TransportationGames: Benchmarking Transportation Knowledge of (Multimodal) Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) and multimodal large language models (MLLMs) have shown excellent general capabilities, even exhibiting adaptability in many professional domains such as law, economics, transportation, and medicine. Currently, many domain-specific benchmarks have been proposed to verify the performance of (M)LLMs in specific fields. Among various domains, transportation plays a crucial role in modern society as it impacts the economy, the environment, and the quality of life for billions of people. However, it is unclear how much traffic knowledge (M)LLMs possess and whether they can reliably perform transportation-related tasks. To address this gap, we propose TransportationGames, a carefully designed and thorough evaluation benchmark for assessing (M)LLMs in the transportation domain. By comprehensively considering the applications in real-world scenarios and referring to the first three levels in Bloom's Taxonomy, we test the performance of various (M)LLMs in memorizing, understanding, and applying transportation knowledge by the selected tasks. The experimental results show that although some models perform well in some tasks, there is still much room for improvement overall. We hope the release of TransportationGames[1] can serve as a foundation for future research, thereby accelerating the implementation and application of (M)LLMs in the transportation domain.

1 Introduction

Large language models (LLMs) are revolutionizing the way humans work by augmenting them in various tasks. As these LLMs, for example GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023), become more sophisticated, they will be able to handle more complex tasks, enabling them to assist and collaborate with humans in a multitude of professional domains (Sanh et al., 2021; Ouyang et al., 2022; Zhang et al., 2022; Shao et al., 2023). Additionally, beyond single-modal LLM, the Multimodal Large Language Model (MLLM) has recently emerged as a popular area of research (Bai et al., 2023b; Ye et al., 2023; Liu et al., 2023a; Zhang et al., 2023a). The MLLM utilizes powerful LLMs to effectively handle multimodal tasks, resulting in versatile problem solvers. To comprehensively and accurately assess the capabilities of (M)LLMs, evaluation benchmarks play a crucial and indispensable role in their development (Hendrycks et al., 2020). By evaluating (M)LLMs using these benchmarks, researchers and developers can gain valuable insights into the strengths and weaknesses of different models, enabling them to identify areas for improvement and innovation.

Currently, many benchmarks have been proposed to assess (M)LLMs on various aspects of universal capabilities, *e.g.*, MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023b), CMMLU (Li et al., 2023), BIG-bench (Srivastava et al., 2023), MMBench (Liu et al., 2023b) and MME (Fu et al., 2023). Moreover, when evaluating (M)LLMs, it is important to not only focus on their general capabilities but also to incorporate domain-specific benchmarks for assessing models specialized in specific fields (Zhao et al., 2023d), because domain-specific benchmarks push (M)LLMs towards tackling the specific challenges and complexities of their target fields, ultimately driving practical progress and responsible implementation. Existing domain-specific benchmarks include LawBench (Fei et al., 2023), LegalBench (Guha et al., 2023), and LAiW (Dai et al., 2023) for the legal domain, MIR-based benchmark (Goenaga et al., 2023) for the medicine domain, ChemLLMBench (Guo et al., 2023) for the chemistry domain, etc. Among various domains, transportation plays a crucial role in modern society as it impacts the

[1]The evaluation method has been released in https://transportation.games.

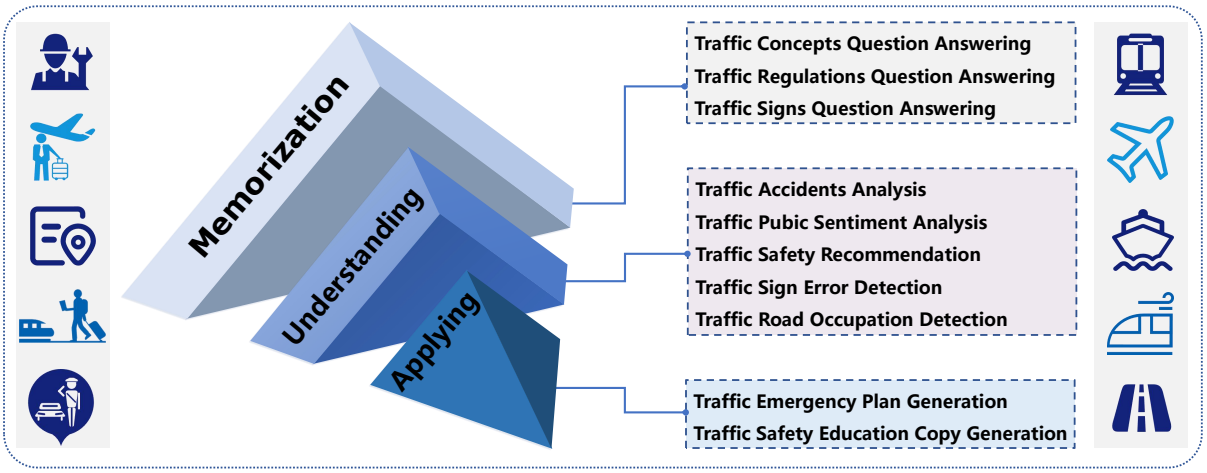


Figure 1: The organization of our TransportationGames. Considering the specific scenarios in the transportation domain, our TransportationGames employs the first three levels in Bloom's Taxonomy, which are Memorization, Understanding, and Applying, to evaluate the (M)LLMs. We select 10 tasks based on diverse sub-domains in the transportation domain such as urban transportation, rail transit, aviation, and maritime transport.

economy, the environment, and the quality of life for billions of people (Taylor, 2015; Koopmans, 1949). However, it is unclear how much traffic knowledge[2] (M)LLMs possess and whether they can reliably perform transportation-related tasks.

To address this gap, we introduce TransportationGames (refer to Figure 1): a thoughtfully designed, all-encompassing evaluation benchmark to accurately evaluate the capabilities of (M)LLMs in executing transportation-related tasks. By comprehensively considering the applications in real-world scenarios, we select 10 varied tasks across 3 types: multiple-choice, "True/False" judge, and text generation, including text and image modality. We categorize these tasks into three skill levels based on widely recognized Bloom's cognitive models (Krathwohl, 2002): (1) Transportation knowledge memorization: whether (M)LLMs can memorize transportation-relevant concepts, facts, regulations, and traffic law articles; (2) Transportation knowledge understanding: whether (M)LLMs can understand, analyze and reasoning based on transportation-domain knowledge; (3) Transportation knowledge applying: whether (M)LLMs can effectively make the necessary logical deductions to solve practical transportation tasks both for public and professionals. Overall, our TransportationGames offers a systematic outline of the skillset necessary for tasks related to transportation.

