STBench: Assessing the Ability of Large Language Models in Spatio-Temporal Analysis

Wenbin Li[1,2], *Di Yao[1], Ruibo Zhao[1,2], Wenjie Chen[1,2], Zijie Xu[1,2], Chengxue Luo[1,2], Chang Gong[1,2], Quanliang Jing[1], Haining Tan[1], *Jingping Bi[1]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[1]
University of Chinese Academy of Sciences, China[2]
{liwenbin20z,yaodi,zhaoruibo23s,chenwenjie23s,xuzijie22s}@ict.ac.cn
{luochengxue21s,gongchang21z,jingquanliang,tanhaining,bjp}@ict.ac.cn

ABSTRACT

The rapid evolution of large language models (LLMs) holds promise for reforming the methodology of spatio-temporal data mining. However, current works for evaluating the spatio-temporal understanding capability of LLMs are somewhat limited and biased. These works either fail to incorporate the latest language models or only focus on assessing a specific dimension of spatio-temporal capabilities, making the evaluation not comprehensive. To address this gap, this paper dissects LLMs' capability of spatio-temporal data into four distinct dimensions: knowledge comprehension, spatio-temporal reasoning, accurate computation, and downstream applications. We curate several natural language question-answer tasks for each category and build the benchmark dataset, namely STBench, containing 15 distinct tasks and over 70,000 QA pairs. Moreover, we have assessed the capabilities of 13 LLMs and experimental results reveal that existing LLMs show remarkable performance on knowledge comprehension and spatio-temporal reasoning tasks, with potential for further enhancement on other tasks through in-context learning, chain-of-thought prompting, and fine-tuning. The code and datasets of STBench are released on https://github.com/LwbXc/STBench.

1 INTRODUCTION

Large language models (LLMs) have shown potential in various domains [11, 24, 26, 31] and one promising direction is enhancing spatio-temporal data analysis with the ability of LLMs [4, 6, 14, 15, 17, 30]. While spatio-temporal data encompasses a variety of datasets crucial for many fields such as geography, meteorology, transportation, and epidemiology, the applicability and effectiveness of LLMs in handling spatio-temporal data remain relatively unexplored.

Related Works. Most existing evaluations of spatio-temporal abilities only focus on a specific dimension. For instance, a series of work [12, 16, 19, 23] focus on evaluating the spatial reasoning capability of LLMs by constructing QA pairs in toy environments without temporal information, and some benchmarks mainly assess the memory ability of spatio-temporal knowledge [9, 20].

In this paper, we propose a framework, namely STBench, to comprehensively evaluate the spatio-temporal capabilities of LLMs. STBench dissects the LLMs' capacity into four distinct dimensions: knowledge comprehension, spatio-temporal reasoning, accurate computation, and downstream applications. **Knowledge Comprehension (KC)** examines the model's capacity to understand and interpret the underlying meaning and context of spatio-temporal entities and concepts. **Spatio-Temporal Reasoning (STR)** evaluates the ability to understand and reason about the spatial and temporal relationships between entities and events. **Accurate Computation (AC)** handles the precise and complex calculations of spatio-temporal data. Moreover, we also employ some **Downstream Applications (DA)** such as trajectory anomaly detection and trajectory prediction to assess the ability of LLMs on practical tasks. We summarize and compare STBench with existing works in Table1.

Table 1: Comparing benchmarks for LLMs' ability in spatio-temporal analysis.

