mathematics

MDPI

Article

Application of Large Language Models and Assessment of Their Ship-Handling Theory Knowledge and Skills for Connected Maritime Autonomous Surface Ships

Dashuai Pei [1], Jianhua He [2,*], Kezhong Liu [1,*], Mozi Chen [1] and Shengkai Zhang [3]

[1] School of Navigation, Wuhan University of Technology, Wuhan 430063, China; pei.dashuai@whut.edu.cn (D.P.); chenmz@whut.edu.cn (M.C.)
[2] School of Computer Science and Electronic Engineering (CSEE), University of Essex, Colchester CO4 3SQ, UK
[3] School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; shengkai@whut.edu.cn
* Correspondence: j.he@essex.ac.uk (J.H.); kzliu@whut.edu.cn (K.L.)

Abstract: Maritime transport plays a critical role in global logistics. Compared to road transport, the pace of research and development is much slower for maritime transport. It faces many major challenges, such as busy ports, long journeys, significant accidents, and greenhouse gas emissions. The problems have been exacerbated by recent regional conflicts and increasing international shipping demands. Maritime Autonomous Surface Ships (MASSs) are widely regarded as a promising solution to addressing maritime transport problems with improved safety and efficiency. With advanced sensing and path-planning technologies, MASSs can autonomously understand environments and navigate without human intervention. However, the complex traffic and water conditions and the corner cases are large barriers in the way of MASSs being practically deployed. In this paper, to address the above issues, we investigated the application of Large Language Models (LLMs), which have demonstrated strong generalization abilities. Given the substantial computational demands of LLMs, we propose a framework for LLM-assisted navigation in connected MASSs. In this framework, LLMs are deployed onshore or in remote clouds, to facilitate navigation and provide guidance services for MASSs. Additionally, certain large oceangoing vessels can deploy LLMs locally, to obtain real-time navigation recommendations. To the best of our knowledge, this is the first attempt to apply LLMs to assist with ship navigation. Specifically, MASSs transmit assistance requests to LLMs, which then process these requests and return assistance guidance. A crucial aspect, which has not been investigated in the literature, of this safety-critical LLM-assisted guidance system is the knowledge and safety performance of the LLMs, in regard to ship handling, navigation rules, and skills. To assess LLMs' knowledge of navigation rules and their qualifications for navigation assistance systems, we designed and conducted navigation theory tests for LLMs, which consisted of more than 1500 multiple-choice questions. These questions were similar to the official theory exams that are used to award the Officer Of the Watch (OOW) certificate based on the Standards of Training, Certification, and Watchkeeping (STCW) for Seafarers. A wide range of LLMs were tested, which included commercial ones from OpenAI and Baidu and an open-source one called ChatGLM, from Tsinghua. Our experimental results indicated that among all the tested LLMs, only GPT-4o passed the tests, with an accuracy of 86%. This suggests that, while the current LLMs possess significant potential in regard to navigation and guidance systems for connected MASSs, further improvements are needed.

Keywords: maritime autonomous surface ships; large language model; ship-handling theory test; mobile edge computing; mobile cloud computing

MSC: 68Txx; 68T40; 68T20; 68T50

Citation: Pei, D.; He, J.; Liu, K.; Chen, M.; Zhang, S. Application of Large Language Models and Assessment of Their Ship-Handling Theory Knowledge and Skills for Connected Maritime Autonomous Surface Ships. *Mathematics* **2024**, *12*, 2381. https://doi.org/10.3390/math12152381

Academic Editor: Jonathan Blackledge

Received: 1 July 2024
Revised: 29 July 2024
Accepted: 30 July 2024
Published: 31 July 2024

1. Introduction

Maritime transport plays a critical role in the global economy and in the movement of goods. It is the backbone of international trade, enabling the efficient and cost-effective transportation of vast quantities of products across the world's oceans [1]. The United Nations Conference on Trade and Development expects maritime trade volume to grow by more than 3% during the 2024–2028 period [2]. However, maritime transport faces several significant challenges, including busy ports, long journeys, major accidents, and greenhouse gas emissions. Additionally, current regional conflicts and escalating international tensions have intensified international shipping demands, thereby exacerbating the issues faced by maritime transport. For example, oil shipments reached record distances in 2022, driven by the devastation of the war in Ukraine [3]. Similarly, grain shipments traveled farther in 2023 than in any previous year, as grain importers were forced to seek alternative exporters, such as the United States and Brazil, necessitating long-distance transportation.

The widespread application of intelligent technologies in the shipping industry, particularly the development of Maritime Autonomous Surface Ships (MASSs), offers promising solutions to these challenges. MASSs can enhance the efficiency and safety of maritime transport by optimizing port operations, reducing travel times, minimizing human error, and lowering emissions [4]. In 2021, the global autonomous ships market had a revenue share of over 89 million USD, and it is projected to grow at a compound annual growth rate of 6.81% through 2031 [5]. As the shipping industry continues to evolve, the integration of autonomous technology stands to address the urgent and complex problems of maritime transport, paving the way for a more resilient and sustainable future [6].

MASSs integrate a variety of advanced technologies, to achieve autonomous navigation and operation. Their core technologies include navigation systems (such as GPS, inertial navigation systems, and electronic chart display and information systems), sensing and recognition technologies (such as radar, LiDAR, and computer vision), communication systems (such as satellite communication and radio communication), data processing and artificial intelligence (such as edge-computing and machine-learning algorithms), and autonomous control systems (such as rudder and propulsion system control and automatic docking systems) [7,8]. These technologies work in concert, to enable MASSs to autonomously navigate under various sea conditions, perform complex navigation tasks, and enhance efficiency, safety, and environmental performance. The extensive application of MASSs in the shipping industry is poised to reduce the involvement of human operators, thereby significantly mitigating the likelihood of human-related maritime accidents [9].

Autonomous navigation technology is the critical core that determines whether MASSs can safely navigate without human intervention. This technology involves the use of sensors, artificial-intelligence algorithms, and automatic control systems to enable a ship to autonomously perceive its environment, plan routes, and execute navigation tasks. Currently, the predominance of deep learning-based autonomous-navigation algorithms is observable [8]. For example, Wright et al. [10] explored the use of deep learning to integrate multiple sensor modalities into autonomous-navigation algorithms for ships, allowing for decision making without human supervision. Han et al. [11] developed deep-learning algorithms for multiple target detection and tracking using sensor fusion to enhance autonomous navigation and collision avoidance for the Unmanned Surface Vehicle (USV) Aragon. However, complex traffic and water conditions, as well as various extreme situations and corner cases, pose significant challenges to deep learning-based autonomous navigation technology. This challenge is known in the deep learning field as the "long tail". The "long tail" refers to the vast number of rare or outlier events that occur infrequently but can significantly impact the performance of a model. These rare scenarios are difficult for the model to handle effectively because the training data often does not adequately cover such infrequent events, leading to issues with generalization and reliability. When these long-tail cases occur, the autonomous navigation system may struggle to respond correctly, potentially resulting in incidents, such as collisions or groundings, and causing significant financial losses.

