

MariQA: A Large Scale Question Answering Dataset in the Domain of Maritime Affairs

Anonymous ACL submission

Abstract

Currently, natural language processing (NLP) is still in the early stages of exploration in one of the world’s oldest industries, maritime, and to date, there is no large-scale dataset available. To fill this gap, we construct the first large scale maritime-focused dataset encompassing eight crew positions with approximately 90,000 question-answer pairs to comprehensively evaluate LLMs’ domain knowledge and response capabilities. Our experiments on this dataset revealed: mainstream LLMs lack maritime knowledge, where even state-of-the-art models like GPT-4o and Qwen-Max achieved only passing scores, showing the significant room for improvement of current LLMs in the domain of maritime affairs. To promote the development of large language models in the maritime field, we will open-sourcing the proposed dataset.

1 Introduction

Question 1: Why build a QA dataset in the domain of maritime affairs?

Answer: The maritime industry plays a crucial role in global trade, transportation, and economic development. Maritime shipping is the most cost-effective and efficient mode of transporting large volumes of goods across continents. As the backbone of international commerce, it facilitates the movement of goods, raw materials, and energy resources across vast distances. Approximately 90% of global trade is carried by sea, making it an essential component of the global supply chain¹. The sector also supports millions of jobs worldwide, ranging from shipbuilding and port management to logistics and navigation. In addition to its economic significance, the maritime industry contributes to the development of global connectivity, enabling nations to maintain interdependent

relationships and promoting international cooperation (Hoffmann et al., 2017).

Recently, there have been several efforts to leverage natural language processing (NLP) techniques to address some issues in the maritime affairs. Specifically, Teske et al. (2018); Mackenzie et al. (2021) leverage document segmentation algorithm and named entity recognition model to extract information about piracy from unstructured maritime news articles. Hodne et al. (2024) employ some NLP techniques, such as dialog management and response generation, to build a conversational user interface, which allows Maritime Autonomous Surface Ships (MASS) to communicate via radio with ships and shore stations. However, NLP is still in the early stages of exploration in one of the world’s oldest industries, maritime, and to date, there is no large-scale dataset available.

Question 2: How to build MariQA, a large-scale QA dataset about maritime affairs?

Answer: To construct such a large-scale dataset, we collect the exam questions for the fitness examination of maritime crew members in the People’s Republic of China and performed corresponding data noise removal and deduplication.

Question 3: What are the characteristics of MariQA?

Answer: (1) **Large-scale:** MariQA comprises 90,000 question-answer pairs; (2) **Well-structured:** MariQA is structured according to a tree-like hierarchical taxonomy. The taxonomy is two-layered: the first layer covers 8 crew positions, and the second layer covers the exam subjects corresponding to each position. (3) **Niche knowledge demand:** To address the question in MariQA, highly specialized knowledge is required, covering all aspects of maritime affairs, such as ship handling and collision avoidance, marine electrical systems and so on. Two examples of MariQA is shown in Figure 1. For the powerful large language models

¹<https://unctad.org/publication/review-maritime-transport-2024>

Question	The main reason why vacuum boiling seawater desalination devices should not use steam for direct heating is to avoid _____.
Options	A: Excessive boiling B: Rapid scaling, forming hard scale C: High heat consumption, uneconomical D: High salt content in the produced water
Analysis	B. Large heat exchange temperature difference, rapid scaling, forming hard scale
Question	When a power-driven vessel underway sees a vessel with both power and sail approaching from the port side astern and posing a risk of collision, the vessel should _____.
Options	A: Turn to starboard and give way B: Stop the engine C: Turn to port and pass astern of the other vessel D: Maintain course and speed
Analysis	D. When a power-driven vessel sees a sailing vessel with machinery approaching from the port side abeam and a risk of collision exists, the correct action is to maintain the current course and speed, i.e., "maintain course and speed." Such action helps avoid exacerbating the risk of collision due to sudden changes in course or speed. Therefore, the correct answer is D, maintain course and speed.

