Towards Benchmarking Situational Awareness of Large Language Models Comprehensive Benchmark, Evaluation and Analysis

Guo Tang[1], **Zheng Chu**[1], **Wenxiang Zheng**[1], **Ming Liu**[1,2*], **Bing Qin**[1,2]
[1]Harbin Institute of Technology, Harbin, China
[2]PENGCHENG Laboratory
{gtang,zchu,wxzheng,mliu,qinb}@ir.hit.edu.cn

Abstract

Situational awareness refers to the capacity to perceive and comprehend the present context and anticipate forthcoming events, which plays a critical role in aiding decision-making, anticipating potential issues, and adapting to dynamic circumstances. Nevertheless, the situational awareness capabilities of large language models have not yet been comprehensively assessed. To address this, we propose SA-Bench, a comprehensive benchmark that covers three tiers of situational awareness capabilities, covering environment perception, situation comprehension and future projection. SA-Bench provides a comprehensive evaluation to explore the situational awareness capabilities of LLMs. We conduct extensive experiments on advanced LLMs, including GPT-4, LLaMA3, Qwen1.5, among others. Our experimental results indicate that even SOTA LLMs still exhibit substantial capability gaps compared to humans. In addition, we thoroughly analyze and examine the challenges encountered by LLMs across various tasks, as well as emphasize the deficiencies they confront. We hope SA-Bench will foster research within the field of situational awareness.

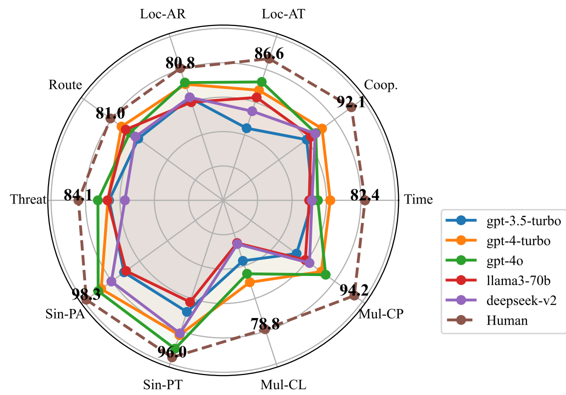

Figure 1: Overall performance of state-of-the-art large language models and humans on SA-Bench.

1 Introduction

Situational awareness (SA) is crucial for facilitating decision-making, foreseeing possible problems, and adjusting to changing environments. It involves perceiving elements in the environment, comprehending their meaning, and projecting their future status (Fracker, 1988). Previous researches category SA into three levels: (Endsley, 1995; Endsley et al., 2000; Endsley, 2012, 2017) **Perception** identifies and records relevant information and dynamics within the environment, such as targets, events, and entities. **Comprehension** synthesizes this information to understand relationships and impacts on the current situation through data analysis and correlation. **Projection** predicts future changes and impacts of these elements, aiding in effective decision-making and planning. Recently, with the rapid scaling of pre-training, LLMs have demonstrated outstanding performance in various downstream tasks (Zhao et al., 2023), possessing excellent text comprehension and logical reasoning capabilities (Yu et al., 2020; Cobbe et al., 2021; Wei et al., 2022; Kojima et al., 2022). With their exceptional capabilities in information extraction, reasoning and decision-making, LLM-based situational awareness has attracted a lot of reaserch interest.

Previous research typically focuses on a single level of SA. For example, information extraction (Yang et al., 2022; Mayhew et al., 2023) focuses on environment perception, link prediction and logical reasoning (Zhang et al., 2024; Zellinger et al., 2024; Zeng et al., 2023) examine situation comprehension, and scenario prediction (Lv et al., 2019; Jin et al., 2020c; Hu et al., 2022) explores future projection. In summary, existing studies focus narrowly and lack comprehensive evaluations of the overall capabilities in SA,

* Corresponding Author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 7904–7928

thereby presenting certain limitations.

To address this, we propose a comprehensive situational awareness benchmark, SA-Bench. Specifically, we category situational awareness into *Evaluation* and *Forecast*. *Evaluation* corresponds to level 1 and level 2 of SA, encompassing two sub-tasks with four diverse scenario categories. We design multi-level knowledge bases and situational questions for various maritime vessels, using a question answering format to evaluate the model's capabilities in information extraction, environmental awareness, and information comprehension. *Forecast* corresponds to the third level of SA, featuring five sub-tasks from different spectrum. We provide collections of real news texts from economic, technological, and other domains in a QA format, requiring models to predict various elements such as time, location, events, and actions based on historical background news. To balance task difficulty, we combine multiple-choice questions (MCQ) and multiple-select questions (MSQ) in different proportions, simulating the decision-making process in real SA tasks as closely as possible. Additionally, we conduct rigorous expert design, multiple independent annotations, and unified verification of the dataset. We validate the data annotations based on Fleiss' Kappa, which indicates that our annotation process is consistent with high annotation quality.

We conduct extensive experiments with advanced LLMs to quanty their capabilities in situational awareness, including GPT3.5 (Ouyang et al., 2022), GPT-4/GPT-4o (Achiam et al., 2023), Gemini-1.5-pro (Reid et al., 2024), LLaMA3 (Touvron et al., 2023), Claude-3 (Anthropic, 2024), DeepSeek-V2 (Bi et al., 2024), Qwen1.5 (Bai et al., 2023; Team, 2024), and Yi (Young et al., 2024). We incorporate chain-of-thought reasoning techniques and evaluate the models under zero-shot and few-shot settings. As shown in Figure 1, the experimental results indicate that GPT-4 and GPT-4o continue to lead, demonstrating excellent capabilities in information extraction, knowledge reasoning, and reasonable prediction. Despite these advancements, a notable disparity persists when compared to human performance. Other advanced models exhibit varying levels of performance across different tasks, but overall, the differences are minor, with all models slightly outperforming GPT-3.5. These findings suggest that while models have room for improvement in information extraction, reasoning, and prediction when handling complex, longer texts, significant progress is still needed. Additionally, we observe that chain-of-thought prompts do not significantly enhance model performance and may even cause slight performance degradation. Finally, we conduct a thorough analysis, providing a detailed examination of the challenges encountered by LLMs in situational awareness tasks.

We summarize our contribution as follows:

(1) We propose SA-Bench, a comprehensive and hierarchical benchmark for assessing the situational awareness capabilities of LLMs.

(2) Our extensive experiments demonstrate that even the state-of-the-art LLMs still have a significant gap compared to human performance in SA tasks, which underscores the need for further investigation and discovery in this field.

(3) Through a thorough analysis, we identify the defects of LLMs in situation awareness tasks defined by Endsley (2012), shedding light for future research.

2 Definition of Situational Awareness

Situational awareness is defined as the perception of elements within a specific time and space, the comprehension of their meaning, and the projection of their future status (Fracker, 1988). Based on previous research, situational awareness can be conceptually divided into three levels (Endsley, 1995; Endsley et al., 2000; Endsley, 2012).

Environmental Element Perception (L1) involves identifying and perceiving relevant information and dynamics within the current environment, encompassing the detection of various elements such as targets, events, and surrounding entities. The system must accurately detect and record information about these elements.

Current Situation Comprehension (L2) requires the system to synthesize the perceived information from various elements, understand the relationships between these elements and their impact on the current situation. This procession necessitates data analysis and correlation.

Future Status Projection (L3) involves predicting future changes and impacts of elements based on the understanding of the current situation. The system must predict the future states and behaviors of environmental elements using current context and historical data, which is crucial for effective decision-making and planning.

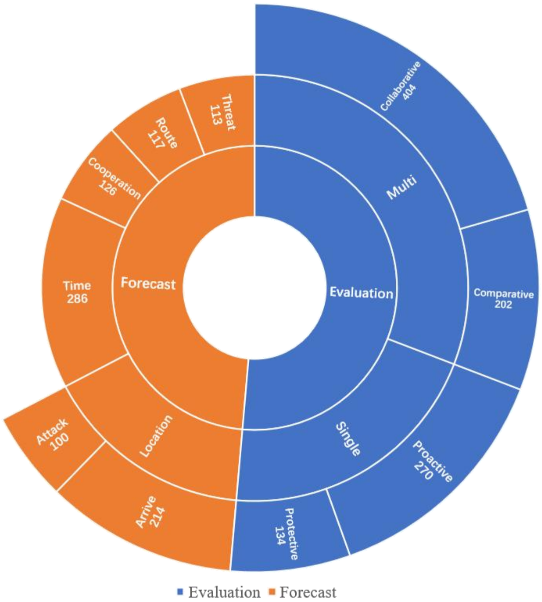

Figure 2: Distribution of categories, tasks and sub-tasks in SA-Bench.

Task	Sub-task	Sub-angle	#	Format
Evaluation	Single	Proactive	270	MCQ+MSQ
		Protective	134	MCQ+MSQ
	Multi	Collaborative	404	MCQ+MSQ
		Comparative	202	MCQ+MSQ
Forecast	Time	-	286	MCQ
	Cooperation	-	126	MCQ
	Location	Attack	100	MCQ
		Arrive	214	MCQ
	Route	-	117	MCQ
	Threat	-	113	MSQ
In total			1966	

Table 1: The statistics of hierarchical tasks, sub-tasks and sub-angles in our SA-Bench.

3 SA-Bench Benchmark

3.1 Benchmark Overview

SA-Bench evaluates different levels of situational awareness capabilities through question-answer across various scenarios. Table 1 shows the statistics of SA-Bench, and Figure 2 presents the distribution of subtasks within SA-Bench. SA-Bench consists of two categories: *Evaluation* and *Forecast*. Examples can be found in Appendix A.1.

Evaluation tasks correspond to the first two levels of SA, where we design scenarios to assess the multi-angle performance of maritime vessels. Models are required to extract pertinent information from complex technical data, integrate this information in accordance with specific scenarios, and provide analytical evaluations.

Forecast tasks align with the third level of SA, necessitating that models predict potential future events by analyzing a provided context of news articles. We proposed five predictive angles in various scenarios, with each testing different aspects of the model's predictive capabilities.

