

Scientific Reasoning with LLMs

Stuttgart Team: Florian Dreyer (3667877) Ekaterina Kolos (3688474)
Daria Matias (3668740)

December 9, 2024

1 Introduction

Scientific machine reasoning is the application of AI techniques to simulate scientific reasoning processes, such as data interpretation, hypothesis generation, and causal inference. It facilitates scientific inquiry and discovery by enabling machines to independently analyze scientific data, generate hypotheses, and make predictions. For our project, we intend to compare different approaches around generative large language models to generate solutions for school science problems. On an existing benchmark with human solutions, we are planning to compare: a pure LLM with zero-shot and few-shot prompting (and possibly some more fine-grained prompting techniques), an agentic approach with ability to query external resources like textbooks in physics, maths etc, and soft-prompting techniques. We then plan to attempt knowledge distillation to a smaller LLM.

2 Background and Related Work

Prompting techniques A prompt is a command a user gives to a generative LLM used to guide its output for a downstream task, which contains several of the following: a role (persona) the LLM has to follow (e.g. "You are a helpful assistant"), a precise task description, examples (exemplars) for analogous learning (few-shot prompting), requirements on the process of reasoning, style instructions, output format requirements, additional information for in-context learning, emotion prompting components (highlighting why the task or a particular requirement is important). Prompts can be combined into sequences or graphs (with complex branching and parallelism). Of particular interest are small requirements on the reasoning process that can significantly improve performance when added to the prompt, such as "Let's think step by step" [kojima2022large]. Chain-of-thought prompts improve reasoning capabilities by asking the model to speak its process of thinking out loud, with variations including self-ask, step-back prompting, thread-of-thought, least-to-most prompting (decomposing then solving), plan-and-solve prompting, tree-of-thought, recursion-of-thought and many more [schulhoff2024prompt]. Soft prompting is an alternative to fine-tuning where the model is frozen while only a small number of "soft" prompt parameters are trained [lester2021prompt]. These parameters are used to guide the model in the right direction.

Agents An important step forward in using LLMs are agent-based architectures [lin2024swiftsage] [ghafarollahi2024sciagents]. They allow the LLM to act as an intelligent being capable of planning a solution of a complex task, while resorting to external resources on the way to acquire in-context the knowledge it does not have from pre-training. The retrieved information, the results of invoking tools, such as calculators, and interactions with the environment are appended to the prompt (or, in a more advanced scenario, to the summarized memory [ge2023context]). Optionally, a reflection step (explicit reasoning on all the accumulated information) is added before allowing further generation of the final response.

Knowledge distillation (KD) allows to obtain smaller models capable of successfully following the behavior of larger teacher models. This is particularly useful for privacy reasons (mobile AI apps) and when access to large models in the cloud is not guaranteed (e.g., in cars). The teacher model can be used to intelligently select examples on which the student model is trained (dataset distillation) [yu2023dataset], or provide negative samples to show the student what incorrect answers or reasoning paths it should avoid to improve task accuracy [li2024turning] (cf. contrastive CoT). Training small models on a CoT reasoning path of a larger model was also shown to be a way to obtain a small student model replicating the reasoning capabilities of the teacher on downstream tasks

[magister2022teaching], which is close to *response-based KD* where the student model mimics the output of the teacher. Alternatively, with *feature-based KD* the knowledge from specific layers of the teacher model is distilled into the student model [sepahvand2022teacher], while the structure of the student model may be a quantized, simplified or condensed version of the teacher’s architecture [gou2021knowledge].

In our project, we plan to develop an extended version of CoT knowledge distillation, where both the reasoning paths and the essence of the external knowledge the teacher agent has retrieved for the tasks are acting as training data for the student model. This way the teacher agent ”defines” the ”protocols” of solving complex tasks, while the student model tries to ”follow” the learned ”instructions” and memorize important information for similar tasks. If we succeed, such small models could be a valuable asset in practice in enterprise environments, to encode important protocols and parts of project documentation, replacing expensive intelligent agents for specific tasks.

3 Methodology

We postulate the following research questions: RQ1: Can we build a LLM agent to improve the LLMs performance on science problems? RQ2: Does Soft Prompting improve the performance of the LLM (agent)? RQ3: Can an LLM agent’s behavior be distilled into a single model?