Our main contributions are three-fold:

- **Systematically-constructed benchmark.** We introduce TransportationGames, a carefully designed and thorough evaluation benchmark for assessing (M)LLMs in transportation-related tasks. It is the first benchmark specifically designed for the transportation domain.

- **Experiments.** We design appropriate rules to accurately extract answers from the model-generated predictions, and employ proper metrics for each task. We conduct extensive testing on 16 widely used (M)LLMs and the evaluation results are presented in Table 3 and Table 4.

- **Analysis.** We observe that although some LLMs perform well in some tasks on text-only knowledge, there is still room for improvement. As for multimodal knowledge, most MLLMs exhibit poor capability. Additionally, we analyze the key factors affecting model performance.

2 Related Work

2.1 Large Language Models

Large language models (LLMs) typically refer to Transforme-based language models encompassing several billion (or more) parameters (Zhao et al., 2023b), such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), Baichuan (Yang et al., 2023), and so on. With the implementation of many training strategies, *e.g.*, model pre-training, instruction tuning, reward model training,

[2]We only focus on Chinese.

and reinforcement learning with human feedback (RLHF) (Zhao et al., 2023c), LLMs can achieve commendable performance on tasks within general domains. To improve the performance of LLMs on more specific domains, more research endeavors increasingly aspire to deploy LLMs across diverse domains, including but not limited to law (Nguyen, 2023; Huang et al., 2023a), medicine (Zhang et al., 2023c,b; Jiang et al., 2023), transportation (Da et al., 2023; Lai et al., 2023; Mo et al., 2023), chemistry (Guo et al., 2023; Ouyang et al., 2023; Wellawatte and Schwaller, 2023), and psychology (Ke et al., 2024; Cho et al., 2023), to proficiently accomplish domain-specific tasks. Moreover, multimodal large language models (MLLMs) have emerged as a recent focal point in the community (Zhao et al., 2023a; Bai et al., 2023b; Ye et al., 2023; Liu et al., 2023a), capitalizing on the prowess of potent large language models to serve as cognitive entities for executing multimodal tasks, thereby exhibiting remarkable emergent capabilities.

In this paper, we focus on the development of (M)LLMs in the transportation domain. There are many (M)LLMs tailored for the traffic domain including TransGPT (Peng, 2023), TrafficGPT[3], MT-GPT[4], and TransCore-M[5]. Among them, TransGPT and TransCore-M have undergone instruction tuning based on traffic domain data.

2.2 Existing Benchmarks

The comprehensive and precise evaluation of the functionalities inherent in (M)LLMs is pivotal and irreplaceable in their development. Evaluation benchmarks assume a critical role, furnishing a standardized framework that facilitates the meticulous measurement and analysis of (M)LLM performance across diverse tasks and domains.

Recently, more and more benchmarks have been developed to evaluate the various capabilities of (M)LLMs. To assess the comprehensive capabilities of LLMs, many benchmarks have been constructed based on knowledge across various disciplines and languages, including MMLU (Hendrycks et al., 2020) and ARC (Clark et al., 2018), which are grounded in English, as well as C-Eval (Huang et al., 2023b) and CMMLU (Li et al., 2023), which are rooted in Chinese. As for MLLMs, there are also many benchmarks with the comprehensive evaluation pipeline, such as MME (Fu et al., 2023) and MMBench (Liu et al., 2023b). In addition, some benchmarks are designed to evaluate the performance of (M)LLMs on some specific domains, *e.g.*, LawBench (Fei et al., 2023), LegalBench (Guha et al., 2023), and LAiW (Dai et al., 2023) for the legal domain, MIR-based benchmark (Goenaga et al., 2023) for the medicine domain, ChemLLMBench (Guo et al., 2023) for the chemistry domain, and so on. However, to the best of our knowledge, there is no systematic evaluation benchmark for the transportation domain, so we propose the TransportationGames for assessing (M)LLMs in transportation-related tasks.

3 Benchmark Construction

In this section, we provide a detailed introduction to the construction of our TransportationGames. Firstly, we elucidate the classification criteria (§3.1) employed in the design of the benchmark, along with the corresponding selection of evaluation tasks (§3.2). Subsequently, we introduce the data collection procedures (§3.3) and the adoption of evaluation metrics (§3.4).

3.1 The Taxonomy of TransportationGames

In the construction of benchmarks, an effective process involves not only evaluating models on multiple sub-tasks but also organizing benchmarks systematically. These benchmarks can be organized based on task difficulty or task categories, which to some extent reflect the models' aptitude. However, such a simplistic classification criterion may not adequately capture the full range of model capabilities.

Inspired by Fei et al. (2023), we adopt Bloom's cognitive model for task classification, aiming to capture the models' capabilities at a higher level. Bloom's Taxonomy system (Anderson et al., 2000), initially introduced by the educational psychologist Benjamin Bloom and his collaborators in 1956, has obtained widespread application and continuous development in subsequent decades. It has proven instrumental in assisting educators in both curriculum design and the evaluation of student learning outcomes. The taxonomy categorizes learning objectives within the cognitive domain into six progressively ascending levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. These

[3] https://github.com/lijlansg/TrafficGPT

[4] https://www.7its.com/?m=home&c=View&a=index&aid=19245

[5] https://github.com/PCIResearch/TransCore-M

Capability Levels	ID	Task	Modality	Type	Metric
Transportation Knowledge Memorization	**T1**	Traffic Concepts Question Answering	Text	TF/MLC	Accuracy
	T2	Traffic Regulations Question Answering	Text	TF/MLC	Accuracy
	T3	Traffic Signs Question Answering	Multimodal	TF/MLC	Accuracy
Transportation Knowledge Understanding	**T4**	Traffic Accidents Analysis	Text/Multimodal	Generation	ROUGE/GPT-4
	T5	Traffic Pubic Sentiment Analysis	Text	Generation	ROUGE/GPT-4
	T6	Traffic Safety Recommendation	Text/Multimodal	Generation	ROUGE/GPT-4
	T7	Traffic Sign Error Detection	Multimodal	Generation	ROUGE/GPT-4
	T8	Traffic Road Occupation Detection	Multimodal	Generation	ROUGE/GPT-4
Transportation Knowledge Applying	**T9**	Traffic Emergency Plan Generation	Text/Multimodal	Generation	ROUGE/GPT-4
	T10	Traffic Safety Education Copy Generation	Text	Generation	ROUGE/GPT-4

Table 1: Task list of TransportationGames. There are 10 tasks corresponding to 3 capability levels: Transportation Knowledge Memorization, Understanding, and Applying, and 2 modalities: Text and Multimodal (text + image), and 3 task types: multiple-choice (MLC), "True/False" judge (TF), and text generation. Additionally, the metrics used for each task are also listed and described in detail in §3.4.

hierarchical levels delineate the depth and intricacy of cognitive learning, providing educators with a structured framework for instructional design and assessment.