Recently, the application of LLMs in autonomous driving has provided inspiration for addressing the aforementioned challenges. These models understand the driving environment in a human-like manner and utilize their reasoning, interpretation, and memorization capabilities to effectively solve long-tail issues. For example, Sha et al. [12] employed LLMs as decision-making components, to enhance autonomous driving systems, particularly in complex scenarios requiring human common-sense understanding. Fu et al. [13] investigated the use of LLMs to understand the driving environment in a human-like manner, emphasizing their ability to solve long-tail issues through reasoning, interpretation, and memorization. Their extensive experiments demonstrated that LLMs exhibit impressive capabilities in handling long-tailed cases, providing valuable insights for developing human-like autonomous driving systems.

Given the notable applications of LLMs in the field of autonomous driving, contemplation of the application of these models within the domain of autonomous ship navigation was inevitable. However, there were two important challenges. Firstly, LLMs necessitate an increased number of parameters, to encapsulate complex patterns within training data, thereby enhancing performance. This requirement results in considerable computational and memory demands. Secondly, the prominence of safety in autonomous ship navigation systems cannot be overstated, with safety expectations surpassing those of human navigation markedly. Despite OOWs being mandated to clear theoretical and practical examinations before certification, LLMs had yet to be subjected to stringent evaluations regarding their automatic navigation capabilities.

To overcome the aforementioned challenges, we investigated a novel method that incorporates LLMs into remote cloud or shore-based systems, to enhance autonomous ship navigation. By employing this strategy, connected MASSs send assistance requests to LLMs. Located onshore or within a remote cloud, the LLMs process these requests and subsequently generate guidance for the MASSs, as illustrated in Figure 1. We aimed to evaluate the theoretical knowledge of the LLMs, similar to the assessment of human OOW. Although practical ship navigation and watchkeeping skills through LLMs are indispensable, we contend that a theoretical examination is equally significant, considering its relative simplicity and controllability. Despite the notable achievements of LLMs across various fields, such as law, education, and economics, the number of reports detailing its performance in ship-handling theory tests is particularly limited.

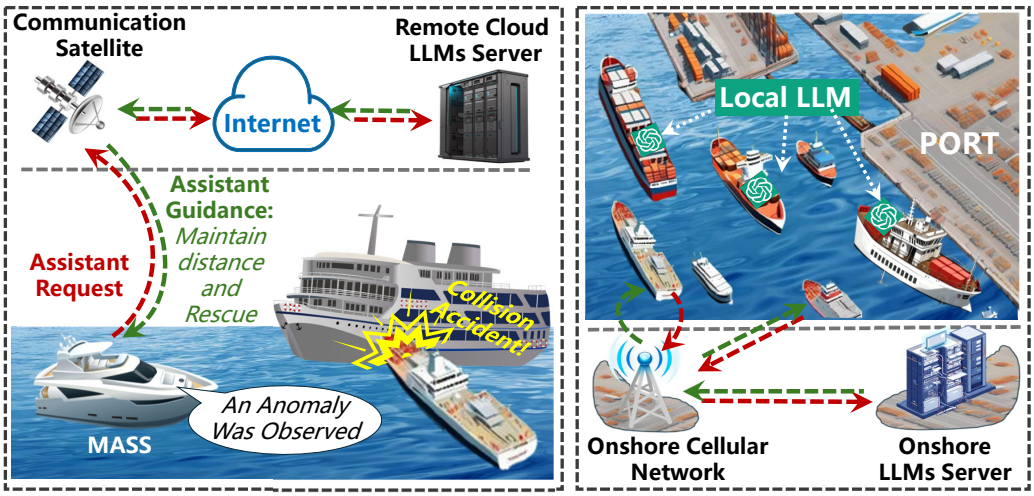

Figure 1. System framework of LLM-assisted navigation system for MASSs. MASSs may receive navigation guidance from LLMs deployed in remote clouds or onshore, and some large vessels can also obtain navigation recommendations from LLMs deployed on board.

In this study, we designed and conducted ship-handling theory tests for 14 LLMs, including GPT-3.5-turbo [14], GPT-4 [15], GPT-4o, ERNIE-4.0-8k [16] and Qwen-turbo [17] et al. Therefore, we developed and implemented ship-handling theory tests comprising over 1500

questions for several LLMs. These questions were analogous to those in the China official theory exam required for seafarers to obtain the Standards of STCW OOW certificate. We evaluated the performance of these LLMs based on accuracy, cost, and processing latency derived from experimental observations. The experimental results indicated that among all the LLMs only GPT-4o achieved a test accuracy rate close to 86%, while all the other models failed the test. In conclusion, although several LLMs showed significant potential for autonomous ship navigation, their performance requires further enhancement to meet the stringent demands of safe navigation. Additional training and fine-tuning are likely necessary. The source code and datasets are available at https://github.com/PeiDashuai/LLMs_Nav, accessed on 1 July 2024.

2. Existing Work

2.1. Autonomous Ship Navigation System

Autonomous Ship Navigation System refers to an integrated framework of sensors, control algorithms, and navigation technologies enabling ships to operate and navigate safely and efficiently without human intervention. Villa et al. [18] investigated the design, modeling, and implementation challenges of a Guidance, Navigation, and Control (GNC) architecture for an autonomous ship navigation system in harbor conditions. They developed a mathematical model validated with field-test data and implemented a line-of-sight guidance system using LiDAR for obstacle avoidance, with their GNC architecture tested in both simulation and field scenarios. Han et al. [11] developed algorithms for multiple target detection and tracking using sensor fusion for the autonomous navigation and collision avoidance system of the USV Aragon. By integrating radar, LiDAR, and cameras, and applying automatic ship-detection algorithms, they achieved persistent and reliable target tracking and designed collision avoidance maneuvers in compliance with the International Regulations for Preventing Collisions at Sea (COLREGs) [19], with validation through field experiments. Kufoalor et al. [20] conducted sea trials for an Autonomous Surface Vehicle (ASV) equipped with a Model Predictive Control (MPC)-based collision avoidance system in the North Sea, to verify compliance with the COLREGs. The trials demonstrated that the MPC approach effectively finds safe solutions in challenging scenarios, often meeting the expectations of experienced mariners, indicating the higher-than-expected technical maturity of autonomous vessels. Kim et al. [21] developed autonomous navigation capabilities for small cruise boats by converting a cruise boat into an ASV with various sensors and actuators. They designed and implemented navigation, object-detection, path-planning, and control algorithms, and they validated the system's performance through field experiments in a canal and surrounding waters.

