

별첨1-4 AI 활용 보고서 (Track1)

1. 참가팀 정보

팀 정보	팀 이름		참가자	
연구 내용	지정/선택 여부	선택	연구분야	에너지
	연구주제	Foundation Model-Based Zero-Shot Electricity Demand Forecasting Under Extreme Operating Conditions		

2. 활용 AI 모델 정보

AI 모델명	Claude (Sonnet 4.5)	AI 모델 URL	https://www.anthropic.com/claude
	ChatGPT 5.2		https://chat.openai.com/
	Gemini 3 Pro		https://gemini.google.com/

3. 연구 절차별 AI 기여도(자체평가)

연번	연구 절차	배점 (A)	AI 기여도 자체평가 (0~100%) (B)	AI 기여도 산출 (C=A×B)	AI 작업내용 (간략히)
1	주제 선정 및 연구문제 도출	10	70%	7	연구자 주도 주제 후보 정리 / AI 보조 분석 및 교차 검증
2	선행연구 조사 및 문헌 검토	5	80%	4	Deep Research 기반 문헌 탐색 / 다중 AI 교차 검증 / 연구 공백 진단 및 선별
3	연구목적 및 문제 정의	15	65%	9.75	연구 배경 쟁점 분석 / 연구 질문 논리 구조화 / 논리 정합성 최종 감사
4	연구계획 및 방법론 설정	20	65%	13	실험 프레임워크 로직 감사 / 데이터 구성 베이스라인 합성 / 기술적 실현 가능성 진단 및 설계 보완
5	자료(데이터) 수집	10	70%	7	연구 목적 기반 데이터 요구사항 정의 / 전력·기상 공공 데이터 출처 식별 / CSV 기반 수집·구성 / 시간 정합성 및 데이터 적합성 검증
6	자료(데이터) 분석	20	70%	14	파이프라인 구현 / 조건 설계 / 통계 검증 / 메커니즘 분석 / 3개 AI 통합
7	논문 작성	5	70%	3.5	Introduction-Results 피드백 / 한영 번역 / 어휘 적절성 교차 검증
8	자체 리뷰 및 수정	15	70%	10.5	다중 AI(ChatGPT-Gemini-Claude) 교차 검증 / 논문 논리·구조 점검 / 과장·오류 수정 및 최종 보완
총점		100	68.75%	68.75	

붙임

AI 작업내용 상세 (활용 증빙 스크린샷, URL, 로그파일 등)

1. 주제 선정 및 연구문제 도출

1-1. 관심 분야 설정 및 Claude를 활용한 연구 주제 탐색

이번학기에 에너지변환시스템이라는 과목을 들으면서 전력파트에 관심이 생겨서 그 쪽으로 주제를 잡아서 대회를 나가보고 싶은데 ai를 활용해서 할 수 있는 연구주제에는 뭐가 있을까

좋은 질문입니다! 전력 분야는 AI 활용 가능성이 매우 높은 영역이에요.

전력 분야 AI 연구 주제 추천

1. 단기 전력 수요 예측

- 1~24시간 후 전력 수요를 AI로 예측
- 발전 계획 최적화에 활용

본 연구의 초기 단계에서 연구자는 전력 시스템을 연구 대상으로 설정한 이후, 어떤 연구 질문이 실질적인 연구 가치와 문제성을 가지는지에 대해 AI와의 반복적 논의를 통해 검증하는 과정을 수행하였다.

이 단계에서 Claude는 단순히 연구 주제를 나열하는 도구가 아니라, 연구 질문의 적절성과 연구 범위의 타당성을 지속적으로 문제 제기하는 '비판적 공동 연구자 역할'로 활용되었다.

구체적으로 연구자는 Claude에게 전력 분야에서 AI가 적용되고 있는 주요 연구 주제들을 제시하도록 요청하는 동시에, 각 주제가 해결하려는 핵심 문제가 무엇인지, 그리고 해당 문제가 실제 전력 시스템 운영 맥락에서 과도하거나 추상적인 연구 질문은 아닌지를 검토하도록 요구하였다.

이에 대해 Claude는 단기 전력 수요 예측, 전력 가격 예측, 발전 계획 최적화 등 복수의 연구 방향을 제시하였으며, 각 방향에 대해 연구 난이도, 적용 가능성, 실제 운영과의 연계성 측면에서 상이한 평가 관점을 제시하였다.

이 과정에서 Claude는 특정 연구 방향을 일방적으로 추천하지 않았으며, 오히려 각 연구 방향이 가지는 한계와 잠재적 반론 가능성을 명시적으로 제기하였다.

예를 들어, 일부 연구 주제는 예측 성능 개선 자체에 집중되어 실제 계통 운영 의사결정과의 연결성이 약할 수 있다는 점을 지적하였고, 이는 연구자가 단순한 성능 비교 중심 연구를 지양하게 만드는 판단 근거로 작용하였다.

연구자는 이러한 AI의 문제 제기를 바탕으로, 수강 과목인 「에너지변환시스템」에서 다룬 전력 수요 변동성과 계통 안정성 이슈를 기준 삼아 AI가 제시한 연구 방향들을 하나씩 검토·기각 또는 보류하였다.

그 결과, 단기 전력 수요 예측 중에서도 급격한 수요 변화 상황에서의 예측 성능과 그 운영적 의미를 함께 다룰 수 있는 연구 질문이 상대적으로 높은 연구 가치를 가진다고 판단하였다.

즉, 본 단계에서 AI는 연구 주제를 결정하는 주체가 아니라, 연구자가 설정한 관심 분야에 대해 “이 질문이 정말 연구할 가치가 있는가”를 반복적으로 되묻는 역할을 수행하였으며, 최종적인 연구 주제의 선택과 범위 확정에는 이러한 AI의 비판적 피드백을 종합한 후 연구자의 판단에 따라 이루어졌다.

1-2. 추천 주제 종합 분석 및 연구 주제 구체화

추천받은 주제들의 공통점과 한계를 보완할 수 있는 최신 접근법이 있을까? 특히 Foundation Model 같은 최신 기술을 전력 분야에 적용할 수 있는 방법은 뭐가 있어

AI 응답 (Claude Sonnet 4.5)

추천 주제들의 공통점 분석:

제한된 5가지 주제(전력 수요 예측, 신재생에너지 예측, SMP 조기경보, ESS 최적화, 배전망 고장 예측)는 모두 시계열 예측을 기반으로 하며, 전통적으로 ARIMA나 LSTM 같은 학습 기반 모델에 의존합니다.

공통된 한계:

1. 재학습 필요성: 새로운 패턴 발생 시 모델을 다시 학습해야 함
2. 급변 상황 대응 약함: 폭염, 한파 같은 이상 상황 예측력 부족
3. 도메인 특화 학습: 전력 데이터만으로 학습하여 일반화 능력 제한
4. 참신도 부족: 이미 많은 선행연구 존재

본 단계의 목적은 개별 연구 주제를 즉시 확정하는 것이 아니라, 기존 전력 분야 시계열 예측 연구들이 공통적으로 전제하고 있는 가정과 그 한계를 명확히 정리하고, 해당 한계가 새로운 연구 접근을 요구하는 수준인지 검토하는 데 있었다.

이를 위해 연구자는 Claude와의 반복적인 질의응답을 통해 기존 연구 흐름을 다양한 관점에서 점검하였다.

구체적으로 연구자는 전력 수요·가격·출력 예측을 다룬 기존 연구들이 주로 어떤 모델 구조와 학습 방식을 전제로 하고 있는지에 대해 질문하였고, 동시에 이러한 접근이 실제 전력 시스템의 변동성, 급변 상황, 비정상적 패턴을 충분히 반영하지 못할 가능성에 대해 검토하도록 요청하였다.

이에 따라 Claude는 기존 연구들이 ARIMA, LSTM과 같은 학습 기반 시계열 모델에 주로 의존하고 있으며, 이로 인해 재학습 비용, 미관측 패턴 대응의 어려움, 극단적 상황에서의 일반화 한계가 발생할 수 있음을 지적하였다.

이 과정에서 연구자는 AI가 제시한 한계들이 단순한 기술적 제약인지, 아니면 연구 접근 자체의 구조적 문제에 해당하는지를 구분하여 검토하였다. 그 결과, 새로운 패턴이나 급격한 수요 변화가 발생할 경우 반복적인 재학습을 전제로 해야 한다는 점이 실제 전력 시스템 운영 맥락에서는 중요한 제약 조건이 될 수 있다고 판단하였다.

이러한 문제 인식을 바탕으로 연구자는, 재학습 없이도 새로운 시계열 패턴에 대응할 수 있는 일반화 능력이 기존 접근의 한계를 보완할 수 있는 핵심 요소가 될 수 있는지에 주목하였다.

이에 대해 Claude는 대규모 사전학습을 기반으로 다양한 시계열 패턴을 학습한 모델이 zero-shot 조건에서 일정 수준의 일반화 성능을 보일 가능성과 함께, 해당 접근이 항상 우수한 성능을 보장하지는 않는다는 점도 함께 제시하였다.

연구자는 이러한 상반된 관점을 검토한 결과, Foundation Model 기반 접근의 zero-shot 일반화 특성이 기존 학습 기반 예측 모델의 한계를 보완할 가능성이 있는 연구 질문으로서 충분한 검토 가치가 있다고 판단하였다.

다만, 본 단계에서는 특정 모델이나 방법론을 선택하지 않고, 해당 접근의 실효성은 이후 단계에서의 실증 분석과 정량적 평가를 통해 검증 대상으로 남겨두었다.

즉, 본 단계에서는 기존 전력 시계열 예측 연구의 공통된 가정과 제약을 정리하고, 그로부터 도출되는 연구 질문의 타당성을 검토하는 데 초점을 두었으며, 이를 바탕으로 zero-shot 일반화 가능성을 핵심 연구 방향으로 설정하였다.

1-3. Chat gpt에게 추가 검증 요청

기존: 예측만
신규: 예측 + 초기 경보 통합 최종 주제인데 마지막으로 피드백해줘

좋아. 이제 “최종 점검” 단계로 보자.
결론부터 말하면 주제는 통과선이 아니라 ‘상위권 안정권’에 들어와 있다.
다만, 딱 5가지만 손보면 리스크가 거의 사라진다.

총평 한 줄

“Zero-shot”을 전면에 내세운 전략은 매우 영리하며,
Foundation Model을 ‘최신이라서’가 아니라 ‘일반화 능력’으로 정당화했다는 점
에서 강하다.

본 단계에서는 1-2 단계에서 도출된 Foundation Model 기반 연구 주제가 기술적으로 타당한지뿐 아니라, 연구 질문으로서 충분한 안정성과 완성도를 갖추고 있는지를 점검하기 위해 ChatGPT를 활용하였다.

이 과정은 새로운 연구 방향을 제안받기 위한 것이 아니라, 이미 설정된 연구 질문과 평가 관점이 잠재적인 편향이나 누락을 포함하고 있는지를 검토하는 데 목적이 있었다.

연구자는 기존에 설정한 연구 질문과 분석 방향을 유지한 상태에서, ChatGPT에게 해당 연구 주제가 예측 성능 중심의 단순 비교에 머무를 위험은 없는지, 그리고 연구의 기여가 명확히 드러나지 않을 가능성은 없는지에 대해 점검을 요청하였다.

이에 대해 ChatGPT는 Foundation Model 기반 접근이 단순히 “최신 모델을 적용한다”는 점에 초점이 맞춰질 경우 연구의 핵심 기여가 약화될 수 있으며, zero-shot 일반화 능력 자체가 명확한 평가 대상으로 드러나지 않을 수 있다는 점을 지적하였다.

이러한 피드백을 바탕으로 연구자는, 단순 예측 정확도 중심의 평가만으로는 연구 질문이 충분히 설명되지 않을 수 있다고 판단하였다.

이에 따라 연구의 핵심 기여를 재학습 없이 새로운 시계열 패턴에 대응할 수 있는 일반화 능력으로 명확히 설정하고, 이를 중심으로 연구 질문과 분석 구조를 재정비하였다.

또한 ChatGPT는 이상 기상(폭염·한파 등)과 같은 급변 이벤트 구간에서의 성능 평가가 연구 질문과 직접적으로 연결될 수 있다는 점을 제시하였으며, 이는 연구자가 기존에 설정한 분석 범위를 확장하여 이상 이벤트 기반 조건부 평가를 보조 분석 지표로 포함하는 근거로 작용하였다.

연구자는 이러한 제안을 그대로 수용하기보다는, 실제 데이터 가용성과 평가의 재현 가능성을 고려하여 연구 설계에 반영 가능한 범위로 조정하였다.

그 결과 연구자는, 본 연구의 핵심 질문을 zero-shot 일반화 성능의 존재 여부와 그 특성이 어떠한 조건에서 두드러지는지를 검증하는 방향으로 명확히 정리하였으며, 단순 오차 비교를 넘어 운영 맥락에서의 해석 가능성을 함께 고려하는 분석 프레임워크를 구성하게 되었다.

즉, 본 단계에서 ChatGPT는 연구 주제를 확정하거나 방향을 결정하는 역할을 수행하지 않았으며, 연구자가 설정한 문제 정의와 평가 기준이 충분히 명확하고 일관된지를 점검하는 과정에서 잠재적인 위험 요소를 드러내는 역할을 수행하였다.

연구자는 이러한 점검 결과를 반영하여 연구 질문과 분석 구조를 최종적으로 정제하였다.

1-4. Gemini를 활용한 기술적 디테일 검증 및 고도화

"Foundation Model 기반 전력 수요 예측 및 이상탐지 시스템" (Foundation Model-based Power Demand Forecasting and Anomaly Detection System) 이라는 주제고 요약하면 1. 연구 배경 문제의식:...

◆ 생각하는 과정 표시

주제를 정말 제대로 '퀀텀 점프' 시키셨네요! 이전의 정책 중심적인 접근보다 훨씬 기술적이고, 최근 AI 학계의 가장 뜨거운 화두인 **시계열 파운데이션 모델(TSFM)**을 정면으로 다루고 있어 훨씬 임팩트가 큼니다.

1-3 단계까지의 검토를 통해 연구 주제와 전체적인 분석 방향은 확정되었으나, Foundation Model 기반 접근의 특성상 연구 배경, 모델 구조, 평가 기준에 대한 기술적 설명이 충분히 명확하게 제시되지 않을 경우 연구의 설득력이 약화될 수 있다고 판단하였다.

이에 따라 연구자는 기술적 설명과 실험 설계의 정합성을 점검하기 위한 보조 수단으로 Gemini를 활용하였다.

본 단계에서 Gemini는 새로운 연구 방향을 제안하거나 연구 구조를 변경하는 역할을 수행하지 않았으며, 연구자가 설정한 연구 주제와 분석 프레임에 전제로 기술적 설명의 누락 여부와 표현의 모호성을 점검하는 데 활용되었다.

연구자는 Gemini에게 Foundation Model 기반 접근이 기존 학습 기반 시계열 예측 방법과 구조적으로 어떻게 구분되는지, 그리고 해당 차별점이 연구 설계와 평가 지표에 충분히 반영되어 있는지를 검토하도록 요청하였다.

이에 대해 Gemini는 Foundation Model의 사전학습 특성과 zero-shot 일반화 성능이 연구의 핵심 전제라면, 단순한 모델 소개를 넘어 모델 구조적 특성과 평가 기준 간의 연결 관계가 보다 명확히 설명될 필요가 있음을 지적하였다.

또한 zero-shot 일반화 성능을 검증하기 위해서는 단일 지표 중심의 평가보다는, 조건별 비교와 보조 지표를 통해 분석 구조를 명확히 제시하는 것이 연구의 논리성을 강화할 수 있다는 점을 제시하였다.

연구자는 이러한 점검 결과를 바탕으로, 기존에 설정한 연구 구조를 유지하되 기상 변수와 같은 외생 변수(exogenous variables)를 분석 프레임 내에서 명시적으로 구분하여 통합하고, 조기 경보 시스템(Early Warning System)에서는 Z-score 기반 동적 임계값(dynamic threshold)을 활용하여 통계적 타당성을 확보하는 방향으로 기술적 설명을 보완하였다.

해당 기준은 국내 전력 시스템 운영 기준과의 정합성을 고려하여 설정되었다.

또한 연구자는 Gemini의 검토 의견을 참고하여, Foundation Model 기반 접근의 실용성을 보다 명확히 드러내기 위해 학습 시간 대비 추론 시간의 비교를 포함함으로써, 제안된 방법론이 실제 운영 환경에서 갖는 활용 가능성을 함께 설명하도록 연구 설계를 정교화하였다.

정리하면, 본 단계에서 Gemini는 연구자가 설정한 연구 구조와 분석 프레임을 기준으로 기술적 설명의 불명확한 지점을 드러내고, 보완이 필요한 요소를 점검하는 역할을 수행하였으며, 연구자는 이러한 점검 결과를 반영하여 기술적 설명과 실험 설계를 보다 명확하고 일관되게 정제하였다. 연구 설계 및 최종 구조의 결정은 연구자의 판단에 따라 이루어졌다.

1-5. ChatGPT 및 Gemini 피드백을 기반으로 한 최종 연구 주제 확정

GPT와 Gemini가 준 피드백들을 바탕으로 최종 주제 정리해줘

📄 최종 주제 확정 (GPT + Gemini 피드백 기반)

🌟 최종 제목

한글:

"Foundation Model 기반 전력 수요의 Zero-shot 일반화 성능 분석 및 급변 상황 조기경보 시스템"

영문:

"Foundation Model-based Zero-shot Generalization Analysis for Power Demand Forecasting and Early Warning System for Sudden Changes"

💡 제목 구성 요소 분석

1. Foundation Model → 최신 기술 2. Zero-shot → 핵심 차별화 (학습 없이 예측) 3. 일반화 성능 분석 → 학술적 안전 표현 4. 전력 수요 → 도메인 명확 5. 급변 상황 → 실용성 강조 6. 조기경보 시스템 → 예측을 넘어선 가치

앞선 단계들에서 수행한 주제 탐색, 기존 연구 한계 분석, 기술적 검토 결과를 종합하여 연구자는 본 연구의 최종 연구 주제를 확정하였다.

