LingoQA: Video Question Answering for Autonomous Driving

Ana-Maria Marcu*, Long Chen*, Jan Hünermann*,
Alice Karnsund*, Benoit Hanotte*, Prajwal Chidananda, Saurabh Nair,
Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Oleg Sinavski*
Wayve
research@wayve.ai
*Equal contributions

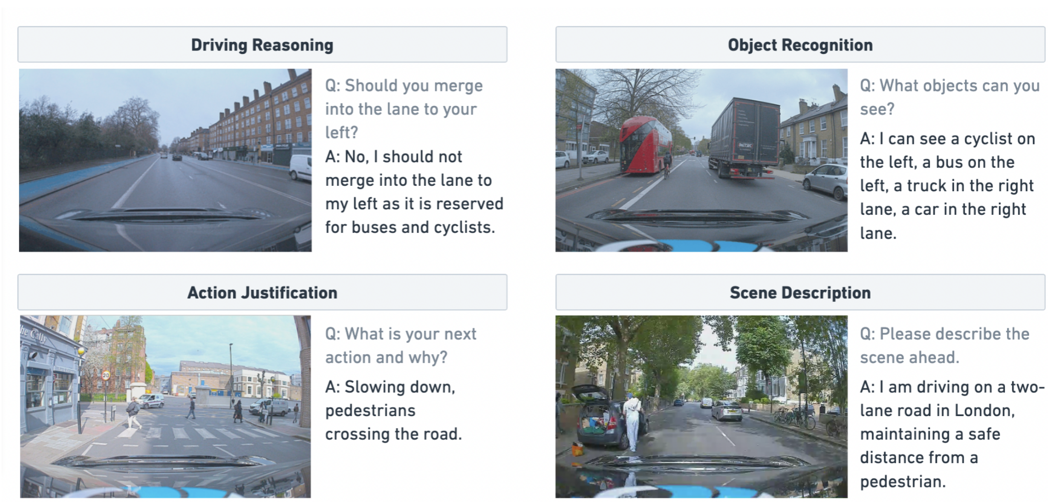

Figure (1) LingoQA is a comprehensive *benchmark* for Video Question Answering in autonomous driving. Our baseline vision-language model on this benchmark, can answer questions related to driving reasoning, object recognition, action justification, and scene description.

Abstract

Autonomous driving has long faced a challenge with public acceptance due to the lack of explainability in the decision-making process. Video question-answering (QA) in natural language provides the opportunity for bridging this gap. Nonetheless, evaluating the performance of Video QA models has proved particularly tough due to the absence of comprehensive benchmarks. To fill this gap, we introduce LingoQA[1]*, a benchmark specifically for autonomous driving Video QA. The LingoQA trainable metric demonstrates a 0.95 Spearman correlation coefficient with human evaluations. We introduce a Video QA dataset of central London consisting of 419k samples that we release with the paper. We establish a baseline vision-language model and run extensive ablation studies to understand its performance.*

[1]https://github.com/wayveai/LingoQA

1. Introduction

Communication plays a pivotal role in naturally fostering trust among individuals. However, establishing trust between users and agents remains a significant challenge within the field of artificial intelligence. Recent studies have indicated that articulating explicit reasoning steps can significantly enhance user confidence [1], in addition to improving the capabilities of Large Language Models (LLMs) [52]. The need for explainability remains critical, particularly in safety-critical domains where technology adoption hinges upon this factor [31].

Consider the domain of end-to-end autonomous driving [11], where the driving policy is often executed through deep neural networks processing camera inputs to generate control commands. Despite their efficiency, these systems have grappled with the persistent challenge of providing transparent and interpretable decisions. Integrating Vision-Language Models (VLMs) into the field of autonomous driving holds the promise of enhancing user trust in these systems. Recent strides in VLMs have solidified transformers as multimodal learners, showcasing remarkable performance in tasks such as Visual Question Answering (VQA) and underscoring their proficiency in acquiring robust representations for complex tasks [14].

Our focus is on vision-only end-to-end autonomous driving, aiming to bridge the gap between data-driven decision-making and user trust. We introduce LingoQA, a benchmark designed for autonomous driving video QA, utilizing a novel dataset comprising more than 419k QA pairs. Distinguished by its free-form approach to questions and answers, this dataset broadens the scope of autonomous driving video QA, encompassing reasoning and action justifications. Additionally, we publish a comprehensive evaluation suite consisting of 1,000 examples. At the core of our benchmark lies a novel evaluation metric based on a learned text classifier called *Lingo-Judge*, inspired by GPT-Judge used in TruthfulQA [35]. We perform rigorous studies correlating automatic metrics to human preferences and find that Lingo-Judge achieves a 0.950 Spearman and 0.993 Pearson correlation coefficient, surpassing existing automated labelling techniques like METEOR [5], BLEU [42], CIDEr [49], and GPT-4 [41] on our benchmark, while being fast enough for frequent runs during training and development. The evaluation code and the weights for the classifier will be released with the paper to support robust benchmarking of autonomous driving explainability.

Equipped with this evaluation toolkit, we conducted a comprehensive empirical study on key components and their ablations in VLMs for autonomous driving. Our findings in Section 5 indicate that the most effective approach involves partially fine-tuning the attention layers of our vision-language model equipped with Vicuna-1.5-7B [13], on both Action and Scenery datasets. This process involves using 5 video frames over 4 seconds and a late video fusion technique. Our collective work, spanning the LingoQA benchmark, the visual instruction-tuning dataset, and the innovative evaluation metric, aims to propel the domain of explainable autonomous driving, laying a robust foundation for subsequent research and development endeavors. To summarise the main contributions of this paper:

- **LingoQA Benchmark**: We introduce a novel benchmark for autonomous driving video QA using a learned text classifier for evaluation. It outperforms existing metrics, including GPT-4, with a Spearman coefficient of 0.950 indicating a strong correlation with human evaluation.
- **LingoQA Dataset**: Our 419.9k QA pair dataset stands out with its free-form questions and answers, covering not just perception but also driving reasoning from the drivers directly, broadening the scope of autonomous driving video QA.
- **LingoQA Baseline**: Through testing of various video-language components on LingoQA, we find that the most effective approach involves partially fine-tuning the attention layers of our vision-language model equipped with Vicuna-1.5-7B [13] and a late video fusion technique. We establish a new baseline for this field with an identified model combination. Example outputs from the model are shown in Figure 1.

2. Related work

2.1. Language in Autonomous Driving

Modern autonomous vehicle software relies heavily on artificial intelligence models [6, 20, 21, 23]. This, together with the increased number of such vehicles on the road, poses a fundamental challenge in terms of interpretability in the decision-making process [4]. Understanding why a decision is made is crucial for understanding areas of uncertainty, building trust, enabling effective human-AI collaboration, and ensuring safety [54]. In a survey conducted by Partners for Automated Vehicle Education (PAVE) in 2020 [1], 60% of participants stated that they would trust AVs more if they better understood the underlying process of the system. To establish trust with the general public, the inner workings of systems must be explained in a human-interpretable way, such as through language and visual explanations.

The field of autonomous driving has been embracing the opportunity to make end-to-end driving models more explainable using language, understanding that methods based purely on attention [31] are not sufficient [27]. The early explorations of GPT3.5 [39, 46] and GPT4-V [53] on autonomous driving scenarios show that LLMs/VLMs demonstrate superior performance in scene understanding and causal reasoning compared to existing autonomous systems. Works such as ADAPT [29] and LLM-Driver [10] propose multi-task learning frameworks for jointly predicting language and control outputs. Inspired by progress in large language models [13,41,47,60], vision-language models [3, 7, 12, 16, 32, 36, 41, 50, 51, 57–59] and multi-modal transformers for robotics [8, 9, 19]. Closely related to our proposed baseline is DriveGPT [56], proposing a multi-modal vision-language-action model that tokenizes videos, as well as text and control actions.

