기술

오픈소스 LLM으로 제조 특화 LLM 만들기

2026년 기준 기술 지도

방동걸 (주)웨이스 대표이사
2026년 8월 17일

공장 데이터는 밖으로 못 나갑니다. 공정 조건, 수율, 설비 파라미터는 그 자체가 영업기밀이라, 제조기업이 AI를 검토할 때 가장 먼저 부딪히는 요구가 "우리 서버 안에서 돌아야 한다"입니다. 이 요구가 오픈소스(정확히는 오픈웨이트) LLM 채택의 1차 동력입니다. 모델 가중치를 내려받아 내부에 두면 데이터가 외부로 나가지 않기 때문입니다. VEXPLOR로 제조 현장 안에 AI를 두는 웨이스가 살펴본 선택지를 지도처럼 펼쳐 놓았습니다.

그런데 "오픈소스 LLM으로 우리 회사 전용 모델을 만들자"는 계획은, 시작하기 전에 알아야 할 것이 몇 가지 있습니다. 2026년 8월 기준의 기술 지형을 실무 순서대로 정리했습니다.



오픈소스라고 다 같은 오픈소스가 아닙니다.

---

1. 첫 번째 확인 — "오픈소스"가 다 같지 않다

가중치를 공개했다고 다 자유롭게 쓸 수 있는 것이 아닙니다. 라이선스가 세 부류로 갈립니다.

제약 없는 부류 (Apache 2.0 / MIT) — 상업 사용·수정·재배포가 자유롭습니다. 알리바바 Qwen3(0.6B~235B, Apache 2.0), DeepSeek V3.1·R1(MIT), OpenAI가 2025년 8월에 공개한 gpt-oss(20B·120B, Apache 2.0), Mistral 최신 오픈웨이트 계열(Apache 2.0), 카카오 Kanana(Apache 2.0)가 여기 속합니다.

조건부 부류 (커뮤니티 라이선스) — Meta Llama 4는 월간 활성 사용자 7억 초과 기업에 별도 계약을 요구하고, 출력물로 다른 모델을 만들 때 파생 모델명에 Llama를 붙이는 등의 조건이 붙습니다. Google Gemma 3는 상업 사용은 허용하되 금지 사용 정책의 하류 전파 의무가 붙습니다. 대부분의 제조기업에는 실질 문제가 없지만, 법무 검토가 필요한 조항이 있다는 뜻입니다.

비상업 부류 — LG EXAONE 4.0은 한국어 성능이 뛰어나지만 연구·교육용 라이선스이고, 상업 사용은 별도 계약 경로입니다. 네이버 HyperCLOVA X SEED(0.5B~3B)는 상업 무료 허용입니다.

제품이나 사내 시스템에 임베드할 계획이라면, 벤치마크 점수보다 이 라이선스 구분이 먼저입니다. 나중에 바꾸려면 파인튜닝부터 다시 해야 합니다.

---

2. "특화"의 4가지 경로 — 무엇을 어디에 쓰나

도메인 특화라고 하면 파인튜닝부터 떠올리지만, 2026년 실무의 합의는 이렇게 요약됩니다. **"지식은 RAG로, 행동은 파인튜닝으로."**

RAG(검색 증강 생성) — 작업표준서, 설비 매뉴얼, 품질 기준처럼 자주 바뀌고 출처가 필요한 지식은 모델에 굽지 않고 검색으로 붙입니다. 문서가 갱신되면 색인만 다시 하면 되고, 답변마다 근거를 보여줄 수 있습니다.

LoRA / QLoRA (어댑터 파인튜닝) — 전체 파라미터의 0.1% 미만만 학습하는 방식으로, 현장 용어·보고서 형식·응답 구조를 가르치는 데 효율적입니다. QLoRA는 베이스 모델을 4비트로 양자화해 소비자급 GPU에서도 파인튜닝을 가능하게 했습니다. 다만 한계도 밝혀져 있습니다 — 새로운 도메인 지식을 주입하는 데는 풀 파인튜닝보다 불리하다는 연구(LoRA Learns Less and Forgets Less, 2024)가 대표적입니다. 어댑터는 "말투와 형식"에 강하고 "새 지식"에 약합니다.

