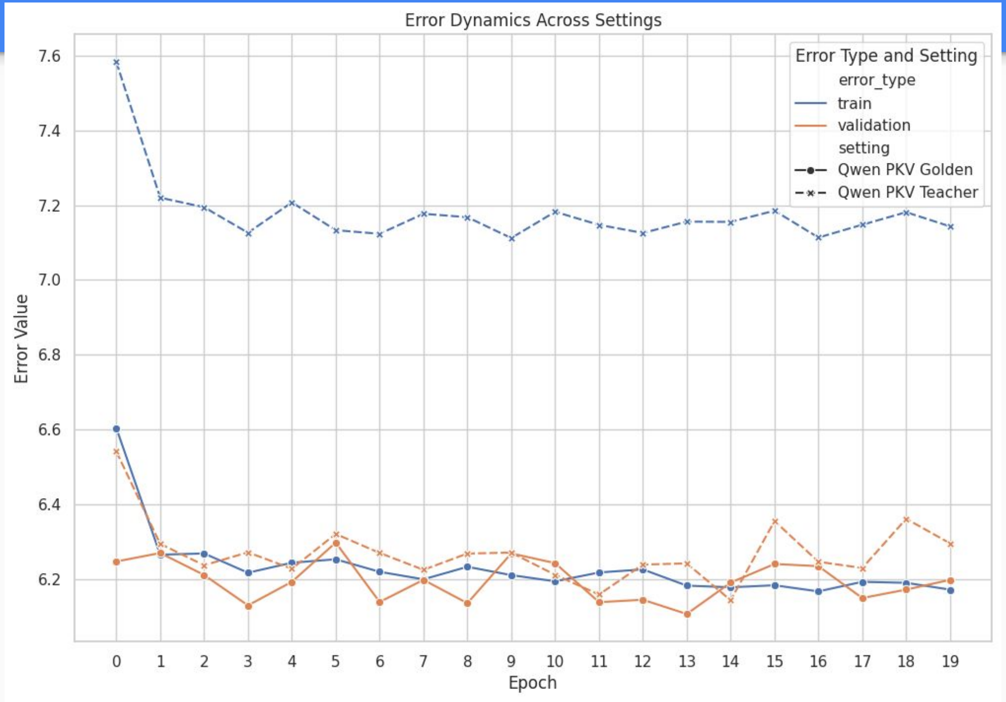

## Paligemma LoRA Teacher



# Appendix
# Training Plots: PEFT Prefix

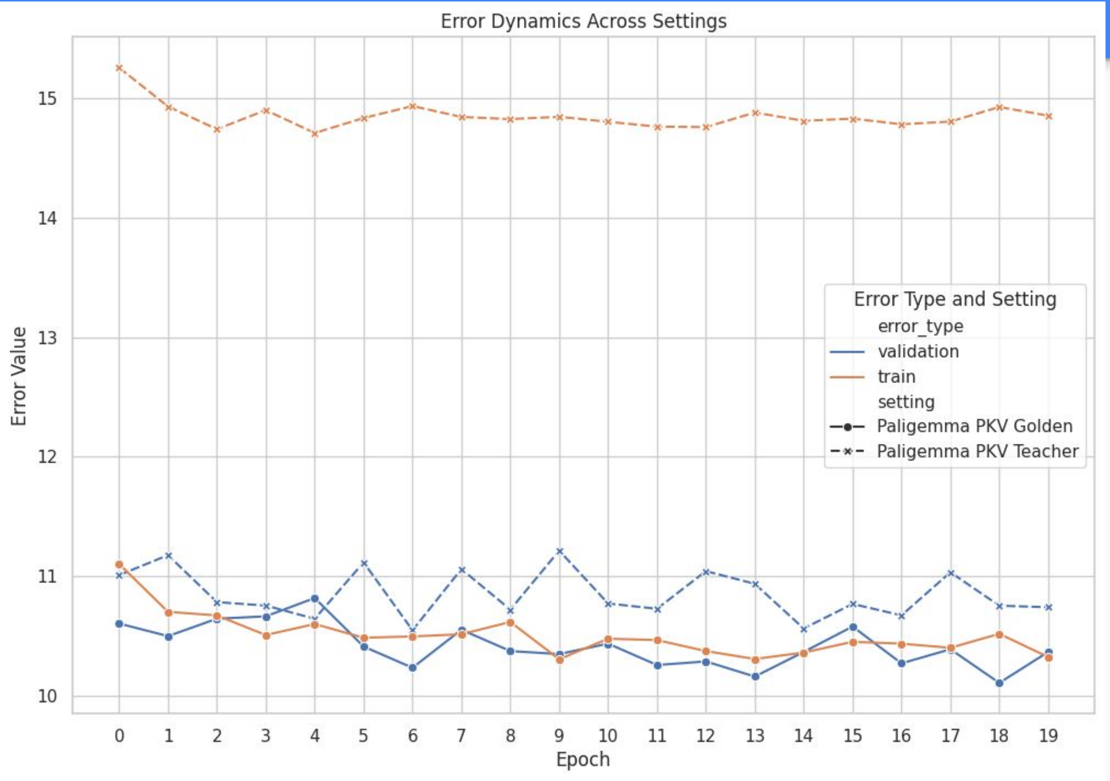


# Appendix
# Training Plots: PKV Qwen




Training on Golden has more stable error on train than training on Teacher outputs. Validation error on Teacher outputs does not really decrease, it does decrease on Golden data (slowly)

# Appendix
# Training Plots: PKV Paligemma



# Appendix
# Error Analysis: Multilinguality

correct solution in Chinese instead of English generated by Qwen2 pretrained: 化学变化 - 经过发酵或分解，可以将有机物（如水果）转化为二氧化碳、水和无机盐。

```
"6"     "Question: What do these two changes have in common?
compost rotting
a piece of apple turning brown
 Task: closed choice
 Choices: ['Both are caused by cooling.', 'Both are only physical changes.', 'Both are caused by heating.', 'Both are chemical changes.']
 Hint:
 Instruction:
Please output the answer in JSON style with an answer and a solution field"     "```json
{
  ""answer"": ""Both are chemical changes."",
  ""solution"": ""化学变化 - 经过发酵或分解，可以将有机物（如水果）转化为二氧化碳、水和无机盐。""
}
"       "natural science"
```

# Appendix
# Error Analysis

Solution generated by MLLM is a description of images instead of reasoning
e.g. sample 117 of validation set

```
"117"   "Question: Which property do these three objects have in common?
 Task: closed choice
 Choices: ['salty', 'scratchy', 'sticky']
 Hint: Select the best answer.