RQ1 We plan to start by prompting a multimodal model with reasoning capabilities with a zero-shot and few-shot prompting settings. The preliminary list of foundation models to be compared as baselines is as follows: CLIP, GPT4o (GPT4-V), T5, BigGAN, Gemini; small models: TinyCLIP, T5-Small; Llama models. This simple generation will be compared with an agent-based approach using the same models, which will now include augmented retrieval of information on the scientific task from domain-specific texts.

RQ2 We plan to use Soft Prompting as one of the parameter-efficient fine-tuning techniques on the LLM used to guide it towards better reasoning. To achieve this we will add learnable prompt parameters to the base LLM we use. While training these parameters the rest of the model will be frozen [lester2021prompt].

RQ3 In order to distill the knowledge from the system obtained at previous steps, we plan to do the model distillation with Chain-of-Thought Prompting for Reasoning approach. Following techniques described in [magister2022teaching] [wei2022chain], the student model will be fine-tuned using these CoT responses together with any additional information the agent has retrieved from external sources to produce intermediate reasoning steps. At the same time, the teacher model generates multiple CoT responses, and the student learns from the aggregate (self-consistent) reasoning paths. The foundation model to serve as student LLM should be small, multimodal, capable of reasoning and capable of acquiring new distilled knowledge. A preliminary list of candidate models: small Llama and LLaVA models, MiniGPT-4, small FLAVA models, distilUNITER, lxmert, distilled ViLT.

4 Dataset

This study is based on the ScienceQA dataset [lu2022learn]. The dataset includes a variety of science-related multimodal multiple-choice questions together with annotations of the answers that provide relevant lectures and explanations in Nature Science, Language Science and Social Science.

Data points are structured as follows:

```
{'image': <PIL.PngImagePlugin.PngImageFile image mode=RGB size=200x202>,
'question': 'Is diorite a mineral or a rock?',
'choices': ['rock', 'mineral'],
'answer': 0,
'hint': 'Diorite has the following properties:\nno fixed crystal structure\nnaturally occurring\nno',
'task': 'closed choice',
'grade': 'grade8',
'subject': 'natural science',
'topic': 'earth-science',
'category': 'Rocks and minerals',
'skill': 'Identify rocks and minerals',
'lecture': 'Minerals are the building blocks of rocks. A rock can be made of one or more minerals.'
```

'solution': 'The properties of diorite match the properties of a rock. So, diorite is a rock.']}

For some datapoints, the image data is missing, while the answer can be deduced from text alone (e.g. questions on 'subject': 'language science', 'topic': 'figurative-language', 'skill': 'Interpret figures of speech' (e.g. irony recognition).

For other datapoints, however, the image data can be missing, although required to make an answer, e.g.

```
{'image': None,
'question': 'What information supports the conclusion that Rick inherited this trait?',
'choices': ["Rick's coworker also has curly hair.",
"Rick's biological father has curly hair.",
"Rick and his biological parents have brown hair."],
'answer': 1,
'hint': 'Read the description of a trait.\nRick has curly hair.',
'task': 'closed choice',
'grade': 'grade7',
'subject': 'natural science',
'topic': 'biology',
'category': 'Genes to traits',
'skill': 'Inherited and acquired traits: use evidence to support a statement',
'lecture': "Organisms, including people, have both inherited and acquired traits. Inherited and acq
'solution': ''}
```

Apart from the image field, missing values can occur in lecture and solution fields. The statistics in table 1 give an overview on the amount of problematic datapoints.

Finally, hint can be missing, e.g. the text-only task 367 on subject language science: "What is the source of allusion in the sentence below? Edward picked up his pace on the trail as his spidey sense began to tingle" with choices ["Italian history", "a comic book"], or the text-only task 3 on the topic of capitalization: "Which correctly shows the title of a play?" with choices ["A breath of Fresh Air", "A Breath of Fresh Air"].

missing	frequency
solution	9.3%
lecture	15.7%
solution and/or lecture	24.2%
image	50.6%

Table 1: Defect datapoints in dataset

Retained data As 50.6% of the datapoints are missing an image, we decide to process both text-only and text+image datapoints similarly with multimodal models, attaching the image to the prompt if it is available. For now, we will only work with datapoints with non-empty solution AND with non-empty lecture. All the other datapoints are disregarded, which results in **3216** datapoints kept in validation data.