2.2. Evaluation Metrics

Progress has been relatively slow for developing vision-language models for autonomous driving, with only a few works aiming to quantitatively improve upon prior work

[29, 30, 56]. A key challenge consists of automated, reproducible evaluation metrics that are highly correlated with human ratings, particularly due to the inherent complexities in assessing natural language. ADAPT [29] reports human feedback in addition to standard natural language generation metrics, while DriveGPT [56] reports ChatGPT ratings. Automated methods such as BLEU [42], METEOR [5], ROUGE [34] show weak alignment with human feedback [49]. CIDEr [49] is also based on *n-gram* level similarity, as opposed to capturing the correctness of an answer based on its meaning. Newer evaluation metrics using ChatGPT have shown improvement in the area of sentence understanding, while still having limitations, such as providing high scores to elaborate, eloquent, but incorrect sentences [2]. Evaluation based on human feedback is subjective and difficult to reproduce. In this work, we address this challenge by introducing a novel video QA benchmark for autonomous driving that checks for factual correctness and is highly correlated to human correctness ratings on our proposed evaluation dataset.

2.3. Datasets for Autonomous Driving

Recent advances in generative AI have been underpinned by training with increasingly large and diverse internet-scale datasets. Nonetheless, when evaluated zero-shot on autonomous driving datasets, many pre-trained *models* such as Flamingo [3] and BLIP-2 [32] fall short of expectations. Consequently, there remains a need for high-quality training datasets for autonomous driving, as well as for reliable benchmarks. Autonomous driving datasets have been focused on commentary, as opposed to question-answer pairs [31, 55]. However, we note that action, justification is a relatively easy task for foundation models, that are capable of efficiently compressing large amounts of data [17]. The advantage of VQA is that related, but slightly different questions can be asked to truly probe the validity of the model representations. Further datasets focus on ranking the importance of elements in the scene and explaining the reasoning behind the choice [38, 45]. The datasets available for VQA have been previously constructed around existing object detection datasets [18, 43]. Datasets such as NuScenesQA [43], while useful for testing perception, contain simple language outputs of on average one word per question that do not tackle the more challenging reasoning problem.

Our proposed dataset LingoQA addresses the existing gap in autonomous driving as it contains a diverse set of questions related to *driving behaviours* and *scenery* in addition to perception questions related to object presence and positioning. This dataset has the strength of being diverse with respect to the language used while being grounded in human reasoning. Examples of the complex questions and answers existent in the dataset are provided in Appendix A. Moreover, a significant advantage of LingoQA is that the intentions and reasoning behind driving decisions are directly labeled by the drivers themselves. This provides an accurate understanding of driving dynamics.

3. LingoQA Benchmark

In this section, we introduce LingoQA, a benchmark to evaluate video question-answering models for autonomous driving.

3.1. Evaluation Dataset

We collected a small, low-density dataset from in-house human labelers, creating both the questions and the answers associated with the short videos. We labeled a small portion of held-out data on 500 human-generated questions using 20+ different evaluators to obtain our test set. Since answers are subjective and noisy, we labeled them twice, making sure the same evaluator does not receive the same question twice. After that, we manually reviewed the answers for semantic disagreements and mistakes. We relabeled such samples two more times and fixed the disagreements, preferring the semantics of the majority of responders but preserving maximal variety in the responses. Finally, we condensed this into 1k high-quality answers to 500 questions, with two correct but diverse answers per question. The dataset evaluates a range of competencies, including action and justification, attention, description, localisation, identification, counting and anticipation, as shown in Figure 2.

3.2. Evaluation Metric

Evaluating open-ended textual dialogues is a challenging task. Quite often the correct answers are ambiguous, subjective, or even not attainable. The most common language-based metrics for evaluating question-answering models in autonomous driving [29, 31, 56] are BLEU [42], METEOR [5] and CIDEr [49], despite their known limitations, such as relying heavily on the *n-gram* frequency as opposed to the underlying meaning of the answer. To address these limitations, we set ourselves the challenge to develop an *automated*, *non-visual* evaluation method for free-form language answers from vision-language models which checks correctness independent of phrasing against a ground truth answer and which is *highly correlated* with human ratings.

GPT-4 based evaluation Inspired by the G-Eval metric [37], we used GPT-4 to evaluate answers on a larger scale. Given a question and answer pair from the test set and a model's answer, we ask GPT-4 to evaluate whether the model's answer corresponds to a human's answer. Notice that it doesn't make use of any visual input. We experimented with prompts and methods achieving good quality of judgements. We achieved the highest accuracy by employing chain-of-thought prompting where we ask GPT-4

	Pearson	Spearman	Val Acc. [%]	Time [sec]
Lingo-Judge	**0.993**	**0.950**	**95.0**	10.5
GPT-4 with CoT	0.990	0.932	91.2	3016.0
GPT-4 [41]	0.988	0.941	90.6	812.4
BLEU [42]	0.881	0.835	-	**0.1**
METEOR [5]	0.891	0.876	-	8.0
CIDEr [49]	0.878	0.853	-	0.2

Table (1) **Lingo-Judge Performance.** Correlation with human ratings, validation accuracy, and time taken to run of our proposed LingoQA evaluation metric compared to previous language-based metrics. All metrics use textual ground truth and have no access to vision information. Further examples are presented in Appendix B.

to first come up with an evaluation strategy before grading a model's answer. However, as shown in Table 1, this leads to increased inference time. Further details are provided in Appendix C. Unfortunately, we found GPT-4 based evaluation impractical to use as a main development and training metric due to the time required to evaluate answers on our relatively small evaluation dataset (from 13min up to 50min for a single evaluation due to the API rate limit).

Lingo-Judge Given these limitations and inspired by TruthfulQA GPT-Judge [35], we pursued an alternative approach using a learned text classifier, dubbed Lingo-Judge, which estimates the correctness of model answers. We measure the correctness of model predictions as an accuracy using a small transformer-based text classifier that takes in a question, the human's, and the model's answer and outputs a probability that the model's answer is correct. Please note, Lingo-Judge does not receive video input and must rely only on the supporting human's answers. For every question, we run Lingo-Judge on all combinations of *(ground-truth answer, predicted answer)* and take the maximum correctness estimate, as shown in Equation 1, where S is the score per sample. We found this recipe yield the best predictive power provided enough diversity of human answers in our evaluation dataset.

$$S = \max_{j \in \{0,1\}} F_{Judge}(pred, ground\_truth[j]) \quad (1)$$

The architecture of the classifier is a DeBERTa-V3 [22] language model, fine-tuned with LoRA [24]. The classification score is predicted using a linear head on top of the class token output. We fine-tuned the model on a diverse dataset of model predictions from early experiments, where questions and ground truth answers come from our evaluation dataset and the correctness target is labeled by human annotators. On top of this initial dataset, we iteratively improved the classifier using active learning by correcting the wrong predictions of discarded models and adding corrections to the training dataset. On a held-out test set, we find that the binary classification accuracy of the classifier is 95%.

	Scenarios	**QA pairs**	**QA per scenario**
Action	24.5k	267.8k	≈ 10.9
Scenery	3.5k	152.5k	≈ 43.6
Eval. Dataset	100	1000	10

Table (2) **Dataset Split.** It consists of three different datasets of varying annotation densities. The *Action* dataset focuses on questions related to driving behaviours, the *Scenery* dataset focuses on perception capabilities, while the evaluation dataset is designed to probe a range of competencies.

In comparison to metrics such as CIDEr, which provide a *system-level* performance metric, the classifier provides a probability of correctness for each of the model predictions, meaning that it provides metrics at the *sample* level. Examples are provided in the Appendix B. This means that 100% classifier accuracy is easy to interpret. The classifier allows us to compute metrics during training, running over our full evaluation dataset in 10 seconds using an A100 GPU.

