OPEN 공공누리 출처표시 상업용금지 공공저작물 자유이용허락

2026-38호
(2026.11.06.)

IITP AI·ICT Brief



CONTENTS

주요 이슈 01
1. 스마트 스케일링, AI 성능 경쟁이 규모에서 효율로 확장 01

주요국 동향 11
1. 프랑스·체코 AI 전략 협정 체결, 기가팩토리 유치 공조 추진 11
2. 중국, 「국가 정보화 15차 5개년 계획」 발표 15

ICT 부문별 모니터링 19

주요 ICT 행사 일정 21

카카오톡 채널 바로가기

IITP 정보통신기획평가원
Institute of Information & Communications Technology Planning & Evaluation

주요 내용

주요 이슈

스마트 스케일링, AI 성능 경쟁이 규모에서 효율로 확장

- 모델 규모를 키울수록 같은 폭의 성능 개선에 드는 연산이 급증하는 반면, 같은 연산으로도 데이터 처리와 학습 설계에 따라 성과가 크게 달라져 자원 활용 능력의 중요성이 확대
- 개발 단계별로 △데이터 다양성·정확성 보완 △대형 학습 사전 예측과 GPU 활용 개선 △작업 기반 강화학습과 증류 △난이도별 추론 연산 배분 등 같은 자원으로 성능을 높이는 기술이 잇따라 등장
- 배포 후 쌓인 업무 경험을 외부 메모리나 작업 절차, 모델 자체에 반영하는 지속 학습도 등장했으나, 경험만으로 업무 규칙을 익히는 능력은 아직 제한적이어서 실제 효과의 검증이 필요
- 스마트 스케일링의 성과는 △실제 환경에서의 재현성 △업무 전체 비용 △배포 후 개선 여부로 판단해야 하며, 기업은 총비용 중심의 AI 운용을, 정부는 성과 중심의 연산 배분을 추진할 필요

주요국 동향

프랑스·체코 AI 전략 협정 체결, 기가팩토리 유치 공조 추진

- 프랑스·체코가 AI 전략 협정을 맺고 EU 기가팩토리 유치에 공조하기로 하면서, 회원국별 단독 응모 중심이던 EU AI 인프라 유치 경쟁에 국가 간 공조 방식이 더해지는 흐름
- 같은 날 체코 CRA와 프랑스 Bull도 체코 기가팩토리 구축 협력 MOU를 체결했으나, 정부 협정과 기업 MOU 모두 비구속 합의 단계로 공모 일정과 인허가·전력망·자금조달이 사업화의 변수

중국, 「국가 정보화 15차 5개년 계획」 발표

- 중국 중앙사이버보안·정보화위원회가 「국가 정보화 15차 5개년 계획」 발표('26.10.), AI 기본법·에이전트 등록제 정비 명시, 2030년 에이전트 보급률 90% 등 12대 지표 제시
- 기술 육성과 규제 정비를 묶어 AI 운영 규칙을 먼저 완성하려는 구도로, 등록·표준 통과가 진입 요건이 되는 만큼 한국도 에이전트 세부 규칙 설계와 국제 표준 논의 참여를 앞당길 필요

문의처 동향분석팀 : 허창회 수석 TEL. 042-612-8243 Email. tareheo@iitp.kr

Ⅰ. 주요 이슈 스마트 스케일링, AI 성능 경쟁이 규모에서 효율로 확장

규모 확대가 이끈 AI 경쟁, 비용·데이터 한계 속에 '스마트 스케일링' 등장

모델 규모를 키울수록 투자 부담은 커지고, 고품질 데이터는 부족해지는 상황

- 지금까지 AI 기업들은 모델의 크기와 학습 데이터, 연산량을 늘릴수록 성능이 높아진다는 '스케일링 법칙'을 근거로, 더 큰 모델을 개발하는 데 투자를 집중

  * 스케일링 법칙: '20년 오픈AI 연구진이 정리한 경험 법칙으로, 자원을 늘렸을 때 성능이 얼마나 향상될지 학습 전에 계산할 수 있어 투자 판단의 근거로 활용

- 그 결과 최첨단 AI 학습은 막대한 자금을 갖춘 소수 기업에 집중되고 있으며, 美 4대 하이퍼스케일러의 '26년 설비투자 계획만 약 7,330억 달러 규모 (Air Street, '26.10.)

  * 하이퍼스케일러: 초대형 데이터센터를 직접 구축·운영하며 전 세계에 클라우드 서비스를 제공하는 대형 IT 기업

- 반면 학습에 적합한 고품질 데이터와 분야별 전문 데이터는 공급이 제한적이어서, 연산 자원과 달리 비용을 더 투입하더라도 필요한 만큼 확보하기 어려운 상황

규모 확대에만 의존하지 않고 성능을 높이는 '스마트 스케일링' 제시

- '24년 12월 일리야 수츠케버 전 오픈AI 수석과학자는 NeurIPS 강연에서 기존 사전학습의 한계를 지적하며, 새로운 AI 성능 향상 방식의 필요성을 강조 (더밀크, '26.10.)

  * 사전학습·사후학습: 사전학습은 대량의 텍스트로 언어의 규칙과 지식을 익히는 첫 학습 단계이며, 사후학습은 그 이후 특정 과업을 잘하도록 모델을 다듬는 단계로 미세조정·강화학습·증류 등을 포함

- 이후 최예진 스탠퍼드대 교수는 '26년 9월 강연에서 '지능의 스케일링은 계속되겠지만 더 똑똑한 방식이어야 한다'며, 규모 중심 개발의 대안으로 '스마트 스케일링'을 강조 (더밀크, '26.10.)

- 실제로 사후학습 기술이 발전하면서, 알리바바의 Qwen 3.8(270억 파라미터)처럼 상대적으로 작은 모델이 일부 평가에서 대형 모델에 근접하는 성능을 보이는 사례도 등장 (MLCommons, '26.9.)

☞ 스마트 스케일링의 정의와 용어 등장 배경


최예진 스탠퍼드대 교수

정의	모델 규모를 계속 키우기보다 데이터와 학습·추론 방법을 개선해, 같은 자원으로 더 높은 성능을 얻으려는 AI 개발 방향		
등장 배경	최예진 스탠퍼드대 교수가 '25년 4월 美 프린스턴대 강연에서 자원 투입에 의존하는 규모 확대 중심 개발의 대안으로 제시 (Princeton AI Lab, '25.5.)		
주요 방법	① 적은 데이터로 빠른 학습 강화학습 등 학습 방식 개선	② 데이터 다양성 확보 부족한 유형의 데이터 합성	③ 추론 시점의 성능 향상 문제를 푸는 중에도 학습

자료 : 더밀크('26.10.), Princeton AI Lab('25.5.) 참고하여 재구성

AI 경쟁의 핵심 질문, '얼마나 크게'에서 '같은 자원으로 얼마나 더'로 이동

- 투자 부담이 커지고 데이터 확보도 어려워지면서, 자원을 더 많이 확보하는 경쟁에 더해 '같은 자원으로 성능을 얼마나 더 끌어낼 수 있는가'라는 새로운 질문이 제기
- 특히 대규모 자원을 갖추지 못한 기업과 국가에는 확보한 자원을 얼마나 효율적으로 활용하느냐에 따라 선도 기업과의 성능 격차를 좁힐 여지가 생긴다는 점에서 의미가 큰 질문
- 다만 효율화에도 추가 비용이 드는 만큼, 실제 업무에서 투입 비용 대비 성능 개선 효과가 나타나는지, 배포 후에도 기존 능력을 유지하면서 지속적으로 성능을 높일 수 있는지 검증할 필요

스마트 스케일링, 자원 배분과 경험 활용이 성능·비용 격차를 좌우

01 (규모 확대의 한계) 성능 개선에 필요한 연산량은 증가하고, 학습 설계에 따른 효율 격차 확대

같은 폭의 성능 개선에 필요한 자원이 갈수록 크게 증가

- 모델 크기와 학습 연산을 늘려도 성능 향상 폭은 점점 줄어, AI가 다음 단어를 잘못 예측하는 정도를 10% 낮추려면 학습 연산을 지금보다 8배 넘게 늘려야 하는 것으로 분석

  * 다른 자원은 충분하다고 가정하고 한 가지 자원만 늘릴 경우, 오차를 10% 낮추려면 학습 데이터는 지금보다 3.03배, 모델 크기(파라미터 수)는 4배, 학습 연산량은 8.23배가 필요 (Springer, '26.10.)

- 또한 기업이 공개하는 학습 비용에는 학습 조건을 찾기 위한 시험 학습과 실패한 학습의 비용이 빠진 경우가 많아, 실제 개발 비용은 공개 수치보다 클 것으로 추정

같은 성능에 필요한 연산도 데이터 처리·학습 설계에 따라 큰 차이

- 반면 같은 연산이라도 데이터 처리 방식과 모델·인프라 설계에 따라 얻을 수 있는 성능이 크게 달라져, 연산 확보와 함께 설계 역량이 성능을 좌우하는 요인으로 주목
- 실제로 마이크로소프트는 데이터·모델 구조·인프라를 함께 개선해, 비교 대상 모델의 성능에 도달하는 데 필요한 학습 연산을 최대 약 12분의 1로 줄일 수 있다고 발표

  * 비교 대상은 딥시크가 공개한 대형 모델(DeepSeek-V4-Pro-Base)이며, 필요 연산이 4.2~12.3배 적다는 수치는 마이크로소프트의 자체 추정치 (Microsoft AI, '26.10.)