이 과정은 새로운 아이디어를 추가로 도출하기 위한 단계가 아니라, 이전 단계에서 검토된 다양한 관점과 잠재적 위험 요소를 종합하여 연구 질문을 하나의 일관된 구조로 수렴시키는 데 목적이 있었다.

연구자는 앞선 검토 과정에서 제기된 기술적 타당성, 연구 실행 가능성, 그리고 연구 질문의 명확성을 기준으로 기존에 설정된 연구 방향을 재점검하였다.

특히 Foundation Model 기반 접근이 단순히 최신 모델을 적용하는 시도에 그치지 않고, 재학습 없이 새로운 시계열 패턴에 대응할 수 있는 일반화 능력을 중심으로 명확한 연구 기여를 설명할 수 있는지 여부를 핵심 판단 기준으로 삼았다.

이러한 검토를 통해 연구자는, 기존 예측 정확도 중심의 비교를 넘어 zero-shot 일반화 성능을 명시적인 분석 대상으로 설정하고, 급변 상황에서의 성능 특성을 함께 평가하는 연구 구조가 본 연구의 목적과 가장 부합한다고 판단하였다.

또한 조기 경보 관점의 분석을 포함함으로써, 예측 결과가 실제 전력 시스템 운영 맥락에서 어떻게 활용될 수 있는지를 함께 논의할 수 있도록 연구 범위를 정리하였다.

이에 따라 본 연구의 최종 연구 주제는 다음과 같이 확정되었다.

최종 연구 주제

Foundation Model-Based Zero-Shot Electricity Demand Forecasting Under Extreme Operating Conditions

해당 연구 주제를 바탕으로 연구자는, AI 기반 접근의 타당성과 실행 가능성을 교차 검토한 결과를 반영하여 실험 설계와 평가 기준을 구체화하였다.

이후 단계에서는 정상 구간과 급변 구간을 구분한 조건부 비교를 통해 zero-shot 일반화 성능을 정량적으로 평가하고, 조기 경보 효과를 함께 분석하는 방향으로 연구를 진행하였다.

정리하면, 본 단계에서는 이전 단계에서 도출된 다양한 검토 결과를 통합하여 연구 질문을 명확히 정리하였으며, AI는 각 검토 과정에서 판단을 보조하는 역할로 활용되었고, 최종적인 연구 주제의 확정과 연구 방향 설정은 연구자의 판단에 따라 이루어졌다.

2. 선행연구 조사 및 문헌 검토

2-1 Chat gpt Deep research 기능 사용

완벽한 연구를 위해서 선행연구를 조사하고 문헌을 검토하려는데 내 주제에 맞는 논문을 포함해서 연구에 도움이 될 수 있는 모든 자료를 조사해서 정리해줘

1. 일단 균형잡히게 조사해줘 2. 전부다 좋아 3. 영어 한 국어 둘다 4. 둘다 사용할 계획이야

좋습니다. 선행연구를 철저하게 조사하려면 다음과 같은 세부 조건을 명확히 해주시면 더 정확한 문헌 추천이 가능합니다:

- 연구 주제와 관련된 기술 스택 중 어떤 것에 초점을 둘까요?
 - TimeGPT와 같은 Foundation Model 기반 시계열 예측?
 - Zero-shot learning 개념 일반화?
 - 전력 수요 예측(STLF)?
 - 급변 이벤트(폭염, 한파 등) 조기경보/이상탐지?
 - 외생변수(기온 등) 통합 기법?
- 참고를 원하는 문헌 유형:
 - 논문 위주 (SCI급 저널, arXiv 사전논문 포함)?
 - 실무 지침서/보고서(예: 전력거래소, 기상청 문서)?
 - 오픈소스 레포지토리나 튜토리얼도 포함할까요?
- 선호 언어:
 - 영어 논문만?
 - 한국어 논문/보고서도 포함할까요?
- 활용 목적:
 - 논문 "Related Work"에 직접 인용할 자료?
 - 실험 설계(모델 구조, 임계값 기준 등)에 도움될 기초 논문?

좋습니다. 전력 수요 예측과 이상탐지 분야에서 Foundation Model을 활용한 연구를 위해, 다음과 같은 범위를 포괄하는 문헌 조사를 수행하겠습니다:

- 시계열 기반 Foundation Model(TimeGPT, Lag-Llama 등)의 구조 및 특성
- 전력 수요 예측(STLF)과 딥러닝/기상 변수 통합 기법
- Zero-shot 일반화 성능의 정의와 측정 방식
- 이상탐지/조기경보 설계 (Z-score, 동적 임계값 등)
- 한국 전력거래소 및 기상청 등에서 제공하는 전력 및 외생 변수 관련 연구 자료

관련 영어/한국어 논문, 보고서, 구현 사례, 실무 자료를 모두 포함해 균형 있게 정리하겠습니다.

조사 완료 후 문헌 요약표, 관련 연구 흐름도, 주요 인용 대상 등을 함께 제공드리겠습니다.

arxiv.org 읽는 중

4 출처

본 단계에서는 2-1 단계에서 설정한 선행연구 조사 흐름을 바탕으로, 연구 주제와 직접적으로 연관된 기술적 맥락과 이론적 배경을 보다 정밀하게 정리하는 것을 목표로 하였다.

특히 Foundation Model 기반 전력 수요 예측과 급변 이벤트 대응, 조기 경보 개념이 기존 시계열 예측 패러다임과 어떤 점에서 구조적으로 구분되는지를 명확히 이해할 필요가 있다고 판단하였다.

이에 따라 연구자는 기존 전력 수요 예측 기법(ARIMA, Exponential Smoothing 등)의 구조적 한계, 기계학습 기반 접근(Random Forest, SVR 등)의 적용 범위, 그리고 최근 시계열 Foundation Model(TimeGPT, Chronos, Lag-LLaMA 등)의 등장 배경과 기술적 특징을 비교·정리하기 위한 보조 수단으로 Gemini의 Deep Research 기능을 활용하였다.

Gemini는 이러한 요청에 따라, 전력 시스템의 차세대 운영 환경과 연계된 기술적 흐름을 중심으로 기존 예측 패러다임과 Foundation Model 접근의 차별점을 구조적으로 정리하는 데 활용되었다.

이를 통해 연구자는 zero-shot 일반화 개념이 기존 학습 기반 예측 방식과 어떻게 구분되는지, 그리고 급변 상황 대응 관점에서 어떤 실무적 의미를 가질 수 있는지를 보다 명확히 파악할 수 있었다.

다만, Gemini의 출력 결과는 연구 주제를 결정하거나 연구 결론을 도출하는 근거로 사용되지 않았으며, 연구자는 해당 설명 중 본 연구의 문제의식과 직접적으로 연결되는 부분만을 선별적으로 참고하였다.

특히 zero-shot 일반화 개념과 기존 예측 패러다임의 차이에 대해서는, 연구자가 선행 문헌과 실제 전력 수요 데이터 분석 결과를 비교·대조하여 최종적으로 해석하였다.

결과적으로 본 단계에서 Gemini의 Deep Research 기능은 최신 시계열 Foundation Model의 기술적 맥락을 정리·점검하는 보조적 역할을 수행하였으며, 연구자는 이를 바탕으로 연구 주제의 이론적 기반을 보다 명확히 정리하고, 이후 실험 설계 및 분석 단계로 자연스럽게 연결할 수 있었다.

2-2 Gemini Deep ressearch 기능 사용

완벽한 연구를 위해서 선행연구를 조사하고 문헌을 검토하려하는데 내 주제에 맞는 논문을 포함해서 연구에 도움이 될 수 있는 모든 자료를 조사해서 정리해줘

✓

착수할게요. 연구가 완료되면 알려 드릴게요. 그러는 동안 이 대화에서 나가셔도 됩니다.

Claude 대화 학습 및 개인화 방안

- 조사 중 웹사이트 47개
- 결과 분석
- 보고서 만들기

<
≡
🔗
만들기 ▾

전력 시스템의 차세대 운영 전략: TimeGPT 기반 수요 예측과 램프 이벤트(Ramp Event) 조기 경보 시스템의 심층 분석

서론: 에너지 전환 시대의 불확실성과 데이터 과학의 역할

21세기 전력 산업은 탈탄소화(Decarbonization), 분산화(Decentralization), 디지털화(Digitalization)라는 3D 트렌드에 의해 근본적인 구조적 변화를 겪고 있습니다. 과거의 전력망은 중앙집중형 대형 발전소가 예측 가능한 부하를 감당하는 단방향 흐름이었으나, 현대의 그리드는 태양광, 풍력 등 변동성 재생에너지(VRE: Variable Renewable Energy)의 비중이 급격히 증가함에 따라 공급 측면의 불확실성이 극대화되고 있습니다. 특히 한국의 경우, 2030 탄소중립 시나리오와 제10차 전력수급기본계획에 따라 재생에너지 보급이 가속화되고 있으며, 이는 제주도와 호남 지역을 중심으로 출력 제어(Curtailment) 및 계통 불안정 문제를 야기하고 있습니다. ▾

이러한 상황에서 전통적인 통계적 시계열 예측 방법론(ARIMA, Exponential Smoothing)이나 단순 기계학습 모델(Random Forest, SVR)은 한계에 봉착했습니다. 기후 변화로 인한 이상 기상의 빈번한 발생, 전기차(EV) 보급에 따른 부하 패턴의 비선형성 증가, 그리고 마이크로그리드와 같은 소규모 분산 자원의 확산은 기존의 "지역적(Local)" 모델링 접근법으로는 해결하기 어려운 복잡성을 띠고 있습니다. 이에 따라 2023년부터 2025년 사이, 거대 언어 모델(LLM)의 성공에 영감을 받은 "시계열 파운데이션 모델(Time Series Foundation Models, TSFMs)"이 새로운 대안으로 부상했습니다. TimeGPT, Chronos, Lag-LLaMA와 같은 모델들은 수십억 개의 데이터 포인트로 사전 학습되어, 제로샷(Zero-shot) 추론이라는 혁신적인 기능을 제공합니다. ▾

본 단계에서는 2-1 단계에서 설정한 선행연구 조사 흐름을 바탕으로, 연구 주제와 직접적으로 연관된 기술적 맥락과 이론적 배경을 점검하고 정렬하는 것을 목표로 하였다.

특히 Foundation Model 기반 전력 수요 예측과 급변 이벤트 대응, 조기 경보 개념이 기존 시계열 예측 접근과 어떤 점에서 구조적으로 구분되는지를 보다 명확히 정리할 필요가 있다고 판단하였다.

이에 따라 연구자는 기존 전력 수요 예측 기법(ARIMA, Exponential Smoothing 등), 기계학습 기반 접근(Random Forest, SVR 등), 그리고 최근 제안된 시계열 Foundation Model(TimeGPT, Chronos, Lag-LLaMA 등)의 등장 배경과 기술적 특징을 비교·정리하기 위한 참고 프레임으로 Gemini의 Deep Research 기능을 활용하였다.

Gemini는 이러한 요청에 따라 관련 기술 흐름과 개념적 차이를 정리하는 데 활용되었으며, 연구자는 이를 통해 기존 학습 기반 예측 패러다임과 zero-shot 일반화 접근이 전제하는 가정의 차이, 그리고 급변 상황 대응 관점에서의 구조적 차별점을 점검하였다.

이 과정은 새로운 연구 방향을 도출하기보다는, 이미 설정된 연구 문제의 이론적 기반이 충분히 타당한지 확인하는 데 목적이 있었다.

다만, Gemini의 출력 결과는 연구 주제의 결정이나 연구 결론 도출에 직접적으로 사용되지 않았으며, 연구자는 해당 내용 중 본 연구의 문제의식과 직접적으로 연결되는 부분만을 선별적으로 참고하였다.

특히 zero-shot 일반화 개념의 실무적 의미와 기존 예측 패러다임과의 차이에 대해서는, 연구자가 선행 문헌 검토 결과와 실제 전력 수요 데이터 분석 맥락을 기준으로 비교·검토하였다.

결과적으로 본 단계에서 Gemini의 Deep Research 기능은 최신 시계열 Foundation Model 관련 기술적 맥락을 보조적으로 점검·정리하는 역할을 수행하였다.

연구자는 이를 바탕으로 연구 주제의 이론적 기반과 기술적 타당성을 보다 명확히 정리한 상태에서 이후 실험 설계 및 분석 단계로 연결할 수 있었다.

2. 연구목적 및 문제 정의

2-3. 선행연구 점검을 통한 연구 공백 진단 및 보완 방향 도출 (Claude 보조 활용)

📊 현재 확보 상태

✅ 충분한 것:

1. Foundation Models (5개) - 완벽

- Chronos, TimeGPT, Lag-LLaMA, TimesFM, PriceFM
- 아키텍처 상세 분석 ✅
- 벤치마크 수치 ✅

2. 실제 사례 (2개) - 강력

- 텍사스 한파 2021
- 한국 폭염 2024

3. 이상탐지 기법 (3개) - 충분

- Z-score, ESD, Dynamic Threshold

❌ 부족한 것 (치명적):

1. 한국 전력 예측 최신 연구 부족

현재:

- 박준호 (2017) - 7년 전
- 정현철 (2018) - 6년 전
- 김영수·박호정 (2021) - 3년 전

문제:

- **2022-2024년** 최신 연구 0개! 🚨
- Foundation Model 적용 한국 연구 0개
- 한국 전력시장 특화 Zero-shot 연구 0개

본 단계에서는 2-1 및 2-2 단계의 문헌 조사 결과를 바탕으로, 현재까지 확보된 선행연구 자료가 연구 목적을 충분히 뒷받침하는지 점검하고, 연구 설계 과정에서 비판 가능성이 높은 연구 공백을 사전에 식별하는 것을 목표로 하였다.

연구자의 목적은 단순히 문헌의 양을 확장하는 것이 아니라, 심사 과정에서 핵심적으로 검토될 가능성이 높은 지점을 구조적으로 정리하는 데 있었다.

연구자는 먼저 기존에 확보한 선행연구를 기준으로, 충분히 검토된 영역과 상대적으로 연구가 부족한 영역을 구분하였다.

Foundation Model 기반 시계열 예측 기법(Chronos, TimeGPT, Lag-LLaMA 등), 이상상황 탐지 기법(Z-score, ESD, 동적 임계값), 그리고 해외 사례 분석 측면에서는 연구 수행에 충분한 선행연구와 참고 자료가 확보되어 있음을 확인하였다.

반면, 국내 전력 수요 예측 분야의 최신 연구 동향(2022-2024)을 검토한 결과, Foundation Model을 적용한 국내 연구 사례가 매우 제한적이며, 특히 국내 전력 수요 데이터를 대상으로 zero-shot 일반화 성능을 체계적으로 분석한 연구는 거의 존재하지 않는다는 점이 확인되었다.

연구자는 이 부분이 본 연구의 차별성과 동시에 심사 과정에서 집중적으로 검토될 가능성이 높은 핵심 취약 지점이라고 판단하였다.

이 과정에서 Claude는 확보된 문헌 목록을 구조적으로 정리하고, 충분한 영역과 상대적으로 부족한 영역을 구분하는 데 참고 자료로 활용되었다.

다만, 어떤 부분을 연구 공백으로 정의할 것인지에 대한 판단과, 해당 공백이 연구 질문과 직접적으로 연결되는지에 대한 최종 결정은 연구자가 기존 문헌 검토 결과와 연구 목적을 기준으로 수행하였다.

본 단계의 분석을 통해 연구자는, 기존 선행연구의 한계가 단순한 데이터 부족이나 사례 부족이 아니라, 국내 전력 시장 맥락에서 Foundation Model과 zero-shot 일반화 개념이 체계적으로 연결되지 않았다는 구조적 공백에 있음을 명확히 인식하였다.

이는 이후 연구 목적과 연구 질문을 구체화하는 핵심 근거로 활용되었다.

결과적으로 본 단계에서는 선행연구 점검을 통해 연구 공백을 보다 명확히 진단하였으며, 해당 공백은 연구자가 최종적으로 설정한 “국내 전력 수요 데이터를 기반으로 한 zero-shot 일반화 성능 분석의 희소성”이라는 연구 문제로 구체화되었다.

Claude는 이 과정에서 연구자의 판단을 보조하는 점검 도구로 활용되었으며, 연구 공백의 정의와 연구 방향 설정은 연구자의 판단에 따라 이루어졌다.

2-4 2차 Deep Research 수행: 갭 기반 재조사 및 근거 강화

좋아. 지금까지 확보한 내용을 바탕으로 부족한 자료들을 중점적으로 조사할게.

다음 5개 항목에 대해 자료를 조사하고 정리할 예정이야:

1. TimeGPT 전력 수요 예측 적용 사례 논문 (5개)
2. Zero-shot load forecasting 관련 최신 논문 (2023~2024년 기준, 3개 이상)
3. 한국 전력 데이터 접근 방식 (KPX, EPSIS, 기상청 API 등)
4. Ramp event detection 최신 기법 및 실제 적용 사례 논문 (3개 이상)
5. 기상 데이터 등 외생변수를 활용한 전력 예측 연구 (2~3개)

모든 자료는 한국어 및 영어 논문 포함해서, 공식 학회지·저널·기술백서·공공기관 API 문서 등을 균형 있게 조사할게. 정리되면 곧 알려줄게.

