Learning Dynamics of Multitask Training Data in Vision Language Models

Tyler Zhu
Princeton University

Koome Murungi*
Swarthmore University

Polina Kirichenko
FAIR at Meta

Olga Russakovsky
Princeton University

Figure 1: **We label each question in the LLaVA dataset as corresponding to one of five skills**, allowing us to understand the learning dynamics of individual skills and their effects on each other. (Left) Example LLaVA questions corresponding to our five target skills. (Right) Model accuracy on training (solid lines) and validation (dashed lines) examples throughout the training process.

Abstract

Vision language models (VLMs) are trained on massive amounts of data to perform many visual tasks simultaneously. Accordingly, many VLM benchmarks have been recently created to properly evaluate the models' capabilities. However, relatively little has been done to understand how and when the model acquires particular skills during training. We evaluate checkpoints throughout a one-epoch VLM training on recently seen and unseen datapoints to capture the generalization dynamics during model learning. We categorize the training data into five broad visual reasoning groups (Bounding, Complex, Object, OCR, and Semantic questions) and observe when these skills are learned. We note for example that despite not being explicitly trained to do OCR, VLMs can quickly learn to perform OCR tasks better than object recognition tasks. Digging deeper, we perform a case study on how VLMs use visual cues to solve OCR questions, indicating a form of shortcut that is not captured by standard VLM benchmarks. In contrast to OCR questions which are quickly learned, bounding capabilities are inefficiently learned due to the the complexity of the bounding box format – despite the fact that bounding box questions comprise the majority of the training data. Our work provides a glimpse into the underlying learning process of VMs on the LLaVA dataset.

1 Introduction

Vision Language Models (VLMs), such as Flamingo and LLaVA [1, 9], achieve impressive performance across diverse visual tasks through massive multi-task pretraining on image-text pairs. These

*Work performed as a research assistant at Princeton University

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Workshop on Space in Vision, Language, and Embodied AI (SpaVLE).

Category	Ratio	Description
Bounding	34.5%	Provide a $[x_1, y_1, x_2, y_2]$ bounding box or describe such a region
Complex	13.8%	Multi-step reasoning and logical deduction, often image-free
Object	19.8%	Relating to specific object(s), e.g., recognizing, counting
OCR	14.2%	Reading printed text and semantic queries about them
Spatial	17.7%	Spatial relationships of object(s) and between them

Table 1: **Category Details**. Ratios averaged over 7 sampled sets and 8 random seeds (gpt-4o-mini).

models can seemingly learn any visual reasoning task, from object detection to OCR [2, 7, 15], given sufficient data. However, the source of this generalization remains poorly understood, and current evaluation methods provide limited insight into how VLMs actually learn during training.

The standard approach of evaluating on held-out validation sets [16] cannot distinguish between what knowledge the training data provides and how well the model learns from it. This is particularly problematic for VLMs trained under large language model loss objectives, where cross entropy differs significantly from model accuracy. We lack metrics to track accurate learning dynamics on the training data itself. Some works are beginning to identify how metrics such as perplexity and math multiple choice accuracy correlate to study the learning dynamics of LLMs for reasoning [4], but similar approaches for VLMs remain unexplored and elusive.We address this gap by developing an evaluation protocol that captures VLM training dynamics in the one-epoch setting. Our approach uses an LLM to judge the accuracy of the model on the open-ended training data, and categorizes visual reasoning skills into five distinct categories (Bounding, Complex, Object, OCR, and Spatial, summarized in Table 1) to identify how different visual skills perform. We track this performance on both recently seen and unseen examples throughout training, demonstrating how much information has been successfully captured from the recently-seen training data and to what extent has this information been generalized to new unseen examples.

Our key findings challenge common assumptions about VLM learning. Despite constituting only 14.2% of training data, OCR questions achieve the highest accuracy by the end of training, while bounding questions, despite being the majority (34.5%) of training data, struggle due to the inefficiency of coordinate-based representations. We demonstrate that VLMs learn OCR through visual shortcuts rather than traditional text recognition, revealing that high accuracy scores can be misleading about the underlying learning mechanisms, and that some skills like Object and Spatial reasoning generalize well while others like Complex reasoning plateau early. These insights provide a foundation for more data-effective VLM training and highlight the importance of evaluating training dynamics directly.

2 Creating categories and an evaluation protocol

We quickly establish the standard procedure for training VLMs on the LLaVA dataset. Our categorization scheme factors the multitask LLaVA data into corresponding visual capabilities to analyze their individual learning dynamics. This is made possible by separating examples into "seen" and "unseen" sets throughout training to measure generalization.