3.2 Evaluation Task (L1 & L2)

3.2.1 Task Definition

Evaluation tasks aim to assess the model's ability to perceive environmental elements and the current situation in specific scenarios involving one or more vessels. The model's capabilities in L1/L2 of SA are evaluated through knowledge reasoning.

3.2.2 Task Categorization

Evaluation tasks are categorized into two distinct types: *Single Evaluation* and *Multi Evaluation*. In *Single Evaluation* tasks, we design assessment questions from proactive and protective perspectives to evaluate seven different aspects of a single vessel's performance. In *Multi Evaluation* tasks, the model is required to extract and understand information about two or more vessels and provide evaluative answers based on the specific context of the questions, which includes collaborative and comparative scenarios.

Single-Proactive evaluates the performance of a single vessel's proactive actions in various areas such as air combat, surface combat, electronic warfare, and personnel transportation.

Single-Protective evaluates the self-defense capabilities of a single vessel, including anti-submarine warfare, reconnaissance, and close-in defense.

Multi-Collaborative involves collaborative scenarios of multiple vessels, which requires understanding the context of the question and reasoning based on background knowledge.

Multi-Comparative involves comparative scenarios of multiple vessels, requiring comparative reasoning based on their characteristics.

3.2.3 Challenges in Evaluation Tasks

We delineate the challenges LLMs encounter when undertaking the evaluation tasks as follows:

Complexity of Multi-Attribute Entity Data. Vessels, as multi-attribute entities, comprise extensive technical data across domains like structure, power, armament, and electronic systems. Effectively organizing and managing this complex data is crucial for accurate application of inference rules to relevant attributes and values.

Design and Application of Inference Rules. Inference rules must derive new attributes from existing ones through intricate logical relationships and calculations. The challenge is to design comprehensive, accurate inference rules and ensure their efficient application in the inference process.

Computational Complexity of the Inference Process. The inference process requires substantial computation and data processing, especially with large knowledge bases of entities and attributes. Optimizing the inference algorithm to perform efficiently and handle significant computational demands is a major challenge.

Dealing with Uncertainty and Ambiguity. Data uncertainty and ambiguity, such as missing or unclear attribute values, are common in practical applications. Managing these uncertainties during the inference process to ensure reliable and accurate results is a significant challenge.

3.3 Forecast Task (L3)

3.3.1 Task Definition

Forecast tasks correspond to Level-3 of situational awareness, requiring the model to predict the possible future times, events, and actions based on historical environment context.

3.3.2 Task Categorization

To evaluate the model's predictive ability from various perspectives, we have divided the *Forecast* tasks into five subtasks. Detailed examples of each subtask can be found in Appendix A.1.

Time aims to predict the specific time or duration of future events based on historical information. **Cooperation** requires LLMs to forecast potential alliances or cooperative relationships between nations based on current geopolitical contexts and historical events. **Location** is divided into *Attack Location*, which involves anticipating the specific locations of future attacks based on historical security threat information, and *Arrive Location*, aiming at predicting the exact venues of future major international conferences or events based on plans and announcements. **Route** focuses on forecasting the primary routes for future actions based on infrastructure developments and geographical information. **Threat** requires anticipating the most likely potential threats to target entities based on environmental changes and reports.

3.3.3 Challenges in Forecast Task

We outline the challenges and obstacles that LLMs may meet in evaluation tasks as follows.

Grasping Temporal and Causal Dynamics Effective prediction hinges on a thorough understanding of how historical events shape future possibilities. It requires models' capability of adeptly reasoning about temporal and causal dynamics within the acquired data.

Predictive Evaluation In addition to gathering and interpreting information, the models are required to generate informed predictions about forthcoming events, a process that entails merging incomplete data and anticipating possible outcomes. The prediction process is inherently speculative and fraught with uncertainty.

Time Limitations The short time for accessing and analyzing articles introduces another layer of difficulty, restricting the availability of complete data and complicating the accuracy of predictions.

Inference Skills Forecasting, unlike other QA tasks that depend on clear-cut answers, necessitates models to infer and forecast. It requires advanced reasoning capabilities to connect historical data with future scenarios.

3.4 Task Format

To balance the difficulty levels across different SA tasks, we design the Evaluation tasks to include both MCQ and MSQ formats, increasing the challenge for models in information extraction and comprehension. For Level 3 SA tasks, except for Threat part which uses MSQ to assess recall, all other sub-tasks use MCQ to narrow the prediction scope and emphasize response accuracy.

3.5 Evaluation Metrics

For the Multiple Choice Questions (MCQ), we utilize accuracy as the primary metric for evaluation. In the case of Multi-Select Questions (MSQ), we adopt option-based Exact Match (EM) and F1 scores to assess performance. When dealing with tasks that incorporate both MCQ and MSQ, we compute the mean of the EM and F1.

4 Benchmark Annotation

4.1 Data Sources

We obtained and integrated technical data of vessels from Wikipedia, Baidu Encyclopedia, and thematic websites. We also acquired our news

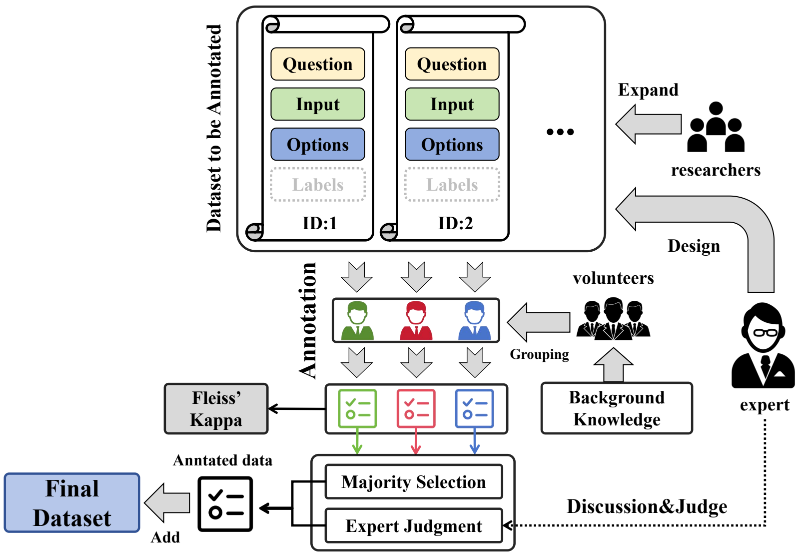

Figure 3: Annotation process of the SA-Bench with majority voting among three experts.

texts from multiple news websites including Baidu News, Tencent News, and China News Service.

4.2 Annotation Process

Our annotation process is shown in Figure 3. We employ a method involving multiple independent volunteers for benchmark annotation. Before the commencement of the formal annotation work, professionals versed in the respective scenario initially devise questions for each task. Subsequently, these questions are expanded based on existing resources to constitute a series of unlabeled datasets, each containing a *context*, *question*, and *options*. These require labeling, with each option being annotated with a label field to denote its correctness or otherwise. The annotators undergo training on the pertinent knowledge pertaining to the scenario. Upon finishing the background knowledge training, the team divides into three groups to individually annotate the answers. After the completion of all data annotations, in cases where the answers do not align completely, the majority answers is deemed the final answer. For questions that elicit three entirely distinct answers, experts convene for a unified discussion and correction, ultimately providing a consistent answer.

4.3 Annotation Rules and Standards

We require annotators to construct a *basis* field for each annotation while providing the answers. This field includes the evidence found in the input context and the reasoning process based on this evidence. This approach provides a reference for subsequent answer verification and correction, ensuring the rigor of the annotation work. After verifying and correcting all answers, we summarize the *basis* for each question, incorporating opinions from annotators and professionals, to provide a rigorous reasoning process for the answers.

4.4 Annotation Consistency Analysis

To assess the quality of the question and option designed in the dataset, we use the Fleiss' Kappa coefficient (κ) (Fleiss, 1971) to evaluate the consistency of the annotations. Fleiss' Kappa is a metric for assessing the agreement among multiple annotators on multi-class tasks. We calculate the corresponding κ for each sub-task of our dataset. The calculation method for the coefficients can be found in Appendix A.2. The results are shown in Table 6. According to the general interpretation from Landis and Koch (1977), most results demonstrate the effectiveness and rigor of our task design strategy. The κ results for the *Route* subtask under *Forecast* are relatively low, which is attributed to the complexity of the task. Each route involves numerous locations, making it challenging to achieve high annotation consistency for this predictive task. Similarly, the κ values for the *Multi Evaluation* subtask are also lower. This is due to the significantly larger knowledge base compared to the *Single* subtask, resulting in a more complex analysis process. Additionally, the higher proportion of MSQ in this subtask increases the difficulty of achieving consistent annotations.[1]

[1]We provide the details of labor costs in Appendix A.7.

5 Experimental Setup

5.1 Models

Considering the complexity and long input context of our tasks, we choose LLMs that can effectively process long texts and perform well in reasoning tasks, including both open-source and proprietary models. The details of LLMs used in experiments can be found in Appendix A.3.

5.2 Implementation Details

We conduct an evaluation with a prompt-based approach, which includes standard prompting and chain-of-thought prompting under zero-shot and few-shot settings, with details in Appendix A.4. The specific prompts used for each task in the experiments will be presented in Appendix A.5.

6 Experimental Results

6.1 Main Results

Overall Results The experimental results of our dataset are the average values of the English and Chinese versions. We list the experimental results of the aligned models in Table 2, covering some of the most popular LLMs. In most tasks, GPT-4o achieves the best performance, ranking first in 10 out of 15 evaluation metrics, while GPT-4 ranks first in the remaining 5 metrics, slightly lower than GPT-4o by 2.9%. All the remaining alignment models achieve at least 74% of GPT-4o's performance. However, the performance of LLMs still lags significantly behind humans, with GPT-4o trailing humans by as much as 18.1%.

