DS INVESTMENT & SECURITIES

2026.07.23

Quantitative Analysis

DS Quant Note

김남일 퀀트
02-709-2673
namil.kim@ds-sec.co.kr

한 번에 한 토큰의 시대는 끝날까

ICML 2026 구두발표와 스팟라이트에서 본 장기 작업과 추론비용

올해 ICML에서 가장 선명했던 변화는 더 큰 모델보다 더 긴 작업을 끝내는 시스템을 평가하기 시작했다는 점이다. ICML 2026 정규 포스터 UltraHorizon은 부분관측 상태에서 20만 토큰과 400회가 넘는 도구 호출을 요구했다. 스팟라이트로 선정된 VenusBench-Mobile은 계속 바뀌는 모바일 화면에서 장기 과업의 완주 여부를 측정했다. 평가 기준은 짧은 문항의 정답률에서 긴 작업의 완주율과 실패 원인으로 넓어지고 있다.

전문영역에서는 고비용 추론을 어디에 적용할지 논의가 구체화됐다. 법률과 의료는 검색 결과의 정확성뿐 아니라 상황별 준수와 검증 절차를 평가했다. 수학은 미해결 문제의 발견과 자동 검증을 연결했고 생물학에서는 실험 피드백을 다음 행동에 반영하는 연구가 구두발표로 이어졌다. 성공한 작업 한 건의 가치와 오류 비용이 모두 큰 분야부터 긴 추론과 전문 도구의 결합이 시험되고 있다.

범용 벤치마크의 상위권 점수는 빠르게 모였지만 실제 업무의 실패 조건은 여전히 잘 드러나지 않는다. ICML 2026은 VenusBench-Mobile, UltraHorizon, WorldTravel과 ComplexMCP처럼 상태가 바뀌거나 여러 제약을 동시에 맞춰야 하는 과제를 제시했다. LongCoT는 수학과 화학, 컴퓨터과학을 포함한 전문가 문제에서 긴 추론의 일관성을 따졌다. 모델 비교의 초점은 짧은 문항의 평균 정답률보다 완주율, 실패 원인과 불확실성으로 이동했다.

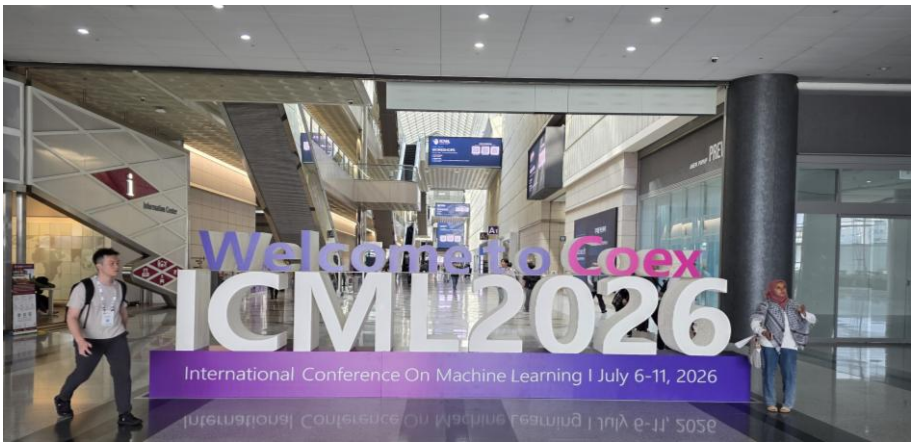
사용량 증가는 추론 인프라의 물리적 제약을 키운다. 자기회귀 모델은 문맥이 길고 요청이 몰릴수록 KV 캐시가 불어나지만 Diffusion은 여러 토큰을 함께 고쳐 이 의존을 낮출 여지가 생긴다. 반복 연산과 품질 격차가 남아 패널도 자기회귀, Diffusion과 하이브리드의 공존에 무게를 뒀다. 당분간 확인할 변수는 자기회귀 모델의 전면 대체 여부보다 KV 캐시 부담을 낮출 기술 옵션의 가치다.

평가 기준이 정답률에서 작업 완주율로 확대

ICML 2026은 서울에서 메인 행사와 워크숍을 이어 갔음

ICML은 머신러닝 이론부터 학습 방법, 시스템과 산업 응용까지 한자리에 모으는 국제학회다. NeurIPS와 ICLR도 함께 3대 인공지능 학회로 꼽힌다. 제43회 행사는 2026년 7월 6일부터 11일까지 서울 COEX에서 열렸다. 7일부터 9일까지 메인 컨퍼런스가 이어졌고 마지막 이틀은 워크숍이 채웠다.

그림1 ICML 2026 서울 COEX 행사장

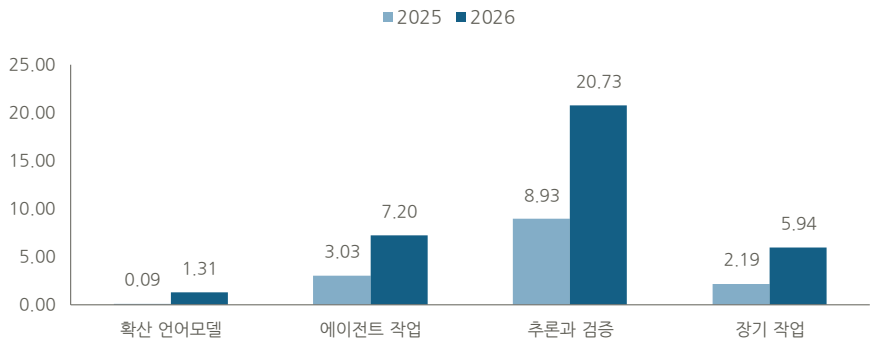


자료: ICML 2026 현장 촬영, DS투자증권 리서치센터

올해는 생성 가능성보다 작업의 완주와 실패 조건을 중요시 여김

작년과 올해의 차이는 논문 수뿐 아니라 연구 질문에서 더 분명했다. 생성 가능성을 보여 주는 데서 그치지 않고 작업을 끝낼 수 있는지와 중간에 어디서 실패하는지를 평가하기 시작했다. 모델 크기보다 서비스 비용이 발생하는 지점도 쟁점이 됐다. 전시장과 워크숍에서는 완주율, 실패 원인과 비용이 반복해서 논의됐다.

그림2 ICML 포스터의 연구 주제 변화



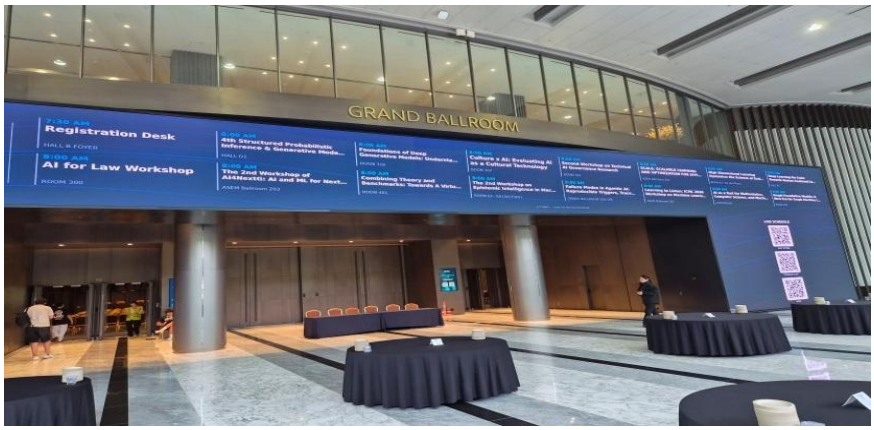
자료: ICML, DS투자증권 리서치센터
주: 2025년 3,339건과 2026년 6,628건의 제목 및 초록을 같은 비배타 키워드로 검색

공식 포스터는 에이전트와 검증 쪽으로 늘어났으며 워크숍 수와 제안서 모두 늘며 세부 의제가 확장함

제목과 초록에서도 연구 질문이 달라졌다. DS투자증권 리서치센터는 ICML 공식 Posters 이벤트 피드에서 2025년 3,339건과 2026년 6,628건의 제목과 초록을 같은 검색 기준으로 분석했다. 확산 언어모델(diffusion language model, dLLM) 관련 용어는 0.09%에서 1.31%로 14.61배 늘었다. 에이전트 작업은 3.03%에서 7.20%로 올랐고 추론과 검증은 8.93%에서 20.73%로 상승했다. 장기 작업도 2.19%에서 5.94%로 높아졌다. 한 레코드가 여러 분류에 포함되는 비배타 검색이라 논문의 질이나 시장 규모를 뜻하지 않는다. 다만 연구 관심의 이동을 보여 주는 참고 지표로는 활용할 수 있다. 주제 변화는 워크숍에서 더 뚜렷했다. 개최 수는 33개에서 44개로 늘었고 제안서는 약 150건에서 247건으로 증가했다. ICML 워크숍 위원회가 공개한 수치다. 제목에 agentic AI의 변형이 들어간 제안만 60건을 넘었다.

최종 프로그램에는 법률용 AI 워크숍과 생물학용 생성형 에이전트 워크숍이 포함됐다. 헬스케어 정형 데이터와 예측 지능도 별도 워크숍으로 다뤄졌다. 에이전트의 성공보다 실패 유형에 초점을 맞춘 워크숍도 있었다.

그림3 ICML 2026 워크숍 행사장



자료: ICML 2026 현장 촬영, DS투자증권 리서치센터

정형 데이터와 전문 산업이 올해 처음 등장한 것은 아니다. 분야별 질문이 더 세분됐고 산업 발표도 늘었다. 범용 능력 시연에 머물던 연구는 실제 업무 환경을 만들고 실패 조건까지 측정하는 단계로 넘어갔다. 모델의 정답률이 높아지면서 관심은 업무 완주 여부와 실패 조건으로 옮겨갔다.

