CONDERI

ACADEMIC & RESEARCH INSIGHT | 대외 공개판

많이 나누는 것과 잘 나누는 것은 다르다 — 에이전트 설계의 최적 입도

Dividing Is Not Designing — Optimal Granularity in Multi-Agent Systems

이 리포트가 다루는 것

지능정보연구 32(3)의 AI 신뢰성 특집에 실린 재난문자(KPAS) 의사결정 지원 연구는 에이전트를 무작정 늘리는 대신 **계층을 만들고 인계 지점을 통제**했을 때 기존 문안 대비 61.0%의 승률을 얻었다고 보고합니다. 단일 에이전트보다 나았고, 과도하게 세분화한 구조보다도 나았습니다. 같은 주 부산에서 열린 한국지능정보시스템학회 추계학술대회는 100여 편 가운데 에이전틱 AI와 AI 거버넌스를 별도 세션으로 배치했습니다. 이 리포트는 이 두 신호와 McKinsey의 조직 연구를 겹쳐, 기업이 에이전트를 어떻게 나눠야 하는지를 프레임워크로 정리합니다.

61.0%
기존 KPAS 문안 대비
종합 평가 승률

100여 편
KMIS 추계학술대회
15개 세션

80% vs 37%
개인 생산성 vs
기업 EBIT 개선

4~6명
에이전트 결합 후
제품팀 인원 예시

I 기술
계층형 구조의
승률 61.0%

II 경영
워크플로를
다시 그린다

III 과학
사람이 남는
지점의 설계

IV 교육
검증하는 사람을
기르는 일

i. 요약 — 성과는 개수가 아니라 인계에서 나온다

Granularity, not headcount: where multi-agent performance comes from

61.0%	100여 편	80%	37%	6%
KPAS 문안 승률 (LLM 쌍대 비교)	KMIS 추계 발표 논문	AI로 개인 생산성 향상 응답	EBIT 긍정 응답	AI 고성과 기업 비율

한 문장

에이전트의 성과는 **개수**가 아니라 **인계 지점의 설계**에서 나온다. 이번 주 문헌은 그 명제를 공공 서비스(재난문자)와 기업 조직(워크플로)이라는 서로 다른 두 현장에서 같은 방향으로 확인했습니다.

논문이 측정한 것

지능정보연구 32(3), 임송재·김진·이정훈(연세대). MCP 기반 이종 데이터 융합과 RAG를 묶은 계층형 멀티에이전트를 LangGraph로 구현하고, 화재 재난 시뮬레이션에서 LLM-as-a-Judge 쌍대 비교와 Human-in-the-Loop 검증을 수행. 기존 KPAS 문안 대비 종합 평가 승률 61.0%.

왜 단일보다 나았나

뉴스·SNS·공공데이터를 병렬로 조율하면 상황 파악이 빨라지고, 법령 지식베이스를 RAG로 붙이면 문안의 법적 근거가 남습니다. 속도와 근거를 동시에 얻는 구조입니다.

왜 과세분화보다 나았나

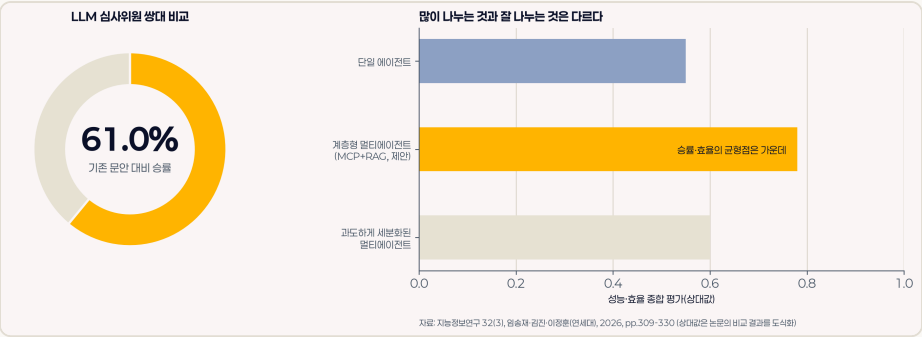
에이전트를 지나치게 쪼개면 인계 횟수가 늘고, 인계마다 맥락이 손실되며 지연이 누적됩니다. 논문은 성능과 효율의 **균형점**이 계층형 구조에 있다고 정리합니다.

이 문서를 읽는 방법

인용한 논문·발표 자료의 수치와 표현은 원문 기준으로 옮겼고, 해석과 프레임워크는 콘데리 AI산업전략연구실의 것입니다. 발표 단계 논문은 게재본과 수치가 달라질 수 있음을 밝힙니다.