Metrics We evaluate model's performance with question answering accuracy domain-wise in order to have a fair comparison with leaderboards. The reasoning steps can be evaluated with semantic similarity metrics (e.g. adopted from machine translation) such as BLEU, METEOR, ROUGE, and cosine similarity and more.

4.1 Multiple choice Evaluation

Owing to the simplicity of the test format, only accuracy score is computed, following original evaluation strategy by [lu2022learn]. It was determined that the computation of the overall score should assign the highest weight (**50%**) to the ability to provide correct answers (accuracy score), as this represents the most critical aspect of performance evaluation.

4.2 Answer Reasoning Evaluation

Due to the peculiarity of scientific texts and approaches to the evaluation of automatically generated texts, the following evaluation approaches were chosen. Our validation process consisted of the metrics introduced in the original paper [lu2022learn] with a few additional metrics and techniques that we consider of high importance in regards to the scientific textual data.

4.2.1 BLEU

BLEU-score ([papineni2002bleu]) can measure how closely the model’s generated explanation aligns with the human-authored explanation. BLEU-1 measures if the model uses the right scientific terms or key words (e.g., "photosynthesis," "temperature"). However, it ignores word order and logical progression, so it can’t evaluate reasoning or explanation quality. That is why BLEU-4 score is also computed in order to capture both vocabulary and the arrangement of words into meaningful phrases, evaluating fluency and coherence. Scientific explanations can often convey the same idea with different wording, making semantic similarity hard to capture. BLEU might penalize valid but rephrased or concise answers and it does not account for deeper meaning or logical correctness, which are crucial in reasoning tasks. Since this metric can give valuable insights on the generated outputs, we do not consider these scores as the most important and both BLEU-1 and BLEU-4 scores have **5%** impact.

4.2.2 ROUGE

ROUGE-L score, following evaluation strategy in [lu2022learn], helps analyze the content recall and sequence preservation of model-generated explanations. They highlight whether a model captures key scientific concepts and produces explanations that align with reference reasoning. However, ROUGE should be used in conjunction with other metrics to fully evaluate the logical and factual quality of generated reasoning. ROUGE-L will highlight the mismatch in logical flow, even though some content overlaps in reference and candidate sentences (1, 2). It lead to the conclusion that this metric would not drastically affect the overall score and contribute (**5%**).

Example 1. • **Reference:** Water evaporates, cools, and condenses to form clouds.

 • **Candidate:** Water cools, evaporates, and forms clouds.

Example 2. • **Reference:** Plants absorb sunlight to produce energy.

 • **Candidate:** Sunlight is absorbed by plants to create energy.

Cosine similarity with Sentence Transformers Cosine similarity using SentenceTransformer embeddings provides a robust method for evaluating model outputs in the ScienceQA dataset, focusing on the semantic alignment of generated explanations with references. It captures nuanced relationships and conceptual understanding, which gives more understanding of the generated outputs in comparison to ROUGE and BLEU scores and that is why it has more importance in final score computation. However, due to domain specificity it can miss some intricate cases like still considering high similarity of two sentences, even though it the reference can contain crucial information about carbon dioxide, while in the candidate sentence the answer was given in relation to oxygen reproduction.

4.2.3 METEOR

METEOR ([banerjee2005meteor]) evaluates text similarity based on precision, recall, and alignment of semantically meaningful components like synonyms and stems, making it well-suited for capturing the nuanced language used in scientific reasoning. By incorporating a weighted F-score and penalties for excessive mismatches, METEOR provides a balanced evaluation of fluency and accuracy, crucial for assessing LLM-generated answers on the Science QA Dataset. Due to these balanced text evaluation, this metric has **10%** impact on the overall score.