Considering the specific scenarios in the transportation domain, we employ the first three levels in Bloom's Taxonomy to assess the (M)LLMs as shown in Figure 1. The detailed introduction is as follows:

Transportation Knowledge Memorization. It tests whether (M)LLMs can memorize and answer basic transportation-related knowledge, such as concepts, facts, regulations, or traffic law articles.

Transportation Knowledge Understanding. The excellent understanding capability generally requires the model to engage in activities such as interpretation, illustration, categorization, summarization, and inference based on transportation-domain knowledge. For example, the models can interpret traffic regulations and rules, compare the applicable conditions of different rules, classify traffic rules based on some features, etc.

Transportation Knowledge Applying. The applying capability is to assess whether the model can flexibly apply acquired knowledge and effectively make the necessary logical deductions to solve practical transportation tasks both for the public and professionals.

3.2 Tasks

The core knowledge areas of the transportation profession generally include transportation infrastructure construction, carrier theory and technical equipment, transportation system planning, port and station hub planning and design, passenger operation organization, cargo operation organization, operation dispatching command, as well as transportation policies and regulations, transportation commerce, transportation economy, transportation safety, modern logistics, and comprehensive transportation. And it mainly involves four sub-domains: road transportation, railway transportation, waterway transportation, and aviation transportation.

During selecting tasks, we take into account diverse sub-domains in transportation and the varying needs of different people, including the general public and industry practitioners, in their day-to-day lives or professional undertakings. Furthermore, we conduct detailed consultations with domain experts to choose the specific tasks. Finally, we select 10 tasks under the aforementioned capability levels and the task list is presented in Table 1. Note that due to the different application scenarios of different tasks, it involves multiple modalities of knowledge, such as text and image modality. The concrete introduction is as follows.

Transportation Knowledge Memorization Tasks

- **Traffic Concepts Question Answering (T1)**: Inquire about common concepts in the field of transportation, formulating queries in either multiple-choice (MLC) or "True/False" judge (TF) formats. In the case of multiple-choice questions, the model is expected to select the correct answer from a set of four options, whereas true/false questions necessitate the model to determine the correctness of a given statement.

- **Traffic Regulations Question Answering (T2)**: Question the model regarding nuanced

components of traffic regulations, including numerical parameters, years, or analogous elements. The question formats are MLC or TF.

- **Traffic Signs Question Answering (T3)**: Given a traffic sign image and a query, test whether the model can memorize the meaning of different traffic signs. The query formats are MLC or TF.

Transportation Knowledge Understanding Tasks

- **Traffic Accidents Analysis (T4)**: Given a photo of a traffic accident scene or a traffic accident process, the model is required to extract and summarize information including the origins, progression, or consequences of the incident.
- **Traffic Public Sentiment Analysis (T5)**: Given the feedback from the public regarding the proposed traffic proposal, the model should analyze, summarize, and discern the authentic demands of the public. This task facilitates a more comprehensive understanding for professionals of public sentiment, enabling targeted actions to be taken to fulfill the public's needs.
- **Traffic Safety Recommendation (T6)**: Given travel plans, such as weather conditions and road conditions, the model needs to provide reasonable safety traffic advice. Additionally, given an image, the model can point out the hidden security risks.
- **Traffic Sign Error Detection (T7)**: Given images containing traffic signs or lines on the road, the model needs to analyze whether the traffic signs are obstructed or defaced, whether traffic lines are designed reasonably, or if these lines need to be redrawn due to damage.
- **Traffic Road Occupation Detection (T8)**: Given images of roads, the model needs to analyze whether there is any illegal occupation of the road.

Transportation Knowledge Applying Tasks

- **Traffic Emergency Plan Generation (T9)**: Given an urgent description of a traffic accident or malfunction, the model should generate targeted emergency response plans.
- **Traffic Safety Education Copy Generation (T10)**: When provided with specific target audiences, the model should generate targeted educational materials.

3.3 Data Collection

In this section, a comprehensive exposition is presented regarding the process of data collection, including the data sources, data processing procedures, and ultimately, culminating in an overview of the acquired data.

Data Sources The aforementioned tasks primarily involve two modalities: text and images. For textual data, the primary source of our dataset is available on the internet. For instance, we have retrieved numerous examination papers related to the field of transportation for the source of question-and-answer data. The accident reports or public sentiment about specific regulations are predominantly sourced from news websites and municipal management platforms. Additionally, specialized articles, such as emergency response plans, are primarily obtained from relevant sections of various institutional websites. As for image data, we employ keyword-based queries to retrieve and select images from online repositories, ensuring conformity with predefined criteria. Simultaneously, the text from image titles or title links is preserved for further analysis.

Data Processing The formats of collected data are various, including Microsoft Word documents, PDFs, PNGs, JPGs, or Web pages. We employ relevant Python toolkits (*e.g.*, `pdfplumber`[6], `pypdf2`[7], `python-docx`[8]) to extract text and preprocess it into the appropriate format for the designated tasks. In cases where automated extraction is not feasible, we seek the relevant professionals to process it manually. Additionally, we take measures to eliminate sensitive information from the data, including but not limited to personal phone numbers, ID numbers, emails, and detailed home addresses, safeguarding privacy. Furthermore, we ensure that each piece of data has undergone meticulous manual verification to guarantee alignment with the specified task, accuracy of answers, and coherence of sentences.

