Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 2, pp. 163-175, April 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.2.163

해양 AIS 데이터 분석을 위한 하이브리드 프롬프트 에이전트: 질문 유형 분류와 동적 라우팅을 통한 성능 최적화

김성진 · 김성일†

울산과학기술원 산업공학과

A Hybrid Prompt Agent for Maritime AIS Data Analysis: Performance Optimization through Query Classification and Dynamic Routing

Seongjin Kim · Sungil Kim

Department of Industrial Engineering, Ulsan National Institute of Science and Technology

While Large Language Model (LLM)-powered agents are transforming conversational data analysis, their efficacy on specialized datasets, such as the Automatic Identification System (AIS), is highly sensitive to prompt engineering strategies. Conventional static prompting approaches fail to adequately address the broad spectrum of user query complexity, which ranges from straightforward fact retrieval to multi-step analytical reasoning. To mitigate this limitation, we propose a Hybrid Prompt Agent architecture. This system employs a preliminary query classification mechanism to dynamically select and apply an optimal prompting template: utilizing compact prompts for high factual precision and Chain-of-Thought (CoT) prompts for in-depth analytical tasks. Empirical evaluation on a custom-built benchmark of 100 diverse AIS-related queries demonstrates that our hybrid approach achieves an effective balance between quantitative accuracy and qualitative contextual relevance compared to single-prompt baselines. These findings validate dynamic prompting architectures as a practical, effective, and resource-efficient alternative to full model fine-tuning for developing specialized domain analysis tools.

Keywords: LLM Agent, Prompt Engineering, AIS Data, Maritime Data Analysis

1. 서 론

선박 자동식별시스템(AIS)은 선박의 식별자(MMSI), 위치, 선속(SOG), 침로(COG) 등을 일정 주기로 송신하는 고차원 시계열 데이터로, 항로 최적화, 연료 효율 분석, 이상 탐지 등 다양한 해양 운영 과제 해결을 위한 필수적인 센서 인프라로 자리 잡았다(Kim *et al.*, 2017; Yang *et al.*, 2019). 그러나 AIS 데이터의 실제적 활용은 단순 조회를 넘어선다. 예를 들어, 해양 운항 관리자는 예측 불가능한 항로 이탈이나 비효율적인 연료 소모 패턴을 신속하게 파악해야 하지만, 방대한 시계열 데이터에서 이상 신호를 실시간으로 탐지하고 그 원인을 해석하는 데는 상당한 시간과 도메인 전문성이 요구된다. 기존의 대시보드나 SQL 기반 시스템은 정해진 규칙의 질의만 가능하여, “왜 이 선박의 속도가 특정 구간에서 급감했는가?”와 같은 탐색적이고 복합적인 질문에 즉각 답을 찾기 어렵다. 이러한 실무적 한계는 자연어 질의만으로 심층 분석까지 수행하는 고도화된 인터

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(No.RS-2023-00218913).
† 연락저자 : 김성일 교수, 44919 울산광역시 울주군 유니스트길 50 울산과학기술원 산업공학과, Tel : 052-217-3195,
E-mail : sungil.kim@unist.ac.kr
2025년 12월 1일 접수; 2026년 2월 8일 수정본 접수; 2026년 2월 13일 게재 확정.

페이스의 필요성을 부각시킨다(Lundström *et al.*, 2025).

1.1 LLM 에이전트를 활용한 대화형 데이터 분석의 부상

최근 대규모 언어모델(LLM)을 활용해 "질의, 분석 계획 수립, 코드 생성 · 실행, 결과 요약"을 일괄 수행하는 에이전트 방식이 주목받고 있다. 에이전트는 프롬프트(Prompt)를 기반으로 분석 계획을 세우고, Python 코드를 자율적으로 생성 · 실행하여 주어진 분석 작업을 완수한다. 예컨대 LangChain (LangChain Authors, 2025)과 같은 프레임워크에서 제공하는 Pandas Agent는 "가장 빠르게 운항한 선박은?", "MMSI 352058000의 최근 속력은?"과 같은 질문에 대해 자동으로 코드를 실행하여 결과를 제시함으로써, 비전문가의 데이터 분석 접근 장벽을 크게 낮추었다(LangChain Authors, 2025; LlamaIndex Authors, 2024). 다만 이러한 성과는 주로 단답형 질의에 집중되어 있으며, AIS 데이터의 복합 요구(여러 단계의 추론과 비교 분석)에는 한계가 따른다(Xie *et al.*, 2024). 해양 데이터 분석 맥락에서 성능을 높이려면 질문 의도에 부합하는 프롬프트 전략을 설계하고, 질의 시점에서 적절히 선택 · 적용하는 체계가 필요하다.

1.2 기존 프롬프트 연구의 한계

프롬프팅 연구는 크게 두 가지로 구분할 수 있다.

(i) 정적 단일 프롬프트: 모든 질문에 동일 구조의 프롬프트를 적용한다. 구현은 간단하지만, 단답형 사실 질의와 복합 추론 질의가 공존하는 환경에서는 성능 상충이 발생한다.

(ii) 적응형 프롬프팅: 질문 특성에 따라 CoT · 예시 · 역할 · 출력 제약 등을 질문별로 조정한다. 여기서 CoT는 최종 답변 이전에 중간 추론 단계를 단계적으로 생성하도록 유도하는 프롬프트 방식으로, 복잡한 추론 과제에서 효과가 보고되어왔다(Kojima *et al.*, 2022; Brown *et al.*, 2020). 한편 CoT의 유효성은 과제 유형 · 모델에 따라 달라지며, 일부 산술 과제에서는 명시적인 지시 없이도 자발적인 중간 추론이 생성되는 등 일관적이지 못한 현상이 관찰되었고 인스턴스 단위의 선택적 CoT 적용이 제안되었다(Schulhoff *et al.*, 2025). 또한, 표 · 텍스트 혼합 질의에서는 증거 검색 단계와 CoT를 결합하는 하이브리드 프롬프트가 보고되기도 했다(Kong *et al.*, 2024).

그러나 과업 목표를 기준으로 질문을 먼저 분류하고, 분류 결과에 따라 상이한 프롬프트를 라우팅하는 경량 구조를 도메인 실제 데이터로 체계 평가한 사례는 드물다. 본 연구는 AIS 질의응답에서 질문 유형 분류 후 목표에 따라 프롬프트를 동적으로 선택하는 아키텍처를 제안하고, 정량 · 정성 · 비용의 균형 관점에서 단일 프롬프트 대비 성능 향상을 검증한다.

1.3 연구 질문

본 연구는 다음의 핵심 질문을 설정한다. 먼저 각 질문 유형에 맞는 최적 프롬프트 구조를 규명(RQ1)하고, 이를 질문 유형 분류와 동적 라우팅으로 구현한 하이브리드 프롬프트 에이전트(HPA)가 단일 전략 대비 우수한지를 검증(RQ2)한다.

RQ1: 해양 AIS 데이터 분석에서, 과업 목표(정량적 정확성과 분석적 품질)에 따라 최적의 프롬프트 구조는 어떻게 다른가?

RQ2: 질문 유형 분류 기반 동적 라우팅은 정적 단일 프롬프트 전략 대비 정확도 · 분석 품질 · 비용의 균형을 유의하게 개선하는가?

1.4 본 연구의 기여

본 연구의 기여는 다음과 같다.

(i) 질문 유형별 프롬프트 효과 검증: 값-중심 프롬프트와 설명-중심 프롬프트를 유형별로 적용하여 정량 · 정성 지표로 효과를 비교하였고, 전자가 정량적 정확성에, 후자가 분석 품질에 각각 유리함을 밝혔다.

(ii) 하이브리드 프롬프트 에이전트: 질문을 분류해 목표에 최적인 프롬프트로 라우팅하는 구조를 설계 · 구현하고, 정량적 정확도와 정성적 분석 품질 간의 균형을 효과적으로 달성함을 입증하였다.

(iii) 전문 도메인 적용 가이드라인 제시: 범용 데이터 분석 에이전트를 해양 분야에 이식할 때, 질문 복잡성 인지와 동적 프롬프팅 적용만으로 모델 수정 없이 성능을 개선할 수 있음을 보이고 구체적 설계 요소를 제시하였다.

본 논문은 2장 선행연구, 3장 방법론, 4장 실험, 5장 결론 순으로 구성된다.