Results on Evaluation Tasks In terms of evaluation results, GPT-4o continues to exhibit near-human-level performance, being the only model to score over 70 in *Evaluation* tasks, leading GPT-4 by 2.4 percentage points. Additionally, Gemini-1.5-Pro performs excellently, ranking in the top three across all performance metrics, demonstrating strong competitiveness in evaluation. Meanwhile, we observe that DeepSeek-V2 ranks third in the em metric for *Protective* tasks, only behind GPT-4o and Gemini-1.5-Pro. We also observe that chain-of-thought prompts have a negative impact on almost all models in *Comparative* tasks. Compared to the forecast results, the evaluation performance of the models has significantly improved, with the best performance observed in *Proactive* tasks. However, the models do not perform as well in *Multi Evaluation* tasks compared to *Single Evaluation* tasks, especially in *Collaborative* tasks, showing a significant gap with other *Evaluation* tasks. Meanwhile, in *Collaborative* tasks, the models' performance shows the most significant gap compared to human performance.

Results on Forecast Tasks Compared to other models, GPT-4o demonstrates strong forecast performance, but the gap with GPT-4 is not significant. GPT-4 ranks first in three out of seven metrics, trailing GPT-4o by only 1.6 percentage points. In *Route* tasks, Llama3 outperforms GPT-4o, ranking third overall in the *Forecast* tasks, showing its strong capability in forecast. The remaining models have similar scores in *Forecast* tasks, all scoring above 50. We notice that although GPT-3.5 does not perform prominently overall, it ranks second in the em metric for *Threat* tasks, only behind GPT-4o. Compared to other tasks, the models perform excellently in *Arrive Location* tasks, *Route* tasks, and *Threat* tasks, but poorly in the *Time* tasks, indicating that LLMs still have significant shortcomings in time forecast. In *Cooperative* tasks, despite the models achieving respectable performance, the gap with human performance is the largest compared to other tasks.

7 Analysis and Discussion

7.1 Chain-of-Thought Reasoning in Situational Awareness

Previous research has demonstrated that chain-of-thought prompts can effectively guide models to reason step-by-step, significantly improving accuracy on complex tasks (Brown et al., 2020; Kojima et al., 2022; Wei et al., 2022). In this study, we incorporate chain-of-thought prompts into SA tasks to investigate whether they can enhance the model's reasoning abilities in Levels 1 and 2, as well as its predictive capabilities in Level 3.

As illustrated in Figure 4, CoT does not achieve the expected preformance improvement. The use of zero-shot CoT prompts results in a 0.7% performance decrease in the Forecast task and a 2.3% decrease in the Evaluation task, with an overall impact of 1.5%. In the few-shot CoT scenario, the performance decline in the Forecast task is more pronounced, with a 1.7% drop, while the Evaluation task shows slightly better performance compared to zero-shot CoT. Overall, few-shot CoT causes a minor 1.3% impact. Both zero-shot and

Method	Forecast							Evaluation								Overall		
	Time	Coop.	Location Atk.	Location Arr.	Route	Threat		Single Proactive		Single Protective		Multi Collaborative		Multi Comparative		Fore.	Eval.	Overall
	Acc	*Acc*	*Acc*	*Acc*	*Acc*	*EM*	*F1*	*EM*	*F1*	*EM*	*F1*	*EM*	*F1*	*EM*	*F1*			
Human	82.4	92.1	86.6	80.8	81.0	67.7	84.1	97.5	98.3	95	96	69.4	78.8	80	94.2	82.1	88.7	85.6
gpt-3.5-turbo	43.0	51.0	38.4	56.0	51.0	29.0	63.2	**66.7**	**71.2**	**60.4**	**68.2**	24.5	32.2	44.0	45.2	47.4	**51.6**	49.6
+ *ZS/CoT*	39.0	**60.0**	41.5	**62.6**	**61.0**	**35.0**	**66.3**	61.1	66.7	56.0	65.1	20.0	28.7	44.0	45.2	**52.2**	48.4	**50.1**
+ *FS/CoT*	**52.0**	52.0	**44.1**	61.0	57.0	25.0	59.7	57.0	58.8	54.0	62.7	**25.0**	**37.0**	**52.0**	**52.7**	50.1	49.9	50.0
gpt-4-turbo	52.0	<u>**71.0**</u>	65.6	70.0	68.0	26.0	60.8	83.3	85.7	**75.4**	**82.4**	42.0	47.4	**69.0**	**70.5**	59.1	**69.5**	**64.6**
+ *ZS/CoT*	53.0	66.0	**67.3**	**71.0**	<u>**73.0**</u>	**33.0**	63.3	78.5	83.7	73.1	80.9	40.0	47.2	67.0	68.5	**60.9**	67.4	64.4
+ *FS/CoT*	<u>**62.2**</u>	65.0	61.3	70.0	60.0	29.8	**67.0**	**85.0**	**87.6**	73.0	79.4	<u>**43.0**</u>	<u>**50.1**</u>	63.0	65.9	59.3	68.4	64.2
gpt-4o	48.5	**66.5**	<u>**72.4**</u>	71.0	66.0	<u>**40.0**</u>	<u>**72.8**</u>	<u>**87.0**</u>	<u>**89.8**</u>	<u>**86.0**</u>	88.2	38.0	40.8	<u>**72.0**</u>	<u>**73.6**</u>	<u>**62.5**</u>	<u>**71.9**</u>	<u>**67.5**</u>
+ *ZS/CoT*	**55.0**	64.0	69.7	70.0	**67.0**	35.0	70.9	81.0	86.6	85.0	<u>**90.7**</u>	37.0	41.5	70.0	72.3	61.7	70.5	66.4
+ *FS/CoT*	54.0	55.4	62.6	<u>**72.0**</u>	59.0	38.4	67.5	81.0	85.9	76.0	83.0	**38.0**	**44.9**	70.0	70.6	58.4	68.7	63.9
gemini-1.5-pro	50.0	48.0	**54.5**	60.0	58.0	24.0	59.9	79.0	81.1	64.0	71.8	30.0	**41.1**	59.0	60.1	50.6	60.8	56.0
+ *ZS/CoT*	47.0	**58.0**	50.5	57.0	53.0	**31.0**	56.5	82.0	82.7	49.0	57.4	30.0	33.8	**67.0**	**68.5**	50.4	58.8	54.9
+ *FS/CoT*	**51.0**	55.0	52.5	**63.0**	**63.0**	**31.0**	**65.9**	**83.0**	**85.5**	**78.0**	**85.6**	**33.0**	39.0	60.0	60.0	**54.5**	**65.5**	**60.4**
llama3-70b	**50.0**	**63.0**	60.6	56.0	58.0	24.2	59.7	**67.8**	**69.8**	67.0	62.2	12.0	16.2	**59.0**	**59.0**	53.1	**51.6**	52.3
+ *ZS/CoT*	45.0	58.0	**62.9**	**60.0**	**70.0**	29.0	**67.2**	62.2	63.7	**68.0**	**64.0**	9.0	13.4	59.0	59.0	**56.0**	49.8	**52.7**
+ *FS/CoT*	45.0	38.0	41.4	55.5	48.0	**31.3**	60.6	62.0	62.3	62.0	62.8	**19.0**	**26.1**	53.0	54.0	45.7	50.2	48.1
claude-3-sonnet	45.0	**61.0**	**56.6**	**67.0**	**63.0**	24.0	62.7	**78.0**	**80.9**	63.0	71.8	32.3	37.4	55.0	57.0	54.2	59.4	**57.0**
+ *ZS/CoT*	**52.0**	60.0	53.2	63.0	59.0	**29.0**	**66.7**	65.0	73.0	65.0	**80.8**	30.0	**39.7**	58.0	59.2	**54.7**	58.8	56.9
+ *FS/CoT*	43.0	57.0	48.5	61.0	57.0	**29.0**	62.8	66.0	69.2	**73.0**	78.2	**33.0**	39.5	**63.0**	**65.0**	51.2	**60.9**	56.3
deepseek-v2	**51.6**	**66.1**	**54.5**	**63.0**	62.0	27.3	56.5	**76.0**	**80.3**	**76.0**	**81.3**	**19.2**	22.9	**62.0**	**62.0**	**54.4**	**60.0**	**57.4**
+ *ZS/CoT*	46.0	54.0	51.5	62.0	61.0	27.3	51.8	69.0	72.5	69.7	74.8	19.0	22.4	58.0	58.8	50.5	55.5	53.2
+ *FS/CoT*	48.0	52.0	54.2	62.0	**63.0**	**31.0**	**57.1**	68.0	70.9	64.0	70.5	19.0	**26.7**	52.0	52.8	52.5	53.0	52.7
qwen1.5-72b-chat	**51.0**	**54.0**	50.0	68.0	**66.0**	23.0	45.2	**75.0**	**82.1**	**70.0**	**77.2**	26.0	33.0	54.0	54.5	**51.0**	**59.0**	**55.3**
+ *ZS/CoT*	46.0	43.0	45.0	**69.0**	58.0	20.0	41.3	62.0	67.2	63.0	70.9	**31.0**	**37.9**	**59.0**	**60.3**	46.0	56.4	51.6
+ *FS/CoT*	47.0	52.0	**51.0**	65.0	56.0	**25.0**	**46.6**	63.0	71.0	67.0	71.7	27.0	31.8	53.0	55.5	48.9	55.0	52.2
yi-large-turbo	47.0	**54.0**	47.0	**67.0**	**56.0**	**25.0**	**62.1**	60.0	67.8	59.0	67.3	21.2	25.9	**55.0**	**55.7**	**51.2**	51.5	51.3
+ *ZS/CoT*	**48.0**	47.0	44.0	62.0	50.0	17.0	46.9	57.0	70.6	60.0	**72.3**	17.5	20.8	48.0	48.8	45.0	49.4	47.3
+ *FS/CoT*	41.0	48.0	**50.0**	65.0	52.0	21.0	57.0	**66.0**	**72.4**	**63.0**	71.2	**27.0**	**29.6**	54.0	54.8	47.7	**54.8**	**51.5**

Table 2: Experimental results of SA-Bench. The best results in each group are **bold** and global best are <u>underlined</u>. Default evaluation is conducted by standard prompting (first row in each group).

few-shot CoT methods do not significantly enhance model performance, but they also do not cause severe degradation.