- 특히 웹 데이터에서 내용이 완전히 겹치는 문서만 걸러내도 동일한 성능에 필요한 연산이 약 100분의 1로 줄어드는 등, 데이터 처리 방식에 따라 필요한 연산이 크게 달라지는 것으로 확인 (Meta 외, '26.3.)
- 또한 동일한 학습 연산량에서도 데이터 처리 방식에 따라 모델 크기 대비 적정 학습 데이터 양의 비율이 최대 182배 차이 나는 것으로 분석돼, 학습 자원 배분의 중요성 부각 (Meta 외, '26.3.)

02 (스마트 스케일링 기술) 데이터·학습·추론 단계별로 같은 자원에서 더 높은 성능 확보

그림 개발 단계별 스마트 스케일링 적용 지점



(1) 데이터: 양을 늘리기보다 부족한 유형을 채우고 틀린 정답을 바로잡아 학습 효과 향상

(다양성) 기존 데이터에 없는 유형의 문제를 채워, 적은 데이터로 높은 성능 달성

- 그간 학습 데이터 확보는 양을 늘리는 데 집중했으나, 유사한 유형의 문제만 반복되면 새로운 유형의 문제에 대응하는 능력을 높이는 데 한계
- 이에 기존 데이터에서 과다한 유형은 줄이고 부족한 유형은 AI로 생성해 보완하는 방식이 주목받으나, 생성한 데이터에도 기존과 겹치는 내용이 많아 이를 걸러내는 별도 연산이 필요
- 데이터 다양성을 높이는 '프리즘 합성' 연구에서는, 상대적으로 작은 모델로 생성한 데이터만으로 학습해도 비교 모델보다 7개 수학 평가 중 6개에서 높은 성능을 기록 (더밀크, '26.10.)

* 프리즘 합성: '25년 세계 최대 AI 학회인 NeurIPS에서 발표된 데이터 합성 기법으로, 300회 넘는 학습 실험을 통해 처음 보는 유형의 문제를 푸는 성능이 데이터 양보다 다양성에 더 크게 영향을 받는다는 점을 확인

(검수) 학습 데이터의 틀린 정답을 바로잡아 특화 모델의 정확도 향상

- 학습 데이터에 틀린 정답이 포함되면 모델이 잘못된 내용을 학습할 수 있으며, 일부 공개 데이터에서도 정답 오류가 발견돼 성능 개선을 저해하는 요인으로 작용
- 이에 AI가 먼저 오류 후보를 찾고 분야 전문가가 이를 확인·수정하는 검수 과정을 별도로 두어, 학습 데이터에 담긴 정답의 정확도부터 높이는 사례가 등장
- 美 일리노이대 연구진은 공개 SQL 학습 데이터(BIRD) 표본 2,500건 중 52.1%에서 정답 오류를 발견, 이를 정제해 학습한 특화 모델은 일부 평가에서 범용 최상위 모델보다 높은 정확도를 기록 (Thinking Machines, '26.8.)

* 특화 모델(ReViSQL): 공개된 AI 모델에 정답 오류를 바로잡은 데이터를 학습시켜 만든 SQL 작성 전용 모델

(2) 학습 시스템: 대형 학습의 실패 위험과 GPU 유휴 시간을 줄여 학습 효율 향상

(확장 예측) 작은 모델로 대형 모델의 성능과 학습 조건을 미리 예측해 실패 비용 절감

- 대형 모델 학습에는 1만 개 이상의 GPU를 한 달 넘게 투입하기도 해, 학습 조건을 잘못 설정하면 대규모 연산 자원이 낭비되고 개발 일정도 지연될 위험
- 이에 크기가 다른 소형 모델 여러 개를 먼저 학습시켜 규모별 성능 변화를 파악한 뒤, 이를 토대로 대형 모델의 성능과 학습 조건을 사전에 예측하는 방법이 활용
- 마이크로소프트는 '스케일링 사다리'를 활용해 2.8조 파라미터 모델의 학습 성능을 사전에 예측하고, 대규모 학습에 적합한 모델 구조와 학습 조건을 선정해 시행착오를 축소 (Microsoft AI, '26.10.)

  * 작은 실험에서도 대형 학습과 같은 횟수만큼 데이터를 반복 학습하도록 조건을 맞춰, 소형 모델의 결과가 대형 학습에서도 그대로 들어맞도록 보정

(구현 최적화) GPU 간 통신·메모리 관리 개선으로 연산 활용률 향상

- 일부 연산량을 줄이는 전문가 혼합(MoE) 구조도 학습 중 전문가별 작업량 불균형과 메모리 관리 부담이 커지면, GPU 활용률이 낮아져 실제 학습 효율이 떨어지는 문제 발생

  * 전문가 혼합(MoE): 모델을 여러 '전문가' 신경망으로 나누고 입력마다 일부만 골라 계산하는 구조로, 전문가들이 여러 GPU에 나뉘어 있어 GPU 간 데이터 전송이 많음

- 이에 전문가별 작업량에 맞춰 메모리를 배분하고 처리 방식을 개선하는 접근이 등장했으나, 모델 구조와 장비 특성에 따른 맞춤 구현이 필요해 전문 인력과 개발 부담이 수반
- 마이크로소프트는 MoE의 메모리 관리와 전문가별 연산 처리를 개선해 GPU 연산 활용률을 24%에서 35%로 높이고, 모든 계층에 MoE 구조를 적용 (Microsoft AI, '26.10.)

(3) 사후학습: 실제 작업 결과로 보상하고, 분야별 능력을 하나로 합치면서 기존 능력 유지

(환경 기반 강화학습) 실제 작업 환경에서 과제를 수행하고 성공 여부를 검증해 학습

- 수학·코딩 문제의 최종 정답을 맞히도록 학습하고 보상하는 강화학습은 파일 열기·명령 실행 등 여러 단계를 거치는 실제 업무의 수행 능력을 평가하고 개선하는 데 한계

  * 보상(Reward): AI가 수행한 작업의 결과를 평가해 부여하는 점수로, 강화학습에서 바람직한 행동을 학습하도록 유도하는 기준

- 이에 '26년 AI 개발사들은 AI가 실제 개발 환경에서 파일을 탐색하고 코드를 수정·실행하도록 한 뒤, 테스트 통과 등 작업 성공 여부에 따라 보상을 부여하는 강화학습 방식이 확산

  * 美 AI 코딩 도구 기업 커서는 자사 코딩 모델을 다양한 개발 환경에서 강화학습시켰으며, 美 엔비디아(Nemotron 3 Ultra)와 中 AI 기업 문샷AI(Kimi K3)·Z.ai(GLM-5)·미니맥스 등도 같은 방식을 도입 (Hugging Face, '26.8.)

- AI 학습 성능을 비교하는 국제 시험인 MLPerf도 '26년 10월 신설한 사후학습 시험에 이 방식(RLVR)을 표준 과제로 채택해, 작업 결과로 보상하는 방식이 성능 평가에도 반영 (MLCommons, '26.9.)

  * RLVR(검증 가능한 보상 기반 강화학습): 사람의 평가 없이 정답 대조나 테스트로 자동 채점할 수 있는 과제에 쓰는 강화학습으로, 수학·코딩 분야에서 주로 활용

(분야별 교사 증류) 분야별 전문 모델의 능력을 한 모델로 모아 기존 능력 저하 방지

- 한 모델을 여러 분야에 걸쳐 순차적으로 강화학습시키면, 새로운 분야를 익히는 과정에서 기존 능력이 손상되는 '재앙적 망각'이 발생할 수 있는 문제

  * 재앙적 망각: 모델 내부의 같은 수치가 여러 지식을 함께 담고 있어, 새 내용을 배우며 그 수치를 고치면 이전 지식까지 흐트러지는 현상

- 이에 수학·코딩 등 분야별 전문 모델(교사)을 각각 강화학습한 뒤, 하나의 모델(학생)이 여러 교사의 출력 결과를 학습해 각 분야의 능력을 통합하는 '증류' 방식이 활용
- '26년 주요 AI 모델 개발에 이 방식이 활용됐으며, 교사는 학생과 비슷한 크기라도 특정 분야에 특화된 학습을 통해 높은 성능을 확보할 수 있는 것으로 분석 (Hugging Face, '26.7.)

(자기 증류) 추가 학습 이전의 모델을 교사로 활용해 기존 능력 회복

- 한편, 분야별 교사 모델을 구축하지 않고 추가 학습 이전의 모델을 활용해 기존 능력을 유지하려는 자기 증류 방식도 등장
- 이는 새로운 분야를 학습하기 이전의 모델을 교사로 활용해 기존의 출력 방식을 다시 익히도록 하는 '자기 증류'로, 추가 학습 과정에서 저하된 능력을 회복하는 방법이 등장
- 美 AI 기업 Thinking Machines는 새 분야의 데이터로 추가 학습한 모델에 이 방식을 적용해, 새로 익힌 지식은 유지하면서 추가 학습 과정에서 저하된 기존 능력을 회복 (Hugging Face, '26.7.)