2. 선행연구

2.1 전통적 기계학습 및 심층학습 기반 분석

LLM 등장 이전의 AIS 데이터 분석 연구는 통계적 기법과 기계학습, 심층학습 모델을 활용한 정량적 패턴 인식에 집중되었다. 초기 연구들은 밀도 기반 클러스터링이나 커널 밀도 추정을 통해 주요 항로와 해상 교통 흐름을 분석하였으며(Pallotta *et al.*, 2013), 이후에는 RNN, LSTM, Transformer 등의 딥러닝 모델, VAE 기반의 Control Chart나 베이지안 부트스트랩 기법 등을 활용하여 경로 예측과 이상 운항을 정밀하게 탐지하려는 연구들이 주를 이루었다(Nguyen *et al.*, 2018; Park and Kim, 2020; Chen *et al.*, 2023; Oh and Kim, 2023; Oh *et al.*, 2024). 그러나 이러한 접근법은 복잡한 전처리와 모델 튜닝을 전제로 하며, 사전 정의된 분석 목적에 의존하는 구조로 인해 비전문가의 활용과 탐색적 질의응답 기반 분석으로의 확장에는 한계가 존재하였다.

2.2 대화형 데이터 분석과 LLM

전통적 분석 방식의 한계를 보완하기 위해, 최근에는 자연어 기반 대화형 데이터 분석이 새로운 연구 흐름으로 부상하고 있다. 초기 연구는 Text-to-SQL 방식에 기반하였으나(Sun *et al.*, 2025), 고정된 스키마와 제한된 표현력으로 인해 복합 분석을 충분히 지원하지 못하였다. 이후 LLM의 발전과 함께, 코드 생성과 실행을 통해 분석을 수행하는 에이전트 기반 방식이 등장하였으며, LangChain(LangChain Authors, 2025), LlamaIndex (LlamaIndex Authors, 2024) 등은 이러한 환경을 체계화하였다.

한편, LLM을 경로 예측이나 이상 탐지와 같은 특정 과업에 적용하는 연구도 보고되고 있으나(Park *et al.*, 2025), 이러한 접근은 사전 정의된 단일 과업에 특화되어 있어 사용자의 다양한 자연어 질의에 실시간으로 대응하는 대화형 분석 환경과는 성격이 다르다. 반면, Merten *et al.*(2025)은 자연어 기반 인터페이스 설계를 탐구하였다. 그러나 기존 연구들은 대부분 단일 프롬프트 또는 제한된 질의 구조에 의존하여, 다양한 분석 요구를 일관되게 처리하는 데 구조적 한계를 지니고 있다.

2.3 범용 에이전트의 한계: 정적 프롬프트 전략

대부분의 LLM 에이전트는 질문 유형과 무관하게 동일 구조의 프롬프트를 적용한다. 이러한 정적 전략의 한계에 대해 Xie *et al.*(2024)는 “모든 데이터셋 유형에 보편적으로 잘 통하는 단일 프롬프팅 전략은 있을 수 없다.”라고 지적하였다. AIS처럼 단순 조회부터 복합 분석까지 넓은 범위를 다루는 도메인에서는 정적 전략을 넘어서는 새로운 접근이 요구된다.

2.4 동적 · 하이브리드 프롬프트 아키텍처

CoT의 효과가 확인되면서(Kojima *et al.*, 2022; Brown *et al.*, 2020), 질문별로 CoT · 예시 · 역할 · 제약을 조정하는 적응형 접근이 제안되었다. 특히 표 기반 복합 추론에서는 증거 검색과 CoT 결합이 성능을 높이는 것으로 보고되었고, 텍스트-테이블 혼합 질의응답에서도 유사한 결합 전략이 탐색되었다(Wei *et al.*, 2022; Kong *et al.*, 2024).

특히, 테이블 형태의 정형 데이터를 다루는 맥락에서 Luo *et al.*(2023)은, 복잡한 데이터 환경에서 표준 CoT 프롬프트가 무관한 정보로 인해 잘못된 추론을 생성하는 문제를 지적했다. 이에 대한 해결책으로 테이블에서 관련 증거를 먼저 검색한 후 이를 CoT와 결합하는 하이브리드 전략을 제안하였다.

이러한 선행 연구들은 질문의 복잡성을 고려한 동적 · 하이브리드 아키텍처의 필요성과 효과를 입증하며 본 연구의 이론적 토대를 제공한다. 본 연구는 이러한 맥락을 확장하여, 질문의 과업 목표(정량적 정확성 대 분석 품질)를 사전 판별하고 그 결과에 따라 프롬프트를 동적으로 라우팅하는 구조를 제안한다. 구체적으로, 정답 일치가 핵심인 질의에는 값-중심 프롬프트(유형 A)를, 근거 제시나 해석이 중요한 질의에는 설명-중심 프롬프트(유형 B)를 적용한다. 라우팅은 연산자, 집계, 비교, 근거 요구 등 질문에 내재된 언어적 · 구조적 특징을 분석하는 경량 LLM 분류기를 통해 수행되며, 모델 구조를 변경하지 않고 프롬프트 선택만으로 정확성-설명성 간의 상충 관계를 완화하는 것을 목표로 한다.

2.5 프롬프트 기법과 평가 지표

본 절은 2.3의 아키텍처 논의와 구분하여, LLM 기반 질의응답에서 널리 쓰이는 프롬프트 기법과 평가 지표를 간략히 정리한다.

(1) 프롬프트 기법

LLM의 성능은 입력을 구성하는 프롬프트의 설계에 크게 의존하며, 이에 따라 프롬프트 엔지니어링은 LLM을 효과적으로 활용하기 위한 핵심 분야로 자리 잡았다. Schulhoff *et al.*(2025)은 프롬프트 기법이 모델의 추론 능력과 도메인 적응력을 극대화하는 데 필수적임을 보고한다. 본 연구는 이러한 기법들 중 LLM 에이전트 설계에 널리 적용되는 다음의 세 가지 주요 전략을 채택 및 활용하였다.

- Few-shot Learning: LLM에 소수의 정답 예시를 함께 제공하는 기법이다. 모델은 이 예시들을 In-Context Learning (Kim and Kim, 2023)의 근거로 삼아, 별도의 가중치 업데이트 없이도 특정 과업의 맥락과 원하는 출력 형식을 빠르게 학습한다(Brown *et al.*, 2020). 본 연구에서는 <Table 1>의 분류기 프롬프트가 이 기법을 활용하여 유형 A와 유형 B의 분류 기준을 명확히 학습하도록 설계되었다.
- Chain-of-Thought: 복잡한 추론 과제(산술 추론, 상식 추론 등)에서 LLM이 최종 답변을 바로 도출하는 대신, 중간 추론 과정을 단계별로 명시적으로 생성하도록 유도하는 프롬프트 패러다임이다(Kojima *et al.*, 2022). 이러한 단계적 사고 과정은 모델이 복잡한 문제를 논리적으로 분해하고 해결책에 접근하도록 보조하여, 최종 답변의 신뢰성과 정확성을 높인다(Wei *et al.*, 2022). 본 연구에서는 유형 B 질문에 대한 프롬프트(<Table 3>)에 CoT를 적용하여 분석 과정의 설득력을 강화했다.
- 역할 부여: “당신은 특정 분야의 전문가입니다”와 같이 LLM에게 구체적인 페르소나 또는 역할을 명시적으로 부여하는 전략이다. 역할 부여는 LLM이 해당 역할에 맞는 어조, 전문 지식, 응답 형식을 일관되게 유지하도록 제어하여, 생성 결과의 도메인 적합성과 품질을 향상시킨다(Kong *et al.*, 2024). 본 연구는 두 프롬프트(<Table 2>, <Table 3>) 모두에 ‘해양 데이터 분석가’ 또는 ‘전문가’라는 역할을 부여하여 답변의 전문성을 확보하고자 했다.

(2) 평가 지표

이러한 전략들의 효과를 측정하기 위한 평가 지표 또한 다각적으로 발전해 왔다. 생성된 답변의 정확도를 측정하기 위한 정량적 지표로는 EM, F1-score(Rajpurkar *et al.*, 2016), BLEU(Papineni *et al.*, 2002), ROUGE-L(Lin, 2004) 등이 활용되며, 답변의 논리적 타당성이나 신뢰도를 평가하기 위한 정성적 지표로는 LLM-as-a-judge 등이 사용된다(Zheng *et al.*, 2023).

