인사이트

AI는 왜 드문 불량에서 자신 있게 틀리는가

공장이 AI의 불량 판정을 쓰기 전에 확인할 4가지

방동걸 (주)웨이스 대표이사
2026년 10월 7일

국내 반도체 대기업들이 공장 운영에 AI를 본격적으로 넣겠다는 목표를 내놓았습니다. SK하이닉스는 2030년까지 자율형 팹을 구축하겠다고 밝혔고, 삼성전자는 같은 해까지 전 세계 제조 현장을 AI가 주도하는 공장으로 바꾸겠다고 발표했습니다. 이 목표가 겨냥하는 일 가운데 하나가 불량이 늘었을 때 원인을 찾고 무엇을 바꿀지 정하는 판단입니다.

이 판단을 AI에게 넘길 때 가장 먼저 부딪히는 문제는 정확도보다 AI가 틀린 답을 자신 있게 낸다는 점입니다. 공장에 AI를 넣는 일을 하는 웨이스가 현장에서 보면 이 문제는 몇 달에 한 번 나오는 불량에서 드러나기 쉽습니다. 웨이스는 제조 AI 운영체제 VEXPLOR를 만들 때도 이 문제부터 다룹니다. 이 글은 공개 연구 결과로 AI가 드문 불량에서 자신 있게 틀리는 이유 3가지를 정리하고, 공장이 AI의 불량 판정을 쓰기 전에 확인할 4가지를 적습니다.

01

웨이퍼 불량 데이터에서 정상은 2만 9천 장, 드문 불량은 49장이었습니다

공장의 불량 기록은 몇몇 유형에 몰려 있습니다. 반도체 웨이퍼 불량 분류 연구(Alawieh·Boning·Pan 2020, 공개 데이터 WM-811K 사용)가 쓴 학습 데이터를 보면 그 차이가 숫자로 드러납니다.

웨이퍼 유형	학습 데이터	비고
정상	29,357장	전체의 대부분
가장자리 고리형 불량	6,802장	불량 중 최다
전면 근접 불량	49장	정상의 약 600분의 1



반도체 웨이퍼 불량 분류 연구의 학습 데이터에서 정상 웨이퍼는 2만 9천 장이 넘었고, 가장 드문 불량은 49장뿐이었습니다

연구진은 「다수 유형만 잘 배운 모델도 전체 정확도는 매우 높게 나올 수 있지만, 수율 분석에 더 중요한 불량 유형에서는 성능이 낮다」고 적었습니다. 실제로 연구진이 비교한 기존 방법은 정상 웨이퍼를 빼고 불량만 놓고 보면 정확도가 44.35%였습니다. 이 수치는 학습 데이터와 시험 데이터의 유형 분포가 서로 다르게 나뉜 조건에서 나온 값이지만, 방향은 분명합니다. 전체 정확도가 높다는 것은 흔한 유형을 잘 맞힌다는 뜻일 수는 있어도, 드문 유형까지 잘 맞힌다는 뜻은 아닙니다.

같은 공개 데이터의 라벨을 다시 살핀 연구(Jia 외 2026, 동료 심사 전 공개본)도 같은 점을 짚었습니다. 라벨이 붙은 웨이퍼의 약 85%가 정상이어서 「전체 정확도는 모델의 행동을 거의 알려 주지 않는다」고 적었습니다. 일반 이미지 데이터에 불균형을 일부러 만들어 체계적으로 실험한 연구(Buda·Maki·Mazurowski 2018)는 한 유형이 드물수록 분류 성능이 떨어지고, 분류할 유형이 많은 과제일수록 그 폭이 커진다고 결론지었습니다. 같은 연구는 이 하락이 학습 사례의 총량이 아니라 유형 사이의 분포에서 온다는 점도 밝혔습니다.

02

AI가 내는 확신 값은 실제 정답률보다 높게 나오기 쉽습니다

드문 불량을 잘 못 맞히는 것만이 문제라면 사람이 결과를 보고 걸러 내면 됩니다. 더 큰 문제는 틀린 판정에도 높은 확신 값이 붙는다는 점입니다.

딥러닝 모델의 확신 값을 조사한 연구(Guo 외 2017)는 「요즘 신경망은 10년 전 모델과 달리 확신 값이 실제 정답률과 맞지 않는다」고 밝혔습니다. 이미지 100종을 분류하는 실험에서 ResNet-110 모델의 확신 값과

실제 정답률은 평균 16.53%포인트 차이가 났습니다. 확신 값이 실제 정답률보다 평균 16.53%포인트 높았다는 뜻입니다. 같은 연구에서 온도 보정이라는 간단한 후처리를 하자 이 차이는 1.26%포인트로 줄었습니다.

