

# 집단지성 vs. 전문가: 저성능 LLM 집단과 단일 고성능 LLM의 주가 반응 예측 정확성 비교

## Collective Intelligence vs. Expertise: Comparing the Stock Reaction Prediction Accuracy of Low-Performance LLM Collectives and a Single High-Performance LLM

### 요약

본 연구는 금융 이벤트 예측에서 전문가 모델(GPT-5)과 페르소나 기반 집단지성 모델(GPT-5-nano)의 성능을 체계적으로 비교·분석하였다. 본 연구의 목적은 복잡한 금융 시장 예측에서 전문가 모델의 깊이 있는 논리와 다양한 관점이 결합된 집단지성 모델 중 어느 것이 더 높은 예측력을 보이는지를 규명하는 것이다. 이를 위해 S&P 100 기업의 주요 이벤트 124 건을 대상으로 주가 반응 예측 실험을 수행하였다. 집단지성 모델은 European Social Survey(ESS) 기반 페르소나 데이터를 활용하여 인구통계학적 요인을 반영한 25개의 가상 투자자 sLLM으로 구성되었으며, 각 모델의 예측 결과를 확신도 가중 평균 방식으로 종합하였다. 한편 전문가 모델에는 Chain-of-Thought(CoT) 프롬프트를 적용하여 단계적 논리 추론을 수행하도록 하였다. 실험 결과, 집단지성 모델의 평균 정답률은 71.29%로 전문가 모델(64.19%)보다 유의하게 높았으며, 가중점수(+40.21 vs. +19.34), 평균 수익률(+7.37% vs. +4.81%)에서도 통계적으로 유의미한 우위를 보였다( $p < 0.01$ ). 이러한 결과는 다양한 관점이 결합된 집단지성 모델이 단일 전문가 모델보다 복잡한 금융 시장의 변동성과 불확실성을 보다 효과적으로 반영함을 시사한다. 본 연구는 LLM을 활용한 금융 의사결정에서 모델의 규모 확대보다 이질적 사고 구조의 통합이 더 큰 예측 효율성을 가져올 수 있음을 실증적으로 보여준다<sup>1</sup>.

### 1. 서론

최근 금융 시장은 방대한 비정형 데이터와 빠른 정보 확산으로 인해 예측을 위해 처리 및 분석해야 하는 데이터의 규모가 커지고 있다. 이러한 변화에 따라 머신러닝과 딥러닝을 넘어, 언어 기반 추론이 가능한 대규모 언어 모델(Large Language Model, LLM)이 금융 의사결정 보조 도구로 주목받고 있다[1].

본 연구는 증권 전문 뉴스의 내용을 통해, 금융 이벤트가 주가에 미치는 영향을 예측하는 과정에서, 하나의 추론 중심 고성능 LLM과 페르소나가 부여된 경량화 언어 모델(smaller Large Language Model, 이하 sLLM) 집단의 성능을 비교·분석한다.

본 연구에서는 비교 대상 모델을 다음과 같이 정의한다. ‘전문가 모델’은 단일 고성능 LLM(GPT-5)에 Chain-of-Thought(이하 CoT) 프롬프트를 적용하여 심층적 논리 추론을 수행하는 모델이다. ‘집단지성 모델’은 sLLM(GPT-5-nano)으로 구성되었으며, 각 모델에 유럽사회조사(European Social Survey, 이하 ESS)<sup>2</sup> 기반 페르소나를 부여하여 실제 투자자들의 이질적 특성을 반영한 판단을 수행하는 모델이다.

본 연구의 목적은 복잡한 금융 시장 예측에서 단일 전문가 모델의 깊이 있는 논리와 다수의 다양한 관점이 결합된 집단지성 모델 중 어느 것이 더 높은 예측력을 보이는지를 규명하는 것이다.

이를 검증하기 위해 S&P 100 지수에 포함된 기업들의 주요 이벤트를 대상으로 예측 실험을 진행하였으며, 두 모델 그룹의 예측 정확도, 확신도 기반 점수를 비교·평가하였다. 또한 통계적 유의성 검정을 통해 결과의 신뢰도를 확보하였다.

### 2. 관련 연구 및 배경지식

집단지성은 다수의 독립적이고 다양한 판단을 합리적으로 결합할 때, 개인 전문가의 판단을 능가할 수 있다는 이론이다. 집단지성의 핵심 원리는 ‘생각의 다양성’과 ‘구성원의 이질성’을 확보하는 것이다[2]. 최근 상이한 12개 LLM으로 구성된 집단을 통해 예측 성능을 향상하는 성과를 달성했다는 연구 결과가 보고되었다[3]. 이 연구는 LLM 간의 사고 경로 다양성이 집단적 성과 향상에 핵심적으로 기여함을 보여준다. 본 연구는 이러한 관점을 확장하여, 다수의 sLLM에 서로 다른 페르소나를 부여함으로써 집단지성의 핵심 조건인 인간 사회의 이질성을 AI 수준에서 구현하고자 한다.