For the base of our studies, we adopt the LLaVA 1.5 dataset [10] using the Prismatic VLM design [5] of a one epoch, Stage 2 only training recipe with a DINOv2-SigLIP vision backbone [13, 17].

LLaVA Dataset. We adopt only the visual instruction tuning portion of the dataset, as these directly correspond to specific visual reasoning skills. This portion consists of image-text pairs derived from five vision datasets: COCO [8], GQA [3], OCRVQA [12], TextVQA [14], and VisualGenome [6]. Each example consists of 0-1 images and an average of five turns of questions and answers which are generally independent. In total, the dataset has 665k examples and 3.4 million question answer pairs.

Visual Skill Categories. While the LLaVA dataset is comprised of multiturn conversations, there are rarely dependencies between turns. We filter out those examples and split the remaining examples by turn. Analyzing the resulting questions, we found 5 overarching task categories: Bounding, Complex, Object, OCR, and Spatial. Examples of each category are shown in Figure 1 and detailed in Table 1.

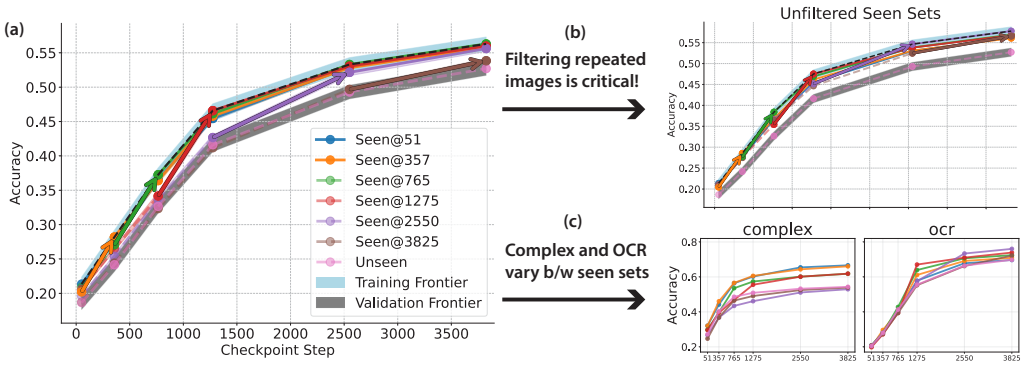

Figure 2: **Our LLaVA training dynamics exhibit a "train" frontier and a "validation" frontier.** As sets become seen for each checkpoint, each line becomes solid and jumps from the unseen level to the seen level. There is still a clear seen-unseen generalization gap, hinting that more learning is possible. Results are averaged over 8 random seeds with 1σ widths.

Evaluation Protocol. Evaluating how well the model answers open-ended questions is a non-trivial task. We adopt the field standard of using an LLM as a judge of the model's correctness [11]. Each ground truth question, answer and model response was provided to text-only gpt-4o-mini, and judgements were calibrated using a rubric which aligned well with human ratings (Appendix A.1).

To model the training dynamics, we fix a set of checkpoint steps as snapshots of different phases of the model's training to evaluate on, and to prevent evaluating a prohibitively large number of questions. These are steps 51, 357, 765, 1275, 2550, and 3825 (out of a total of 5120 steps).

However, designing training and validation sets in a one-epoch regime requires care. A standard multi-epoch setup has a static training and validation set, the former being trained on in its entirety right before validation while the latter is never. In the one-epoch regime, we never see the same training example twice within a training run. Therefore, to mirror the standard setup as closely as possible, we create one seen set for each checkpoint consisting of the 5,000 most recently seen samples to ensure the freshest evaluation of capabilities, denoted as Seen@k for step k. Our validation set, or unseen set, is sampled from beyond the last selected checkpoint to ensure the questions are unseen. In this way, Seen@k is a validation set for checkpoints before k and a training set afterwards.

3 Results

Emerging Training and Validation Frontiers. Our LLaVA training dynamics reveal two distinct curves (Figure 2a): a "training" frontier tracking examples seen during training, and a "validation" frontier tracking unseen examples. These are highlighted as bands based on the highest seen and unseen accuracies for each checkpoint. Notably, there is an observable train-val gap, offering insight into how much the model generalizes from the data.

As validation of our methodology, we observe the following. For each Seen@k set, prior to checkpoint k, performance matches the validation frontier as expected for an untrained set. Once the model trains on the set, performance jumps to the training frontier, remaining consistent across all seen sets. We mark this transition with arrows in Figure 2a. The last checkpoint is an exception, which we attribute to the low learning rate near training end due to cosine annealing.