 Instruction:
Please output the answer in JSON style with an answer and a solution field"     "```json
{
  ""answer"": ""sticky"",
  ""solution"": [""honey"", ""wet glue"", ""chocolate milkshake""]
}
"        "natural science"
```

# Appendix
# Error Analysis: Output format

## Model generates output of arbitrary format instead of following expected JSON structure, e.g.:

"Question: Based on the arrows, which of the following organisms is a consumer?  Task: closed choice Choices: ['mushroom', 'lichen'] Hint: Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem. Instruction: Please output the answer in JSON style with an answer and a solution field"


```
{
  ""foodweb"": [
    {
      ""primary"": ""bear"",
      ""consumers"": [
        ""collared lemming"",
        ""bilberry""
      ],
      ""consumed_by"": [
        ""arctiid earthworm"",
        ""arctic fox""
      ]
    },
    {
      ""primary"": ""collared lemming"",
      ""consumers"": [
        ""bear sedge""
      ],
      ""consumed_by"": [
        ""owls""
      ]
    },
    {
      ""primary"": ""owls"",
      ""consumers"": [
        ""owls""
      ],
      ""consumed_by"": [
        ""owls""
      ]
    },
    {
      ""primary"": ""isonote"",
      ""consumers"": [
        ""mushroom"",
        ""collared lemming""
      ],
      ""consumed_by"": [
        ""mushroom"",
        ""collared lemming""
      ],
      ""within"": []
    }
```