Correlation to human ratings We studied empirical *correlation* of various metrics with human judgments. Several human annotators assigned a scalar score [0, 1] to the inference outputs of 17 different models which can be interpreted as the probability that the response correctly addresses the question [35]. Notably, this process takes several days, highlighting the need for an automated evaluation metric that provides faster development feedback. The final human score of each model is the average of all inference output scores. Further details regarding the methodology for the correlation analysis are in the Appendix D.

The *Spearman rank correlation* coefficient of our automated metric, Lingo-Judge, with human scores is 0.95, and the *Pearson correlation* coefficient is 0.993. These values are notably higher compared to other natural language evaluation metrics and GPT-4, as detailed in Table 1. Our analysis demonstrates that Lingo-Judge accurately mirrors human judgments, outperforming existing metrics such as BLEU, METEOR, and CIDEr, as well as GPT-4 with and without chain-of-thoughts prompting. This indicates that Lingo-Judge can effectively serve as a proxy for human labelling, which is particularly significant given the stagnant nature of metrics in autonomous driving since the introduction of the CIDEr metric in 2015. Notably, despite their limitations, prominent models like ADAPT [29] and DriveGPT [56] still use BLEU, METEOR, and CIDEr metrics and report ChatGPT ratings without analyzing their correlation to human preferences. Our work fills this gap by providing a reliable benchmark that better reflects human preferences.

	Scenarios	**Annotations**	**QA**	**Captioning**	**Video length [sec]**
Rank2Tell [45]	118	>118	✗	✓	20
BDD-OIA [55]	22.9k	35k	✗	✓	5
BDD-X [31]	6.9k	26k	✗	✓	40
NuScenesQA [43]	34k	460k	✓	✗	20
DriveLM [15]	30k	360k	✓	✓	20
LingoQA	28k	419.9k	✓	✓	4

Table (3) **Dataset Features.** The dataset that we introduce alongside our benchmark consists of questions related to object presence, as well as action, justification, attention, localisation, counting, anticipation, and counterfactuals. In total, it has a similar size to other driving-related datasets such as NuScenesQA, while having a much higher diversity and not being limited to questions related to object positioning.

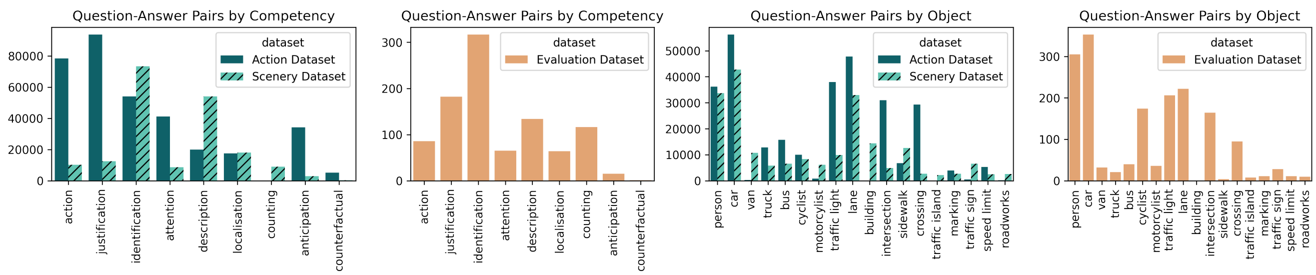

Figure (2) **Dataset Statistics.** Dataset split by the number of question-answer pairs for the competencies covered and for the objects referred. One question-answer pair might cover more than one competency or object, hence the total is higher than the size of the datasets. The *Action* and *Scenery* datasets have complementary strengths, with one focused more on action-justification competencies and one more on description and localisation.

3.3. Datasets

We created a collection of datasets for explainable driving. The total dataset size is 419.9k question-answer pairs, where a single data sample consists of a 4-second video clip at 1Hz. The total size of the dataset is about 10x larger than BDD-X [31], as shown in Table 3. Compared to prior datasets such as NuScenesQA [43], our dataset contains reasoning pairs in addition to object presence, description, and localisation. The answers are also more free-form and more complex, with an average answer length of 17.2 words versus 1.0 words in NuScenesQA. Examples of question answers pairs from LingoQA are shown in Appendix A.

Our labeled autonomous driving training dataset consists of two complementary parts: the *action* dataset and the *scenery* dataset.

Action dataset. This dataset was made from a recorded driving corpus of interesting events where the car's behavior changes, such as decelerations, accelerations, lane changes, narrow gaps, and turns. Such events were succinctly labeled by driving operators with very short high-level descriptions of the situations and behavioral policies (e.g. "following lane, pedestrian on a zebra crossing, should stop"). Additionally, we added metadata for such events from various perception systems, such as traffic light presence, vehicles and pedestrian visual detectors, weather descriptors, as well as other metadata (speed, steering wheel position, and road type from the map data). Using this data, we developed prompt templates for (1) describing the current action and its justification and (2) a set of example questions and hints about what the answer should mention. Next, we used those prompts with GPT-3.5 to rephrase, answer, and extend the example questions using the provided action description and answer hints. We rebalanced events by bucketing by actions and behavioral policies and sampled up to 500 events from each bucket without replacement, leading to 24,577 video snippets with 167,774 question/answer pairs.

Scenery dataset The scenery dataset was built to complement the action dataset by focusing on perception-related question in addition to driving behaviours. The dataset was made by densely and thoroughly labelling three 30-minute driving sessions with the ELAN video annotation software [40]. For the entire duration of the driving sessions, we provided short captions in about 15 different categories:

- Driver's actions, and their justifications
- Driver's attention
- Observations about relevant vehicles, pedestrians, and other road actors with their visual descriptions
- Observations about relevant static road elements such as traffic lights, traffic islands, lane and intersection structures

- Miscellaneous observations about the environment, such as weather, tube stations, and buildings.

Then, for every keyframe every second (1fps), we collect all annotations around this frame and build a textual description containing the driver's actions and their justifications, the objects requiring the driver's attention, and the observations. As opposed to the Action dataset, where recommended questions were provided to GPT-3.5 for rephasing, for the Scenery dataset, we asked GPT-4 to generate questions and answers using a set of generic prompts, but also using a prompt with the chain of thought specifically targeting perception questions. This forced GPT-4 to generate many diverse questions and answers. This led to a high-quality diverse dataset with about 43 QA-pairs per video.

Our training dataset covers 9 different competencies: action (what the vehicle is doing), justification (why the action is taken), attention (what should be paid attention to in the current situation), identification (identifying an object given its description), localisation, description, counting, anticipation and reasoning given counterfactuals. The questions also cover a diverse set of objects, such as pedestrians, vehicles, cyclists, buildings, road infrastructure, signs, markings. In Figure 2, we present the number of question and answer pairs for each of the 9 competencies above, as well as for the referred objects, for our two datasets, namely Action and Scenery. The complementary strengths of the datasets are apparent, with one focused on driving behaviours and one on perception tasks.

4. Model Methodology

We propose LingoQA Baseline, a vision language model for autonomous driving based on Vicuna v1.5 [13] with 7B parameters that takes explainability beyond simple image captioning and can answer reasoning questions grounded in video outputs. We train a model that consumes a short video segment and produces answers to autonomous driving-related questions.

4.1. Architecture

The LingoQA Baseline model architecture is based on recent VLMs [19, 32, 36] but enhances them by incorporating a video encoding strategy to process multiple frames from a video snippet, as is shown in Figure 3.