Why does CoT fail to improve performance? We conduct a combined analysis of the experimental results and output cases, attributing the lack of improvement to two main reasons: (1) Most of the alignment models used in the experiments are already well-trained, and the baseline method suffices for most knowledge reasoning and prediction tasks. Adding CoT may lead to over-reasoning by the model, thus reducing performance. (2) The variety of questions within the same sub-task is extensive, but the number and diversity of provided examples are insufficient. This limits the model's ability to learn a broad range of reasoning and prediction capabilities.

Performance gaps across different tasks. To observe the impact of CoT on each subtask more thoroughly, we calculate the performance differences between CoT and SP across various subtasks, as shown in Figure 5. In the zero-shot scenario, proprietary models exhibit performance improvements in *Forecast* tasks, while showing minimal overall enhancement in *Evaluation* tasks. In the few-shot setting, proprietary models demonstrate significant improvements in the *Time* and *Threat Forecast* tasks, with noticeable gains in the *Multi* subtask of the *Evaluation* tasks.

By analyzing the dataset properties and output examples, we conclude that in the *Evaluation* tasks, few-shot CoT has a greater impact on complex knowledge reasoning tasks (such as *Multi-CL*). Providing examples that demonstrate information extraction and knowledge-based multi-hop implicit reasoning helps the model understand the task requirements better. For the *Forecast* tasks, making predictions requires the model to autonomously extract and integrate key background information before making a reasonable forecast. Although providing prediction examples does not effectively enhance the model's ability to extract key information, it can broaden the model's predictive capabilities. This explains the notable im-

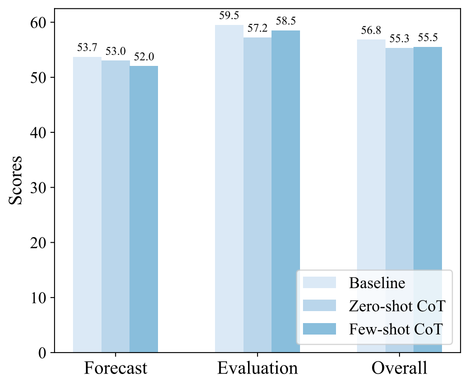

Figure 4: Performance gap with and without CoT. The results are averaged over different models.

provements seen with CoT in the *Threat* task, which requires multiple answers.

7.2 Thorough Analysis on Sub-tasks

As shown in Table 3, we select various LLMs to assess their performance across various sub-tasks. If an model exceeds its overall score on a particular sub-task, we consider it to have a better grasp in that aspect, and vice versa. Based on our observations, we draw the following conclusions:

Models excel at single elements compared to complex ones. From the results of the *Single* sub-task in the evaluation category, it is evident that models achieve high accuracy in extracting and understanding information about single entities. They can identify the required elements based on the problem context and perform simple analysis or calculations. However, as the information volume and complexity increase in the *Multi* tasks, models struggle to efficiently integrate information from multiple entities. They find it challenging to precisely extract useful information from complex multi-level knowledge and conduct comprehensive analysis. This indicates that the models' abilities in Levels 1 and 2 of SA significantly decline as the information volume increases.

Models are better at predicting specific entities. In the *Forecast* tasks, models perform better in predicting concrete entities such as countries, locations, and routes. However, their ability to understand abstract concepts like time and potential threats significantly decreases. We attribute this to two main reasons. Firstly, LLMs are typically trained on data that contains a substantial amount of content related to specific entities. Information about countries, locations, and routes frequently appears in news, books, and other texts, allowing models to learn relatively accurate and specific knowledge from these rich contexts. This enables the models to rely on an extensive knowledge base and clear semantic associations for reasoning and judgment when predicting these concrete entities. Secondly, abstract concepts like time and potential threats often involve complex logical reasoning and background knowledge, posing a greater challenge for LLMs. The expression of abstract concepts is varied and often lacks clear context, making it difficult for models to accurately capture their meanings. Additionally, abstract concepts such as potential threats require implicit background information and multi-layered logical reasoning, which may be insufficiently represented or ambiguously presented in the training

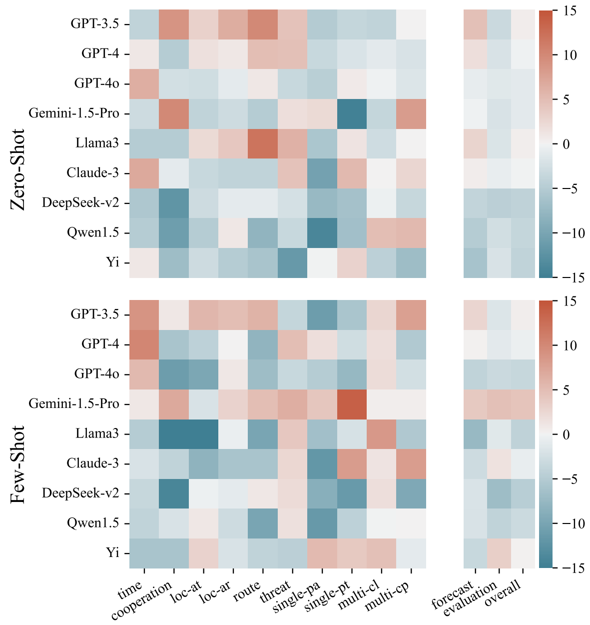

Figure 5: ΔScore between the CoT and standard prompting. **Top**: zero-shot results, **Bottom**: few-shot results, **Right**: averaged variation in two sub-tasks.

Model	Time	Coop.	Location	Route	Threat	Single	Multi	Avg.
GPT-3.5	43.0	51.0	47.2	51.0	46.1	66.6	36.5	49.6
GPT-4	52.0	71.0	67.8	68.0	43.4	81.7	57.2	64.6
GPT-4o	48.5	66.5	71.7	66.0	56.4	87.8	56.1	67.5
Llama3	50.0	63.0	58.3	58.0	42.0	66.7	36.6	52.3
DeepSeek-V2	51.6	66.1	58.8	62.0	41.9	78.4	41.5	57.4
Qwen1.5	51.0	54.0	59.0	66.0	34.1	76.1	41.9	55.3

Table 3: Results on SA-Bench's sub-tasks. The scores for each task are derived from either accuracy or the average of EM and F1 scores. Scores below the overall average are highlighted in green, while scores above the overall average are highlighted in red.

data, increasing the difficulty for understanding.[2]

8 Related Work

8.1 Situational Forecast

In recent years, the importance of situational forecast has been highlighted in multiple fields (Munir et al., 2022; Zhang et al., 2023; Shao et al., 2021; Stavroulakis and Papadimitriou, 2017). Early research (Adya et al., 2000) demonstrated the possibility of using rules and pattern recognition to make prediction lacking domain knowledge. Technologically, numerous studies (Jin et al., 2020a; Ju et al., 2024; Cheng et al., 2024; Xie et al., 2024; Fernandes, 2024; Ayoub et al., 2024) have been devoted to developing new predictive and processing models using neural networks. Moreover, next-generation computational methods (Chepuri et al., 2024) have also been applied in predicting dynamic systems. Therefore, the significance of deep learning (Jin et al., 2020b; Cheng et al., 2024) and large language models (Halawi et al., 2024) in understanding and predicting complex events has become increasingly evident.

8.2 Large Language Model

Recently, LLMs have showcased remarkable performance among various NLP tasks (Zhao et al., 2023; Yang et al., 2024). Research indicates that using zero-shot and few-shot examples (Brown et al., 2020; Kojima et al., 2022) as well as the chain-of-thought prompting method can significantly enhance performance on new, complex tasks without requiring fine-tuning (Wei et al., 2022; Kojima et al., 2022; Sahoo et al., 2024). In the realm of reasoning, LLMs have achieved remarkable results in various complex tasks such as mathematical reasoning (Cobbe et al., 2021; Mishra et al., 2022) and logical reasoning (Yu et al., 2020; Ho et al., 2022; Friedman, 2023).

These efforts demonstrate that LLMs possess the foundation to handle increasingly complex and dynamic situational awareness tasks.

9 Conclusion

Situational awareness is fundamental for effective decision-making and response to dynamic situations, yet the capabilities of LLMs in this area remain largely unexplored. To fill this gap, we introduce SA-Bench, a comprehensive benchmark designed to assess LLMs across three levels of situational awareness: environment perception, situation comprehension, and future projection. We evaluates various advanced LLMs, such as GPT-4, Claude-3, LLaMA3, and Qwen1.5, revealing significant gaps in their abilities compared to humans. Additionally, we provide a further analysis about obstacles and challenges in situational awareness, providing insights for future research.

Limitations

We present SA-bench, a comprehensive evaluation benchmark for Situation Awareness (SA), encompassing tasks across three levels from two categories, assessing various capabilities such as information extraction, multi-level knowledge reasoning, and plausible inference. Despite its comprehensive assessment situational awareness capabilities of LLMs, there are still some limitations:

Problem Design SA is a broad concept, encompassing research directions in diverse fields such as military operations, aviation, and cybersecurity. In terms of problem design, we have only selected a limited number of topics, primarily constructing multi-level knowledge bases and related scenarios from the perspective of maritime vessels. This approach introduces certain design limitations. In future work, we plan to design SA-related evaluation problems from a broader range of disciplines to ensure the comprehensiveness of SA-bench.

Task Segmentation SA tasks are typically divided into three levels. For the first level, element perception, considering the close relationship between information acquisition and understanding, SA-bench does not separately evaluate the model's information extraction capabilities but instead incorporates Levels 1 & 2 into the *Evaluation* subtask. Although this approach ensures the coherence of knowledge reasoning, it does not allow for the independent quantification of the model's information extraction capabilities. In future research, we will collect relevant evaluation data for Level 1 to enhance the granularity of task segmentation in SA-bench.

Task Formats To balance the difficulty of different SA sub-tasks, SA-bench primarily employs Multiple Choice Questions (MCQ) and Multiple Selection Questions (MSQ). These predefined options limit the model's divergent thinking abilities

[2] We provide an error analysis in Appendix A.6.

and the evaluation metrics of experiments. In future iterations, we plan to introduce tasks such as Constrained Text Generation and Free-form Reading Comprehension to assess the model's performance under increased task openness.