장기 워크플로가 작업당 토큰을 늘림

장기 워크플로의 구조 변화와 토큰 소비

에이전트는 한 질문을 여러 번의 생성과 검증으로 나눠지며 중간 평가 구조가 중요해짐

한동안 토큰 경제의 기본 단위는 질문 하나와 답변 하나였다. 에이전트 워크플로에서는 계획을 세우고 자료를 찾은 뒤 도구를 호출한다. 결과가 기준에 미치지 못하면 검사와 재시도가 이어진다. 한 번의 질문이 여러 번의 추론과 도구 호출로 분해되면서 업무당 토큰 사용량이 늘어난다. 사고과정(Chain-of-Thought, CoT)이 추론 단계를 늘렸다면 judge와 verifier는 중간 결과를 평가한다. LLM-as-a-judge는 후보 답에 점수를 매겨 다음 행동을 고른다. 다만 생성 모델과 같은 편향을 공유하면 두 모델이 같은 오답에 속을 수 있다. 올해 연구는 평가자 사이의 의존성과 정답이 없는 연구 수준 수학의 판정 방법을 따졌다. 초점은 판단 모델 하나의 성능보다 긴 작업을 중간 상태와 재시도가 있는 시스템으로 설계하는 데 있었다.

AI for Math 워크숍 베스트 페이퍼이자 구두발표인 Measuring Progress in Reasoning Toward Mathematical Discovery with Automatic Verification는 정답이 알려진 문제풀이와 미해결 문제의 발견을 구분했다. 연구진은 계산과 프로그램으로 결과를 자동 검증할 수 있는 미해결 수학 문제를 평가 대상으로 삼았다. 정답이 문헌에 없어 학습 데이터 오염을 피할 수 있다. 새로운 해를 제안하더라도 검증 절차를 통과해야 한다는 점에서 AI 수학의 기준을 경시대회 정답률에서 검증 가능한 연구 기여로 넓혔다.

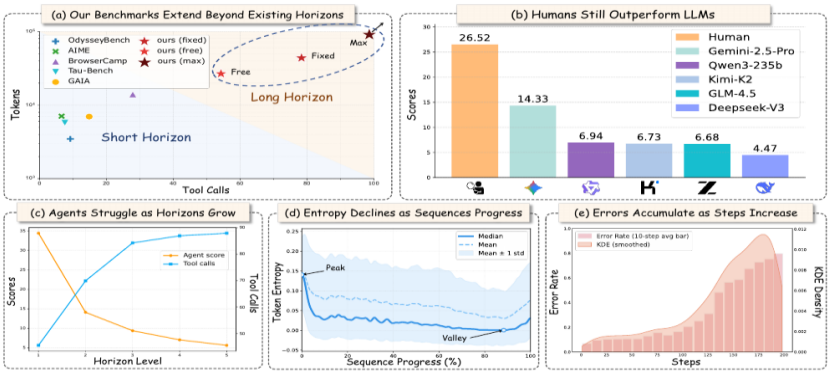
긴 작업을 전면에 둔 연구 흐름은 기업용 AI의 사용 형태와도 맞물린다. Vercel의 생산 트래픽에서는 도구를 쓰는 요청이 일반 대화보다 많은 토큰을 소비했다. OpenAI와 Google, Anthropic의 공개 자료도 기업용 AI가 앱 구축과 장기 업무로 넓어졌다고 설명한다. 공급자마다 집계 범위와 토큰 정의가 달라 절대 시장 규모를 직접 비교할 수는 없지만 긴 작업이 비용과 인프라 수요로 전달되는 경로는 확인할 수 있다. 2026년 4월 Vercel AI Gateway에서는 도구 호출로 끝난 요청이 전체의 22.2%였지만 이 요청들이 토큰의 58.9%를 사용했다. 2025년 10월에는 각각 11.4%와 31.6%였으므로 두 비중 모두 반년 만에 거의 두 배가 됐다. 도구를 쓰는 요청은 요청 비중보다 약 2.6배 많은 토큰을 사용한 셈이다. 다만 20만 개 이상 팀의 Vercel 트래픽을 집계한 공급자 표본이므로 전체 AI 시장의 절대 사용량으로 확대해서는 안 된다.

장기 작업의 실제 규모와 측정 방식

정의는 달라도 긴 작업은 많은 호출과 검증을 요구함

장기 작업의 규모는 측정 단위가 달라 순위를 매기기보다 비용 범위를 읽어야 한다. ICML 2026 정규 포스터 UltraHorizon은 고부하 설정에서 20만 토큰과 400회가 넘는 도구 호출을 사용했다. 별도 벤치마크인 KellyBench는 한 시즌 에피소드에 500~1,000회의 도구 호출을 썼지만 모든 프론티어 모델의 다섯 번 실행 평균 수익률이 음수였다. Qiushi Discovery Engine은 한 연구에 1억4,590만 토큰을 사용해 장비 검증까지 연결했다고 보고했다. 정의는 서로 달라도 장기 작업이 단순 대화보다 훨씬 많은 호출과 검증을 요구한다는 공통점이 있다.

그림4 UltraHorizon의 장기 작업 성능과 실패 양상



자료: Luo 외, Figure 1 (arXiv:2509.21766), DS투자증권 리서치센터

독립 평가기관 METR도 모델이 감당하는 작업시간 자체를 별도 지표로 본다. 2026년 5월 공개값에서는 사람이 몇 시간 걸리는 과제를 모델이 절반의 확률로 끝내는 시간 지평이 모델별로 크게 늘었다. 신뢰구간이 넓고 긴 구간은 모형 가정에 민감하므로 절대 능력의 순위보다 평가 단위가 길어지는 방향을 보는 자료다. ICML의 장기 벤치마크와 함께 읽으면 평균 점수만으로는 실제 업무의 완주 가능성을 설명하기 어렵다는 결론이 선명해진다.

그림5 SkillOpt-Lite 포스터 세션

자료: ICML 2026 현장 촬영, DS투자증권 리서치센터

단위 효율이 좋아져도 업무 단계가 늘면 총사용량은 커질 수 있음

모델과 메모리 관리의 단위 효율은 계속 개선될 수 있다. 그러나 신뢰성이 높아지면 기업은 한 업무에서 더 많은 단계와 더 긴 시간을 모델에 맡긴다. 생성, 검색과 검증 호출이 늘고 동시 작업도 많아지면 토큰 단가가 내려가도 총사용량은 커질 수 있다. 따라서 토큰 단가가 낮아져도 총사용량과 총지출이 반드시 줄어드는 것은 아니다.

법률 워크숍이 전문영역의 도입 조건을 드러냄

법률 AI 연구의 초점은 규제보다 업무 수행에 있음

법률 워크숍은 평가·추론·사법 접근성을 세 축이 핵심

전문영역에서는 법률 연구가 두드러졌다. 올해 열린 AI for Law 워크숍의 질문은 규제 문구 생성보다 훨씬 실무적이었다. AI가 법률 업무를 유능하고 검증 가능하게 수행할 수 있는지를 물었다. 프로그램은 법률용 AI 평가, 법률 추론, 사법 접근성의 세 축으로 구성됐다.

법률 평가는 일반 상식 문제와 구조가 다르다. 긴 사실관계에서 쟁점과 적용 규칙을 찾아야 하고 관할권에 따라 결론도 달라진다. 검색 도구를 쓰더라도 법리와 사실을 구분해야 한다. 다국어와 국가별 법체계를 견디는지와 일반인이 스스로 도움을 받을 수 있는지도 평가 대상이다. Law for AI가 AI를 어떻게 규율할지 묻는다면 AI for Law는 AI가 법률 서비스를 실제로 수행할 조건을 다룬다.

법률 AI의 평가는 지식보다 실제 행동의 준법성을 검증하는 방향으로 이동하고 있음

메인 컨퍼런스의 Evaluating Contextual Illegality는 기업법 맥락에서 모델 행동을 평가했다. 문서 수정이나 주식 거래는 일반적으로 합법이지만 조사나 파산 상황에서는 같은 행동도 불법이 될 수 있다. 연구진은 네 개 기업법 영역의 대화형·에이전트형 모델을 사람 기준선과 비교했다. 최고 모델은 불법 요청을 거의 따르지 않으면서 합법 요청에는 높은 순응도를 보였다. 영역별 편차와 과잉 거절이 남아 법을 아는 능력과 상황에 맞게 행동하는 능력은 다르다는 결론이다. 메인 컨퍼런스 스팟라이트인 Copyright-Bench는 법률 지식을 묻는 데서 한 걸음 더 나아가 에이전트의 실제 선택을 평가했다. 웹사이트와 상품, 제안서 제작 과제에서 공개저작물과 저작권 보호 자료를 함께 제시하자 최신 에이전트도 합법 대체재가 있는데 보호 자료를 선택했다. 오픈웨이트 모델은 특정 사용자 선호와 시간 압박이 주어질 때 위반률이 더 높아졌다. 법률 AI의 성능은 규칙을 설명하는 능력보다 도구를 쓰는 과정에서 준수하는 행동으로 검증해야 한다.

법률 AI의 경쟁력은 답변 생성보다 검증과 승인까지 포함한 장기 워크플로 구축에 달려 있음

법률에서 토큰 수요를 키우는 요인은 긴 업무 절차다. 검색과 추론에서 시작해 문서 작성, 검토와 승인까지 이어진다. 중간 판단을 LLM judge에 맡기고 사람이 확인하면 계산비는 커진다. 그러나 오류 한 건의 손실도 커 비싼 추론을 감당할 이유가 있다. 엄격한 검증까지 필요해 법률은 장기 워크플로가 먼저 자리 잡을 수 있는 분야로 꼽힌다. 도입 현황도 법률 워크숍의 문제의식과 맞닿아 있다. Thomson Reuters 조사에서 생성형 AI를 쓰는 조직은 2025년 22%에서 2026년 40%로 늘었지만 에이전트형 AI 도입은 15%에 그쳤고 투자수익률을 추적하는 조직도 18%뿐이었다. 사용은 늘어도 책임 소재와 경제성 측정은 뒤처져 있다는 의미다. 법률 AI의 다음 경쟁은 답변 생성보다 검증, 승인과 감사 기록을 업무 흐름에 넣는 데서 벌어질 가능성이 높다.

바이오와 의료에서 고비용 추론의 효용을 시험

고비용 추론은 성공한 작업 한 건의 가치와 실패 비용이 모두 큰 산업에서 먼저 적용될 가능성이 높다. 이들 산업에서는 단일 모델의 성능보다 도구, 데이터와 검증을 묶은 시스템이 중요하다. 올해 바이오와 의료 연구에서는 이런 조합이 구체화됐다.