	Format	Temporal	KC	STR	AC	DA
[23]	QA	✗	✗	✓	✗	✗
[19]	QA	✗	✗	✓	✗	✗
[12]	QA	✗	✗	✓	✗	✗
[5]	QA	✗	✓	✗	✗	✗
[16]	QA	✓	✓	✗	✗	✓
[20]	QA	✓	✓	✓	✓	✗
[9]	Probing	✓	✓	✗	✗	✗
[8]	QA	✓	✓	✓	✗	✓
Ours	QA	✓	✓	✓	✓	✓

For each dimension, we design several tasks and construct QA pairs to qualitatively assess the ability of LLMs. We have curated a benchmark dataset, STBench, which contains over 70,000 QA pairs and 15 distinct tasks covering the four dimensions. Furthermore, we evaluated the latest 13 LLMs, including GPT-4o[1], Gemma [18], Llama2 [25], and provide a detailed evaluation report[2]. The results show that existing LLMs show remarkable performance on knowledge comprehension and spatio-temporal reasoning tasks, with potential for further enhancement on other tasks through in-context learning, chain-of-thought prompting, and fine-tuning.

The contributions of this paper are summarized as following:

- This paper presents STBench, a comprehensive benchmarking framework designed to evaluate the spatio-temporal analysis capabilities. For a comprehensive evaluation, STBench categorizes spatio-temporal abilities into four dimensions, each with multiple tasks tailored to various data types, including POI, trajectory, region and traffic flow.

* Corresponding authors.

[1]https://platform.openai.com/docs/models/gpt-4o
[2]https://arxiv.org/abs/2406.19065

- We assessed 13 LLMs and conducted a detailed analysis of their performance. The results highlight the remarkable performance of LLMs in knowledge comprehension and spatio-temporal reasoning tasks, while also identifying areas for improvement in accurate computation and downstream applications.
- Multiple enhancement methods, including in-context learning, chain-of-thought and supervised fine-tuning, are incorporated to enhance LLMs' ability. The results reveal LLMs' great potential in spatio-temporal analysis.

2 BENCHMARK CONSTRUCTION

Data Format. To make the output of LLMs controllable and to make it easier to identify the final answer, all data samples in STBench are in the form of text completion rather than chatting. We provide the question and several options to the model, and expect it to generate an option number by completing the text "The answer is <Option>".

Knowledge Comprehension. We design four tasks for this aspect: (1) *POI Category Recognition (PCR)* evaluates LLMs' understanding of POI semantics by asking them to predict the POI category according to the coordinates and related comments. (2) *POI Identification (PI)* provides coordinates and comments of two POIs and asks LLMs to determine if they are the same POI. (3) *Urban Region Function Recognition (URFR)* requires LLMs to predict the urban region function according to the boundary lines and the POIs located in the region, which evaluates LLMs' understanding of urban regions. (4) *Administrative Region Determination (ARD)* refers to determining which administrative region a coordinate falls in, which involves relevant knowledge of the administrative regions and the ability to associate it with geographical coordinates.

Spatio-Temporal Reasoning. Four tasks are designed to assess this ability of LLMs: (1) *Point-Trajectory Relationship Detection (PTRD)* is to determine whether a trajectory passes through a point; given a point and several regions, (2) *Point-Region Relationship Detection (PRRD)* aims to infer which region the point falls in; (3) *Trajectory-Region Relationship Detection (TRRD)* provides a trajectory and some regions, and asks LLMs to determine which regions the trajectory has passed through chronologically; (4) *Trajectory Identification (TI)* requires LLMs to determine if two point sequences are sampled from the same trajectory. These tasks evaluate the model's ability to infer spatio-temporal relationships between point, line and surface.

Accurate Computation. There are three tasks for this dimension: (1) *Direction Determination (DD)* is to determine the direction between two geographical points, which involves calculating the corresponding azimuth according to the coordinates; (2) *Navigation (NAV)* requires LLMs to plan a shortest route from a source point to a destination point based on a given road network; (3) *Trajectory-Trajectory Relationship Analysis (TTRA)* asks LLMs to calculate the number of times two trajectories encounter each other simultaneously in time and space.