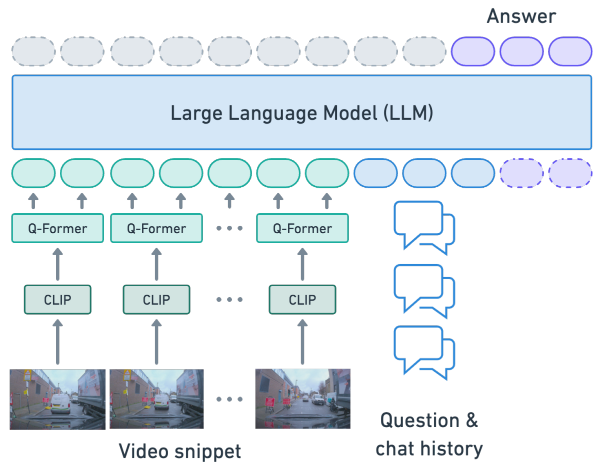

Figure (3) **LingoQA Baseline model architecture.** We first encode individual frames using CLIP and Q-Former. The Q-Former outputs tokens and we feed the tokens from all frames along with chat history and questions into the LLM, which then predicts an answer.

Vision encoder. We use CLIP [44], a Vision Transformer (ViT) pre-trained contrastively on image-language pairs, to encode images into features. The inputs to the vision encoder are RGB images from the front camera. We squash the input images to a size of 224×224 as opposed to cropping them in order to keep the full image context. Subsequently, we pass the features through a transformer network, the Querying Transformer (Q-Former), that akin to BLIP-2 [32] acts as a bridge between the vision and language feature spaces. The embeddings are then projected into the large language model (LLM) space using a linear projection layer. We repeat this process for $T = 5$ frames of the input video and concatenate the tokens from each image.

Large language model. We leverage pretrained LLMs to give LingoQA Baseline the ability to answer general questions related to both driving scenes, as well as general knowledge. We use Vicuna v1.5 [13] with 7B parameters built on top of Llama-2 [47]. The language model is autoregressive and hence can be conditioned on textual inputs, as well as image tokens. The training objective is to predict the next language token in a sequence. We mask all tokens from the training loss that belong to the text prompt, including question and chat history.

4.2. Training Recipe

Our training uses a two-step approach to better utilise video features and improve learning in answering questions based on video data. Through this two-step training, we aim for a better understanding and use of video data in addressing challenges in autonomous driving. During both stages, we train only the attention layers of both CLIP and the large language model, as well as the parameters of the Q-Former and language projection layer, while keeping all other parameters frozen. Further details regarding training parameters are presented in Appendix E.

Stage 1: Pre-training for feature alignment. In the first stage, we pre-train the model on the GQA and SVIT datasets to align image features with the embedding space of the pretrained LLM. The GQA dataset [25] contains more than 22M questions over 113k images. The recently introduced SVIT dataset [61] contains 4.2M question-answer pairs over 108.1k images. We leverage initial weights from different models to accelerate the training process. We initialise the vision encoder using publicly available weights of OpenCLIP [26], the Q-Former from BLIP2 weights [33], and language model from Vicuna v1.5 [13].

Stage 2: Fine-tuning for video QA. In the second stage, we fine-tune the model on our video question-answering Action and Scenery datasets described in Section 3.3. During the fine-tuning phase, each sample is composed of 5 frames taken from a 4-second span of video, accompanied by a QA-pair. To facilitate further exploration of autonomous driving QA, we open-source the dataset used to fine-tune LingoQA Baseline.

5. Empirical Evaluation on LingoQA

With the highly modular architecture of VLMs, the question remains what architectural components of the LingoQA Baseline model and dataset composition contribute the most to its performance? We conduct several ablation studies around the architecture and training paradigm described in Section 4. To this end, we investigate variations to the *training strategy*, *training data composition*, *frame count*, *video fusion methods*, and the use of different *large language models*. The results are obtained by having each model generate one answer per question and then compare the predicted answer to the two ground truth answers. Examples of comparisons between our baseline model's answers and answers from other models from the ablations are presented in Appendix F.

5.1. Training Strategy

The aim of the training strategy experiments is to understand how much the pre-training and the fine-tuning steps contribute to performance. It becomes apparent that fine-tuning is essential to yield answers relevant to autonomous driving. The model fine-tuned on the LingoQA dataset has nearly double the performance of the dataset that is pre-trained on generic VQA datasets. This shows that while generalised pre-training leads to improvements, task-specific fine-tuning is still required for optimal performance.

5.2. Training Datasets Mixture

Table 4 shows the contributions of our two datasets, *Action Dataset* and the *Scenery Dataset*. Both datasets proved influential in improving model performance. We show that fine-tuning on the LingoQA dataset that we open source leads to a considerable improvement compared to general pre-training only.

5.3. Impact of Frame Count

We want to investigate the variation in VQA performance with decreasing and increasing the number of video frames fed into the model. The base model contains 5 frames over a 4-second context. The performance declined when shifting from multi-frame video to a single image representation, which can be explained by the model not getting enough information to answer questions where temporal information is crucial. The performance when increasing the number of frames to 3, 5 and 7 remains relatively consistent. We hypothesise this is due to the lack of effective video fusion and leave it to future work to investigate video-language encoders further,

5.4. Impact of Video Fusion Strategy

Given how crucial temporal context is for scenario understanding, this study explores three methods for integrating video frames into the LLM: *early-fusion*, *mid-fusion*, and *late-fusion*. The *early-fusion* method employs average pooling to condense features from the vision encoder prior to their incorporation into the Q-Former, producing a unified visual feature vector for language space projection. In contrast, the *mid-fusion* approach, merges video features into fixed-size tokens within the Q-Former with the cross-attention mechanism. The *late-fusion* method, where individual frame embeddings from Q-Former output are fed into the LLM, allowing it to resolve temporal relationships. Our findings demonstrate that both *mid-fusion* and *late-fusion* are effective methods for incorporating video content into the model. *Mid-fusion* allows for a greater number of context tokens through the use of a predetermined number of video embeddings. Conversely, *late-fusion* shows a slightly enhanced performance by providing comprehensive frame information to the LLM.

5.5. Impact of Large Language Model

We investigate the impact that different Large Language Models have on the overall performance of our vision-language model. As shown in Table 4, the best score is achieved by the Vicuna-1.5-7B [13] that our base model uses. In the same model family Llama-2-7B [48] achieves comparable, but slightly lower performance. Mistral-7B [28], despite its promise of superior performance over Llama-2, proved less effective in our fine-tuning task. No-

	Ablation	Lingo-Judge [%] ↑	BLEU ↑	METEOR ↑	CIDEr ↑
	LingoQA Baseline	**60.80**	**15.00**	18.56	**65.61**
Training recipe	No fine-tuning	33.60	8.33	14.33	39.16
Instead of pre-train and fine-tune	No pre-training	56.60	13.53	17.91	57.98
Fine-tuning dataset	Action only	53.80	11.65	17.68	46.50
Instead of action and scenery	Scenery only	55.40	13.00	18.38	55.88
Frame count	Single frame	57.00	14.21	18.40	59.46
Instead of 5 frames	3 frames	59.80	14.61	18.44	62.61
	7 frames	60.60	14.46	**18.61**	61.82
Video fusion	Early-fusion	48.40	13.98	17.61	61.42
Instead of late-fusion	Mid-fusion	59.20	14.44	18.47	63.05
Language model	OPT-7B [60]	50.00	14.98	15.99	60.08
Instead of Vicuna-1.5-7B [13]	Llama-2-7B-Chat [48]	59.20	13.52	18.43	59.87
	Mistral-7B-Instruct [28]	58.00	13.80	18.33	64.21

Table (4) **Empirical Evaluation on LingoQA.** Ablation study highlighting the impact of various modifications in training recipes, dataset composition, frame count, video processing techniques and language model.

tably, OPT-7B [60] demonstrates significantly lower performance, despite having a similar number of parameters. This discrepancy underscores the crucial role of the pretraining phase in the base language model's effectiveness.