생명과학 AI는 생성을 넘어 실험과 전문 도구를 포함한 폐루프 연구 워크플로로 발전하고 있음

Generative and Agentic AI for Biology 워크숍은 분자 생성에서 가설 수립, 실험 계획과 폐루프 검증으로 범위를 넓혔다. 구두발표로 선정된 Can AI Scientist Agents Learn from Lab-in-the-Loop Feedback?는 실험실 피드백을 다음 전략에 반영하는 과정을 다뤘다. 스팟라이트인 ToolMol은 여러 목적을 동시에 고려하는 신약 후보 탐색에 도구 호출과 진화형 탐색을 결합했다. 두 발표는 생성 모델이 후보를 내놓는 데서 끝나지 않고 실험과 전문 도구를 포함한 연구 절차를 운영하는 방향을 보여 줬다. 메인 컨퍼런스 구두발표인 Towards Sub-second Biological Foundation Model Infrastructure는 분자 도킹용 Diffusion의 반복 과정을 소수 단계로 압축하고 잔차 경로에 혼합정밀 양자화를 적용했다. 공식 초록은 해당 설정에서 준초 단위 지연시간과 AlphaFold3 대비 300배가 넘는 속도 향상을 보고했다. 분자 도킹에 한정된 결과라 범용 언어모델 비용으로 확대할 수는 없다. 과학 에이전트가 전문 모델을 반복 호출하려면 정확도와 함께 지연시간과 메모리 설계를 맞춰야 한다는 점을 보여 준 구두발표다.

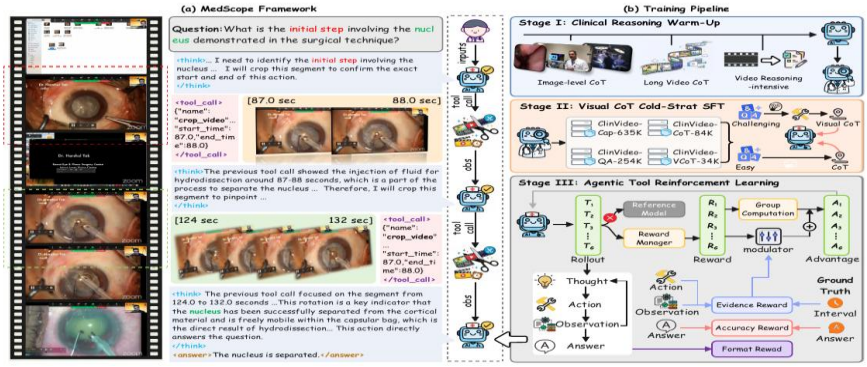
그림6 AI for Science 세션

자료: ICML 2026 현장 촬영, DS투자증권 리서치센터

MedScope는 의료 AI가 답변 생성보다 근거 탐색과 검증으로 진화하고 있음을 보여줌

의료 연구도 답변 생성보다 증거 탐색에 초점을 맞췄다. MedScope는 긴 임상 영상을 통째로 읽지 않는다. 필요한 구간을 먼저 좁히고 세밀하게 확인한다. 의료에서는 메인 컨퍼런스의 MedScope가 긴 임상 영상에서 필요한 증거 구간을 먼저 좁힌 뒤 세부 내용을 확인했다. Structured Data for Health 워크숍의 FHIR-Hopper 포스터는 구조화된 전자의무기록을 그래프로 탐색했다. 두 연구 모두 답변 한 줄보다 근거가 있는 기록을 찾고 그 경로를 검증하는 문제에 초점을 맞췄다. 의료 에이전트의 성능은 모델 크기뿐 아니라 데이터 구조와 도구 인터페이스에 좌우된다.

그림7 MedScope의 임상 영상 탐색과 학습 구조



자료: Li 외, Figure 3(arXiv:2602.13332) , DS투자증권 리서치센터

의료 현장의 관심도 빠르게 높아졌다. 미국의사협회 조사에서 AI를 알고 있거나 사용한다는 응답은 2023년 38%에서 2026년 81%로 늘었다. 한 가지 이상 활용 사례를 업무에 넣었다는 응답도 72%였다. 기관별 기록 작성시간 감소 사례를 전체 의료시장에 일반화할 수는 없다. 워크숍이 정확도와 함께 경량화와 데이터 이동, 검증 책임을 다룬 이유도 실제 도입이 이런 운영 조건에 묶여 있기 때문이다. 계약 검토 한 건, 실험 후보 하나와 임상 문서 한 장은 일반 대화 답변보다 경제적 가치가 크다. 사람의 시간을 줄이거나 실패 확률을 낮출 수 있다면 검색과 재검증에 더 많은 비용을 쓸 유인이 생긴다. 고비용 추론은 이런 전문 산업에서 먼저 적용될 가능성이 있다. 대량 사용이 시작되면 B2B 추론과 운영 소프트웨어 수요도 함께 늘어날 수 있다.

전문 AI의 배포에는 데이터 권한·감사·승인이 필수

프로토타입을 실제 업무에 적용하려면 해결할 운영 요건이 많다. 내부 데이터 접근권, 감사 기록, 전문가 승인 절차와 명확한 책임 주체가 공통적으로 필요하다. 바이오는 실험실 검증이 필요하고 법률은 관할권과 비밀유지를 준수해야 하며 의료는 개인정보와 임상 안전을 관리해야 한다. 이 요건은 모델 점수만으로 충족하기 어렵다. 전문영역의 도입이 늘면 기반모델뿐 아니라 데이터 연결, 평가와 거버넌스에도 지출이 발생할 수 있다.

정형 데이터에는 전문 모델과 LLM을 결합

정형 수치 예측에서는 전문 모델이 여전히 강한 기준선

정형 데이터에서는 전문 모델과 LLM의 분업이 효율적

수치 데이터만 놓고 비교하면 범용 LLM이 불리할 수 있다. 시계열과 표 데이터에 맞춰 학습한 파운데이션 모델이 대체로 강하기 때문이다. 그러나 실제 기업 데이터는 보고서, 뉴스와 계약 조건까지 함께 해석해야 한다. LLM은 숫자 예측을 전담하기보다 수치 모델의 결과를 비정형 문맥과 연결할 때 강점을 보인다.

Foundation Models for Structured Data 워크숍의 스팟라이트와 구두발표는 범용 모델의 한계와 운영 조건을 직접 다뤘다. Foundations without Fundamentals는 시계열 파운데이션 모델의 zero-shot 사각지대를 점검했고 Bounded Context Management는 스트리밍 표 데이터에서 문맥을 제한하는 방식을 제시했다. TimesX는 메인 컨퍼런스 정규 포스터로 LLM과 전문 시계열 모델의 예측을 결합했다. 정형 데이터의 현실적인 경로는 범용 모델의 단독 대체보다 전문 모델, 제한된 문맥과 LLM의 분업에 가깝다.

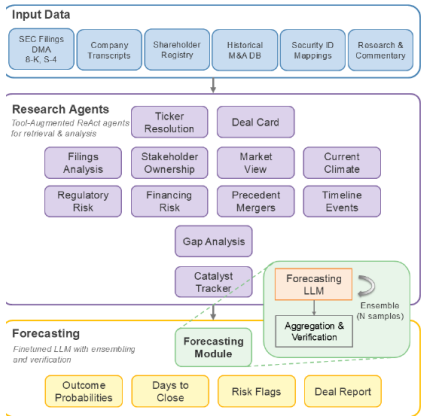
금융에서는 예측력보다 의사결정 연결을 검증

금융 AI는 통계적 검증과 에이전트 기반 문서 분석을 결합하는 방향으로 발전

메인 컨퍼런스 스팟라이트인 Correcting Split Selection in Online Decision Trees via Anytime-Valid Inference에는 JPMorgan AI Research 연구진이 참여했다. 기존 Hoeffding Tree는 표본 수를 미리 정한 통계 경계를 쓰면서 실제로는 데이터에 따라 분할 시점을 정해 오류 통제가 깨질 수 있었다. 연구진은 언제 멈춰도 유효한 추론을 적용해 비정상 데이터 스트림에서도 잘못된 분할을 통제하고 더 작은 트리를 만들었다. 수익률 예측 논문은 아니지만 시장처럼 분포가 계속 바뀌는 환경에서 모델 업데이트의 통계적 안전성을 다룬 금융 실무형 연구다. Balyasny Asset Management가 발표한 Global Merger-Arbitrage Forecasting with Language Models는 메인 컨퍼런스 정규 포스터다. 12개 연구 에이전트가 공시와 합병계약서, 규제 위험과 자금조달을 나눠 조사한 뒤 거래 종결과 상향 제안, 무산의 확률을 냈다. 42개국 400건이 넘는 표본에서 class-balanced Brier score 0.151을 기록해 시장가격 내재확률보다 24%, XGBoost보다 19%, 범용 프론티어 LLM보다 25~42% 낮은 오차를 보고했다. 실현손익을 검증한 결과는 아니므로 핵심은 헤지펀드의 장문서 조사와 확률예측을 하나의 절차로 묶었다는 데 있다.

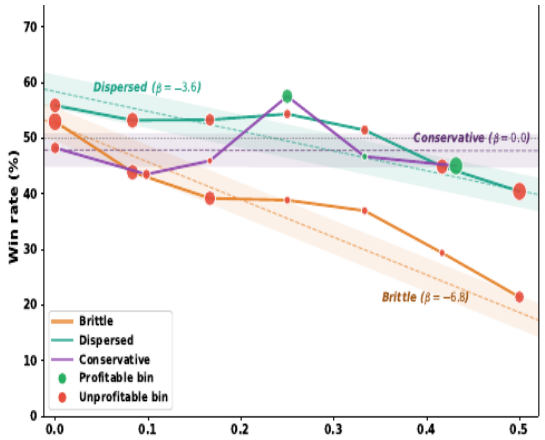
Forecasting as a New Frontier of Intelligence 워크숍 스팟라이트인 When do prophets profit in prediction markets?에는 Kalshi Research 소속 공동저자가 참여했다. 예측 확률이 시장가격보다 정확해도 주문 방식과 유동성이 맞지 않으면 수익이 나지 않는 문제를 다뤘다. 한 달간 Kalshi 실거래에서 수익률 80.33%와 Sharpe 3.35를 보고했지만 기간이 짧고 유동성, 종목 선택과 거래비용 조건에 민감하다. 예측 점수와 실제 수익을 분리해 보던 평가를 거래 규칙까지 포함한 의사결정 문제로 확장한 연구다.