3. 방법론(Hybrid Prompt Agent, HPA)

본 연구는 해양 AIS 데이터에 대한 자연어 질의응답 성능을 극대화하기 위해, 질문 유형에 따라 최적의 프롬프트를 동적으로 선택 및 적용하는 하이브리드 프롬프트 에이전트(HPA)를 제안한다. <Figure 1>은 HPA의 전체적인 구조를 나타낸다.

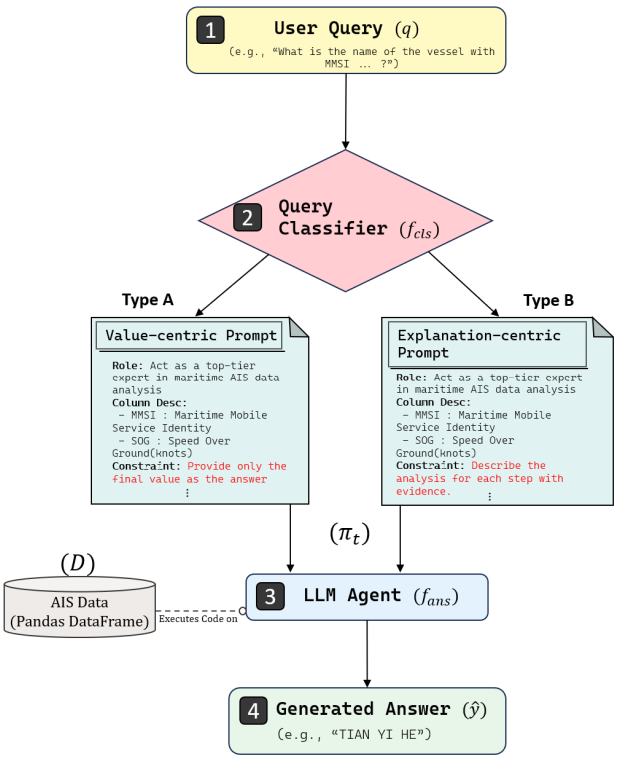

Figure 1. The Proposed Hybrid Prompt Agent (HPA): a classifier f_{cls} routes queries to type-specific prompts π_t; the agent f_{ans} generates/executes code on D and returns $\hat{y}$.

HPA는 질문 유형 분류 LLM인 f_{cls}와 그 질문에 답변하는 LLM 에이전트 f_{ans}으로 구성된다. 사용자 질문 q가 주어지면, 먼저 f_{cls} 가 질문의 유형 t를 판별하고, 해당 유형에 최적화된 프롬프트 템플릿 Φ_t를 선택하여 프롬프트 π_t를 생성한다. 최종적으로 f_{ans}은 생성된 프롬프트 π_t와 AIS 데이터 D를 입력받아 답변 $\hat{y}$를 도출한다. 이 과정은 다음과 같이 공식화할 수 있다.

$$t = f_{cls}(q) \in \{A,B\}$$
$$\pi_t = \Phi_t(q,S)$$
$$\hat{y} = f_{ans}(\pi_t;D)$$

여기서 D는 전처리된 AIS 데이터, S는 도메인 지식(컬럼 설명 등), Φ_t는 유형 t에 대한 프롬프트 템플릿이다.

Table 1. Prompt Template for the Question Type Classifier

Component	Content
Role & Instruction	# [System] You are a lightweight classifier that analyzes user questions into two types. - TYPE_A: Questions asking for a single value like a specific fact, number, or name (What, Who, Which, How many...). - TYPE_B: Questions requiring analysis or reasoning, such as reasons, methods, comparisons, or detailed explanations (Why, How, Compare...).
Examples	# [Examples] Q: What is the name of the vessel with MMSI 352058000? A: TYPE_A Q: Compare the operational patterns of the two vessels and explain the differences. A: TYPE_B
Question	# [Question] {{ question }}
Output	# [Output] - Only output 'TYPE_A' or 'TYPE_B'. - Do not add any other explanations.

3.1 질문 유형 분류

HPA 구조의 첫 단계는 사용자 질문의 근원적인 목표를 판별하는 것이다. 실제 AIS 데이터 분석 시나리오에서는 특정 값을 빠르고 정확하게 확인해야 하는 요구(특정 선박의 현재 속도 등)와, 현상의 원인에 대한 깊이 있는 해석을 요구하는 상황(운항 효율성 비교 분석 등)이 혼재한다. 단일 프롬프트 전략은 이 두 가지 상충된 목표(정확성과 설명력)를 동시에 만족시키기 어렵다는 본질적 한계가 있다. 따라서 본 연구에서는 질의를 답변의 최종 목표에 따라 두 가지 유형으로 분류하는 접근 방식을 채택했다.

<Table 1>은 f_{cls}가 사용하는 Few-shot 프롬프트 템플릿의 구성을 보여준다. 이 템플릿은 명확한 역할 정의("You are a lightweight classifier..."), 구체적인 분류 예시("Q: What is the name...; A: TYPE_A"), 간결한 출력 형식 지시("Only output

'TYPE_A' or 'TYPE_B'.")를 포함하여, LLM이 빠르고 정확하게 유형을 판단하도록 유도한다. 각 유형에 대한 구체적인 정의는 다음과 같다.

- 유형 A (값-중심 질의): 최종 답변으로 단일 값 또는 값의 목록을 간결하게 도출하는 것이 핵심 목표인 질문 유형으로 정의한다. 이 유형의 질문에 대한 평가는 정량 정확성에 초점을 맞춘다.
- 유형 B (설명-중심 질의): 단순한 값 도출을 넘어, 데이터에 기반한 분석 과정, 비교, 원인 추론 등 논리적 설명을 답변에 포함해야 하는 질문 유형으로 정의한다. 이 유형의 답변은 분석 품질을 핵심적으로 평가한다. 예를 들어, 특정 항로 선박들의 평균 속력을 묻더라도, 그 계산 과정과 통계적 의미에 대한 해석까지 함께 요구한다면 유형 B로 분류될 수 있다.

3.2 동적 프롬프팅 및 응답 생성

각 유형에 맞게 설계된 프롬프트를 LLM 에이전트에 선택적으로 적용하였다.

(1) 유형 A를 위한 값-중심 프롬프트(Value-Centric Prompt, Value-CP)

유형 A 질문의 목표는 단일 값 또는 값의 목록을 간결하게 도출하는 것, 즉 정량 정확성을 극대화하는 것이다. 이를 달성하기 위한 핵심은, 모델이 정답 값 이외의 장문의 설명이나 불필요한 부가 정보를 생성하여 정량적 평가의 정확도를 저해할 가능성을 방지하는 것이다. 따라서 <Table 2>에 제시된 값-중심 프롬프트는 모델이 불필요한 추론이나 부연 설명을 생성할 여지를 원천적으로 차단하도록 설계되었다. 이 프롬프트는 정량적・통계적 답변을 정확히 도출하는 해양 AIS 데이터 분석가(maritime AIS data analyst)로 비교적 중립적인 역할로 정의한다. 지시사항에서는 "Provide only the value as the answer, without any other text"라고 명시적으로 지시하여, 최종 출력을 단일 값으로 제한하고 결과의 재현성과 검증 가능성을 높였다.

Table 2. Key Instructions for the Type A Prompt

Component	Key Instructions
Role	Act as a maritime AIS data analyst. Derive the quantitative/statistical answer accurately from the entire DataFrame (df).
Context	Provided column descriptions defining quantitative thresholds (e.g., SOG < 1.0 knots for stationary status) for accurate state classification and a one-shot example (Q: ... ; A: ...).
Instruction	1. Provide only the value as the answer, without any other text. 2. The final result must be a single-line, short answer (e.g., a number or a string). 3. The answer should match the value retrieved from the df.