[6] https://pypi.org/project/pdfplumber/
[7] https://pypi.org/project/PyPDF2/
[8] https://pypi.org/project/python-docx/

Model	Parameters	SFT	RLHF	Access	BaseModel
Large Language Models					
ChatGLM3-6B (Zeng et al.)	6B	✓	✗	Weights	ChatGLM
Qwen-7B-Chat (Bai et al.)	7B	✓	✗	Weights	Qwen-7B
Qwen-14B-Chat (Bai et al.)	14B	✓	✗	Weights	Qwen-14B
Baichuan2-13B-Chat (Baichuan)	13B	✓	✗	Weights	Baichuan2-13B-Base
InternLM-Chat-7B (Team)	7B	✓	✓	Weights	InternLM-7B
InternLM-Chat-20B (Team)	20B	✓	✓	Weights	InternLM-20B
Yi-6B-Chat	6B	✓	✗	Weights	Yi-6B
LLaMa2-Chinese-13B-Chat-ms	13B	✓	✗	Weights	LLaMa2-13B
GPT-4	/	✓	✓	API	/
Multimodal Large Language Models					
VisualGLM (Zeng et al.)	7.8B	✓	✗	Weights	ChatGLM-6B + BLIP2-Qformer
mPLUG-Owl2 (Ye et al.)	8.2B	✓	✗	Weights	LLaMa-7B + CLIP ViT-L/14
Qwen-VL-Chat (Bai et al.)	9.6B	✓	✗	Weights	Qwen-7B + ViT-G/16
Chinese-LLaVa-Cllama2	7.3B	✓	✗	Weights	LLaVa + Chinese-LLaMa2-7B
Chinese-LLaVa-Baichuan	7.3B	✓	✗	Weights	LLaVa + Baichuan-7B
InternLM-XComposer-7B (Zhang et al.)	8B	✓	✗	Weights	InternLM-Chat-7B + EVA-CLIP
LLaVa-v1.5-13B (Liu et al.)	13.4B	✓	✗	Weights	Vicuna-v1.5-13B + CLIP ViT-L/14
Transportation-domain Models					
TransGPT (Peng)	7B	✓	✗	Weights	ChatGLM2-6B
TransCore-M	13.4B	✓	✗	Weights	PCITransGPT-13B + CLIP ViT/L-14

Table 2: Models tested on TransportationGames. We classify these models by different modalities and we list the open-source models TransGPT and TransCore-M in the transportation domain separately.

Data Overview Following data processing and manual verification, we obtain the final dataset corresponding to each task. Due to variations in task difficulty, the amount of data instances is different across tasks. A detailed data distribution is shown in Appendix A. Additionally, according to the involved modalities of different tasks (refer to the fourth column in Table 1), the entire dataset can be divided into two parts, the text-only dataset and the multimodal dataset, which will be utilized to evaluate LLMs and MLLMs respectively. The input for the text-only dataset is a text question and the input for the multimodal dataset is an image with a question. We have listed some examples for each task in Appendix B.

3.4 Evaluation

For the evaluation of each task, we first extract the answer from the model-generated prediction and then compute the corresponding metric values according to the golden answer.

Answer Extraction For questions with the type of MLC and TF, some models generate answers that include content other than "A/B/C/D" or "True/False". It is imperative to extract the options from the generated answers in such cases before calculating metrics. Moreover, we do not conduct extraction for other question types.

Different Metrics

- **Accuracy:** For MLC and TF, there are the gold answers for each query (T1∼T3). Therefore, we calculate the accuracy of the extracted answer according to the gold answer. Additionally, we also calculate the format error rate of model-generated answers.
- **ROUGE:** For the questions of Generation type (T4∼T10), we calculate the ROUGE-Chinese-L[9] score between the predicted answer and the reference answer. ROUGE-L is a commonly used metric in generation tasks.
- **GPT-4-Eval:** Since the reference answers for some tasks (T4∼T10) are not unique, we also utilize GPT-4[10] to evaluate the model-generated answers for accuracy, redundancy, fluency, and completeness. The example instruction that we designed is presented in Appendix C.

4 Experiments

4.1 Selected Models

We evaluate a substantial number of models (listed in Table 2) on our TransportationGames. According to modalities and domains, they are primarily

[9] https://pypi.org/project/ROUGE-chinese/
[10] The 0613 version.

Models	T1	T2	T4	T5	T6	T9	T10	SUM
GPT-4	**81.33**$_{(\mathbf{0.00})}$	80.89$_{(\mathbf{0.00})}$	21.2/**88.6**	44.3/**99.5**	10.6/97.6	19.4/**93.6**	**18.1/95.4**	**750.52**
Qwen-14B-Chat	80.12$_{(0.36)}$	84.89$_{(0.22)}$	20.2/82.6	39.2/97.5	12.6/96.0	20.8/87.7	16.4/89.4	727.34
Yi-6B-Chat	79.16$_{(11.1)}$	**87.78**$_{(7.11)}$	14.8/85.6	39.5/97.8	7.5/**98.0**	17.3/85.4	11.4/92.7	717.00
Baichuan2-13B-Chat	69.04$_{(\mathbf{0.00})}$	77.11$_{(\mathbf{0.00})}$	**22.9**/83.8	35.9/97.3	9.0/97.3	18.8/93.0	13.8/93.9	711.72
Qwen-7B-Chat	71.81$_{(4.94)}$	82.22$_{(2.67)}$	17.7/79.7	39.5/97.2	12.7/96.9	19.9/83.4	16.5/84.9	702.44
ChatGLM3-6B	63.98$_{(7.95)}$	71.56$_{(7.56)}$	21.0/83.5	36.1/96.4	9.1/95.1	19.0/89.6	14.9/89.1	689.43
TransGPT	62.05$_{(33.3)}$	69.78$_{(27.3)}$	16.1/84.6	38.4/97.2	10.3/96.2	19.3/88.9	15.0/90.2	688.03
InternLM-Chat-20B	62.89$_{(\mathbf{0.00})}$	76.44$_{(\mathbf{0.00})}$	11.0/50.8	**49.6**/95.7	12.1/96.9	**22.2**/90.4	17.2/92.0	677.21
InternLM-Chat-7B	62.65$_{(0.12)}$	66.00$_{(\mathbf{0.00})}$	18.7/72.7	37.8/87.6	**15.4**/88.0	19.9/81.1	17.5/89.6	656.81
LLaMa2-Chinese-13B-Chat-ms	49.64$_{(2.05)}$	62.89$_{(3.33)}$	16.1/75.5	35.6/94.0	10.1/88.3	20.4/84.1	14.1/77.1	627.65