Downstream Applications. We assess this ability of LLMs through four tasks: (1) *Flow Prediction (FP)* is to predict traffic inflows and outflows based on the historical inflows and outflows, which requires LLMs to model the periodicity and trends of traffic changes; (2) *Trajectory Anomaly Detection (TAD)* asks LLMs to detect anomalous trajectories, which requires LLMs to infer the underlying route and shape from trajectory data; (3) *Trajectory Classification (TC)* aims to infer the category of a trajectory, which requires the model to comprehensively consider the coordinates, length, speed and other relevant information; (4) *Trajectory Prediction (TP)* is to predict the next point based on the historical points of a trajectory, involving the ability to model the trajectory patterns and the moving speed.

3 EXPERIMENTS

We conduct extensive experiments on STBench to evaluate the spatial-temporal ability of LLMs and to investigate if in-context learning, chain-of-thought and fine-tuning can improve the performance.

3.1 Experimental setup

Evaluated models. We evaluate the performance of two closed-source model, *i.e.*, **ChatGPT** and **GPT-4o**, and a set of open-source models: **Llama-2** [25], **Vicuna**[3], **ChatGLM2**, **ChatGLM3** [7, 29], **Gemma** [18], **Phi-2**, **Mistral** [10], **Falcon** [1], **Deepseek** [3], **Qwen** [2] and **Yi** [28]. The detailed versions of the evaluated models can be found in [13].

Metrics. We adopt accuracy for tasks other than trajectory prediction and flow prediction. For trajectory prediction, we report absolute error, *i.e.*, the distance in meters between the predicted coordinates and ground truth. For flow prediction, we adopt MAE and RMSE as the metrics.

Experimental details. In our experiments, we adopt the precision of FP32 for all LLMs. For other hyperparameters, we adopt the default value of each model. All experiments of open source models are conducted on two NVIDIA H100.

3.2 Main results

To investigate the spatio-temporal ability of LLMs, we conduct experiments to evaluate the performance of all models on each task. The main results are shown in Table 2. Detailed results on all tasks can be found in [13]. Some key findings are shown as follows:

Model size is important for knowledge comprehension. For knowledge comprehension, GPT-4o performs better than ChatGPT on all tasks, and ChatGPT outperforms other models on most tasks. The possible reason is that LLMs rely on sufficient parameters to compress and store knowledge, and ChatGPT/GPT-4o has more parameters than other evaluated open-source models. We also observe that Gemma-7B, with the same technology as Gemma-2B but more parameters, performs better on knowledge comprehension tasks than Gemma-2B. It also supports our conclusion.

The evaluated models have difficulty in multi-step reasoning. The performance of most models on PRRD is much higher than TRRD. Note that TRRD can be achieved by performing PRRD for each point in the trajectory, thus it is a multi-step reasoning task. Although models such as ChatGPT, GPT-4o, and Gemma-7B can achieve high performance on each step, their performance on this multi-step task is poor.

[3]https://lmsys.org/blog/2023-03-30-vicuna/

Table 2: The performance of *ACC*, *MAE* and absolute error (in meters) on different tasks (bold: best closed-source LLM; underline: best open-source LLM). '-' denotes the model failed to answer most questions.