Chain of Thought Prompting In few-shot scenarios, all examples are manually set. When selecting representative question-reasoning pairs, the diversity of questions and subjective human judgment make it challenging to ensure that the reasoning paths are sufficiently detailed and effective for all questions in the task. This limitation poses a challenge to fully realizing the potential of few-shot CoT prompting in SA-bench. In future research, we will focus on providing natural language reasoning processes tailored for model learning for each question, thereby maximizing the task value of SA-bench.

Ethical Considerations

The news materials and technical data used for our benchmark are open-source, containing no classified or personal information. All annotated content has been manually verified to ensure accuracy and prevent any potential harm to individuals or organizations. The experimental data should be only used for analysis/research purposes.

Acknowledgement

The research in this article is supported by the National Science Foundation of China (U22B2059, 62276083).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Adya, J. Scott Armstrong, Fred Collopy, and Miles Kennedy. 2000. An application of rule-based forecasting to a situation lacking domain knowledge. *International Journal of Forecasting*, 16:477–484.

Anthropic. 2024. Introducing the next generation of claude.

Omran Ayoub, Davide Andreoletti, Aleksandra Knapiska, Róa Gocie, Piotr Lechowicz, Tiziano Leidi, Silvia Giordano, Cristina Rottondi, and Krzysztof Walkowiak. 2024. Liquid neural network-based adaptive learning vs. incremental learning for link load prediction amid concept drift due to network failures. *arXiv preprint arXiv:2404.05304*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mingyue Cheng, Jiqian Yang, Tingyue Pan, Qi Liu, and Zhi Li. 2024. Convtimenet: A deep hierarchical fully convolutional model for multivariate time series analysis. *Information Fusion*.

Ravi Chepuri, Dael Amzalag, Thomas Jr. Antonsen, and Michelle Girvan. 2024. Hybridizing traditional and next-generation reservoir computing to accurately and efficiently forecast dynamical systems. *arXiv preprint arXiv:2403.18953*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Mica R Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1):32–64.

Mica R Endsley. 2012. Situation awareness. *Handbook of human factors and ergonomics*, pages 553–568.

Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors*, 59(1):5–27.

Mica R Endsley, Daniel J Garland, et al. 2000. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21.

Pedro Afonso Fernandes. 2024. Forecasting with neuro-dynamic programming. *arXiv preprint arXiv:2404.03737*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Martin L Fracker. 1988. A theory of situation assessment: Implications for measuring situation awareness. In *Proceedings of the Human Factors Society Annual Meeting*, volume 32, pages 102–106. SAGE Publications Sage CA: Los Angeles, CA.

Robert Friedman. 2023. Large language models and logical reasoning. *Encyclopedia*, 3(2):687–697.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2022. What happens next? future subevent prediction using contextual hierarchical lstm. *Proceedings of the AAAI Conference on Artificial Intelligence*.

G. Jin, Q. Wang, C. Zhu, Y. Feng, J. Huang, and J. Zhou. 2020a. Addressing crime situation forecasting task with temporal graph convolutional neural network approach. In *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 1–6.

Guangyin Jin, Qi Wang, Cunchao Zhu, Yanghe Feng, Jincai Huang, and Xingchen Hu. 2020b. Urban fire situation forecasting: Deep sequence learning with spatio-temporal dynamics. *Applied Soft Computing*, 97:106730.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2020c. Forecastqa: A question answering challenge for event forecasting with temporal text data. *arXiv preprint arXiv:2005.00792*.

Wei Ju, Yusheng Zhao, Yifang Qin, Siyu Yi, Jingyang Yuan, Zhiping Xiao, Xiao Luo, Xiting Yan, and Ming Zhang. 2024. Cool: A conjoint perspective on spatio-temporal graph neural network for traffic forecasting. *Information Fusion*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Shangwen Lv, Wanhui Qian, Longtao Huang, Jizhong Han, and Songlin Hu. 2019. Sam-net: Integrating event-level and chain-level attentions to predict what happens next. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 6802–6809.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, et al. 2023. Universal ner: A gold-standard multilingual named entity recognition benchmark. *arXiv preprint arXiv:2311.09122*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.

Arslan Munir, Alexander Aved, and Erik Blasch. 2022. Situational awareness: Techniques, challenges, and prospects. *AI*, 3(1):55–77.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Wei Shao, Arian Prabowo, Sichen Zhao, Piotr Koniusz, and Flora D. Salim. 2021. Predicting flight delay with spatio-temporal trajectory convolutional network and airport situational awareness map. *Neurocomputing*.

Peter J. Stavroulakis and Stratos Papadimitriou. 2017. Situation analysis forecasting: the case of european maritime clusters. *Maritime Policy & Management*, 44(6):779–789.

Qwen Team. 2024. Introducing qwen1.5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wanli Xie, Ruibin Zhao, Zhenguo Xu, and Tingting Liang. 2024. Grey-informed neural network for time-series forecasting. *Computer Science > Machine Learning*.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Lena Zellinger, Andreas Stephan, and Benjamin Roth. 2024. Counterfactual reasoning with knowledge graph embeddings. *arXiv preprint arXiv:2403.06936*.

Zefan Zeng, Qing Cheng, and Yuehang Si. 2023. Logical rule-based knowledge graph reasoning: A comprehensive survey. *Mathematics*, 11(21):4486.

Wen Zhang, Yajing Xu, Peng Ye, Zhiwei Huang, Zezhong Xu, Jiaoyan Chen, Jeff Z Pan, and Huajun Chen. 2024. Start from zero: Triple set prediction for automatic knowledge graph completion. *IEEE Transactions on Knowledge and Data Engineering*.

Xiaojian Zhang, Xilei Zhao, Yiming Xu, Daniel Nilsson, and Ruggiero Lovreglio. 2023. Situational-aware multi-graph convolutional recurrent network (sa-mgcrn) for travel demand forecasting during wildfires. *AI*, 3(1):1–32.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Appendix

A.1 Examples

We provide several examples of different sub-task in SA-Bench. Tables 7 and 8 present examples of the *Evaluation* sub-tasks, and Table 9 shows examples of the *Forecast* sub-tasks.

A.2 Fleiss' Kappa

Fleiss' Kappa is a statistical measure for assessing the reliability of agreement between multiple annotators on categorical items (Fleiss, 1971). It extends Cohen's Kappa (Cohen, 1960), making it suitable for measuring agreement among more than two annotators. Below are the steps for calculating Fleiss' Kappa when annotating a dataset:

Data Preparation

- Assume there are N questions, each with k possible options, and n annotators who have labeled the answers for each question.

Calculate the Observer Agreement for Each Question P_i

- For each question i, count the number of annotations for each option n_{ij} (i.e., the number of times the j-th option is selected for the i-th question).
- Calculate the observer agreement for question i:

$$P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \right) \quad (1)$$

  where n is the number of annotators, and k is the number of options.

Calculate the Overall Observer Agreement $\bar{P}$

- Compute the mean agreement across all questions:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \quad (2)$$

Calculate the Expected Agreement P_e

- Determine the proportion of all annotations that fall into each category across all questions p_j:

$$p_j = \frac{\sum_{i=1}^{N} n_{ij}}{N \times n} \quad (3)$$

  For questions with fewer options, treat the counts for non-existent options as 0.

Fleiss' Kappa Value	Interpretation
< 0	Poor agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Table 4: General interpretation of Fleiss' Kappa values.

- Calculate the expected agreement:

$$P_e = \sum_{j=1}^{k} p_j^2 \quad (4)$$

Calculate Fleiss' Kappa

- The formula for Fleiss' Kappa is:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e} \quad (5)$$

  where $\bar{P}$ is the overall observed agreement, and P_e is the expected agreement.

Following these steps, the resulting κ value represents the quality of our dataset annotation. The value ranges from -1 to 1. Table 4 provides a general interpretation of the κ values (Landis and Koch, 1977).

A.3 Details of Models

GPT-3.5-turbo (Ouyang et al., 2022) GPT-3.5, developed by OpenAI, utilizes the transformer architecture (Vaswani et al., 2017) and is trained on diverse internet text to generate coherent and contextually relevant responses (Brown et al., 2020). We use *GPT-3.5-Turbo* via the OpenAI API, which supports a maximum input length of 16K tokens.

GPT-4/GPT-4o (Achiam et al., 2023) Upgraded from the ChatGPT baseline, GPT-4 excels in text and image understanding and generation, performing at a human level on various benchmarks. GPT-4o, an optimized version, offers enhanced performance for specific tasks. We utilize *GPT-4-Turbo-2024-04-09* and *GPT-4o-2024-05-13* through the OpenAI API, with a maximum input of 128K tokens.

Gemini-1.5-Pro (Reid et al., 2024) is a LLM developed by Google, capable of processing over one million tokens.

Llama3 (Touvron et al., 2023) Our experiments employ *llama3-70b-8192* through the GroqCloud API[3], featuring 70 billion parameters and supporting 8K context inputs.

Claude-3 We use *Claude-3-Sonnet-20240229*[4], which has 3 trillion parameters and supports up to 200K context inputs.

DeepSeek-V2 (Bi et al., 2024) We access *deepseek-chat-v2*[5]. The model has 236 billion parameters, with the open-source version supporting 128K context inputs and the API version supporting 32K context inputs.

Qwen1.5 (Bai et al., 2023; Team, 2024) We utilize *Qwen1.5-72B-chat* through Alibaba Cloud's DashScope[6], supporting up to 32K context inputs.

Yi (Young et al., 2024) We access the *yi-large-turbo* model via the ZeroOne Universe platform[7], supporting a context length of up to 16K tokens.