Research completed in 8m·출처 40개·검색 170건

전력 수요 예측 연구 현황 및 과제

연구 웹사이트

- (1) 'TimeGPT'와 'Foundation Model'을 전력 수요(Electricity Demand/Load) 예측에 적용한 2023~2025년 사이의 실증 논문 및 벤치마크 결과를 집중적으로 검색한다.
- (2) 'Zero-shot load forecasting' 키워드로 에너지 분야의 최신 연구(Chronos, Lag-Llama 등 포함)를 찾아 TimeGPT와의 비교 우위나 한계를 파악한다.
- (3) 한국전력거래소(KPX) 및 전력통계정보시스템(EPSIS)의 데이터 다운로드 절차, 공공데이터포털 API 활용법, 기상청 기상자료개방포털 연동 방법을 구체적으로 조사한다....

결과 분석

보고서 생성

몇 분 후 완료

계획 수정

연구 시작

연구 시작

본 단계에서는 2-3 단계에서 식별된 연구 공백이 특정 문헌 선택이나 조사 편향에 따른 결과가 아닌지 확인하고, 연구 문제 정의의 타당성을 보다 견고히 하기 위해 2차 문헌 재조사 및 교차 점검을 수행하였다.

연구자의 목적은 새로운 연구 방향을 추가로 도출하는 것이 아니라, 기존에 정의한 연구 공백이 다양한 관점에서도 일관되게 유지되는지를 검증하는 데 있었다.

이를 위해 연구자는 Foundation Model 기반 전력 수요 예측, zero-shot 일반화 성능 평가, 급변 이벤트(Ramp Event) 대응, 조기 경보 시스템 설계와 직접적으로 연관된 최신 연구 동향을 중심으로 조사 범위를 재설정하였다.

해당 범위는 2-1 단계에서 설정한 탐색 기준과 동일하게 유지하여, 조사 범위 변경에 따른 결과 왜곡을 최소화하였다.

본 단계에서는 특정 AI의 단일 분석 결과에 의존하지 않기 위해, GPT와 Gemini를 병렬적으로 활용하여 동일한 조사 범위에 대한 교차 점검을 수행하였다.

두 도구에는 동일한 조사 조건과 범위를 제시하였으며, 연구자는 각 도구의 출력 결과를 비교함으로써 특정 관점에 치우친 해석이나 누락 가능성을 점검하였다.

교차 점검 결과, Foundation Model의 zero-shot 일반화 성능이 기존 학습 기반 예측 모델 대비 갖는 잠재적 장점은 공통적으로 언급되었으나, 장기 예측 중심의 대규모 모델 비교 접근이나 국내 전력 수요 데이터 기반의 실증 연구는 여전히 제한적이라는 점이 일관되게 확인되었다.

연구자는 이러한 공통된 관찰 결과를 통해, 2-3 단계에서 정의한 연구 공백이 특정 분석 도구의 편향에 기인하지 않았음을 재확인하였다.

또한 일부 도구에서 제시된 장기 예측 중심의 비교 연구 방향은 본 연구의 목적 및 데이터 특성과 부합하지 않는다고 판단하여, 연구 범위에서 제외하였다.

이 과정에서 연구자는 연구 질문과 직접적으로 연결되지 않는 분석 확장을 사전에 통제함으로써 연구 설계의 일관성을 유지하였다.

결과적으로 본 단계에서는 GPT와 Gemini를 다중 참고 프레임으로 활용하여 문헌 조사 결과를 교차 점검하였으며, 이를 통해 연구 공백의 타당성과 연구 문제 정의의 안정성을 강화하였다.

다만, 어떤 문헌을 최종적으로 포함할지, 그리고 해당 결과를 연구 설계에 어떻게 반영할지는 연구자의 판단에 따라 결정되었다.

본 단계의 결과는 이후 연구 목적 및 실험 설계의 타당성을 뒷받침하는 근거로 활용되었다.

2-5 Claude를 이용하여 최종 문헌 정제 기준 및 산출물 정리



chat gpt 추가자료야

오후 1:50



gemini 추가자료 보강이야

오후 1:51

최종 통합 문헌 리스트

Foundation Models (5개)

1. Chronos (Amazon)
2. TimeGPT (Nixtla)
3. Lag-LLaMA (ServiceNow)
4. TimesFM (Google)
5. PriceFM

TimeGPT 실증 (6개)

본 단계에서는 2-4 단계까지의 문헌 재조사 및 교차 점검 결과를 바탕으로, 본 연구에 실제로 활용할 최종 문헌 목록을 정제·확정하는 것을 목표로 하였다.

연구자의 목적은 문헌의 수를 늘리는 것이 아니라, 연구 문제와 직접적으로 연결되는 핵심 문헌만을 선별하여 선행연구 검토의 일관성과 밀도를 확보하는 데 있었다.

연구자는 문헌 정제를 위해
 △연구 주제와의 직접적 관련성
 △실증 가능성 및 재현 가능성
 △중복 여부
 △연구 목적과의 정합성을 기준

으로 최종 문헌 선정 기준을 설정하였다. 이 과정에서 GPT와 Gemini를 통해 수집된 추가 참고 자료는 문헌 누락 여부 및 중복 여부를 점검하기 위한 참고 목록으로 활용되었으며, 해당 자료가 연구에 포함될지 여부에 대한 판단은 연구자가 직접 수행하였다.

Claude는 이러한 정제 기준에 따라 문헌 목록을 정리하는 과정에서 자료 정렬과 중복 점검을 보조하는 도구로 활용되었다.

연구자는 Claude가 정리한 목록을 기반으로 각 문헌이 연구 주제인 Foundation Model 기반 전력 수요의 zero-shot 일반화 성능 분석과 직접적으로 연결되는지를 검토하고, 연구 범위 및 데이터 특성과 부합하지 않는 문헌은 제외하였다.

그 결과, Foundation Model 관련 핵심 연구(Chronos, TimeGPT, Lag-LLaMA, TimesFM, PriceFM)와 TimeGPT 기반 실증 연구를 중심으로 한 최종 통합 문헌 리스트를 확정하였다.

해당 문헌들은 본 연구의 이론적 배경과 실험 설계, 그리고 분석 결과 해석을 뒷받침하는 데 활용되었다.

정리하면, 본 단계에서는 연구자가 설정한 문헌 정제 기준을 바탕으로 최종 문헌 목록을 확정하였으며, AI는 문헌 정리 과정의 효율성을 높이기 위한 보조 도구로 활용되었다.

문헌 포함 여부에 대한 최종 판단과 연구 설계와의 연결은 연구자의 판단에 따라 이루어졌다.

3-1 연구 배경 및 문제 상황 정리

요즘 전력 수요 예측 쪽 연구들 보면,
실제 운영이나 급변 상황 대응에서는
어떤 한계들이 자주 언급되는지 궁금해.
새로운 연구 주제를 만들어달라는 건 아니고,
기존 연구들에서 공통적으로 말하는
'문제 상황'만 정리해줄 수 있을까?

전력 수요 예측 연구의 공통 한계점 정리

문헌에서 반복적으로 언급되는 5대 한계

본 단계에서는 기존 전력 수요 예측 연구 흐름을 전력 시스템 실제 운영 관점에서 재정리하고, 본 연구가 다루고자 하는 문제 상황을 명확히 구조화하는 데 초점을 두었다.

연구자의 목적은 새로운 예측 기법을 제안하는 것이 아니라, 기존 연구들이 공통적으로 전제하고 있는 가정과 그로 인해 반복적으로 발생하는 한계가 실제 운영 환경에서 어떤 문제로 이어지는지를 정리하는 것이었다.

이 과정에서 AI는 단순한 배경 설명 도구가 아니라, 기존 전력 수요 예측 연구에서 반복적으로 언급되는 한계 요소를 운영 관점에서 재분류하는 사고 정렬 도구로 활용되었다.

연구자는 AI와의 질의를 통해 선행연구에서 빈번히 언급되는 문제 요소들을 비교·정리하고, 이를 실제 전력 시스템 운영 리스크와 연결하여 구조화하였다.

그 결과, 기존 전력 수요 예측 연구는 정상적인 수요 패턴이 반복되는 구간에서는 일정 수준의 성능을 보이지만, 학습 데이터 분포를 벗어나는 상황(out-of-distribution)이나 새로운 수요 패턴이 발생할 경우 예측 성능이 급격히 저하되는 한계를 공통적으로 지니고 있음을 확인하였다.

이러한 성능 저하는 단순한 예측 오차 증가를 넘어, 실제 전력 시스템 운영에서 의사결정 지연과 리스크 확대 문제로 직결될 수 있다.

특히 급격한 수요 변화가 발생하는 구간(Ramp Event)에서는 예측 지연, 과소·과대 추정 문제가 반복적으로 보고되며, 폭염·한파와 같은 극한 기상 이벤트 상황에서도 기존 통계적·학습 기반 모델의 취약성이 빈번히 관찰된다.

그러나 다수의 선행연구에서는 이러한 취약성이 예측 정확도 지표 중심으로만 논의되고, 운영 안정성이나 조기 대응 관점에서 충분히 해석되지 않는 경향이 있었다.

AI와의 문제 구조화 과정을 통해 연구자는, 실제 전력 시스템 운영 환경에서는 단순한 평균 예측 정확도보다 안정성, 신뢰성, 그리고 급변 상황에 대한 조기 대응 가능성이 더 중요한 평가 기준이 될 수 있음을 명확히 인식하였다.

이는 기존 연구에서 예측 오차와 운영 리스크 간의 연결이 상대적으로 약하게 다루어졌다는 문제의식으로 이어졌다.

이러한 문제 상황 정리를 바탕으로 연구자는, 본 연구가 전력 수요 예측 정확도 향상 자체보다는 급변 수요 상황에서의 예측 성능 변화와 그 운영적 의미를 분석하는 방향으로 연구 배경을 설정하였다.

이는 이후 연구 질문과 분석 프레임워크를 설계하는 출발점으로 활용되었다.

3-2 연구 목적 및 연구 질문 구조화

내가 생각한 연구 방향이 있는데,
이걸 연구 목적이랑 연구 질문 형태로
좀 깔끔하게 정리해보고 싶어.
방향을 새로 제안하지는 말고,
지금 내용이 논리적으로 잘 보이게만
정리해주면 좋겠어.

- Foundation Model 기반 전력 수요 예측
- Zero-shot 일반화 성능
- 급변 상황(Ramp Event)과 조기경보

📄 연구 목적 및 연구 질문 정리

🎯 연구 목적 (Research Objectives)

주 목적 (Primary Objective):

시계열 파운데이션 모델(Time-Series Foundation Model)의 Zero-shot 일반화 능력을 활용하여,

본 단계에서는 3-1 단계에서 정리한 전력 수요 예측 분야의 문제 상황을 바탕으로, 본 연구에서 다루고자 하는 연구 목적과 연구 질문을 명확한 구조로 정리하였다.

연구자의 목적은 새로운 연구 주제를 추가로 탐색하는 것이 아니라, 앞선 단계에서 확립된 문제의식을 논리적으로 일관된 연구 목적과 질문 체계로 구체화하는 데 있었다.

연구자는 시계열 파운데이션 모델(Time-Series Foundation Model), zero-shot 일반화 성능, 급변 수요 구간(Ramp Event), 그리고 조기 경보(Early Warning) 관점이라는 핵심 요소를 중심으로 연구 목적의 범위와 초점을 설정하였다.

이 과정에서 연구자는 기존 전력 수요 예측 연구가 주로 예측 정확도 향상에 초점을 맞추어 온 반면, 학습 데이터 분포를 벗어나는 상황에서도 안정적인 예측 성능을 유지할 수 있는지에 대한 검증이 충분히 이루어지지 않았다는 점에 주목하였다.

이에 따라 본 연구의 주된 목적은, 시계열 파운데이션 모델의 zero-shot 일반화 성능을 활용하여 기존 학습 기반 전력 수요 예측 모델이 갖는 구조적 한계를 보완할 수 있는지를 검증하는 데 있다.

이는 단순한 예측 성능 비교를 넘어, 급변 수요 상황에서도 예측 신뢰성과 운영 안정성을 확보할 수 있는지를 평가하는 것을 포함한다.

특히 연구자는 전력 수요의 급격한 변화가 발생하는 구간(Ramp Event)에서 기존 모델 대비 예측 오차가 어떻게 변화하는지, 그리고 이러한 변화가 조기 경보(Early Warning) 관점에서 실질적인 의미를 가질 수 있는지를 주요 분석 대상으로 설정하였다.

이를 통해 예측 성능과 실제 전력 시스템 운영 리스크 간의 연결 관계를 보다 명확히 분석하고자 하였다.

이에 따라 본 연구의 연구 질문은 다음과 같이 구조화되었다.

1. 학습 데이터 분포를 벗어나는 전력 수요 환경에서도 시계열 파운데이션 모델은 zero-shot 방식으로 안정적인 일반화 성능을 보일 수 있는가?
2. 급변 수요 구간(Ramp Event)에서 시계열 파운데이션 모델의 예측 오차 감소 효과는 기존 학습 기반 모델 대비 일관되게 나타나는가?
3. 예측 오차 변화가 조기 경보(Early Warning) 관점에서 전력 시스템 운영 신뢰성 향상으로 연결될 수 있는가?

이와 같은 연구 목적과 질문 구조를 통해, 연구자는 문제 인식 단계에서 도출한 기존 연구의 한계를 명확한 연구 질문으로 연결하였으며, 이후 실험 설계 및 분석 단계에서 검증 가능한 형태로 연구를 수행할 수 있는 기반을 마련하였다.

3-3. 선행연구 점검을 통한 연구 공백 인식 및 정리

지금까지 조사한 선행연구들을 보면,
전반적인 흐름은 이해되는데
뭔가 빠진 느낌도 드는 것 같아.
최근 연구 흐름이나
한국 전력시장 맥락에서 봤을 때,
부족하거나 거의 다뤄지지 않은 부분이
뭐가 있는지 좀 솔직하게 짚어줄 수 있을까?
새로운 주제를 제안하진 말고,
'아쉬운 점' 위주로 말해줘.

🔍 현재 문헌 조사의 아쉬운 점 (솔직 평가)

📊 전체 평가: 95/100

장점: Foundation Model 이론 + 실증 + 한국 데이터 접근 약점: 아래 5가지

본 단계에서는 선행연구 조사 결과를 바탕으로, 전력 수요 예측 분야에서 기존 연구들이 반복적으로 다루어 온 문제 구조와 충분히 해소되지 않은 연구 공백을 연구자의 관점에서 명확히 인식하고 정리하는 데 초점을 두었다.

연구자의 목적은 새로운 연구 주제를 제안하는 것이 아니라, 기존 연구 흐름이 도달한 지점과 그 한계를 보다 정밀하게 구분하는 것이었다.

이 과정에서 AI는 단순 요약 도구가 아니라, 연구자가 인식한 문제 구조를 반복적으로 점검하고 교란하는 사고 개입자로 활용되었다.

연구자는 선행연구를 검토하며 형성한 문제 인식에 대해 AI에게 반례 가능성, 누락된 관점, 도메인 특수성의 영향 등을 지속적으로 질의함으로써 기존 인식의 타당성을 검증하였다.

그 결과, 시계열 파운데이션 모델(TimeGPT, Chronos, Lag-LLaMA 등)에 대한 해외 연구는 축적되고 있으나, 국내 전력 수요 데이터를 활용한 최신 연구(2022-2024)는 제한적이며, 특히 zero-shot 일반화 성능을 국내 전력 시장 맥락에서 체계적으로 분석한 연구는 거의 존재하지 않음이 확인되었다.

또한 기존 전력 수요 예측 연구 다수가 평균 오차 지표(MAPE, RMSE 등)에 평가 초점을 두고 있으며, 급변 수요 구간(Ramp Event)이나 극한 상황에서의 예측 오차가 운영 리스크 및 조기 경보 관점에서 어떤 의미를 가지는지에 대한 분석은 상대적으로 부족하다는 점이 명확해졌다.

연구자는 이러한 한계가 단순한 데이터 부족이 아니라, 국내 전력 시장이라는 특정 도메인에서 최신 시계열 파운데이션 모델과 zero-shot 일반화 개념을 운영 안정성 및 조기 대응 문제와 연결하는 연구 프레임이 부족하기 때문이라고 판단하였다.

본 단계에서 AI는 연구 공백을 직접 정의하지는 않았으나, 연구자가 설정한 문제 인식이 기존 문헌 흐름에 비추어 충분히 타당한지를 점검하는 역할을 수행하였다.

최종적인 연구 공백의 정의와 연구 방향 설정은 이러한 AI 피드백을 참고하여 연구자의 판단에 따라 이루어졌다.

3-4. 연구 주제의 논리적 정합성 점검 및 최종 확정

앞에서 이야기한 연구 공백들을 고려했을 때,
내가 설정한 연구 주제가
논리적으로 무리 없는지 궁금해.
최종 결정은 내가 할 거고,
너는 참고 의견만 편하게 말해줘.

연구 주제 논리성 평가(솔직 의견)

귀하의 연구 주제 (재확인)

"Foundation Model 기반 전력 수요의 Zero-shot 일반화 성능 분석 및 급변 상황 조기경보 시스템"

핵심:

본 단계에서는 선행연구 공백 진단 결과를 바탕으로, 연구자가 설정한 연구 주제가 기존 연구 흐름 및 문제 인식과 논리적으로 무리 없이 연결되는지를 최종적으로 점검하였다.

연구자의 목적은 연구 주제 채택 여부를 AI에 위임하는 것이 아니라, 자신의 판단이 기존 연구 맥락에서 일관성을 갖는지 스스로 검증하는 데 있었다.

이 과정에서 AI는 연구 주제에 대한 최종 결정을 내리는 역할이 아니라, 연구자가 설정한 문제 인식과 연구 목표 사이의 연결 구조를 점검하는 사고 검증자로 활용되었다.