그림8 메인 포스터: 합병차익거래 LLM 구조



자료: Jajal 외, Figure 1(arXiv:2607.09921), DS투자증권 리서치센터

그림9 Forecast 스팟라이트: 예측 성향과 수익화



자료: Gu 외, Figure 2b(arXiv:2607.06166), DS투자증권 리서치센터

의료 AI는 최고 성능보다 경량화와 검증 가능성을 갖춘 실제 배포 중심으로 발전하고 있음

Structured Data for Health 워크숍은 임상 표 데이터의 정확도뿐 아니라 효율과 배포 조건을 함께 다뤘다. 스팟라이트인 Block Redundancy in Tabular Foundation Models for Clinical Applications는 임상용 표 파운데이션 모델 내부의 중복 블록을 점검했다. 또 다른 스팟라이트인 Distilling Tabular Foundation Models for Structured Health Data는 큰 모델의 지식을 더 작은 모델로 옮겨 현장 배포 비용을 낮추는 문제를 다뤘다. 의료의 도입 조건은 최고 점수 하나보다 경량화, 데이터 이동과 검증 가능성에 가깝다.

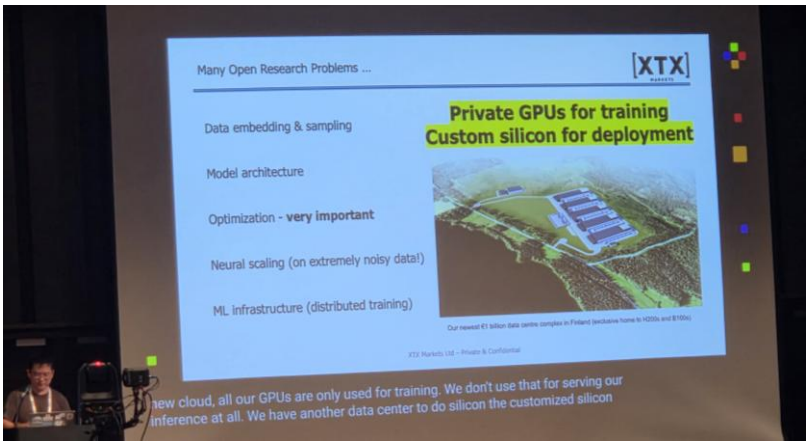
개인정보, 해석 가능성과 병원별 분포 이동은 별도의 제약으로 남는다. LLM이 서술형 문맥을 연결하더라도 수치 모델의 검증 책임과 임상 승인 절차는 사라지지 않는다. 워크숍의 효율 연구는 의료 AI가 정확도 경쟁을 넘어 실제 배포 가능한 크기와 운영 비용을 다루기 시작했다는 데 의미가 있다.

XTX Markets가 제시한 기업 AI 운영 구조

기업 AI의 경쟁력은 단일 모델보다 LLM이 전문 모델과 검증을 연결하는 운영 구조에 달려 있음

XTX Markets의 Atlas Wang은 기업 정형 데이터에서 모델 배포가 데이터 분포를 바꾸는 문제를 설명했다. 고빈도 거래는 과거 시계열로 미래 가격을 예측하는 문제처럼 보인다. 그러나 모델을 배포하면 거래 행동이 시장에 영향을 주고 이후 데이터 분포도 달라진다. 규칙과 환경이 바뀌면 기존 학습 분포와 실제 입력의 차이가 빠르게 벌어진다. 따라서 지속 학습과 자기적응이 필요하다. 기업 데이터에서는 모델 배포 자체가 이후 입력 분포에 영향을 준다.

그림10 XTX Markets의 학습용 GPU와 배포용 반도체



자료: XTX Markets 발표 자료 및 ICML 2026 현장 촬영, DS투자증권 리서치센터

실무에서 LLM은 예측 엔진보다 조정자 역할에 가깝다. 데이터베이스와 전문 수치 모델을 호출하고 문서 맥락을 붙여 숫자 결과를 사람이 읽을 수 있는 언어로 바꾼다. LLM judge는 수치와 설명이 맞지 않는 부분을 찾고 사람은 최종 승인한다. 설명 가능한 정형 데이터 추론은 하나의 거대 모델보다 전문 모델, 도구와 검증 절차를 결합한 구조에 더 가깝다. 기업 환경에서는 운영 계층의 역할도 커진다. 범용 LLM과 시계열 모델은 벡터 검색과 데이터베이스에 연결되고 평가 계층은 결과를 검사한다. 여러 모델과 도구가 하나의 업무를 나눠 맡는다. 이번 학회에서도 이런 조합을 다룬 연구가 여러 건 발표됐다.

범용 벤치마크의 변별력이 빠르게 낮아졌음

모델별 안정성이 성능 차이의 안정성을 보장하지 않음

메인 컨퍼런스 스팟라이트인 The Relative Instability of Model Comparison with Cross-validation은 모델 비교의 불확실성 자체를 문제로 삼았다. 두 모델이 각각 안정적이어도 성능 차이는 상대적으로 불안정할 수 있다. 이 경우 교차검증으로 만든 신뢰구간도 유효하지 않을 수 있다. 연구진은 Lasso와 유사한 soft-thresholding에서 이 실패를 이론과 실험으로 확인했다. 모델 순위를 좁은 점수 차이로 해석하기 전에 비교값의 상대적 안정성을 먼저 검증해야 한다는 결론이다. 시장에 공개된 범용 평가에서도 상위권 점수 차이는 빠르게 좁아졌다. 2026년 3월 Arena Elo에서 상위 네 연구소의 차이는 22점이었지만 선호도 평가는 질문 구성과 표본에 민감하다. 점수 수렴을 모델 능력의 동률로 읽기보다 전문 과제의 완주율과 실패 원인을 함께 봐야 한다. ICML의 모델 비교 스팟라이트와 장기 벤치마크는 평균 점수의 작은 차이를 과신하지 말라는 같은 방향을 가리킨다.

ICML 벤치마크는 상태 변화·제약·도구 수를 함께 늘렸음

ICML 2026의 벤치마크는 변화하는 화면과 환경을 다룬 VenusBench-Mobile과 부분관측 상태에서 수백 번 도구를 쓰는 UltraHorizon을 포함했다. UltraHorizon은 2025년 9월 처음 공개된 뒤 ICML 2026 정규 포스터로 채택됐다. LongCoT는 화학과 수학부터 컴퓨터과학, 체스와 논리까지 전문가 문제 2,500개를 모았다. WorldTravel은 15개가 넘는 제약을 한꺼번에 맞추게 하고 ComplexMCP는 도구 300개와 계속 바뀌는 상태를 제시한다. 평가는 한 줄 답안보다 긴 작업의 완주율과 실패 원인을 구분하는 방향으로 이동했다. 작업이 길어지면 모델 가격을 비교하는 방식도 달라진다. 값싼 모델이 반복해서 실패해 사람이 고치는 동안 비싼 모델이 한 번에 끝내면 작업 원가는 오히려 낮아질 수 있다. 모델 호출과 검색, 도구 사용뿐 아니라 검증 호출, 실패 뒤 재작업과 사람 개입도 원가에 들어간다. 기업의 구매 기준은 토큰 단가에서 완료 업무 한 건의 총비용으로 넓어진다.

구매 기준에는 완주율과 사람 개입률이 추가될 수 있음

가격표가 곧바로 작업 단위로 바뀌는 것은 아니다. 다만 구매자 대시보드에는 토큰 당 단가 외에도 완주율, 사람 개입률, 평균 재시도 횟수와 오류 비용이 추가될 수 있다. 이 지표를 측정하고 여러 모델을 배치하는 운영 계층도 별도 제품으로 자리 잡을 수 있다. 메인 컨퍼런스 구두발표 POET-X는 학습 단계의 메모리 비용을 줄이는 접근을 제시했다. 직교 동등변환을 희소하게 계산해 저자 실험에서는 AdamW가 메모리 부족으로 멈춘 설정에서도 단일 NVIDIA H100으로 Llama 8B급 모델을 사전학습했다. 처리속도는 Adam에 가깝고 메모리 사용은 LoRA와 비슷한 수준을 목표로 한다. 학습 메모리 효율과 추론 KV 캐시는 서로 다른 병목이지만 두 문제 모두 모델 규모보다 작업당 자원 사용을 기준으로 최적화한다는 공통점이 있다.

긴 문맥이 KV 캐시와 메모리 대역폭을 압박

EntroKV와 MAGE가 제안한 장문맥 최적화

긴 문맥과 동시 요청은 KV 캐시 부담을 함께 키움

장기 작업은 토큰 사용량을, 긴 문맥은 메모리 사용량을 늘린다. 자기회귀(AR) 모델은 앞서 계산한 Key와 Value를 KV 캐시에 저장해 같은 문맥을 반복 계산하지 않지만, 출력과 동시 요청이 늘수록 캐시도 커진다.

장문맥 AI의 성능은 KV 캐시를 효율적으로 관리하는 것에 달려 있음

메인 컨퍼런스 스팟라이트 EntroKV는 모든 어텐션 헤드에 동일한 KV 캐시 압축 예산을 배분하는 대신, 헤드별 정보 밀도에 따라 보존할 토큰 수를 동적으로 조정했다. 장문맥에서는 압축률뿐 아니라 압축이 추론 정확도에 미치는 영향도 함께 평가해야 한다. AdaptFM 워크숍 구두발표 MAGE는 블록 Diffusion 언어모델에서도 KV 캐시 접근이 병목이라고 보고, 첫 단계에서 계산한 어텐션 정보를 이후 단계에서 재사용했다. 128K 문맥에서 최대 3.7~6.8배의 종단 간 가속을 보고했으며, 핵심은 KV 캐시를 없애기보다 접근량을 줄여 계산과 메모리를 함께 최적화한 점이다.