4.2.4 Cosine similarity with Sentence Transformers

Sentence Transformer models are optimized for sentence-level embeddings, which capture the overall contextual similarity of answers more effectively, making them suitable for general-purpose evaluation tasks like scientific reasoning. Unlike SciBERT, which is tailored to scientific text and could have been used as sentence embedder, Sentence Transformers are pre-trained on diverse datasets, enabling them to generalize better to variations in phrasing and style present in the Science QA Dataset. Their ability to

handle broader linguistic nuances ensures that the evaluation accounts for both semantic correctness and the logical flow of responses, which is essential when assessing reasoning across a wide range of scientific topics. That is the reason for using Sentence Transformer model for measurements of cosine similarity between the lecture material and annotated explanations and models’ outputs with impact weight of 15%.

5 Experiments

5.1 Zero-shot and Few-shot Benchmarking

We start by zero-shot benchmarking existing multimodal models in three simple prompt settings, partly repeating with the official evaluation suite:

1. **question - choices - hint - image - task**, with the following prompt:

```
{f'Question: {question}\n Task: {task}\n Choices: {choices}\n Hint: {hint}']}
```

2. **question - choices - hint - image - task + lecture**, with the following prompt:

```
{f'Question: {question}\n Task: {task}\n Choices: {choices}\n Hint: {hint}\n Lecture: {lecture}]}
```

3. **question - choices - hint - image - task + lecture + solution**, with the following prompt:

```
{f'Question: {question}\n Task: {task}\n Choices: {choices}\n Hint: {hint}\n Lecture: {lecture}\n Solution: {solution}]}
```

For the first two the model has to output **answer - solution**, for the third - **answer** (the easiest set up meant to ensure the model can deduce the correct answer from an already provided correct solution).

We evaluate: OpenAI, Gemini, MistralAI, as well as small local models.

6 Benchmark

The primary focus in identifying the optimal model for comparative analysis is Setting 1, which contains no other additional information except for the question, option choices, hint, image, and the task itself. This approach aims to evaluate at what extent models are capable of leveraging external knowledge from pre-trained data rather than relying on the input-specific information.

6.1 Accuracy

The highest performance was observed in settings that included solutions, indicating that the models were generally capable of extracting relevant information effectively. Among these, settings combining both lecture and solution information typically yielded slightly better results than those with solutions alone. ‘GPT-4’ consistently received highest ratings, especially in scenarios in which lecture and/or solution information was included. The Gemini family of models showed better robustness in many scenarios and came in second overall. Notably, Gemini models achieved an average accuracy advantage of 5% over GPT-family models in the “pure” task context (without lecture or solution information). The ‘LLaVAModel1.5-7b’, on the other hand, demonstrated the lowest average accuracy scores, particularly in environments without lecture and solution data.

6.2 Cosine similarity

Cosine similarity highlights the semantic alignment between the model outputs and the correct answers.

Gemini_{1.5} consistently achieves the highest similarity scores, particularly in settings 3 and 4, indicating strong alignment with 4 variants maintain steady scores across all settings, demonstrating robustness, while mistralai_p and llama₁₅₇ show more variation.

6.3 BLEU-1

Across settings, higher BLEU-1 scores were observed in settings 3 and 4 (those incorporating solutions and/or lecture information), indicating that models performed better with richer contextual data. The ‘GPT-4o-mini’ model demonstrated superior performance, surpassing ‘GPT-4’, in achieving the highest results. Lower scores in settings 1 and 2 suggest that models, particularly ‘LLaVAModel1.5-7b’, may have had trouble correctly aligning their outputs with reference words when given less auxiliary information, as indicated by lower scores in settings 1 and 2.