2.2. LLMs-Based Autonomous Driving

Many studies have evaluated the potential and challenges of LLMs in autonomous driving. Cui et al. [22] proposed a novel framework for LLMs to enhance decision making in autonomous vehicles by integrating their language and reasoning capabilities. Their research demonstrated that LLMs can influence driving behavior through real-time personalized tasks and ongoing verbal feedback, improving safety and effectiveness in autonomous driving. Sha et al. [12] employed LLMs as decision-making components, to enhance autonomous driving systems, particularly in complex scenarios requiring human common-sense understanding. Their approach integrated LLM decisions with low-level controllers, demonstrating superior performance and improved handling of complex driving behaviors through experiments, highlighting the potential of LLMs for advancing autonomous driving capabilities. Duan et al. [23] proposed a hybrid end-to-end learning framework for autonomous driving by integrating LLMs with visual and LiDAR sensory input, aiming to correct mistakes and handle complex scenarios. Their methodology achieved a driving score of 49.21% and a route completion rate of 91.34% in offline evaluations, comparable to state-of-the-art driving models. Huang et al. [24] explored the application of LLM-based voice assistants, such as ChatGPT-4, to mitigate passive driving fatigue and

enhanced driving performance and safety. Their empirical study, using the voice assistant "Driver Mate", revealed that low-complexity, high-frequency conversations improve driver alertness and acceptance, while low-complexity, low-frequency interactions enhance driving performance.

2.3. Evaluating LLMs with Multiple-Choice Questions

Numerous studies employ multiple-choice questions (MCQs) to evaluate the capabilities of LLMs. It has been proved that the use of MCQs is one of the effective means by which to evaluate the capability of LLMs [25,26]. Zhang et al. [27] introduced SafetyBench, a comprehensive benchmark designed to evaluate the safety of LLMs, using 11,435 diverse multiple-choice questions across seven safety concern categories. Their extensive tests on 25 popular Chinese and English LLMs revealed significant performance advantages for GPT-4, highlighting the need for further safety improvements in current models. Huang et al. [28] presented C-Eval, the first comprehensive Chinese evaluation suite designed to assess the advanced knowledge and reasoning abilities of LLMs, using multiple-choice questions across four difficulty levels and 52 diverse disciplines. Their comprehensive evaluation revealed that only GPT-4 achieved an average accuracy above 60%, highlighting the need for further improvement in current LLMs. Wu et al. [29] investigated the medical knowledge capabilities of multiple LLMs by comparing their performance on nephrology MCQs from the Nephrology Self-Assessment Program. The study revealed significant performance differences, with open-source LLMs scoring between 17.1% and 30.6% correct answers, while the proprietary models GPT-4 and Claude 2 achieved 73.3% and 54.4%, respectively, highlighting notable gaps in zero-shot reasoning ability among LLMs. Xu et al. [30] evaluated the performance of two state-of-the-art LLMs, ChatGPT and Microsoft Bing AI Chat, on a dataset of 200 high school chemistry MCQs, to assess their educational potential and challenges. The study found that both LLMs struggled with application- and high-application-level questions, performing worse than Vietnamese students, indicating a need for further development to improve their capabilities.

3. System Framework of LLM-Assisted Navigation for MASSs

For this study, we innovatively constructed a framework that uniquely integrates the power of LLMs, to enhance the performance of MASSs navigation systems. To the best of our knowledge, this is the first attempt to apply LLMs to assist in ship navigation, pioneering this field. This framework allows MASSs to intelligently interact with LLMs located in remote clouds or onshore bases via satellite links or advanced 5G mobile networks. It also supports MASSs in consulting LLMs deployed on the ship directly for immediate navigation strategies. For example, MASSs navigating crowded inland waterways or port areas can communicate with land-based LLMs through nearby cellular network access points. In contrast, ships traversing vast open seas can seek complex navigation decisions via satellite communications with LLM servers deployed in remote clouds. Furthermore, large freight or luxury cruise ships can deploy LLMs directly on board, to achieve real-time, efficient navigation recommendations. The maneuvering inertia of ships is substantial, and collision avoidance actions typically begin when two ships are several kilometers or even tens of kilometers apart. This provides sufficient time for communication and course adjustments within the radar monitoring range; thus, autonomous ship navigation does not require extremely high LLM response speeds. Standard cellular or satellite communication speeds are entirely sufficient to support remote LLM response rates.

To meet the needs of ship navigation, we will meticulously customize and fine-tune existing top-tier LLM models, such as GPT-4o, Meta-Llama-3-70B, and Qwen-turbo. These models were originally built upon vast and diverse datasets, with a deep knowledge base and excellent generalization capabilities. After specialized tuning, these LLMs deeply understand maritime domain knowledge and can be precisely applied in practice, providing reliable and flexible auxiliary decision-making services for MASSs.

The integration of LLM technology in MASSs systems heralds a new phase for autonomous ship navigation applications, encompassing but not limited to real-time exchange of vessel status, environmental perception, and navigation intent information within maritime areas through Automatic Identification Systems (AISs), Very High Frequency (VHF), and cellular network technologies. This promotes collaborative environmental perception, multi-ship cooperative navigation, automatic fleet formation navigation, and other advanced functions. This advancement not only significantly enhances navigation efficiency and safety but also lays a solid foundation for future intelligent and networked maritime traffic management.

Figure 1 illustrates a vivid example of how an LLM can be utilized to assist navigation. In the scenario, a sudden collision accident occurs between a passenger ship and another vessel in a specific water area, causing the passenger ship to capsize. At this moment, a MASS equipped with an LLM-assisted navigation system is passing by. Its advanced sensing system immediately detects the abnormal situation and automatically initiates a navigation assistance request to the LLM deployed on a remote cloud server. Upon receiving the signal, the LLM server rapidly analyzes the situation and guides the MASS to take action, such that it maintains a safe navigation distance while urgently deploying lifeboats and related rescue equipment, to quickly participate in the rescue of people in the water.