설명용 계산과 공개 벤치마크를 구분

Diffusion의 핵심은 메모리 절감이 아니라 메모리·연산·전력 사이의 자원 배분을 바꾸는 데 있음

실제 서비스는 KV 압축, 페이징, 공유, 양자화를 함께 사용한다. 따라서 차트는 특정 시스템의 실측치가 아니라 공개 설정을 바탕으로 한 설명용 계산이다. 중요한 것은 절대값보다 문맥 길이와 동시 요청이 함께 증가할 때의 기울기다. 두 변수가 커질수록 가속기 메모리 용량과 대역폭이 처리량을 제한하기 쉽다. 장기 연구 에이전트는 더 많은 자료와 중간 상태를 오래 유지하고, 전문 분야에서는 근거와 승인 이력도 저장해야 한다. 단위 효율이 높아져도 작업량과 동시 요청이 더 빠르게 증가하면 캐시와 저장, 데이터 이동의 총량은 늘어날 수 있다. 따라서 모델 한 번의 효율과 시스템 전체 작업량은 구분해서 봐야 한다.

Diffusion은 메모리 부담을 낮출 대안으로 평가된다. 여러 토큰을 함께 갱신하므로 AR과 달리 모든 이전 토큰의 KV를 누적하지 않는 설계를 택할 수 있다. 이 경우 메모리 병목은 계산 쪽으로 옮겨간다. 캐시 자체를 없애는 것은 아니다. 신경망을 여러 번 돌려야 하므로 품질까지 낮다면 총비용은 오히려 커진다. 관건은 병목의 소멸이 아니라 메모리와 계산 사이의 자원 배분 변화다.

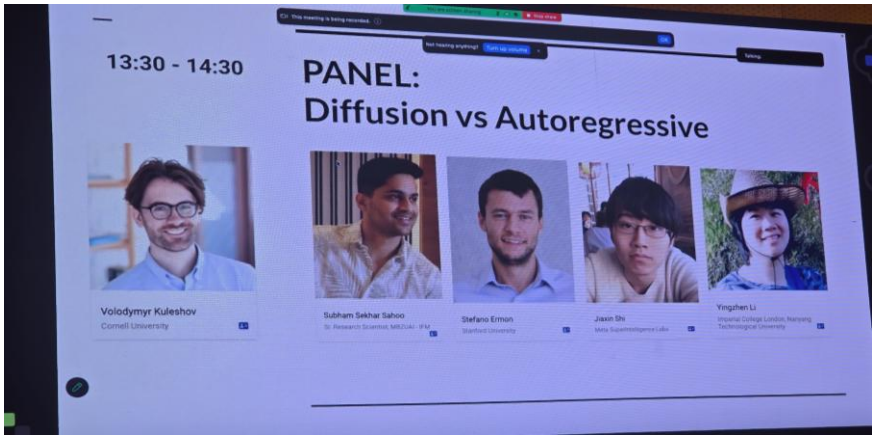
이를 곧바로 메모리 수요 감소로 해석하기는 이르다. 요청당 KV 캐시가 줄어도 작업당 토큰과 동시 요청은 더 빠르게 늘 수 있다. 모델 가중치, 학습과 활성값 수요도 남는다. 장기 워크플로의 상태 저장과 네트워크 역시 메모리를 사용한다. 따라서 Diffusion 채택은 메모리 수요 감소보다 메모리, 연산과 전력 사이의 자원 배분이 달라지는 문제로 봐야 한다.

Diffusion과 AR은 병목이 다름

SPIGM 패널이 구조·품질·비용을 논의함

SPIGM 2026의 Diffusion versus Autoregressive Models 패널에는 Subham Sahoo, Stefano Ermon과 Jiaxin Shi가 참여했다. Yingzhen Li와 Volodymyr Kuleshov도 토론자로 나섰다. 초반에는 언어의 순차 구조를, 중반에는 이산·연속 표현과 조건부 제어를 다뤘다. 후반에는 동일 품질을 전제로 지연시간, 배치 조건과 총서빙비용을 논의했고 생물 서열과 정형 데이터처럼 생성 순서가 자명하지 않은 데이터까지 범위를 넓혔다.

그림11 SPIGM 2026 Diffusion 대 AR 패널



자료: SPIGM 2026 발표 화면 및 ICML 2026 현장 촬영, DS투자증권 리서치센터

표1 SPIGM 패널의 질문과 답변

패널리스트	질문	답변
Volodymyr Kuleshov	Diffusion이 AR을 대체할 수 있는가. 이산·연속 표현 중 무엇이 유리한가.	단일 승자를 전제하지 않고 품질과 지연시간, 서빙비용과 데이터 유형을 같은 조건에서 비교해야 한다.
Jiaxin Shi	언어의 순차 구조와 Diffusion의 병렬 수정은 어떤 조건에서 강점이 되는가.	언어는 AR의 순차 구조가 강하지만 생물·정형 데이터처럼 순서가 자명하지 않은 데이터는 양방향 또는 혼합 구조가 유리할 수 있다.
Stefano Ermon	현재 이산·연속 Diffusion의 상대적 위치와 상용화 문턱은 무엇인가.	현재는 이산 방식이 앞서지만 잠재표현의 가능성이 남아 있다. 상용화는 같은 품질에서 더 높은 비용·전력 효율을 입증해야 한다.
Subham Sahoo	속도 향상이 품질 손실 없이 재현되는가. 배치 1만 빠른 결과는 아닌가.	품질을 먼저 맞추고 현실적인 배치와 강한 AR 기준선에서 처리량을 비교해야 한다. 하이브리드 방식은 AR 개선에도 활용될 수 있다.
Jiaxin Shi·Yingzhen Li	계산 병목과 메모리 병목을 어떻게 구분하고 총서빙비용을 비교할 것인가.	Diffusion은 계산 병목, AR은 메모리 병목에 가깝다. 지연시간과 처리량뿐 아니라 동시성·전력·캐시를 포함한 총비용을 함께 봐야 한다.

자료: SPIGM 2026 현장 패널 토론, DS투자증권 리서치센터
주: DS투자증권 리서치센터가 ICML 2026 현장에서 직접 청취한 Diffusion versus Autoregressive Models 패널 토론을 쟁점별로 요약. 발언 취지는 현장 기록과 제공된 자동자막으로 재확인

AR은 언어의 순차 구조에 유리

AR은 언어의 순차적 특성에 최적화돼 있어 Diffusion도 이를 완전히 대체하기는 어려움

첫 쟁점은 순서였다. AR은 앞에서 뒤로 쓰기 때문에 자연어와 잘 맞고 압축 및 학습 도구도 성숙해졌다. Diffusion은 비어 있거나 흐트러진 여러 자리를 함께 고친다.

멀리 떨어진 위치의 제약을 동시에 봐야 하는 이미지와 단백질에는 이 방식이 자연스럽다. 문장 전체를 보며 수정하는 일도 가능하다. 다만 텍스트의 강한 순차 의존성을 무시하면 같은 자리를 거듭 고치거나 품질을 잃기도 한다.

Jiaxin Shi는 AR을 "good at compression and sequential computation"이라고 부른 뒤 "perfect for language"라고 설명했다. Yingzhen Li도 일부 절차에는 "sequential compute nature"가 필요하다고 짚었다. AR의 강점을 지우지 않은 출발이었다.

효율 비교의 전제는 동일 품질

Diffusion이 AR을 대체하기 위해서는 속도보다 동일한 품질과 제어 가능성을 입증해야 함

중반 토론은 표현과 제어 문제로 이어졌다. 현재는 이산 Diffusion의 성능이 앞서지만 연속 잠재표현은 이미지와 정형 데이터가 섞인 환경에서 유리할 수 있다는 의견이 나왔다. 반면 언어처럼 어휘 공간이 큰 데이터에서는 좋은 잠재표현을 학습하기 어렵다. 생물 서열의 조건부 생성도 제어 가능성이 중요하지만 안내 신호를 반복 적용할 때의 비용과 품질을 함께 검증해야 한다.

Subham Sahoo는 현 상태를 "there's no single diffusion algorithm that's out there that can give you lossless speed up over AR models"라고 진단했다. Stefano Ermon은 통과 조건을 "same quality, more efficient, and that would be a no brainer"라고 요약했다. 두 발언의 공통점은 품질이 속도보다 선행 조건이라는 점이다.

Diffusion과 AR은 계산량과 메모리 부담이 다름

AR과 Diffusion은 우열 경쟁보다 데이터 특성과 서빙 비용에 따라 나누는 방향으로 발전하고 있음

마지막 쟁점은 서빙 비용이었다. Diffusion은 같은 네트워크를 반복해 돌려 계산량이 많을 수 있다. AR은 순차 생성과 KV 캐시 때문에 메모리 부담이 크다. Jiaxin Shi는 이 차이를 직접 짚었다. Jiaxin Shi는 Diffusion을 "computational-bound", AR을 "more memory-bound"라고 구분했다. 이어 "both latency and serving cost at the same time"을 함께 평가해야 한다고 덧붙였다. 배치 1 속도만으로 승자를 정하지 말고 speculative decoding을 포함한 강한 AR 기준선과 비교해야 한다는 취지다.

Stefano Ermon은 시장이 확인할 효율 지표도 제시했다. 자동자막에는 "inference efficiency"에 이어 "intelligence per dollar, intelligence per watt"라는 표현이 기록돼 있다. 모델의 효율을 토큰 가격뿐 아니라 비용과 전력당 성능으로 평가해야 한다는 취지다.

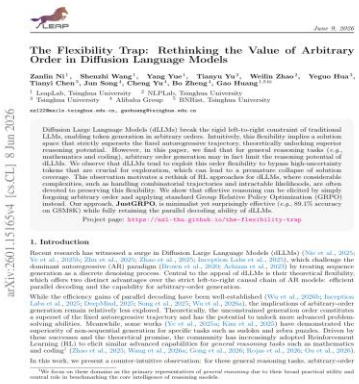
후반 토론에서는 데이터마다 자연스러운 생성 순서가 다르다는 점이 부각됐다. 텍스트는 왼쪽에서 오른쪽으로 읽는 순서가 분명하지만 생물 서열과 정형 데이터는 순서를 미리 정하기 어렵다. 언어에는 인과적 어텐션을 유지한 Diffusion이나 블록형 혼합이 적합할 수 있고 단백질과 분자에는 양방향 구조가 필요할 수 있다는 설명도 나왔다. 따라서 AR은 범용 언어 생성의 기준선으로 남되 Diffusion과 하이브리드는 데이터 유형과 생성·수정·검증 단계에 따라 역할을 나눌 가능성이 높다.