Model	Question and Explanation	Model's Output
GPT-4	Question: <i>What does the idiom in this text suggest? Sophia submitted her suggestions to Chad; now the ball is in his court.</i> Choices: (0) 'Chad needs to act next.', (1) 'Chad feels like playing or relaxing, not working.'	The idiom "the ball is in his court" is used to suggest that it is now Chad's turn to take action or make a decision after Sophia has done her part. It is commonly used in situations where one person has done all they can and now it's up to the other person to respond or act. Therefore, the idiom in this text suggests that Chad needs to act next, which is why option 0 is the correct answer.
GPT-4o-mini	Explanation: The text uses an idiom, an expression that cannot be understood literally. The idiom the ball is in his court suggests that Chad needs to act next. In tennis, when the ball is in a player's court, it is that person's turn.	The idiom "the ball is in his court" suggests that it is now Chad's responsibility to take action or make a decision regarding the suggestions submitted by Sophia. This phrase typically indicates that one party has completed their part and it is now up to the other party to respond or act. Therefore, the correct choice is that Chad needs to act next, rather than feeling like playing or relaxing.
Gemini-1.5-flash		The idiom "the ball is in his court" means that it is Chad's turn to take action. This suggests that Sophia has done her part and now it's Chad's responsibility to decide what to do with her suggestions.
Gemini-1.5-flash-8B		The idiom "the ball is in his court" means that it is now Chad's turn to take action or make a decision. It's not about playing or relaxing.
Llava-1.5-7B		The idiom 'the ball is in his court' suggests that Chad is the one who needs to act next, as he is the one who is responsible for the next step in the process. This implies that Sophia has completed her part and now it is Chad's turn to take action.
Pixtral-12B		The idiom 'the ball is in someone's court' means that it is their turn to make a move or do something.

Table 2: Example outputs of the tested models on Setting 1.

6.4 BLEU-4

Scores were consistently higher in settings 3 and 4, particularly for models, demonstrating their ability to produce more contextually precise outputs when both lecture and solution information were provided. Gemini models had higher scores in comparison to other models, which represents a better capability of information extraction. 'LlaVAModel', at the same time, gained the highest score in the setting without solutions, which indicates the ability of extracting knowledge from massive textual data.

6.5 ROUGE

The highest ROUGE scores were observed in setting 3, where both lectures and solutions were available, highlighting the importance of comprehensive input for producing informative responses. Setting 1 showed lower ROUGE scores across all models, reflecting limited informativeness when models were provided with minimal context. 'LlaVAModel' and 'GPT-4' had highest ROUGE-scores on settings without solutions. Outputs of 'GPT-4' without knowledge of the correct answer had higher ROUGE score than 'GPT-4o-mini' given solutions in the input.

6.6 METEOR

METEOR scores reveal how well the models capture fluency, grammar, and word-level alignment with the references. Similar to cosine similarity, gemini_{1.5} leads with the highest METEOR scores, particularly in setting 4 (0.318678), level text coherence.

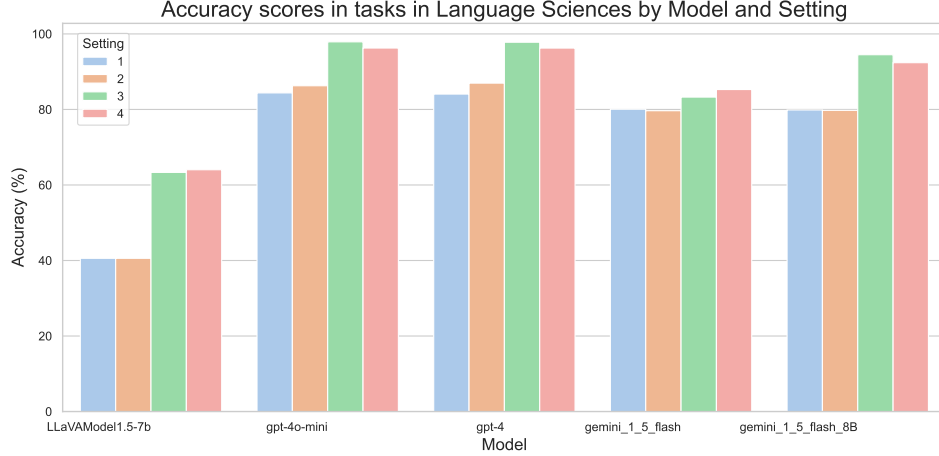


Figure 1: Performance metrics for various LLMs on the Science QA dataset.

6.7 Cosine similarity

Semantic alignment with the reference is assessed by the similarity metric. Gemini and its variants outperformed other models, demonstrating their great capacity to capture the semantic structure of enriched input. Similarity scores were highest in settings 3 and 4. LLaVAModel1.5-7b, on the other hand, received the lowest score in setting 1, indicating that it mostly depends on lecture or solution data to generate outputs that are semantically accurate. This result emphasises how more data affects the resulting content’s semantic integrity. ‘GPT-4o-mini’ and ‘LLaVAModel’ performed better than other models in environments lacking solution information, indicating that they can produce reliable results even with limited contextual guidance.