	Knowledge Comprehension			Spatio-temporal Reasoning			Accurate Computation		Downstream Application		
	PCR	URFR	ARD	PTRD	PRRD	TRRD	DD	NAV	FP(*MAE*)	TAD	TP(meters)
ChatGPT	0.7926	0.3978	0.8358	**0.7525**	**0.9240**	0.0258	0.1698	0.4384	**37.33**	0.5382	-
GPT-4o	**0.9588**	**0.6026**	**0.9656**	-	0.9188	**0.1102**	**0.5434**	**0.7552**	43.25	**0.6016**	-
ChatGLM2	0.2938	0.2661	0.2176	0.2036	0.5216	0.2790	0.1182	0.2924	63.72	0.5000	231.2
ChatGLM3	0.4342	0.2704	0.2872	0.3058	0.8244	0.1978	0.1156	0.2576	59.24	0.5000	224.5
Phi-2	-	-	0.2988	-	-	-	0.1182	0.2912	34.82	0.5000	206.9
Llama-2-7B	0.2146	0.2105	0.2198	0.2802	0.6606	0.2034	0.1256	0.2774	53.79	0.5098	189.3
Vicuna-7B	0.3858	0.2063	0.2212	0.3470	0.7080	0.1968	0.1106	0.2588	48.19	0.5000	188.1
Gemma-2B	0.2116	0.1989	0.1938	0.4688	0.5744	0.2014	0.1972	0.2592	41.79	0.5000	207.7
Gemma-7B	0.4462	0.2258	0.2652	0.3782	0.9044	0.1992	0.1182	0.3886	31.85	0.5000	139.4
DeepSeek-7B	0.2160	0.2071	0.1938	0.2142	0.6424	0.1173	0.1972	0.3058	56.89	0.5000	220.8
Falcon-7B	0.1888	0.1929	0.1928	0.1918	0.4222	0.2061	0.1365	0.2610	62.52	0.5000	3572.8
Mistral-7B	0.3526	0.2168	0.3014	0.4476	0.7098	0.0702	0.1182	0.3006	42.59	0.5000	156.8
Qwen-7B	0.2504	0.2569	0.2282	0.2272	0.5762	0.1661	0.1324	0.3106	53.49	0.5049	205.2
Yi-6B	0.3576	0.2149	0.1880	0.5536	0.8264	0.1979	0.1284	0.3336	52.03	0.5000	156.2

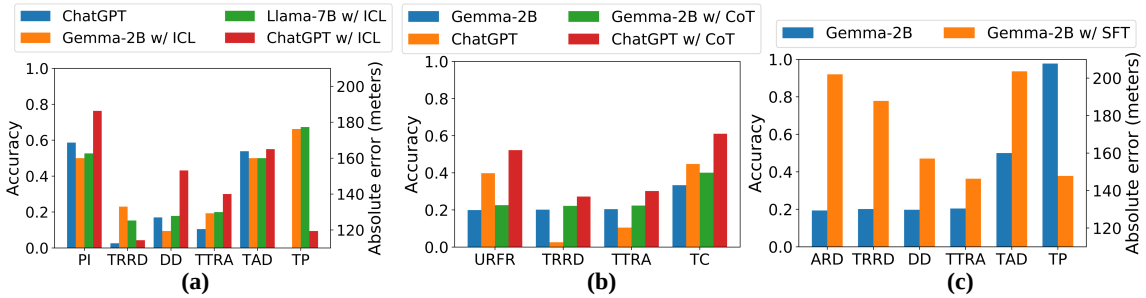

Figure 1: The performance of ACC and absolute error (in meters) in (a) in-context learning evaluation, (b) chain-of-thought evaluation, (c) fine-tuning evaluation.

Accurate computation and downstream tasks are more challenging. As shown in Table 2, the accuracy of all models except GPT-4o is below 45% on accurate computation tasks, which is because LLMs are mainly trained on nature language corpus and are not good at computation. We also find that GPT-4o outperforms other LLMs by a large margin, which is consistent with the significant improvement in mathematical ability of GPT-4o. Moreover, the performance of evaluated models is also poor on downstream tasks, due to the lack of task-specific expert knowledge.

A suitable model is more important than larger parameters for spatio-temporal mining. We observe ChatGPT and GPT-4o perform poorer than most open-source models on TRRD and TI, despite having a larger number of parameters. For FP, the lightweight model, Phi-2, with only 2.7B parameteres, performs better than all models except Gemma-7B. Although LLMs have the potential to analyze spatio-temporal data, not all models have been adequately trained on relevant corpora and learned corresponding spatio-temporal ability, regardless of the model size. It leads to a significant difference in performance between different models for many spatio-temporal tasks.

3.3 In-context learning evaluation

Whereas the results in many scenarios are poor, we conduct experiments to investigate if in-context learning can improve the performance of LLMs on STBench. Specifically, we select six tasks where the evaluated models performed poorly and we adopt two-shot prompting. Due to the heavier computation cost caused by the longer context, we only evaluate one closed-source model, ChatGPT, and two open-source models with different model sizes, *i.e.*, Gemma-2B and Llama-2-7B. The results are shown in Fig. 1(a).