(4) 추론: 문제 난이도에 맞춰 생각하는 양을 조절하고, 어려운 문제는 푸는 도중에 추가 학습

(사고 강도 조절) 쉬운 질문에는 적게, 어려운 문제에는 많이 생각하도록 추론 연산 배분

- 기존에는 질문 난이도에 따라 추론 연산량을 세밀하게 조절하기 어려워, 쉬운 문제에 불필요한 연산이 투입되거나 복잡한 문제에 충분한 연산을 배분하지 못하는 비효율 발생
- 이에 추론 과정에 투입할 연산량을 '사고 강도' 설정으로 조절해, 간단한 질문에는 연산을 줄이고 복잡한 문제에는 더 많은 연산을 배분하는 방식이 확산

  * 사고 강도: 답하기 전에 모델이 얼마나 오래 생각할지(추론에 쓸 연산량)를 낮음·중간·높음 등 단계로 정하는 설정

- 앤트로픽은 Sonnet 5.5에 사고 강도 조절 기능을 적용했으며, 일부 성능 평가에서 이전 모델의 최고 점수를 약 10분의 1 비용으로 넘어섰다고 발표 (Anthropic, '26.9.)

(테스트타임 학습) 문제 해결 과정에서 모델을 추가 학습시켜 새로운 해법 탐색

- 추론 연산을 늘리면 다양한 해법을 탐색할 수 있지만, 모델의 학습된 능력 자체는 바뀌지 않아 새로운 해법이 필요한 과학·공학 난제 해결에는 한계
- 이에 문제 해결 중 다양한 해법을 시도하고 그 결과를 활용해 모델을 추가 학습시킴으로써, 해당 문제에 더 적합한 해법을 찾아가는 '테스트타임 학습(TTT)'이 등장

  * 테스트타임 학습(TTT): 문제를 푸는 시점에 그 문제만을 위해 추가 학습하는 방식

- TTT-Discover 연구진은 공개 AI 모델에 테스트타임 학습을 적용해, 문제당 수백 달러의 비용으로 수학·알고리즘 설계·생물학 분야의 여러 난제에서 기존 최고 성과를 경신 (더밀크, '26.10.)

표 스마트 스케일링 기술별 개선 내용과 추가 비용

구분	개선 내용	추가 비용·부담	적합한 분야
데이터	데이터 품질·다양성 개선으로 학습 효율 향상	데이터 선별·합성 연산, 전문가 검수	데이터가 부족한 소수 언어·전문 분야
학습 시스템	대형 학습 실패 감소, GPU 활용률 향상	장비·소프트웨어별 맞춤 구현, 전문 인력·인프라	대형 모델을 직접 학습하는 기업
사후학습	업무 수행 능력 강화, 기존 능력 유지	분야별 교사 모델·학습 환경 구축, 채점 기준 검증	결과 검증이 가능한 코딩·수학·업무 과제
추론	난이도별 연산 배분, 난제의 해법 탐색	적정 연산 수준 설정, 난제당 높은 추론·학습 비용	난이도 차이가 큰 업무, 검증 가능한 과학·공학 난제

자료 : 참고자료 종합

03 (지속 학습) 자원 활용 효율화가 배포 이후로 확대되며, 업무 경험을 성능 개선에 활용

서비스 운영 중 축적되는 업무 경험이 배포 이후 성능을 높이는 새로운 학습 자원으로 주목

- 배포된 모델의 지식은 배포 시점에 고정되고 테스트타임 학습으로 얻은 개선도 해당 문제에만 적용되어, 새로운 업무나 바뀐 환경에 맞추려면 모델을 다시 학습시켜야 하는 한계
- 이에 서비스 운영 중 축적되는 작업 기록과 사용자 수정 내역, 실패 사례를 후속 업무에 활용해, 배포 이후에도 성능을 개선하는 '지속 학습'이 새로운 대안으로 주목
- 지속 학습은 업무 경험을 외부 메모리·실행 틀(하네스)·모델 가중치 중 어디에 반영하느냐에 따라 구분되며, 모델 자체를 변경하는 방식일수록 기존 능력에 영향을 미칠 위험이 증가 (AWS, '26.7.)

  * 하네스(Harness): 지시문·도구·작업 순서 등 모델 바깥에서 AI의 작업 수행을 이끄는 실행 틀

(메모리) 업무 중 확인한 정보와 처리 절차를 모델 밖에 저장해 다음 업무에 재사용

- 모델은 그대로 둔 채 업무 중 확인한 정보와 처리 절차를 외부 저장소에 기록해 다음 업무에 다시 쓰는 방식으로, 지속 학습 가운데 적용 부담이 가장 적어 실제 서비스에도 도입되는 추세

- 다만 한 번의 작업에서 얻은 정보는 틀렸거나 시간이 지나 바뀔 수 있어, 메모리에 저장하기 전에 실제 환경에 비추어 내용의 정확성과 최신성을 검증하는 과정이 중요
- 마이크로소프트는 실제 데이터베이스 조회를 통해 메모리의 정확성을 검증하는 방식을 적용해, 실험에서 과제 통과율을 39%에서 73%로 높이고 과제당 비용을 절반으로 절감 (Microsoft, '26.9.)

(하네스 개선) 실패 기록을 분석해 지시문·도구·작업 순서 등 업무 수행 방식을 수정

- 정보를 저장하는 것만으로는 잘못된 업무 처리 방식이 개선되지 않아, 실패 기록을 분석해 지시문·도구 선택·작업 순서 등 하네스를 수정하는 방식이 활용
- 에이전트의 작업 실패는 모델 자체의 성능뿐 아니라 부적절한 지시문이나 도구 선택에서도 발생해, 모델을 재학습하지 않고 하네스를 개선하는 것만으로도 성능을 높일 가능성
- GEPA 연구에서는 실행 기록을 바탕으로 지시문을 개선한 결과, 비교 대상 강화학습 방식보다 성능이 약 6% 높았으며 평가에 필요한 시도 횟수는 약 35분의 1 수준으로 감소 (AWS, '26.7.)

  * 이렇게 개선한 작업 절차는 업무별 '스킬' 문서로 정리해 두면 새로운 업무에도 재사용 가능

(가중치 갱신) 업무 경험을 모델 내부에 직접 학습시켜 새로운 능력을 내재화

- 외부 메모리나 하네스 개선만으로는 새로운 지식과 능력을 모델 자체에 반영하기 어려워, 축적된 업무 경험을 활용해 모델 가중치를 추가 학습하는 방식도 활용

  * 가중치: 학습으로 정해지는 모델 내부 수치(파라미터)의 값으로, 모델이 아는 내용과 행동 방식을 결정

- AWS는 가중치 변경에 따른 기존 능력 저하 위험을 고려해, 메모리와 하네스에서 충분히 검증된 경험만 골라 주기적으로 모델에 반영하는 단계적 접근을 권고 (AWS, '26.7.)
- 구글은 사용 중 축적한 정보를 별도 학습 단계에서 추가 파라미터에 반영하는 '수면 프레임워크'를 제시하고, 새로운 언어를 순차 학습하는 실험에서 기존 성능이 거의 유지됨을 확인

  * 수면 프레임워크: 사람이 잠자는 동안 기억을 정리하듯, 서비스 중 쌓인 정보를 따로 모아 두었다가 별도 단계에서 모델에 학습시키는 방식 (Google Research, '26.10.)

표 지속 학습 방식별 변경 대상과 운영 특성 비교

구분	바뀌는 대상	반영 주기	구현 비용	오류 발생 시 영향·복구
메모리	외부 저장소의 기록	작업마다	낮음	해당 메모리 이용 업무에 영향, 기록 수정·삭제로 복구
하네스 개선	지시문·도구·작업 순서	배포 단위	중간	해당 하네스 이용 업무에 영향, 이전 버전으로 복구
가중치 갱신	모델 가중치	주기적	높음	변경 모델 이용 업무에 영향, 이전 모델로 복구

자료 : AWS('26.7.) 참고하여 재구성

04 (실효성) 실제 업무 성능·전체 비용·배포 후 개선 여부로 스마트 스케일링의 성과 판단

(실제 성능) 연구에서 확인된 효과가 실제 개발 환경과 업무에서도 재현되는지 검증

- 스마트 스케일링의 성과는 특정 모델 크기·데이터·평가 과제에서 확인된 결과가 많아, 실제 개발 환경에서도 같은 수준의 효과가 나타나는지 추가 검증할 필요
- 실제로 소형 모델에서 효과적이었던 데이터 처리 방식이 모델 규모가 커지면 다른 방식보다 성능이 낮아지는 등, 모델 규모와 자원 조건에 따라 효과가 달라지는 것으로 확인 (Meta 외, '26.3.)
- 또한 높은 평가 점수가 실제 업무 성능을 보장하지는 않으며, 작은 점수 차이를 위해 과도한 연산이 투입될 수도 있어 실제 업무에서 목표 성능을 달성하는지 검증할 필요

(전체 비용) 이용 단가가 아닌 업무 하나를 완수하는 데 드는 전체 비용으로 효율 판단

- AI 서비스는 운영 기간이 길어질수록 반복적인 추론 비용이 초기 학습 비용을 넘어설 수 있어, 모델 개발·학습부터 실제 서비스 운영까지 전체 비용을 고려할 필요 (Springer, '26.10.)
- 같은 성능을 내는 데 드는 추론 비용은 매년 약 13분의 1로 낮아지고 있으나, 에이전트가 AI를 여러 번 호출해 처리하는 업무 하나의 비용은 '28년까지 5배 넘게 늘어날 전망 (Epoch AI, '26.9.)

  * 가트너는 단가 하락이 더 비싼 고성능 모델과 복잡한 에이전트 활용을 늘려 총비용을 키우는 현상을 '추론 역설'로 명명 (Gartner, '26.8.)

- 독일 SAP는 업무에 필요한 성능을 충족하는 저비용 AI 모델로 일부 고가 모델을 대체하는 등 비용 최적화를 추진하면서도, 전체 AI 투자 규모는 확대할 것으로 전망 (Reuters, '26.10.)