4. Research Methodology

4.1. OOW Theory Examination

The OOW is responsible for watchkeeping, navigation, communication, log-keeping, and emergency responses, all of which are critical for ensuring safe navigation. This role is assigned to a sufficiently qualified deck officer and involves various duties, including ensuring the ship operates in accordance with regulations and company procedures, maintaining the ship's equipment and machinery, and ensuring the crew effectively carries out their duties. To apply for the OOW role, candidates must meet specific eligibility criteria and possess the required certifications. The Standards of Training, Certification, and Watchkeeping (STCW) Training Convention for seafarers [31] outlines general requirements and certifications by rank. For OOW, the Convention specifies requirements concerning age, seagoing service, bridge watchkeeping, radio duties, and education and training.

Typically, after completing academic studies and gaining the necessary seafaring experience, a crew member must pass a written and practical assessment, to obtain an OOW license. The specific assessments may vary by country and the type of license sought. In China, the OOW examination encompasses core subjects, such as maritime English, ship steering and collision avoidance, navigation, ship structure and cargo handling, and ship management. The exam caters to various tonnage levels (e.g., 500 gross tons and above, 3000 gross tons and above, less than 500 gross tons) and navigational areas (unlimited and coastal) for positions like captain, chief mate, second mate, and third mate. The total score and passing score vary depending on the subject and the ship's tonnage, ensuring that OOWs possess the professional knowledge and skills necessary to fulfill their duties. Each subject is scored out of 100 points and primarily consists of approximately 160 MCQs. The passing score is 80 points for ship steering and collision avoidance, while it is 70 points for the other subjects. This paper primarily considered the subjects directly related to ship handling.

4.2. Test Datasets

The competency tables outlined in the STCW Convention detail the content of training programs for seafarers, the criteria for evaluating competencies, and the standards of competence that students must demonstrate. Relevant authorities have developed test questions based on the STCW Convention and practical maritime experience. Due to the unavailability of official questions, we collected test questions from Chinese public websites. The questions were meticulously selected and processed, to ensure their relevance and quality. After removing duplicates, we compiled 706 Chinese and 814 English MCQs. Each MCQ included multiple answer options, with only one correct answer. The Chinese MCQs and English MCQs

we collected were not different language versions of the same questions. They contained different questions and were sourced from different websites. In Figure 2, we present two examples of the MCQs. However, we did not find any test questions that included traffic scenario videos or images, which are crucial components of theoretical tests. Future iterations will include multimedia questions, to capture a broader range of navigational scenarios. All the data we collected can be found at our open-source project address.

English MCQ Example	Chinese MCQ Example
Question 1:	**Question 1:**
When approaching a traffic separation scheme, a vessel shall:	*在追越过程中，被追越船的协助避让行动为()。*
Options:	**Options:**
A. do so at right angles to the general direction of traffic flow	*A. 只要航道情况和周围环境允许，就应同意追越船追越*
B. seek permission to do so from all other vessel in the vicinity	*B. 尽可能让出部分航道，适当减速，减少两船并行时间，使追越船迅速通过*
C. do so only in a case of an emergency or to engage in fishing within the zone	*C. 前方发现情况及时通知追越船的注意*
D. do so at as small an angle as possible as nearly as practical	*D. 以上都是*
Correct Answer: D	**Correct Answer: D**

Figure 2. Two examples of MCQs for MASSs asking LLMs.

4.3. Prompt Design

For this section, we designed the prompts used in our experiments, based on prompt engineering.

4.3.1. Instructing the LLMs to Role-Play and Demonstrate Specific Skills

Instead of having the model directly answer our MCQs, we instructed it to assume the role of an experienced OOW, to respond to our inquiries. Role-playing is considered effective in prompt engineering, as it helps set the overall behavior of the assistant. This enables the model to understand user requirements and provide appropriate responses based on those needs. On the other hand, clearly specifying the skills that the model should possess can significantly enhance its performance. Precisely describing the required skills not only guides the LLMs to generate more relevant and high-quality responses but also improves the accuracy and effectiveness of tasks. We demonstrated the effectiveness of this approach in improving accuracy through continuous iterative optimization of prompts.

4.3.2. Providing Example MCQs and Answers

Providing examples to LLMs can be considered a form of "few-shot learning", enabling the models to utilize these demonstrations for analogical reasoning when generating responses, thereby improving accuracy. Including examples of questions and answers in the prompt also helps establish the model's expected behavior, allowing it to understand the question format and response style, thus enhancing accuracy and consistency. Furthermore, these examples reduce ambiguity in the model's interpretation, making it more precise in identifying patterns and the logic of correct answers. Overall, this approach ensures that LLMs not only predict possible answers based on the questions themselves but also understand the structure and logic through provided examples, leading to more accurate responses. This method is crucial for improving the accuracy and reliability of LLMs in handling MCQs.

4.3.3. Designing Structured Prompts

Designing structured prompts is crucial for querying LLMs, as it ensures clarity, consistency, focus, and improved accuracy in the responses. In our design, we structured the prompts to include five parts: role, skills, action, output format and constraints, and example. For the role, we instructed the LLMs to role-play as an experienced OOW. In the skills section, we required the LLMs to excel in ship handling, to be well-versed in the STCW Convention, to have extensive experience in theoretical exams, and to be proficient in selecting the most accurate option from multiple candidates based on the question's intent. For the action part, we asked the LLMs to answer our MCQs.

Regarding output format and constraints, we instructed the LLMs to output only the option letter of the MCQs. Finally, we provided an MCQ and answer as an example in the

prompts. We have provided examples of prompts that we designed in both Chinese and English, as shown in Figure 3. They convey the same meaning, but are written in different languages, to facilitate testing different LLMs.