The performance of ChatGPT has been greatly improved with in-context learning. For instance, its accuracy on PI and DD has increased from 58.64% to 76.30%, and from 16.98% to 43.16%, respectively. Moreover, the two-shot prompting also constrains the output, *e.g.*, ChatGPT refuses to answer the questions of trajectory prediction in Table 2, but its absolute error is only 119.4 meters with two-shot prompting. However, in-context learning is useless for Gemma-2B and Llama-2-7B, which is consistent with the phenomenon that in-context learning is less effective for smaller LLMs [27].

3.4 Chain-of-thought evaluation

We further conduct experiments to verify if chain-of-thought (CoT) is effective on STBench. Specifically, we evaluate ChatGPT and Gemma-2B with CoT prompting on several tasks that involve multi-step reasoning: URFR, TRRD, TTRA and TC. For each task, we add two samples with a detailed reasoning process in the context. The results are shown in Fig. 1(b).

Table 3: The performance of *MAE* and *RMSE* on traffic prediction (bold: best; underline: runner-up).

		Gemma-2B w/ SFT	STID	PatchTST
Inflow	MAE	<u>26.79</u>	38.57	**24.43**
	RMSE	<u>30.87</u>	43.62	**28.28**
Outflow	MAE	<u>25.91</u>	36.96	**23.49**
	RMSE	<u>29.87</u>	42.04	**27.25**

We observe the performance of ChatGPT increases significantly in all selected tasks. For instance, its accuracy with CoT prompting is 52.20% on URFR, much better than 39.78% in Table 2. For Gemma-2B, the performance on all selected tasks is slightly improved. The results demonstrate the effectiveness of CoT prompting in spatio-temporal analysis.

3.5 Fine-tuning Evaluation

We conduct experiments to investigate if supervised fine-tuning (SFT) can significantly improve the performance of smaller LLMs. Due to the high computational cost and memory usage, we only fine-tune a 2B model, *i.e.*, Gemma-2B. To compare the fine-tuned LLM with existing supervised methods, we train two effective flow prediction method, *i.e.*, STID [22] and PatchTST [21], on the same dataset. The results are shown in Fig. 1(c) and Table 3.

The performance on all tasks in Fig. 1 is significantly improved after fine-tuning. For instance, the accuracy on ARD and DD increased from 19.89% to 91.98%, and from 19.72% to 47.08%, respectively. This confirms LLMs' great potential in spatial-temporal analysis. We also observe that the zero-shot capability of LLMs is surprising that Phi-2 (without fine-tuning and few-shot prompting) can surpass the supervised method STID. While Gemma-2B performs poorer than both STID and PatchTST, it outperforms STID and achieved comparable performance to PatchTST after supervised fine-tuning. Overall, the experimental results reveal the bright prospects of LLMs in spatio-temporal data analysis.

4 CONCLUSION

In this paper, we propose STBench to assess LLMs' ability in spatio-temporal analysis. STBench consists of 15 tasks and over 70,000 QA pairs, systematically evaluating four dimensions: knowledge comprehension, spatio-temporal reasoning, accurate computation, and downstream applications. We benchmark 13 latest LLMs and the results show their remarkable performance on knowledge comprehension and spatio-temporal reasoning tasks. Our further experiments with in-context learning, chain-of-thought prompting and fine-tuning also prove the great potential of LLMs on other tasks.

ACKNOWLEDGEMENTS

This work has been supported by the Natural Science Foundation of China under Grant No. 62472405. POI information from the Yelp dataset[4], trajectories from Geolife[5] and DiDi[6], and urban region information from the New Orleans dataset[7] are incorporated to generate the data samples of STBench.