A.4 Methodology

Standard Prompting Our evaluation tasks consist of both Multiple Choice Question (MCQ) and Multi-Select Question (MSQ) based on the input text, using the standard input-output format. The prompt format is presented below, where I represents *instruction*, C represents *context*, Q represents *question*, and O represents *options*.

$$prompt^{sp} = \{I\}\,\{C\}\,\{Q\}\,\{O\} \tag{6}$$

Chain-of-Thought Prompting We utilize identical instructions for standard prompting. In the zero-shot chain-of-thought scenario, as described by Kojima et al. (2022), we append a reasoning trigger prompt "*Let's think step by step*" after the instructions to enable CoT reasoning. Our prompt format will be designed as below where T refers to *trigger*.

$$prompt_{zs}^{cot} = \{I\}\,\{T\}\,\{C\}\,\{Q\}\,\{O\} \tag{7}$$

For the few-shot CoT scenario, we manually design rationales for each task to demonstrate step-by-step reasoning to the model.

$$prompt_{fs}^{cot} = \{I\}\,\{C\}\,\{Q_1\}\,\{O_1\}\,\{R_1\}\,..\,\{Q\}\,\{O\} \tag{8}$$

[3]https://console.groq.com/docs/models
[4]https://www.anthropic.com/news/claude-3-family
[5]https://platform.deepseek.com/api-docs/zh-cn/
[6]https://help.aliyun.com/zh/dashscope/developer-reference/tongyi-qianwen-7b-14b-72b-quick-start
[7]https://platform.lingyiwanwu.com/docs

Error Category	**Ratio**
Extracting incorrect information	0.22
Unit conversion mistake	0.08
Lack of commonsense knowledge	0.15
Over-Reasoning	0.26
Context understanding errors	0.29

Table 5: Error categories and corresponding ratios in SA-bench. The error samples are obtained through random sampling.

A.5 Prompts

Our standard prompts are shown in Figure 6. The few-shot demonstrations for each task will be presented from Figure 7 to 15.

A.6 Error Analysis

As illustrated in Table 5, we manually analyze 100 errors categorize them into five categories.

Extracting incorrect information. 22% of errors occur in information extraction. For example, when a question involves the maximum distance an aircraft can operate from the carrier, the required information is "Combat Radius", but the model incorrectly extracts "Maximum Range". This may happen because the model fails to distinguish between similar knowledge and does not understand the implicit meaning of "operating" as participating in combat, overlooking that the payload during a mission affects the flight distance.

Unit conversion mistake. 8% of errors arise from confusion over units conversion. For instance, the model might misinterpret *nautical miles* as *kilometers*, leading to calculation errors.

Lack of commonsense knowledge. 15% of errors are due to the model's lack of basic temporal or geographical common sense. For example, the model might confuse "9 AM" with "9 PM" due to context omission or include overly distant locations when selecting a route, indicating that the model sometimes lacks the grasp of common knowledge necessary to make accurate judgments.

Over-Reasoning. 26% of errors result from the model over-reasoning, leading to answers beyond the actual scope of the question. For instance, in evaluation tasks, when calculating the number of targets a ship can hit, the model only needs to count the number of missiles equipped but instead considers additional factors like missed targets, resulting in incorrect answers.

Context understanding errors. 29% of errors stem from the model misinterpreting or omitting context. The model might miss key information in long texts or fail to correctly link information across different paragraphs. For example, when predicting the time an event will occur, the model might overlook contextual constraints mentioned earlier in the text , such as *within two years*, leading to inaccurate responses.

A.7 Management and Costs of Annotation Tasks

We hired 3 experts to design question templates and verify the final answers. We recruited 120 volunteers for data annotation. The cost for each data entry is around $2. The entire annotation process took two months to complete.

Multiple Choice Question
I will provide you with some text. Please read the text and answer the questions based on the content of the text. After each question, there will be several options, only one of which is correct. Please select the most correct option from the ones given. Text Material:{ } Question:{ } Options:{ }

Multi-Select Question
I will provide you with some text. Please read the text and answer the questions based on the content of the text. After each question, there will be several options, at least one of which is correct. Please select all potentially correct options from the ones given. Text Material:{ } Question:{ } Options:{ }

Figure 6: Zero-shot instructions and input formats for two kinds of questions.For zero-shot CoT prompting,we put the reasoning trigger *Let's think step by step* after our instruction and add *Provide your reasoning process* to ask for explicit process.

CoT Demonstration of Time Forecast (3-shot)	
Context: {News} ...... **Question**:When is the newly procured naval ship expected to be in service? **Options**:A.Mid-2024 B.End of 2024 C.Mid-2025 D.End of 2025 **Answer**:According to the context provided, the country's navy was authorized to procure 10 new ships in the 2023 National Defense Authorization Act.Typically, the construction and testing of new ships take around two years. If the procurement began in 2023, the completion and service entry time would be around the end of 2025. Therefore, the most appropriate answer is D.End of 2025. **Question**:At what time is the attacked merchant ship expected to resume sailing on the same day? **Options**:A.10:00 AM B.12:00 PM C.3:00 PM D.8:00 PM **Answer**:According to the context provided, the attack occurred at 5:00 AM on March 15, 2024. The article mentions that the estimated repair time is about 5 hours. Therefore, the most reasonable time forthe ship to resume sailing would be 10:00 AM on the same day. The correct answer is A.10:00 AM.	**上下文**: {新闻} ...... **问题**: 新采购的海军舰艇预计何时投入使用? **选项**: A.2024 年中 B.2024 年末 C.2025 年中 D.2025 年末 **答案**: 根据提供的上下文，国家海军被授权在 2023 年《国防授权法案》中采购 10 艘新舰艇。通常，新舰艇的建造和测试大约需要两年时间。如果采购从 2023 年开始，那么完成和投入使用的时间大约在 2025 年末。因此，最合适的答案是 D. 2025 年末。 **问题**: 被袭击的商船预计在同一天何时恢复航行? **选项**: A. 上午 10:00 B. 中午 12:00 C. 下午 3:00 D. 晚上 8:00 **答案**: 根据提供的上下文，袭击发生在 2024 年 3 月 15 日上午 5:00。文章提到，预计维修时间大约是 5 小时。因此，船只在同一天恢复航行的最合理时间是上午 10:00。正确答案是 A. 上午 10:00。

Figure 7: Chain-of-thought demonstrations of Time Forecast.We present a bilingual example in both Chinese and English.

CoT Demonstration of Cooperation Forecast (3-shot)	
Context: {News} ...... **Question**:Which countries and organizations are likely to be the main partners in the new international maritime cooperation initiative launched at the 2023 "Belt and Road" summit? **Options**:A. Egypt, United Nations B. India, World Bank C. Brazil, Asian Development Bank D. South Africa, World Wildlife Fund **Answer**:According to the fifth news item in the context, the international maritime cooperation initiative focuses on the sustainable use of ocean resources and biodiversity conservation. As announced by China at the "Belt and Road" summit, the main partners include African countries and global organizations. Egypt, as an important country along the "Belt and Road," especially with the strategic Suez Canal, is a natural maritime cooperation partner. The United Nations plays a significant role in global environmental and sustainable development issues. Involving the UN can enhance the international recognition and implementation of the initiative. India, Brazil, and South Africa have less cooperation with the "Belt and Road" initiative, while the World Bank and Asian Development Bank primarily focus on economic development. Therefore, the most appropriate answer is A. **Question**:Which countries are most likely to participate in the AUMX maritime exercise? **Options**:A. Thailand B. Indonesia C. Cambodia D. China **Answer**:According to the first news item in the context, the 2019 AUMX exercise involved ASEAN member countries and the United States, focusing on enhancing maritime security cooperation. The exercise was conducted in the waters between Thailand and Singapore. From this, it can be inferred that the most likely participants are ASEAN member countries, particularly those near the exercise area. Therefore, the most likely answer is A.	**上下文**: {新闻} ...... **问题**: 在 2023 年“一带一路”峰会上启动的新国际海事合作倡议中，哪些国家和组织最有可能成为主要合作伙伴? **选项**: A. 埃及，联合国 B. 印度，世界银行 C. 巴西，亚洲开发银行 D. 南非，世界自然基金会 **答案**: 根据上下文中的第五条新闻，国际海事合作倡议的重点是海洋资源的可持续利用和生物多样性保护。中国在“一带一路”峰会上宣布，主要合作伙伴包括非洲国家和全球组织。埃及作为“一带一路”沿线的重要国家，尤其是具有战略意义的苏伊士运河，是自然的海事合作伙伴。联合国在全球环境和可持续发展问题上发挥着重要作用，参与联合国可以增强该倡议的国际认可度和实施力度。印度、巴西和南非与“一带一路”倡议的合作较少，而世界银行和亚洲开发银行主要关注经济发展。因此，最合适的答案是 A。 **问题**: 哪些国家最有可能参加 AUMX 海上演习? **选项**: A. 泰国 B. 印度尼西亚 C. 柬埔寨 D. 中国 **答案**: 根据上下文中的第一条新闻，2019 年的 AUMX 演习涉及东盟成员国和美国，重点是加强海上安全合作。演习在泰国和新加坡之间的海域进行。因此，可以推断最有可能的参与者是东盟成员国，特别是靠近演习区域的国家。因此，最可能的答案是 A。

Figure 8: Chain-of-thought demonstrations of Cooperation Forecast.We present a bilingual example in both Chinese and English.

CoT Demonstration of Location Forecast (3-shot)	
Context: {News} ...... **Question**:Which ports are primarily connected by the new maritime freight route? **Options**:A. New York, Vancouver, Cancún B. Los Angeles, Montreal, Veracruz C. Miami, Toronto, Acapulco D. Houston, Halifax, Mexico City **Answer**:According to the news, in 2024, the NAFTA member countries—the United States, Canada, and Mexico—launched a new maritime freight route aimed at optimizing trade logistics in North America. The major ports are typically chosen based on the density of economic and trade activities. Los Angeles, Montreal, and Veracruz are key trade ports in North America, aligning with the goal of optimizing trade logistics. Therefore, the correct answer is B. **Question**:Where is the joint maritime exercise between India and ASEAN most likely to take place? **Options**:A. Black Sea B. South China Sea C. Red Sea D. Indian Ocean **Answer**:According to the context, during the first joint maritime exercise (AIME) held by India and ASEAN in 2023, the port phase of the exercise took place at Singapore's Changi Naval Base. Geographically, the maritime phase is most likely to take place in the South China Sea, which is a significant strategic region. This aligns with the goal of India and ASEAN to strengthen maritime security cooperation. Therefore, the most likely answer is B.	**上下文**: {新闻} ...... **问题**: 新的海运货运路线主要连接哪些港口? **选项**: A. 纽约，温哥华，坎昆 B. 洛杉矶，蒙特利尔，韦拉克鲁斯 C. 迈阿密，多伦多，阿卡普尔科 D. 休斯顿，哈利法克斯，墨西哥城 **答案**: 根据新闻，2024 年，北美自由贸易协定（NAFTA）成员国——美国、加拿大和墨西哥——推出了一条新的海运货运路线，旨在优化北美的贸易物流。主要港口通常根据经济和贸易活动的密度来选择。洛杉矶、蒙特利尔和韦拉克鲁斯是北美的重要贸易港口，符合优化贸易物流的目标。因此，正确答案是 B。 **问题**: 印度和东盟之间的联合海上演习最有可能在哪里进行? **选项**: A. 黑海 B. 南海 C. 红海 D. 印度洋 **答案**: 根据上下文，2023 年印度和东盟举行的首次联合海上演习（AIME）期间，演习的港口阶段在新加坡樟宜海军基地进行。从地理位置来看，海上阶段最有可能在南海进行，这是一个具有重要战略意义的地区。这与印度和东盟加强海上安全合作的目标一致。因此，最有可能的答案是 B。

Figure 9: Chain-of-thought demonstrations of Location Forecast.We present a bilingual example in both Chinese and English.