(2) 유형 B를 위한 설명-중심 프롬프트(Explanation-centric Prompt, Explanation-CP)

유형 B 질문의 목표는 단순한 값 도출을 넘어, 데이터에 기반한 분석 과정, 비교, 원인 추론 등 논리적 설명을 제공하는 것, 즉 논리적이고 설득력 있는 분석을 제공하는 것이다. 이를 위해 <Table 3>의 설명-중심 프롬프트는 CoT 기법을 핵심 전략으로 채택했다. 이 프롬프트는 <Table 2>와 명확히 대비된다. 역할부터 단순 분석가가 아닌 해양 AIS 및 운항 패턴 분석 최고 전문가로 격상시켜 더 깊이 있는 해석을 유도한다. 가장 큰 차이점인 지시사항에서는, 값만 출력하도록 제시했던 것과 정반대로 [Analysis Process]라는 항목을 명시하여, 복잡한 분석 과제를 논리적인 하위 단계로 분해하고 각 단계의 결론이 데이터에 기반하도록 강제한다. 또한, 해양 도메인의 특수성을 반영하여 리샘플링과 항차 구분 방법을 명시하였으며, 운항 상태의 자의적 해석을 방지하기 위해 '속력 1노트 미만'과 같은 실무적 임계값을 분석의 근거로 사용하도록 가이드하였다. 이를 통해 최종 답변의 신뢰성을 높였다.

Table 3. Key Instructions for the Type B Prompt

Component	Key Instructions
Role	Act as a top expert in maritime AIS and vessel operation pattern analysis. Provide a detailed explanation based on the data and cite evidence.
Context	Provided detailed column descriptions and operational criteria for vessel status judgement and voyage separation rules to support (e.g., "Separate voyages based on time ..." and "SOG < 1.0 is stationary"), along with a one-shot example (Q: ... ; A: ...).
Instruction	1. [Analysis Process]: First, list the 3-5 key steps to solve the problem. 2. [Detailed Explanation]: Describe the analysis for each step. All claims must be based on data, with evidence (values, stats) in parentheses. 3. [Final Summary]: Conclude with a 2-3 sentence summary.

3.3 LLM 에이전트의 작동 방식

본 연구에서 제안하는 HPA의 f_{ans}은 LangChain 프레임워크의 Pandas DataFrame Agent와 같은 LLM 기반 코드 생성 에이전트의 작동 방식을 따른다. 이 에이전트는 (1) 선택된 프롬프트(π_A 또는 π_B)와 (2) AIS 데이터가 담긴 Pandas DataFrame(D)을 입력으로 받는다. 그 후, 에이전트는 프롬프트의 지시에 따

라 질문을 해결하기 위한 Python 코드를 자율적으로 생성하고, 실행하여 DataFrame을 분석한다. 마지막으로, 코드 실행 결과를 바탕으로 최종 자연어 답변 $\hat{y}$를 생성하여 사용자에게 반환한다.

3.4 해양 질의 데이터셋

단순한 임의 질의만으로는 제안 모델의 강건성과 다면적 능력을 객관적으로 측정하기 어렵다. 이에 본 연구는 체계적이고 재현 가능한 평가를 수행하고자, 실제 AIS 데이터 분석 시나리오와 '인간 중심 - 탄소 중립 글로벌 공급망 연구센터'(https://scsc.pusan.ac.kr/) 소속 연구원들의 도메인 지식을 반영하여 총 100개의 질의응답 쌍으로 구성된 벤치마크 데이터셋을 자체 제작하였다.

이는 Merten *et al.*(2025)에서 제시된 AIS 데이터 분석 프레임워크를 참고하여, 에이전트의 다면적인 능력을 정밀하게 측정할 수 있도록 설계되었다. 구체적으로 질문의 복잡도와 요구되는 분석 능력에 따라 세 가지 대분류를 설정하였다. 첫째, '단순 사실 추출(Simple Fact Extraction)'은 데이터에서 특정 정보를 직접 찾는 유형이다. 둘째,'계산 및 집계(Calculation & Aggregation)'는 통계적 연산을 요구하는 유형이다. 마지막으로,'추론 및 분석(Inference & Analysis)'은 데이터 간의 관계를 파악하고 복합적인 결론을 도출해야 하는 가장 높은 난이도의 유형이다. 각 대분류는 다시 <Table 4>와 같이 총 8개의 소분류로 세분화하여 에이전트의 다양한 분석 능력을 정밀하게 측정하고자 하였다.

벤치마크 구축에 사용된 데이터는 2023년 1월 1일부터 5월 23일까지의 부산항(북항 KRPUS, 신항 KRBNP)에서 입출항한 선박 기록으로, 고유 선박 170척, 총 680개 항차(voyage) 기록을 포함하며, MMSI, 타임스탬프, 위경도, SOG, COG 등의 주요 컬럼으로 구성된다. (1) 시간 정렬 · 중복/이상치 제거, (2) 좌표 보정 · 선형 보간(간격 임계 내), (3) 항차 분할 순의 전처리 과정도 포함되었다.

본 데이터셋은 정의된 능력들을 측정하기 위한 평가용 벤치마크로 문항의 총량보다 분석 범위의 다양성을 확보하는 데 초점을 두었다. 본 연구에서 사용한 질의응답 데이터셋은 질문의 복잡도에 따라 세 가지 대분류로 구성되지만, 에이전트가 달성해야 할 핵심 목표는 정량적 정확성 확보(값-중심), 논리적 분석 품질 확보(설명-중심)의 두 가지로 수렴한다. 실제 해양 실무에서 단순 조회 및 집계 질의가 복합 추론보다 빈번하게 발생하는 특성을 고려하여 3:1의 비율로 구성하였다. 본 벤치마크 데이터셋은 관련 연구의 재현성 및 발전을 위해 공개하였다(http://bit.ly/49Z9IgX).

4. 실 험

4.1 실험 환경

상용 모델 중 코드 생성 및 추론 능력이 뛰어난 OpenAI의 GPT-4o를 API 호출하여 사용하였으며, 질문 유형 분류기와 응답 생성 모두 동일 모델을 기반으로 작동한다. LLM 에이전트 구현을 위해 LangChain 프레임워크를 사용하였고, 상세한 환경은 <Table 5>와 같다.

Table 4. Question Taxonomy and Examples

Major Category	Subcategory	#Items	Description	Example Question
Simple Fact Extraction (35)	Static attributes	15	Fixed vessel specifications and identifiers	"What is the name of the vessel with MMSI 352058000?"
	Time-specific attributes	10	Dynamic state at a specific times tamp	"What was the last recorded speed of MMSI 352058000?"
	Conditional attributes	10	State when a condition is satisfied	"Where was the vessel when its speed first exceeded 15 kn?"
Calculation & Aggregation (40)	Global / group statistics	20	Dataset level or groupwise aggregates	"How many unique vessels are included in the dataset?"
	Per-vessel / time-series stats	20	Statistics from a single vessel's trajectory	"What is the maximum speed of MMSI 352058000?"
Inference & Analysis (25)	Pattern & anomaly detection	10	Detect abnormal patterns or irregular reporting	"What are the characteristics of vessels with abnormally long reporting intervals?"
	Comparison & performance analysis	10	Compare behaviors and explain efficiency differences	"Which vessel operated most efficiently, and why?"
	Hypotheses & data quality	5	Discuss data limitations or form hypotheses	"What are the limitations of this dataset?"

Table 5. Experimental Environment

Model	GPT-4o (API inference, July 2025)
Agent Framework	LangChain 0.2.1
Environment	Python 3.11
Specification	Processor: AMD Ryzen 5 5600X (6 Core) Memory: 32 GB RAM

4.2 평가 프레임워크

LLM 에이전트의 다면적 성능을 평가하기 위해, 본 연구에서는 (i)정량 정확성, (ii)분석 품질, (iii)비용 효율성의 세 가지 축으로 평가 프레임워크를 설계했다. 정량 정확성 평가를 위해서는 여러 상호보완적인 지표를 채택했다. EM은 답변의 완전한 일치 여부를 측정하는 가장 엄격한 지표이며, 토큰 수준의 정밀도와 재현율을 함께 고려하는 F1-score를 통해 부분 정답을 평가했다. 또한, 답변의 구조적 유사성을 평가하기 위해 최장공통부분수열(LCS)에 기반한 ROUGE-L 지표를 활용했다. 분석 품질은 정량 지표만으로 측정하기 어려운 논리적 타당성을 평가하며, 이를 위해 최근 자연어 생성 모델 평가에 널리 활용되는 LLM-as-a-judge 방식을 채택했다.