또한, 투자 의사결정은 합리적 경제 주체 가정과 달리 다양한 심리적 편향에 의해 체계적으로 왜곡된다. 이러한 편향과 왜곡은 시장의 변동성을 심화시킨다. 이러한 행동경제학적 통찰을 반영하기 위해 ESS의 인구통계학적·심리적 변수를 활용하여 다수의 페르소나를 구성하였다. 이를 통해, 서로 다른 성향의 sLLM들이 동일한 금융 이벤트를 어떻게 해석하고 의사결정을 내리는지를 분석한다.

한편, 전문가 모델 기반 접근에서는 CoT 프롬프트가 복잡한 논리적 추론을 가능하게 하는 핵심 기법으로 활용된다. CoT 프롬프트가 중간 추론 단계를 명시함으로써 수학적·논리적 문제 해결 능력을 크게 향상시킨다는 연구 결과가 보고되었다[4]. 금융 도메인에서는 CoT를 적용한 전문가 모델이 뉴스, 기업 실적, 거시경제 지표 등 다양한 요인을 종합적으로 고려해 일관된 추론을 수행할 수 있다는 장점이 있다. 그러나 단일 모델은 특정 추론 경로에 고착되거나, 학습 데이터의 편향을 벗어나기 어렵다는 구조적 한계를 가진다.

따라서 본 연구는 행동경제학적 이질성을 갖춘 집단지성 모델과 CoT 기반 전문가 모델의 금융 시장 예측 효율성을 비교함으로

1) 관련 코드는 오른쪽 링크에 있습니다.<https://github.com/yujinworld/collective-vs-expert-llm>

2) European Social Survey, Round 11 (2022).

써, 집단적 합리성이 개인의 전문적 추론 대비 어떤 시장 조건과 예측 환경에서 우위를 점하는지를 체계적으로 규명하고, AI 기반 의사결정 모델 설계에 실질적인 함의를 제공하고자 한다.

### 3. 실험

#### 3.1 데이터

본 연구는 주가 반응 예측 실험을 위해 금융 데이터와 ESS 기반 페르소나 데이터를 활용하였다.

**주식 관련 데이터** 분석 대상은 S&P100 지수에 포함되어 있는 국제 주요 기업들로 한정하였다. 예측에 투입된 데이터는 1) 개별 종목 일 종가, 2) 관련 뉴스 제목 및 기사가 있다. 일 종가 데이터의 경우 Yahoo Finance API를 활용하여 데이터를 불러온 뒤, 날짜, 종가를 추출하였다. 관련 뉴스의 제목 및 본문의 경우 Google RSS 기반 오픈소스 라이브러리 GNews를 통해 URL을 수집한 뒤, Selenium 및 Newspaper 라이브러리를 활용하여 데이터를 수집하였다. 수집된 데이터는 학습용과 테스트용으로 분할하였으며, 모델의 학습 데이터에 포함되지 않은 새로운 이벤트만 테스트셋에 포함시켜 data leakage를 차단하였다.

**페르소나 관련 데이터** 집단지성 기반 실험을 위해 ESS 데이터를 활용하여 100개의 가상 투자자 페르소나를 생성하였다. ESS 원본 변수는 인구통계, 사회경제적 지위, 가치관, 신뢰 수준, 정치 성향, 정보 습관 등 광범위한 개인 특성을 포함한다. 본 연구에서는 이를 네 가지 핵심 차원으로 재분류하였다. 첫째, Identity 차원은 국적, 연령, 성별, 교육 수준을 포함한다. 둘째, Socio-Economics 차원은 직업, 소득 분위, 정치적 성향을 반영한다. 셋째, Outlook and Trust 차원은 행복도, 국가 경제 전망, 가계 재정 만족도, 대인 및 제도 신뢰도로 구성된다. 마지막으로 Information Habit 차원은 정치 관심도, 뉴스 및 인터넷 사용 빈도, 최근 선거 참여 여부를 포함한다.

페르소나의 분포는 표 1과 같다. 이러한 분포를 통해 실제 사회의 다양한 인구통계학적 특성을 반영하여 집단 내 이질성을 확보하고자 하였다.

#### 3.2 LLM 모델 구성 및 평가 전략

전문가 모델과 집단지성 모델을 비교하기 위해 두 가지 대조적인 LLM 그룹을 설정하였다. 각 모델은 매수 또는 비매수의 투자 결정과 함께 해당 판단에 대한 확신도를 0에서 100 사이의 값으로 보고하였다.