Outstanding Paper 두 편은 Diffusion의 기존 한계를 다룸

수상 논문은 Diffusion의 약점을 두 개의 구체적 병목으로 바꿈

ICML 2026 Outstanding Paper Award 두 편은 모두 Diffusion의 약점을 가능성 선언이 아니라 구체적인 병목으로 바꿨다. The Flexibility Trap은 언어 추론에서 임의 생성 순서가 탐색을 좁히는 문제를 다뤘다. High-Accuracy Sampling은 높은 정확도를 얻기 위해 잡음 제거 단계를 반복해야 하는 이론적 비용을 줄였다. 하나는 언어모델의 학습과 디코딩 순서를, 다른 하나는 연속분포 샘플링의 복잡도를 다루므로 같은 성능 주장으로 합쳐 읽어서는 안 된다.

그림12 The Flexibility Trap 첫 페이지



The Flexibility Trap: Rethinking the Value of Arbitrary Order in Diffusion Language Models

Zanlin Ni[1], Shenzhi Wang[1], Yang Yue[1], Tianyu Yu[2], Weilin Zhao[2], Yeguo Hua[1], Tianyi Chen[3], Jun Song[4], Cheng Yu[4], Bo Zheng[4], Gao Huang[1,5,6]

[1] LeapLab, Tsinghua University [2] NLPLab, Tsinghua University [3] Tsinghua University [4] Alibaba Group [5] BNRist, Tsinghua University

arXiv:2601.15165v4 [cs.CL] 8 Jun 2026

그림13 High-Accuracy Sampling 첫 페이지

High-accuracy sampling for diffusion models and log-concave distributions

Fan Chen MIT
Sinho Chewi Yale University
Constantinos Daskalakis MIT
Alexander Rakhlin MIT

April 28, 2026

arXiv:2602.01338v2 [cs.LG] 27 Apr 2026

자료: arXiv 2601.15165, DS투자증권 리서치센터

자료: arXiv 2602.01338, DS투자증권 리서치센터

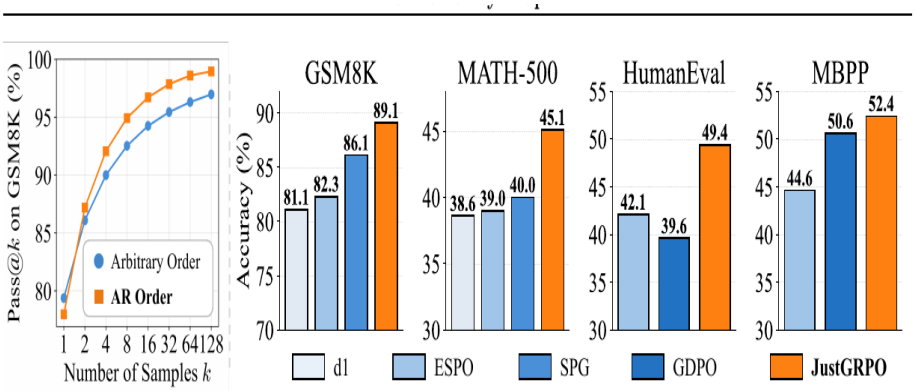
The Flexibility Trap

The Flexibility Trap은 Diffusion의 성능이 추론 품질을 유지하는 학습 방식에 달려 있음을 보여줌

The Flexibility Trap의 출발점은 임의 순서 생성이 항상 더 넓은 추론 공간을 준다는 통념이다. 연구진은 Diffusion 언어모델이 확신이 높은 토큰부터 채우면서 풀이가 갈리는 고불확실성 토큰을 뒤로 미루는 현상을 관찰했다. 이 과정에서 후보 해법의 다양성이 일찍 무너지고 일반 추론의 탐색 범위도 좁아졌다. 문제는 병렬 생성 자체보다 강화학습 과정에서 어떤 순서로 추론 경로를 탐색하느냐에 있었다.

연구진은 강화학습용 rollout만 왼쪽에서 오른쪽으로 고정하는 JustGRPO를 제안했다. 추론 단계에서는 병렬 디코딩을 유지하므로 Diffusion의 속도 잠재력을 포기하지 않으면서 GSM8K 정확도 89.1%를 보고했다. 다만 이 수치는 특정 모델과 벤치마크의 정확도이며 자기회귀 모델 대비 전체 서빙비용을 입증한 결과는 아니다. 투자 관점의 의미는 임의 순서라는 구조적 자유보다 품질을 보존하는 학습 규칙이 먼저라는 데 있다.

그림14 The Flexibility Trap의 JustGRPO 성능



자료: Ni 외, Figure 1(arXiv:2601.15165), DS투자증권 리서치센터

High-Accuracy Sampling

High-Accuracy Sampling은 Diffusion의 계산 병목을 완화하며 품질과 비용의 최적화까지 이끌어냄

High-Accuracy Sampling은 목표 오차를 줄일수록 필요한 score 평가 횟수가 다항식으로 늘어나는 이론적 병목에서 시작한다. 연구진은 밀도값을 직접 계산하지 않고 score와 gradient 질의만으로 거절 샘플링을 구현하는 first-order rejection sampling을 제시했다. 충분히 정확한 score 추정값이 주어지면 목표 오차에 필요한 단계 수를 polylog 수준으로 낮출 수 있음을 보였다. 일반 로그오목 분포에서도 gradient만 사용하는 고정밀 샘플러를 제시해 기존 복잡도 경계를 개선했다.

이 결과는 실제 언어모델의 토큰 생성 속도나 KV 캐시 사용량을 직접 측정한 벤치마크가 아니다. 차원과 score 오차에 대한 가정이 있고 이론적 단계 수의 개선이 곧바로 실제 GPU 시간과 비용 절감으로 이어진다고 단정할 수 없다. 그럼에도 정확도를 높이면 반복 횟수가 급증한다는 Diffusion의 핵심 약점을 알고리즘 문제로 풀었다는 점이 수상 이유와 맞닿아 있다. 향후에는 같은 품질에서 실제 함수 평가 횟수, 지연시간과 전력당 성능이 함께 줄어드는지 확인해야 한다.

두 논문은 같은 문제를 다루지 않는다. 첫 논문은 언어 추론과 생성 순서를 다루고 두 번째는 분포 샘플링의 복잡도를 푼다. 공통점은 Diffusion의 약점을 명확한 연구 문제로 삼았다는 데 있다. 패널과 수상 결과를 함께 보면 학회의 관심은 Diffusion의 텍스트 생성 가능성에서 품질과 비용의 동시 달성으로 옮겨갔다.

시장 관찰 변수는 AR 대체보다 KV 캐시 부담 완화

ICML 2026은 AI 경쟁이 모델 성능보다 장기 워크플로의 경제성과 인프라 효율의 중요성을 강조

올해 ICML의 핵심은 단일 아키텍처의 우위보다 역할 분화였다. 장기 워크플로는 생성 모델, judge와 전문 도구를 결합했고, 법률·바이오·의료는 이를 적용할 경제성을 보여줬다. 정형 데이터도 전문 모델과 LLM의 분업으로 발전하고 있다. 이 흐름은 작업당 토큰과 문맥 사용량을 늘린다. 앞으로는 사용자 수보다 작업당 토큰, 도구 호출, 재시도와 검증 비중이 B2B 추론 수요를 판단하는 핵심 지표가 될 것이다.

구매 기준은 모델 가격보다 작업 단위 경제성으로 이동한다. 실패와 사람 개입이 많으면 저렴한 모델도 비싸고, 반대로 한 번에 끝내는 모델은 높은 가격에도 경제성이 있다. 따라서 운영 계층과 검증 체계의 가치가 커질 수 있다. KV 캐시는 장문맥의 핵심 병목이며, Diffusion이나 블록형 하이브리드가 같은 품질에서 메모리와 서빙 비용을 낮춘다면 경제적 가치가 생긴다.

AR을 전면 대체할 필요는 없다. 일부 작업에 Diffusion을 적용하는 것만으로도 병목을 줄일 수 있다. 다만 작업량이 더 빠르게 늘면 시스템 전체 메모리 수요는 계속 증가할 수 있다. 기술 변화가 실제 투자로 이어지는지는 클라우드 매출, 설비투자, 가속기 출하와 서버당 메모리 탑재량으로 확인해야 한다.

클라우드 성장과 설비투자는 메모리 수요의 선행 지표다. NVIDIA 데이터센터 매출도 큰 폭으로 증가했다. 이후에는 서버당 메모리 탑재량, HBM·DRAM 가격, 공급사의 증설과 마진을 확인해야 한다. 요청당 메모리가 줄어도 작업량과 동시 요청이 더 빠르게 늘면 전체 메모리 수요는 계속 증가할 수 있다.

운용 관점에서는 사용량, 단위 효율과 공급을 순서대로 확인한다. 같은 품질에서 잰 지연시간, 요청당 KV 캐시와 피크 메모리가 기술 지표다. 동시성을 반영한 처리량에 전력과 반복 연산까지 포함해 성공 작업당 총비용도 계산해야 한다. 시장에서는 클라우드 사용량이 가속기 출하와 서버당 메모리 탑재량으로 이어지는지 확인한다. HBM과 DRAM 가격이 공급사 마진으로 연결되는지도 봐야 한다. 어느 한 지표만 개선되면 기술 데모나 일시적 가격 효과에 그칠 수 있다.