추가 사전학습(Continued Pretraining) — 도메인 문서 수억 토큰을 통째로 더 학습시키는 정공법입니다. 대신 원래 알던 것을 잊는 catastrophic forgetting이 최대 위험이라, 일반 데이터를 5~20% 섞고 학습률을 원래의 5분의 1에서 10분의 1로 낮추는 것이 통용되는 완화책입니다. 비용이 커서 데이터가 충분히 쌓인 뒤의 선택지입니다.

증류(Distillation) — 큰 모델의 출력으로 작은 모델을 가르치는 경로입니다. DeepSeek-R1이 자기 출력 80만 샘플로 1.5B부터 70B까지 6종의 소형 모델을 만들어 공개한 것이 대표 사례로, 현장 배치용 소형 특화 모델을 만드는 표준 경로로 자리 잡았습니다.

---

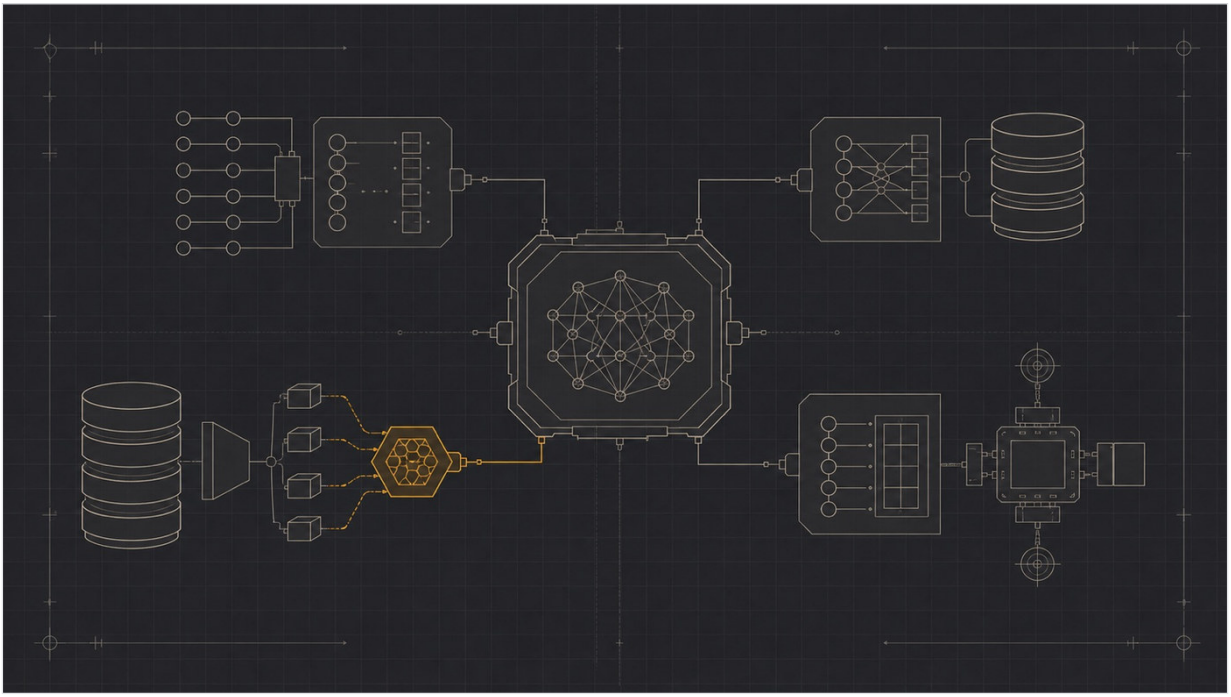
3. 제조 도메인이 유독 어려운 이유

일반 챗봇과 제조 AI의 차이는 데이터와 리스크 양쪽에 있습니다.