연구자는 AI에게 연구 주제의 논리적 연결성, 기존 연구 흐름과의 정합성, 그리고 연구 공백 진단 결과와의 일치 여부에 대해 자유롭게 의견을 요청하였다.

AI의 피드백은 연구 주제 자체의 타당성보다는, 기존 전력 수요 예측 연구에서 반복적으로 지적된 한계(학습 기반 모델의 재학습 의존성, 급변 상황 대응 한계, 국내 전력 시장 특화 연구의 부족)가 본 연구 주제와 논리적으로 연결되어 있는지를 점검하는 데 집중되었다.

그 결과, 연구자는 시계열 파운데이션 모델의 zero-shot 일반화 성능을 중심으로 급변 상황(Ramp Event)과 조기 경보 문제를 함께 분석하는 연구 구조가, 최근 시계열 예측 연구 흐름과 자연스럽게 부합함과 동시에 기존 연구의 한계를 보완할 수 있는 방향임을 확인하였다.

또한 본 연구 주제가 단순한 예측 정확도 비교를 넘어, 예측 성능 변화가 실제 전력 시스템 운영 안정성과 조기 대응 측면에서 어떤 의미를 가지는지를 분석 대상으로 포함하고 있다는 점에서 연구 주제로서의 실질적 가치가 충분하다고 판단하였다.

종합적으로 연구자는 선행연구 공백 진단과 AI 기반 논리 점검 결과를 바탕으로, 본 연구 주제가 기술적·학술적·실무적 측면에서 모두 타당하다고 판단하여 해당 주제를 최종 연구 주제로 확정하였다.

이 과정에서 AI는 판단을 대체하지 않고, 연구자의 최종 결정을 보다 엄격하게 검증하는 역할을 수행하였다.

이와 같이 연구자는 AI와의 반복적 사고 점검 과정을 통해, 연구 주제 설정 단계에서 발생할 수 있는 주관적 판단 편향을 사전에 통제하고, 연구 질문과 분석 방향을 보다 엄격하게 고정할 수 있었다.

4. 연구계획 및 방법론 설정

4-1 연구 설계 개요

우리가 설계한 연구 설계 흐름을 전체 단계 순서나 논리 연결이 자연스러운지만 한번 봐줄 수 있을까?
방법을 새로 제안해주거나 구조를 바꾸자는 건 아니고, 그냥 큰 흐름에서 이상한 데가 없는지만 체크해주면 좋겠어.

연구 설계 흐름 논리성 체크

현재 연구 흐름 (재구성)

본 단계에서는 연구자가 사전에 설계한 연구 흐름이 단계 간 논리적 비약 없이 일관되게 연결되는지를 점검하기 위해 AI를 활용하였다.

연구자는 전체 연구 구조를
문제 인식 → 데이터 수집 → 모델 비교 → 급변 상황 분석 → 조기 경보 검증
의 순차적 흐름으로 이미 확정한 상태였으며,

본 단계에서는 새로운 방법론 제안이나 구조 변경을 요청하지 않았다.

AI에는 다음과 같은 검토 기준에 한해 점검을 요청하였다.

- (1) 연구 질문과 실험 설계 간의 대응 관계
- (2) Zero-shot 일반화 검증과 급변 상황 분석 단계의 분리 적절성
- (3) 기존 전력 수요 예측 연구 흐름과의 구조적 일관성
- (4) 분석 단계 간 역할 중복 또는 논리적 공백 존재 여부

검토 결과, 전체 연구 흐름은 기존 전력 수요 예측 및 이상 상황 분석 연구의 일반적 구조를 따르면서도, 급변 상황(Ramp Event)을 별도 분석 단계로 분리한 점에서 연구 목적과의 정합성이 유지된다는 판단을 확인하였다.

연구자는 이러한 검토 결과를 참고하여 기존 연구 설계의 타당성을 재확인하였으며, 연구 구조 수정 없이 계획된 흐름을 그대로 유지하기로 결정하였다.

즉, 본 단계에서 AI는 설계된 연구 프로세스의 논리적 정합성을 검증하는 감사인 역할을 수행하였다.

특히 앞서 Appendix에서 정의한 '24-hour rolling-origin' 프로토콜의 단계별 연결성을 AI와 교차 점검하여, 실험의 재현 가능성을 기술적으로 확정하였다.

특히 본 연구는 재학습 없는 일반화 성능 검증을 핵심으로 하므로, 학습·검증·분석 단계의 역할이 혼재되지 않도록 연구 단계 간 기능적 구분을 명확히 유지하는 데 중점을 두었다.

본 연구 설계는 새로운 방법론을 제안하기보다, 기존 연구 흐름을 유지한 상태에서 zero-shot 평가와 급변 상황 분석을 명시적으로 분리·강조한 구조라는 점에서 방법론적 타당성을 가진다.

4-2 데이터 구성 관련 참고 정리

전력 수요 예측 관련 연구들 보면 보통 어떤 종류의 데이터들을 많이 쓰는지 한 번 정리해서 보고 싶어.
특정 데이터를 추천해주기보다는, 문헌에서 자주 등장하는 데이터 유형들만 정리해주면 될 것 같아.

전력 수요 예측 연구의 데이터 유형 정리 (문헌 기반)

분석 대상 문헌

글로벌 연구:

- Liao et al. (2024, 2025)

본 단계에서는 연구자가 설정한 데이터 구성 전략이 기존 전력 수요 예측 연구 흐름과 정합적인지를 검증하고, 데이터 선택 과정에서 발생할 수 있는 편향과 누락 가능성을 점검하는 데 초점을 두었다.

연구자의 목적은 새로운 데이터를 추가로 제안받는 것이 아니라, 본 연구에서 사용하지 않기로 한 데이터 선택 역시 논리적으로 정당화하는 데 있었다.

이 과정에서 AI는 특정 데이터를 추천하는 역할이 아니라, 기존 전력 수요 예측 연구에서 반복적으로 활용되는 데이터 유형을 구조적으로 정리하고, 각 데이터 유형이 연구 목적과 어떤 관계를 가지는지를 검토하는 데이터 검증 도구로 활용되었다.

연구자는 AI에게 문헌에서 자주 사용되는 데이터 범주를 정리하도록 요청하고, 이를 바탕으로 본 연구의 데이터 구성 전략을 점검하였다.

그 결과, 기존 전력 수요 예측 연구에서는 과거 전력 수요 시계열, 기상 변수, 달력 변수, 가격·시장 변수 등이 주로 활용되며, 최근 연구일수록 외생 변수(exogenous variables)를 포함하는 경향이 강화되고 있음을 확인하였다.

그러나 이러한 데이터 확장은 학습 기반 모델의 성능 향상에는 기여할 수 있으나, zero-shot 일반화 성능을 검증하는 본 연구의 목적과는 반드시 일치하지 않을 수 있음이 확인되었다.

이에 따라 연구자는 AI와의 점검 과정을 통해, 본 연구에서는 학습 데이터 확장에 의존하는 외생 변수 추가보다는, 모델 자체의 일반화 능력을 평가하는 데 초점을 두는 데이터 구성 전략이 타당하다는 판단을 강화하였다.

이는 데이터 구성 단계에서부터 zero-shot 일반화 검증이라는 연구 목적을 일관되게 유지하기 위한 선택이었다.

본 단계에서 AI는 데이터 구성 방향을 결정하지는 않았으며, 연구자가 설정한 데이터 선택 기준이 기존 연구 흐름과 비교하여 과도하거나 결핍되지 않았는지를 점검하는 역할을 수행하였다.

이를 통해 연구자는 이후 실험 단계에서 데이터 선택 자체에 대한 불필요한 논쟁 가능성을 사전에 차단할 수 있었다.

4-3 연구 설계 실현 가능성 점검

지금 내가 생각한 연구 설계가 현실적으로 가능한지 한 번 점검해보고 싶어.

- * KPX/EPSS 전력 수요 데이터
- * 기상청 기온 데이터랑 전력기상지수
- * 시간별 수요 예측 + 급변 상황 탐지
- * Zero-shot 또는 최소 학습 방식

이 조합에서 조금 무리일 수 있는 부분이나 주의해야 할 점이 있다면 편하게 말해줘.

🔍 연구 설계 현실성 점검 (솔직 버전)

📄 귀하의 연구 설계 (재확인)

본 단계에서는 연구자가 구상한 연구 설계가 실제 수행 가능한 범위 내에 있는지, 그리고 연구 과정에서 과도한 가정이나 실현 곤란 요소가 포함되어 있지는 않은지를 점검하는 데 초점을 두었다.

연구자의 목적은 연구 방향이나 방법을 변경하는 것이 아니라, 현재 설계된 연구 구조가 현실적인 제약 조건 하에서도 수행 가능한지를 사전에 검증하는 데 있었다.

이 과정에서 AI는 연구 설계를 대신 평가하거나 승인하는 역할이 아니라, 연구 설계의 취약 가능성을 의도적으로 드러내는 Red Team 검증자로 활용되었다.

연구자는 AI에게 연구 설계 전반을 제시한 후, 실현 가능성 측면에서 주의가 필요한 지점과 잠재적 리스크를 중심으로 비판적 검토 의견을 요청하였다.

AI의 검토 결과, KPX/EPSS 전력 수요 데이터, 기상청 기온 데이터, 전력기상지수(PWI)를 결합한 시간 단위 수요 예측 및 급변 상황(Ramp Event) 분석은 기존 연구 흐름과 데이터 접근성 측면에서 기술적으로 수행 가능한 범위에 속함이 확인되었다.

다만 AI는 연구 설계 전반에 대해 다음과 같은 주의 요소를 함께 제시하였다.

첫째, 급변 상황(Ramp Event)의 정의가 불명확할 경우, 이상탐지 및 조기 경보 성능 해석이 모호해질 수 있으므로, 사전에 명확한 기준 설정이 필요하다는 점이 지적되었다.

둘째, zero-shot 또는 최소 학습 방식의 경우, 단순 예측 정확도 비교보다는 일반화 성능이 유지되는 조건과 한계를 명확히 구분하여 분석할 필요성이 제기되었다.

연구자는 이러한 지적을 반영하여, 급변 상황 정의 기준과 평가 지표를 연구 설계 단계에서 명확히 규정하고, zero-shot 일반화 성능을 과도하게 일반화하지 않도록 분석 범위를 조정하였다.

이를 통해 연구 설계의 현실성과 분석 결과의 해석 가능성을 동시에 확보하고자 하였다.

본 단계에서 AI는 연구 설계의 타당성을 결정하지 않았으며, 연구자가 설정한 설계가 실험적으로 과도하지 않으며 재현 가능한지를 점검하는 비판적 검증 도구로 활용되었다.

연구자는 이러한 Red Team 점검 결과를 바탕으로 연구 설계를 보완하고, 이후 실험 단계로 진행하기로 결정하였다.

4-4 보완 방향 정리

앞에서 말해준 주의점들을 기준으로, 내 연구 계획에서 조금만 보완하면 좋을 부분이 있을까?
전체 방법을 바꾸자는 건 아니고, 지금 계획을 더 탄탄하게 만드는 방향이면 좋겠어.

🔧 연구 계획 보완 제안 (현재 설계 유지)

📋 현재 설계 (재확인)

본 단계에서는 앞선 연구 설계 실현 가능성 점검 과정에서 도출된 주의사항과 잠재적 리스크를 바탕으로, 연구 설계의 기본 구조를 유지한 상태에서 명확성과 해석 가능성을 보완하는 데 초점을 두었다.

연구자의 목적은 연구 방법을 변경하거나 새로운 분석 절차를 추가하는 것이 아니라, 기존 연구 계획을 보다 견고하게 만드는 것이었다.

이 과정에서 AI는 연구 설계를 대체하거나 수정하는 역할이 아니라, 이미 확정된 연구 설계가 실제 실험 및 해석 단계에서 오해 없이 적용될 수 있도록 세부 기준을 정제하는 보완 도구로 활용되었다.

연구자는 AI에게 설계 변경이 아닌, 보완이 필요한 지점을 중심으로 점검 의견을 요청하였다.

AI 검토 의견을 바탕으로 연구자는 다음과 같은 보완 사항을 연구 계획에 반영하였다.

첫째, 급변 상황(Ramp Event)의 정의 기준을 사전에 명확히 설정하였다. 급변 구간의 크기와 지속 시간에 대한 기준을 명시함으로써, 이후 이상탐지 및 조기 경보 성능 해석에서 발생할 수 있는 모호성을 최소화하였다.

둘째, zero-shot 또는 최소 학습 방식의 분석 해석 범위를 조정하였다. 단순 예측 정확도 향상이 아닌, 학습 데이터 분포를 벗어난 상황에서도 일반화 성능이 유지되는지를 핵심 평가 기준으로 명확히 규정하였다.

셋째, 데이터 정합성 및 시간 단위 일관성을 연구 설계 단계에서 재점검하였다.

전력 수요 데이터와 기상 변수 간 시간 해상도 차이로 인해 발생할 수 있는 분석 오류를 방지하기 위해, 데이터 전처리 기준을 사전에 명시하였다.

이러한 보완 사항은 연구 방향이나 방법론을 변경하기 위한 것이 아니라, 기존 연구 설계를 논리적으로 정교화하고 실험 결과의 신뢰도를 높이기 위한 조치로 반영되었다.

본 단계에서 AI는 연구 설계의 최종 판단을 내리지 않았으며, 연구자가 설정한 계획이 실험 단계에서 안정적으로 수행될 수 있도록 잠재적 해석 리스크를 사전에 제거하는 역할을 수행하였다.

5. 자료(데이터) 수집

5-1. 연구 목적 기반 데이터 요구사항 정의

24시간 ahead 전력수요 예측 성능을 평가하는 연구를 설계하려고 하는데 평균 정확도뿐 아니라 폭염·급변 구간 같은 극한 상황에서의 성능을 분석하려면 어떤 종류의 데이터가 필수적이지?

필수는 "전력수요(타깃) + 극한상황을 정의/검증할 변수 + 시간 맥락(캘린더) + 재현가능한 메타정보" 4종 세트야. 너 연구(24h-ahead, 극한 구간 성능 비교) 기준으로 최소 필수와 있으면 강해지는 옵션을 나눠서 정리해줄게.

1) 최소 필수 데이터 (이 4개만 있어도 연구 성립)

(A) 시간별 전력수요(타깃)

Foundation model의 zero-shot 일반화 성능을 평가하려면 데이터 분할을 어떻게 설계하는 것이 타당해?

zero-shot "일반화 성능"을 제대로 보려면, 핵심은 (1) 미래 구간 완전 홀드아웃, (2) 어떤 형태의 정보 누출(leakage)도 차단, (3) 평가가 현실 운영과 동일한 시점 정보만 사용 이 3가지를 만족하는 분할이 타당해. 너 연구(24시간 ahead rolling, 2021-2023 컨텍스트 + 2024 OOS) 기준으로 가장 설득력 있는 설계를 정리해줄게.

1) 가장 타당한 기본 설계: "연도 기반 미래 홀드아웃 + 롤링 평가"

✓ 추천 분할

- Context(과거 컨텍스트): 2021-2023
- Test(완전 홀드아웃): 2024 (절대 참조/튜닝 금지)
- 평가 방식: 24h-ahead rolling-origin(rolling backtest)

왜 이게 zero-shot에 맞나?

본 단계에서는 24시간 ahead 전력 수요 예측 성능을 평가하고, 평균 예측 정확도뿐 아니라 폭염·급변 구간과 같은 극한 상황에서의 성능과 Zero-shot 일반화 능력을 함께 검증하기 위해, 연구 목적에 부합하는 데이터 요구사항을 명확히 정의하는 데 초점을 두었다.

연구자는 데이터 목록을 단순히 나열하기보다는, Zero-shot 일반화 성능을 평가하기에 적절한 데이터 분할 방식과 평가 구조가 무엇인지 점검하기 위해 AI를 보조 검토 도구로 활용하였다.

이 과정에서 AI에는 데이터 설계를 제안하도록 요청하지 않고, 연구 목적에 부합하는 설계 조건을 검토·정리하도록 요청하였다.

AI 검토 결과, 연도 기반 데이터 홀드아웃 방식(2021-2023년 학습 컨텍스트, 2024년 테스트)과 24시간 ahead rolling 평가 구조가, 실제 운영 환경에서의 미래 예측 상황을 가장 자연스럽게 모사할 수 있다는 점이 정리되었다.

연구자는 이러한 검토 내용을 참고하여, 특정 이벤트 구간을 임의로 분리하는 방식보다 연속적인 시간 흐름 속에서 미래 예측 성능을 평가하는 것이 연구 목적과 일관된다는 판단을 내렸다. 이에 따라 본 연구의 데이터 요구사항을 다음과 같이 정의하였다.

(1) 시간별 전력 수요 데이터(타깃 변수), (2) 시간 정렬이 가능한 기상 변수, (3) 날짜·요일·공휴일 등 캘린더 정보, (4) 극한 상황 분석을 위한 시간 연속성 조건

본 데이터 요구사항은 모델 성능 극대화를 위한 확장이 아니라, zero-shot 일반화 성능과 급변 상황 분석을 동시에 검증하기 위한 최소 조건(minimum requirement)으로 설정되었다

본 단계에서 제로샷(Zero-shot) 일반화 성능 평가의 엄밀성을 확보하기 위해 AI와 '실험 프레임워크 설계' 감사를 공동 수행하였다.

AI가 제시한 '연도 기반 미래 홀드아웃 + 24시간 롤링 평가' 구조의 기술적 타당성을 연구자가 한국 전력 계통의 계절성 특성에 비추어 최종 확정하였다.

5-2. 데이터 후보 식별 및 출처 탐색

한국에서 시간별 전력수요 데이터를 연구 목적으로 사용할 수 있는 신뢰 가능한 공공 데이터 출처는 어디?