I. 기술 — 계층이 성과를 만든 이유

Why hierarchy beat both the single agent and the swarm



자료: 지능정보연구 32(3), 임송재·김진·이정훈(연세대), 2026, pp.309-330

연구 설계	내용
문제	재난문자(KPAS)는 가장 즉각적인 공공경보 수단이지만 발송 관리자의 수동 판단에 의존해 정보 취합 지연, 인지적 과부하, 경직된 표준 문안, 경보 피로가 반복
구조	MCP 기반 이종 데이터 융합 + RAG. LangGraph 위에서 뉴스·SNS·공공데이터 전문 에이전트를 병렬 조율하고, 행정안전부 예규를 지식베이스로 법적 준거성 강화
검증	화재 재난 중심 트레이스 기반 시뮬레이션. LLM-as-a-Judge 쌍대 비교 + Human-in-the-Loop 검증
결과	기존 KPAS 문안 대비 종합 평가 승률 61.0%. 단일 에이전트와 과도하게 세분화된 멀티에이전트 모두와 비교해 성능·효율 균형이 가장 좋은 구성

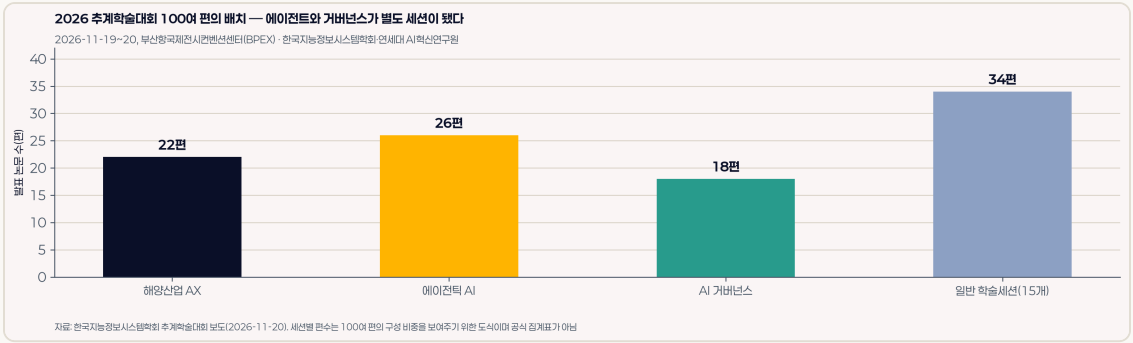
61.0%를 읽는 법

이 숫자는 정답률이 아니라 **선호 판정의 승률**입니다. 심사위원(LLM)에게 두 문안을 나란히 주고 어느 쪽이 나은지 고르게 한 비율이 61.0%라는 뜻입니다. 더 중요한 것은 이 연구가 **비교 대상**을 설정한 방식입니다. 단일 에이전트와 과세분화 구조를 함께 놓고 "더 많이 나누면 좋아지는가"라는 질문에 답했고, 결과는 아니오였습니다.

기술 에이전트 수를 늘리는 결정은 성능 결정이 아니라 비용·지연·책임 결정이다

II. 학회 — 에이전트와 거버넌스가 별도 트랙이 됐다

When agentic AI and governance get their own sessions



자료: 한국지능정보시스템학회 2026 추계학술대회(2026-11-19~20, BPEX)

학회가 세션을 나눈 방식이 곧 연구 의제다

한국지능정보시스템학회는 11월 19~20일 부산항국제전시컨벤션센터에서 연세대 AI혁신연구원과 함께 추계학술대회를 열었습니다. 주제는 "북극항로 시대를 여는 해양산업의 인공지능 대전환"이었고, 해양산업 AX·에이전틱 AI·AI 거버넌스가 특별세션으로, 15개의 일반 학술세션과 함께 100편 이상의 논문이 발표됐습니다. 눈에 띄는 것은 **에이전틱 AI와 AI 거버넌스가 별도 트랙**이 됐다는 점입니다. 2년 전이라면 같은 세션에 섞여 있던 주제가 이제 각자의 심사 기준과 발표 형식을 가집니다.