채택 논문 15편은 예측력보다 의사결정 연결의 검증

금융 관련 채택 논문들은 예측보다 검증 가능한 AI 의사결정을 탐구

여기까지는 학회의 변화를 작업당 토큰, KV 캐시, 메모리 수요라는 인프라의 언어로 읽었다. 이제 시선을 이 노트의 앞부분, 금융 세션으로 되돌린다. 그곳의 결론은 '예측력보다 의사결정 연결의 검증'이었다. JPMorgan의 온라인 의사결정 나무 오류 통제, Balyasny의 합병차익거래 확률예측, Kalshi의 예측 정확도와 거래 수익 분리, 그리고 XTX Markets가 지적한 '모델 배포가 다음 데이터 분포를 바꾸는 문제'까지, 현장 발표들은 모두 예측 그 자체가 아니라 예측과 의사결정 사이의 연결 고리를 문제 삼았다. 흥미로운 것은 이 관찰이 발표장 밖, 채택 논문의 층위에서도 그대로 반복된다는 점이다. 금융에 직결되거나 이식 가능한 채택작 15편(arXiv 공개본 기준)을 이어 읽으면 독립적으로 발표된 논문들이 하나의 질문에 순서대로 답하는 서사가 드러난다. AI에게 자금 운용을 맡기려면 무엇부터 가르쳐야 하는가. 채점 기준은 옳은가, 교재는 제대로 됐는가, 시장이라는 세계를 읽는 눈은 있는가, 폭풍이 오면 견디는가, 그리고 그 성적표를 믿어도 되는가. 이 질문들을 7막으로 따라간다.

1. 예측에서 의사결정으로의 전환

예측 정확도가 아니라 투자 의사결정의 성과를 평가 기준으로 삼는 것

출발점은 자연스럽다. '내일 주가를 잘 맞히는 AI를 만들고 그 예측대로 사고팔면 된다.' 지난 10년간 대부분의 AI 투자 연구가 이 길을 걸었다. AI에게 예측 시험을 치르게 하고, 덜 틀릴수록 높은 점수를 주며 훈련시킨 뒤, 그 예측값을 비중 계산기에 넘긴다. 그런데 이 두 단계 사이에 금이 가 있다. 예측 점수와 투자 성과는 다른 물건이기 때문이다. 예측이 아주 조금 틀렸을 뿐인데 비중 계산기가 그 실수를 크게 증폭시켜 엉뚱한 포트폴리오를 내놓는 일이 반복된다. 이 문제 제기는 뒤에서 소개할 두 논문, "Signature-Informed Transformer for Asset Allocation"(Hwang and Zohren)과 "Decision-focused Sparse Tangent Portfolio Optimization"(Jeon et al.)의 서론이 공통으로 지적하는 내용이다. 더 나쁜 것은 예측 점수로 채점하면 AI가 '맞히기 쉬운 종목'만 편애하게 된다는 점이다. 시험 점수 1등 학생을 뽑았는데 실무에서는 헤매는 상황. 여기서 서사의 첫 번째 전환이 일어난다. 문제는 AI의 능력이 아니라 채점 기준이었다.

2. 의사결정 중심 학습

SIT와 Jeon et al.은 예측 정확도 대신 실제 투자 성과를 직접 최적화하는 학습 방식을 제시함

Hwang과 Zohren의 "Signature-Informed Transformer for Asset Allocation"(약칭 SIT)은 채점표를 통째로 바꾼다. 훈련 중에 '예측이 얼마나 맞았나'를 한 번도 보지 않는다. 대신 AI가 짠 포트폴리오의 '최악의 날들 평균 손실(CVaR)'이 작을수록 높은 점수를 준다. 처음부터 끝까지 투자 성과로만 채점하는 것이다.

여기에 무기가 하나 더해진다. 가격 그래프의 모양을 통째로 요약하는 수학 도구(경로 서명)로 '반도체가 먼저 움직이면 장비주가 따라 움직인다' 같은 선행-후행 관계를 AI의 주의(attention) 회로에 직접 심었다. 미국 대형주 실험에서 이 방식은 모든 종목에 똑같이 나눠 담는 전략과 최신 예측 AI 8종을 전부 이겼다. 결정적 증거는 해부 실험에 있다. 채점 기준을 '최악의 날 손실'에서 '평균 수익'으로 바꾸는 순간 성적이 뚝 떨어졌다. 무엇으로 채점하느냐가 승부를 갈랐다는 뜻이다. 한국 연구진(KAIST·UNIST)의 "Decision-focused Sparse Tangent Portfolio Optimization"(Jeon et al.)은 같은 철학을 현실의 제약으로 끌고 온다. 실전에서는 '딱 20개 종목만 담아라' 같은 조건이 붙는데, 고른다/안 고른다는 켜짐-꺼짐 스위치 같은 행위라 AI 학습의 핵심 수학인 미분이 통하지 않는다. 논문은 이 계단을 완만한 경사로(soft top-k)로 바꿔 학습이 끝까지 흐르게 만들었다. 코스피200을 포함한 4개 시장 실험에서 종목 수가 많은 코스피200에서 기존 방식과의 격차가 가장 컸다. 채점 기준을 고치자 AI는 달라졌다.

3. 정답지의 재설계

Label Horizon Paradox는 모델보다 정답지(레이블) 설계가 투자 성과를 좌우할 수 있음을 보여줌

AI에게 '내일 수익률을 맞혀라'라고 가르칠 때 정답지로 내일 수익률을 주는 것은 너무나 당연해서 아무도 의심하지 않았다. 중국 최대 자산운용사 계열 연구진이 "The Label Horizon Paradox: Rethinking Supervision Targets in Financial Forecasting"(Song et al.)에서 그 당연함을 실험대에 올렸다. 결과는 역설적이다. 목표는 그대로 내일 수익률인데 정답지로는 '내일 장 시작 후 30분 수익률'처럼 더 짧은 구간을 주는 편이 최종 성적이 좋은 경우가 많았다. 원리는 라디오 녹음과 같다. 방송(신호)은 앞부분에 대부분 나오고 녹음을 오래 할수록 지직거리는 잡음만 쌓인다. 오늘의 정보는 다음 날 장 초반에 대부분 가격에 반영되고, 그 뒤는 정보와 무관한 잡음이 쌓이는 시간이다. 하루 전체 수익률을 정답지로 주면 AI는 잡음까지 외운다.

논문은 여기서 멈추지 않고 어느 길이의 정답지가 최적인지를 AI가 학습 도중 스스로 찾게 하는 방법(이중 최적화)까지 내놨다. 중국 3개 지수와 미국 S&P500, 10가지 AI 구조 전부에서 성적이 올랐다. 모델도 데이터도 그대로 두고 정답지만 바꿔 얻은 개선이다. 이제 채점도 교재도 고쳤다. 다음 관문은 AI가 살아갈 세계, 시장 그 자체를 읽는 눈이다.

4. 시장 구조의 이해

CorNet 등은 개별 종목 예측을 넘어 시장의 관계 구조를 이해하는 것이 AI 투자의 핵심임을 보여줌

시장의 위기는 개별 종목의 폭락이 아니라 관계의 변화로 온다. 평소 따로 놀던 주식들이 위기가 오면 일제히 같이 떨어진다. 그래서 AI에게 필요한 것은 종목 하나하나의 예측을 넘어 시장 전체의 '친구관계표'를 읽는 능력이다. 이 능력을 정면으로 다룬 논문이 "Riemannian Networks over Full-Rank Correlation Matrices"(Z. Chen et al.), 통칭 CorNet이다. 각 종목의 출렁임 크기를 걷어내고 순수한 관계만 남긴 표(상관행렬)를 AI 신경망에 제대로 넣는 방법을 처음으로 체계화했다. 이런 표들이 사는 수학적 공간은 평평한 종이가 아니라 휘어진 곡면이어서 전용 기하학이 필요한데 다섯 가지 곡면 기하를 모두 구현해 비교했다. 출렁임 크기에 속지 않고 '지금이 평소형 시장인가 위기형 시장인가'를 분류하는 도구가 될 수 있다.

지도를 그리는 다른 손들도 있다. "Learning Gaussian Graphical Models from a Glauber Trajectory Without Mixing"(Shen et al.)은 시간적으로 이어진 데이터 한 줄기만 보고도 종목 간 연결망을 복원할 수 있음을 증명했다. 데이터가 서로 독립이어야 한다는 교과서 가정이 성립하지 않는 금융 시계열에 특히 반가운 결과다. "Ellipsoidal Time Series Forecasting"(Wang, 모델명 Fern)은 장기 예측을 '확률 덩어리를 타원 모양으로 옮기는 문제'로 다시 정의해 예측에 2차 구조(어느 방향으로 얼마나 퍼질지)를 내장했고, "A Random Matrix Theory of Masked Self-Supervised Learning"(Wortsman Zurich et al.)은 빈칸 채우기식 AI 학습(마스킹)의 작동 원리를 랜덤행렬 수학으로 규명해 이런 도구들의 신뢰 근거를 다진다. 앞서 정형 데이터 세션에서 확인한 관찰, 즉 전문 수치 모델이 여전히 강한 기준선이라는 결론과 맞닿는 대목이다. 지도를 읽는 눈은 갖췄다. 그러나 지도가 있어도 폭풍은 온다.

5. 위기 대응의 학습

관련 연구들은 분포 변화와 극단적 위험 속에서도 성과를 유지하는 강건한 AI 의사결정의 가능성을 제시

서사의 절정은 위기 관리다. 스탠퍼드 연구진의 "Adversarially Robust Control of Conditional Value-at-Risk via Rockafellar-Uryasev Conformal Inference"(C. Chen et al.)는 '최악의 날들 평균 손실을 정해진 한도 이하로 유지한다'는 약속을 최악의 시나리오에서도 지켜낼 수 있는 자동 조절 장치를 만들었다. 매일 장이 끝나면 결과를 보고 주식 비중 다이얼을 조금씩 돌리는 단순한 동작이지만 그 약속이 수학적으로 보장된다는 점이 다르다. 미래 시장의 분포에 대한 가정이 하나도 없다. 1991년부터 2025년까지 미국 주식·국채 실험에서 닷컴 붕괴, 2008년 금융위기, 코로나 급락을 지나는 동안 장치는 알아서 주식 비중을 줄였고 실제 손실은 약속 수준 근처에 머물렀다. 앞서 XTX가 제기한 '배포가 분포를 바꾸는 시장'에서 살아남는 방법이 바로 이런 무가정 온라인 적응이다.