Figure 1: Two examples of MariQA. In the dataset, Analysis includes the correct answers and the evidence to answer the question.

(LLMs) of today, MariQA is an excellent testing ground to evaluate the breadth of knowledge of these LLMs.

Question 4: What are the findings from the experiments?

Answer: We test 6 mainstream LLMs on MariQA, including GPT-4o, GPT-3.5 turbo, Qwen2.5-Max and so on. Passing all 90,000 questions and their corresponding options to these LLMs, we find that Qwen2.5-max can achieve the best results, slightly higher than the passing score (70%), and meanwhile, most other LLMs struggle. The results reveal that currently mainstream LLMs lack maritime knowledge, as maritime knowledge rarely appears during both pre-training and supervised fine-tuning stages.

2 MariQA Dataset

MariQA is a specialized dataset for evaluating LLMs' maritime knowledge. It spans 8 professions (e.g., Captain and Chief Officer) and 40 subjects, comprising 90,000 question-answer pairs.

2.1 Data Construction

To construct such a large-scale dataset, we collect the exam questions for the fitness examination of maritime crew members in the People's Republic of China. The fitness examination is designed to assess and verify whether the crew members possess the professional knowledge, skills, and abilities required to perform maritime navigation tasks. Besides, the fitness examination is conducted in accordance with the relevant regulations of the In-

ternational Maritime Organization (IMO) and the requirements of national maritime authorities, with the aim of ensuring the safe navigation of ships, preventing accidents, and safeguarding the safety of crew members and the environment.

After collection, we further leverage string similarity to remove duplicates and avoid data repetition. The data in MariQA are presented in the form of multiple-choice questions. Each question-answer pair consists of an ID, topic content, options, answer, and analysis. The topic content refers to questions related to a specific position and subject, while the options provide several potential answers. The answer represents the correct solution to the question. In addition, we have collected explanations for each question from various resources, which are stored in the analysis section.

2.2 Data Taxonomy

In maritime operations, the responsibilities and required knowledge vary significantly across different seafarer roles. To address this, we have organized our dataset based on a tree-like hierarchical taxonomy, which is shown in the Figure 2. Besides, the statistical information about the taxonomy is shown in the Appendix Table 2.

Position Division. The maritime occupations we selected encompass a variety of critical roles essential to the efficient operation of a vessel, including Captain, Chief Officer, Chief Engineer, Second Engineer, Third Officer, Third Engineer, Able Seaman and Electrician. These positions represent key responsibilities in the maritime industry, each contributing to the safe navigation, technical

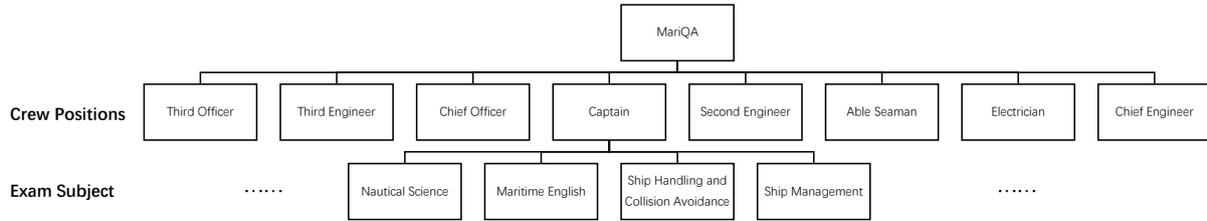


Figure 2: The tree-like hierarchical taxonomy of MariQA. Note that, there are a total of 40 exam subjects, due to page limitations, we only present 4 subjects corresponding to the captain position."