Table 3: The evaluation results of LLMs on the text-only dataset of our TransportationGames. For **T1** and **T2** tasks, the values of Accuracy are listed and the format error rate is placed in the bottom right corner marker. "xx/yy" in the T4∼T10 columns represents the values of the "ROUGE/GPT-4-Eval" metrics. The larger the value of all metrics except the format error rate, the better the performance. "SUM" is the sum of all values of different tasks, and we list all results according to the value of "SUM" from largest to smallest. Results highlighted in **bold** represent the best result in each column.

categorized into three groups: Large Language Models (LLMs), Multimodal Large Language Models (MLLMs), and Transportation-domain Models (T-LMs). Specifically, for LLMs, we select some common models that support Chinese, such as ChatGLM3 (Zeng et al., 2023), Qwen-7/14B-Chat (Bai et al., 2023a), Baichuan2-13B-Chat (Baichuan, 2023), InternLM-Chat-7/20B (Team, 2023), Yi-6B-Chat[11], and LLaMa2-Chinese-13B-Chat-ms[12]. We also evaluate GPT-4[13] on our TransportationGames. For MLLMs, we pick out some models that also support Chinese, such as VisualGLM (Zeng et al., 2023), mPLUG-Owl2 (Ye et al., 2023), Qwen-VL-Chat (Bai et al., 2023b), Chinese-LLaVa-Cllama2[14]/Baichuan[15], InternLM-XComposer-7B (Zhang et al., 2023a), and LLaVa-v1.5-13B (Liu et al., 2023a). Moreover, we also evaluate TransGPT (Peng, 2023) and TransCore-M[16], the open-sourced models in the transportation domain. The more detailed information about these models is shown in Table 2.

4.2 Experimental Settings

We set the input token length limit to 2048 and the output token length to 1024. Right truncation is performed for input prompts exceeding the length limitation. For all open-sourced models, we set the officially recommended decoding strategy for each model. Additionally, we evaluate all models in the zero-shot setting. We utilize the text-only dataset and the multimodal dataset to evaluate LLMs and MLLMs respectively.

4.3 Main Results

The evaluation results of the selected models on our TransportationGames are shown in Table 3 and Table 4. Next, we will introduce the performance of LLMs and MLLMs separately.

Large Language Models Table 3 presents the evaluation results of LLMs on the text-only dataset of our TransportationGames. The values of "SUM" in the last column show that GPT-4 obtains the best performance and Qwen-14B-Chat ranks second. Yi-6B-Chat also achieves outstanding performance on many tasks, such as the **T2** and **T6** tasks, ranking third. Overall, it is promising that some LLMs perform well in memorizing, understanding, and applying transportation knowledge, but there's still room for improvement on many tasks.

Multimodal Large Language Models The evaluation results of MLLMs on the multimodal dataset shown in Table 4 present that Qwen-VL-chat performs excellently on the majority of tasks and ranks first as a whole. InternLM-XComposer-7B ranks second and LLaVa-v1.5-13B ranks third. However, even the top-performing model in the **T3** task, Qwen-VL-chat, achieves only 54.47% accuracy, indicating the poor capability of MLLMs in the multimodal transportation domain.

[11] https://www.modelscope.cn/models/01ai/Yi-6B-Chat/summary

[12] https://www.modelscope.cn/models/modelscope/Llama2-Chinese-13b-Chat-ms/summary

[13] https://chat.openai.com/

[14] https://huggingface.co/LinkSoul/Chinese-LLaVA-Cllama2

[15] https://huggingface.co/LinkSoul/Chinese-LLaVA-Baichuan

[16] https://huggingface.co/PCIResearch/TransCore-M

Models	T3	T4	T6	T7	T8	T9	SUM
Qwen-VL-Chat	**54.47**$_{(0.00)}$	9.3/75.1	**15.3/86.7**	7.4/**70.5**	20.6/**85.9**	14.4/64.5	**504.15**
InternLM-XComposer-7B	48.94$_{(0.00)}$	8.9/77.9	16.1/86.4	**10.5**/56.4	**32.7**/67.7	**19.7**/77.6	502.76
TransCore-M	46.81$_{(0.00)}$	8.0/**79.3**	11.6/82.1	7.2/60.8	13.2/80.3	19.1/77.6	486.01
LLaVa-v1.5-13B	48.94$_{(1.28)}$	10.3/67.4	14.0/79.3	6.5/54.4	15.9/67.6	18.3/77.9	460.51
Chinese-LLaVa-Baichuan	20.43$_{(80.85)}$	6.9/73.5	9.9/84.6	4.2/60.5	10.3/73.4	14.0/**82.0**	439.80
VisualGLM-6B	26.38$_{(79.15)}$	10.1/73.0	11.6/77.6	7.4/64.0	8.8/75.2	14.6/65.6	434.18
mPLUG-Owl2	40.43$_{(0.43)}$	**11.6**/64.0	14.8/71.1	8.8/48.3	22.7/60.8	14.9/70.4	427.66
Chinese-LLaVa-Cllama2	8.09$_{(88.94)}$	7.6/65.5	10.3/83.5	4.5/54.1	9.3/74.7	12.2/79.5	409.39

Table 4: The evaluation results of MLLMs on the multimodal dataset of our TransportationGames. The values of Accuracy are listed for **T3** task and the values of "ROUGE/GPT-4-Eval" metrics are present for the **T4**∼**T9** tasks. And other pattern introduction is the same as Table 3.

4.4 Analysis

Different models have different instruction-following capacities in T1/T2/T3 tasks. According to the format error rate of T1/T2/T3 tasks listed in Table 3 and Table 4, we observe that the format error rate of GPT-4 and the InternLM series models are all zero, demonstrating the excellent instruction-following ability. We speculate that the reason may be that these models have been trained with RLHF.

There is still much room for improvement for some tasks. Due to the varying difficulty of different tasks, the performance of the models also varies. Overall, the model performs poorly on difficult tasks, especially in all tasks of multimodal scenarios as shown in Table 4. This provides a guiding direction for the model to further adapt to the transportation field.