Determining what constitutes an "unseen" example proves critical to observing this behavior. A natural criteria would be to consider each $(\mathcal{I}, \mathcal{T})$ image-text pair jointly, making every example different. However, images are used in an average of 3 different conversations, and seeing an image once is enough to alter our validation frontier (Figure 2b). The same is not true of text, likely due to the vast pre-training the LLM has already received.

OCR learns quickly while Bounding struggles and Complex plateaus early. Using our categorization scheme, we analyze each category's performance throughout training. A clear pattern emerges (Figure 1, right). Object, Spatial, and Complex questions all start with similar accuracy at

step 51, while OCR is slightly lower and Bounding is nearly 0%. Over time, Object, Spatial and OCR questions improve to around 65% accuracy, with OCR surprisingly becoming the best performer. Complex questions plateau quickly, suggesting the model cannot learn these effectively. Bounding improves slowly, which is disappointing given that 34.5% of questions are Bounding questions, though this is not unexpected due to the difficulty of the bbox format and pseudo-regression objective.

Examining generalization gaps, Object and Spatial questions show acceptable gaps, while OCR has a large gap between seen and unseen. Complex demonstrates odd behavior, sometimes performing better on unseen than seen sets, suggesting the model fails to learn the data or that patterns are weak in this category. Bounding performs similarly between seen and unseen, which makes sense given the specificity of object detection tasks leaves little to overfit on.

Complex and OCR dynamics differ throughout training. One concern with our protocol is that seen sets may drift from the overall data distribution, as we are resampling and filtering them for each checkpoint. To investigate this, we plot each category's learning dynamics across every seen set (Figure 2c). Only Complex and OCR show strong variance with seen sets, while Bounding, Object, and Spatial questions remain tightly grouped without temporal artifacts from the one-epoch setting. Complex questions perform worse for later sets than earlier ones without much difference between checkpoints, which we attribute to a lack of strong language-centric data in LLaVA causing the model to relinquish complex text understanding abilities and plateau early. OCR exhibits different dynamics where the best performing set for each checkpoint is the corresponding seen set, followed by previous ones in order of recency, pointing to strong memorization and weak generalization.

VLMs take surprising paths to learn Bounding and OCR. We conclude with an enlightening case study of Bounding and OCR questions. Bounding questions divide into two forms: "Describe" questions where the model describes a provided bounding box region, and "Bbox" questions where the model returns bounding box coordinates for a specified description.

Intuitively, one might expect describe questions to perform much better due to stronger text supervision, but both types perform poorly. As shown in Figure 3, the model performs slightly better when asked to return a region description as expected. However, the performance gap is small because describe questions also involve floating coordinates.

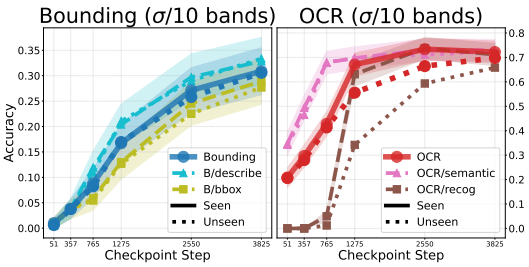

Figure 3: **Bounding and OCR have counter-intuitive learning dynamics** when investigating their subcategories.

OCR questions in the LLaVA dataset are nearly entirely about books and divide into two subcategories: recognition and semantic questions. Recognition questions require direct transcription of the title or author. Semantic questions require external knowledge or logical reasoning, such as determining a book's genre or whether it's for children.

Surprisingly, VLMs learn semantic questions much faster and earlier than recognition questions, circumventing the expected learning order entirely! This shows that recognition questions are difficult, especially given the large generalization gap, and that the VLM likely relies on other visual cues to solve semantic questions rather than first learning to read text.

4 Conclusion

We have presented a new methodology for evaluating the training dynamics of VLMs in a one-epoch setting. Under this setting, we have shown that two distinct frontiers emerge which can capture generalization: a "training" frontier, which tracks examples seen during training, and a "validation" frontier, which tracks unseen examples. We take this one step further by categorizing the questions to understand the learning dynamics of each category and how some categories learn faster than others. Limitations of our method are that we only perform analysis on a single type of VLM architecture and on one dataset, albeit one of the most popular ones. Additionally, LLM-as-a-judge is expensive and not feasible for development, so an equally informative signal or automatic metric could make such understanding more widespread. This is a promising direction for future work.