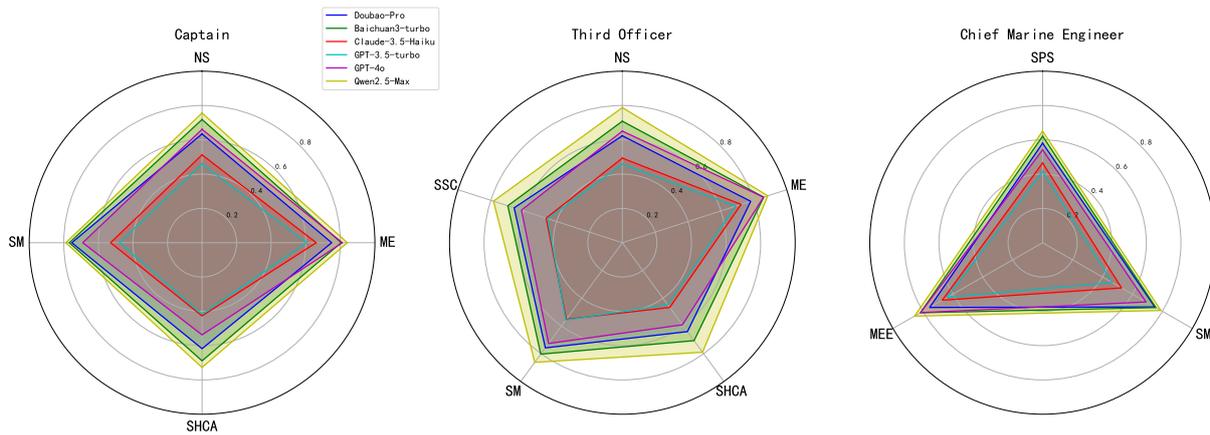


Figure 3: Accuracy of all tested LLMs on the Captain and Chief Mate test dataset.

144 maintenance, and overall functionality of a ship.
 145 Acquiring a comprehensive understanding of the
 146 duties and skills required for these roles is crucial
 147 for individuals aspiring to work in maritime sec-
 148 tors.

149 **Subject Division.** We also categorized the sub-
 150 jects required for each occupation. For example,
 151 the subjects that a captain needs to study include
 152 nautical science, maritime English, ship handling
 153 and collision avoidance, and ship management,
 154 among others. Such categorization helps to clarify
 155 the educational path and training required for each
 156 occupation, ensuring a comprehensive understand-
 157 ing of the maritime industry.

158 3 Experiments

159 3.1 Tested LLMs

160 We evaluate 6 LLMs capable of handling Chi-
 161 nese input on MariQA, including GPT-4o (OpenAI
 162 et al., 2024), GPT-3.5 turbo², Qwen2.5-Max (Team,
 163 2024), Claude 3.5 Haiku³, Baichuan3-Turbo and
 164 Doubao-pro-128k⁴.

²<https://platform.openai.com/docs/models/gpt-3-5-turbo>

³<https://www.anthropic.com/claude/haiku>

⁴<https://www.volcengine.com/product/doubao>

The input prompt consists of a question and its
 corresponding four options, and the tested LLM is
 required to output the corresponding answer based
 on the prompt.

169 3.2 Experimental Results

170 Table 1 shows the overall performance of tested
 171 LLMs on MariQA. From the table, we could find
 172 that: (1) Qwen2.5-max performs the best, scoring
 173 78.59% accuracy on the dataset. The reason Qwen
 174 performs so well could be that it incorporated some
 175 maritime data during the pre-training or supervised
 176 fine-tuning process; (2) GPT-3.5 turbo performs the
 177 worst, only achieving 50.51% accuracy. Consider-
 178 ing that sailors typically need to score above 70%
 179 to pass the exam, there is still significant room for
 180 improvement in the performance of current LLMs.