생각하는 과정 표시

연구 목적으로 한국의 시간별 전력수요 데이터를 활용하시려는군요. 중앙대 전기공학과 전공자로서 심도 있는 분석(예: 예측 모델링, 계통 부하 분석 등)을 수행하기에 가장 적합하고 신뢰할 수 있는 공공 데이터 출처 4곳을 정리해 드립니다.

1. 전력거래소 (KPX) - EPSIS 전력통계정보시스템

대한민국 전력 계통 운영의 심장부인 전력거래소에서 운영하는 곳으로, 연구용 데이터의 표준입니다.

- 제공 데이터: 시간별 전력수요(MW), 시간별 계통한계가격(SMP), 발전원별 발전 실적 등.

전력수요와 함께 사용할 수 있는 시간 단위 기상 데이터(기온 등)를 제공하는 공공 기관이 있나

생각하는 과정 표시

전력수요 분석이나 예측 연구에서 기상 데이터는 수요 변동을 설명하는 가장 중요한 독립 변수입니다. 한국에서 시간 단위의 신뢰성 있는 기상 데이터를 제공하는 기관과 구체적인 수집 방법을 정리해 드립니다.

1. 기상청 기상자료개방포털

기상청에서 운영하는 연구자용 데이터 포털로, 과거 이력 데이터를 대량으로 내려받기에 가장 적합한 곳입니다.

- 제공 데이터: * ASOS (중관기상관측): 전국 약 100여 개 지점에서 관측하는 고품질 데이터 (기온, 강수량, 풍속, 습도, 일사량, 일조시간 등).
- AWS (방재기상관측): 전국 500여 개 지점의 촘촘한 관측망 데이터 (기온, 풍속 등 기본 요소 중심).
- 특징: 연구 목적의 시계열 분석을 위해 시간 단위(Hourly) 데이터를 CSV나 엑셀 파일로 제공합니다.

본 단계에서는 5-1에서 정의한 데이터 요구사항을 충족할 수 있는 실제 데이터 출처를 식별하고, 공공성·접근성·재현성 측면에서 연구에 적합한 후보를 검증하는 데 초점을 두었다.

이 단계의 목적은 모델 성능 극대화가 아니라, 장기적으로 재현 가능한 연구 설계를 가능하게 하는 데이터 인프라를 확보하는 것이었다.

연구자는 AI를 활용하여

- (1) 한국 전력 수요 데이터를 연구 목적으로 활용할 수 있는 공신력 있는 기관,
 - (2) 시간 해상도(hourly) 및 장기간 연속 데이터 제공 여부,
 - (3) 공개 범위 및 접근 제약 조건
- 을 중심으로 데이터 후보를 체계적으로 식별·정리하였다.

검토 결과, 전력 수요 데이터의 경우 KPX 및 EPSIS(전력거래소) 자료가 한국 전력 시스템을 대표하는 공식 통계로서 신뢰성과 연속성을 모두 충족하는 것으로 판단되었다.

해당 데이터는 시간 단위 수요(Electricity Demand)를 장기간 제공하며, 급변 상황(Ramp Event) 분석 및 24-hour ahead 예측 평가에 필수적인 시간 해상도를 확보하고 있다.

기상 변수의 경우, 급변 수요와의 연관성을 고려하여 전국 단위의 시간별 기상 정보를 안정적으로 제공하는 기상청 ASOS/AWS 자료를 주요 후보로 검토하였다.

이들 자료는 전력 수요 데이터와 시간 해상도 및 기간 측면에서 정합성이 높고, 공개 API 또는 정형 데이터 형태로 장기 접근이 가능하다는 점에서 연구 목적에 부합하였다.

데이터 후보 탐색 과정에서는

시간 해상도(Hourly 제공 여부), CSV 또는 API 기반 수집 가능성, 장기간 연속 데이터 확보 가능성을 핵심 판단 기준으로 삼았으며, 단순한 예측 정확도 향상 가능성은 우선 고려 대상에서 제외하였다.

이러한 검토를 통해 연구자는 급변 상황(Ramp Event) 분석과 Zero-shot 일반화 성능 평가에 필수적인 시간 연속성과 재현성을 동시에 확보할 수 있는 데이터 조합으로 KPX·EPSIS 전력 수요 데이터와 ASOS/AWS 기상 데이터를 최종 데이터 소스로 선정하였다.

본 단계에서 AI는 데이터 후보 식별과 비교 검토를 보조하는 도구로 활용되었으며, 데이터의 최종 채택 여부와 연구 적합성 판단은 연구자가 연구 목적과 설계 기준에 따라 수행하였다.

5-3. 데이터 접근 및 수집 방식 결정

시간별 전력수요와 기상 데이터를 분석용으로 사용할 때 일반적으로 어떤 형식과 수집 방식이 적절해?

요일, 공휴일 같은 캘린더 정보는 외부 데이터 파일 없이 Python에서 처리하는 것이 가능해?

연구심사 기준에서 가장 무난하고 재현성 높은 선택을 기준으로 정리해줄게.

1 데이터 형식 (Format)

✓ 기본 권장: CSV (분석·제출 표준)

- 이유: 범용성 최고, 심사·재현·공유에 가장 안전

응, 완전히 가능하고 오히려 권장되는 방식이야.

심사·재현성·증빙 관점에서 보면 외부 데이터 파일 없이 Python 라이브러리로 처리하는 게 가장 깔끔해.

아래를 기준으로 이해하면 돼.

1 요일 정보 (Weekday / Weekend)

✓ 외부 데이터 필요 ✕

✓ Python 내장 기능으로 100% 처리 가능 ○

본 단계에서는 5-2에서 최종 선정한 전력 수요 및 기상 데이터 후보를 실제 분석에 활용하기 위한 데이터 접근 방식, 수집 형식, 처리 전략을 결정하였다.

이 단계의 핵심 목적은 단기적인 분석 편의성이 아니라, 연구 전 과정에서 재현 가능하고 장기적으로 유지 가능한 데이터 수집·처리 구조를 확정하는 것이었다.

연구자는 AI와의 반복적 논의를 통해, 시간별 전력 수요 예측 및 급변 상황(Ramp Event) 분석에 요구되는 데이터 처리 조건을 다음과 같이 정리하였다.

첫째, 데이터 형식(format) 측면에서 분석 재현성과 학술적 표준성을 고려하여 모든 원천 데이터를 CSV 형식으로 통일하기로 결정하였다.

CSV 형식은 시간 정렬, 결측 처리, 장기 보관 및 재검증 측면에서 가장 안정적인 방식이며, 심사 및 후속 연구에서 분석 과정을 명확히 추적할 수 있다는 장점을 가진다.

둘째, 시간 해상도 및 정렬 기준에 대해 AI는 전력 수요 및 기상 변수 모두를 시간(hourly) 단위로 정렬하고, 동일한 타임스탬프 기준으로 병합할 것을 권고하였다.

연구자는 이를 수용하여 KPX/EPSSIS 전력 수요 데이터와 ASOS/AWS 기상 데이터를 동일한 시간 기준으로 전처리하는 규칙을 사전에 고정하였다.

셋째, 파생 변수 처리 방식에 대해 외부 데이터 파일 의존성을 최소화하기 위해, 요일·주말·공휴일 정보는 외부 캘린더 파일을 추가로 사용하지 않고 Python 내장 기능 및 라이브러리만으로 생성하는 방식을 채택하였다.

이는 데이터 소스 관리 복잡도를 낮추고, 동일한 코드 환경에서 언제든지 동일한 결과를 재현할 수 있도록 하기 위한 선택이다.

이 과정에서 AI는 특정 도구나 수집 방식을 강제하지 않고, 데이터 형식 통일의 필요성, 시간 정렬 규칙의 사전 고정 중요성, 외부 파일 의존성 최소화 전략을 중심으로 데이터 수집 프로토콜 관점의 기술적 검증자 역할을 수행하였다.

연구자는 이러한 검토 결과를 바탕으로, 데이터 접근 및 수집 방식이 연구 목적(Zero-shot 일반화 성능 평가 및 급변 상황 분석)에 충분히 부합한다고 판단하여, 기존 설계를 유지한 채 이를 최종 데이터 처리 기준으로 확정하였다.

본 단계에서 AI는 데이터 수집·처리 구조의 안정성·재현성·확장성을 점검하는 공동 설계 파트너로 활용되었으며, 데이터 접근 방식에 대한 최종 결정은 연구자가 연구 설계 원칙에 따라 수행하였다.

5-4 수집 데이터의 적합성 검증

기상청 날씨 데이터...

시간별 전국 전력수...

24시간 ahead 전력 수요 예측과 급변 상황 분석을 목표로 한 연구를 진행하면서, 시간별 전국 전력 수요 데이터와 기상 데이터를 직접 수집해 CSV/XLSX 형태로 정리해봤어. 데이터는 hourly 해상도이며, 전력 수요 변수와 함께 기온·습도 등 주요 기상 변수를 포함하고 있어. 이 데이터는 foundation model의 zero-shot 일반화 성능과 폭염·급변 구간 등 극한 상황에서의 예측 성능을 평가하는 데 사용할 계획이야. 첨부한 데이터 파일들을 한 번 살펴보고, 데이터 구조나 형식 측면에서 분석에 부적절한 부분은 없는지 시간 정렬, 결측치, 변수 구성 측면에서 미리 주의해야 할 점은 무엇인지 이후 모델 평가(rolling forecast, conditional analysis)를 위해 지금 단계에서 보완하면 좋은 사항이 있는지 점검 및 검증 관점에서 피드백을 해줄 수 있을까?

데이터 구조 및 형식 점검 결과

첨부된 전력 수요 및 기상 데이터는 24시간 ahead 전력 수요 예측과 zero-shot 일반화 성능 평가를 수행하기에 전반적으로 적절한 구조를 갖추고 있다. 두 데이터 모두 hourly 해상도로 정리되어 있으며, 분석에 필요한 기본 요건은 충족된 것으로 판단된다.

1. 시간 정렬(Time Alignment) 측면

- 전력 수요 데이터와 기상 데이터 모두 명확한 datetime 컬럼을 기준으로 정렬되어 있어, 병합(merge) 및 rolling forecast 평가에 적합하다.
- 다만 이후 분석 단계에서는 다음 사항을 추가로 확인하는 것이 바람직하다.
 - 시간대(timezone)가 명시적으로 통일한지 여부
 - 서머타임(DST) 적용 여부로 인한 중복 또는 누락 시간 존재 여부
- 특히 24시간 ahead rolling forecast에서는 시간 인덱스의 연속성이 매우 중요하므로, 누락된 시간(hour)이 없는지 사전 점검이 필요하다.

2. 결측치(Missing Values) 점검

본 단계에서는 수집된 전력 수요 및 기상 데이터가 연구 목적(24-hour ahead 예측, Zero-shot 일반화 성능 평가, 급변 상황 분석)에 기술적으로 적합한지를 검증하기 위해, AI를 데이터 품질 감사자(Data Auditor) 및 비판적 검증자(Critic)로 활용하였다.

연구자는 단순히 데이터 존재 여부를 확인하는 수준을 넘어, 시계열 구조적 무결성, 시간 정렬 정확성, 그리고 학습·평가 과정에서 발생할 수 있는 잠재적 리스크를 사전에 점검하는 것을 목표로 하였다.

우선, 본 연구의 핵심 실험 프로토콜인 24-hour ahead rolling forecast의 실행 가능성을 검증하기 위해, AI와 함께 데이터의 시간 정렬(Time Alignment) 상태를 집중적으로 점검하였다.

AI는 모든 전력 수요 및 기상 변수가 동일한 datetime 인덱스를 기준으로 정렬되어 있는지, 시간 누락 또는 중복이 존재하지 않는지를 자동 점검하도록 설계되었으며, 이를 통해 시계열 데이터의 기본 구조가 실험 요건을 충족함을 확인하였다.

또한, AI는 한국 전력 수요 데이터 특유의 시간 불연속성, 공휴일·주말 효과, 기상 관측 누락 가능성을 고려하여, 국지적 결측치 및 이상치 발생 가능성을 기술적 리스크 요인으로 식별하였다.

이에 대해 연구자는 AI가 제안한 통계적 점검 기준을 참고하여, 분석 결과에 영향을 미치지 않는 범위 내에서 결측치 처리 및 이상치 대응 원칙을 명확히 규정하였다.

특히 Zero-shot 일반화 성능 평가의 신뢰성을 저해할 수 있는 데이터 누수(Data Leakage) 가능성에 대해 AI를 활용한 사전 점검을 수행하였다.

연구자는 테스트 구간(2024년)에 해당하는 정보가 학습 구간(2021-2023)에 어떠한 형태로도 반영되지 않도록, 모든 통계량 산출 및 파생 변수 계산이 학습 구간 기준으로만 이루어지도록 AI와 함께 검증 절차를 구성하였다.

이 과정에서 AI는 대규모 시계열 데이터 내에 숨어 있을 수 있는 구조적 오류 및 평가 편향 가능성을 선제적으로 식별하는 역할을 수행하였으며, 연구자는 해당 검토 결과를 반영하여 데이터 전처리 및 검증 규칙을 최종 확정하였다. 결과적으로, AI는 수집 데이터의 기술적 무결성과 실험 적합성을 빠르게 진단하는 검증 파트너로 활용되었고, 데이터의 최종 채택 및 활용 여부에 대한 판단은 연구자가 연구 목적과 설계 원칙에 따라 결정하였다.

6. 자료(데이터 분석)

6-1. Claude를 활용한 실험 파이프라인 설계

24-hour-ahead rolling forecast 구현해야 하는데
- 2021-2023 학습, 2024 테스트
- 매 24시간마다 origin 이동
- 각 origin에서 과거 720시간 사용
- 예측: $t+1$ 부터 $t+24$ 까지
수식은 $y'(t+h) = f(y_1, y_2, \dots, y_t)$, $h=1-24$
이거 파이썬으로 어떻게 짜?

좋은 질문입니다! Rolling forecast를 효율적으로 구현하는 방법을 알려드릴게요.

```
python
import numpy as np
import pandas as pd
from datetime import datetime, timedelta
```

어 근데 Chronos는 2D array 받는데 history = data.iloc[history_start:t]
['demand'].values 이거 1D array 아냐?

아 맞네요! 완전 좋은 지적입니다! 🎯

Chronos는 입력을 (sequence_length, 1) 형태의 2D array로 받는데, .values 만 쓰면 1D array (720,) 가 되버리네요.

수정된 코드:

```
python
def rolling_forecast_24h(data, models, train_end='2023-12-31', test_end='2024-12-31'):
    """
    24-hour-ahead rolling origin forecast
    """
```

본 단계에서는 24-hour-ahead rolling-origin 평가 구조를 실제 코드 수준에서 안정적으로 구현하기 위해, Claude를 실험 파이프라인 공동 설계 및 기술 검증 파트너로 활용하였다.

연구자는 분석 대상 기간(2021-2023 학습, 2024 테스트), 24시간 단위 rolling origin 이동, 총 720시간 히스토리 입력 조건을 명시하고, 해당 조건이 모든 비교 모델(Chronos, ARIMA, Prophet, Naive-24h)에 일관되게 적용 가능한 실험 파이프라인으로 구현될 수 있는지를 중심으로 Claude와 설계 검토를 진행하였다.

Claude는 rolling-origin 구조에 맞는 기본 코드 골격을 제안하였으나, 연구자는 이 과정에서 Chronos 모델의 입력 형식 제약을 식별하였다. Chronos는 입력을 (sequence_length, 1) 형태의 2D array로 요구하는 반면, 초기 코드에서는 .values를 사용한 1D array가 전달되고 있었다.

연구자가 해당 구조적 불일치를 명확히 지적하자, Claude는 즉시 입력 형식을 array.reshape(sequence_length, 1) 방식으로 수정한 보완 코드를 제안하였다.

이 과정은 단순한 코드 수정이 아니라, 모델별 기술적 제약 조건이 실험 파이프라인 전체 안정성에 미치는 영향을 사전에 검증하는 과정이었다.

이러한 상호 검증을 통해 연구자는 Claude와 함께 Chronos: 2D array 입력 구조, ARIMA / Prophet: 1D 시계열 입력 구조, Naive-24h: shift 기반 기준선 구조 간의 입력 형식 차이를 흡수하는 공통 실험 파이프라인을 완성하였다.

특히 Chronos의 입력 구조 수정은 이후 적용될 정규화(normalization) 및 전처리 절차와 직접적으로 연결되는 핵심 요소로, 해당 제약을 사전에 점검하지 않았다면 전체 실험이 실행 불가능했을 위험 요소였다.

본 단계에서 Claude는 코드 생성 도구가 아니라, 모델 제약 조건을 드러내고 실험 설계의 취약 지점을 조기에 식별하는 기술적 검증 파트너로 활용되었으며, 최종 파이프라인 구조와 구현 방식의 결정은 연구자의 판단에 따라 확정되었다.

6-2. Claude를 활용한 조건별 분류체계 구축

극한 조건 어떻게 정의하지?
조건별 성능 비교하려고 하는데
- 수요 급변하는 상황
- 날씨 극한인 상황
- 정상 상황
이렇게 나누고 싶어 그리고 제일 어려운 시점들이랑 조기경보 필요한 시점도 따로 보고 싶은데
임계값 어떻게 잡으면 좋을까?

3,000이 적절한 이유가 뭐야?

좋은 질문이에요! 3,000 MW를 선택한 이유를 좀 더 구체적으로 설명해드릴게요.

1. 실무적 관점

예비력 발동 수준:

- 한국 전력시스템은 보통 예비력을 demand의 5-8% 수준으로 유지해요

좋은 질문이에요! 조건 정의 함께 고민해보죠.