(배포 후 개선) 운영 중 성능이 실제로 오르면서도 기존 성능은 유지되는지 확인

- 지속 학습은 축적된 경험이 실제 성능 개선으로 이어지는지가 중요하나, 에이전트가 별도 설명 없이 업무 경험만으로 새로운 규칙을 익히는 능력은 아직 제한적 (CMU 외, '26.10.)

  * CMU·아마존 연구진의 비교 실험에서 업무 규칙을 알려 준 에이전트의 평균 정답률은 95.4%였으나, 경험만으로 익혀야 하는 가장 어려운 과제에서는 최고 59.6%, 중앙값 22%에 그침

- 또한 새로운 경험을 반영하는 과정에서 기존 업무의 성능이 저하될 수 있어, 변경 후 기존 업무를 다시 시험하는 '회귀 평가'를 통해 성능 유지 여부를 확인할 필요
- 아울러 한 번의 성능 시험만으로는 운영 중 개선 여부를 알 수 없어, 새로운 업무나 바뀐 규칙에 얼마나 빨리 적응하는지를 일정 기간에 걸쳐 따로 측정할 필요

시사점

자원 배분과 지속 개선 역량 중심의 국내 AI 대응 방향

국내 기업·기관은 보유 자원·데이터에 맞춰 직접 개발과 외부 모델 활용 구분

- 스마트 스케일링은 AI 경쟁의 기준을 연산 보유량에서 같은 자원으로 얻는 성능과 배포 후 개선 속도까지 넓히는 변화로, 연산 격차를 자원 활용 방식으로 일부 보완할 여지를 제시
- 선도 기업도 효율화 기술을 대형 모델 개발에 함께 활용해 범용 대형 모델의 격차는 좁히기 어려우나, 정답을 검증할 수 있는 과업에서는 데이터를 정교하게 다듬은 특화 모델로 경쟁할 기회 존재
- 전문 데이터가 부족한 업무는 외부 모델을 과업에 맞게 골라 쓰는 편이 현실적인 만큼, 보유한 자원과 데이터를 따져 직접 개발할 영역과 외부 모델을 활용할 영역을 먼저 구분할 필요

AI 도입 기업은 과업별 총비용으로 모델을 운용하고 학습 자산을 축적

- 기업은 AI 성과를 벤치마크 점수가 아닌 업무당 총비용과 품질로 측정하고, 업무 난이도에 맞춰 모델과 사고 강도를 정해 고비용 모델은 필요한 업무에만 배정할 필요
- 특화 모델을 개발할 때는 현업 전문가가 정답 검수에 직접 참여해, AI가 찾아낸 오류 후보를 확인하고 학습 데이터에 담긴 정답의 정확도를 높이는 것이 중요
- 지속 학습은 되돌리기 쉬운 메모리·하네스부터 적용해 검증된 내용만 가중치에 반영하고, 업무 기록과 평가셋은 모델을 바꿔도 쓸 수 있게 직접 관리하는 편이 바람직

정부는 성과 중심으로 연산을 배분하고 데이터·평가 기반을 확충

- 정부는 연산 자원을 지원할 때 △과제별 목표 △필요 연산량 △데이터 확보 수준 △성과 확인 방법을 따져 배분하고, 지원 후 성과를 점검하는 기준을 마련할 필요
- 한국어·산업 분야처럼 데이터를 구하기 어려운 영역일수록 정제와 검수가 성과를 좌우하는 만큼, 분야별 데이터 정제·검수 역량 확보와 데이터 공동 구축을 지원할 필요
- 공공 지원사업과 조달 평가에서도 벤치마크 순위에 더해 △목표 성능까지의 총비용 △배포 후 개선 속도 △기존 성능 유지 여부를 평가 항목에 반영하는 방안을 검토

참고자료

1) Discover Artificial Intelligence(Springer), 'Energy constraints in large scale artificial intelligence and insights from human cognition for efficient large language models', (2026.10.12.) https://link.springer.com/article/10.1007/s44163-026-02369-8
2) Air Street Press, 'The State of AI Report 2026', (2026.10.08.) https://press.airstreet.com/p/state-of-ai-2026
3) OECD, 'Data Spaces between Co-ordination Challenges, Financial Viability and Public Policy', (2026.10.20.) https://www.oecd.org/en/publications/data-spaces-between-co-ordination-challenges-financial-viability-and-public-policy_0066f11b-en.html
4) 더밀크, '알트만은 틀렸다, 이제는 스마트 스케일링 시대: 최예진 교수의 소버린 AI 해법', (2026.10.15.) https://www.themiilk.com/articles/a67647cdc
5) MLCommons, 'MLPerf Training v6.1: First LLM Post-Training Benchmark', (2026.09.24.) https://mlcommons.org/2026/09/mlperf-training-llm-post-training/
6) Meta FAIR 외, 'How Text Quality Interventions Reshape Neural Scaling Laws for LLMs: Empirical Study', ICLR 2026, (2026.03.) https://openreview.net/pdf/5411d7dab79df39e9b10b42a603387f16249a2ab.pdf
7) Microsoft AI, 'Frontier Pre-training at Microsoft AI', (2026.10.15.) https://microsoft.ai/news/frontier-pre-training-at-microsoft-ai/
8) NVIDIA, 'Synthetic Data Generation for Financial AI Research with NVIDIA NeMo', (2026.07.09.) https://developer.nvidia.com/blog/synthetic-data-generation-for-financial-ai-research-with-nvidia-nemo/
9) NVIDIA, 'NVIDIA Nemotron 3 Ultra Powers Faster, More Efficient Reasoning for Long-Running Agents', (2026.06.04.) https://developer.nvidia.com/blog/nvidia-nemotron-3-ultra-powers-faster-more-efficient-reasoning-for-long-running-agents/
10) Hugging Face, 'Distillation in 2026 (so far): which frontier models use it and how', (2026.07.08.) https://huggingface.co/blog/sergiopaniego/distillation-2026
11) Hugging Face, 'How frontier models train on outcomes in 2026', (2026.08.10.) https://huggingface.co/blog/sergiopaniego/agentic-rl-2026
12) Thinking Machines, 'Putting Task Expertise into RL Achieves State-of-the-Art Performance on Text-to-SQL', (2026.08.27.) https://thinkingmachines.ai/news/putting-task-expertise-into-rl/
13) Anthropic, 'Introducing Claude Sonnet 5.5', (2026.09.28.) https://www.anthropic.com/claude-sonnet-5-5
14) Google Research, 'The sleep framework: A new way for LLMs to retain knowledge in continual learning', (2026.10.27.) https://research.google/blog/the-sleep-framework-a-new-way-for-llms-to-retain-knowledge-in-continual-learning/
15) AWS Builder Center, 'Continual Learning: Building Agents That Improve Through Experience', (2026.07.22.) https://builder.aws.com/content/3Gol5xmJgmcZI8qDNC7Z485VI0R/continual-learning-building-agents-that-improve-through-experience
16) Microsoft, 'Grounding Agent Memory: Environment-Probing Curation for Enterprise Agents', arXiv, (2026.09.10.) https://arxiv.org/html/2609.11060v1
17) CMU·Amazon, 'ServeLearnBench: How Well Can Agents Self-Improve from Serving Experience?', arXiv, (2026.10.06.) https://arxiv.org/html/2610.07792v1
18) Epoch AI, 'The plunging price of thought', (2026.09.22.), https://epoch.ai/publications/the-plunging-price-of-thought
19) Gartner, 'Gartner Predicts AI Inference Costs Per Agentic Workflow Will Increase More Than Fivefold Through 2028', (2026.08.17.) https://www.gartner.com/en/newsroom/press-releases/2026-08-17-gartner-predicts-ai-inference-costs-per-agentic-workflow-will-increase-more-than-fivefold-through-2028
20) Reuters, 'SAP swaps costly AI models for cheaper alternatives, targets over 2 bln euros in gains', (2026.10.28.) https://www.reuters.com/world/china/sap-swaps-costly-ai-models-cheaper-alternatives-targets-over-2-bln-gains-2026-10-28/

Ⅱ. 주요국 동향

1 프랑스·체코 AI 전략 협정 체결, 기가팩토리 유치 공조 추진

EU 기술주권 확보를 위한 AI 인프라 확충, 프랑스·체코 전략 협정 체결

양국, 유럽 AI 주권 확보 위한 전략 협정 체결 및 기가팩토리 유치 공조

- 프랑스 AI·디지털 담당 장관과 체코 제1부총리 겸 산업통상부 장관이 AI·디지털 분야 전략 협력 협정에 서명('26.10.21.), '주권적·경쟁력 있는·개방형 유럽 AI' 구축 목표 제시
- 협정은 EU AI 기가팩토리 공모에 대한 양국 공조를 주요 목표로 제시했으며, 같은 날 체코 CRA와 프랑스 컴퓨팅 기업 Bull도 별도 MOU를 체결해 체코 측 유치 사업 협력 추진

  * AI 기가팩토리 : 최첨단 AI 모델의 학습·운영을 위한 초대형 연산 시설로, EU가 지역 거점으로 운영 중인 기존 AI 팩토리보다 3~4배 이상 많은 최신 AI 프로세서를 갖춘 차세대 AI 인프라

- 체코는 유치 대상 기가팩토리를 프라하 AI 선언에 참여한 중·동유럽 국가의 기업에도 개방해, 자국 수요 중심의 시설을 넘어 지역 AI 거점으로 키운다는 구상을 제시

  * 프라하 AI 선언 : '26.9.3 체코·슬로바키아·폴란드·크로아티아·헝가리 등 중·동유럽 9개국이 체결한 협력 선언으로, EU AI 정책 공조와 연산 인프라 연계, 모범사례 공유 등을 약속 (Bull, '26.10., Prague Daily News, '26.9.)