181 Furthermore, we make a statistic on the perfor-
 182 mance of these LLMs across various positions and
 183 subjects. The results are shown in the Figure 4 and
 184 Appendix Table 3. From the results, we find that:
 185 the models exhibit varying levels of accuracy, but
 186 overall, none of them significantly exceed the pass-
 187 ing threshold. For instance, in the subjects of Nauti-
 188 cal Science and Maritime English for the Captain
 189 profession, Qwen2.5-Max achieves accuracy rates

Question	Chemical reaction films are suitable for _____ working conditions.
Options	A. Moderate load, speed, and temperature B. Normal temperature, low speed, and light load C. Heavy load, high speed, and high temperature D. Normal temperature, medium speed, and high temperature
Analysis	C. Chemical reaction films, such as oxide films and phosphating films, possess exceptional wear resistance, high-temperature resistance, and load-bearing capabilities due to their unique chemical properties and structure. Under extreme working conditions involving heavy loads, high speeds, and high temperatures, these films effectively protect the substrate material, reduce wear and frictional heat, and ensure stable operation and extended service life of equipment. Therefore, chemical reaction films are particularly suitable for heavy load, high speed, and high-temperature working conditions.
Qwen2.5-Max	C. Chemical reaction films are typically suitable for heavy load, high speed, and high-temperature working conditions, as such films form protective layers through chemical reactions and are ideal for harsh environments. ✔
Baichuan3-Turbo	C. Heavy load, high speed, and high temperature. Chemical reaction films are usually formed under extreme working conditions, such as heavy load, high speed, and high temperature, which promote chemical reactions and result in the formation of protective film layers. Such films provide better wear resistance and corrosion resistance. ✔
Claude 3.5 Haiku	A. Chemical reaction films are generally suitable for moderate load, speed, and temperature working conditions, as these conditions ensure the stability and performance of the film. Option A best aligns with the general application characteristics of chemical reaction films. ✘
GPT-4o	A. Chemical reaction films are typically suitable for moderate load, speed, and temperature conditions, as such conditions ensure the stability and performance of the film. ✘
Doubao-pro-128k	A. Chemical reaction films are suitable for moderate load, speed, and temperature working conditions, as such conditions facilitate the formation of stable chemical reaction films, providing effective lubrication and protection. ✘
GPT-3.5 turbo	B. Chemical reaction films are typically suitable for normal temperature, low speed, and light load conditions. ✘

Figure 4: A case in the subject of Main Propulsion Plant of testing Third Engineer. From the case, we could find that Qwen and Baichuan give the correct answer, while the other LLMs fail. But when analyzing the evidence of the result provided by LLMs, we find that current LLMs still cannot match the performance of professionals on MariQA.

LLMs	Accuracy(%) [↑]
GPT-4o	68.39
GPT-3.5 turbo	50.51
Qwen2.5-Max	78.57
Baichuan3-Turbo	73.82
Claude 3.5 Haiku	54.37
Doubao-pro-128k	70.06

Table 1: Overall Performance of tested LLMs on MariQA

of 75.61% and 84.06%, respectively. Baichuan3-turbo ranks second in these subjects, with accuracy rates of 71.83% and 80.51%, while GPT-4o performs poorly, with accuracy rates of only 66.11% and 81.21% for the aforementioned subjects. However, GPT-4o’s accuracy in Maritime English ranks higher than in other subjects. For example, it ranked fourth in Ship Management, and Ship Handling & Collision Avoidance, but ranked second in Maritime English. Details are listed in the Table 3. Additionally, GPT-3.5 turbo and Claude 3.5 Haiku exhibit poor performance, with average accuracy rates are 50.51% and 54.37%. This observation suggests that as the model size increases, the per-

formance in handling specific, specialized tasks improves (Dong et al., 2024). In the maritime domain, the task performance varies across different crew position and exam subjects. Larger, updated models generally demonstrate greater advantages, but this also indicates that existing LLMs still have shortcomings in the domain of maritime affairs.