[4]https://www.yelp.com/dataset
[5]https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/
[6]https://gaia.didichuxing.com/
[7]https://catalog.data.gov/dataset/zoning-district-9939c

REFERENCES

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, et al. 2023. The Falcon Series of Open Language Models. *CoRR* abs/2311.16867 (2023).
[2] Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen Technical Report. *CoRR* abs/2309.16609 (2023).
[3] Xiao Bi, Deli Chen, Guanting Chen, et al. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *CoRR* abs/2401.02954 (2024).
[4] Mark Chen, Jerry Tworek, Heewoo Jun, et al. 2021. Evaluating Large Language Models Trained on Code. *CoRR* abs/2107.03374 (2021).
[5] Rémy Découpes, Roberto Interdonato, Mathieu Roche, et al. 2024. Evaluation of Geographical Distortions in Language Models: A Crucial Step Towards Equitable Representations. *CoRR* abs/2404.17401 (2024).
[6] Cheng Deng, Tianhang Zhang, Zhongmou He, et al. 2024. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. In *WSDM 2024*. 161–170.
[7] Zhengxiao Du, Yujie Qian, Xiao Liu, et al. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *ACL 2022*. 320–335.
[8] Jie Feng, Jun Zhang, Junbo Yan, et al. 2024. CityBench: Evaluating the Capabilities of Large Language Model as World Model. *CoRR* abs/2406.13945 (2024).
[9] Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *ICLR 2024*.
[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023).
[11] Ming Jin, Qingsong Wen, Yuxuan Liang, et al. 2023. Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook. *CoRR* abs/2310.10196 (2023).
[12] Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024. Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark. In *AAAI 2024*. 18500–18507.
[13] Wenbin Li, Di Yao, Ruibo Zhao, et al. 2024. STBench: Assessing the Ability of Large Language Models in Spatio-Temporal Analysis. *CoRR* abs/2406.19065 (2024).
[14] Zhonghang Li, Lianghao Xia, Jiabin Tang, et al. 2024. UrbanGPT: Spatio-Temporal Large Language Models. In *KDD 2024*. 5351–5362.
[15] Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. 2023. GPT-ST: Generative Pre-Training of Spatio-Temporal Graph Neural Networks. In *NeurIPS 2023*.
[16] Gengchen Mai, Weiming Huang, Jin Sun, et al. 2024. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). *ACM Trans. Spatial Algorithms Syst.* 10, 2 (2024), 11.
[17] Rohin Manvi, Samar Khanna, Marshall Burke, et al. 2024. Large Language Models are Geographically Biased. In *ICML 2024*.
[18] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR* abs/2403.08295 (2024).
[19] Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. In *EMNLP 2022*. 6148–6165.
[20] PETER MOONEY, WENCONG CUI, BOYUAN GUAN, et al. 2023. Towards Understanding the Spatial Literacy of ChatGPT. In *SIGSPATIAL 2023*.
[21] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, et al. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR 2023*.
[22] Zezhi Shao, Zhao Zhang, Fei Wang, et al. 2022. Spatial-Temporal Identity: A Simple yet Effective Baseline for Multivariate Time Series Forecasting. In *CIKM 2022*. 4454–4458.
[23] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *AAAI 2022*. 11321–11329.
[24] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, et al. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.
[25] Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023).
[26] Lei Wang, Chen Ma, Xueyang Feng, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 1–26.
[27] Jason Wei, Yi Tay, Rishi Bommasani, et al. 2022. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.* 2022 (2022).
[28] Alex Young, Bei Chen, Chao Li, et al. 2024. Yi: Open Foundation Models by 01.AI. *CoRR* abs/2403.04652 (2024).
[29] Aohan Zeng, Xiao Liu, Zhengxiao Du, et al. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *ICLR 2023*.
[30] Yifan Zhang, Cheng Wei, Shangyou Wu, et al. 2023. GeoGPT: Understanding and Processing Geospatial Tasks through An Autonomous GPT. *CoRR* abs/2307.07930 (2023).
[31] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2023. A Survey of Large Language Models. *CoRR* abs/2303.18223 (2023).