Acknowledgements

All experiments, data collection, and processing activities were conducted at Princeton University. Meta was involved solely in an advisory role and no experiments, data collection or processing activities were conducted on Meta infrastructure. We thank William Yang and Honglu Zhou for detailed comments and feedback.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. NeurIPS, 2022. 1

[2] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852, 2021. 2

[3] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR, 2019. 2

[4] Katie Kang, Amrith Setlur, Dibya Ghosh, Jacob Steinhardt, Claire Tomlin, Sergey Levine, and Aviral Kumar. What do learning dynamics reveal about generalization in llm reasoning? arXiv preprint arXiv:2411.07681, 2024. 2

[5] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In ICML, 2024. 2

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73, 2017. 2

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022. 2

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023. 1

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/. 2

[11] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In ACL, 2024. 3, 7

[12] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In ICDAR, 2019. 2

[13] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. In TMLR, 2023. 2

[14] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In CVPR, 2019. 2

[15] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860. 2

[16] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567, 2024. 2

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In ICCV, 2023. 2

A Technical Appendices and Supplementary Material

A.1 LLM-as-Judge

To evaluate the performance during training, we utilized text-only GPT-4o-mini as a judge. For each question, we pass in a tuple consisting of the question, ground truth answer, and predicted answer, and ask the LLM to decide if the question was answered correctly or not. Following Video-ChatGPT [11], we create a rubric that guides the model to first provide an overall score before returning a yes/no prediction for better calibration. See Figure 5 for the rubric.

	Acc	Prec	Recall	F_1 Score
LLM (w/ rubric)	**89.4**%	93.0%	**89.8**%	**0.91**
LLM (w/o rubric)	85.1%	**97.9**%	78.0%	0.87

Table 2: **An LLM** (gpt-4o-mini) **serves as a suitable judging replacement for humans.** Our rubric helps improve the overall accuracy and consistency as shown by the better F_1 score.

The rubric consists of three scores ranging from 1-5 which help the LLM better judge under the text-only setting: Missed details, Hallucinations, and Major Subjects. Missed details describes how many ground truth answer details were left out in the predicted answer. Hallucinations describes how plausible the added details (hallucinations) are in the predicted answer. Major subjects compares the major subjects in ground truth and predicted answer. A score of 5 is the most accurate, and score of 1 represents a bad and weak answer. Each score is averaged to get a general prediction score for each question-answer pair.

We conducted a small human study to validate both the use of GPT-4o-mini as a judge, as well as our rubric for improved calibration. The human study was conducted on 4 individuals, each receiving the same 100 question-answer-prediction tuples as the judge. The subjects were asked to rate if each prediction was plausibly correct given the ground truth, and a threshold of 75% agreement was used for the final human ratings. As shown in Table 2, we found that the LLM on its own was a suitable replacement for humans, achieving 85.1% accuracy. However, including the rubric boosts its accuracy by 4.3% to 89.4%, and improves its overall precision-recall tradeoff as shown by the 4 point increase in F_1 score. We adopt this rubric evaluation scheme by default for our methodology.

For completeness, we also include the prompt used to categorize the training data examples into categories in Figure 6.

A.2 Full Figures

We present the full plots of Figure 2(c) here in Figure 4, showing that the other three categories are consistent across seen sets while Complex and OCR vary.