4 Conclusion

We propose MariQA, a specialized dataset designed to assess the proficiency of LLMs in maritime knowledge. Our dataset covers 8 crew positions across approximately 40 distinct subjects, containing roughly 90,000 question-answer pairs. Our experiments on MariQA show that Qwen2.5-max can achieve the best results, slightly higher than the passing score (70%), and meanwhile, most other LLMs struggle. The results reveal that currently mainstream LLMs lack maritime knowledge, as maritime knowledge rarely appears during both pre-training and supervised fine-tuning stages. This work highlights the current limitations of LLMs in maritime knowledge and underscores the need to improve both training data and model architectures for better domain-specific understanding.

228 **Limitations**

229 Although this study offers valuable contributions,
230 we acknowledge the following limitations:

- 231 1. In this article, we only evaluated six main-
232 stream LLMs and analyzed their performance,
233 without proposing new methods to improve
234 the performance of LLMs on this dataset.
- 235 2. The data we used are mainly collected from
236 the China Maritime Safety Administration,
237 which may encounter compatibility issues
238 when extended to maritime issues in other
239 countries.

240 **References**

241 Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng
242 Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng
243 Yuan, Chang Zhou, and Jingren Zhou. 2024. [How
244 abilities in large language models are affected by
245 supervised fine-tuning data composition.](#)

246 Philip Hodne, Oskar K. Skåden, Ole Andreas Alsos,
247 Andreas Madsen, and Thomas Porathe. 2024. [Con-
248 versational user interfaces for maritime autonomous
249 surface ships.](#) *Ocean Engineering*, 310:118641.

250 Jan Hoffmann, Gordon Wilmsmeier, and YH Venus Lun.
251 2017. [Connecting the world through global shipping
252 networks.](#)

253 Andrew Mackenzie, Alexander Teske, Rami Abielmona,
254 and Emil Petriu. 2021. [Maritime incident informa-
255 tion extraction using machine and deep learning tech-
256 niques.](#) In *2021 IEEE Symposium Series on Compu-
257 tational Intelligence (SSCI)*, pages 01–06.

258 OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher,
259 Adam Perelman, Aditya Ramesh, Aidan Clark,
260 AJ Ostrow, Akila Welihinda, Alan Hayes, Alec
261 Radford, Aleksander Mądry, Alex Baker-Whitcomb,
262 Alex Beutel, Alex Borzunov, Alex Carney, Alex
263 Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex
264 Renzin, Alex Tachard Passos, Alexander Kirillov,
265 Alexi Christakis, Alexis Conneau, Ali Kamali, Allan
266 Jabri, Allison Moyer, Allison Tam, Amadou Crookes,
267 Amin Tootoochian, Amin Tootoonchian, Ananya
268 Kumar, Andrea Vallone, Andrej Karpathy, Andrew
269 Braunstein, Andrew Cann, Andrew Codispoti, An-
270 drew Galu, Andrew Kondrich, Andrew Tulloch, An-
271 drey Mishchenko, Angela Baek, Angela Jiang, An-
272 toine Pelisse, Antonia Woodford, Anuj Gosalia, Arka
273 Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver,
274 Barret Zoph, Behrooz Ghorbani, Ben Leimberger,
275 Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin
276 Zweig, Beth Hoover, Blake Samic, Bob McGrew,
277 Bobby Spero, Bogo Giertler, Bowen Cheng, Brad
278 Lightcap, Brandon Walkin, Brendan Quinn, Brian