EU AI 연산 역량 열세 속 기가팩토리 추진, 회원국 간 공조 중요성 확대

- EU의 AI 연산 인프라 용량은 '26년 세계의 약 5%로 미국(78%)·중국(11%)에 크게 뒤처지며, 유럽 클라우드 인프라 시장의 약 70%를 미국 3대 기업이 차지하는 등 역외 기술·서비스 의존도가 높은 상황

  * AI 연산 인프라 용량 비중은 Bruegel의 2026년 추정치로, 주요 AI 데이터센터의 전력용량을 기준으로 산정하며 일부 기존 클라우드 시설의 연산 자원은 제외 (Bruegel, '26.9.)

- EU는 역내 연산 역량 확충을 위해 '25년 InvestAI로 기가팩토리 구축에 200억 유로 투자 동원 계획을 발표, '26년 7월 EuroHPC 주관 공모를 개시해 최대 7곳 구축 추진
- 기가팩토리는 민감 데이터 보호와 역외 의존 완화를 노리는 기술주권의 핵심 시설로, EU 지원과 회원국 예산이 함께 투입되는 만큼 유치 과정에서 회원국 간 지지 확보가 중요

표 프랑스·체코 협정 체결까지의 주요 흐름

시점	주요 내용
2025년 2월	EU, InvestAI 발표 및 AI 기가팩토리 기금 200억 유로 조성 계획
2025년 4월	EU, AI 대륙 행동계획 발표 및 기가팩토리 사전 참여의향 조사 개시
2025년 6월	체코 CRA, AIGF CZ 유치 구상 제출 및 정부 지지 확보
2026년 7월	EU 공동사업단(EuroHPC), AI 기가팩토리 공식 공모 개시(최대 7곳 선정)
2026년 10월	프랑스·체코 정부 간 협정 및 CRA-Bull 기업 간 MOU 체결

자료 : EU 집행위('25.2.11·'25.4.9·'26.7.30), EuroHPC('25.4.9), Lupa.cz('25.6.20), 프랑스·체코 정부 보도자료('26.10.) 종합

프랑스·체코 협정 기반 AI 기가팩토리 유치 공조 및 산업 협력 구체화

(1) 프랑스·체코 협정, AI 기가팩토리 유치 공조와 유럽 AI 생태계 협력 기반 마련

양국, 상대국 기가팩토리 사업 지지 및 상시 협력체계 구축

- 양국 협정은 정부 간 협력 강화·소버린 AI 개발·EU 기가팩토리 공조·상설 협력 채널 구축 등 4대 우선순위를 설정하고, 이 중 기가팩토리 공조는 공모 단계의 이행 방식까지 명시
- 양국은 EuroHPC 공모에서 상대국 영토의 기가팩토리 사업까지 지지하고 EU 내 입장도 함께 조율하기로 했으며, 연산 자원을 공공·국가안보·전략연구와 기업 수요에 우선 공유하는 방향을 발표
- 양국 AI 팩토리 간 협력은 기가팩토리 선정 결과와 관계없이 추진하기로 했으며, 체코와 프랑스가 각각 상시 연락관을 지정해 전문가 협의 및 후속 협력 과제 발굴을 지원하도록 규정

* AI 팩토리 : 기가팩토리보다 작은 규모로 EuroHPC 슈퍼컴 기반 스타트업·중소기업·연구기관에 AI 연산과 지원 서비스를 제공하는 EU 지역 거점으로 프랑스 AI2F와 체코 CZAI를 포함해 현재 19곳 운영(Bruegel, '26.9.)

AI 공급망·인재 양성·공공 활용을 포괄하는 10대 협력 의제 구체화

- 협정은 AI 생태계 전반의 10대 협력 분야를 규정해, 기가팩토리 유치 공조 외에도 인프라 구축에 필요한 공급망과 이를 활용할 인재·공공 수요까지 협력 범위에 포함한 점이 특징

* 10대 협력 분야는 ① AI 교육·역량 ② 신뢰 AI ③ 공공·경제 도입 ④ AI·HPC 인프라 ⑤ 고등교육·R&D ⑥ AI 공급망 ⑦ 기술주권 ⑧ EU 협상 공조 ⑨ EU·국제 협의 ⑩ 산업·상업 협력으로 구성(프랑스·체코 협정문, '26.10)

- 공급 측면에서는 첨단 반도체·컴퓨팅 하드웨어에 대한 접근성 확보, 유럽 내 AI 시스템 제조·통합, 공급망 회복력·보안 강화 등을 명시, 중소기업·스타트업의 공급망 참여 지원도 약속
- 수요 측면에서는 국민 대상 AI 리터러시와 공무원 교육, 공공 서비스의 AI 활용 사례 공유, 중소기업 AI 도입 지원 등을 포함하고 EU AI 법 이행 과정에서 양국의 규제 샌드박스 운영 경험도 교환하기로 합의

(2) CRA-Bull MOU로 체코 기가팩토리 사업 구상 제시, 공모 일정·인프라 여건은 과제

CRA-Bull MOU, 인프라·AI 시스템 역량 결합을 통한 체코 기가팩토리 유치·구축 추진

- 양국 장관은 협정 서명 자리에서 CRA-Bull MOU를 양자 협력의 첫 구체 성과로 함께 발표, 두 기업도 이를 정부 협정에 따른 첫 산업 협정으로 규정(French Ministry of Economy·Bull·CRA, '26.10.)
- MOU는 CRA의 데이터센터·네트워크와 Bull의 AI 서버·SW를 결합해 중·동유럽에 소버린 AI 서비스를 제공하는 구상으로, CRA는 Bull을 서버 주 공급사·시스템 통합사로 검토
- CRA는 Bull을 AI 서버 주 공급사 및 시스템 통합사로 검토 중이며, 양사는 인프라 운영과 AI 연산 시스템 공급을 중심으로 역할을 분담할 계획

- Bull은 자체 AI 플랫폼(BullSequana AI) 공급과 Foxconn 협력을 통한 NVIDIA 차세대 칩 기반 서버의 유럽 역내 생산을 추진하며, 향후 유럽산 반도체 상용화 시 서버에 통합할 방침(Bull·Lupa.cz, '26.10.)

공모 일정 지연과 인허가·전력·자금조달 부담, 사업화 변수로 부상

- 양국 협정은 법적·재정적 구속력이 없으며, CRA-Bull MOU 역시 비구속·비독점 합의로 구체적인 협력 범위와 기술·상업 조건은 후속 협상 및 별도 계획을 통해 확정할 예정
- EuroHPC는 회원국 예산 절차에 필요한 시간을 확보하기 위해 체코가 응모한 중형(Lot 1)의 공모 마감을 '27년 2월로 연기한 반면, 대형(Lot 2)은 '26년 12월 마감을 유지

  * 사업 선정 후 계약을 체결하면 최대 18개월 이내 시설 가동을 목표로 하는 일정(EuroHPC, '26.7./EC, '26.7.)

- 공모 일정이 지연되는 가운데 CRA 측은 사업 추진의 예측 가능성과 명확한 규칙 마련을 요구, 선정 이후에도 인허가·전력망 접속·자금조달 등이 사업 실행의 주요 변수로 지목(Bruegel·Lupa.cz, '26.9.)

시사점

EU 기가팩토리 유치, 회원국 간 공조와 기업 협력을 결합하는 방향으로 확대

회원국 간 연합을 통한 기가팩토리 유치 및 EU 정책 공조 강화

- 프랑스·체코 협정은 기가팩토리 유치를 단독 응모에서 상대국 사업 지지와 EU 내 입장 조율을 결합한 회원국 간 공조로 넓힌 사례로, 상시 연락관을 통해 EU 정책 협상까지 공조 범위 확대
- 체코가 프라하 AI 선언을 바탕으로 기가팩토리를 중·동유럽 기업에도 개방하는 구상을 제시하는 등, 유치 경쟁에서 주변국과의 연계를 앞세운 지역 거점 전략도 함께 등장하는 모습

기업 간 역할 분담 중심 사업 구조와 후속 계약·인프라 구축 과제

- CRA-Bull 사례처럼 기가팩토리 사업은 데이터센터 운영사와 AI 서버·시스템 기업이 역할을 나누는 구조로 추진되나, 실제 사업화는 후속 계약과 인허가·전력망·자금조달 해결에 좌우
- 국내 AI 서버·데이터센터 기업은 EU 기가팩토리 참여를 검토할 때, 단독 공급보다 역내 운영사·시스템 통합사가 주도하는 사업 구조 안에서의 역할을 중심으로 진입 경로를 판단할 필요