CoT Demonstration of Route Forecast (3-shot)

Context: {News}
......
Question:In December 2025, Holland America Line's MS Koningsdam will conduct a series of cruises. Which itinerary includes a round trip from Fort Lauderdale to San Juan?
Options:A. 21-day, Eastern and Western Caribbean, Mexico B. 11-day, Eastern Caribbean Windward Islands C. 14-day, Eastern Caribbean Bahamas and San Juan Holiday D. 10-day, Eastern Caribbean Antilles Grand Island
Answer:Based on the options and the background news, the 21-day itinerary in option A is too long and involves more destinations than just a round trip. Options B and D, the 11-day and 10-day itineraries, do not mention the Bahamas and San Juan. The 14-day itinerary in option C explicitly includes the Bahamas and a round trip to San Juan. Therefore, the correct answer is C.

Question:In December 2025, the Diamond Princess plans to depart from Singapore for a Southeast Asia cruise. Which ports are most likely included in this itinerary?
Options:A. Singapore, Malaysia, Vietnam, Thailand B. Singapore, Philippines, Indonesia, Thailand C. Singapore, Myanmar, Cambodia, Vietnam D. Singapore, India, Sri Lanka, Maldives
Answer:The Diamond Princess's Southeast Asia itinerary typically includes major countries and cities in Southeast Asia. Based on geographic knowledge and background news, the Diamond Princess will travel from south to north on a short journey. Option B's route from Singapore to the Philippines is too far. Option C includes Myanmar and Cambodia, which are not on the same route. Option D's endpoint, the Maldives, is not part of Southeast Asia. Option A, which includes ports in Malaysia, Vietnam, and Thailand, is a typical Southeast Asia cruise route and the most reasonable answer. Therefore, the correct answer is A.

上下文: {新闻}
......
问题: 2025 年 12 月，荷美邮轮的 MS Koningsdam 将进行一系列航行。哪条行程包括从劳德代尔堡到圣胡安的往返航程?
选项: A. 21 天，东加勒比和西加勒比，墨西哥 B. 11 天，东加勒比迎风群岛 C. 14 天，东加勒比巴哈马和圣胡安假日 D. 10 天，东加勒比安的列斯大岛
答案: 根据选项和背景新闻，选项 A 中的 21 天行程太长，涉及的目的地不仅仅是往返航程。选项 B 和 D 的 11 天和 10 天行程中都没有提到巴哈马和圣胡安。选项 C 中的 14 天行程明确包含巴哈马和往返圣胡安。因此，正确答案是 C。

问题: 2025 年 12 月，钻石公主号计划从新加坡出发进行东南亚邮轮航行。哪些港口最有可能包含在这条航线上?
选项: A. 新加坡，马来西亚，越南，泰国 B. 新加坡，菲律宾，印尼，泰国 C. 新加坡，缅甸，柬埔寨，越南 D. 新加坡，印度，斯里兰卡，马尔代夫
答案: 钻石公主号的东南亚行程通常包括东南亚的主要国家和城市。根据地理知识和背景新闻，钻石公主号将从南向北进行短途旅行。选项 B 从新加坡到菲律宾的路线太远。选项 C 包含的缅甸和柬埔寨不在同一路线上。选项 D 的终点马尔代夫不属于东南亚。选项 A，包括马来西亚、越南和泰国，是典型的东南亚邮轮航线，也是最合理的答案。因此，正确答案是 A。

Figure 10: Chain-of-thought demonstrations of Route Forecast.We present a bilingual example in both Chinese and English.

<table>
<tr><th colspan="2">CoT Demonstration of Threat Forecast (3-shot)</th></tr>
<tr><td>

Context: {News}
......
Question:As mentioned in the news, which areas might pose higher risks when the cruise ship passes through the South China Sea region?
Options:A. Strait of Malacca B. English Channel C. Panama Canal D. Suez Canal
Answer:According to the context, the Strait of Malacca in the South China Sea region is considered high risk due to frequent pirate attacks, threatening vessels navigating this area. Therefore, the Strait of Malacca is the area with higher risks. The correct answer is A.

Question:For the merchant ships mentioned in the news, which areas could pose potential threats?
Options:A. Strait of Hormuz B. Strait of Malacca C. Panama Canal D. Suez Canal
Answer:According to the context, a 2022 report from the U.S. Maritime Administration indicates that the Strait of Hormuz is a strategic chokepoint, with military activities and political tensions being major threats, making it a potentially dangerous area for vessels. Additionally, the Strait of Malacca in the South China Sea region is listed as high risk due to frequent pirate attacks. Therefore, the potential threats to the merchant ships mentioned in the news include the Strait of Hormuz and the Strait of Malacca. The correct answers are A and B.

</td><td>

上下文: {新闻}
......
问题: 根据新闻提到的内容，邮轮经过南海区域时哪些区域可能会带来更高的风险?
选项: A. 马六甲海峡 B. 英吉利海峡 C. 巴拿马运河 D. 苏伊士运河
答案: 根据上下文，南海区域的马六甲海峡由于频繁的海盗袭击，被认为是高风险区域，对在此航行的船只构成威胁。因此，马六甲海峡是风险较高的区域。正确答案是 A。

问题: 对于新闻中提到的商船，哪些区域可能构成潜在威胁?
选项: A. 霍尔木兹海峡 B. 马六甲海峡 C. 巴拿马运河 D. 苏伊士运河
答案: 根据上下文，美国海事管理局 2022 年报告指出，霍尔木兹海峡是一个战略要地，军事活动和政治紧张局势是其主要威胁，使其对船只来说可能是危险区域。此外，南海区域的马六甲海峡因频繁的海盗袭击被列为高风险区域。因此，对新闻中提到的商船构成潜在威胁的区域包括霍尔木兹海峡和马六甲海峡。正确答案是 A 和 B。

</td></tr>
</table>

Figure 11: Chain-of-thought demonstrations of Threat Forecast.We present a bilingual example in both Chinese and English.

CoT Demonstration of Single-PA Evaluation (3-shot)

Context: {Technical Data}
......
Question:What is the maximum strike speed of the air defense weapons onboard HMS Prince of Wales?
Options:A. Mach 4+ B. Mach 2 C. Mach 1.6 D. 25 knots
Answer:According to the text under the "Weapons" section, the "Missiles" subsection mentions the "Evolved Sea Sparrow Missile" with a "technical specification" indicating a speed of "Mach 4+," which is the highest strike speed. Other systems, such as the "Aster anti-aircraft missile," have a speed of "greater than Mach 2," the F-35B Lightning II fighter has a "maximum speed" of "Mach 1.6+," and the "MK-15 Block 1B Phalanx Close-In Weapon System" has a "rate of fire" of "3000-4500 rounds/min." Since the "maximum speed" typically refers to the measure of speed and not the rate of fire, these options are incorrect. Therefore, the correct answer is A. Mach 4+.

Question:What is the maximum personnel transport distance of the JS Hyga?
Options:A. 860 kilometers B. 500 kilometers C. 1200 kilometers D. 700 kilometers
Answer:The MCH-101 mine countermeasure helicopter, one of the shipborne helicopters, has a maximum range of 860 kilometers. Considering that this helicopter's cabin is used for "transport," it can be inferred that this is the maximum personnel transport distance the JS Hyga can execute. Therefore, the correct answer is A. 860 kilometers.

上下文: {技术数据}
......
问题: 威尔士亲王号上防空武器的最大打击速度是多少?
选项: A. 4 倍音速以上 B. 2 倍音速 C. 1.6 倍音速 D. 25 节
答案: 根据“武器”部分下的“导弹”子部分，技术规格中提到“进化型海麻雀导弹”的速度为“4 倍音速以上”，这是最高的打击速度。其他系统，如“紫菀防空导弹”速度为“大于 2 倍音速”，“F-35B 闪电 II 战斗机”最大速度为“1.6 倍音速以上”，“MK-15 Block 1B 密集阵近程防御武器系统”的射速为“每分钟 3000-4500 发”。由于“最大速度”通常指速度而不是射速，因此这些选项是错误的。因此，正确答案是 A. 4 倍音速以上。

问题: 日向号直升机护卫舰的最大人员运输距离是多少?
选项: A. 860 公里 B. 500 公里 C. 1200 公里 D. 700 公里
答案: 机载的 MCH-101 扫雷直升机的最大航程为 860 公里。考虑到该直升机的机舱用于“运输”，可以推断这是日向号执行人员运输任务的最大距离。因此，正确答案是 A. 860 公里。

Figure 12: Chain-of-thought demonstrations of Single-Proactive Evaluation.We present a bilingual example in both Chinese and English.

CoT Demonstration of Single-PT Evaluation (3-shot)

Context: {Technical Data}
Question:What is the altitude range that the reconnaissance system on the USS Dwight D. Eisenhower can detect?
Options:A. 12,100 meters B. 35,000 meters C. 84,390 meters D. 45,720 meters
Answer:The radar data for the reconnaissance system mentions that the "AN/SPS-49(V)" radar is an "air search radar" with a "coverage range" indicating an "altitude" of "45,720 meters." Therefore, the correct answer is D. 45,720 meters.