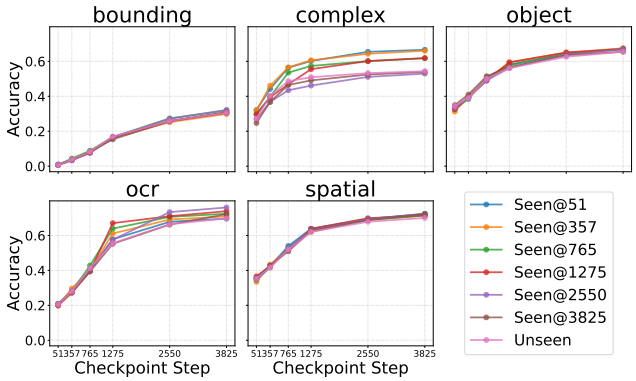

Figure 4: **Complex and OCR questions vary with seen set, while other categories are consistent.** Complex questions show degraded performance in later sets, suggesting interference from newly acquired visual knowledge. OCR performance indicates memorization, as seen accuracy never drops below unseen performance.

```
System: You are an intelligent chatbot designed for evaluating the correctness of
generative outputs for question-answer-prediction tuples. Your task is to
grade the pred answer given a correct answer and a rubric, then provide a
final score.
------
##Rubric:
Do not grade too harshly
In example sentence, 'a woman in a brown shirt, holding a white purse, is
hugging a bear', major subjects are 'a woman' and 'a bear'. The details
would include everything related to major subjects. Such as 'in a brown
shirt', 'hugging', 'holding a white purse'.
Also subjects can be the same/similar but have different names. For example,
a bag and suitcase would be the same subject.
When comparing, focus on the meaning rather than exact language.

Major Subjects: Graded 1 - 5, with 5 being the highest score. Give a score of:
1 if no major subjects remotely similar.
2 if a few (>2) subjects are not the same.
3 if some (2) subjects are not the same.
4 if subjects that are different are also semi-plausible given context of answer
5 if subjects that are different are also plausible given context of answer

Missed Details: Graded 1 - 5, with 5 being the highest score. Give a score of:
1 if many (>5) key details missed that change the meaning of the answer
2 if some (~5) key details missed
3 if some (2 to 4) non-important details missed
4 if few (2) non-important details are missed
5 if very few to none (<2) non-important details are missed

Hallucinations: Graded 1 - 5, with 5 being the highest score. Give a score of:
1 if all details that are added are not plausible
2 if some (>3) details added that are not plausible
3 if few (~3) details added or if added details are semi-plausible
4 if very few (1 to 3) details are added or that details added are plausible
5 if no details added or details added are plausible

Please evaluate the following image-based question-answer-prediction tuples
with the given rubric.

Question: ..., Correct Answer: ..., Predicted Answer: ...
Provide your evaluation only as a yes/no, score for Major Subjects, score for
Hallucinations, and a score for Missed Details where both scores are an
integer value between 1 and 5, with 5 indicating the highest meaningful match.
If the pred answer could be true given the context of the correct answer, then
evaluation should be yes, otherwise it should be no.
Please generate the response in the form of a Python dictionary string with
keys 'p', 'MS','MD', and 'H', where value of 'p' is a string of 'yes' or 'no'
and values of 'MS','MD', and 'H' are in INTEGER, not STRING.
p stands for prediction, MS for Major Subjects, H for Hallucinations and MD
for Missed Details
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python
dictionary string.
For example, your response should look like this:
{'p': 'yes', 'MS': 5, 'MD': 4, 'H': 3}.
```

Figure 5: **The rubric used to judge the VLM responses.** We use the scores as a way to calibrate the LLM's responses before asking it to give a final score. The score is generally calibrated with the accuracies, so that the accuracies within each score band are roughly proportional to the score.

```
System: You are an intelligent chatbot designed for
categorizing different types of questions. Your task is to group questions/tasks
into 4 categories: Object Analysis, Spatial Analysis, Bounding Box, and Complex
Reasoning.
------
##INSTRUCTIONS:
- Focus on the goal of the questions.
- Look at the examples of each category to aid in categorization. Examples of
each category are the following:

Object Analysis:
Which kind of appliance is it?
Are there either any doors or windows that are made of metal?
Is this a transportation engineering book?
What color are the umbrellas in the picture?

Spatial Analysis:
Are there any players to the left of the helmet on the right?
What kind of device is to the left of the computer monitor?
What is near the bottle of alcohol?
A. bunny
B. toilet
C. whistle
D. man
Answer with the option's letter from the given choices directly.
Who is in the water on the beach?

Bounding Box:
Please provide the bounding box coordinate of the region this sentence
describes: man with royal blue and white toothbrush in mouth.
Please provide a short description for this region: [0.06, 0.78, 0.93, 0.83].

Complex Reasoning:
Why are the cats resting?
What can we infer about the elephants' social behavior from this scene?
What potential reasons might explain the unattended devices in this scene?
What is the genre of this book?
What activities might someone enjoy in this well-lit room?

Please categorize the following question into the 4 categories: Object Analysis,
Spatial Analysis, Bounding Box, and Complex Reasoning. DO NOT PROVIDE ANY OTHER
OUTPUT TEXT OR EXPLANATION. Return O if Object Analysis, S if Spatial Analysis,
B if Bounding Box, or C if Complex Reasoning.
```

Figure 6: **The prompt used to categorize the training data examples into categories.** We use the scores as a way to calibrate the LLM's responses before asking it to give a final score. The score is generally calibrated with the accuracies, so that the accuracies within each score band are roughly proportional to the score.

A.3 Societal Impacts

Like any work involving generative models, specifically LLMs, there are potential harms without significant safeguards. The models we use should have been tuned to be safe, but there are always risks. The datasets we use are among some of the most widely used, so they have been through much scrutiny. In general, we hope that our work helps to shed a light on how exactly these models gain their capabilities, which previously has been a relatively opaque process.

A.4 Compute Resources

All experiments were run on a cluster of 48GB L40s. The storage of the checkpoints was the limiting factor, which required some smart management of the checkpoints.