(1) 지표 정의

본 절에서는 평가에 사용되는 지표를 정의한다. 모델을 $m \in M$으로 표기하고, Q_A는 유형 A 문항 집합으로, Q_B은 유형 B 문항 집합을 의미한다. 각 문항 집합의 수는 $N_A = |Q_A|$, $N_B = |Q_B|$로 표기하며, 각 문항 i의 정답은 y_i, 모델 m의 출력은 $\hat{y}_{m,i}$ 으로 표기한다. $\mathrm{Set}(\cdot)$은 공백 기준 토큰 집합 변환을, $\mathrm{Seq}(\cdot)$은 공백 기준 토큰 시퀀스 변환을 의미하며, $J_j \in \{1, \ldots, 5\}$은 문항 $j \in Q_B$의 LLM-as-a-judge 점수를 나타낸다. 평균 지연(ms)은 $\overline{L}_m$로, 총 토큰 수는 $\overline{T}_m$로 표기한다. $w_A, w_R \in [0,1]$, $w_A + w_R = 1$은 성능 결합 가중치이며, 본 연구에서는 $w_A = 0.5$, $w_R = 0.5$를 사용하였다.

(i) 정량 정확성(A, Accuracy)

정답과 예측의 일치·포함을 토큰 수준에서 측정한다.

- Exact Match

$$\mathrm{EM_m} = \frac{1}{\mathrm{N_A}} \sum_{\mathrm{i} \in \mathrm{Q_A}} \mathbf{1}\left[\hat{\mathrm{y}}_{\mathrm{m,i}} = \mathrm{y_i}\right]$$

- Partial Match

$$\mathrm{PM_m} = \frac{1}{\mathrm{N_A}} \sum_{\mathrm{i} \in \mathrm{Q_A}} \frac{\left|\mathrm{Set}(\hat{\mathrm{y}}_{\mathrm{m,i}}) \cap \mathrm{Set}(\mathrm{y_i})\right|}{\mathrm{Set}(\mathrm{y_i})}$$

F1 (토큰 정밀·재현 기반)

$$\mathrm{F1_m} = \frac{1}{\mathrm{N_A}} \sum_{\mathrm{i} \in \mathrm{Q_A}} \frac{2 \cdot \left|\mathrm{Set}(\mathrm{y_i}) \cap \mathrm{Set}(\hat{\mathrm{y}}_{\mathrm{m,i}})\right|}{\left|\mathrm{Set}(\mathrm{y_i})\right| + \left|\mathrm{Set}(\hat{\mathrm{y}}_{\mathrm{m,i}})\right|}$$

ROUGE-L (LCS 기반 재현율)

$$\mathrm{ROUGE\_L_m} = \frac{1}{\mathrm{N_A}} \sum_{\mathrm{i} \in \mathrm{Q_A}} \frac{\mathrm{LCS}(\mathrm{Seq}(\mathrm{y_i}), \mathrm{Seq}(\hat{\mathrm{y}}_{\mathrm{m,i}}))}{\left|\mathrm{Seq}(\mathrm{y_i})\right|}$$

Accuracy (A; 정량 정확성 종합)

$$A_m = \frac{\mathrm{EM_m} + \mathrm{PM_m} + \mathrm{F1_m} + \mathrm{ROUGE\_L_m}}{4}$$

(ii) 분석 품질(R, Analysis)

Inference 문항의 LLM-as-a-judge의 5점 척도 점수를 [0,1]로 정규화한 평균을 사용하였다.

$$R_m = \frac{1}{N_B} \sum_{\mathrm{j} \in \mathrm{Q_B}} \frac{\mathrm{J_j(m)}}{5}$$

(iii) 종합 성능(P, Performance)

정량 정확성과 분석 품질의 균형적인 성능을 위해 기하평균으로 결합하였다. 유형 B가 유형 A 대비 난이도가 높고 문항 수가 불균형(3:1)인 것을 고려해, 동일한 가중치를 채택하였다.

$$P(m) = A^{w_A} \cdot R^{w_R}$$

(iv) 비용

지연 시간과 총 토큰 수를 모델 간 중앙값으로 정규화한 가중 기하평균으로 정의하였다. 가중치는 서비스 품질에 지연 시간이 끼치는 영향이 더 큰 것을 고려하여 토큰 수보다 지연 시간에 더 높은 가중을 두었다.

$$\overline{L}(m) = \frac{1}{N} \sum_i \mathrm{latency_i(m)} \times 1000\ [\mathrm{ms}]$$

$$\overline{\mathrm{T}}(\mathrm{m}) = \sum_{\mathrm{i}} (\mathrm{prompt\_tokens_i} + \mathrm{completion\_tokens_i})$$

$$\mathrm{Cost} = \sqrt[\mathrm{w_1}+\mathrm{w_2}]{\overline{\mathrm{L}}(\mathrm{m})^{\mathrm{w_1}} \times \overline{\mathrm{T}}(\mathrm{m})^{\mathrm{w_2}}}$$

(2) 비교 대상

제안하는 HPA의 성능을 검증하기 위해, Baseline 프롬프트, 값-중심 프롬프트, 설명-중심 프롬프트와 비교하였다.

- Baseline: 특정 과업을 위한 프롬프트 기법이 적용되지 않은 최소 프롬프트로, 체계적인 프롬프트 설계의 필요성을 입증하는 역할을 한다.
- 값-중심 프롬프트 & 설명-중심 프롬프트: 각각 정량적 정확성과 분석 품질이라는 단일 목표에 최적화된 프롬프트 베이스라인이다. 이는 3.2.1절과 3.2.2절에서 각각 설명된 프롬프트에 해당한다.
- Hybrid (Ours; HPA): 질문 유형 분류를 통해 값-중심 프롬프트와 설명-중심 프롬프트를 동적으로 라우팅하여 앞서 제시된 성능 상충을 완화하는 본 연구의 제안 모델이다.

4.3 결과

(1) 성능·효율 지표

<Table 6>과 <Figure 2>는 비교 실험의 결과를 보여준다.

Table 6. Performance & Efficiency Comparison across models

Model	EM	PM	F1	ROUGE-L	Accuracy (A)	Analysis (R)	Performance (P)	Latency (ms)	Tokens
Baseline	0.067	0.573	0.084	0.084	0.202	0.552	0.334	3,184	103,240
Value-CP	0.507	**0.653**	**0.535**	0.533	**0.557**	0.361	0.448	2,653	136,099
Explanation-CP	0.000	0.520	0.003	0.003	0.132	0.643	0.290	8,637	269,973
Hybrid(Ours)	**0.520**	0.640	0.534	**0.534**	**0.557**	**0.736**	**0.640**	4,565	155,601

Bold indicates the top-performing model (1st) in each metric, and underlined text indicates the second-best (2nd).

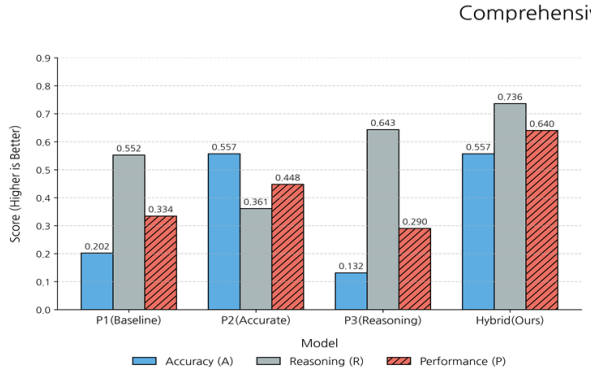

Figure 2. Key Performance Metric Comparison

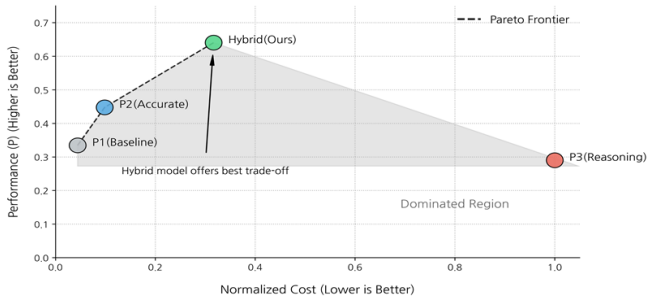

Figure 3. Performance-Cost Pareto Frontier

Value-CP는 정량 정확성(A) 하위 지표(EM 0.507, F1 0.535 등)에서 우수한 성능을 보였으나, 분석 품질(R=0.361)은 저조했다. 이는 값-중심 답변을 강제하는 프롬프트가 단답형 질의에 대해 강점이 있지만, 분석 질의에는 불리함을 시사한다. 설명-중심 프롬프트는 R=0.643로 높은 분석 품질을 보였으나, F1=0.003 등 정량 정확성은 가장 낮고 비용도 가장 높았다. Baseline은 전반적으로 저조한 성능을 보여, 체계적인 프롬프트 설계의 필요성을 재확인시킨다. HPA는 정량 정확성(0.557)과 분석 품질(0.736)을 동시에 향상시켜 종합 성능 P=0.640를 기록하여 가장 우수한 결과를 달성하였다. <Figure 3>을 보면, 값-중심 프롬프트는 정확도와 비용 면에서 경량 대안으로 의미가 있고, 설명-중심 프롬프트는 높은 분석 품질에도 불구하고 정량 정확성과 비용 측면에서 파레토 열위에 놓인다. 결론적으로, HPA는 성능과 효율성의 균형을 가장 잘 맞춘 안정적인 접근법임을 실험적으로 입증했다.