English Prompt Example	Chinese Prompt Example
# Role You are an experienced Officer of the Watch (OOW). # Skills • Extensive experience in ship navigation and is familiar with the International Convention on Standards of Training, Certification, and Watchkeeping for Seafarers Convention • Proficient in taking theoretical exams • Skilled at selecting the most accurate option from multiple choices based on the question's meaning # Action Answer multiple-choice questions about ship handling theory and experience. # Output Format and Constraints These questions have multiple candidate answers, but only one answer is correct. Your response should only include the initial letter of the chosen option, such as A or B. Do not add any additional content or punctuation marks; only output the initial letter of the chosen option. *# Example* *Question:* *You are making way in restricted visibility when you hear the sound of a fog signal forward of your beam. You are required to reduce speed to:* *Options:* *A. a moderate speed commensurate with conditions* *B. the minimum where your vessel can be kept on course* *C. half speed if proceeding at a higher speed* *D. a safe speed in relation stopping distance* *<Assistant answer> B*	# Role 你是一个经验丰富的船舶值班驾驶员 # Skills • 拥有丰富的船舶驾驶经验，并熟知海员培训、发证和值班标准国际公约公约 • 具有丰富的参加理论考试的经验 • 擅长根据题目的含义从多个候选答案中选出最正确的一个选项 # Action 我需要你去回答一些船舶操纵理论和经验方面的选择题 # Output Format and Constrains 这些题目有多个候选答案，但是仅有一个答案是正确的，你的回答只需要包含候选答案的首字母，例如A或B，不要增加任何额外的内容或标点符号，仅需要输出候选答案的首字母。 *# Example* *Question:* *在能见度受限的情况下航行时，如果你听到船首前方有雾笛声，你需要减速至?* *Options:* *A、与当前情况相适应的中等速度* *B、能让船舶保持航向的最低速度* *C、如果当前速度较高，则减半* *D、与安全停车距离相对应的安全速度* *<Assistant answer> B*

Figure 3. Two examples of prompts.

4.4. LLMs Used in Theory Test

There are several powerful LLMs from leading companies, such as Alibaba, Google, Baidu, and OpenAI, et al. Several LLMs were chosen for the ship operation theory test, which were among the best-performing ones. The fourteen chosen models have different capabilities and price points. ERNIE-4.0-8k and Qwen-turbo are two leading LLMs developed by Baidu and Alibaba, respectively. GPT-3.5-turbo, GPT-4, and GPT-4o are LLMs developed by the well-known OpenAI. GLM-3-turbo, GLM-4, and GLM-4-Air [15,32] are jointly open-sourced by Zhipu AI and the Tsinghua University. The Qianfan-Chinese-Llama-2 series models were fine-tuned by Baidu's Qianfan team based on the open-source Llama 2 model from Meta AI [33,34], optimizing its support for Chinese. Gemma-7B [35] is an open-source LLM developed by Google, with 7 billion parameters. The Meta-Llama-3 [36] series models were developed by Meta AI. Table 1 shows information about the employed LLMs.

Table 1. Information about the LLMs used in the experiment.

Model Name	Prices ($)/1 k Tokens		Model Size	Version	Creators
	Input	Output			
Qwen-turbo	0.0145	0.0435	undisclosed	\	Alibaba Cloud
ERNIE-4.0-8k	0.871	0.871	undisclosed	0329	Baidu
GPT-3.5-turbo	0.0005	0.0015	undisclosed	\	Open AI
CPT-4	0.03	0.06	undisclosed	\	
GPT-4o	0.005	0.015	undisclosed	\	
GLM-3-turbo	Open Source		undisclosed	\	Tsinghua and Zhipu
GLM-4-Air			undisclosed	\	
GLM-4			9B	0520	
Qianfan-Chinese-Llama-2-7B	0.029	0.029	7B	\	Qianfan
Qianfan-Chinese-Llama-2-13B	0.044	0.044	13B	v1	
Qianfan-Chinese-Llama-2-70B	0.254	0.254	70B	\	

Table 1. *Cont.*

Model Name	Prices ($)/1 k Tokens		Model Size	Version	Creators
	Input	Output			
Meta-Llama-3-8B	Open Source		8B	Instruct	Meta AI
Meta-Llama-3-70B			70B	Instruct	
Gemma-7B-it			7B	Instruct	Google

5. Experiments

5.1. Experiments Settings

Implementation Details. In our experiments, Meta-3-Llama-3-8B and Meta-3-Llama-3-70B were deployed on a server equipped with an L20 (48GB) GPU, 20 vCPUs of Intel(R) Xeon(R) Platinum 8457C, and 100 GB of RAM for inference. The testing tasks for the remaining models, accessed via API, were conducted on a laptop equipped with an i9-13950HX CPU, Nvidia GeForce RTX 4060 GPU, and 16 GB of RAM. The parameter settings of all the tested models are shown in Table 2. The temperature parameter determined whether the output was more random or more predictable. A lower temperature resulted in a higher probability, leading to a more predictable output. The top_p parameter affected the diversity of the output text generated by the LLMs; the larger the value, the greater the diversity of the generated text. For all Chinese MCQs, we tested using both Chinese prompts and English prompts. For all English MCQs, we tested using only English prompts. The max_output_tokens parameter specified the maximum number of tokens that the model could output. The parameters that we do not mention in the table were the default settings for the LLM creators.

Table 2. The parameter settings of the tested LLMs.

Model Name	Temperature	Top_p	# Max Output Tokens
Qwen-turbo	0.5	0.7	100
ERNIE-4.0-8k	0.5	0.7	100
GPT-3.5-turbo	0	1	100
CPT-4	0	1	100
GPT-4o	0	1	100
GLM-3-turbo	0.5	0.7	100
GLM-4-Air	0.5	0.7	100
GLM-4	0.5	0.7	100
Qianfan-Chinese-Llama-2-7B	0.5	0.7	100
Qianfan-Chinese-Llama-2-13B	0.5	0.7	100
Qianfan-Chinese-Llama-2-70B	0.5	0.7	100
Meta-Llama-3-8B	0.5	0.7	100
Meta-Llama-3-70B	0.5	0.7	100
Gemma-7B-it	0.5	0.7	100

Evaluation Protocols. Considering that we required the tested models to output only the correct answer option for the MCQs, we used ***Accuracy*** as the sole evaluation metric. ***Accuracy*** was defined as the number of correctly answered MAQs by the tested LLM divided by the total number of tested MCQs.