6.8 Summary

Overall, the findings demonstrate how important enriched contextual data, such as lectures and solutions, in enhancing model performance across all metrics. The GPT-4o-mini model outperformed GPT-4 in multiple instances, indicating the efficacy of smaller, more targeted designs, even if GPT-4 continuously received excellent accuracy and informativeness scores. While LLaVAModel1.5-7b failed without extra input, highlighting its dependence on extensive contextual knowledge to function well, the Gemini family of models demonstrated remarkable robustness and semantic alignment, particularly in enriched situations. Lecture information doesn’t really affect models’ performance, it can even decrease the ratio of correct answers.

6.9 Overall score

6.9.1 Local Models from HuggingFace

Pre-experiments We were not successful with processing with small non-chat models. For the given text + image, the models would return short non-meaningful answers, like “Yes” to a “X or Y” question, or “no idea” to a completely defined task with all relevant context included. These were multimodal models below 1B on Huggingface, not tuned to be used in a chat scenario (and prepending a “You are a helpful assistant...” to the beginning of the prompt was not enough to fix this). We are not sure how to use those models for our purposes as of now.

MiniGPT-4 This is a small multimodal model designed from ?? [zhu2023minigpt]. It offers to build the model based on Llama or Vicuna. **Unsuccessful experiment:** We tried to resort to an unofficial implementation: wangrongsheng/MiniGPT-4-LLaMA-7B. Loading the model was not successful (no preprocessor.config.json). Another issue: using BlipForConditionalGeneration and BlipProcessor: “You are using a model of type llama to instantiate a model of type blip. This is not supported for all configurations of models and can yield errors.”. Model size: 10GB. Processor: 3.6GB.

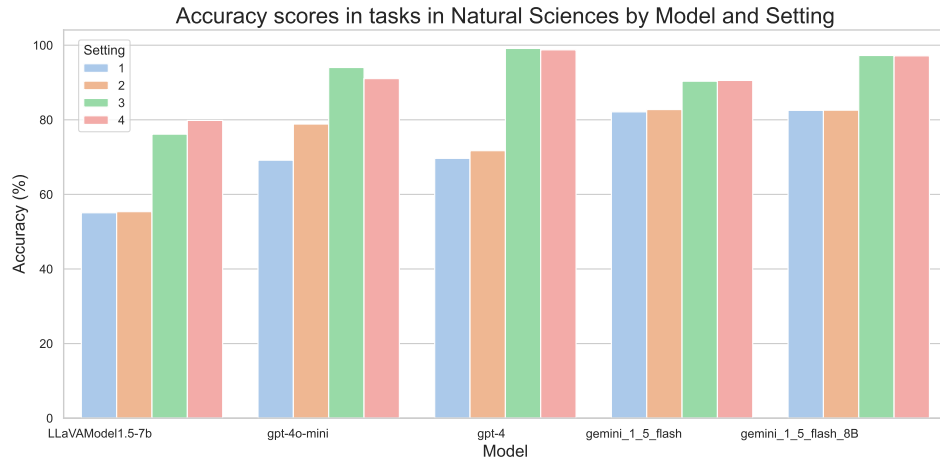


Figure 2: Performance metrics for various LLMs on the Science QA dataset.