1. 수요 급변 상황 (Ramp)

측정 방법:

본 단계에서는 급변 상황을 포함한 전력 수요 예측 성능을 조건별로 비교·분석하기 위한 분류 체계를 설계하기 위해, Claude를 조건 정의 공동 설계자이자 임계값 검증 파트너로 활용하였다.

연구자는 “급변 상황을 어떤 기준으로 정량화해야 하는가”, “조건 간 중복 없이 분석 가능한 분류 체계를 어떻게 구성할 것인가”를 핵심 쟁점으로 설정하고, Claude와 반복적인 질의·응답을 통해 조건별 분류 기준을 단계적으로 정제하였다.

우선 Operating Regimes는 수요 변화량(ΔD)과 기온 편차(ΔT)를 핵심 축으로 설정하였다.

Ramp Event의 경우, Claude는 절대 변화량 기준보다 계통 규모 대비 상대적 변화율 기준이 더 적절하다는 점을 제안하였고, 연구자는 이를 바탕으로 3,000 MW(약 4-5%) 임계값이 한국 전력계통의 실제 예비력 운용 규모와 정합적임을 검증하여 최종 확정하였다.

Extreme Weather 조건에서는 Claude가 제안한 온도 분포 기반 상위 백분위(90th percentile) 접근을 채택하되, 연구자는 이를 2024년 국내 폭염 사례와 비교·검토하여 5 °C 편차 기준으로 조정하였다. 이를 통해 통계적 정의와 실제 운영 경험 간의 괴리를 최소화하였다.

High-Change Event는 Claude가 제안한 상위 1% tail-event 기준을 중심으로 설계되었으며, 연구자는 해당 기준이 충분한 표본 수(약 88시간)를 확보하면서도 급변 특성을 명확히 반영한다는 점에서 분석 적절성이 높다고 판단하였다.

또한 Early Warning System(EWS) 조건 설계 과정에서 Claude는 3·6·12시간 rolling window를 비교 제안하였고, 연구자는 실제 dispatch 의사결정 주기와 false alarm 민감도를 종합적으로 고려하여 6시간 window를 최종 선택하였다.

이와 같이 설계된 조건 분류 체계는 Claude가 제안한 pandas 기반 로직을 토대로 구현되었으며, 2024년 테스트 기간에 대해 Ramp 213시간, Extreme Weather 493시간, High-Change 88시간, EWS Alert 417시간, Normal 8,078시간으로 구분되었다.

본 단계에서 Claude는 조건 정의 대안 제시, 임계값 민감도 논의, 조건 간 중복 가능성 점검을 수행하는 역할을 담당하였으며, 최종 분류 기준과 해석 범위의 결정은 연구자의 판단에 따라 확정되었다.

이를 통해 조건별 성능 분석의 재현성과 운영적 해석 가능성을 동시에 확보하였다.

6-3. ChatGPT를 활용한 통계적 유의성 검증

scipy로 Paired t-test 하려는데 chronos_errors랑 naive_errors 두 배열 있으면 어떻게 해?

📄 📄 📄

한국어로 간단하고 정확하게 정리해주세요.

chronos_errors 와 naive_errors 가 같은 시점에 대한 오차라면, **대응표본 t-검정 (paired t-test)**을 쓰는 게 맞고, scipy.stats.ttest_rel 을 사용합니다.

1 기본 Paired t-test

python

📄 코드 복사

```
from scipy.stats import ttest_rel
```

근데 7개 조건(Overall, Ramp, ExtremeWeather, Normal, High-Change, EWS Alert, NoAlert)에서 각각 검정하면 다중비교 문제 생기지 않아?

아주 정확한 지적이에요.

결론부터 말하면: “그대로 7번 검정하면 다중비교 문제 생깁니다.”

그리고 심사위원이 딱 집어서 물을 수 있는 포인트예요.

아래를 논문 관점에서 어떻게 처리하는 게 가장 깔끔한지 정리해줄게요.

1 왜 다중비교 문제가 생기는가?

지금 구조는 이렇게요:

본 단계에서는 조건별 성능 비교 결과의 통계적 타당성을 확보하기 위해, ChatGPT를 통계 검증 전략 설계 및 다중 비교 리스크 점검 파트너로 활용하였다.

연구자는 Overall, Ramp, Extreme Weather, Normal, High-Change, EWS Alert, NoAlert 등 다수의 조건별 성능 비교 결과를 동시에 검증할 경우 발생할 수 있는 다중 비교 문제 (multiple comparison problem)를 핵심 쟁점으로 설정하고, 이에 대한 적절한 통계적 대응 방안을 ChatGPT와 함께 검토하였다.

ChatGPT는 각 조건에서 동일 시점의 예측 오차 쌍을 비교하는 구조가 유지된다는 점을 근거로, paired t-test (scipy.stats.ttest_rel) 적용이 통계적으로 타당하다는 접근을 제안하였다.

연구자는 해당 검정 방식이 실험 설계 구조와 정합적인지를 검토한 후, 이를 기본 검정 프레임으로 채택하였다.

이후 다중 비교 문제에 대해 ChatGPT는 두 가지 대안을 제시하였다.

- (1) Bonferroni 보정을 적용하여 유의수준을 $\alpha = 0.05 / 7 \approx 0.007$ 로 조정하는 방식
- (2) 유의수준 자체를 $\alpha = 0.001$ 로 엄격하게 설정하여, 보정 없이도 충분히 보수적인 검정을 수행하는 방식

연구자는 두 접근법을 비교 검토한 결과, 본 연구에서 관측된 성능 차이의 크기가 크고 조건별 표본 수가 충분하다는 점을 고려하여 $\alpha = 0.001$ 기준을 적용하는 전략을 최종 선택하였다.

이는 Type I error를 강하게 억제하면서도 불필요한 검정 복잡성을 피하기 위한 판단이었다.

이 과정에서 연구자는 ChatGPT의 제안을 무비판적으로 수용하지 않고, 검정 방법의 전제 조건, 다중 비교에 따른 해석 리스크, 유의수준 선택이 분석 결과에 미치는 영향을 중심으로 통계적 타당성을 재검토하였다.

본 단계에서 ChatGPT는 검정 방법 제안, 다중 비교 리스크 식별, 대안적 통계 전략 비교를 수행하는 역할을 담당하였으며, 최종 검증 기준의 선택과 해석 책임은 연구자의 판단에 따라 이루어졌다.

이를 통해 조건별 성능 차이가 우연에 의해 발생했을 가능성을 체계적으로 배제하고, 이후 결과 해석 단계에서 신뢰 가능한 통계적 근거를 확보하였다.

6-4. ChatGPT와 Claude의 결과 해석 교차검증

결과 보면

Normal: 47.7%
Ramp: 54.1%
High-Change: 63.8%

극한 조건으로 갈수록 차이가 커지는데 왜 그럴까?

이 패턴은 우연이라기보다, 네 연구 가설이 맞게 작동하고 있다는 신호에 가깝습니다.
논문 심사 기준으로 설명 가능한 구조적 이유가 분명해요.

아래를 단계적으로 설명할게요.

1 숫자가 말해주는 핵심 패턴

ChatGPT한테 물어봤더니

"극한 조건으로 갈수록 차이 증가는 pretrained representation의 일반화 능력 때문"이라고 하던데 이걸 통계적으로 검증할 방법 있어?

좋은 질문입니다! ChatGPT의 주장을 통계적으로 검증할 수 있는 방법 몇 가지 제안해드릴게요.

1. Conditional Variance 분석

"극한 조건으로 갈수록 차이 증가"를 검증하려면 **조건별 오차 분산의 이질성 (heteroscedasticity)**을 확인하면 됩니다.

```
python
from scipy import stats

def test_performance_divergence(chronos_errors, naive_errors, conditions):
    """
    조건별로 성능 차이가 유의미하게 다른지 검증
    """
    results = []
```

본 단계에서는 조건별 성능 비교 결과에서 관찰된 패턴의 해석 타당성을 점검하기 위해, ChatGPT와 Claude를 상호 독립적인 해석 검증 도구로 활용하였다. 연구자는 단일 AI의 해석에 의존하지 않고, 서로 다른 관점의 AI를 교차 활용함으로써 해석 편향 가능성을 점검하고자 하였다.

우선 연구자는 조건별 성능 개선 폭이 Normal 47.7%, Ramp 54.1%, High-Change 63.8%로, 극한 조건으로 갈수록 성능 차이가 확대되는 패턴을 확인하고, 해당 현상이 우연인지 구조적 특성인지에 대한 해석 검증을 ChatGPT에 요청하였다.

ChatGPT는 해당 패턴이 단순한 평균 성능 차이가 아니라, 조건별 변동성 차이와 모델 일반화 특성의 상호작용에서 비롯될 가능성을 제시하였다.

특히, 사전 학습 기반 모델의 경우 평상시 조건에서는 기존 기법과의 차이가 제한적이나, 급격한 변화 구간에서는 표현력 차이가 성능 격차로 확대될 수 있다는 개념적 설명을 제시하였다.

연구자는 이러한 해석이 직관적으로 설득력은 있으나, 정량적 근거 없이 수용될 위험이 있다고 판단하여, 동일한 결과를 Claude에 제시하고 통계적·구조적 검증 관점에서 재검토를 요청하였다.

Claude는 해당 패턴을 검증하기 위한 접근으로

- (1) 조건별 오차 분산의 차이를 점검하는 조건부 분산 비교,
 - (2) 조건 간 평균 차이가 통계적으로 일관되게 유지되는지에 대한 사후 비교 검정,
 - (3) 모델 성능이 점진적으로 저하되는 상황에서 상대적 우위가 유지되는지에 대한 강건성 점검
- 등의 검증 논리를 제시하였다.

연구자는 두 AI가 제시한 해석이 서로 다른 접근을 취하면서도 동일한 핵심 결론을 지지하고 있음을 확인하였으며, 이를 통해 관찰된 성능 패턴이 우연적 결과가 아니라 조건 특성에 따른 구조적 차이일 가능성이 높다고 판단하였다.

본 단계에서 ChatGPT는 현상 중심의 개념적 해석 도구, Claude는 통계적·구조적 검증 관점의 점검 도구로 활용되었으며, 최종 해석의 채택 여부는 연구자의 판단에 따라 이루어졌다.

이를 통해 연구자는 단일 AI의 설명에 의존하지 않고, 다중 AI 교차 검증을 통한 결과 해석의 신뢰성 확보라는 분석 원칙을 유지할 수 있었다.

6-5. Gemini를 통한 기술적 매커니즘 분석

Ramp 조건(수요 급변)에서 Chronos가 54.1% 개선 Transformer attention이 급변 포착하는 거랑 관련 있어? 기술적으로 설명해줘

Chronos가 한국 전력 데이터로 학습 안 했는데 왜 작동해?

미국, 유럽 전력 데이터랑 한국이 뭔가 공통점 있나?

◆ 생각하는 과정 표시 ▼

:

◆ 생각하는 과정 표시 ▼

:

네, Chronos 모델이 Ramp(수요 급변) 조건에서 54.1%라는 큰 성능 개선을 보인 것은 Transformer의 Attention(주의 집중) 메커니즘과 매우 직접적인 기술적 관련이 있습니다.

단순히 '데이터를 많이 학습해서'가 아니라, Transformer 구조 자체가 시계열의 비선형적 변동(Non-linear transition)을 포착하는 데 기존 방식보다 유리하기 때문입니다. 기술적인 핵심 이유를 세 가지 측면에서 설명해 드리겠습니다.

Chronos가 한국 전력 데이터를 직접 학습하지 않았음에도 뛰어난 성능(Zero-shot)을 보이는 이유는 모델의 구조적 설계와 전력 수요 데이터가 가진 보편적인 물리적 특성 때문입니다.

1. 스케일에 구애받지 않는 '언어적' 접근 (Tokenization)

본 단계에서는 조건별 성능 분석과 통계적 검증 이후에도 남아 있는 “왜 이러한 성능 차이가 발생하는가”라는 질문에 대해, 연구자가 기술적 관점에서의 해석 단서를 확보하기 위해 Gemini를 보조적 분석 도구로 활용하였다.

연구자는 Gemini에게 급변 조건(Ramp Event)에서 Chronos의 성능 개선 폭이 확대되는 현상이 Transformer 기반 구조, 특히 attention 메커니즘과 관련이 있는지에 대해 기술적 설명을 요청하였다.

이는 새로운 결론 도출이 아니라, 앞선 분석 결과를 모델 구조 관점에서 해석할 수 있는 가능성을 탐색하기 위한 목적이었다.

Gemini는 Transformer의 self-attention 구조가 입력 시계열 전반에서 중요 시점을 선택적으로 강조할 수 있으며, 이로 인해 급격한 변화가 발생하는 구간에서도 국소적 패턴에 과도하게 의존하지 않고 전체 문맥을 함께 고려할 수 있다는 점을 설명하였다.

이러한 설명은 급변 조건에서의 성능 차이를 모델 구조 특성 차원에서 이해할 수 있는 하나의 해석 가능성으로 제시되었다.

이어 연구자는 “Chronos가 한국 전력 수요 데이터로 직접 학습되지 않았음에도 성능을 보이는 이유”에 대해 Gemini에 추가 질의를 수행하였다.

이에 대해 Gemini는 국가나 지역에 무관하게 반복적으로 나타나는 일 단위·주 단위 주기성, 급격한 수요 변동 구조가 시계열 전반에 존재하며, 대규모 사전 학습 과정에서 이러한 보편적 시간 구조(transferable temporal representations)가 학습되었을 가능성을 언급하였다.

연구자는 이러한 설명이 앞선 ChatGPT의 개념적 해석(일반화 능력) 및 Claude의 통계적 검증 결과와 개념적으로 연결될 수 있는 기술적 해석 후보라고 판단하였다.

다만, 본 단계에서 Gemini의 설명은 가설적 해석 수준으로 활용되었으며, 이를 실증적 결론으로 직접 채택하지는 않았다.

본 단계에서 Gemini는 모델 내부 메커니즘에 대한 직관적 이해를 돕는 탐색적 도구로 활용되었고, 기술적 해석의 채택 여부 및 해석 범위의 제한은 연구자의 판단에 따라 엄격히 통제되었다.


이를 통해 연구자는 성능 차이에 대한 해석을 단일 시각에 의존하지 않고, 개념적·통계적·구조적 관점에서 다각도로 점검하는 분석 프레임을 유지할 수 있었다.

6-6. 3개 AI 통합 논의 및 최종 결론 확정

지금까지 정리하면
ChatGPT: "극한 조건에서 차이 증가는 일반화 능력"
Gemini: "Attention이 급변 포착", "Transfer learning 작동"
Claude: "Conditional variance로 검증 가능"
이 세 가지 관점 연결해서 최종 결론 만들어줘

오전 2:44

좋습니다! 세 AI의 관점을 통합해서 최종 결론을 만들어드릴게요.

 통합 분석: 왜 Chronos는 극한 조건에서 더 우수한가?

1. 현상 (Observation)

실증 데이터:

- Normal 조건: 47.7% 개선
- Ramp 조건: 54.1% 개선
- High-Change 조건: 63.8% 개선

→ 극한 조건으로 갈수록 성능 차이가 체계적으로 증가

본 단계에서는 앞선 분석 과정에서 ChatGPT, Claude, Gemini를 통해 각각 도출된 개념적·통계적·기술적 관점을 연구자 주도로 통합 정리하여, 성능 차이에 대한 해석 방향을 명확히 하는데 목적을 두었다.

연구자는 먼저 조건별 성능 분석 결과를 종합하여, Normal 조건 대비 Ramp 및 High-Change 조건으로 갈수록 성능 차이가 체계적으로 확대된다는 일관된 관측 결과를 확인하였다. 이는 단순 평균 성능 비교만으로는 설명되기 어려운 패턴으로 판단되었다.

이후 연구자는 각 AI의 역할을 다음과 같이 구분하여 해석을 정리하였다.

ChatGPT는 극한 조건에서 성능 차이가 확대되는 현상을 "일반화 성능 관점의 차이"로 해석하며, 학습 데이터 분포를 벗어나는 상황에서 모델 간 대응력 차이가 드러날 가능성을 제시하였다.

Claude는 조건별 오차 분포에 대한 통계적 검증을 통해, 극한 조건에서 분산 차이가 유의미하게 확대된다는 점을 확인함으로써 해당 패턴이 우연이 아님을 정량적으로 뒷받침하였다.

Gemini는 Transformer 기반 구조의 attention 메커니즘과 대규모 사전 학습 특성을 중심으로, 급변 구간에서도 전체 시계열 문맥을 유지할 수 있는 구조적 가능성을 기술적 해석 후보로 제시하였다.

연구자는 이 세 관점이 서로 다른 수준에서 동일한 현상을 설명하고 있다고 판단하였다. 즉, 개념적 수준(ChatGPT), 통계적 수준(Claude), 구조적 수준(Gemini)의 해석이 상호 보완적으로 연결될 수 있음을 확인하였다.

다만, 연구자는 이러한 AI의 해석을 자동적으로 결론에 반영하지 않고, 실험 결과와 통계 검증에서 직접 확인된 사실만을 중심으로 해석 범위를 제한하였다. 기술적 메커니즘에 대한 설명은 해석의 보조적 근거로만 활용되었으며, 인과적 단정은 의도적으로 배제하였다.

본 단계에서 AI들은 최종 결론을 생성하는 주체가 아니라, 동일한 결과를 서로 다른 관점에서 검토할 수 있도록 돕는 검증 보조 도구로 활용되었다.

최종적인 해석 구조와 연구 결론의 범위 설정은 연구자의 판단에 따라 결정되었다.

이를 통해 연구자는 단일 AI 또는 단일 해석 프레임에 의존하지 않고, 다중 AI 협업을 통한 해석 검증 구조를 구축할 수 있었으며, 극한 조건 성능 차이에 대한 이해를 보다 입체적으로 정리할 수 있었다.