(2) 모델 일반화 검증

본 연구는 제안하는 HPA 아키텍처가 특정 모델(GPT-4o)에 국한되지 않고 다양한 LLM에서도 범용적으로 작동하는지 검증하기 위해, 특성이 다른 모델 군을 대상으로 확장 실험을 수행하였다. Qwen3-80B-Instruct(Qwen Team, 2025)는 GPT-4o 모델 대비 최신 모델로서, 오픈소스 기반임에도 뛰어난 성능을 보

Table 7. Performance & Efficiency Comparison Across LLMs and Models

LLM	Model	Accuracy (A)	Analysis (R)	Performance (P)	Latency (ms)	Tokens
GPT-4o	Baseline	0.202	0.552	0.334	3,184	103,240
	Value-CP	**0.557**	0.361	0.448	2,653	136,099
	Explanation-CP	0.132	0.643	0.290	8,637	269,973
	Hybrid(Ours)	**0.557**	**0.736**	**0.640**	4,565	155,601
	Unified	0.199	0.576	0.339	10,681	419,787
Qwen3-80B Instruct	Baseline	**0.184**	0.464	0.292	9,238	81,477
	Value-CP	**0.184**	0.456	0.290	13,587	131,743
	Explanation-CP	0.183	0.712	0.361	25,681	397,456
	Hybrid(Ours)	0.180	0.696	0.354	21,264	210,522
	Unified	0.180	**0.736**	**0.364**	18,748	411,729

Bold indicates the top-performing model (1st) in each metric, and underlined text indicates the second-best (2nd).

이며 상대적으로 적은 파라미터 수(80B)로도 효율적인 추론이 가능함을 확인하기 위해 선정하였다. 실험 결과(<Table 7> 참조), HPA 전략은 GPT-4o뿐만 아니라 Qwen3 환경에서도 단일 프롬프트 대비 전반적으로 정확도와 분석 품질의 균형을 개선하는 경향을 보였다. 다만 모델의 기초 추론 능력이나 컨텍스트 처리 특성에 따라 성능 향상의 폭에는 다소 차이가 있었다.

(3) 통합 프롬프트 전략과의 비교

제안하는 HPA 구조의 핵심인 질문 유형 분류 방식의 실효성을 검증하기 위해, 해당 단계를 생략하고 모든 분석 지침을 하나의 프롬프트 안에 통합한 '통합 전략(Unified)'과의 비교 실험을 수행하였다. 이는 단일 통합 프롬프트 방식 대비 HPA가 갖는 비용 및 성능 효율성을 객관적으로 평가하기 위함이다.

비교 분석 결과, 통합 전략은 HPA 대비 명확한 비용적 한계를 드러냈다. 통합 전략은 단순한 사실 조회 질의를 처리할 때에도 심층 분석을 위한 복잡한 지침과 예시를 문맥에 포함해야 하므로, 불필요한 연산 자원을 소모한다. 실험 결과, 통합 전략은 HPA 대비 평균 약 2.7배의 토큰을 추가로 사용하였다(<Table 7>: GPT-4o 기준: HPA 155,601 토큰 vs. Unified 419,787 토큰).

LLM 모델에 따른 변동성 또한 관찰되었다. Qwen3 같은 모델 환경에서는 통합 전략이 HPA보다 높은 종합 성능을 보이기도 하였다. 그러나 이는 2배 이상의 과도한 비용을 지불하여 얻은 결과임에도, 여전히 GPT-4o 기반 HPA(0.640)의 성능에는 크게 미치지 못한다는 한계가 있다. GPT-4o의 경우 통합 전략에서 과도한 지침이 혼재된 프롬프트가 모델의 주의 집중을 분산시켜 오히려 종합 성능이 하락(0.640 → 0.339)하는 상반된 결과가 나타났다.

요약하면, 비록 특정 모델에서 통합 전략이 점수 우위를 점하는 경우가 있었으나, 이를 위해 2배 이상의 비용을 지불해야 하고, 절대적 성능 자체가 다른 모델에 비해 낮다는 한계가 있다. HPA는 통합 전략 대비 현저한 토큰 절감 효과를 제공하며, 비용 대비 성능을 고려할 때 다양한 모델 환경에서 비용 효율적인 대안임을 확인하였다.

(4) 모델별 답변의 질적 비교

정량적 성능 지표에서 나타난 모델 간의 차이는 실제 생성된 답변의 질적 수준에서 명확하게 드러난다. <Table 8>은 각 질문 유형별로 Baseline 모델과 HPA가 생성한 답변을 비교하여 HPA의 우수성을 보여주는 대표적인 사례이다. 구체적으로

Table 8. Examples for comparison between the baseline model and HPA model (GPT-4o)

Query	Baseline		HPA		Status
	Response	Metrics	Response	Metrics	
Simple Fact Extraction What is the ship name of MMSI 636018101?	The ship name of MMSI 636018101 is NAVIOS UNISON.	EM: 0.0 PM: 1.0	NAVIOS UNISON	EM: 1.0 PM: 1.0	Baseline contains the correct value but is verbose; HPA is an exact match.
Computation and Aggregation What is the average speed of MMSI 636020476 when the vessel is under way (SOG > 0.1)? Round to the third decimal place.	It's 10.158 knots.	EM: 0.0 PM: 0.0	10.16	EM: 1.0 PM: 1.0	Baseline gives an incorrect rounded value (hallucinated precision); HPA exactly matches the ground truth.
Inference & Analysis Analyze whether there are common segments where the ship speed sharply drops among vessels sailing along the route CNSHA → KRPUS.	To analyze speed drop, the route data should be filtered and SOG variation examined.	R: 0.40	(summary) Conclusion: multiple speed-drop segments; >5-knot decreases frequent. Evidence: repeated sharp drops (e.g., 11.4→0 kn), including HMM PROMISE and HYUNDAI COURAGE. Interpretation: drops mainly occur near port approach/anchoring; may reflect congestion, weather, or mechanical factors.	R: 1.00	**Baseline** is procedural-only (**no findings**); **HPA** provides **concrete findings** with domain-consistent reasoning.

‘단순 사실 추출’ 질의에서, Baseline은 정답(‘NAVIOS UNISON’)을 포함했음에도 “The ship name of ...”와 같은 불필요한 서술어를 추가하여 가장 엄격한 지표인 EM 점수를 획득하지 못했다. 반면 HPA는 ‘값-중심 프롬프트’의 지시사항(Instruction)에 따라 정확한 값만 출력하여 EM 1.0을 달성했다. ‘계산 및 집계’ 질의에서는 Baseline이 부정확한 소수점 처리(10.158)를 보인 반면, HPA는 정확한 값(‘10.16’)을 도출했다.

가장 큰 차이를 보인 ‘추론 및 분석’ 질의에서, Baseline은 “데이터를 필터링해야 한다”는 절차만 언급했을 뿐 실제 분석을 수행하지 못했다(R: 0.40). 그러나 HPA는 ‘설명-중심 프롬프트’의 CoT 지침에 따라 구체적인 근거(e.g., “HMM PROMISE”, “11.4 →0 kn”)를 제시하며 도메인 지식에 부합하는 심층 분석을 제공했다(R: 1.00). 이는 HPA가 라우팅을 통해 단순 질의의 정확성과 복합 질의의 분석 품질을 모두 효과적으로 확보함을 질적으로 입증한다.

(5) 질문 유형 분류기 성능 분석

본 연구에서 제안한 질문 유형 분류기는 100개의 테스트 케이스에 대해 98%의 평균 분류 정확도를 기록하였다. 이러한 높은 정확도는 ‘단순 사실(What/How many)’과 ‘추론(Why/How)’이라는 과업의 언어적 특징이 비교적 뚜렷하여, 경량 모델로도 효과적인 분류가 가능했음을 시사한다.