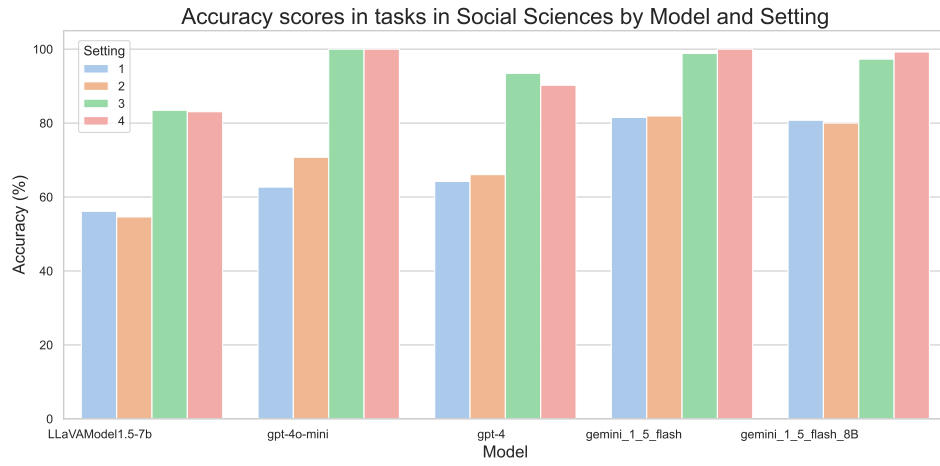


Figure 3: Performance metrics for various LLMs on the Science QA dataset.

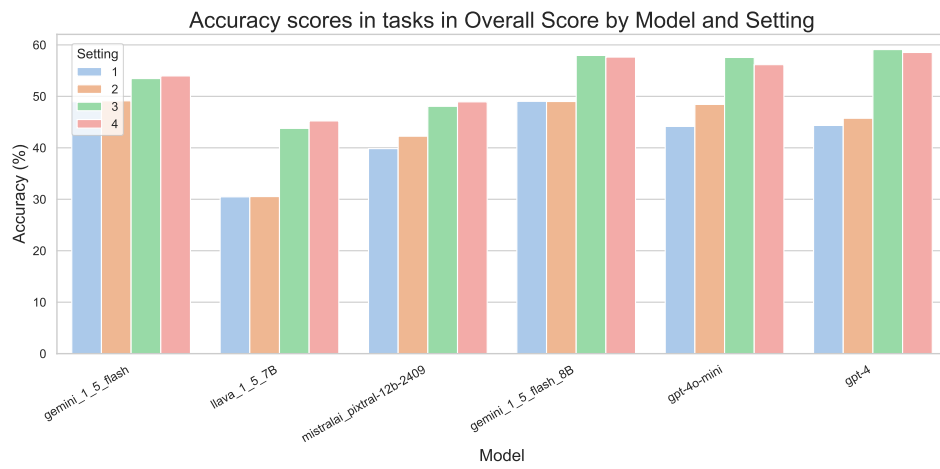


Figure 4: Overall metrics.

Model name	llava-1.5-7b-hf
Size	7B
Published	09.2023
Type	SLM
Chat model	yes
Parent models	LLaMA, Vicuna
Data	GPT-generated multimodal instruction-following data

Table 3: Model Comparison: Benchmarked Models

LLaVa The exact model used is: `llava-hf/llava-1.5-7b-hf` (and there is a larger sister 13b model). It is based on LLaMA and (?) Vicuna fine-tuned on GPT-generated multimodal instruction-following data. It expects inputs as conversation, allowing the preprocessor to convert textual input into format "USER: request ASSISTANT: " where the assistant has to fill in the rest. We encountered the following issues while applying this model version on ScienceQA data:

- although prompted like larger models, this model returned answer strings and not choice index; this may additionally raise an issue of the returned answer not being entirely part of the choice list;
- although prompted to generate JSON, the model would sometimes generate a different string, e.g. simply repeating part of the task. Unexpected format happened in ?? % of all answers. We disregarded these answers, considering them all wrong, which results in a bit more pessimistic estimation of the model's performance. The metrics on the correctly formatted answers alone are also given.
- the model currently does not accept inputs without an image (text-only); according to a huggingface discussion¹, the authors claim to have accidentally removed this possibility in a recent update. We worked around this by producing placeholder empty 10x10 white images and passing them as part of the input when the image in the dataset was missing. We also appended the suffix "Ignore image" to the end of the prompt, because the model would otherwise try to hallucinate a relation between the text and the image (e.g. "Why is the sky blue" + placeholder image results in "The sky is blue on the picture because ...");
- the model does not really seem to support batch processing; although officially the technical possibility exists, it has been reported online ² that the model would only take into account the first image in the batch, which makes the processing useless. We did not risk to attempt the supported or custom batching and processed the filtered validation dataset 1 datapoint by 1, which took 10+ hours on GPU for inference.