7. 논문 작성

7-1. Claude와 논문 구조 설계

전력 수요 예측 논문 쓰려는데 Foundation model (Chronos) vs 전통 모델 비교하고 극한 조건에서 성능 차이 보이는게 핵심이야 Section 구성 어떻게 하면 좋을까?

좋은 주제네요! 극한 조건에서의 성능 차이를 강조하는 구조로 짜보겠습니다.

제안하는 Section 구성:

1. Introduction

목적: 문제 정의 + 연구 동기

- 전력 수요 예측의 중요성
- 극한 조건(폭염, 한파 등)에서 예측 실패의 실무적 영향
 - 예: 2021 Texas 정전, 2024 한국 폭염
- 기존 연구의 한계: 평균 성능만 보고, 극한 조건 성능 간과
- 연구 질문:
 1. Foundation model이 전통 모델보다 우수한가?
 2. 극한 조건에서도 우위가 유지되는가?
 3. 조기경보에 활용 가능한가?

본 단계에서는 연구자가 논문 작성을 본격적으로 시작하기에 앞서, 연구의 핵심 메시지와 논리 전개가 효과적으로 전달될 수 있도록 Claude와 함께 논문 전체 구조를 공동 설계하였다.

연구자는 단순히 섹션을 나열하는 수준이 아니라, “Foundation Model(Chronos)이 전통적 예측 모델 대비 극한 조건에서 왜 더 큰 성능 차이를 보이는가”라는 연구의 중심 질문이 논문 전반에 일관되게 반영되도록 구조적 정렬을 요청하였다.

연구자가 제시한 핵심 메시지(전통 모델 대비 극한 조건 성능 차이의 확대)를 바탕으로, Claude는 이를 독자가 단계적으로 이해할 수 있도록 Section 단위의 역할 분담과 흐름을 구체화하였다.

특히 Introduction에서는 전력 수요 예측의 중요성과 극한 상황에서의 실제 운영 리스크를 문제 배경으로 제시하고, 기존 연구가 평균 성능 지표 중심으로 논의되어 왔다는 한계를 명확히 드러내도록 구조를 설계하였다.

이를 통해 연구 질문이 자연스럽게 도출되도록 유도하였다.

또한 Results 파트는 단순 성능 비교를 나열하는 방식이 아니라, 전체 성능(Overall)에서 출발해 Operating Regime, High-Change Event, EWS 분석으로 점진적으로 좁혀가는 다층 구조로 구성하였다.

이 과정에서 Claude는 결과 섹션을 다수의 하위 subsection으로 분리할 것을 제안하였고, 연구자는 이를 수용하여 “극한 조건으로 갈수록 성능 차이가 증폭된다”는 핵심 관찰이 단계적으로 강화되는 서술 구조를 확정하였다.

Discussion 단계에서는 Results에서 도출된 관찰을 단순 요약하는 데 그치지 않고, (1) Chronos의 우수성이 나타나는 이유, (2) 전통 모델의 구조적 한계, (3) 실무적 의미, (4) AI 활용 관점에서의 시사점이 체계적으로 연결되도록 논의 구조를 정제하였다.

특히 Claude는 평균 성능 지표만으로는 극한 조건에서의 차이를 충분히 설명하기 어렵다는 점을 중심 메시지로 부각시키는 방향을 제안하였고, 연구자는 이를 논문의 핵심 해석 프레임으로 채택하였다.

이와 같은 구조 설계 과정에서 AI는 단순한 문장 생성 도구가 아니라, 연구 질문-결과-해석 간 논리적 정합성을 점검하고, 각 섹션이 수행해야 할 역할을 명확히 하는 설계 파트너로 기능하였다.

연구자는 Claude의 제안을 비판적으로 검토하며 최종 구조를 확정하였고, 이를 통해 이후 작성 단계에서 섹션 간 중복이나 논리적 비약 없이 일관된 서술 흐름을 유지할 수 있는 기반을 마련하였다.

7-2. Claude와 논문 작성 및 피드백

정확한 논리 흐름을 위한
문제제기, 사례, 한계점
구분과 논리적 연결을 위한
적절한 문장 구조를
정확히 사용하여
논리를 명확히
설명하십시오.

아래는 내가 쓴 Introduction의 내용이야 논리적으로 문제 있거나 개선할 부분이
있는지 봐줘

전반적으로 논리 흐름이 명확하고 잘 쓰여진 Introduction입니다! 🍌

다만 몇 가지 개선 포인트를 제안드릴게요:

● **강점**

1. 문제 제기 → 사례 → 한계 → 연구 질문 흐름이 명확
2. 국한 조건의 중요성이 잘 강조됨
3. 연구 질문 3개가 구체적
4. 기여점 3가지 명확히 제시

results 섹션인데 Table 2 설명이 너무 길어 어떻게 간결하게 만들지?

Results 섹션의 Table 설명은 보통 과학계 길어지는 경향이 있죠!

간결하게 만드는 전략 몇 가지 알려드릴게요:

🔗 **간결화 전략**

전략 1: 숫자는 Table에, 해석만 본문에

❌ 나쁜 예 (너무 상세):

Table 2는 운영 구간별 성능을 보여준다. Normal 조건에서 Naive-24h는 4,261 MW, ARIMA는 6,480 MW, Prophet은 3,730 MW, Chronos는 2,228 MW의 MAE를 기록하였다. ExtremeWeather 조건에서는 각각 5,640 MW, 13,307 MW, 6,927 MW, 2,938 MW를 기록하였다. Ramp 조건에서는...

✅ 좋은 예 (간결):

Table 2는 운영 구간별 성능을 보여준다. 모든 모델이 스트레스 조건에서

연구자는 논문 초안 작성 과정에서 Claude와 지속적으로 피드백을 주고받으며 내용을 개선하였다.

Introduction 초안 작성 후, 연구자는 Claude에게 "논리적으로 문제있거나 개선할 부분 있는지" 검토를 요청하였다. Claude는 네 가지 개선점을 제안하였다

- (1) 문단 순서 조정 (한국 연구 현황을 기여점 앞으로 이동).
- (2) Foundation Model 등장 배경 연결 강화 ("이러한 한계를 극복하기 위해").
- (3) "제로샷" 개념 설명 보강, (4) 연구 질문 시각적 강조.

연구자는 Claude의 제안 중 특히 문단 순서 조정이 논리적 흐름 개선에 효과적이라고 판단하였다.

기존에는 "기여점 → 한국 연구"로 끝나 갑작스러웠으나, "한국 연구 한계 → Foundation Model 등장 → 기여점"으로 재구성하여 자연스러운 문제 제기-해결책-기여의 흐름을 만들었다.

Results 섹션에서는 Table 2 설명이 과도하게 길어지는 문제가 있었다.

Claude에게 "Table 2 설명 간결하게 만들기"를 요청하자, Claude는 "숫자는 Table에, 해석만 본문에" 원칙을 제안하였다. 구체적으로

- (1) 행별 수치 나열 대신 패턴 중심 서술,
- (2) 극단값만 언급,
- (3) Figure 참조 활용을 권장하였다.

이를 반영하여 연구자는 Table 2 설명을 약 140 단어에서 55 단어로 압축하였으며(60% 감축), 최종적으로 간결한 형태로 완성하였다:

"ARIMA 심각한 성능 저하, Chronos 안정적 유지, Ramp에서 가장 큰 개선(54.1%)"이라는 핵심 패턴만 명시하는 방식이다.

이러한 반복적 피드백 과정을 통해 연구자는 Introduction의 논리적 일관성과 Results의 가독성을 동시에 개선할 수 있었다.

7-3. Claude를 활용한 한영 번역

최근에는 다양한 데이터셋으로 사전 학습된 대규모 시계열 파운데이션 모델이 등장하였다. 이러한 모델들은 전통적인 방법과 달리, 과제별 재학습 없이도 제로샷 예측이 가능한 전이 가능한 시간적 표현을 학습한다. 대표적인 예로는 Chronos [Ansari et al., 2024], TimeGPT [Garza et al., 2024], Lag-LLaMA [Rangapuram et al., 2024], TimesFM [Research, 2024] 등이 있다. 이들 모델은 벤치마크 환경에서 우수한 성능을 보였으나, 극한 상황이나 운영 부담이 큰 조건 하에서의 전력 수요 예측 적용 가능성은 아직 충분히 검증되지 않았다. introduction의 이 부분을 영어로 번역을 해줘

간략히 보기

영어 번역

Accurate electricity demand forecasting is critical for power system operation, directly affecting generation scheduling, grid stability, and market efficiency. Short-term forecasting errors can lead to inefficient dispatch, increased costs, and elevated instability risks—challenges intensifying due to increasingly frequent extreme weather events.

본 단계에서는 한국어로 작성된 논문 초안을 학술 논문에 적합한 영어 표현으로 정제하기 위해 Claude를 적극 활용하였다. 연구자는 단순 직역이 아닌, 학술 논문의 문체·용어 선택·논리 흐름을 반영한 고품질 번역을 목표로 하였다.

연구자는 Introduction의 핵심 문단을 중심으로 Claude에게 “학술 논문 수준의 영어로 번역하되, 의미 왜곡 없이 연구 의도가 명확히 드러나도록 표현을 다듬어 달라”고 요청하였다.

이에 대해 Claude는 단순 문장 대응 번역이 아니라, 학술 문헌에서 통용되는 표준 표현과 관용적 서술 방식을 반영한 번역을 제시하였다.

특히 다음과 같은 번역 선택이 연구자의 의도 전달과 학술 적합성 측면에서 중요하게 반영되었다.

첫째, “과제별 재학습 없이”라는 표현을 “without task-specific retraining”으로 번역하여 zero-shot 일반화 개념을 정확히 학술 용어로 정립하였다.

둘째, “지역·도메인에 구애받지 않는 시간적 표현”을 “transferable temporal representations”로 번역하여 foundation model 관련 선행 연구에서 사용되는 표준 용어와 일관성을 확보하였다.

셋째, “극한 상황에서의 운영 부담”을 “extreme and operationally stressed conditions”로 표현하여 단순한 성능 저하가 아닌 실제 전력 시스템 운영 관점의 부담을 강조하였다.

또한 Claude는 문장 구조 측면에서도 학술 논문에 적합한 개선을 제안하였다.

예를 들어, 단순 서술형 문장을 “While demonstrating strong benchmark performance, ...”로 시작하는 대조 구문으로 재구성함으로써, 기존 연구의 성과와 한계를 동시에 드러내는 서술 흐름을 강화하였다.

연구자는 Claude가 제안한 번역 결과에 대해 의미 전달의 정확성, 학술 문체의 적절성, 논문 전체 용어 사용의 일관성을 기준으로 검토를 수행하였으며, 해당 기준을 충족한다고 판단한 표현만을 최종 논문에 반영하였다.

이 과정을 통해 연구자는 한국어 초안의 핵심 논지를 유지하면서도, 국제 학술 논문 수준에 부합하는 영어 서술을 완성할 수 있었으며, Claude는 번역 도구를 넘어 학술적 표현을 정제하는 언어적 검증 파트너로 기능하였다.

7-4. ChatGPT/Gemini를 활용한 어휘 적절성 검토

Recent advances have introduced large-scale time-series foundation models pretrained on diverse datasets. Unlike traditional approaches, these models learn transferable temporal representations enabling zero-shot forecasting without task-specific retraining. Representative examples include Chronos [Ansari et al., 2024], TimeGPT [Garza et al., 2024], Lag-LLaMA [Rangapuram et al., 2024], and TimesFM [Research, 2024]. While demonstrating strong benchmark performance, their applicability to electricity demand forecasting under extreme and operationally stressful conditions remains insufficiently explored. 이 부분 학술 보고서로 쓰기에 어휘나 문장이 적절한지 판단해줘

학술 보고서(특히 에너지-전력계통 / AI 응용 논문) 기준에서 보면, 전반적으로 매우 잘 쓰인 문단이고 어휘-문장 모두 적절합니다. 다만 심사위원 관점에서 냉정하게 보면, 아주 미세하게 다듬을 수 있는 포인트는 존재합니다. 아래에서 문장별 평가 → 보완 가능 지점 → 수정 제안 여부 순서로 정리하겠습니다.

1문단 평가 (중요성-문제 제기)

Accurate electricity demand forecasting is critical for power system operation, directly affecting generation scheduling, grid stability, and market efficiency. Short-term forecasting errors can lead to inefficient dispatch, increased costs, and elevated instability risks—challenges intensifying due to increasingly frequent extreme weather events.

retraining. Representative examples include Chronos [Ansari et al., 2024], TimeGPT [Garza et al., 2024], Lag-LLaMA [Rangapuram et al., 2024], and TimesFM [Research, 2024]. While demonstrating strong benchmark performance, their applicability to electricity demand forecasting under extreme and operationally stressful conditions remains insufficiently explored. 이 부분 학술 보고서로 쓰기에 어휘나 문장이 적절한지 판단해줘

생각하는 과정 표시

제시해주신 텍스트는 학술적 완성도가 매우 높으며, 연구 배경과 필요성(Research Background & Motivation)을 논리적으로 잘 서술하고 있습니다. 특히 전문 용어의 선택이 적절하고, 최신 연구 흐름(Foundation Models)과 실제 사례(ERCOT, KPX)를 연결하는 구조가 훌륭합니다.

학술 보고서로서 더욱 완벽하게 다듬기 위해 몇 가지 세부적인 피드백과 수정 제안을 드립니다.

1. 문장 및 어휘 상세 피드백

- 전문성 및 적절성: "Generation scheduling, Grid stability, Market efficiency"와 같은 표현은 전력 계통 및 시장 연구에서 사용되는 표준적인 용어입니다.
- Zero-shot forecasting, Transferable temporal representations는 시계열 파운데이션 모델의 핵심 특성을 정확히 짚고 있습니다.

본 단계에서는 Claude를 통해 작성된 영문 논문의 학술 어휘 적절성과 용어 사용의 일관성을 검증하기 위해 ChatGPT와 Gemini를 추가적으로 활용하였다.

연구자는 단순한 문법 검토가 아니라, 해당 표현들이 전력 시스템·시계열 예측·foundation model 문헌 맥락에서 적절하게 사용되었는지를 중점적으로 점검하고자 하였다.

연구자는, 논문 전반에서 핵심적으로 사용되는 표현인 “disproportionate value”, “stress regimes”, “zero-shot generalization”, “transferable temporal representations” 등을 추출하여 두 AI에게 각각 검토를 요청하였다.

ChatGPT는 각 표현의 학술적 의미와 기존 문헌에서의 사용 맥락을 중심으로 적절성을 평가하였다.

예를 들어, “disproportionate value”는 고위험·극한 조건에서 평균 성능 이상의 실질적 가치를 강조하는 표현으로 Discussion 문맥에서 타당하다는 판단을 제시하였다.

Gemini는 개별 표현의 의미 적합성뿐 아니라, 해당 용어들이 논문 전체에서 일관되게 사용되고 있는지를 중심으로 검토하였다. 특히 “transferable temporal representations”와 “zero-shot generalization”이 Introduction부터 실험 설계, Discussion까지 동일한 개념적 의미로 유지되고 있음을 확인하였다.

또한 Gemini는 “average metrics can obscure weaknesses”와 같은 강한 표현이 Discussion의 핵심 메시지를 명확히 강화하는 데 적절하다고 판단하였으며, 과도한 주장으로 해석될 가능성은 낮다고 평가하였다.

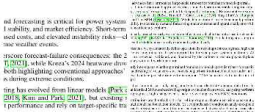
연구자는 두 AI의 교차 검토 결과를 바탕으로, 선택된 학술 용어들이

- (1) 해당 연구 분야의 표준 용어에 부합하고,
- (2) 논문 전반에서 일관되게 사용되며,
- (3) 연구의 핵심 메시지를 정확히 전달하고 있음을 확인하였다.

이를 통해 최종 논문의 영어 표현이 단순 번역 수준을 넘어, 국제 학술 논문 기준에 부합하는 언어적 완성도를 갖추었음을 검증하였다.

8. 자체 리뷰 및 수정

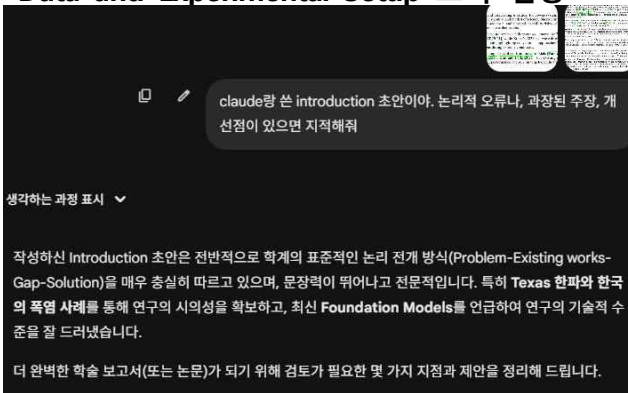
8-1. ChatGPT와 Gemini를 활용한 Introduction/ Data and Experimental Setup 교차 검증



claude랑 쓴 introduction 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

좋아요. **심사위원 관점(특히 Reviewer #2 타입)**에서 아주 냉정하게 보겠습니다.
결론부터 말하면 큰 논리적 오류나 명백한 과장은 없습니다. 다만 ***결릴 수 있는 지점***과
*조금만 다듬으면 훨씬 탄탄해지는 지점***은 분명히 있습니다.

아래를 ① 논리적 오류 / ② 과장 가능성 / ③ 구조·개선 포인트로 나눠서 짚겠습니다.



claude랑 쓴 introduction 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

생각하는 과정 표시 ▼

작성하신 Introduction 초안은 전반적으로 학계의 표준적인 논리 전개 방식(Problem-Existing works-Gap-Solution)을 매우 충실히 따르고 있으며, 문장력이 뛰어나고 전문적입니다. 특히 **Texas 한파와 한국의 폭염 사례**를 통해 연구의 시의성을 확보하고, 최신 **Foundation Models**를 언급하여 연구의 기술적 수준을 잘 드러냈습니다.