5.2. Experimental Results and Discussions

Table 3 presents the experimental results using Chinese prompts to query multiple LLMs with Chinese Ship Handling and Navigation Theory MCQs. A total of 706 MCQs were employed, with GPT-4o achieving the best performance, attaining an accuracy of 60.76% and the lowest time consumption. It is important to note that the time reported here does not refer to the absolute inference time of the LLM, as network latency from API access can influence the time statistics. In our prompt, we instructed the tested models to output only the letter corresponding to the correct option, without any additional explanations or symbols, akin to actual human theoretical tests. However, some models still generated explanations beyond the letter, resulting in a significant increase in output tokens. This indicates that certain LLMs need to improve their understanding of prompts. Furthermore, the variability in accuracy and time consumption among the models highlights differences in their architectures and training methodologies. For instance, models like GPT-3.5-turbo and GPT-4o not only provided high accuracy but also demonstrated efficient processing times, suggesting their robustness in understanding and responding to Chinese prompts. On the other hand, while models such as Meta-Llama-3-70B and Qianfan-Chinese-Llama-2-70B exhibited competitive accuracy, their higher time consumption could be attributed to more complex processing requirements or network-related delays. This suggests a trade-off between accuracy and computational efficiency that should be considered based on specific application needs. Additionally, the significant differences in the number of output tokens across models suggests variations in their adherence to prompt instructions. For example, Qianfan-Chinese-Llama-2-70B and Meta-Llama-3-70B generated a large number of output tokens, indicating a propensity to provide additional explanations beyond the required answer letter. This behavior could be detrimental in scenarios where concise responses are crucial. Moreover, the models developed by Chinese companies, such as Qwen-turbo and ERNIE-4.0-8k, demonstrated promising results. Their performance was comparable to GPT-4o, which achieved the best results.

Table 3. Using Chinese prompts to query multiple LLMs with Chinese ship-handling theory MCQs.

Model	# Ques.	# Corr.	Acc.	Time (s)	# Total Tokens	
					# Input Tokens	# Output Tokens
Qwen-turbo	706	423	59.92%	636.6	247,105	739
ERNIE-4.0-8k	706	412	58.36%	2921.65	214,026	6925
GPT-3.5-turbo	706	316	44.76%	415.86	415,338	711
CPT-4	706	389	55.10%	529.42	415,338	733
GPT-4o	706	429	60.76%	339.01	292,675	706
GLM-3-turbo	706	340	48.16%	940.27	239,117	2134
GLM-4-Air	706	352	49.86%	995.76	230,267	2122
GLM-4	706	371	52.55%	1110.49	230,267	2133
Qianfan-Chinese-Llama-2-7B	706	273	38.67%	2520.14	230,108	2,010
Qianfan-Chinese-Llama-2-13B	706	317	44.90%	6994.99	230,108	111,678
Qianfan-Chinese-Llama-2-70B	706	398	56.37%	5510.65	230,108	93,757
Meta-Llama-3-8B	706	283	40.08%	2235.81	230,108	709
Meta-Llama-3-70B	706	313	44.33%	9015.19	230,108	116,630
Gemma-7B-it	706	282	39.94%	2779.97	230,108	30,907

In Table 4, we present the results of using English prompts to test the same set of Chinese MCQs, providing an evaluation of the impact of language on the accuracy of large language models.

Table 4. Using English prompts to query multiple LLMs with Chinese ship-handling theory MCQs.

Model	# Ques.	# Corr.	Acc.	Time (s)	# Total Tokens	
					# Input Tokens	# Output Tokens
Qwen-turbo	706	333	47.17%	631.79	232,279	734
ERNIE-4.0-8k	706	398	56.37%	3129.01	218,968	8784
GPT-3.5-turbo	706	315	44.62%	442.89	260,724	713
CPT-4	706	380	53.82%	498.75	260,724	799
GPT-4o	706	435	61.61%	329.97	237,607	706
GLM-3-turbo	706	336	47.59%	1519.96	232,763	2375
GLM-4-Air	706	356	50.42%	1387.82	223,913	2116
GLM-4	706	367	51.98%	1391.58	223,913	2365
Qianfan-Chinese-Llama-2-7B	706	312	44.19%	1980.27	224,460	4461
Qianfan-Chinese-Llama-2-13B	706	337	47.73%	4285.56	224,460	113,071
Qianfan-Chinese-Llama-2-70B	706	396	56.09%	5289.31	224,460	93,379
Meta-Llama-3-8B	706	289	40.93%	1843.25	224,460	706
Meta-Llama-3-70B	706	359	50.85%	8680.55	224,460	116,637
Gemma-7B-it	706	294	41.64%	2,803.91	224,460	31,128

Figure 4 compares two datasets, showing that the nine models developed by Meta, OpenAI, and Google exhibited a slight advantage when using English prompts over Chinese prompts. Conversely, the five LLMs developed by Chinese companies, such as Qwen-turbo, ERNIE-4.0-8k, and ChatGLM, demonstrated better performance with Chinese prompts. This difference may be attributed to variations in the corpora used by different companies in training their base models. Additionally, we observed that the number of parameters in the tested models exhibited a linear relationship with accuracy. As the number of model parameters increased, accuracy improved. For instance, GPT-4o, with its higher parameter count, consistently outperformed other models in both scenarios. The results underscore the significant influence of language on model performance. Models like GPT-4o and GPT-3.5-turbo demonstrated high adaptability, maintaining robust performance across both English and Chinese prompts, which is essential for applications requiring multilingual support. However, certain models exhibited a marked preference for prompts in their native language. For example, Qwen-turbo achieved accuracy of 59.92% with Chinese prompts but dropped to 47.17% when using English prompts. This suggests that these models may have been predominantly trained on Chinese corpora, optimizing their performance for Chinese prompts. Time consumption data reveal that models like GPT-4o not only provided high accuracy but also demonstrated efficient processing times, particularly with English prompts. The significant differences in the number of output tokens generated by the models suggest variations in their adherence to prompt instructions. While some models, such as Qianfan-Chinese-Llama-2-70B, generated excessive tokens when using English prompts, indicating the inclusion of unnecessary explanations, others like GPT-4o adhered strictly to the prompt requirements, thereby enhancing their overall efficiency.

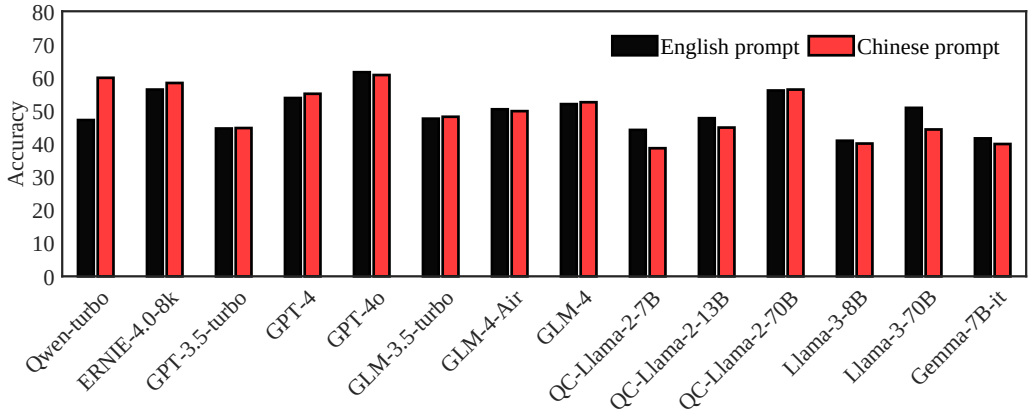

Figure 4. Results of testing the same set of Chinese ship-handling theory MCQs using Chinese and English prompts, respectively.