6. Discussion

Limitations of Lingo-Judge. This work still has a few limitations that we discuss and provide guidelines as to how their effects may be mitigated. The primary limitation of our proposed Lingo-Judge is that we have not studied its ability to *generalise* to other datasets and to other domains. The classifier has been trained for autonomous driving evaluation and has been shown effective for this purpose. Hence, the classifier can be used for evaluation when paired with the benchmark dataset that we release with the paper but would need to be re-trained if employed on a new dataset. Second, we optimized the classifier to evaluate responses in the style provided by human annotators in the evaluation dataset. The same response style is adopted in the LingoQA training sets and the models. We leave generalization to different answering styles to future work. Notably, Lingo-Judge does rate human labels as highly truthful, providing confidence in its ability to rate typical human answers. Third, as the classifier is only trained to predict factual correctness, it cannot discern which answer of two equally correct answers humans prefer. We leave it to future work to extend the classifier to predict human preference scores directly instead of factual correctness.

Strengths of Lingo-Judge. The strength of our contribution comprises proposing a classifier that is *highly correlated* with human inputs and *efficient* to run. In conjunction with the evaluation dataset that we propose, it becomes a useful tool for benchmarking vision-language models for autonomous driving on the video question answering task, which has been historically challenging to evaluate in a consistent fashion. With this contribution, autonomous driving research can be accelerated by providing a reliable, efficient, and easy-to-interpret benchmark.

Dataset and model limitations. One of the primary constraints is that our model operates on relatively short video segments and few frames, limiting the contextual understanding of scenarios. We also do not test for driving decisions and attention mechanisms, focusing on question-answering abilities only. We did not test the scaling in our models and focused on the most practical 7B parameter LLMs only. Our dataset and baseline are limited to information from a single front-facing car camera, excluding additional sensory inputs like LiDAR that could enrich the model's understanding of the driving environment. Expanding the model to address the short video context, as well as adding action prediction and evaluation to the dataset and the benchmark would result in a more robust and versatile system for autonomous driving.

7. Conclusion

In this paper, we introduced a novel benchmark for Video Question Answering for autonomous driving. The benchmark consists of a evaluation dataset, learned classifier-based metric Lingo-Judge that is highly correlated with human evaluation, a comprehensive high-quality training dataset for autonomous driving. The fast feedback from employing Lingo-Judge facilitates effective exploration in the video QA field. Additionally, the comprehensive experiments on different model combinations presented in this paper can become a foundation for further improvement of explainability in end-to-end autonomous driving systems. The LingoQA benchmark is openly released to spur further community research, providing a reliable and highly correlated evaluation method to human ratings.

Acknowledgements

This work was possible through the help of many colleagues across Wayve. We would like to thank Elahe Arani for the valuable discussions and feedback on the report. In particular we would like to acknowledge the support from: Anthony Hu, Miguel Sanchez Lohff, Lorenzo Bertoni, Charlie Lyons-Rothbart, Emma Wang, Harriett-Rose Follas, Kyle Esecson, Ben Foxall, Naama Zahavi, Ruben Diaz, Rudi Rankin, Tilly Pielichaty, Sarah Belghiti, Giulio D'Ippolito, Dan Reisman, Alex Persin, Fergal Cotter, Przemyslaw Mazur, Thomas Sajot, Giacomo Gallino, Alex Garcia Mayans, Tim Geypens, Robin Tweedie, Rebecca Hills.

References

[1] Partners for automated vehicle education. pave poll 2020. https://pavecampaign.org/pave-poll-americans-wary-of-avs-but-say-education-and-experience-with-technology-can-build-trust/. Accessed: 2023-10-12. 1, 2

[2] What's going on with the open llm leaderboard? https://huggingface.co/blog/evaluating-mmlu-leaderboard. Accessed: 2023-10-22. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. 2

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 2, 3, 4

[6] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018. 2

[7] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. 2

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 2

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023. 2

[10] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving, 2023. 2

[11] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers, 2023. 1

[12] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, and Weicheng Kuo. Pali: A jointly-scaled multilingual language-image model. *International Conference on Learninig Representation*, 2023. 2

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2, 6, 7, 8

[14] Pranav Singh Chib and Pravendra Singh. Recent advancements in end-to-end autonomous driving using deep learning: A survey, 2023. 2

[15] DriveLM Contributors. Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM, 2023. 5

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[17] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression, 2023. 3

[18] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2088–2098, 2019. 3

[19] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 2, 6

[20] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020. 2

[21] "Jeffrey Hawke, Haibo E, Vijay Badrinarayanan, and Alex Kendall ". "reimagining an autonomous vehicle", 2021. 2

[22] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. 4

[23] Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20703–20716. Curran Associates, Inc., 2022. 2

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 4

[25] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 7

[26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 7

[27] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 2

[28] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 7, 8

[29] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer, 2023. 2, 3, 4

[30] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention, 2017. 3

[31] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles, 2018. 1, 2, 3, 5

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2022. 2, 3, 6

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 7

[34] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 3

[35] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. 2, 4

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 6

[37] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. 3

[38] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving, 2022. 3

[39] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. 2023. 2

[40] The Language Archive Nijmegen: Max Planck Institute for Psycholinguistics. Elan, 2023. 5

[41] OpenAI. Gpt-4 technical report, 2023. 2, 4

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine

translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 2, 3, 4

[43] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023. 3, 5

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[45] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Behzad Dariush, Chiho Choi, and Mykel Kochenderfer. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning, 2023. 3, 5

[46] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving, 2023. 2

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2, 6

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 7, 8

[49] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 2, 3, 4

[50] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[51] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, and Yuan Cao Yulia Tsvetkov. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learining Representation*, 2022. 2

[52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1, 12

[53] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, and Botian Shi. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving, 2023. 2

[54] Wei Xu. From automation to autonomy and autonomous vehicles: Challenges and opportunities for human-computer interaction. *Interactions*, 28(1):48–53, dec 2020. 2

[55] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles, 2020. 3, 5

[56] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2023. 2, 3, 4

[57] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022. 2

[58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2

[59] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 2022. 2

[60] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. 2, 8

[61] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning, 2023. 7

A. LingoQA Dataset Examples

Further examples on the capabilties existent in the training and the evaluation datasets are shown in Figure 4. The *scenery* dataset contains highly descriptive elements, such as object colours, junction type, construction zones, traffic lights, and the road layout. The *action* dataset is complementary and focused on driving competencies, such as the impact of traffic lights on driving and interactions with other road agents. The *evaluation* dataset contains a broad range of questions aimed to test competencies relevant for autonomous driving.

B. Lingo-Judge Examples

We present additional qualitative examples from our evaluation dataset in Table 5, alongside predictions from our base model and corresponding metrics for each individual sample. Metrics based on *n-gram* matching such as CIDEr tend to be error-prone. For example, expressions that have the same meaning, but entirely different words, are marked as not similar at all, such as *"None"* and *"There are no cars."*. Sentences with minor but significant differences are graded as highly similar, despite having opposite meanings, such as *"The traffic lights are showing green"* and *"The traffic lights are showing red"*. Lingo-Judge demonstrates robustness against these varied expressions and subtle changes. Lingo-Judge also has limitations, primarily seen when establishing the correctness of the answer would require extra context from the videos. These examples can be seen in Table 6.

We qualitatively compare our classifier to GPT-4 ratings. These examples are shown in Figure 5. In this situation, GPT-4 is misled by the fact that the model answer contains partially correct information. The GPT-4 assessment states that *"The student correctly identified the presence of a traffic light"* and, despite the colours not being correct, further explains that *"and accurately stated its colour"*. This highlights some challenges faced by GPT-4 when trying to rate the truthfulness of an answer. Lingo-Judge correctly identifies that the statements described by the model are false.