There is another available model LLaVAnext, which recently started to support batching, and should have better performance.

Note: I later (after running the model) understood that the original model is this one: 'liuhaotian/llava-v1.5-7b'. The results may thus be different from what would be expected from the model's paper.

BLIP-2

LLaMA models

Gemma2

Qwen models e.g. Qwen2.5-Coder

The metrics used for evaluation at this point are: accuracy, BLEU-1, BLEU-4, ROUGE, similarity.

¹<https://huggingface.co/llava-hf/llava-1.5-7b-hf/discussions/38>

²<https://huggingface.co/llava-hf/llava-1.5-13b-hf/discussions/10>

7 Approximate Timeline

1. Now - mid November:

- Preliminary baseline tests: compare "pure LLM"s relying only on commonsense reasoning / only on pretraining (zero-shot, few-shot).
- Stronger baseline – same LLM with better prompting techniques like CoT/Self-Ask.

2. mid November - mid December:

- Construct agentic / RAG pipeline with "big smart" model + additional resources.
- Optional: soft-prompting for better workflow to distill from.

3. mid December - mid January:

- Knowledge distillation experiments.

Model	Setting	Question and Explanation	Model's Output
GPT-4	1	Question: <i>What does the idiom in this text suggest? Sophia submitted her suggestions to Chad; now the ball is in his court.</i> Choices: (0) 'Chad needs to act next.' (1) 'Chad feels like playing or relaxing, not working.' Explanation:	The idiom "the ball is in his court" is used to suggest that it is now Chad's turn to take action or make a decision after Sophia has done her part. It is commonly used in situations where one person has done all they can and now it's up to the other person to respond or act. Therefore, the idiom in this text suggests that Chad needs to act next, which is why option 0 is the correct answer.
	2	The text uses an idiom, an expression that cannot be understood literally. The idiom the ball is in his court suggests that Chad needs to act next. In tennis, when the ball is in a player's court, it is that person's turn.	The phrase "the ball is in his court" is a commonly used idiom in English. It is borrowed from tennis, where it means it's your opponent's turn to hit the ball. In a general context, it means it is now up to that person to take the next step or make the next decision. Therefore, in the given context, it means that Chad needs to act next following Sophia's suggestions.
	3		The idiom "the ball is in his court" is used to imply that it's Chad's turn to take action or make the next move. This expression originates from tennis, where a player must respond when the ball is in their court. Therefore, in the context of the sentence, the idiom suggests that Chad needs to act next. Option 1 is incorrect because this idiom doesn't imply anything about Chad's mood or willingness to work.
	4		The idiom "the ball is in his court" is used to denote that it is now Chad's responsibility to take the next step or make a decision. It does not imply that Chad feels like playing or relaxing, not working. This idiom is often used in a context where someone has made a move or decision, and now it's another person's turn to respond. Therefore, the correct answer is option 0 - Chad needs to act next.
GPT-4o-mini	1		The idiom "the ball is in his court" suggests that it is now Chad's responsibility to take action or make a decision regarding the suggestions submitted by Sophia. This phrase typically indicates that one party has completed their part and it is now up to the other party to respond or act. Therefore, the correct choice is that Chad needs to act next, rather than feeling like playing or relaxing.
	2		The idiom "the ball is in his court" suggests that it is now Chad's responsibility to take action regarding Sophia's suggestions. This phrase indicates that the next move is up to him, implying that Sophia has done her part by submitting her suggestions. The other choice, which suggests Chad feels like playing or relaxing, does not align with the meaning of the idiom and misinterprets the context of the situation.
	3		The idiom "the ball is in his court" suggests that it is now Chad's responsibility to take action or make a decision following Sophia's suggestions. This expression originates from tennis, where the player must respond when the ball is in their side of the court, indicating that the next move is up to them. Therefore, the correct interpretation aligns with the idea that Chad needs to act next, making option (0) the accurate choice.
	4		The text uses an idiom, an expression that cannot be understood literally. The idiom the ball is in his court suggests that Chad needs to act next. In tennis, when the ball is in a player's court, it is that person's turn.

Table 4: Comparison of GPT-4 outputs under different settings.