Question:What is the maximum target detection accuracy error of the reconnaissance system on the USS Carl Vinson?
Options:A. 0.9 meters B. 1.5 meters C. 5 meters D. 10 meters
Answer:The description of the "AN/SPQ-9B" fire control radar in the text mentions "range accuracy: 0.9m ś 0.025%," indicating that its detection accuracy error is within 0.9 meters. Therefore, the correct answer is A. 0.9 meters.

Question:Considering the carrier-based early warning aircraft, what is the maximum reconnaissance distance of the Charles de Gaulle?
Options:A. 366 km B. 250 km C. 183 km D. 50 km
Answer:The text provides the operating ranges of multiple reconnaissance systems. The "DRBJ11D/E" radar has an air search range of "366 km," which is the furthest distance among all reconnaissance systems mentioned. Although the "E-2 Airborne Early Warning Aircraft" is also reconnaissance equipment, the text does not specify its exact range. Therefore, its reconnaissance capability cannot be directly added to the radar data. Thus, the maximum reconnaissance distance is based on the longest radar range provided in the text. The correct answer is A. 366 km.

上下文: {技术数据}

问题: 艾森豪威尔号的侦察系统可以检测的高度范围是多少?
选项: A. 12,100 米 B. 35,000 米 C. 84,390 米 D. 45,720 米
答案: 侦察系统的雷达数据提到"AN/SPS-49(V)"雷达是"空中搜索雷达",其"覆盖范围"指示"高度"为"45,720 米"。因此,正确答案是 D. 45,720 米。

问题: 卡尔 · 文森号的侦察系统的最大目标检测精度误差是多少?
选项: A. 0.9 米 B. 1.5 米 C. 5 米 D. 10 米
答案: 文本中对"AN/SPQ-9B"火控雷达的描述提到"距离精度:0.9 米 ś 0.025%",表明其检测精度误差在 0.9 米以内。因此,正确答案是 A. 0.9 米。

问题: 考虑舰载预警机的情况下,戴高乐号的最大侦察距离是多少?
选项: A. 366 公里 B. 250 公里 C. 183 公里 D. 50 公里
答案: 文本提供了多个侦察系统的操作范围。"DRBJ11D/E"雷达的空中搜索范围为"366 公里",这是提到的所有侦察系统中最远的距离。尽管"E-2 空中预警机"也是侦察设备,但文本未具体说明其确切范围,因此其侦察能力不能直接加到雷达数据中。因此,最大侦察距离基于文本中提供的最长雷达范围。正确答案是 A. 366 公里。

Figure 13: Chain-of-thought demonstrations of Single-Protective Evaluation.We present a bilingual example in both Chinese and English.

CoT Demonstration of Multi-CL Evaluation (3-shot)	
Context: {Technical Data} ...... **Question**:When the USS Carl Vinson and HMS Prince of Wales both use all their RIM-7 Sea Sparrow surface-to-air missiles and Aster anti-aircraft missiles to counter long-range incoming targets, how many targets can they intercept at most? **Options**:A. 32 targets B. 24 targets C. 36 targets D. 48 targets **Answer**:The USS Carl Vinson is equipped with three 8-cell RIM-7 Sea Sparrow missile launchers, carrying a total of 24 missiles. The HMS Prince of Wales is equipped with 4 to 8 Aster anti-aircraft missiles. Therefore, the maximum number of targets they can intercept together is 24 + 8 = 32 targets. The correct answer is A. 32 targets.	**上下文**: {技术数据} ...... **问题**: 当卡尔 · 文森号和威尔士亲王号都使用所有的 RIM-7 海麻雀防空导弹和紫菀防空导弹来对抗远程来袭目标时，它们最多能拦截多少目标? **选项**: A. 32 个目标 B. 24 个目标 C. 36 个目标 D. 48 个目标 **答案**: 卡尔 · 文森号装备有三个 8 单元的 RIM-7 海麻雀导弹发射器，总共携带 24 枚导弹。威尔士亲王号装备有 4 到 8 枚紫菀防空导弹。因此，它们最多可以共同拦截的目标数量是 24 + 8 = 32 个目标。正确答案是 A. 32 个目标。

Figure 14: Chain-of-thought demonstrations of Multi-Collaborative Evaluation.We present a bilingual example in both Chinese and English.

CoT Demonstration of Multi-CP Evaluation (3-shot)	
Context: {Technical Data} ...... **Question**:Compared to the six-barrel Sadral air defense system equipped on the Charles de Gaulle, which weapon system on the USS Dwight D. Eisenhower has a longer range? **Options**:A. RIM-7 Sea Sparrow B. RIM-116 RAM C. Phalanx CIWS MK15 D. Mistral air defense missile **Answer**:The Mistral air defense missile equipped on the Charles de Gaulle has a maximum range of 6 kilometers. The RIM-7 Sea Sparrow equipped on the USS Dwight D. Eisenhower has an operational range of 10 nautical miles (approximately 19 kilometers). Therefore, the RIM-7 Sea Sparrow on the USS Dwight D. Eisenhower has a longer range. The correct answer is A. RIM-7 Sea Sparrow.	**上下文**: {技术数据} ...... **问题**: 相比于装备在戴高乐号上的六管 Sadral 防空系统，艾森豪威尔号上的哪种武器系统射程更远? **选项**: A. RIM-7 海麻雀 B. RIM-116 RAM C. 密集阵 CIWS MK15 D. 西北风防空导弹 **答案**: 戴高乐号上的西北风防空导弹最大射程为 6 公里。装备在艾森豪威尔号上的 RIM-7 海麻雀导弹的作战射程为 10 海里（约 19 公里）。因此，艾森豪威尔号上的 RIM-7 海麻雀导弹射程更远。正确答案是 A. RIM-7 海麻雀。

Figure 15: Chain-of-thought demonstrations of Multi-Comparative Evaluation.We present a bilingual example in both Chinese and English.

Evaluation Metric	Forecast					Evaluation			
	Time	Cooperation	Location	Route	Threat	Single		Multi	
						Proactive	Protective	Collaborative	Comparative
κ	0.79	0.85	0.77	0.67	0.79	0.83	0.80	0.62	0.68

Table 6: Fleiss' Kappa values for each annotation task.

Single-PA

C:*{... "Ship Type": "Aircraft Carrier", "Ship Name Origin": "President Abraham Lincoln", ... "Carrier Aircraft": [{"Name": "F/A-18 'Hornet'", "Performance Data": {... "Combat Radius": "460 miles", ... }, ... }], ... }*

Q: *In full load condition, how far can the Lincoln's carrier-based fighter jets operate from the ship?*
O: <u>740 kilometers</u>, 350 kilometers, 900 kilometers, 1200 kilometers

Single-PT

C: *{... "Ship Type": "Aircraft Carrier", "Ship Name Origin": "Carl Vinson", "Weapons": [... , {"Missiles": [{"Model": "RIM-116 RAM", "Type": "Close-In Interceptor Missile", "Technical Specifications": {... "Maximum Speed": "Mach 2.0+ ", ... }}]}], ... }*

Q: *What is the maximum speed of incoming targets that the Carl Vinson's close-in defense system can intercept?*
O: <u>2.0+ Mach</u>, 58.3 km/h, 1.8 Mach, 648 km/h

Table 7: MCQ examples of Single Evaluation tasks.The correct answer is <u>underlined</u>.

Multi-CL

C: *Wales-{"Model":"Shirne Air Defense Missile",... ,"Quantity":"4-8 rounds",... }*
Carl Vinson-{"Model":"RIM-7 Sea Sparrow Missile",... ,"Quantity":"3 sets of 8",... }

Q:*During an exercise involving the Wales and the Carl Vinson, how many long-range aerial targets can they intercept at most?*
O:<u>32</u>, 24, 36, 48

Multi-CP

C: *deGaulle-{"System":"SADRAL","Missiles":[., {"Model":"Mistral",... ,"Effective Range":"6 km",... }...]*
Carl Vinson-{"Model":"RIM-7 Sea Sparrow Missile",... ,"Range":"10 nautical miles",... }

Q:*Compared to the SADRAL air defense system, which weapon system on the Carl Vinson has a longer range?*
O:<u>RIM-7 Sea Sparrow</u>, RIM-116 RAM, Phalanx CIWS MK15, Mistral Air Defense Missile

Table 8: MCQ examples of Multi Evaluation tasks.The correct answer is <u>underlined</u>.

Subtask	Question	Sentence
Time	*When will the next significant policy change in US-China trade relations occur?*	***Recent discussions*** *between trade representatives have highlighted the urgency of reaching a new agreement, with many experts predicting* ***a resolution by the end of Q1 2024****. (12/15/23)*
Cooperation	*Which countries are likely to form an alliance in response to recent geopolitical tensions in Asia by mid-2024?*	***Japan and South Korea*** *have held multiple high-level meetings in recent months to discuss mutual security concerns. (5/15/23)*
Location - Attack	*Where is the next significant cyber-attack likely to occur in 2024?*	***Security analysts*** *have identified* ***Silicon Valley*** *as a prime target due to the concentration of high-tech firms and recent* ***increased activity from hacker groups****. (10/22/23)*
Location - Arrive	*Where will the next major international summit be held in 2024?*	*The* ***United Nations*** *has announced plans to host its next major summit in* ***Geneva****, Switzerland. (12/01/23)*
Route	*What will be the primary trade route for oil shipments by 2025?*	*The new* ***pipeline project*** *will connect the* ***Caspian Sea*** *with* ***major ports*** *in the* ***Mediterranean****, passing through cities like* ***Baku*** *and* ***Ceyhan****. (11/18/23)*
Threat	*What is the most likely environmental threat to coastal cities in 2024?*	*Reports indicate that* ***rising sea levels*** *and* ***increased storm activity*** *pose significant risks to* ***coastal cities****. (10/10/23)*

Table 9: Examples of *Forecast* tasks.**Sentence** refers to sentences in the news corpus that provide inferential evidence for the corresponding question. The bolded parts are the inferential evidence.