Guarraci, Brian Hsu, Bright Kellogg, Brydon East-
280 man, Camillo Lugaresi, Carroll Wainwright, Cary
281 Bassin, Cary Hudson, Casey Chu, Chad Nelson,
282 Chak Li, Chan Jun Shern, Channing Conger, Char-
283 lotte Barette, Chelsea Voss, Chen Ding, Cheng Lu,
284 Chong Zhang, Chris Beaumont, Chris Hallacy, Chris
285 Koch, Christian Gibson, Christina Kim, Christine
286 Choi, Christine McLeavey, Christopher Hesse, Clau-
287 dia Fischer, Clemens Winter, Coley Czarnecki, Colin
288 Jarvis, Colin Wei, Constantin Koumouzelis, Dane
289 Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy,
290 David Carr, David Farhi, David Mely, David Robin-
291 son, David Sasaki, Denny Jin, Dev Valladares, Dim-
292 itris Tsipras, Doug Li, Duc Phong Nguyen, Duncan
293 Findlay, Edede Oiwoh, Edmund Wong, Ehsan As-
294 dar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow,
295 Eric Kramer, Eric Peterson, Eric Sigler, Eric Wal-
296 lace, Eugene Brevdo, Evan Mays, Farzad Khorasani,
297 Felipe Petroski Such, Filippo Raso, Francis Zhang,
298 Fred von Lohmann, Freddie Sulit, Gabriel Goh,
299 Gene Oden, Geoff Salmon, Giulio Starace, Greg
300 Brockman, Hadi Salman, Haiming Bao, Haitang
301 Hu, Hannah Wong, Haoyu Wang, Heather Schmidt,
302 Heather Whitney, Heewoo Jun, Hendrik Kirchner,
303 Henrique Ponde de Oliveira Pinto, Hongyu Ren,
304 Huiwen Chang, Hyung Won Chung, Ian Kivlichan,
305 Ian O’Connell, Ian O’Connell, Ian Osband, Ian Sil-
306 ber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya
307 Kostrikov, Ilya Sutskever, Ingmar Kanitscheider,
308 Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub
309 Pachocki, James Aug, James Betker, James Crooks,
310 James Lennon, Jamie Kiros, Jan Leike, Jane Park,
311 Jason Kwon, Jason Phang, Jason Teplitz, Jason
312 Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Var-
313 avva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui
314 Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang,
315 Joaquin Quinonero Candela, Joe Beutler, Joe Lan-
316 ders, Joel Parish, Johannes Heidecke, John Schul-
317 man, Jonathan Lachman, Jonathan McKay, Jonathan
318 Uesato, Jonathan Ward, Jong Wook Kim, Joost
319 Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross,
320 Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao,
321 Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai
322 Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin
323 Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu,
324 Kenny Nguyen, Keren Gu-Lemberg, Kevin Button,
325 Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle
326 Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lau-
327 ren Workman, Leher Pathak, Leo Chen, Li Jing, Lia
328 Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lil-
329 ian Weng, Lindsay McCallum, Lindsey Held, Long
330 Ouyang, Louis Feувrier, Lu Zhang, Lukas Kon-
331 draciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz,
332 Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine
333 Boyd, Madeleine Thompson, Marat Dukhan, Mark
334 Chen, Mark Gray, Mark Hudnall, Marvin Zhang,
335 Marwan Aljubeih, Mateusz Litwin, Matthew Zeng,
336 Max Johnson, Maya Shetty, Mayank Gupta, Meghan
337 Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao
338 Zhong, Mia Glaese, Mianna Chen, Michael Jan-
339 ner, Michael Lampe, Michael Petrov, Michael Wu,
340 Michele Wang, Michelle Fradin, Michelle Pokrass,
341 Miguel Castro, Miguel Oom Temudo de Castro,
342 Mikhail Pavlov, Miles Brundage, Miles Wang, Mi-