C. GPT-4 Grading

In this section we provide an overview of the implementation details for the evaluation method using GPT-4 with and without Chain-of-Thought (CoT) [52] prompting.

GPT-4 with CoT. In order to evaluate a model's answer with GPT-4 and CoT prompting, we first provide GPT-4 with the question and one or more valid answers for the questions, and ask it to come up with a strategy to evaluate new answers to this question. We then provide GPT-4 with the model's answer and ask it to evaluate the answer using the strategy it proposed in the previous step. Finally, we ask GPT-4 to give the model a grade between 0 and 5, where 5 means the answer is perfect. The prompt used is shown in Figure 6.

GPT-4 without CoT. When evaluating model outputs without CoT prompting, we provide GPT-4 with the question, one or more valid answers for the questions, and the model predictions and we directly ask GPT-4 to give the model a grade between 0 and 5, without the intermeidate reasoning steps. The prompt used is shown in Figure 7.

We emit concurrent requests to our Azure's GPT-4 deployment in order to max-out the limit of 40k tokens per minute. GPT-4 without CoT prompting required more than 13 minutes to perform the evaluation, and GPT-4 with CoT prompting requires more than 50 minutes.

D. Lingo-Judge Correlation Study

We show that Lingo-Judge exhibits a higher correlation to human judgment than commonly-used language-based metrics, and than GPT-4. To do so, we computed the scores of 15 different models and 2 groups of human labellers on the questions in our evaluation dataset using Lingo-Judge, GPT-4, BLEU4, METEOR and CIDEr. These scores are reported in Table 7.

We then computed the Pearson and Spearman correlation coefficients between these metrics and the human evaluation. The **Pearson correlation coefficient** measures the strength of the linear correlation between the human evaluation and a metric score, while the **Spearman rank correlation coefficient** measures the monotonic relationship between the human evaluation and the metric. The higher the Spearman coefficient, the better a metric is at ranking answers in the same order as our human evaluators. To compute the confidence intervals, we use the Fisher transformation with a 95% confidence level.

In Figure 8, the metric scores are plotted against the human evaluators' grades (from 0 to 1). In red is the least-squares regression of the linear relationship between the metric and the human-assigned grades. Figure 9 shows the value of both correlation coefficients for each of the 5 metrics, as well as their confidence interval bounds. We note than not only does Lingo-Judge provided higher correlation, it also provides tighter confidence intervals than the other metrics.

E. Training Parameters

In this sections we present further details on the training parameters used for the LingoQA Baseline. The training process consists of a pre-training stage, and a fine-tuning stage. Table 8 shows the parameters for pre-training and fine-tuning respectively. The datasets are sampled with equal weight for both pre-training and fine-tuning. The overall training time was 20h for pre-training and 5h for fine-tuning on an NVIDIA A100 8GPU 80GB machine.

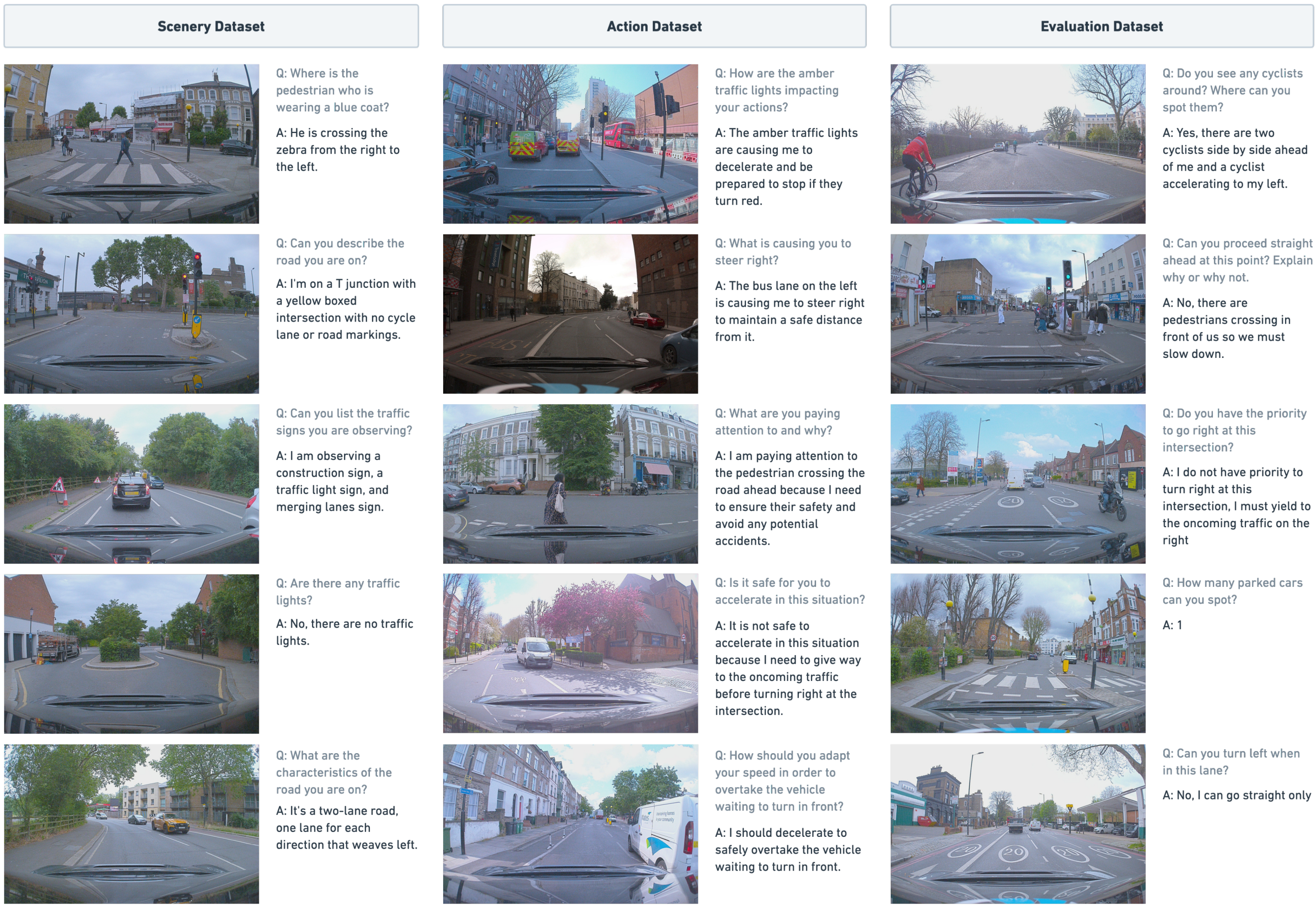

Figure (4) **LingoQA dataset examples.** From left to right: *scenery* dataset, *action* dataset, and *evaluation* dataset.

F. LingoQA Baseline Examples

We qualitatively showcase the impact of our proposed LingoQA dataset. Figure 10 compares three models: a model that is not fine-tuned on any LingoQA datasets, one that is fine-tuned on the *action* dataset only, one on the *scenery* dataset only, and the baseline that is trained with both. Two questions are asked, one focused on perception only, and one focused on action justification. The action only model performs well at answering action-related questions, but not perception. The scenery only model performs well at perception tasks, but not action justification. The baseline exhibits good performance on both.