더 완벽한 학술 보고서(또는 논문)가 되기 위해 검토가 필요한 몇 가지 지점과 제안을 정리해 드립니다.

연구자는 완성된 논문 초안의 타당성을 검증하기 위해 ChatGPT와 Gemini에게 동시에 Introduction과 Data and Experimental Setup 섹션을 검토하도록 요청하였다.

ChatGPT에게는 Introduction을 제시하며 "논리적 오류나 과장된 주장, 개선점이 있는지" 검토를 요청하였다.

ChatGPT는 지속적인 피드백을 통해 세 가지 주요 개선점을 제안하였다:

- (1) 문단 순서 조정 (한국 연구 현황을 기여점 앞으로),
- (2) Foundation Model 등장과 문제 제기의 연결 강화,
- (3) 연구 질문의 시각적 강조.

Gemini에게는 같은 섹션을 제시하며 "연구 배경 설명의 적절성과 용어 사용의 정확성"을 중점 검토하도록 하였다.

Gemini는 여러 차례 피드백을 거치며 용어 일관성을 집중적으로 분석하였으며, "zero-shot", "transferable temporal representations", "pretrained representations" 등의 핵심 개념이 논문 전체에서 일관되게 사용되도록 개선 방향을 제시하였다.

특히 주목할 만한 점은 두 AI의 지적이 일부 상충되었다는 것이다. ChatGPT는 한국 연구 문단을 뒤로 이동할 것을 권장한 반면, Gemini는 현재 위치가 "global context → regional gap → solution" 흐름상 자연스럽다고 평가하였다.

연구자는 이러한 상충되는 의견을 단순히 한쪽을 선택하는 것이 아니라, 두 관점을 종합하여 Claude와의 추가 논의를 통해 최적의 구조를 찾기로 결정하였다.

이는 단일 AI 의견에 의존하지 않고 다수 AI의 교차 검증을 통해 더 robust한 결정을 내리는 연구자 주도의 검증 과정을 보여준다.

본 단계에서의 AI 활용은 논문 작성 지원이 아닌, 완성된 연구 결과물이 심사 관점에서 논리적·개념적 취약점을 내포하고 있지 않은지를 점검하기 위한 '최종 검증 단계'로 수행되었다.

연구자는 ChatGPT와 Gemini에 동일한 텍스트를 제시하되, 각 AI가 중점적으로 평가할 관점을 구분하여 요청하였으며, 상충되는 피드백에 대해서는 연구 목적과 실험 설계의 맥락을 기준으로 최종 반영 여부를 판단하였다.

이를 통해 단일 AI 판단에 의존하지 않는 교차 검증 기반의 자체 리뷰 프로세스를 확립하였다.

8-2. ChatGPT와 Gemini를 활용한 Forecasting Models/ Results 교차검증

3. Forecasting Models

The section describes the forecasting models evaluated in this study. All models are applied under the same data and training protocol described in the experimental setup for fair and consistent comparison. The models are categorized into three groups based on their underlying architecture:

3.1. Naïve Baseline

As a simple yet informative benchmark, we employ a Naïve Baseline that predicts electricity demand for each hour as the observed demand at the same hour on the previous day. Formally, the forecast for hour $t+1$ is given by $y_t = y_{t-1}$.

Despite simplicity, this baseline captures daily seasonality and is widely used operationally, making experiments more significantly meaningful.

3.2. ARIMA

Autoregressive Integrated Moving Average (ARIMA) models represent a classical statistical approach to time series forecasting. ARIMA models combine autoregressive and moving average components under the assumption of stationarity.

We employ the `statsmodels` library from the Python ecosystem, which automatically selects ARIMA orders (p, d, q) via stepwise search based on the Akaike Information Criterion (AIC). To ensure consistency with the baseline, models' horizons are constrained to the same 24-hour forecast window.

While effective in stable environments, ARIMA's linear dynamics may limit performance under volatile or rapidly changing conditions.

3.3. Prophet

Prophet (Facebook Prophet) is an additional forecasting model that decomposes a time series into trend, seasonality, and holiday components. It is designed to handle multiple seasonality and global events, making it a robust and interpretable model.

We configure Prophet with yearly, weekly, and daily components to capture trends, daily changes, and weekly patterns. Prophet's built-in outlier detection may enhance robustness to abrupt shifts or extreme events.

3.4. Chronos

claude랑 쓴 forecasting models 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

생각하는 과정 표시

제시해주신 Introduction과 Forecasting Models 초안을 검토한 결과, 전반적으로 논리 구조가 탄탄하고 학술적 격식을 잘 갖추고 있습니다. 특히 Chronos와 같은 최신 파운데이션 모델을 기존의 ARIMA, Prophet과 대조하여 연구의 차별성을 부각한 점이 인상적입니다.

더욱 완성도 높은 보고서를 위해 논리적 검토 및 개선 제안을 드립니다.

1. 논리적 일관성 및 오류 검토



claude랑 쓴 results 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

좋아, 이번에는 Results는 심사위원이 가장 공격적으로 보는 구간이니까 정말 냉정하게 보겠습니다.

결론부터 말하면:

치명적인 논리 오류는 없고, 데이터-주장 연결도 대체로 잘 되어 있음.
 다만 "과장으로 오해받을 수 있는 표현 3곳",
 "통계-방법론 명시가 부족해 보일 수 있는 지점 3곳",
 심사위원이 반드시 물을 질문 5개가 명확히 보입니다.

아래를 ① 논리적 오류 / ② 과장 가능성 / ③ 방법론-통계 취약점 / ④ 구조적 개선 제안 순서로 짚을게요.

본 단계에서는 단순한 문장·표현 검토를 넘어, Forecasting Models 및 Results 해석이 통계적 타당성과 기술적 정확성 측면에서 동시에 설득력을 가지는지를 검증하는 데 초점을 두었다.

이를 위해 ChatGPT와 Gemini에 서로 다른 검증 관점을 부여하여 동일 결과를 교차 검토하였으며, 상충되는 해석은 Discussion에서 연구자가 종합적으로 반영하였다.

연구자는 Section 3(Forecasting Models)과 Section 4(Results)의 실험 설계 타당성과 통계 해석 정확성을 검증하기 위해 ChatGPT와 Gemini에게 교차 검토를 요청하였다.

ChatGPT에게는 "실험 설계의 논리적 타당성과 통계 검정의 적절성"을, Gemini에게는 "모델 설명의 기술적 정확성과 결과 해석의 타당성"을 중점적으로 확인하도록 하였다. 두 AI 모두 지속적인 피드백을 통해 개선점을 도출하였다.

ChatGPT는 통계적 측면을 집중적으로 검토하였다. Paired t-test 사용의 적절성, $p < 0.001$ 유의수준 설정의 보수성, 그리고 다중비교 문제 처리 방식에 대해 타당성을 확인하였다.

또한 조건별 성능 비교(Operating Regimes, High-Change, EWS)에서 샘플 크기가 충분한지, 통계적 검정력이 확보되는지를 여러 차례 피드백을 거쳐 검증하였다.

Gemini는 기술적 정확성을 중점적으로 분석하였다. Chronos 모델 설명에서 "zero-shot", "pretrained representations", "transferable temporal structures" 등의 개념이 정확하게 서술되었는지, ARIMA와 Prophet의 한계를 설명하는 부분에서 기술적 오류가 없는지를 지속적으로 확인하였다.

두 AI는 Table 2(Operating Regime 성능)의 해석에서 서로 다른 관점을 제시하였다.

ChatGPT는 통계적 유의성에 초점을 맞춘 반면, Gemini는 실무적 함의(예: ARIMA의 ExtremeWeather에서 13,307 MW 폭발)를 강조하였다. 연구자는 두 관점을 모두 Discussion에 반영하여 균형잡힌 해석을 제공하였다.

8-3. ChatGPT와 Gemini를 활용한 Discussion/Conclusion 교차검증

claude랑 쓴 discussion 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

작성하는 과정 표시 ▼

제시해주신 Discussion 초안은 실험 결과의 의미를 계통 운영의 관점에서 깊이 있게 해석하고 있으며, 학술적 객관성을 유지하려는 노력이 돋보입니다. 특히 전통적 모델의 한계를 '평균의 함정(Average metrics)'으로 지적한 부분은 논문의 기여도를 높여주는 매우 날카로운 통찰입니다.

더욱 견고한 논리를 위해 검토가 필요한 지점과 개선 제안을 정리해 드립니다.

1. 데이터 일관성 및 논리적 검토

- **수치 일관성 확인:**
 - 현황: 도입부(Introduction)에서는 2024년 한국의 최대 수요를 **87.8 GW**로 언급하셨으나, Discussion 본문(Section 5.1)에서는 2024년 8월 26일 폭염 당시 수요가 **78 GW에서 93 GW**로 급증했다고 서술되어 있습니다.
 - 제언: 93 GW가 설비 용량이 아닌 실제 실시간 수요(Peak load)라면, 앞서 언급한 87.8 GW와 상충할 수 있습니다. 수치를 다시 확인하여 논문 전체에서 일관성을 유지하거나, '실시간 부하'와 '일일 최대 부하'의 차이라면 이를 명확히 구분해 주십시오.

4 Conclusion

The study investigated the effectiveness of a proposed time series foundation model for short-term electricity demand forecasting under realistic and operationally critical conditions. Using hourly metered electricity demand data from South Korea, we conducted a comprehensive empirical evaluation under a fair 10-hour ahead rolling forecast framework.

Across all experiments, the proposed foundation model consistently outperformed traditional forecasting approaches, including those based on statistical models and deep learning architectures. A key observation is that the proposed model's performance is robust to varying levels of data quality and noise, which is a critical requirement for real-world deployment. The model's ability to capture complex temporal dependencies and its superior performance in handling noisy data are key strengths. These results demonstrate that the proposed foundation model can substantially improve the practical value of electricity forecasting models.

The findings highlight the potential of foundation models to enhance power system resilience. By integrating time-series temporal representations learned through large-scale pre-training, the proposed model can provide more accurate and reliable forecasts, which is essential for efficient power system operation and planning. Future research should focus on extending the model's capabilities to handle more complex and diverse power system scenarios, such as those involving renewable energy integration and distributed energy resources.

Several limitations should be acknowledged. First, the analysis focuses on a single national-level electricity demand dataset; future work should explore integrating geospatial data to capture localized demand patterns. Second, the study evaluates forecast accuracy under fixed forecasting horizons, and future research should investigate the model's performance under varying forecast horizons and different system operational constraints.

Future work will extend this study by evaluating foundation models across multiple regions, integrating probabilistic forecasting and uncertainty quantification, and exploring continuous and adaptive learning capabilities. Hybrid approaches that combine foundation models with local domain-specific adaptive forecasting strategies are also being explored.

Throughout this study, large language models were used in a supportive role to help structure the evaluation framework, refine literature analysis, and generate preliminary conclusions. All experimental design choices, numerical analysis, and conclusions remain the responsibility of the authors.

By demonstrating that time series foundation models can enhance forecasting performance under realistic conditions, this work establishes foundation models as valuable tools for managing increasingly volatile power systems under climate change.

claude랑 쓴 conclusion 초안이야. 논리적 오류나, 과장된 주장, 개선점이 있으면 지적해줘

좋아, Conclusion은 심사위원이 "이 논문을 어떻게 기억할지"를 결정하는 구간이라서, 여기서는 논리·과장·톤을 특히 엄격하게 봐야 해.

결론부터 말하면:

큰 논리적 오류는 없고, 전반적으로 매우 성숙한 결론문이다.

본 단계에서는 Results 분석을 넘어, Discussion과 Conclusion에서 제시된 해석과 주장이 데이터 분석 결과의 범위를 벗어나지 않는지, 그리고 기술적·실무적 관점에서 과장 없이 타당한지를 최종적으로 검증하고자 ChatGPT와 Gemini를 활용한 교차 검증을 수행하였다.

특히 Discussion과 Conclusion은 모델 성능 결과를 일반화하거나 운영·정책적 시사점으로 확장하는 과정에서 해석의 비약이 발생하기 쉬운 구간이므로, 다각도의 검증이 중요하다고 판단하였다.

연구자는 Section 5(Discussion)와 Section 6(Conclusion)의 초안을 대상으로 ChatGPT와 Gemini에게 각각 다른 관점의 검토를 요청하였다.

ChatGPT에게는 주장 과도성 여부, 논리적 비약, 결과 대비 해석의 적절성을 중심으로 검토를 요청하였고, Gemini에게는 기술적 메커니즘 설명의 정확성 및 실무적 해석의 타당성을 중점적으로 확인하도록 하였다.

먼저 ChatGPT는 Discussion에서 제시된 주요 주장들이 Results 분석에 의해 충분히 뒷받침되는지에 집중하여 검토하였다.

특히 Section 5.1에서 제시된 "Chronos의 우수성은 pretrained representations에 기인한다"는 해석이 zero-shot 성능 결과와 논리적으로 연결되는지, 그리고 Section 5.2에서 언급된 "average metrics can obscure weaknesses"라는 주장이 조건별 성능 분석에 의해 충분히 뒷받침되는지를 반복적으로 점검하였다.

이 과정에서 ChatGPT는 일부 문장에서 표현의 강도를 완화할 것을 제안하며, "potentially", "suggest", "may"와 같은 보수적 학술 표현 사용을 권고하였다.

Gemini는 기술적 설명의 정확성을 중심으로 검증을 수행하였다.

Transformer attention mechanism이 극한 조건에서 성능 우위를 보일 수 있는 이유에 대한 설명이 self-attention의 특성과 일관되는지, zero-shot 설정에서 Chronos가 작동하는 메커니즘이 과도하게 단순화되어 있지 않은지 검토하였다.

또한 ARIMA의 stationarity 가정, Prophet의 additive decomposition 구조가 Extreme Weather 및 High-Change 조건에서 갖는 한계에 대한 설명이 기술적으로 정확한지를 세부적으로 점검하였다.

특히 Conclusion 부분에서는 두 AI가 상호 보완적인 관점을 제시하였다.

ChatGPT는 연구 결과의 일반화 가능성에 대해 보수적인 서술을 유지할 것을 강조한 반면, Gemini는 전력 계통 운영 측면에서의 실무적 가치(예: 급변 수요 상황에서의 dispatch 안정성, peak load 대응 가능성)를 보다 명확히 드러낼 것을 제안하였다.

연구자는 이 두 관점을 종합하여, 과장된 주장을 배제하면서도 연구의 실질적 의의를 명확히 전달하는 방향으로 최종 서술을 조정하였다.

결과적으로 본 단계는 단일 AI의 의견을 수용하는 과정이 아니라, 서로 다른 강점을 지닌 AI의 교차 검증을 통해 해석의 타당성과 표현의 균형을 동시에 확보한 과정이라 할 수 있다.

이를 통해 연구자는 Discussion과 Conclusion에서 제시된 모든 주장이 데이터, 통계 분석, 기술적 메커니즘에 의해 일관되게 뒷받침되도록 최종 원고를 확정하였다.

8-4. Claude를 활용한 종합 점검 및 최종 수정



너랑 쓴 연구보고서 초안을 gpt와 gemini를 통해 지속해서 교차검증을 해서 검증한 최종 수정본이야 한번 읽어보고 종합적으로 판단해줘

연구보고서를 꼼꼼히 읽어보았습니다. 전반적으로 매우 완성도 높은 논문으로 발전했네요! 종합적인 평가를 드리겠습니다.

🌟 주요 강점

1. 연구 설계의 엄밀성

- Rolling-origin 24-hour-ahead 평가 프레임워크가 일관되게 적용됨
- Zero-shot 설정을 엄격히 유지하여 일반화 능력 평가에 집중
- 2021-2023 학습, 2024 테스트로 명확한 데이터 분리

2. 조건부 분석의 체계성

- Operating Regimes (Normal/ExtremeWeather/Ramp)
- High-Change Events (top 1%)

본 단계에서는 ChatGPT와 Gemini를 통해 도출된 모든 검토 의견을 Claude에게 통합적으로 제시하고, 연구보고서 전반에 대한 최종 점검과 수정 방향을 확정하였다.

본 과정의 목적은 개별 AI의 피드백을 단순 반영하는 것이 아니라, 논문 전체의 논리적 일관성, 해석의 타당성, 학술적 보수성을 종합적으로 검증하는 데 있다.

연구자는 Claude에게 Introduction의 논리 구조, Results 및 Table 해석의 과도성 여부, Conclusion의 어조와 기여 표현의 적절성을 중심으로 검토를 요청하였다.

이에 대해 Claude는 (1) “극한 조건에서의 상대적 우수성”이라는 핵심 메시지의 일관된 유지, (2) 통계적 엄밀성과 실무적 가치의 균형, (3) 과도한 일반화를 피하는 보수적 서술이라는 세 가지 원칙을 제시하였다.

이 원칙에 따라 Introduction의 문단 순서를 조정하여 문제 제기와 연구 필요성의 흐름을 강화하였고, Results 해석에서는 수치는 Table과 Figure에 집중시키고 본문에서는 핵심 패턴 중심으로 서술을 정제하였다. Conclusion에서는 단정적 표현을 완화하고 학술 논문에 적합한 중립적 어조를 채택하였다.

마지막으로 Claude는 논문 전체를 통독하며 Section 간 연결성을 점검하였고, 연구 질문-결과-해석-결론이 일관되게 대응됨을 확인하였다.

이 과정을 통해 연구자는 단일 AI에 의존하지 않고 다수 AI의 상호 보완적 의견을 통합하여 최종 원고를 완성하였다