데이터부터 다릅니다. 센서 시계열, 정비 일지, 작업표준서, 도면, 품질 기록 — 형식이 제각각인 데이터가 섞여 있고, 같은 부품을 라인마다 다른 코드로 부르는 용어 불일치가 흔합니다. 텍스트 말뭉치 중심으로 학습된 LLM에게 시계열과 도면은 여전히 어려운 입력입니다.

리스크는 더 다릅니다. 사무용 챗봇의 환각은 틀린 문장으로 끝나지만, 공정 맥락의 환각은 그럴듯하지만 물리 제약과 모순되는 진단이 되어 설비와 안전으로 이어질 수 있습니다. 그래서 제조 특화 LLM은 "얼마나 잘 답하나"만큼 "틀린 답이 실행되지 않게 어떻게 막나"가 설계의 절반입니다.

마지막으로 배치 환경입니다. 현장에는 GPU 서버실이 없는 경우가 많아, 1~7B급 소형 모델(SLM)을 엣지 장비에서 돌리는 흐름이 뚜렷합니다. gpt-oss 20B가 16GB 메모리에서 돌고, EXAONE 1.2B나 Gemma 3 1B·4B가 이 자리를 노립니다.



모델을 바꾸는 길과 근거를 묶는 길은 비용 구조가 다릅니다.

---

4. 서빙 스택 — 돌리는 것도 기술이다

모델을 만들었으면 돌려야 합니다. 2026년 기준 통용 스택은 이렇습니다.

- **프로덕션 서빙**: vLLM 또는 SGLang. 여러 사용자의 요청을 이어 붙여 처리하는 continuous batching이 핵심이고, vLLM은 베이스 모델 하나에 어댑터 여러 개를 요청별로 갈아 끼우는 multi-LoRA 서빙을 지원합니다 — 부서별 특화 모델을 GPU 한 장으로 운영할 수 있다는 뜻입니다.
- **개발·엣지**: Ollama, llama.cpp. 혼자 쓰기엔 간편하지만 동시 사용자가 붙으면 처리량이 급락하므로 프로덕션에는 부적합합니다.
- **양자화**: GPU 서빙은 AWQ·GPTQ, CPU·엣지는 GGUF가 표준 포맷입니다. 4비트 양자화로 모델 크기를 약 4분의 1로 줄입니다.

GPU 감각은 대략 이렇습니다. 7B 모델은 4비트로 5GB 안팎이라 소비자급 GPU 한 장에 여유 있게 들어가고, 32B는 4비트로 18~~20GB라 24GB급 한 장에 빠듯하게, 70B는 4비트로 35~~40GB라 48GB급 한 장 또는 24GB 두 장이 필요합니다.

---

5. 평가와 안전 — 만드는 것보다 믿는 것이 어렵다

범용 리더보드 점수는 우리 공장에서의 성능을 말해주지 않습니다. 실무 패턴은 현장 질의를 모아 자체 평가셋을 만들고, 전문가가 작성한 채점 기준(루브릭)으로 LLM이 채점하는 방식입니다. 항공 제조 도메인에서 QA를 자동 생성해 평가한 연구가 제조 쪽 선례로 참고할 만합니다.

안전 쪽은 레이어를 겹칩니다. RAG로 근거에 묶는 것이 1차 방어선이고, 그 위에 가드레일(NVIDIA NeMo Guardrails, Meta Llama Guard, 국내는 카카오가 공개한 Kanana Safeguard)을 얹어 입력과 출력을 거릅니다.