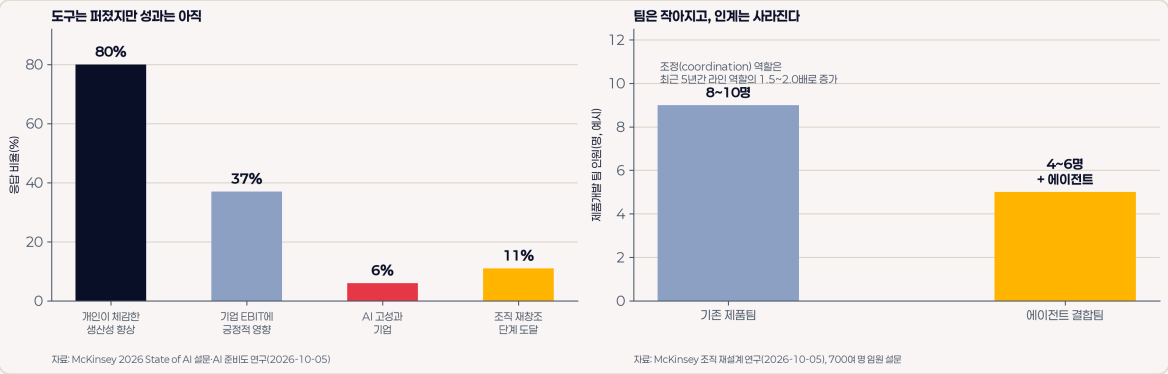
산업별 적용이 먼저 온다

해양산업이라는 특정 도메인을 전면에 세운 것도 시사적입니다. 항만·조선·해운은 규정이 많고 문서가 길며, 실수의 비용이 큰 산업입니다. 이 조건은 재난문자 연구의 조건과 같습니다. 규정 준수가 필요한 곳에서 에이전트가 먼저 쓰이고, 그곳에서 검증 방법이 먼저 만들어집니다. 범용 챗봇의 성능 경쟁이 아니라 **도메인 문서와 규정을 얼마나 정확히 다루는가**가 연구의 중심으로 옮겨간 것입니다.

학회 에이전트 연구의 무게중심은 모델 성능에서 규정 준수·검증·책임으로 이동했다

III. 경영 — 조직도를 오케스트레이션 차트로

From org chart to orchestration chart



자료: McKinsey 2026 State of AI 설문·조직 재설계 연구(2026-10-05), 임원 7000여 명

같은 격차가 기업에서도 확인된다

McKinsey의 2026년 조사에서 응답자의 80%는 AI가 개인 생산성을 높였다고 답했지만, EBIT에 긍정적 영향이 있다고 답한 비율은 37%였고 AI 고성과 기업은 6%에 그쳤습니다. 고성과 기업의 약 4분의 3이 워크플로를 크게 재설계했다고 답한 반면, 나머지 기업은 4분의 1 수준이었습니다. 격차를 만든 것은 모델이 아니라 **일이 흐르는 방식**입니다.

팀의 크기가 바뀐다

보고서는 한 조직의 제품개발 조직이 8~10명 단위 팀에서 4~6명에 에이전트를 붙인 하이브리드 팀으로 옮겨간 사례를 소개합니다. 사라진 것은 사람이 아니라 **인계**입니다. 같은 보고서는 조정(coordination) 역할이 최근 5년간 라인 역할의 1.5~2.0배 속도로 늘었다고 지적합니다. 조직이 복잡해질수록 결정을 옮기는 사람이 늘었고, AI는 그 이동을 줄일 기회입니다. 단, 잃어버린 조정을 더 나은 정보와 명확한 결정 권한으로 대체할 때만 그렇습니다.

경영 AI를 기존 프로세스 위에 얹으면 복잡성만 한 겹 늘어난다. 재설계는 역할·권한·인계를 함께 바꾸는 일이다

IV. 과학·교육 — 검증하는 사람을 기르는 일

Human in the loop, by design

사람이 남는 지점을 설계하는 일

재난문자 연구에서 가장 눈여겨볼 대목은 성능 수치가 아니라 **검증 절차**입니다. 연구진은 LLM이 두 문안 중 나은 것을 고르게 하는 자동 평가와, 사람이 최종 확인하는 절차를 함께 두었습니다. 자동 평가는 대량 비교를 가능하게 하고, 사람의 확인은 그 비교가 놓친 것을 잡습니다. 이 조합은 인간-AI 협업 연구에서 반복적으로 확인되는 패턴입니다. AI는 후보를 만들고 좁히는 데 강하고, 사람은 **맥락과 책임**을 판단하는 데 강합니다.

자동 평가의 한계

LLM 심사위원은 일관성이 높지만, 심사 기준 자체가 학습 데이터의 편향을 물려받습니다. 그래서 승률 61.0%는 "기존 문안보다 낫다"는 신호이지 "충분히 좋다"는 증명이 아닙니다. 논문이 Human-in-the-Loop을 함께 둔 이유입니다.