참고자료

1) Ministère de l'Économie, des Finances et de la Souveraineté industrielle, énergétique et numérique, La France et la Républiq ue tchèque signent un accord stratégique sur l'IA et le numérique au service de la souveraineté européenne (2026.10.20.) https://presse.economie.gouv.fr/la-france-et-la-republique-tcheque-signent-un-accord-strategique-sur-lia-et-le-numerique-au-service-de-la-souverainete-europeenne/
2) Ministry of Industry and Trade of the Czech Republic, France and the Czech Republic sign a strategic agreement on AI and digital technology to serve European sovereignty (2026.10.22.) https://mpo.gov.cz/en/guidepost/for-the-media/press-releases/-france-and-the-czech-republic-sign-a-strategic-agreement-on-ai-and-digital-technology-to-serve-european-sovereignty--295089/
3) Bull, Bull and CRA join forces to build the Czech AI Gigafactory for Central and Eastern Europe (2026.10.21.) https://www.bull.com/en/press-releases/bull-and-cra-join-forces-to-build-the-czech-ai-gigafactory-for-central-and-eastern-europe
4) České Radiokomunikace, České Radiokomunikace a Bull podepsaly strategické memorandum o budoucí spolupráci (2026.10.22.) https://www.cra.cz/tiskove-centrum/datova-centra/ceske-radiokomunikace-a-bull-podepsaly-strategicke-memorandum-o-budouci-spolupraci
5) European Commission, EU launches AI Gigafactories call to boost Europe's computing capacity (2026.7.30.) https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1708
6) Bruegel, How can Europe address its pressing AI compute infrastructure shortfall? (Policy Brief 18/2026) (2026.9.10.) https://www.bruegel.org/policy-brief/how-can-europe-address-its-pressing-ai-compute-infrastructure-shortfall
7) Lupa.cz, Obří české AI datacentrum bude používat servery vyrobené v Pardubicích a software z Prahy a Brna (2026.10.22.) https://www.lupa.cz/aktuality/obri-ceske-ai-datacentrum-bude-pouzivat-servery-vyrobene-v-pardubicich-a-software-z-prahy-a-brna/
8) Lupa.cz, Sliby vs. realita: Evropa chce dohnat USA v AI, ale gigatovárnu pro ČR dál brzdí byrokracie (2026.10.28.) https://www.lupa.cz/aktuality/sliby-vs-realita-evropa-chce-dohnat-usa-v-ai-ale-gigatovarnu-pro-cr-dal-brzdi-byrokracie/
9) Prague Daily News, Nine European Countries Agree Closer AI Cooperation (2026.9.4.) https://www.praguedaily.news/2026/09/04/cee-ai-summit-in-prague-nine-european-countries-agree-closer-cooperation-on-artificial-intelligence/

2 중국, 「국가 정보화 15차 5개년 계획」 발표

AI 육성과 규제 정비를 한 문서에 담은 中 5개년 정보화 구상

中 최고 네트워크 정책기구, 2030년을 향한 국가 정보화 설계도 제시

- 중국 공산당 직속 사이버보안 정책 총괄 기구인 중앙사이버보안·정보화위원회가 「국가 정보화 15차 5개년 계획」을 발표, 각 지역·부문의 디지털 전환과 정보화 발전을 이끄는 행동지침으로 자리매김 ('26.10.23.)
- '3대 체계 구축·5개 수준 제고·2개 기반 강화'의 총괄 구도 아래 10대 중점 임무와 7대 전문 프로젝트를 배치, 인프라·과학기술·산업부터 민생·안보·법제까지 망라한 실행 계획 성격
- △AI 분야 기초 입법 △생성형 AI 등록제 개선 △에이전트 등 신형 AI 서비스의 평가·등록 체계 정비를 공식 과제로 명시, 기술 육성과 규제 정비를 동시에 추진하는 구조가 특징

2030년까지 에이전트 보급률 90% 확대 등 목표 제시

- 이번 계획은 2030년 목표로 6G 상용화의 적기 개시와 국가 블록체인망 기본 구축을 제시, 12대 주요지표는 시장 주도 달성을 유도하는 '예상 목표치'로 설정돼 투자 방향을 알리는 신호 역할 수행
- 특히 AI 에이전트 등 신세대 지능형 애플리케이션 보급률을 '25년 40%에서 '30년 90% 이상으로 설정, 12대 지표 중 상승 폭이 가장 커 5년 안에 두 배 이상의 보급 확대를 겨냥 (CAC, '26.10.)
- 에이전트 보급률이 인터넷 보급률·5G 기지국 수와 같은 반열의 국가 관리 지표로 편입, AI 에이전트를 사회 기반시설처럼 다루겠다는 선언으로 이번 계획의 차별 지점

기본법 제정·등록제 확장·데이터 질서 재편, 中 AI 제도화의 골격

(1) (AI 기본법 제정) '기초 입법' 과제로 격상된 AI 법제화

AI, 디지털경제·통신과 나란히 '기초 입법' 과제로 격상

- 계획 말미의 법제 정비 장은 AI를 디지털경제와 통신, 사이버범죄 및 사이버폭력 대응과 나란히 '기초 입법' 추진 분야로 명시, 개별 규정 중심이던 AI 규율을 기본법 차원으로 격상
- 자율주행·저공경제 등 신흥 분야의 입법 연구와 함께, 사이버보안·데이터보안·개인정보보호·정보화 발전 등 기존 법규 체계 정비와 새로운 기술·서비스 형태에 맞춘 제도 공급 확대까지 담은 구성
- AI 기본법이 제정되면 흩어져 있던 기술별 규정을 아우르는 상위 틀이 생겨, 중국 내 AI 사업을 하는 기업 입장에서는 법적 예측 가능성과 준수 부담이 동시에 커지는 이중 효과 전망

입법에서 사법해석까지, 공백 없이 움직이는 AI 법제 정비

- AI와 개인정보보호, 인터넷 공간 질서 같은 쟁점에 대해서는 사법해석을 연구해 발표하겠다는 계획을 함께 명시, 법 제정을 기다리지 않고 법원의 판단 기준부터 먼저 제시하는 구도

* 사법해석(司法解释) : 최고인민법원 등이 법률 적용 기준을 구체화해 내놓는 유권해석으로, 하급 법원을 구속해 사실상 규범처럼 작동. 신기술 분야에서는 입법 공백을 메우는 수단으로 활용

- 국경 밖 쟁점을 다루는 대외 네트워크 법치의 확장도 과제로 제시, 자국 법제의 역외 적용과 국제 규칙 대응까지 시야에 넣어 외국 기업과 관련된 분쟁이나 집행 기준이 구체화될 가능성
- 입법과 사법해석, 대외 법제가 하나의 계획 안에서 함께 움직여 제도 공백기를 줄이는 구성으로, AI 기본법 제정이 단발성 입법이 아니라 법제 정비 패키지의 일부임을 시사

표준으로 시장 규격 선점, 국제 규범화까지 겨냥

- 성숙 분야의 표준 체계를 다듬는 한편 AI와 에이전트 등 신흥 분야의 표준 제정을 서두르고, 양자기술과 뇌-컴퓨터 인터페이스 등 미래 기술 표준까지 미리 배치하는 단계별 로드맵 제시
- 에이전트는 식별 ID 표준과 기술성숙도 평가체계, 에이전트 간 통신 표준이 산업과 인프라 장에 나뉘어 제시돼 한 대상을 여러 갈래로 규격화, 등록제와 짝을 이뤄 사실상 인증 요건으로 작동할 전망
- 나아가 국가표준의 국제표준 전환과 다자 틀에서의 규칙 제정 참여까지 명시, 국내에서 법과 등록제 및 표준을 완성한 뒤 국제 규범으로 확산하는 수순으로 표준 주도권 경쟁 예고

(2) (에이전트 등록·평가제) 생성형 AI 등록제를 에이전트까지 확장

모델 중심 심사에서 서비스 전반 등록으로 관리 틀을 재편

- 안전보장 장은 기본법의 틀을 집행으로 잇는 장치로, 이미 운영 중인 생성형 AI 등록 제도를 손질하는 동시에 △에이전트 △온디바이스 대형모델 △의인화 서비스 등 새 유형의 평가·등록 정비를 명시

  * 등록(備案·비안) 제도 : 서비스 출시 전후에 당국에 서비스 내용과 안전조치를 신고해 명부에 올리는 제도

- 등록 대상이 모델 중심에서 서비스 형태 전반으로 넓어지는 변화로, 챗봇 심사에 맞춰져 있던 기존 관리 틀을 스스로 판단하고 작업을 수행하는 에이전트 시대에 맞게 고치는 작업
- 에이전트를 등록 대상으로 명시한 국가 단위 계획은 세계적으로 드물어, 아직 규제 공백 상태인 주요국과 달리 중국은 서비스가 확산되기 전에 관리 제도부터 갖추는 모습

수시로 조이고 푸는 'AI 민첩 거버넌스' 채택

- 중점 분야 응용에는 안전위험평가 제도를 세우고 전체 운영은 'AI 민첩 거버넌스' 기제로 꾸린다고 명시, 기술 변화 속도에 맞춰 규칙을 수시로 조정하는 방식 채택
- 대규모 모델을 활용한 보안 취약점 탐지와 방어 기술의 개발 응용, AI 암호 방호체계 구축까지 함께 명시, 공격을 막는 쪽에도 AI를 투입해 창과 방패를 같이 키우는 양면 전략
- 사전 등록과 수시 평가, 기술 방어가 한 세트로 설계된 구성으로, 앞서 본 에이전트 보급률 90% 목표 같은 확산 드라이브를 제도가 뒤에서 받치는 안전판 역할

생태 규칙을 기반으로 정부가 첫 고객이 되는 확산 설계

- 산업 생태 장에서는 에이전트 사이의 상호신뢰와 상호연결, 상호운용 규칙을 먼저 마련하고, 다중 에이전트 기술 육성과 중점 분야별 수요 지침까지 제시해 보급의 순서와 방향을 설계