In Table 5, we present the results of testing 814 English MCQs using English prompts. These data allowed us to evaluate the performance of LLMs across different languages and question types. All models showed significant improvements in accuracy when queried with English prompts. Among them, GPT-4o achieved an outstanding accuracy rate of over 85%, making it the only LLM likely to pass the test for ship handling, watchkeeping, and navigation theory. Regarding cross-language adaptability, models developed by Meta, OpenAI, and Google, such as GPT-4o and GPT-3.5-turbo, exhibited high adaptability across languages. For instance, GPT-4o maintained high performance across all prompt types, achieving 86.00% accuracy with English prompts and slightly lower, yet still impressive, accuracy with Chinese prompts. This adaptability is essential for applications requiring multilingual support. Chinese LLMs, such as Qwen-turbo and ERNIE-4.0-8k, demonstrated a strong preference for their native language prompts. Qwen-turbo exhibited a drop in accuracy from 59.92% with Chinese prompts to 55.41% with English prompts in the English MCQ scenario. This suggests that these models may be more optimized for their native language, due to the training corpus. There is a clear linear relationship between the number of model parameters and accuracy. As observed, models with higher parameter counts, such as GPT-4o and Meta-Llama-3-70B, consistently outperformed others in both test scenarios. This indicates that larger models tend to better handle complexity in multiple languages. The number of output tokens varied significantly across models and prompt languages. Models such as Qianfan-Chinese-Llama-2-70B and Meta-Llama-3-70B tended to generate excessive tokens when using English prompts, suggesting a need for better prompt adherence. Conversely, GPT-4o adhered closely to prompt instructions, boosting its overall efficiency. While our results show that most models improved their performance with English prompts, this cannot conclusively demonstrate that LLMs perform better on English MCQs than on Chinese MCQs. The datasets contain different content, which likely influences the performance of the models.

Table 5. Using English prompts to query multiple LLMs with English ship-handling theory MCQs.

Model	# Ques.	# Corr.	Acc.	Time (s)	# Total Tokens	
					# Input Tokens	# Output Tokens
Qwen-turbo	814	451	55.41%	767.11	244,152	845
ERNIE-4.0-8k	814	549	67.44%	3788.96	237,234	6011
GPT-3.5-turbo	814	467	57.37%	352.25	243,151	832
CPT-4	814	613	75.31%	547.76	243,151	814
GPT-4o	814	700	86.00%	366.41	243,703	814
GLM-3-turbo	814	476	58.48%	1533.61	252,842	2481

Table 5. *Cont.*

Model	# Ques.	# Corr.	Acc.	Time (s)	# Total Tokens	
					# Input Tokens	# Output Tokens
GLM-4-Air	814	531	65.23%	1250.09	239,865	2442
GLM-4	814	553	67.94%	1523.23	239,878	2443
Qianfan-Chinese-Llama-2-7B	814	341	41.89%	2364.15	242,846	2709
Qianfan-Chinese-Llama-2-13B	814	393	48.28%	4775.56	242,846	123,678
Qianfan-Chinese-Llama-2-70B	814	486	59.71%	5511.66	242,846	93,846
Meta-Llama-3-8B	814	407	50.00%	2475.05	242,846	814
Meta-Llama-3-70B	814	542	66.58%	8699.69	242,846	113,623
Gemma-7B-it	814	361	44.35%	3588.27	242,846	32,208

6. Conclusions

In this study, we explored the use of LLMs to support navigation and guidance in MASSs. Given the significant computational requirements of LLMs, we proposed a framework for LLM-assisted navigation for connected MASSs, wherein LLMs are deployed onshore or in remote clouds to facilitate navigation and provide guidance services. Additionally, certain large oceangoing vessels can deploy LLMs locally, to obtain real-time navigation recommendations. MASSs units transmit assistance requests to LLMs, which process these requests and return guidance.

To assess the LLMs' knowledge and suitability for the navigation assistance system, we designed and conducted navigation theory tests comprising over 1500 multiple-choice questions, similar in format to the official exams for the OOW certificate under the STCW. Our experiments evaluated the performance of 14 LLMs, including GPT-3.5-turbo, GPT-4, GPT-4o, ERINE-4.0-8k, and Qwen-turbo, among others. The performance metrics included accuracy, cost, and processing latency.

Among all the tested models, only GPT-4o achieved a passing score with an accuracy of 86%, suggesting its potential for supporting autonomous ship navigation and guidance systems. Although the experimental results indicate that most large language models (LLMs) perform relatively poorly on multiple-choice questions related to ship navigation knowledge, the success of GPT-4o highlights the promise of LLMs in ship navigation tasks. These findings underscore the necessity for further fine-tuning and optimization of model architectures to enhance the capabilities of LLMs in navigation tasks.

As the maritime industry moves towards greater automation and intelligence, ensuring the safety and reliability of LLM-assisted systems is crucial. Therefore, future work should focus on advancing LLM capabilities to meet the stringent demands of safe ship navigation. At the same time, we will also address the ethical, safety, and privacy issues associated with the application of LLMs in ship navigation. We plan to tackle these issues by clarifying responsibilities, ensuring transparency in decision making, maintaining system reliability, and implementing data protection and privacy measures.

Author Contributions: Conceptualization, J.H.; Methodology, D.P.; Validation, D.P.; Formal analysis, D.P.; Investigation, D.P.; Writing—original draft, D.P.; Writing—review & editing, J.H., M.C. and S.Z.; Supervision, J.H. and K.L.; Project administration, J.H. and K.L.; Funding acquisition, J.H. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was conducted by Dashuai Pei during his visit to the University of Essex, supported by the China Scholarship Council. Additionally, the research received funding from the Natural Science Foundation of Hubei Province, China, under Grant No. 2021CFA001. This work was also funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No. 824019 and No. 101022280, the Horizon Europe

MSCA programme under grant agreement No. 101086228, the EPSRC with RC Grant reference EP/Y027787/1, and the EPSRC/UKRI with grant reference RCP 15831/DCM4480.

