Time-LLM 프레임워크를 통한 시계열 예측 성능에 대한 LLM의 영향 분석

공무경, 임수환, 방대호, 박민수, 강재모*

경북대학교

{rhdanrud1000, tngks9317, dkrlwnstn1, operioncwal, *jmkang}@knu.ac.kr

Impact of LLM Backbones on Time Series Forecasting Performance: A Study Based on the Time-LLM Framework

Kong Mu Kyeong, Im Su Hwan, Bang Dae Ho, Park Min Soo, Kang Jae Mo*

Kyungpook National Univ.

요 약

시계열 예측은 금융, 에너지 등 다양한 분야에서 효율적 자원 분배와 위험 관리에 필수적이다. 그러나 ARIMA나 LSTM과 같은 기존 방법은 데이터 부족 상황에서 일반화 성능에 한계를 보인다. 이에 최근 방대한 사전학습을 통해 강력한 패턴 인식과 제로샷 능력을 지닌 대규모 언어 모델(LLM)을 활용하려는 연구가 주목받고 있다. 본 연구는 Time-LLM 프레임워크에서 서로 다른 LLM 백본을 교체하며 시계열 예측 성능에 미치는 영향을 분석하였다. 첫 번째 실험에서는 Qwen3 계열(4B, 8B, 14B)을 사용해 파라미터 수의 효과를 검증했고 두 번째 실험에서는 BBH 점수가 다른 Qwen3-8B와 Llama3-8B를 비교하여 추론 능력의 기여도를 평가하였다. ETTh1 데이터셋을 대상으로 동일한 학습 환경에서 MSE와 MAE로 성능을 측정한 결과, Qwen3-4B가 가장 낮은 오차를 기록하여 단순히 파라미터 규모가 성능 향상으로 이어지지 않음을 확인하였다. 반면 Qwen3-8B는 Llama3-8B보다 현저히 우수한 성능을 보여, 모델의 추론 능력(BBH)과 아키텍처 등 계열 고유의 특성이 성능에 중요한 영향을 미침을 시사한다.

Ⅰ. 서 론

시계열 예측은 금융, 에너지 등 다양한 산업에서 효율적인 자원 분배 및 위험 관리와 같은 합리적 의사결정에 필수적이다. 기존 ARIMA[1]나 LSTM[2]과 같은 방법론들은 특정 도메인에 특화되어야 하고 데이터가 부족할 경우 일반화 성능에 한계를 보여왔다. 이러한 문제를 해결할 대안으로 최근 방대한 데이터로 사전 학습되어 뛰어난 패턴 인식과 제로샷 학습 능력을 갖춘 대규모 언어 모델(LLM)을 활용한 연구가 활발히 진행되고 있다. LLM은 사전 학습된 지식을 바탕으로 데이터가 부족한 환경에서도 높은 일반화 성능을 기대할 수 있다는 점에서 기존 모델 대비 명확한 이점을 가진다.

그러나 LLM을 시계열 예측에 적용하는 데에는 근본적인 도전 과제, 즉 모달리티 간극이 존재한다. LLM은 이산적인 텍스트 토큰을 처리하도록 설계된 반면 시계열 데이터는 본질적으로 연속적인 숫자 값으로 이루어져 있기 때문이다. Time-LLM[3], Chronos[4], TimeCMA[5] 와 같은 최신 모델들은 이러한 간극을 해소하기 위해 연속적인 시계열을 LLM이 이해할 수 있도록 변환하는 다양한 '입력 표현 전략'을 제시한다. 특히 Time-LLM은 시계열 데이터를 LLM이 이해할 수 있도록 변환하는 Patch Reprogramming 과 도메인·태스크·통계 정보를 접두사로 제공하는 Prompt-as-Prefix(PaP) 를 핵심 아키텍처로 사용한다. 이를 통해 동결된 LLM 백본을 그대로 활용하면서도 시계열 예측이 가능하다.

본 논문은 Time-LLM 프레임워크에서 LLM 백본의 선택이 시계열 예측 성능에 중대한 영향을 미친다는 가설을 검증하고자 한다. 이러한 영향은 모델의 파라미터 규모를 넘어 특정 벤치마크로 측정된 모델의 고유 역량과도 상관관계가 있을 것으로 가정한다. LLM의 파라미터 규모와 추론 능력이 시계열 예측 정확도에 미치는 영향을 분석하여 향후 최적의 LLM 백본을 선택하는 데 기여하고자 한다.