Question	Label	Prediction	CIDEr	GPT4	GPT4-CoT	L-J Prob.	L-J Class.
How many pedestrians are crossing the road?	Zero pedestrians	There are no pedestrians crossing the road.	23.18	5	5	0.96	True
What is the road speed limit?	20 mph - it is written on the road	The road speed limit is 20 mph.	97.91	4	3	0.95	True
How many cars are driving in your direction?	None	There are no cars driving in my direction.	0.11	4	3	0.96	True
Which vehicle should you follow if any?	The motorcyclist.	If any, I should follow the motorcycle ahead.	0.42	4	5	0.95	True
What is the current action and its justification? Answer in the form "action, justification"	Slow down, there is a stationary van infront of us	I am decelerating because of the stationary truck ahead.	34.79	5	5	0.96	True
What is the current action and its justification? Answer in the form "action, justification"	Stop, Red light	I am stopping because the traffic lights to go straight are red.	18.62	5	5	0.95	True
How many cyclists can you see?	I can see 3 cyclists	I can see two cyclists.	150.29	1	2	0.05	False
What color are the traffic lights showing?	The traffic lights are showing green	The traffic lights are showing red.	329.36	0	1	0.05	False
What action are you taking with respect to the cyclist?	Overtaking them on the right and keeping the speed	I am overtaking the cyclist on the left.	349.52	2	2	0.10	False
In which direction is the bus driving?	The bus is driving in the opposite direction	The bus is driving in the oncoming direction.	404.65	4	5	0.31	False
Are there any parked car on the side of the road?	Yes, there are two cars parked on the right of the road	No, there are no parked cars on either side of the road.	142.40	0	0	0.05	False
Is acceleration necessary in this situation? If so, provide the reason.	No. We should decelerate in this situation because there is a vehicle stopping ahead of us.	No, acceleration is not necessary in this situation as I am already driving at the speed limit.	177.96	3	3	0.31	False

Table (5) **Qualitative comparison of metrics.** Questions and labels from our evaluation dataset along with exemplary predictions and their corresponding sample-level metrics. Notable is that the classifier (here "L-J") can pick up slight but crucial differences between ground truth and prediction to correctly identify a wrong prediction, such as "The traffic lights are green" and "The traffic lights are red", in which case CIDEr still shows high agreement. Green color indicates agreement with human judgement, while orange and red show disagreement.

Question	Label	Prediction	CIDEr	GPT4	GPT4-CoT	L-J Prob.	L-J Class.
Is any vehicle stopped at the zebra crossing?	There is no zebra crossing.	No, there are no vehicles stopped at the zebra crossing.	106.19	4	2	0.93	True
What elements are capturing your attention and what is the reason behind it?	I am making sure there is no vehicle on the left lane I am making my lane change to. I am also looking at the traffic light to make sure it stays green while I am crossing the intersection.	I am paying attention to the bus lane on the left because I need to maintain a safe distance from it.	0.91	3	1	0.57	True
What is the current action and its justification? Answer in the form ”action, justification”	We are taking a left turn, because we are at a t-junction. Then we keep a safe distance from the cyclists in front.	I am decelerating to keep a safe distance from the cyclist ahead of me.	46.45	2	2	0.32	False

Table (6) **Failure Cases of Lingo-Judge.** Examples where Lingo-Judge makes a wrong judgement about the correctness of the model prediction. Green color indicates agreement with human judgement, while orange and red show disagreement.

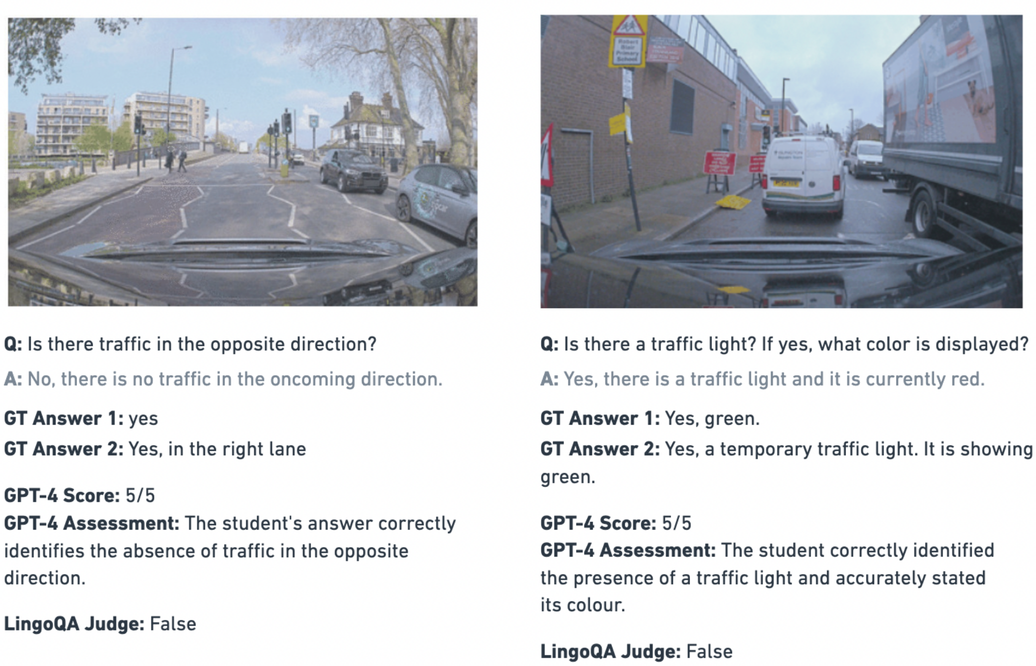

Figure (5) **Classifier examples.** Examples of Lingo-Judge outputs compared to GPT-4.

System Prompt:

You are compassionate but fair AI which task is to grade students.

First Prompt:

The following question is asked to students during a test: **"{question}"**

A quorum of teachers decided that the following answers are the acceptable answers to this question:
- **"{ground_truth_1}"**
- **"{ground_trurh_2}"**

Both of the teachers' answers are acceptable but it would be unfair to expect students' answers to contain the information in both answers.As long as their answer matches one of the acceptable answers, they should be awarded full points.

In order to check whether a student's answer correctly answers the questions and matches the information present in at least of the acceptable answers, how should you test an answer?

Second Prompt:

A student answered the following: **"{model_answer}"**
What can you infer from the above regarding the level of accuracy of the student's answer?

Third Prompt:

You must now grade the student between 0 and 5, where 0 means the student's answer is completely false and 5 means they have accurately answered the question.
You should show empathy for the students and award them points if their answer is partially correct.
It's ok to award them any grade between 0 and 5 to reward them for a partial answer.
You must answer in the following format: "<one-line explanation for the score>. Score: <score from 0 to 5>."

Figure (6) **GPT-4 with Chain of Thought (CoT) prompting.** First, GPT-4 is provided with the question and ground truth answers, and asked to come up with a strategy for testing the answer. Second, GPT-4 is provided with the model answer and is prompted to evaluate the accuracy of the response based on the previously defined strategy. Finally, GPT-4 is asked to provide a grade for the student.

System Prompt:

You are compassionate but fair AI which task is to grade students.

First Prompt:

The following question is asked to students during a test: **"{question}"**

A quorum of teachers decided that the following answers are the acceptable answers to this question:
- **"{ground_truth_1}"**
- **"{ground_truth_2}"**

Both of the teachers' answers are acceptable but it would be unfair to expect students' answers to contain the information in both answers. As long as their answer matches one of the acceptable answers, they should be awarded full points.

A student answered the following: **"{model_answer}"**

You must grade the student between 0 and 5, where 0 means the student's answer is completely false and 5 means they have accurately answered the question.
You should ignore the style of the answer and focus only on the accuracy of the information it contains.
You should show empathy for the students and award them points if their answer is partially correct.
It's ok to award them any grade between 0 and 5 to reward them for a partial answer.

Figure (7) **GPT-4 without Chain of Thought (CoT) prompting.** GPT-4 is provided with a prompt that contains the question, the ground truth answers, and the model response, and is requested to directly provide a grade for the student.