343 nal Khan, Mira Murati, Mo Bavarian, Molly Lin,
344 Murat Yesildal, Nacho Soto, Natalia Gimelshein, Na-
345 talie Cone, Natalie Staudacher, Natalie Summers,
346 Natan LaFontaine, Neil Chowdhury, Nick Ryder,
347 Nick Stathas, Nick Turley, Nik Tezak, Niko Felix,
348 Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel
349 Bundick, Nora Puckett, Ofir Nachum, Ola Okelola,
350 Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins,
351 Olivier Godement, Owen Campbell-Moore, Patrick
352 Chao, Paul McMillan, Pavel Belov, Peng Su, Pe-
353 ter Bak, Peter Bakkum, Peter Deng, Peter Dolan,
354 Peter Hoeschele, Peter Welinder, Phil Tillet, Philip
355 Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming
356 Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Ra-
357 jan Troll, Randall Lin, Rapha Gontijo Lopes, Raul
358 Puri, Reah Miyara, Reimar Leike, Renaud Gaubert,
359 Reza Zamani, Ricky Wang, Rob Donnelly, Rob
360 Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchan-
361 dani, Romain Huet, Rory Carmichael, Rowan Zellers,
362 Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan
363 Cheu, Saachi Jain, Sam Altman, Sam Schoenholz,
364 Sam Toizer, Samuel Miserendino, Sandhini Agar-
365 wal, Sara Culver, Scott Ethersmith, Scott Gray, Sean
366 Grove, Sean Metzger, Shamez Hermani, Shantanu
367 Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shi-
368 rong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay,
369 Srinivas Narayanan, Steve Coffey, Steve Lee, Stew-
370 art Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao
371 Xu, Tarun Gogineni, Taya Christianson, Ted Sanders,
372 Tejal Patwardhan, Thomas Cunningham, Thomas
373 Degry, Thomas Dimson, Thomas Raoux, Thomas
374 Shadwell, Tianhao Zheng, Todd Underwood, Todor
375 Markov, Toki Sherbakov, Tom Rubin, Tom Stasi,
376 Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce
377 Walters, Tyna Eloundou, Valerie Qi, Veit Moeller,
378 Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne
379 Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra,
380 Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian,
381 Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen
382 He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and
383 Yury Malkov. 2024. [Gpt-4o system card](#). *Preprint*,
384 arXiv:2410.21276.

385 Qwen Team. 2024. Qwen2.5 technical report. *arXiv*
386 *preprint arXiv:2412.15115*.

387 Alexander Teske, Rafael Falcon, Rami Abielmona, and
388 Emil Petriu. 2018. Automatic identification of mar-
389 itime incidents from unstructured articles. In *2018*
390 *IEEE Conference on Cognitive and Computational*
391 *Aspects of Situation Management (CogSIMA)*, pages
392 42–48. IEEE.

Role	Subject	Total Num.
Captain	Nautical Science	5373
	Maritime English	2102
	Ship Handling and Collision Avoidance	8419
	Ship Management	10704
Chief Officer	Nautical Science	7363
	Maritime English	4987
	Ship Handling and Collision Avoidance	11798
	Ship Management	6378
	Ship Construction and Cargo	8926
Chief Engineer	Marine Power Plant	14116
	Ship Management	4402
	Marine Engineering English	3554
Second Engineer	Main Propulsion Plant	9564
	Marine Electrical and Automation	5714
	Ship Management	8265
	Marine Auxiliary Machinery	5632
	Marine Engineering English	3493
Third Officer	Nautical Science	11519
	Maritime English	2530
	Ship Handling and Collision Avoidance	9696
	Ship Management	5066
	Ship Construction and Cargo	7083
Third Engineer	Main Propulsion Plant	5054
	Marine Engines	944
	Marine Electrical and Automation	5446
	Ship Management	6175
	Marine Auxiliary Machinery	6134
	Marine Engineering English	624
Able Seaman	Motorman Watchkeeping Duties	1133
	Able Seaman Watchkeeping Duties	1617
	Watchkeeping Duties for Able Seaman (Under 500 GT)	605
	Watchkeeping Duties for Motorman (Under 750 KW)	692
	Electro-Technical Officer Duties	1128
Electrician	Information Technology and Communication Navigation	1199
	Electro-Technical Officer English	749
	Ship Engine Room Automation	1381
	Marine Electrical Systems	2906
	Ship Management	2821

Table 2: The achievement rate of abstract goals, the achievement rate of specific goals and the preference index of each model in inductive reasoning (evaluated by humans).