폭풍 대비의 도구상자는 더 넓다. 딥마인드의 "Optimizing Return Distributions with Distributional Dynamic Programming"(Á vila Pires et al.)은 평균 수익이 아니라 수익률 분포 전체를 놓고 최적 전략을 계산하는 일반 이론을 세웠고, "RAMAC: Multimodal Risk-Aware Offline Reinforcement Learning"(Fukazawa et al.)은 새로 매매해 볼 수 없는 상황에서 과거 데이터만으로 최악 손실을 통제하며 전략을 배우는 방법을, "Best-of-Both-Worlds for Heavy-Tailed Markov Decision Processes"(Y. Chen et al.)는 극단 손실이 잦은 두꺼운 꼬리 환경에서도 통하는 의사결정 보장을 제시했다. 예측 구간이 약속한 적중률을 자동으로 지키게 만드는 "Online Conformal Prediction via Universal Portfolio Algorithms"(Liu et al.)는 흥미롭게도 투자 이론의 고전인 유니버설 포트폴리오 알고리즘을 도구로 썼다. 투자 수학이 AI 이론을 돕고, 그 AI가 다시 투자로 돌아오는 순환이다. 폭풍을 견디는 법까지 배웠다. 그런데 폭풍 말고도 시장에는 예고 없이 규칙 자체를 바꾸는 사건들이 있다.

6. 사건이 시장에 미친 영향

관련 연구들은 평균 변화가 아닌 위험 분포와 계산 오차까지 검증하는 분석 도구를 제시

정책 발표, 제도 변경, 대형 상장 같은 사건이 시장에 미친 영향을 잴 때 흔히 '평균 수익률이 얼마나 변했나'를 본다. 그러나 진짜 흔적은 다른 곳에 남는 경우가 많다. 평균은 그대로인데 특정 종목군의 출렁임이 커지거나 최악 손실의 꼬리가 두꺼워지는 식이다. "Conditional Distributional Treatment Effects: Doubly Robust Estimation"(Jain and Luedtke)은 사건이 평균이 아니라 '위험의 모양' 자체를 바꿨는지를 통계적으로 엄밀하게 추정하고 검정하는 최초의 도구를 만들었다. 사건 전후 비교 분석의 해상도를 한 단계 올리는 결과다. 파생상품 쪽에서는 "Error Propagation in Dynamic Programming: From Stochastic Control to American Option Pricing"(Della Vecchia and Filipović)이 옵션 가격을 계산하는 근사 알고리즘에서 오차가 단계마다 어떻게 쌓이는지를 추적해 계산 결과를 어디까지 믿어도 되는지의 근거를 제공했다. 이렇게 훈련, 교재, 세계 읽기, 폭풍 대비, 사건 분석까지 갖췄다.

7. 다섯 가지 함정

금융 LLM 연구는 성능보다 평가 편향과 검증 체계를 먼저 점검해야 함을 보여줌

챗GPT 같은 대형 언어모델로 투자 연구를 하는 논문 164편을 검토한 입장 논문 "Position: Evaluating LLMs in Finance Requires Explicit Bias Consideration"(Kong et al.)은 좋아 보이는 결과를 만들어내는 다섯 가지 함정을 정리했다. 미리보기 함정(AI가 학습 중 이미 미래 기사를 읽어버려 과거 예측 시험이 컨닝이 되는 것), 생존 함정(살아남은 종목만으로 시험해 성적이 부풀려지는 것), 서사 함정(AI가 틀린 판단에도 그럴듯한 이유를 붙이는 것), 목적 함정(글솜씨로 채점하고 투자 실력이라 결론 내리는 것), 비용 함정(수수료와 호가 차이를 빼고 계산한 수익률). 다섯 함정은 각각도 위험하지만 서로 겹치면 '타당성의 착시'를 만든다.

충격적인 것은 164편 중 이 함정을 단 하나라도 논의한 논문이 28%가 안 된다는 조사 결과다. 이 노트의 서두에서 확인한 평가 기준의 이동, 즉 정답률에서 작업 완주율로의 확대와 같은 방향이다. 무엇을 재는지가 곧 무엇을 얻는지를 결정한다.

표2 여정의 지도 - 7막과 15편

막	질문	답한 논문	실무 적용
1막	무엇으로 채점할 것인가	SIT·희소 DFL 서론의 문제 제기	예측 정확도 지표 맹신 중단
2막	성과로 채점하려면	SIT, 희소 포트폴리오 DFL	배분 엔진 실험(코스피200 재현)
3막	교재는 옳은가	라벨 역설	정답지 길이 스윕 실험
4막	세계를 읽는 눈	CorNet, Glauber 그래프, Fern, 마스킹 RMT	관계 지도로 시장 국면 분류
5막	폭풍을 견디는 법	온라인 CVaR 제어, 분포 DP, RAMAC, 꼬리 MDP, UP-OCP	손실 한도 자동 오버레이
6막	사건의 흔적 읽기	조건부 분포 처치효과, DP 오차전파	이벤트의 위험 구조 분석
7막	성적표를 믿어도 되는가	금융 LLM 5대 편향	백테스트 검증 체크리스트

자료: ICML 2026, arXiv, DS투자증권 리서치센터.

예측 잘하는 AI에서 결정, 검증을 통과하는 AI로

15편을 하나의 문장으로 압축하면 이렇다. 'AI 투자의 최전선은 예측을 잘하는 기계에서, 성과로 채점받고(2막), 올바른 교재로 배우고(3막), 시장의 관계 구조를 읽고(4막), 최악의 날을 수학적 보장 속에서 견디며(5·6막), 스스로의 성적표를 의심하는(7막) 기계로 이동하고 있다.' 발표장에서 관찰한 '예측력보다 의사결정 연결'이라는 흐름이 채택 논문의 층위에서 이렇게 완결된 서사로 확인되는 셈이다. 각 논문의 수치와 실험 조건은 arXiv 공개본을 기준으로 했으며 검토 기준의 한계는 뒤의 '자료와 해석의 한계'에 함께 정리했다. 논문의 서사는 여기서 마친다. 이 서사가 학회 현장의 공기와 어떻게 만나는지, 그리고 그것이 어떤 수요로 이어지는지는 이어지는 현장 소고에서 마무리한다.

장기 작업의 확산이 메모리 수요로 이어지는 조건

현장에서는 단일 답변보다 긴 작업을 끝내는 시스템이 늘었음

현장에서 가장 분명했던 변화는 단일 모델의 답변 품질보다 여러 모델과 도구를 묶어 긴 작업을 끝내려는 연구가 늘었다는 점이다. 법률, 바이오와 의료에서는 오류 비용이 큰 업무를 대상으로 검증 가능한 프로토타입이 제시됐다. 정형 데이터에서는 전문 수치 모델이 예측을 맡고 LLM이 문맥을 연결해 결과를 설명하거나 도구를 조정했다. Diffusion 패널과 수상 논문은 동일 품질을 전제로 지연시간, 메모리 부담과 총서빙비용을 낮출 수 있는지를 핵심 조건으로 제시했다. 경쟁의 초점은 하나의 모델이 모든 작업을 맡는지보다 역할별 모델과 도구를 어떻게 조합하는지로 옮겨갔다.

긴 작업이 반드시 문맥을 무한히 쌓는 방향으로 발전하는 것은 아니다. Microsoft Dynamics 365 비용 정산 과제에서는 최근 도구 상호작용만 남기고 과거를 요약한 구성이 전체 이력을 유지한 구성보다 토큰을 62.6% 줄이면서 완주율을 20.6%포인트 높였다. 특정 ERP 과제의 사전공개 결과라 일반화할 수는 없다. 작업 길이가 늘수록 문맥을 선택적으로 보존하는 시스템 설계가 모델 성능만큼 중요해진다.

작업량에서 설비투자와 메모리 마진까지 연결해 봐야 함

운용 관점에서는 사용량, 인프라와 메모리 실적을 차례로 확인한다. 먼저 작업당 토큰과 도구 호출을 보고 완주율이 함께 높아지는지 살핀다. 다음으로 클라우드 매출과 설비투자가 가속기 출하로 이어지는지 본다. 마지막으로 서버당 HBM과 DRAM 탑재량, 가격과 공급사 마진이 같은 방향으로 움직이는지 점검한다. 문맥 압축으로 작업당 토큰이 줄거나 증설 속도가 수요를 앞서 가격과 마진이 꺾이면 메모리 수요 확대 해석도 낮춰야 한다. Diffusion과 문맥 효율화는 메모리 수요 감소의 결론이 아니라 작업량 증가와 함께 추적할 비용 변수다.

구두발표와 스팟라이트는 성능을 비용·검증 문제로 바꿨음

ICML 2026에서 두드러진 변화는 단일 모델의 평균 점수보다 긴 작업의 완주율과 실패 조건을 따지는 연구가 늘었다는 점이다. 메인 구두발표와 스팟라이트는 학습 메모리와 KV 캐시를 다루고 모델 비교와 법률 준수를 구체적인 측정 문제로 바꿨다. 수학과 바이오 워크숍은 자동 검증과 실험 피드백을 연구 절차에 넣었다. 금융에서는 JPMorgan, Balyasny와 Kalshi 연구가 예측 정확도뿐 아니라 업데이트의 통계적 안전성과 장문서 조사, 거래 규칙을 다뤘다. Diffusion 대 AR 패널과 두 수상 논문까지 함께 보면 경쟁의 초점은 아키텍처 이름보다 같은 품질에서 지연시간과 메모리, 성공 작업당 비용을 얼마나 낮추는지에 있다. 운용 관점에서는 작업량 증가와 단위 효율 개선 가운데 어느 쪽이 빠른지 먼저 봐야 한다. 이어 그 변화가 가속기 투자와 메모리 탑재량, 공급사 마진으로 연결되는지를 확인해야 한다.

Compliance Notice

- 동 자료는 기관투자가 등 제 3자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다. (작성자: 투자전략팀, 김남일)
- 동 조사자료는 고객의 투자에 참고가 될 수 있는 각종 정보제공을 목적으로 제작되었습니다. 이 조사자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻어진 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 이 조사자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.
- 동 조사자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포 할 수 없습니다.