The BaseModel is a key factor affecting model performance. The selection of BaseModel is critical to the overall model performance, as the model learns large-scale knowledge during the pre-training phase. We can observe from Table 3 and Table 4 that the performance of some small-scale models can even outperform that of many large-scale models, such as Yi-6B-Chat surpassing InternLM-Chat-20B, Qwen-7B-Chat surpassing LLaMa2-Chinese-13B-Chat-ms, Qwen-VL-Chat surpassing LLava-v1.5-13B, and so on. Additionally, due to the limited amount of Chinese corpus learned by LLaMa during the pre-training stage, the performance of the LLaMa series models is unsatisfactory such as LLaMa2-Chinese-13B-Chat-ms and Chinese-LLaVa-Cllama2. These results further demonstrate the importance of the BaseModel, which almost determines the upper limit of model performance.

Scaling up the model size improves the performance with the similar BaseModel. The results in Table 3 showcase that Qwen-14B-Chat exceeds Qwen-7B-Chat and InternLM-Chat-20B exceeds InternLM-Chat-7B, which indicates that expanding the model scale will further improve the model performance when the BaseModel is the model of the same series.

5 Conclusion

In this work, we propose TransportationGames, a carefully designed and thorough evaluation benchmark for assessing (M)LLMs in the transportation domain. By comprehensively considering the applications in real-world scenarios, we select 10 varied tasks including the text and image modality. Referring to the first three levels in Bloom's Taxonomy, we categorize these tasks into three skill levels to test the performance of various (M)LLMs in memorizing, understanding, and applying transportation knowledge. The experimental results show that although some models perform well in some tasks, there is still much room for improvement overall. Additionally, we analyze the key factors affecting model performance, which is helpful for how to further improve model performance. We hope the release of TransportationGames can serve as a foundation for future research, thereby accelerating the implementation and application of (M)LLMs in the field of transportation.

Furthermore, due to the need to connect to external databases for some scenarios in the transportation domain, such as real-time road condition queries and traffic flow prediction, our TransportationGames does not include these complex tasks. In future work, we will further test the ability of (M)LLMs as an agent to call relevant interfaces to achieve specified tasks.

Limitations

First, the biggest limitation is data leakage as our data is collected from the Internet. Although the original format of the data is complex and various, it is still difficult to ensure that existing (M)LLMs have not been directly trained on relevant data. We will explore more effective methods to prevent data leakage and strive for a more fair evaluation.

Second, the evaluation of long text generation tasks is very difficult, and we used ROUGE-L and GPT-4-Eval to evaluate the model-generated predictions together in our work. Due to the non-uniqueness of the answers, it is still difficult to ensure that the same effect as manual evaluation can be achieved.

Moreover, due to time constraints and the large amount of existing open-source models, we only test a small portion of common models in this work. We will test more models in the future.

References

Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2000. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on llm for high-functioning autistic adolescent psychological counseling.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Longchao Da, Kuanru Liou, Tiejin Chen, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, and Hua Wei. 2023. Open-ti: Open traffic intelligence with augmented language model.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: A chinese legal large language models benchmark (a technical report).

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. Explanatory argument extraction of correct answers in resident medical exams.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023a. Lawyer llama technical report.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2023. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review.

Tjalling C Koopmans. 1949. Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, pages 136–146.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2023. Large language models as traffic signal control agents: Capacity and opportunity.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player?

Baichuan Mo, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo, and Jinhua Zhao. 2023. Large language models for travel behavior prediction.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Jiawei Han, and Lianhui Qin. 2023. Structured chemistry reasoning with large language models.

Wang Peng. 2023. Duomo/transgpt.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Nan Shao, Zefan Cai, Chonghua Liao, Yanan Zheng, Zhilin Yang, et al. 2023. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

George R Taylor. 2015. *The transportation revolution, 1815-60*. Routledge.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

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. 2023. Llama: Open and efficient foundation language models.

Geemi P. Wellawatte and Philippe Schwaller. 2023. Extracting human interpretable structure-property relationships in chemistry using xai and large language models.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan

Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiaocheng Zhang, Zonghai Yao, and Hong Yu. 2023b. Ehr interaction between patients and ai: Noteaid ehr interaction.

Xiaodan Zhang, Sandeep Vemulapalli, Nabasmita Talukdar, Sumyeong Ahn, Jiankun Wang, Han Meng, Sardar Mehtab Bin Murtaza, Aakash Ajay Dave, Dmitry Leshchiner, Dimitri F. Joseph, Martin Witteveen-Lane, Dave Chesla, Jiayu Zhou, and Bin Chen. 2023c. Large language models in medical term classification and unexpected misalignment between response and reasoning.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023a. Mmicl: Empowering vision-language model with multi-modal in-context learning.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023c. A survey of large language models.

Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Li Yun, Hejie Cui, Zhang Xuchao, Tianjiao Zhao, et al. 2023d. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

A Data Distribution

The detailed data distribution is shown in Figure 2. All data in our TransportationGames has undergone meticulous manual verification.

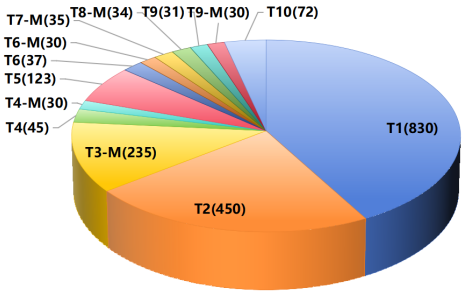


Figure 2: The distribution of data amounts for different tasks. "-M" means the multimodal dataset.

B Examples of Tasks

We list some examples for each task in Figure 3, Figure 4, and Figure 5.

C An Example Instruction for GPT-4-Eval

We utilize GPT-4 to evaluate the model-generated answers for accuracy, redundancy, fluency, and completeness. The English version of the instruction is "Below, I will give a question and a standard answer to the question, as well as an answer generated by the question-and-answer model. Since the answer is not unique, please judge the rationality of the answer generated by the question-and-answer model according to the reference answer given and combined with the actual situation, and it is necessary to consider the logic/accuracy/redundancy/fluency/integrity of the generated answer. The returned format is JSON, and the field is gpt4-score: The value is a decimal in the range of 0 to 1. Three decimal places are reserved after the decimal point. Question: xxx. Standard answer: xxx. The answer generated by the question-and-answer model: xxx."