주목할 점은 유형 B를 값-중심 유형으로 오분류한 2건의 사례가 오히려 분석 품질 점수 상승으로 이어졌다는 것이다. 상세 분석 결과, 이는 ‘가장 효율적인 선박은 무엇인가?’와 같이 분석적 추론 과정이 선행되어야 함에도 산출되는 답변은 특정 값을 위주로 설명되는 일부 경계성 질의에서 관찰되었다. 이러한 문항의 정답지는 논리적 전개 과정을 풍부하게 포함하는 일반적인 추론형 질의와 달리, 분석의 최종 산출물인 핵심 대상(선박 명칭)과 근거 수치(평균 속력 등)의 정확성 확인에 더 큰 비중을 두고 작성되었다. 결과적으로 LLM 평가 모델은 장황한 서술보다 핵심 결과를 명료하게 제시한 오분류된 답변이 정답지의 형식에 더 부합한다고 판단한 것이다.

즉, 분류기의 오판단이 해당 질의의 특수성과 맞물려 역설적으로 점수 상승을 유발하였다. 결과적으로 본 실험 환경에서 분류기의 오류는 전체 시스템의 성능 붕괴로 이어지지 않았으며, 이는 제안 모델이 경계성 질의에 대해 일정 수준의 강건성을 지니고 있음을 보여준다. 다만, 향후 이러한 모호한 경계의 질문들을 정교하게 식별할 수 있도록 데이터셋을 확충하고 분류 기준을 고도화할 필요가 있다.

5. 결 론

5.1 연구 요약 및 시사점

본 연구는 해양 AIS 데이터 분석에서 사용자의 다양한 질의 의도에 단일 프롬프트 전략이 효과적으로 대응하지 못하는 실무적 한계를 해결하고자 하였다. 이를 위해, 질문의 목표를 정량적 정확성과 분석 품질로 분류하고, 이에 최적화된 프롬프트를 적용하는 하이브리드 프롬프트 에이전트를 제안하였다. 본 연구의 핵심적인 기여점과 시사점은 다음과 같이 요약할 수 있다.

첫째, 상충 관계의 효과적 해결책을 제시하였다. 제안된 에이전트는 사용자의 질의 의도에 따라 경로를 동적으로 변경함으로써, 정량적 데이터 추출의 정확성과 해석적 분석의 품질이라는 두 가지 목표를 동시에 달성할 수 있음을 확인하였다.

둘째, 성능과 비용 효율성의 균형을 입증하였다. 실험 결과, 제안 모델은 기존의 단일 전략 대비 우수한 성능을 보이면서도 불필요한 연산을 줄여 비용 효율성 측면에서도 가장 균형 잡힌 접근법임을 증명하였다.

셋째, 프롬프트 아키텍처 중심의 설계 패러다임을 제시하였다. 본 연구는 고비용의 모델 미세조정 과정 없이, 프롬프트 아키텍처 설계만으로도 특정 도메인의 정형 데이터 분석 성능을 유의미하게 향상시킬 수 있음을 보여주었다. 이는 해양 분야뿐만 아니라 금융이나 제조 품질 관리 등 복잡한 데이터 분석 요구가 혼재하는 타 산업 분야에서도 실용적인 대규모 언어 모델 기반 시스템을 구축하는 데 기여할 수 있다.

5.2 한계 및 향후 연구 방향

본 연구는 다양한 언어 모델 환경에서 하이브리드 프롬프트 에이전트의 유효성을 검증하고 질문 유형 분류기의 높은 정확도를 확인했으나, 여전히 몇 가지 한계가 존재한다.

첫째, 기초 모델의 추론 능력에 따른 성능 편차이다. 실험 결과, 제안하는 아키텍처가 전반적인 성능을 향상시키는 경향성은 뚜렷했으나, 모델이 가진 고유한 지시 이행 능력이나 문맥 처리 역량에 따라 개선 폭에는 차이가 있었다. 이는 에이전트가 경량화된 모델이나 오픈소스 모델에서도 일관된 고성능을 보장하기 위해서는 추가적인 프롬프트 최적화가 필요함을 시사한다.

둘째, 벤치마크 데이터셋의 정답지 구성 한계이다. 4.3.5절에서 논의한 바와 같이, 일부 분석형 질의의 정답지가 핵심 결과 위주로 구성되어 있어, 에이전트의 상세한 추론 과정을 정량적으로 평가하는 데 한계가 있었다. 향후 연구에서는 질문의 개수를 늘리고, 추론 과정을 더 구체적으로 명시한 정답지를 구성하는 등 정답지의 서술 방식을 다양화하여, 분류기의 민감도와 라우팅 효과를 더욱 정밀하게 검증할 계획이다.

셋째, 데이터 범위의 제한이다. 본 실험은 부산항의 특정 기간 데이터에 한정되어 수행되었기에, 기상 변화나 실시간 스트리밍 환경과 같은 동적 변수들에 대한 추가적인 검증이 요구된다. 이를 보완하고 발전시키기 위한 향후 연구 방향은 다음과 같다.

(1) 다단계 라우팅 및 외부 지식 연동을 통한 전문가 에이전트로의 진화

현재의 이진 분류 체계를 넘어, 질문의 의도를 더욱 세분화

하여 처리하는 다단계 라우팅으로 분류 과정을 고도화할 수 있다. 나아가 검색 증강 생성 기술을 결합하여 국제해사기구 규정이나 최신 해사 안전 문서와 같은 외부 도메인 지식을 실시간으로 참조하도록 확장할 수 있다. 이는 에이전트가 단순 분석을 넘어 규정 준수 여부까지 판단하는 해양 도메인 전문가로 진화하는 기반이 될 것이다.

(2) 도메인 특화 도구 결합을 통한 분석 능력 확장

실제 분석 환경에서 필수적인 항차 분할 등의 반복적인 전처리 작업을 언어 모델이 매번 코드로 생성하는 것은 비효율적일 수 있다. 따라서 이러한 핵심 기능들을 사전 정의된 API 형태의 도구로 제공하고, 에이전트가 필요에 따라 이를 동적으로 호출하는 도구 기반 에이전트 시스템으로 확장할 수 있다. 이를 통해 특정 구간의 이상 감속 식별이나 평균 속력 계산과 같이 더욱 복잡하고 실무적인 문제를 단일 질의로 해결하는 시스템을 구축할 수 있을 것이다.

참고문헌

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., *et al.* (2020), Language Models are Few-Shot Learners, *Advances in Neural Information Processing Systems*, **33**, 1877-1901.

Chen, L., Chen, P., and Wu, H. (2023), Vessel Trajectory Prediction Based on Transformer with Introduction of Attention Mechanism, *Ocean Engineering*, **280**, 114624.

Kim, H. and Kim, H. (2023), Contextual Anomaly Detection for High-Dimensional Data Using Dirichlet Process Variational Autoencoder, *IISE Transactions*, **55**(5), 433-444.

Kim, S., Kim, H., and Park, Y. (2017), Early Detection of Vessel Delays Using Combined Historical and Real-Time Information, *Journal of the Operational Research Society*, **68**(2), 182-191.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022), Large Language Models are Zero-Shot Reasoners, arXiv preprint arXiv:2205.11916.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. (2024), Better Zero-Shot Reasoning with Role-Play Prompting, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4099-4113.

LangChain Authors (2025), Pandas DataFrame Agent - Documentation (create_pandas_dataframe_agent), Available at: https://python.langchain.com.

Lin, C.-Y. (2004), ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 74-81.

LlamaIndex Authors (2024), Pandas Query Engine Documentation. Available at: https://www.llamaindex.ai.

Lundström, O., Nilsson, A., and Johansson, M. (2025), *Large Language Models (LLMs) in Maritime Data Analysis and Decision Support*, Lighthouse Report.

Luo, T., Lei, F., Lei, J., Liu, W., He, S., Zhao, J., and Liu, K. (2023), HRoT: Hybrid Prompt Strategy and Retrieval of Thought for Table-Text Hybrid Question Answering, arXiv preprint arXiv:2309.12669.

Merten, G., Dejaegere, G., and Sakr, M. (2025), Using LLMs for Analyzing AIS Data, arXiv preprint arXiv:2504.07557.

Nguyen, D. L., Vadaine, R., Haj-Yihia, G., *et al.* (2018), A Multi-Task Deep Learning Architecture for Maritime Surveillance Using AIS Data Streams, *IEEE Transactions on Intelligent Transportation Systems*, **19**(11), 3429-3440.