- 정부 업무에 AI를 접목하는 'AI+행정(政务)' 행동에 따라 정부가 에이전트의 첫 대규모 수요처가 되고, 교육과 의료 등 민생 서비스와 도시 관리가 그 뒤를 받치는 확산 경로
- 식별 ID와 성숙도 평가가 신원 확인과 품질 검증 장치로 붙고, 에이전트를 운영하고 관리하는 '지능대리' 같은 신직업 육성까지 언급돼 규제와 육성, 고용 제도가 한 묶음으로 설계된 점이 특징

⑶ (데이터·개인정보 거버넌스) 데이터 국외 이전의 촉진과 관리를 병행

데이터 국외 이전의 선별적 개방을 허용

- 법·등록에 이어 규율의 대상은 데이터로, 계획은 안전평가·표준계약·인증 세 제도를 통해 국외 이전을 촉진하면서 규범화한다고 명시해 수년간 조여온 통제의 축을 선별적 개방으로 이동
- 전담 과제인 '데이터 국외 이전 편리화 행동'에는 업종별 준수 지침과 자유무역시험구의 네거티브 리스트 제정, 정책 해설부터 신고 지원까지 한곳에서 돕는 서비스센터 구상까지 포함
- 금지 목록에 없으면 내보낼 수 있는 네거티브 방식이 자유무역시험구를 시작으로 자리 잡으면, 중국에 진출한 외국 기업 입장에서는 데이터 반출의 예측 가능성이 높아지는 효과

등급 관리와 보호 기술로 내부 보안을 보강

- 중점 업종의 데이터를 중요도에 따라 나눠 관리하는 분류분급 체계를 강화하고, 데이터 보안에는 암호 보호를 적용해 관리의 기본 단위를 '업종과 등급'으로 정리
- 유통과 거래 단계에는 블록체인과 프라이버시 보호 컴퓨팅 같은 보호 기술을 붙이고, 보안 준수 실험 플랫폼을 세워 평가와 감사에 쓰일 도구와 소프트웨어까지 개발
- 반출의 문을 열어주는 대신 안쪽의 등급 관리와 기술 방비를 촘촘히 하는 교환 구조로, 데이터를 더 내보내되 더 정밀하게 지키는 개방과 보안의 맞교환 설계가 특징

신원 인증과 개인정보 보호로 에이전트 시대의 신뢰 기반을 구축

- 국가 온라인 신원인증 공공서비스를 보급하고 블록체인 기반 분산형 신원의 상호 인정을 넓힌다고 명시, 온라인에서 본인을 증명하는 기반을 국가 단위로 통일하는 작업
- 스마트 기기의 개인정보 보호에 대한 검사와 평가를 강화하고, 얼굴 정보 같은 생체정보의 기술 활용 관리도 함께 강조해 수집 단계부터 고삐를 조이는 구성
- 에이전트가 사람 대신 일하려면 누구의 위임인지 확인할 신원 기반이 필수라는 점에서, 앞서 본 에이전트 식별 ID와 짝을 이루는 기초 인프라 성격

시사점

中 AI 제도화가 산업 생태계에 던지는 의미와 과제

기술 경쟁을 넘어 AI 운영 규칙을 먼저 완성하려는 선언

- 이번 계획은 법과 등록, 표준, 신원 인증을 하나의 제도 묶음으로 설계했다는 점에서, AI 경쟁의 축이 기술력만이 아니라 제도를 완성하는 속도로 넓어졌음을 보여주는 문서
- 같은 대상을 국가 지표로 끌어올리는 동시에 관리 대상에 올리는 구성은 확산과 통제를 분리하지 않는 중국식 운영의 전형으로, 향후 중국 AI 시장의 규칙은 이 문서를 기준으로 짜일 전망
- 국내에서 완성한 규칙을 국제 표준 논의로 들고 나가는 수순까지 담겨 있어, 계획의 영향 범위가 중국 내수를 넘어 각국이 수용 여부를 선택해야 하는 글로벌 규범 경쟁으로 번질 가능성

에이전트 제도 공백 점검과 표준 선점 경쟁 대응이 과제

- 중국 시장을 겨냥한 기업에는 등록과 표준 통과가 사실상의 진입 요건이 되는 만큼, 제품 기획 단계부터 제도 일정과 심사 요건을 기술 로드맵과 같은 비중의 사업 변수로 관리할 필요
- 한국의 과제는 법 제정 이후의 각론으로, 에이전트에 특화된 세부 규칙의 설계 순서를 정하고 관련 국제 표준 논의에 참여하는 시점을 앞당기는 일이 종속을 피하는 경로
- 주요국의 AI 규율이 아직 형성 중인 가운데 중국이 먼저 기준을 내놓는 구도여서, 글로벌 사업자에게는 권역별로 갈라지는 규제에 맞춰 대응 체계를 나눠 갖추는 부담이 현실화될 전망

참고자료

1) 安徽师范大学(안후이사범대학), '国家信息化"十五五"规划' (2026.10.28.)
https://imc.ahnu.edu.cn/info/1407/7502.htm
2) 中国网信网(중앙사이버보안·정보화위원회 판공실), '中央网络安全和信息化委员会印发《国家信息化"十五五"规划》' (2026.10.23.)
https://www.cac.gov.cn/2026-10/23/c_1794509044378569.htm
3) 21世纪经济报道(21세기경제보도), '详解"十五五"信息化蓝图 智能体等应用普及率超90%' (2026.10.27.)
https://m.21jingji.com/article/20261027/fae4aaef14cdf5af7156a0790061effc.html
4) 中国网信网(중앙사이버보안·정보화위원회 판공실), '国家信息化"十五五"规划' (2026.10.23.)
https://www.cac.gov.cn/2026-10/23/c_1794509028407372.htm
5) 经济参考报(경제참고보), '深读"十五五"专项规划 | 《国家信息化"十五五"规划》印发 到2030年新一代通信网全面建成' (2026.10.26.)
https://www.jjckb.cn/20261026/fa8313173734483db39a64b6d63681f8/c.html

Ⅲ. ICT 부문별 모니터링

구 분	품 목	주요 이슈	영 향
통신 & 네트워크	네트워크	**O2-노키아, 상용 5G망에 AI-RAN 적용해 주파수 효율 20% 개선** 영국 통신사 O2와 노키아가 런던 패딩턴 상용 5G 기지국에 AI 기반 RAN 기능을 적용한 영국 최초 실환경 배치를 완료하며 주파수 효율 최대 약 20% 개선 달성. 무선망에 내장된 예측형 AI가 트래픽 변화를 학습·예측해 주파수와 용량 할당을 자동 최적화하며, 추가 하드웨어 없이 기존 인프라로 구현. 에너지 소비 증가나 지연시간 악화 없이 피크 시간대 처리량이 향상되었으며, O2는 7억 파운드 규모 모바일 혁신 계획의 일환으로 확대 적용 방침. AI-RAN이 실험 단계를 넘어 상용망 성과 창출 단계로 진입함에 따라 삼성전자 등 장비사 간 AI 네이티브 네트워크 기술 경쟁 심화 전망	중립적
방송 & 전파 & 위성	디스플레이	**애플, 탠덤 OLED·터치 탑재한 신형 맥북 프로 공개** 애플이 탠덤 OLED 기반 'PROLED XDR' 디스플레이와 맥 최초 터치 기능을 적용한 신형 맥북 프로를 공개하며 노트북 디스플레이의 OLED 전환 본격화. 최대 2,000니트 밝기, 1,000만:1 명암비, 120Hz 가변 주사율, P3·Adobe RGB 색역을 지원하며 터치 센서를 패널에 내장해 지연시간 최소화. OLED의 효율적 적층 구조로 무게를 최대 0.7파운드 줄였으며 M5 프로·맥스 칩으로 대형 LLM 온디바이스 구동 지원. 프리미엄 노트북의 OLED 채택 확산으로 IT용 고부가 OLED 패널 수요 증가가 예상되나, 2,899달러부터 시작하는 높은 가격이 판매 확대의 변수로 작용	중립적
SW & 자율주행	자율주행차	**코디악 AI, 공공 고속도로용 무인 트럭 첫 상업 공급 계약 체결** 자율주행 트럭 기업 코디악 AI가 미국 중견 운송사 DTL 트랜스포트와 2028년부터 무인 트럭 75대 공급 계약을 체결하며 공공 고속도로 대상 첫 상업 계약 확보. DTL은 댈러스-엘패소, 댈러스-애틀랜타 등 주간 고속도로 노선에 투입할 예정이며, 코디악은 12월 댈러스-휴스턴 구간에서 안전요원 없는 무인 운행 개시 계획. 이번 계약은 자사 보유·운영 방식(TaaS)에서 기술 라이선스 기반 마일당 과금 방식(DaaS)으로의 전환을 의미하며, 전량 인도 시 연 1,120만 달러 이상 반복 매출 창출 추정	중립적
인공지능 & 데이터	초거대 인공지능	**구글, '제미나이 4 카본' 공개하며 AI 성능 세계 1위 탈환** 구글이 사이버보안·소프트웨어 엔지니어링 성능을 강화한 프런티어 AI 모델 '제미나이 4 카본(Carbon)'을 공개하며 주요 벤치마크 세계 1위 탈환. 독립 평가기관 아티피셜 애널리시스 지능지수 v5.0에서 59점으로 클로드 오퍼스 5.5(58점), GPT-6 아스트라(57점)를 앞섰으며, 자체 비교 17개 항목 중 15개에서 최고 성적 기록. 가격은 전작 아르곤과 동일하게 책정했으나 초기에는 사이버보안 방어 기관 대상 '페어윈드 프로그램'으로만 접근 제한. 다만 사실 정보 평가에서 환각률이 4%에서 11%로 상승하는 등 한계도 확인되어 정식 출시 시점과 API 사양 공개가 향후 변수	중립적
	초거대 인공지능	**구글 딥마인드, 실시간 오디오·장면 제어 지원 월드모델 '지니 4' 공개** 구글 딥마인드가 실시간 오디오 생성과 시뮬레이션 중 프롬프트 기반 장면 수정 기능을 갖춘 차세대 월드모델 '지니 4(Genie 4)'를 연구 프리뷰로 공개. 날씨·카메라 앵글 변경, 인물·사물 추가 등 실시간 편집이 가능하며 장면 일관성 유지 시간은 최대 2분으로 전작 대비 2배 확대. 제미나이가 작업 계획·검증을 담당하는 하이브리드 구조로 로봇 시나리오 사전 시뮬레이션, 영화 프로토타이핑 등 활용 가능성 제시. 런웨이의 후속 모델 출시, AMD의 월드랩스 인수 등 업계 경쟁과 통합이 가속화되는 가운데, 장시간 구동 시 객체 왜곡과 오디오 정확도 문제는 향후 과제로 잔존	중립적