**단일 LLM: 전문가 모델** 전문가 모델 실험에는 고성능 추론 모델인 GPT-5를 활용하였다. 금융 전문가의 심층적 분석 능력을 모방하기 위해 CoT 기반 프롬프트 구조를 시스템 지침으로 적용하였다. 이를 통해 모델은 단순히 최종 결론만을 제시하는 것이 아니라, 이벤트 해석, 유사 사례와의 비교, 잠재적 요인 도출, 최종 결론 도출이라는 명시적인 단계별 추론 과정을 거치도록 유도되었다. 이러한 구조화된 추론 과정은 모델이 복합적인 금융 정보를 체계적으로 분석하고 논리적 근거를 제시하도록 설계되었다.

표 1: 페르소나 데이터 요약 통계

구분	항목	비율(%)
성별	남성	58.0
	여성	42.0
연령대	10대	4.0
	20대	9.0
	30대	11.0
	40대	24.0
	50대	24.0
	60대	15.0
	70대	10.0
	80대	3.0
	초등/중등 이하	24.0
교육 수준	고등/직업교육	60.0
	학사	14.0
	석사/박사	2.0
	낙관적	45.0
가계 재정 만족도	보통	40.0
	비관적	15.0
	높음	29.0
정치 관심도	중간	63.0
	낮음	8.0
	낮음	29.0

**집단 sLLM: 집단지성 모델** 집단지성 모델 실험에서는 25개의 sLLM(GPT-5-nano)을 독립된 에이전트로 구성하였다. 각 모델은 앞서 생성된 100개의 ESS 기반 페르소나 중 하나를 무작위로 할당받아, 해당 페르소나의 인구통계학적 특성과 심리적 성향에 따라 금융 이벤트를 해석하도록 설계되었다. 최종 집단 의사결정은 단순 다수결 방식이 아닌, 각 모델이 보고한 확신도를 가중치로 활용한 가중 평균 방식으로 산출하였다. 이러한 접근은 높은 확신을 가진 개별 판단이 집단의 최종 결정에 더 큰 영향을 미치도록 함으로써, 실제 시장 참여자들의 집단 의사결정 과정을 보다 사실적으로 반영하고자 하였다.

#### 3.3 결과 및 분석

예측 이벤트로는 주가 등락폭이 10% 이상인 경우를 사용하였다. 총 124개의 테스트 사례를 대상으로 이벤트 발생 시점으로부터 1, 3, 7, 15, 30일이 경과한 시점에서의 예측을 실험하였다. 예측일이 휴장일일 경우, 가장 가까운 미래의 영업일을 기준으로 예측을 수행하였다. 따라서 각 모델은 620회의 예측을 시행하였다. 단순 정답 기준 평가에서 전문가 모델의 정답률은 64.19%, 집단지성 모델의 정답률은 71.29%로 나타났다. 즉, 집단지성 모델이 약 7.1%p 높은 정확도를 기록하며 전반적인 우위를 보였다.

또한 모든 예측 시점에서 집단지성 모델이 일관된 우위를 보였다. 특히 1일 후 예측에서는 전문가 모델의 정답률 65.32% 대비 집단지성 모델이 75.00%를 기록하며 약 9.7%p 높은 정확도를 보였다. 가중점수 또한 전문가 모델(+25.85)과 집단지성 모델 (+46.58)은 유의미한 차이를 보였다. 3일, 7일, 15일 구간에서도

집단지성 모델의 우세가 유지되었으며, 30일 구간에서는 두 모델의 정답률이 동일하였으나 집단지성 모델의 가중점수가 더 높게 나타났다.

표 2: 예측 시점별 성능 비교

예측시점	전문가 정답률 / 가중점수	집단지성 정답률 / 가중점수
1일 후	65.32% / +25.85	75.00% / +46.58
3일 후	63.71% / +18.93	73.39% / +44.41
7일 후	62.10% / +15.32	72.58% / +43.39
15일 후	62.90% / +15.55	68.55% / +35.68
30일 후	66.94% / +21.06	66.94% / +30.97

표 3의 종합 점수 기준으로, 전문가 모델의 평균 가중점수는 +19.34, 집단지성 모델은 +40.21로 나타났다. 이는 집단지성 모델이 단순 정답률뿐 아니라 확신도 기반 평가에서도 일관된 우위를 보였음을 의미한다. 결론적으로, 집단지성 모델은 전 구간

표 3: 종합 성능 비교

모델	정답률(%)	평균 가중점수
전문가	64.19	+19.34
집단지성	<b>71.29</b>	<b>+40.21</b>

에서 전문가를 상회하였으며, 특히 단기(1-7일) 예측 정확도에서 두드러진 차이를 보였다. 반면 장기(30일) 구간에서는 양측의 예측이 수렴하는 양상을 보였다. 이러한 현상은 효율적 시장 가설의 관점에서 해석할 수 있다. 즉, 시간이 지남에 따라 시장 참여자들이 정보를 충분히 반영하게 되면서 개별 판단의 차이가 줄어든 것으로 이해할 수다.