Data Availability Statement: The data supporting the reported results, including the publicly archived datasets analyzed or generated during the study, can be found at https://github.com/PeiDashuai/LLMs_Nav, accessed on 1 July 2024. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma, S. *Economics of Maritime Business*; Routledge: London, UK, 2020.
2. UNCTAD. *Review of Maritime Transport 2023*, 2023rd ed.; United Nations: San Francisco, CA, USA, 2023.
3. OECD. *Impacts of Russia's War of Aggression against Ukraine on the Shipping and Shipbuilding Markets*; OCED: Paris, France, 2023.
4. de Vos, J.; Hekkenberg, R.G.; Banda, O.A.V. The impact of autonomous ships on safety at sea—A statistical analysis. *Reliab. Eng. Syst. Saf.* **2021**, *210*, 107558. [CrossRef]
5. StraitsResearch. Global Autonomous Ships Market to Expand at a CAGR of 6.81% by 2031. 2024. Available online: https://straitsresearch.com/press-release/global-autonomous-ships-market-outlook (accessed on 29 July 2024)
6. Fenton, A.J.; Chapsos, I. Ships without crews: IMO and UK responses to cybersecurity, technology, law and regulation of maritime autonomous surface ships (MASS). *Front. Comput. Sci.* **2023**, *5*, 1151188. [CrossRef]
7. Thombre, S.; Zhao, Z.; Ramm-Schmidt, H.; García, J.M.V.; Malkamäki, T.; Nikolskiy, S.; Hammarberg, T.; Nuortie, H.; Bhuiyan, M.Z.H.; Särkkä, S.; et al. Sensors and AI techniques for situational awareness in autonomous ships: A review. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 64–83. [CrossRef]
8. Qiao, Y.; Yin, J.; Wang, W.; Duarte, F.; Yang, J.; Ratti, C. Survey of Deep Learning for Autonomous Surface Vehicles in Marine Environments. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 3678–3701. [CrossRef]
9. Issa, M.; Ilinca, A.; Ibrahim, H.; Rizk, P. Maritime autonomous surface ships: Problems and challenges facing the regulatory process. *Sustainability* **2022**, *14*, 15630. [CrossRef]
10. Wright, R.G. Intelligent autonomous ship navigation using multi-sensor modalities. *Transnav Int. J. Mar. Navig. Saf. Sea Transp.* **2019**, *13*, 503–510. [CrossRef]
11. Han, J.; Cho, Y.; Kim, J.; Kim, J.; Son, N.s.; Kim, S.Y. Autonomous collision detection and avoidance for ARAGON USV: Development and field tests. *J. Field Robot.* **2020**, *37*, 987–1002. [CrossRef]
12. Sha, H.; Mu, Y.; Jiang, Y.; Chen, L.; Xu, C.; Luo, P.; Li, S.E.; Tomizuka, M.; Zhan, W.; Ding, M. Languagempc: Large language models as decision makers for autonomous driving. *arXiv* **2023**, arXiv:2310.03026.
13. Fu, D.; Li, X.; Wen, L.; Dou, M.; Cai, P.; Shi, B.; Qiao, Y. Drive like a human: Rethinking autonomous driving with large language models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024; pp. 910–919.
14. Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv* **2023**, arXiv:2303.10420.
15. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
16. Tang, Z.; Shen, K.; Kejriwal, M. An Evaluation of Estimative Uncertainty in Large Language Models. *arXiv* **2024**, arXiv:2405.15185.
17. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
18. Villa, J.; Aaltonen, J.; Koskinen, K.T. Path-following with lidar-based obstacle avoidance of an unmanned surface vehicle in harbor conditions. *IEEE/ASME Trans. Mechatron.* **2020**, *25*, 1812–1820. [CrossRef]
19. Cockcroft, A.N.; Lameijer, J.N.F. *Guide to the Collision Avoidance Rules*; Elsevier: Amsterdam, The Netherlands, 2003.
20. Kufoalor, D.K.M.; Johansen, T.A.; Brekke, E.F.; Hepsø, A.; Trnka, K. Autonomous maritime collision avoidance: Field verification of autonomous surface vehicle behavior in challenging scenarios. *J. Field Robot.* **2020**, *37*, 387–403. [CrossRef]
21. Kim, J.; Lee, C.; Chung, D.; Cho, Y.; Kim, J.; Jang, W.; Park, S. Field experiment of autonomous ship navigation in canal and surrounding nearshore environments. *J. Field Robot.* **2024**, *41*, 470–489. [CrossRef]
22. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Receive, Reason, and React: Drive as You Say, With Large Language Models in Autonomous Vehicles. *IEEE Intell. Transp. Syst. Mag.* **2024**, *4*, 81–94. [CrossRef]
23. Duan, Y.; Zhang, Q.; Xu, R. Prompting Multi-Modal Tokens to Enhance End-to-End Autonomous Driving Imitation Learning with LLMs. *arXiv* **2024**, arXiv:2404.04869.
24. Huang, S.; Zhao, X.; Wei, D.; Song, X.; Sun, Y. Chatbot and Fatigued Driver: Exploring the Use of LLM-Based Voice Assistants for Driving Fatigue. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–8.
25. Li, W.; Li, L.; Xiang, T.; Liu, X.; Deng, W.; Garcia, N. Can multiple-choice questions really be useful in detecting the abilities of LLMs? *arXiv* **2024**, arXiv:2403.17752.

26. Zhang, Z.; Xu, L.; Jiang, Z.; Hao, H.; Wang, R. Multiple-Choice Questions are Efficient and Robust LLM Evaluators. *arXiv* **2024**, arXiv:2405.11966.
27. Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; Huang, M. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv* **2023**, arXiv:2309.07045.
28. Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 62991–63010.
29. Wu, S.; Koo, M.; Blum, L.; Black, A.; Kao, L.; Fei, Z.; Scalzo, F.; Kurtz, I. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI* **2024**, *1*, AIdbp2300092. [CrossRef]
30. Dao, X.Q.; Le, N.B.; Ngo, B.B.; Phan, X.D. LLMs' Capabilities at the High School Level in Chemistry: Cases of ChatGPT and Microsoft Bing Chat. *ChemRxiv* **2023**. [CrossRef]
31. Sadek, A. The Standards of Training, Certification and Watchkeeping for Seafarers (STCW) Convention 1978. In *The International Maritime Organisation*; Routledge: London, UK, 2024; pp. 194–213.
32. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. CogVLM: Visual Expert for Pretrained Language Models. *arXiv* **2023**, arXiv:2311.03079.
33. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
34. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
35. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295.
36. AI@Meta. Llama 3 Model Card. 2024. Available online: https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3/ (accessed on 29 July 2024).