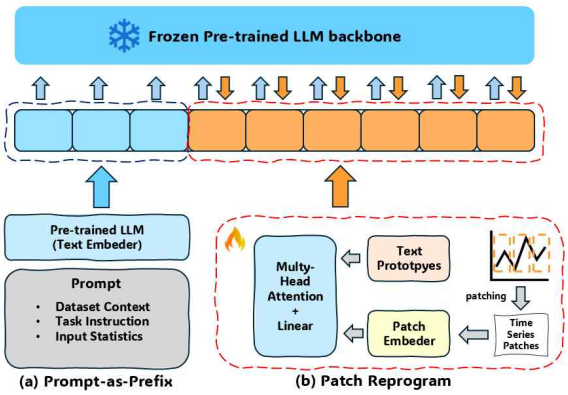

그림 1 Time-LLM 핵심 아키텍처

Ⅱ. 모델 선정 및 실험 설계

2.1. 모델 선정

본 연구는 LLM 백본 모델의 고유 역량이 시계열 예측 성능에 미치는 영향을 검증하기 위해 설계되었다. 이를 위해 모달리티 간극을 효과적으로 해소하면서도 LLM 자체는 수정하지 않는 Time-LLM 프레임워크를 실험 모델로 채택했다. Time-LLM은 LLM의 가중치를 frozen 상태로 유지하므로 백본 모델을 교체했을 때 나타나는 성능 변화를 LLM의 순수한 역량 차이로 귀속시켜 분석하기에 최적의 환경을 제공한다.

첫 번째 실험에서는 LLM의 일반화 능력이 시계열 예측 성능에 미치는 영향을 평가하기 위해 설계한다. 이를 위해 모델 선정 기준으로 파라미터 수를 채택하였다. 파라미터 규모는 모델이 학습 데이터에서 복잡한 패턴을 포착하고 새로운 상황에 적용할 수 있는 잠재력과 직접적으로 연관된다고 알려져 있다[6]. 따라서 동일한 아키텍처 내에서 파라미터 수가 다른

모델들을 백본으로 교체하여 규모 차이가 실제 예측 정확도에 어떠한 변화를 가져오는지를 분석하기 위하여 Qwen3[7] 계열의 4B, 8B, 14B 모델을 사용하였다.

두 번째 실험에서는 LLM의 추론 능력이 예측 정확도에 미치는 기여도를 검증하기 위해 설계되었다. 이를 위해 널리 사용되는 표준 벤치마크인 BIG-Bench Hard (BBH)[8] 점수를 모델 선정 기준으로 활용하였다. BBH는 논리, 상식, 알고리즘적 사고 등 단순한 패턴 학습을 넘어서는 복합적인 문제 해결 능력을 평가하는 지표로서 LLM의 범용적 추론 역량을 반영한다. 따라서 표 1에서 볼 수 있듯 파라미터 규모는 유사하게 유지하되 BBH 점수에서 큰 차이를 보이는 Qwen3-8B(78.40%)와 Llama3-8B[9](57.50%)를 비교 대상으로 선정하였다.

Model	Parameters (B)	BBH Score (%)
Qwen3-4B	4.02B	72.59
Qwen3-8B	8.19B	78.40
Qwen3-14B	14.8B	81.07
Llama3-8B	8.03B	57.70

표 1 실험에 사용된 백본 LLM의 파라미터 수와 BBH점수

2.2. 실험 설계

본 연구의 실험은 Time-LLM 프레임워크를 기반으로 하며 백본 LLM을 교체하는 방식으로 모델 간 성능을 비교하였다. 데이터셋은 ETT-small에 포함된 ETTh1 (Electricity Transformer Temperature, hourly) 시계열을 사용하였다. ETTh1은 전력 변압기의 시간 단위 온도 데이터를 포함한 다변량 시계열로 복잡한 계절성과 추세 패턴을 지니고 있어 시계열 예측 모델의 성능을 평가하는 표준 벤치마크로 널리 활용된다.

본 연구에서는 Time-LLM 프레임워크를 유지한 채 백본 LLM을 Qwen3-4B, Qwen3-8B, Qwen3-14B, Llama3-8B의 네 가지 모델로 교체하여 시계열 예측 성능을 비교 평가하였다. 모든 실험에서 LLM의 가중치는 동결(frozen) 상태로 유지하였으며 동일한 학습 환경 하에서 백본 교체로 인한 순수한 모델 역량 차이를 분석하였다.

성능 평가는 시계열 예측에서 표준적으로 사용되는 평균제곱오차(Mean Squared Error, MSE)와 평균절대오차(Mean Absolute Error, MAE)를 지표로 삼았다. MSE는 큰 예측 오차에 높은 페널티를 부여하여 모델이 이상치와 큰 편차에 얼마나 민감한지를 측정할 수 있고 MAE는 평균적인 예측 오차 크기를 직관적으로 보여준다. 두 지표를 함께 사용함으로써 모델의 정밀성과 안정성을 균형 있게 평가할 수 있도록 설계하였다.