Role	Subject	Doubao	Claude	GPT-4o	GPT-3.5	BC	Qwen
Captain	Nautical Science	63.43	51.23	66.11	45.99	71.83	75.61
	Maritime English	75.01	66.06	81.21	60.52	80.51	84.06
	SH & CA	61.78	42.67	53.75	41.74	68.86	72.73
	Ship Management	75.37	52.92	69.05	47.84	76.79	78.84
Chief Officer	Nautical Science	63.38	50.00	64.57	46.05	69.14	74.02
	Maritime English	74.30	66.48	80.64	63.77	79.67	82.78
	SH & CA	63.20	43.87	54.91	41.74	67.03	75.09
	Ship Management	73.16	53.35	69.01	50.20	73.46	80.49
	SC & CH	56.79	41.19	52.83	39.96	59.48	68.66
Chief Engineer	Marine Power Plant	58.02	46.48	54.18	41.47	62.15	65.06
	Ship Management	74.88	52.73	69.09	46.54	75.67	78.98
	Marine Engineering English	75.49	67.03	81.54	63.79	81.67	85.60
Second Engineer	Main Propulsion Power Plant	65.88	46.75	62.60	44.84	69.58	75.32
	ME & AS	73.21	56.19	70.01	51.36	75.51	81.21
	Ship Management	73.93	52.82	68.60	51.32	73.03	79.46
	Marine Auxiliary Machinery	66.39	46.42	62.32	43.42	69.21	74.93
	Marine Engineering English	76.07	66.45	81.25	65.47	81.02	85.09
Third Officer	Nautical Science	62.35	49.30	65.09	45.89	70.85	78.77
	Maritime English	78.09	72.27	85.88	68.98	85.98	88.50
	SH & CA	64.08	46.67	59.16	44.71	70.58	78.92
	Ship Management	75.80	55.17	72.61	55.75	80.30	86.19
	SC & CH	65.83	46.37	61.39	45.36	69.69	78.40
Third Engineer	Main Propulsion Power Plant	69.14	49.26	63.45	45.21	72.12	81.69
	Marine Main Engine	69.25	48.25	63.94	49.05	68.33	76.72
	ME & AS	79.83	58.05	76.23	54.38	80.29	86.87
	Ship Management	75.48	52.59	69.38	50.45	76.79	84.41
	Marine Auxiliary Machinery	71.11	48.28	66.03	46.78	72.57	82.27
	Marine Engineering English	81.89	73.35	86.63	72.71	86.04	89.11
Able Seaman	Watchkeeping Engineer Duties	65.87	51.16	63.64	43.38	71.77	73.65
	Watchkeeping Seaman Duties	51.76	39.83	51.40	37.91	54.88	56.65
	WSDV Under 500 GT	55.87	47.00	54.96	41.16	57.88	61.76
	WEDV Under 750 KW	51.81	38.17	48.86	33.98	55.83	58.23
	Electro-Technical Officer Duties	88.09	68.37	87.31	62.30	88.36	92.36
Electrician	IT & CN	73.76	61.25	74.96	51.83	78.57	82.55
	Electro-Technical Officer English	84.91	87.11	85.52	82.04	86.85	91.58
	Electrical Engineering English	85.96	88.42	85.77	84.65	86.69	91.90
	Electrical Appliances English	87.61	89.42	86.59	85.60	88.65	93.12
	Electro-Technical Officer Skills	89.15	86.88	87.53	85.99	88.46	94.15

Table 3: The table above shows the performance of various models across different subjects for maritime roles, with scores representing the proficiency of each model in specific areas. Below is an explanation of the abbreviations:

Abbreviation	Explanation
SH&CA	Ship Handling and Collision Avoidance
SC&CH	Ship Construction and Cargo Handling
ME&AS	Marine Electrical and Automation Systems
WSDV	Watchkeeping Seaman Duties for Vessels
WEDV	Watchkeeping Engineer Duties for Vessels
IT & CN	Information Technology and Communication Navigation
BC	Baichuan