(T1) Traffic Concepts Question Answering

Question: 请判断下面的说法是否正确？只需要回答正确或错误即可；初次取得的机动车驾驶证的有效期为10年。
Please determine whether the following statement is correct. Just answer right or wrong.
A motor vehicle driving license obtained for the first time is valid for 10 years.

→Answer: 错误
Wrong

Example-1

Question: 请从下面的A、B、C、D四个答案中给问题选出一个正确答案；只需要回答选项名A或B或C或D即可；
交通标志版面安装时应平整完好，在标志面任何一处面积为500mm×500mm的范围内不得存在总面积大于多少平方毫米的一个或一个以上气泡？
A.10 B.20 C.15 D.25
Please choose a correct answer from the four answers A, B, C and D below. Just answer the option name A or B or C or D;
When the traffic sign layout is installed, it should be smooth and intact, and there should not be one or more bubbles with a total area greater than how many square millimeters within the range of any part of the sign surface with an area of 500mm×500mm?
A.10 B.20 C.15 D.25

→Answer: A

Example-2

(T2) Traffic Regulations Question Answering

Question: 请判断下面的说法是否正确？只需要回答正确或错误即可；绿色方向指示信号灯的箭头方向向上，表示准许车辆直行。
Please determine whether the following statement is correct. Just answer right or wrong. The arrow of the green direction indicator light points upward, indicating that the vehicle is allowed to go straight.

→Answer: 正确
Right

Example-1

Question: 请从下面的A、B、C、D四个答案中给问题选出一个正确答案；只需要回答选项名A或B或C或D即可；
汽车遇雨天，能见度在50米以内时，最高时速不准超过多少？
A.30公里 B.45公里 C.50公里 D.60公里
Please choose a correct answer from the four answers A, B, C and D below. Just answer the option name A or B or C or D; When the car meets a rainy day, the visibility is within 50 meters, the maximum speed is not allowed to exceed how much?
A.30 km B.45 km C.50 km D.60 km

→Answer: A

Example-2

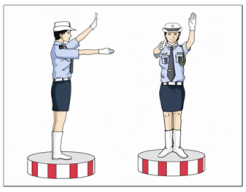
(T3) Traffic Signs Question Answering

Question: 结合所给图片，请判断下面的说法是否正确？只需要回答正确或错误即可；
图中是车辆易滑标志。
Combined with the pictures given, please judge whether the following statement is correct. Just answer right or wrong. The picture shows the ease of vehicle slip sign.

→Answer: 正确
Right

Example-1

Question: 请结合所给图片，从下面的A、B、C、D四个答案中给问题选出一个正确答案；
只需要回答选项名A或B或C或D即可；
图中警察手势为什么信号？
A.靠左停车 B.停止 C.右转弯 D.靠边停车
Please choose A correct answer from the following four answers A, B, C and D. Just answer the option name A or B or C or D;
What signal does the policeman's hand signal in the picture signify??
A. Stop to the left B. stop C. Turn right D. Pull over to the side of the road



→Answer: D

Example-2

Figure 3: There are some examples for the **T1/T2/T3** tasks. The blue text represents the answer, and the gray text is the version of English.

(T4) Traffic Accidents Analysis

Question: 请从下面给出的交通事故发生经过中总结出事故的起因；2022年3月17日6时34分许，王*驾驶豫*号重型特殊结构货车由金志路方向沿高腾大道越过道路中心双黄实线左转往金曾路方向行驶，车行驶至重庆市高新区高腾大道金曾路口人行横道线处时，其车辆车头将由其车行方向由右至左由人行横道线处横过道路的行人覃**碰撞倒地后，车辆右侧车轮碾压覃**，造成行人覃**当场死亡的交通事故。

Please summarize the cause of the accident from the following traffic accident events; At 6:34 on March 17, 2022, Wang * drove Yu * heavy truck with special structure along Gaoteng Avenue from the direction of Jinzhi Road, crossed the double Yellow Solid line of the road center and turned left to the direction of Jinzeng Road. When the car reached the pedestrian crossing line at Jinzeng Intersection of Gaoteng Avenue, High-tech Zone, Chongqing, The front of the vehicle will be from the direction of the vehicle from right to left by the pedestrian crossing line across the road after the collision, the right wheel of the vehicle crushed Tan **, resulting in the pedestrian Tan ** on the spot death of the traffic accident.

→Answer: 王*驾驶机动车行驶过程中违反交通标线通行，遇行人正在通过人行横道时未停车让行，是本次事故发生的直接原因，覃**在事故中无过错行为。

Wang * driving motor vehicles in the process of violation of traffic markings, pedestrians are passing through the crosswalk did not stop to give way, is the direct cause of the accident, Tan ** in the accident without fault behavior.

Example-1

Question: 请分析图中可能造成交通事故发生的原因。

Please analyze the possible causes of traffic accidents in the figure.

→Answer: 图中银色车辆非法变道与相邻车道的车辆相撞。

The silver vehicle in the picture illegally changed lanes and collided with the vehicle in the adjacent lane.

Example-2

(T5) Traffic Pubic Sentiment Analysis

Question: 下面我会给出公众提出的一些问题或者给出的一些建议，请从中概括出用户的真实诉求或建议；
真南路现状是连接嘉定区和上海市中心的主干道，同时真南路也是一条货物运输的主要道路，规划真南路在众仁路口改道截断，并缩窄红线宽度至20米，对真南路车流影响很大，降低真南路道路通行能力，造成比现状更加拥堵的情况，同时会产生噪音、污染、尘土、扰民等问题，小区居住品质下降。

Below I will give some questions or suggestions raised by the public, please summarize the real demands or suggestions of users;
The current status of Zhennan Road is the main road connecting Jiading District and downtown Shanghai. At the same time, Zhennan Road is also a major road for cargo transportation. The planned diversion and truncation of Zhennan Road at the entrance of Zhongren Road and the narrowing of the red line width to 20 meters will greatly affect the traffic flow of Zhennan Road and reduce the traffic capacity of Zhennan Road, resulting in more congestion than the current situation. At the same time, it will produce noise, pollution, dust, nuisance and other problems, and the living quality of the community will decline.