		Lingo-Judge [%] ↑	BLEU ↑	METEOR ↑	CIDEr ↑	GPT-4 ↑	Human ↑
Models	Model A	59.6	15.45	18.36	66.32	3.23	0.571
	Model B	59.6	15.16	18.84	65.11	3.16	0.564
	Model C	57.4	14.87	18.52	65.49	3.08	0.563
	Model D	58.2	14.51	18.59	66.02	3.15	0.559
	Model E	59.0	14.42	18.58	66.95	3.14	0.553
	Model F	58.0	14.82	18.89	65.43	3.11	0.552
	Model G	54.8	14.41	17.86	64.67	2.98	0.529
	Model H	50.0	13.29	17.44	59.87	2.88	0.520
	Model I	53.0	14.63	17.98	64.45	2.96	0.510
	Model J	52.6	12.17	17.59	50.45	3.00	0.509
	Model K	53.0	13.20	18.03	54.90	3.04	0.500
	Model L	51.2	14.69	17.83	64.51	2.91	0.485
	Model M	43.2	13.76	17.37	60.36	2.67	0.371
	Model N	35.8	13.18	15.67	56.07	2.41	0.361
	Model O	33.6	8.33	14.33	39.16	2.07	0.279
Humans	Human labellers group A	96.6	81.04	52.92	361.77	4.68	0.934
	Human labellers group B	91.2	61.72	42.57	267.87	4.3	0.894

Table (7) **Correlation study metrics.** Metrics from different models on our evaluation dataset used in the correlation study in Table 1. For reference, we also present metrics for answers provided by human labellers. "Human" is the average of inference output scores in range $[0, 1]$ where 0 is worst and 1 is best, as described in section 3.2.

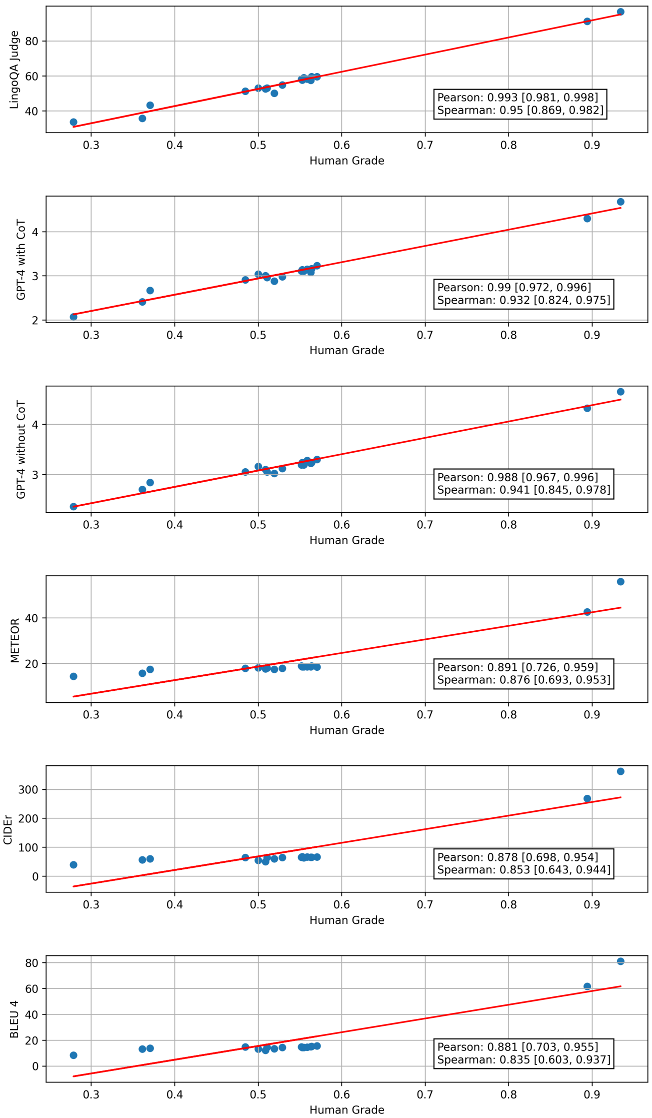

Figure (8) **Correlation trends.** Correlation trends of the average grade of models compared to the average human-grades, for different metrics.

Parameter	Pre-training	Fine-tuning
Precision	bf16	bf16
Warm-up steps	1000	1000
Maximum steps	100000	10000
Batch size	6	8
Gradient acc. steps	1	1
Learning rate	$5 * 10^{-5}$	$5 * 10^{-5}$
Learning rate scheduler	cosine	cosine
Weight decay	0.1	0.1

Table (8) **Training parameters.** This table shows the training parameters utilised for the pre-training and for the fine-tuning stages respectively.

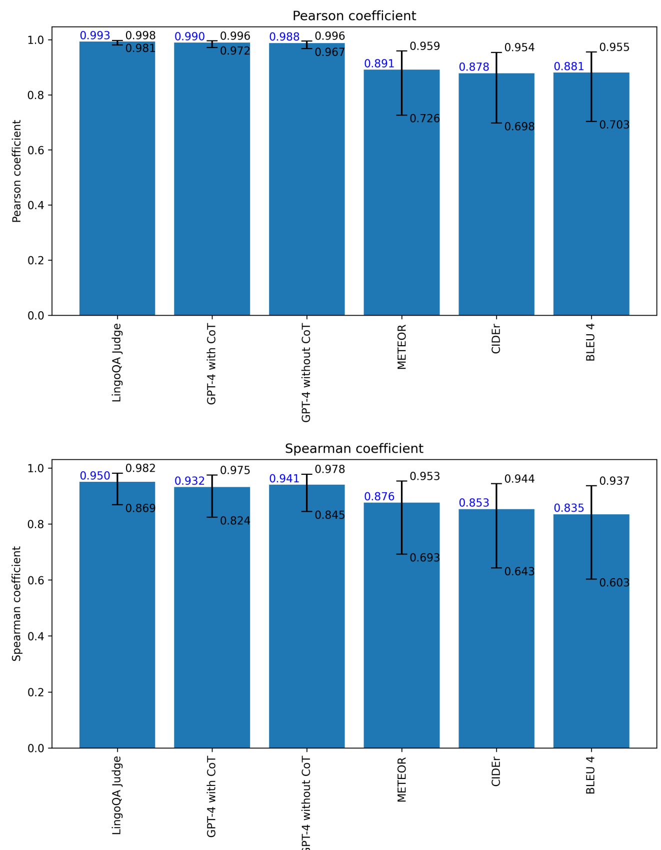

Figure (9) **Correlation coefficients**. Correlation coefficients of the average grade of different models vs. the average human-grades, for different metrics.

Question	Answer	Model	Correct
Is there a traffic light? If yes, what color is displayed?	Yes, there is a traffic light and it is displaying a red light.	*No fine-tuning*	✗
	Yes, there is a traffic light and it is red.	*Action only*	✗
	Yes, there is a traffic light and it is currently green.	*Scenery only*	✓
	Yes, there is a traffic light and it is currently green.	*Baseline*	✓

Question	Answer	Model	Correct
What is the current action and its justification? Answer in the form "action, justification"	A man is crossing the street, justified by the presence of a crosswalk.	*No fine-tuning*	✗
	I am stopping to let the pedestrian safely cross on the zebra or pedestrian crossing ahead because pedestrians have priority on the crossing.	*Action only*	✓
	I am accelerating because the traffic light has turned green and there are no obstacles in my way.	*Scenery only*	✗
	I am stopping because there is a pedestrian crossing on a zebra crossing ahead.	*Baseline*	✓

Figure (10) **Examples of model outputs on the LingoQA benchmark.** We compare the baseline with a model that has not been fine-tuned on the LingoQA dataset, a model fine-tuned on the action dataset only, and a model fine-tuned on the scenery dataset only. This shows qualitatively how the baseline can handle both action justification as well as descriptive tasks by combining the strengths of both datasets.