사람이 개입할 세 지점

첫째, 입력 단계에서 무엇을 근거로 삼을지 정합니다. 둘째, 인계 단계에서 무엇을 넘길지 화이트리스트로 정합니다. 셋째, 출력 단계에서 공개 여부와 책임자를 정합니다. 세 지점 중 하나라도 비어 있으면 자동화는 책임을 흐립니다.

교육·훈련에 옮기면

이번 주 공개된 문헌 가운데 AI 리터러시 교육을 직접 실증한 연구는 확인하지 못했습니다. 다만 위의 두 연구가 공통으로 말하는 것은 분명합니다. 교육의 목표를 "도구를 쓸 수 있다"에서 **"도구의 답을 검증할 수 있다"**로 옮겨야 합니다. 재난문자 문안을 심사하는 능력, 에이전트가 넘긴 맥락에서 빠진 것을 찾는 능력이 그 사례입니다. 기업 교육의 성과 지표도 수료율보다 검증 비율로 바뀌어야 같은 방향이 됩니다.

과학·교육 자동화의 마지막 단계는 언제나 사람이다. 그 사람이 무엇을 검증할지 정하는 것이 설계의 핵심이다

V. 실무 프레임워크 — 네 개의 결정

Four decisions before adding an agent

결정	질문	권장 기본값
① 입도	이 일을 몇 개의 에이전트로 나눌 것인가	업무 단계 수만큼 나누지 말고, **판단이 바뀌는 지점**에서만 나눈다
② 인계	단계 사이에 무엇을 넘길 것인가	대화 전체가 아니라 결정에 필요한 필드만 화이트리스트로 넘긴다
③ 검증	누가 무엇을 확인하는가	자동 평가 + 사람 확인의 2단. 확인 항목은 3개 이하로 고정한다
④ 책임	결과의 책임자는 누구인가	에이전트가 아니라 직책으로 지정한다. 로그에 남긴다

도입 심사에서 먼저 볼 세 숫자

정확도를 첫 장표에 올리지 않는 것이 좋습니다. 대신 ① 출처가 붙은 문장의 비율, ② 인계 지점에서 걸러진 오류 건수, ③ 사람이 최종 확인한 비율을 봅니다. 이 세 숫자가 오르지 않으면 에이전트 수를 늘려도 결과는 같습니다. McKinsey가 말한 "워크플로 재설계"도 결국 이 세 숫자를 올리는 일입니다.

이번 주 문헌이 남긴 문장

지능정보연구 32(3)의 결론은 절제되어 있습니다. 계층형 구조가 성능과 효율의 균형을 달성한다는 것, 그리고 그 균형은 **설계 선택**이라는 것입니다. 모델을 바꾸는 것으로 얻을 수 없는 개선이 설계에는 남아 있습니다.

실무 에이전트를 늘릴지 말지가 아니라, 어디서 나누고 무엇을 넘길지를 먼저 정하라

English Summary

For global readers and archive

English Summary

A study in the Journal of Intelligence and Information Systems (vol. 32, no. 3, 2026, pp. 309-330) by Songjae Lim, Jin Kim and Jeonghun Lee of Yonsei University evaluates a hierarchical multi-agent system for Korea's public alert service (KPAS). The system combines Model Context Protocol based data fusion with retrieval-augmented generation, orchestrating news, social media and public data agents in parallel on LangGraph, and grounds drafts in the Ministry of the Interior's alert regulations. In a trace-based fire scenario, with pairwise evaluation by an LLM judge and human-in-the-loop verification, the proposed system won 61.0 percent of overall comparisons against existing KPAS drafts. It also outperformed both a single-agent architecture and an over-fragmented multi-agent design, which the authors describe as the balance point between performance and efficiency.

The same signal from industry and from the field

The Korean Intelligent Information Systems Society held its autumn conference in Busan on November 19 and 20 with Yonsei's AI Innovation Institute, running dedicated sessions on agentic AI and AI governance alongside 15 academic sessions and more than 100 papers. In management research, McKinsey's 2026 survey finds 80 percent of respondents reporting individual productivity gains from AI but only 37 percent reporting positive EBIT impact and 6 percent qualifying as high performers; nearly three quarters of high performers report substantially redesigning workflows, against about a quarter of others. One product organisation moved from pods of eight to ten people to hybrid teams of four to six supported by agents, and coordination roles have grown at 1.5 to 2.0 times the rate of line roles over five years.

What to measure first

Before adding agents, measure three numbers: the share of sentences with a cited source, the number of errors caught at handoffs, and the share of outputs a human verified. If those do not rise, more agents will not change the outcome. The design decisions that matter are where to split, what to pass, who verifies and who is accountable.

Takeaway **Dividing work is not the same as designing it. The gain comes from controlled handoffs**