→Answer: 保留真南路的通行能力，避免改道截断和缩窄红线，以维护交通流畅和改善居住环境。

To preserve the capacity of True South Road, avoid diversion to cut off and narrow the red line, in order to maintain traffic flow and improve the living environment.

Example-1

(T6) Traffic Safety Recommendation

Question: 你是交通领域的专家，请回答我下面提出来的问题；若某地有较强台风预警，某人计划出行，对于交通出行他应考虑哪些事项？同时评估该路段的交通事故风险。

You are an expert in the field of transportation, please answer the questions I put forward below; If there is a strong typhoon warning in a place and someone plans to travel, what should he consider about transportation? At the same time, the risk of traffic accidents is assessed.

→Answer: 台风天气，请确保车辆停放在安全的地方，并注意避开易发生水灾的低洼地区。该路段交通事故的可能性较高，请尽量避免出行。

In typhoon weather, please ensure that your vehicle is parked in a safe place, and take care to avoid low-lying areas prone to flooding. There is a high probability of traffic accidents on this section of road, please try to avoid traveling.

Example-1

Question: 分析图中的交通隐患并给出相应的安全出行建议。

Traffic hazards in the figure are analyzed and corresponding safe travel suggestions are given.

→Answer: 电动车超载会增加操控难度，甚至出现前轮翘起、转弯困难的情况，并造成制动距离延长等问题。在进入弯道、横向盘山坡道或制度时，极易产生侧滑或原地侧翻、掉沟等严重事故。安全出行电动车出行应严禁超载超员驾驶。

Overload of electric vehicles will increase the difficulty of handling, and even appear front wheel warping, difficult turning, and cause problems such as prolonged braking distance. Serious accidents such as sideslip, rollover in place, ditching, etc. are likely to occur when entering a curve, lateral mountain ramp or system. Safe travel Electric vehicle travel should be strictly prohibited overloading and overdriving.

Example-2

Figure 4: There are some examples for the **T4/T5/T6** tasks.

(T7) Traffic Sign Error Detection

Question: 请分析以下图片中交通标线存在的不合理的地方。
Please analyze the unreasonable place of traffic marking in the following picture.

→Answer: 新旧交通标志线并存，容易造成干扰。
The coexistence of old and new traffic signs is easy to cause interference.

(T8) Traffic Road Occupation Detection

Question: 请分析图片中存在什么非法占道现象。
Please analyze the picture for illegal traffic.

→Answer: 图片中存在非法占用消防通道的现象。
There is illegal occupation of fire escape in the picture.

(T9) Traffic Emergency Plan Generation

Question: 你是交通领域的专家，请回答下面我提出的问题；地铁安全事故的应急处理方案是？
You are an expert in the field of transportation. Please answer the following questions.
What is the emergency treatment plan for subway safety accidents?

→Answer: （1）发生地铁安全事故后，地铁工作人员应立即通知调度中心，并启动应急预案。
（2）地铁工作人员应迅速疏导乘客，确保人员安全疏散，并组织人员进行抢救和急救工作。
（3）地铁工作人员应紧急关闭事故车站，并对周边区域进行交通管制，以确保现场安全。
（4）地铁工作人员应与相关部门密切合作，展开事故调查工作，并采取措施减轻事故影响，如及时修复设备、恢复运营等。
（5）地铁工作人员应向媒体和公众及时发布准确信息，避免造成恐慌和不必要的传言，维护社会稳定。
（6）地铁公司应建立健全地铁安全监测预警系统，加强事故防范和应急管理，提高应对突发事件的能力和效率。
(1) After a subway safety accident occurs, the subway staff shall immediately notify the dispatch center and initiate an emergency plan.
(2) Subway staff should quickly guide passengers, ensure the safe evacuation of personnel, and organize personnel to rescue and first aid work.
(3) Subway staff should urgently close the accident station and conduct traffic control in the surrounding area to ensure the safety of the scene.
(4) Subway staff should work closely with relevant departments to investigate the accident and take measures to mitigate the impact of the accident, such as timely repair of equipment and resumption of operations.
(5) Subway staff should timely release accurate information to the media and the public to avoid panic and unnecessary rumors and maintain social stability.
(6) The subway company shall establish and improve the subway safety monitoring and early warning system, strengthen accident prevention and emergency management, and improve the ability and efficiency of responding to emergencies.

Example-1

Question: 上图中大巴车在高速路段发生车祸，造成2死数伤，堵车4公里，道路管理部门对此事故的应急处理方案应该是什么？
In the above picture, the bus in the high-speed section of the accident, resulting in 2 deaths and several injuries, 4 km traffic jam, the road management department of this accident emergency treatment plan should be what?

→Answer: 1.在来车方向设置警示标志；
2.组织有行动能力的人员迅速撤离到公路护栏外安全位置；
3.确定受伤人员和死亡人员的数量和情况，及时处理受伤人员，并组织救援人员进行援救；
4.必要时拆解事故车辆，以尽快完成救援并恢复交通通畅。
1. Set up warning signs in the direction of incoming vehicles.
2. Organize personnel with the ability to quickly evacuate to a safe position outside the highway guardrail.
3. Determine the number and condition of the injured and the dead, treat the injured in a timely manner, and organize rescue workers to carry out rescue.
4. Disassemble the accident vehicle if necessary to complete rescue and restore traffic flow as soon as possible.

Example-2

(T10) Traffic Safety Education Copy Generation

Question: 你是交通领域的教育专家，请帮我给学生设计公共出行主题的交通安全教育材料
You are an education expert in the field of transportation. Please help me design traffic safety education materials on the theme of public travel for students

→Answer: 1、不在机动车道等公交车和追逐车辆。
2、排队候车，有序上下车。
3、乘坐时坐稳扶好，在车内不玩耍打闹，不饮食，不把头、手等身体任何部位伸出窗外。
4、不乘坐超载或无载客许可证的车辆。
5、车子停稳后从右侧下车。
1. Do not wait for buses and chase vehicles in the motorway.
2. Queue up and get on and off the bus in an orderly manner.
3. Sit firmly and hold well when riding, do not play in the car, do not eat, do not stick out of the window any part of the body such as the head, hands.
4. Do not take overloaded or no passenger permit vehicles.
5. Get off from the right side after the car has stopped.

Figure 5: There are some examples for the **T7/T8/T9/T10** tasks.