---

6. 실제로 만든 곳들

- **SemiKong** — 세계 첫 오픈소스 반도체 특화 LLM. Llama 3.1을 베이스로 전문 서적 129권, 에칭 논문 708편, 특화 지시 데이터 5만 건(총 5억 토큰 규모)을 학습시켰고 Apache 2.0으로 공개됐습니다. 도메인 코퍼스 구성의 실물 견본입니다.
- **Siemens Industrial Copilot** — 자연어로 PLC 코드(SCL)를 생성하는 엔지니어링 보조. 2025년 Hermes Award를 받았고 티센크루프 등이 적용했습니다.
- **국내** — NC AI 컨소시엄이 제조·국방 특화 모델 VAETKI를 2025년 말 오픈소스로 공개했고(포스코DX

등 14개 기관 참여), 포스코DX는 오픈소스 파인튜닝 기반 사내 모델 P-GPT를 운영하며 제조 특화 버전을 계획하고 있습니다. 정부 독자 AI 파운데이션 모델 사업도 진행 중입니다.

7. 우리가 택한 길 — 모델을 바꾸는 길과 근거를 묶는 길

여기까지가 "모델을 특화하는" 기술이라면, 마지막으로 실무에서 마주친 관점 하나를 보태겠습니다.

저희(VEXPLOR)는 제조 특화의 무게중심을 모델이 아니라 **근거와 검증 구조**에 두었습니다. 위 2절의 합의("지식은 RAG로")를 제조 환경에 맞게 밀어붙인 형태입니다. 그래프·SQL·벡터·문서 4개 경로에 멀티홉 추론을 더한 하이브리드 GraphRAG로 답의 근거를 여러 경로에서 교차 확인하고, ERP·MES처럼 서로 말이 다른 시스템의 코드를 온톨로지(311개 이상의 시맨틱 관계)로 묶어 "어느 시스템에 물어도 같은 실체"를 보장합니다. 답변은 다경로 일치 기반으로 검증해 환각을 방지하는데, 이 중 그래프 검색 증강 구조는 특허(제10-3003470호)로 등록돼 있습니다. 그리고 제어로 이어지는 판단은 SHACL 물리 제약 검증을 거쳐 위반 명령이 실행되지 않게 막습니다.

이렇게 접근하면 베이스 모델을 갈아 끼워도 제조 지식 층이 그대로 남습니다. 오픈소스 모델의 세대교체가 반년 주기로 일어나는 지금, 특화 자산을 모델 안이 아니라 모델 밖(지식 그래프·검증 레이어)에 쌓는 것이 갈아타기 비용을 줄이는 길이라는 것이 저희 판단입니다. 물론 용어·형식 수준의 어댑터 파인튜닝과 이 구조는 배타적이지 않고, 실무에서는 결국 함께 갑니다.

08

시작 순서 제안

정리하면 이렇게 권합니다. ① 라이선스부터 확인하고(임베드 계획이 있으면 Apache 2.0/MIT 우선) ② RAG로 시작해 지식을 모델 밖에 두고 ③ 형식·용어가 아쉬우면 그때 어댑터 파인튜닝을 더하고 ④ 제어와 닿는 지점에는 반드시 검증 레이어를 세우십시오. 만드는 것보다 믿을 수 있게 만드는 것이 제조 특화 LLM의 본론입니다.

참고 자료: Qwen3 GitHub · DeepSeek-R1 GitHub · OpenAI gpt-oss 발표(2025-08) · Meta Llama 4 라이선스 · Google Gemma Terms · LG EXAONE 4.0 · 카카오 Kanana · "LoRA Learns Less and Forgets Less"(arXiv 2405.09673) · Databricks "RAG vs Fine-tuning" · SemiKong 케이스 스터디(llama.com) · Siemens 보도자료(2025) · NC AI VAETKI·포스코DX P-GPT 보도(2025-12) — 2026-08-17 확인 기준

웨이스가 하는 일이 궁금하시면: wace.me

WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.

원문 blog.vexplor.com/post/building-manufacturing-llm-open-source
발행 2026년 8월 17일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702-1708호 (영천동, 실리콘앨리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc

원문 홈페이지 링크드인