구 분	품 목	주요 이슈	영 향
인공지능 & 데이터	엣지 컴퓨팅	**오페라, 브라우저 내 로컬 AI 실행 웹 표준 'WebAI' 공개** 브라우저 기업 오페라가 클라우드 없이 브라우저에서 AI 모델을 직접 실행하는 웹 표준 'WebAI'를 공개하고 '오페라 원'에 조기 프리뷰로 적용. 구글 젬마 등 다양한 오픈웨이트 모델을 지원하고 맥에서는 애플 온디바이스 모델과도 연동되며, 향후 오페라 GX·네온으로 확대 예정. 개발자는 AI 인프라 운영이나 요청별 비용 없이 기능을 구현하고, 사용자는 구독료 없이 개인정보를 기기 내에 유지하며 텍스트 생성·분류 등 활용 가능. AI 비서의 개인정보 접근 확대에 따른 프라이버시 우려 속에 로컬 AI 대안으로 주목되나, 기기 사양에 따른 성능 편차는 한계로 지적	중립적
콘텐츠 & 미디어	메타버스	**엑스리얼, 안드로이드 XR 기반 스마트 글래스 '아우라' 출시** 구글의 스마트 글래스 파트너 엑스리얼이 안드로이드 XR과 퀄컴 스냅드래곤 리얼리티 엘리트 기반의 유선형 스마트 글래스 '엑스리얼 아우라' 출시. 소니 마이크로 OLED(1200p·120Hz·시야각 95도)를 탑재해 화면 사양은 메타 레이밴 디스플레이를 상회하나, 무게가 95g으로 무겁고 연산용 별도 기기 휴대 필요. 구글 플레이 앱과 XR 전용 콘텐츠, 제미나이 AI 비서를 지원하며 한국 등 4개국에서 1,099달러부터 판매 개시. 메타의 시장 선점에 맞서 구글이 삼성·엑스리얼 등 파트너 중심의 안드로이드 XR 생태계 확장을 추진하는 가운데, 높은 가격과 수익성 확보가 대중화의 과제로 지목	중립적
	IT 플랫폼	**구글, 검색·제미나이에 외부 앱 연동하는 '커넥티드 앱 플랫폼' 공개** 구글이 검색·제미나이에 외부 기업 앱을 연동할 수 있는 '커넥티드 앱 플랫폼'을 공개하며 AI 서비스 생태계 개방 추진. 개발자는 MCP(모델 컨텍스트 프로토콜) 서버 또는 안드로이드 AppFunctions API로 앱을 연동하고, 전용 콘솔에서 등록·테스트 및 트래픽·오류율 모니터링 가능. 제미나이에는 어도비·펠로톤 등 50여 개 앱이 연동되어 있으며 수주 내 검색으로 확대 적용 예정. 이용 행태가 개별 앱 실행에서 AI 비서 중심으로 전환되는 흐름 속에 SEO와 유사한 '에이전틱 웹' 노출 경쟁의 출발점이 될 전망이나, 현재 미국 개발자 대상 베타에 한정되어 정식 확대 시점은 미정	중립적
반도체 & 양자	지능형 반도체	**오라클, 구글 외 하이퍼스케일러 최초로 OCI에 구글 TPU 도입** 오라클과 구글이 구글의 AI 전용 가속기 TPU를 오라클 클라우드 인프라(OCI)에 도입한다고 발표하며 구글 외 하이퍼스케일러 최초의 TPU 제공 사례 기록. OCI가 전력·냉각·네트워킹을 담당하고, 고객은 기존 GPU 등과 함께 워크로드별 성능·비용에 맞춰 TPU 선택 가능. 학습용 'TPU 8t'와 추론용 'TPU 8i'로 이원화된 8세대 TPU는 포드당 연산 성능을 전 세대 대비 약 3배 높이며 MoE·추론·에이전틱 AI 수요에 대응. 2027년부터 다년간 단계적 공급 확대 예정으로 GPU 중심 가속기 시장의 다변화와 함께 HBM 등 연관 부품 수요 확대 전망	중립적
ICT 융합기술	글로벌 패권 경쟁 및 공급망 재편	**앤듀릴-퀄컴, 미국산 부품 기반 노트북 '블루칩' 공개** 앤듀릴이 퀄컴과 협력해 부품 단위까지 미국 내 생산을 지향하는 견고형 노트북 '블루칩(BlueChip)'을 공개하고 2027년 사전 주문 개시 예정. 1세대 제품은 TSMC 애리조나 공장에서 생산된 스냅드래곤 X CPU를 탑재하며 부품의 약 90%를 미국 및 동맹국에서 조달하고, 장기적으로 디스플레이·금속·자성 소재까지 100% 미국산 전환 목표. 자국산 컴퓨터 조달 규정에 대응해 군·정부·치안 기관 대상 공급망 추적·감사가 가능한 보안 특화 제품으로 포지셔닝. 다만 디스플레이 등 일부 부품은 여전히 중국 의존 구조이며, 고사양 PC 대비 약 20% 높은 가격으로 민수 시장 확대에는 한계 존재	중립적

Ⅳ. 주요 ICT 행사 일정

■ 해외 전시회, ■ 국내 전시회

11/2	11/3	11/4	11/5	11/6
2026 중국 홍콩 핀테크 위크·스타트미업HK 전시회 [HKFTW x StartmeupHK] (ICT 융합 서비스), 중국 홍콩				
	2026 스페인 바르셀로나 스마트시티 세계 전시회 [SCEWC] (ICT 융합 서비스), 스페인 바르셀로나			
	2026 AIoT 국제전시회 [AIoT Korea] (IoT), 서울 코엑스			
		2026 로보월드 [ROBOTWORLD] (로봇), 고양 킨텍스		
11/9	11/10	11/11	11/12	11/13
2026 포르투갈 리스본 글로벌 기술 전시회 [Web Summit] (디지털 전환), 포르투갈 리스본				
2026 미국 유타 클라우드 네이티브 기술 전시회 [KubeCon + CloudNativeCon] (클라우드), 미국 유타				
	2026 독일 뮌헨 국제 전자부품·시스템 전시회 [electronica] (디바이스), 독일 뮌			
	2026 독일 뮌헨 유럽 반도체 전시회 [SEMICON Europa] (반도체), 독일 뮌헨			
11/16	11/17	11/18	11/19	11/20
2026 미국 시카고 슈퍼컴퓨팅 전시회 [SC26] (장비), 미국 시카고				
	2026 남아공 케이프타운 아프리카 기술 전시회 [Africa Tech Festival] (ICT 융합 서비스), 남아공 케이프타운			
	2026 미국 샌프란시스코 마이크로소프트 클라우드·AI 전시회 [Microsoft Ignite] (클라우드 서비스), 미국 샌프란시스코			
		2026 일본 오사카 IT 종합 전시회 [Japan IT Week] (디지털 전환), 일본 오사카		
		2026 국제 스마트제조·AI 기술전 (인공지능), 수원 컨벤션센터		
11/23	11/24	11/25	11/26	11/27
	2026 독일 뉘른베르크 스마트 생산 솔루션 전시회 [SPS] (디지털 전환), 독일 뉘른베르크			
		2026 스마트건설·안전·AI 엑스포 (디지털 전환), 고양 킨텍스		
		2026 코리아 핀테크 위크 (ICT 융합 서비스), 서울 aT센터		
			2026 중국 선전 하이테크 페어 [CHTF] (ICT 융합 서비스), 중국 선전	

AI·ICT Brief

2026-38호 (2026.11.06)



본 저작물은 정보통신기획평가원에서 작성하여 공공누리 '제2유형(출처 표시+상업적 이용금지)', 'AI 유형(인공지능 학습용)'으로 개방하였으며, 기관 홈페이지(www.iitp.kr)에서 무료로 다운로드 받으실 수 있습니다.

본 내용은 정보통신기획평가원 동향분석팀에서 수행하는 「ICT 동향분석 및 정책지원」 사업의 연구 결과물입니다.

본 내용은 작성자의 주관적인 의견으로 정보통신기획평가원의 공식적인 입장이 아님을 밝힙니다.

정보통신기획평가원
Institute of Information & Communications Technology Planning & Evaluation