### 3.4 통계 검정

추가적으로, 예측의 품질을 다각도로 평가하기 위해 통계 검정을 수행하였다. 집단지성 모델의 Precision, Recall, F1-score는 각각 0.7036, 0.7129, 0.7041로, 전문가 모델의 0.6511, 0.6419, 0.6456 대비 전반적으로 우수하였다. 또한 평균 수익률은 집단지성 모델이 +7.37%로 전문가 모델(+4.81%) 대비 약 2.56%p 높았으며, T-검정( $p = 0.0081$ ) 및 Mann-Whitney U 검정( $p = 0.0124$ ) 결과 모두에서 집단지성 모델의 수익률 우위가 통계적으로 유의미함을 확인하였다. 승률(수익으로 이어진 예측 비율) 또한 집단지성 모델 71.29%, 전문가 모델 64.19%로, 카이제곱 검정 ( $\chi^2 = 6.82, p = 0.009$ ) 결과 유의한 차이를 보였다. 확신도 수준 역시 집단지성 모델이 평균 89.6점으로 전문가 모델의 60.8점보다 유의미하게 높았으며( $p < 0.001$ ), 확신도가 높을수록 수익률이 증가하는 양의 상관관계(전문가  $r = 0.23$ , 집단지성  $r = 0.17$ )가 관찰되었다. 전반적으로 집단지성 모델은 단순 정답률뿐 아니라 예측 신뢰도, 수익률, 통계적 유의성 전 항목에서 일관된 우위를 보였다.

## 4. 결론

본 연구는 전문가 모델과 집단지성 모델의 금융 예측 성능을 비교 분석하여, 인공지능을 활용한 금융 의사결정의 새로운 패러다임을 탐색하고자 하였다. 실험 결과, 다양한 페르소나를 기반으로 한 집단지성 모델이 핵심 평가 지표에서 전문가 모델을 통계적으로 능가하며 그 우수성을 입증하였다.

이러한 결과는 금융 시장 예측에서 단일화된 심층 분석보다 다양화된 관점의 통합이 더 강력한 예측력을 가질 수 있음을 실증적으로 보여준다. 전문가 모델이 CoT를 통해 구조적이고 논리적인 분석을 수행했음에도 불구하고, 복잡하고 때로는 비합리적인 요소가 작용하는 금융 시장의 본질을 완벽히 포착하는 데에는 한계가 있었을 수 있다. 반면, 집단지성 모델은 ESS 데이터를 기반으로 한 다양한 인구통계학적, 심리적 특성을 반영함으로써 시장 참여자들의 이질적인 편향, 기대, 정보 해석 방식을 종합적으로 모사했다. 결과적으로 다양한 관점들이 서로의 맹점을 보완하며 만들어 낸 종합적 판단이 시장의 복잡성을 더 효과적으로 투영한 것으로 해석된다.

본 연구는 더 큰 모델을 향한 경쟁을 넘어, 다수의 소형 에이전트들을 효과적으로 구성하고 그들의 집단적 지성을 활용하는 ‘AI 오케스트레이션’의 중요성을 부각했다는 점에서 의미가 크다. 이는 LLM을 활용한 금융 의사결정에서 집단 지성, 군중의 지혜가 실질적인 성과로 이어질 수 있음을 입증한 것이다.

본 연구는 모델 간 상호작용이 없는 독립적 구조를 채택하였으나, 실제 집단지성의 본질은 토론과 정보 교환을 통한 협의 과정에 있다. 향후 연구에서는 모델 간 상호작용이 가능한 토론형 협의 구조나 계층적 집단 프레임워크로의 확장이 필요하다. 또한 전문가 모델의 논리적 일관성과 집단지성 모델의 다양성을 결합한 하이브리드 예측 프레임워크에 대한 연구 또한 기대된다.

## 참고 문헌

- [1] A. Lopez-Lira and Y. Tang, “Can chatgpt forecast stock price movements? return predictability and large language models,” 2024.
- [2] F. Galton, “Vox populi,” 1907.
- [3] P. Schoenegger, I. Tuminauskaite, P. S. Park, R. V. S. Bastos, and P. E. Tetlock, “Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy,” *Science Advances*, vol. 10, no. 45, p. eadp1528, 2024.
- [4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.