Ⅲ. 실험 결과 및 결론

Model	MSE ↓	MAE ↓
Qwen3-4B	0.3756	0.4027
Qwen3-8B	0.3941	0.4133
Qwen3-14B	0.3821	0.4033
Llama3-8B	0.4731	0.4597

표 2 LLM 백본 모델별 시계열 예측 성능 비교 결과

3.1. 실험 결과

표 2는 Time-LLM 프레임워크에서 서로 다른 백본 LLM을 적용했을 때의 성능 비교 결과를 보여준다. 전체적으로 Qwen3 계열이 Llama3-8B보다 안정적으로 우수했으며, 특히 Qwen3-4B가 가장 낮은 MSE와 MAE를 기록하여 Qwen3-14B와 같이 파라미터 수가 상대적으로 큰 모델과 동등하거나 더 나은 성능을 달성하였다. 이는 모델 규모의 증가가 성능 향상으로 단조적으로 이어지지 않음을 의미한다. 반면, 동일한 8B 규모에서 Qwen3-8B는 Llama3-8B를 크게 상회하여 모델 계열과 추론 능력(BBH)이 성능 결정의 핵심 요인임을 확인할 수 있었다.

3.2. 결론

본 연구는 Time-LLM 프레임워크에서 LLM 백본 교체가 시계열 예측 성능에 미치는 영향을 분석했다. 실험 결과, 파라미터 수의 증가가 예측 정확도의 향상으로 직결되지 않는다는 점이 확인되었다. Qwen3-4B는 8B와 14B와 비교하여 가장 낮은 예측 오차를 기록하였고 이는 일반적으로 모델 규모가 커질수록 성능이 향상된다는 경향이 시계열 예측과 같은 특수한 태스크에서는 그대로 적용되지 않음을 보여준다. 다만 이는 실험에 사용된 ETTh1 데이터셋의 패턴이 비교적 단순하여 4B 모델의 용량으로도 충분히 분석 가능했기 때문일 수 도 있다. 즉, 이번 실험에서는 대규모 모델이 제공하는 고도의 패턴 인식 및 추론 능력이 요구되지 않았으며 단순한 모델 규모 확장이 항상 최적의 전략은 아님을 시사한다.

반면, 동일한 파라미터 규모를 가진 모델 간 비교에서는 뚜렷한 성능 차이가 관찰되었다. Qwen3-8B는 Llama3-8B 대비 현저히 우수한 성능을 보였으며, 이는 두 모델의 파라미터 수가 유사함에도 불구하고 BBH 점수로 측정되는 추론 능력과 계열적 특성의 차이가 실제 예측 정확도에 영향을 미친다는 점을 시사한다.

따라서 본 연구는 시계열 예측에서 LLM의 성능을 결정짓는 요인이 단순한 파라미터 규모가 아니라 모델 계열과 내재된 추론 능력임을 강조한다. 특히, ETTh1과 같은 단순한 데이터셋에서는 중간 규모 모델이 오히려 최적의 선택일 수 있음을 확인하였다. 그러나 파라미터 규모의 효과를 보다 정확하게 이해하기 위해서는 다양한 난이도와 복잡성을 가진 데이터셋을 활용한 추가적인 분석이 필요하다. 이러한 결과는 향후 LLM 기반 시계열 예측 시스템을 설계할 때, 모델 크기만을 기준으로 선택하기보다는 추론 능력, 계열적 특성, 데이터 특성을 종합적으로 고려한 전략적 접근이 필요함을 보여준다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2020-II201808)

참 고 문 헌

[1] Box, George EP, et al. Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[2] Graves, Alex. "Long short-term memory." Supervised sequence labelling with recurrent neural networks (2012): 37-45.

[3] Jin, Ming, et al. "Time-llm: Time series forecasting by reprogramming large language models." arXiv preprint arXiv:2310.01728 (2023).

[4] Ansari, Abdul Fatir, et al. "Chronos: Learning the language of time series." arXiv preprint arXiv:2403.07815 (2024).

[5] Liu, Chenxi, et al. "Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 18. 2025.

[6] Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).

[7] Yang, An, et al. "Qwen3 technical report." arXiv preprint arXiv:2505.09388 (2025).

[8] Srivastava, Aarohi, et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." Transactions on machine learning research (2023).

[9] Grattafiori, Aaron, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).