Oh, Y. and Kim, S. (2023), Grid-Based Bayesian Bootstrap Approach for Real-Time Detection of Abnormal Vessel Behaviors from AIS Data in Maritime Logistics, *IEEE Transactions on Automation Science and Engineering*, **21**(4), 6680-6692.

Oh, Y., Yoon, K., Park, J., and Kim, S. (2024), Comparative Evaluation of VAE-Based Monitoring Statistics for Real-Time Anomaly Detection in AIS Data, *Maritime Policy & Management*, **52**(4), 609-626.

Pallotta, G., Vespe, M., and Bryan, K. (2013), Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction, *Entropy*, **15**(6), 2218-2245.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002), BLEU: A Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318.

Park, H., Jung, J., Seo, M., Choi, H., Cho, D., Park, S., and Choi, D.-G. (2025), AIS-LLM: A Unified Framework for Maritime Trajectory Prediction, Anomaly Detection, and Collision Risk Assessment with Explainable Forecasting. arXiv preprint arXiv:2508.07668.

Park, J. and Kim, S. (2020), Maritime Anomaly Detection Based on VAE-CUSUM Monitoring System, *Journal of the Korean Institute of Industrial Engineers*, 46(4), 432-442.

Qwen Team (2025), Qwen3 Technical Report, arXiv preprint arXiv:2505.09388.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016), SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv preprint arXiv:1606.05250.

Schulhoff, S., Ilie, M., *et al.* (2025), The Prompt Report: A Systematic Survey of Prompt Engineering Techniques, arXiv preprint arXiv:2406.06608.

Sun, S., Zhao, L., Deng, M., and Fu, X. (2025), VTS-LLM: Domain-Adaptive LLM Agent for Enhancing Awareness in Vessel Traffic Services through Natural Language, arXiv preprint arXiv:2505.00989.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., *et al.* (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv preprint arXiv:2201.11903.

Xie, J., Zhang, K., Chen, J., Yuan, S., Zhang, K., Zhang, Y., and Li, L. (2024), Revealing the Barriers of Language Agents in Planning. arXiv preprint arXiv:2410.12409.

Yang, D., Wu, L., Wang, S., Jia, H., and Li, K. X. (2019), How Big Data Enriches Maritime Research: A Critical Review of AIS Data Applications, *Transportation Reviews*, **39**(6), 755-773.

Zheng, L., Chiang, W. L., Sheng, Y., *et al.* (2023), Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

부록 A: 출력 형식 정제

Table A1. Post-processing Answer Reduction (Rule, GPT-4o)

Model (Rule-based)	EM	PM	F1	ROUGE-L	Accuracy (A)	Analysis (R)	Performance (P)
Baseline	0.387	0.392	0.394	0.394	0.391 (+0.189)	0.360 (-0.192)	0.375 (+0.041)
Value-CP	0.493	0.496	0.499	0.499	0.496 (-0.061)	0.336 (-0.025)	0.408 (-0.040)
Explanation-CP	0.240	0.241	0.241	0.241	0.241 (+0.109)	0.560 (-0.083)	0.367 (+0.077)
Hybrid(Ours)	**0.507**	**0.507**	**0.507**	**0.507**	**0.507 (-0.050)**	**0.577 (-0.159)**	**0.540 (-0.100)**

Table A2. Post-processing Answer Reduction (LLM, GPT-4o)

Model (LLM-based)	EM	PM	F1	ROUGE-L	Accuracy (A)	Analysis (R)	Performance (P)
Baseline	0.173	0.195	0.195	0.193	0.187 (-0.015)	0.544 (-0.008)	0.287 (-0.047)
Value-CP	0.507	0.529	0.533	0.533	0.523 (-0.034)	0.384 (+0.023)	0.448 (+0.000)
Explanation-CP	0.213	0.225	0.224	0.224	0.221 (+0.089)	0.585 (-0.058)	0.370 (+0.080)
Hybrid(Ours)	**0.520**	**0.536**	**0.536**	**0.536**	**0.531 (-0.026)**	**0.585 (-0.151)**	**0.557 (-0.083)**

Values in parentheses denote the change(Δ) related to <Table 6>.

설명-중심 프롬프트는 정답 정보를 포함하더라도, CoT 기법의 특성상, 부연 설명 등 형식적 불일치로 인해 EM과 같은 정형 지표에서 성능이 과소평가될 수 있다. 이러한 현상을 분석하고 후처리의 실효성을 검증하기 위해, 원본 출력 $\hat{y}_{i,m}$에서 핵심 답변만을 추출하는 정제과정을 설계하였다. 정제 방식으로는 (i) 규칙 기반 정제와 (ii) LLM 기반 정제를 각각 적용하였다. 먼저 규칙기반 정제는 답변에서 수치와 단위 패턴(예 12.5 knots)을 최우선으로 탐색하여 추출한다. 만약 수치 패턴이 부재할 경우 인용구(“”)로 강조된 식별자를 찾고, 텍스트의 마지막 줄을 최종 답변으로 간주하는 로직을 적용했다. 반면, LLM 기반 정제는 경량 모델(gpt-4o-mini)을 활용하여 원본 출력 내에 존재하는 정답 문자열을 그대로 가져오도록 프롬프팅하였다. 이후 추출된 답변이 실제 원본 텍스트에 존재하는지 여부를 검증하여 모델의 환각 가능성을 차단하였다.

<Table A1>, <Table A2>는 각각 규칙 기반 정제 방식과 LLM 기반 정제 방식의 후처리를 적용한 실험 결과이다. 설명-중심 프롬프트에 후처리를 적용했을 때 정량 정확도(A)가 유의미하게 상승하였다. 규칙 기반 정제 방식 적용 시 A는 0.241, LLM 기반 정제 방식 적용 시 0.221로, 원본(A=0.132) 대비 각각 큰 폭으로 개선되었다. 이는 정답을 포함하고도 장황한 형식 때문에 감점되었던 다수의 사례가 후처리를 통해 복원되었음을 시사하였다. 그러나 분석 품질(R)은 변하지 않은 상태에서 종합 성능(P)은 LLM 기반 정제 방식 기준 0.370에 그쳐, Hybrid(P=0.640) 및 값-중심 프롬프트(P=0.448)의 성능에는 미치지 못하였다.

오히려 본 연구가 제안하는 HPA는 후처리를 적용했을 때 성능이 하락하거나(규칙 기반 정제 방식, P=0.540), 원본 출력(P=0.640)과 비교하여 유의미한 이점을 보이지 않았다. 이는 하이브리드 전략이 생성 단계에서부터 유형별로 최적화된 간결한 출력을 생성하므로, 추가적인 후처리가 불필요하거나 오히려 성능 저하를 야기할 수 있음을 의미하였다.

이상의 결과를 종합하면, 출력 후처리는 장황한 응답의 형식적 문제를 일부 교정하여 정량적 정확도(A) 지표를 개선할 수는 있었다. 그러나 이 과정을 거치더라도 기존 모델 대비 성능 우위를 확보하지 못했다. 따라서, 생성 이전에 질문 유형을 먼저 분류하고 최적의 프롬프트 전략을 선택 적용하는 본 연구의 모델이 성능과 비용 효율성에서 우위임을 재확인했다.

부록 B: Baseline 프롬프트

Table B. Baseline Prompt Template

Component	Key Instructions
Role	You are a helpful assistant.
Context	Provided column descriptions containing standard physical definitions of AIS variables (e.g., defining SOG as Speed Over Ground, knots) to understand the dataset structure.
Instruction	Please answer the question ({{ question }}) using the provided information.

저자소개

김성진: 울산과학기술원(UNIST) 산업공학과에서 2024년 학사 학위를 취득하고, 현재 동 대학 산업공학과 석사과정에 재학 중이다. 주요 연구분야는 다변량 시계열 이상탐지 및 도메인 특화 LLM 응용이다.

김성일: 연세대학교 정보산업공학과에서 2005년 학사, Georgia Tech에서 2007년 산업공학 석사, 2011년에 산업공학 박사 학위를 취득하였다. 미국 Terra Technology와 삼성 SDS연구소에서 근무하고, 2016년부터 울산과학기술원 산업공학과 교수로 재직 중이다. 연구 분야는 산업통계, 품질 공학, 산업 인공지능이다.