공장이 이 결과에서 얻을 것은 두 가지입니다. 화면에 뜨는 확신 값을 그대로 믿으면 안 된다는 것, 그리고 확신 값이 실제 정답률과 맞는지는 따로 확인해 보정할 수 있다는 것입니다.

03

AI는 처음 보는 입력에도 높은 확신으로 답합니다

확신 값을 보정해도 남는 문제가 있습니다. 학습 때 본 적 없는 입력이 들어오는 경우입니다. 공장에서는 원자재 공급처가 바뀌거나, 조명을 교체하거나, 설비 부품이 마모되기 시작할 때 이런 일이 생깁니다.

신경망의 오분류와 분포 밖 입력을 다룬 연구(Hendrycks·Gimpel 2017)는 손글씨 숫자를 분류하도록 학습한 모델에 아무 의미 없는 무작위 잡음 이미지를 넣었습니다. 모델은 평균 91% 확률로 그 잡음을 어떤 숫자라고 답했습니다. 그래서 연구진은 이 확률을 「확신의 직접적인 표현으로 보면 안 된다」고 적었습니다. 다만 같은 연구는 이 확률의 통계를 쓰면 처음 보는 입력을 어느 정도 걸러 낼 수 있다는 점도 보여 주었습니다.

앞의 웨이퍼 연구에서도 같은 일이 확인됩니다. 연구진이 한 불량 유형을 학습에서 일부러 빼고 시험에만 넣자, 보류 기능을 빼고 보면 모델은 그 불량을 하나도 맞히지 못했습니다(재현율 0%). 모델은 반드시 8개 유형에서 답을 골라야 했으므로 처음 보는 불량을 이미 아는 다른 유형으로 분류했습니다.

언어 모델도 산업 불량 판정에서는 정확도가 아직 낮습니다. 산업 이상 탐지 평가(Jiang 외 2025, MMAD)에서 GPT-4o의 정확도는 74.9%로 사람 전문가보다 12%포인트 낮았고, 연구진은 「산업 현장 요구에 크게 못 미친다」고 평가했습니다. 반도체 노광 불량 평가(LDU-Bench 2026, 동료 심사 전 공개본)에서도 최신 멀티모달 언어 모델들은 불량이 있는지는 비교적 잘 가려냈지만, 원인을 짚는 과제의 최고 점수는 1점 만점에 0.409였습니다. 다만 두 평가는 확신 값을 따로 보지 않았으므로, 이 결과만으로 언어 모델이 자신 있게 틀린다고 말할 수는 없습니다.

04

「모른다」고 답하게 하면 자동으로 처리하는 판정의 정확도가 올라갑니다

세 가지 이유는 모두 AI가 반드시 하나의 답을 골라야 하는 구조에서 나옵니다. 그래서 공개 연구들이 내놓은 방법은 AI에게 답하지 않을 선택지를 주는 것입니다. AI는 확신이 기준에 못 미치는 판정을 보류하고 사람에게 넘깁니다.



AI는 확신이 기준에 못 미치는 판정을 자동으로 처리하지 않고 보류해 사람에게 넘깁니다

앞의 웨이퍼 연구는 이 방법의 효과를 숫자로 보여 줍니다.

자동 처리 비율	보류된 판정	자동 처리분 정확도
전체(보류 없음)	0%	94%
약 58%	약 42%	99.0%

연구진이 따로 한 실험에서는 한 불량 유형을 학습에서 빼고 자동 처리 비율 목표를 50%로 정했습니다. 그러자 모델은 그 유형의 웨이퍼를 모두 보류했습니다. 틀린 답을 자신 있게 내는 대신 사람에게 넘긴 것입니다. 연구진은 이 방식이 「사람의 검사 인력을 필요한 곳에 배분하게 돕는다」고 정리했습니다. 다만 보류가 늘어난 만큼 자동으로 처리하는 비율은 줄어듭니다. 정확도와 자동 처리 비율 사이에서 기준을 어디에 둘지는 공장이 정해야 합니다.

보류 방식의 효과는 웨이퍼 데이터에만 해당하지 않습니다. 일반 이미지 인식에서 이 방식을 다룬 연구(Geifman·El-Yaniv 2017)는 애매한 입력에 모델이 답하지 않게 하자, 이미지 인식 상위 5개 후보 오류를 2%로 묶은 채 시험 이미지의 약 60%를 여전히 자동으로 처리할 수 있음을 보였습니다. 나머지 약 40%는 사람이나 다른 시스템이 맡아야 한다는 뜻입니다. 여기서 한 걸음 더 나아간 연구(Madras 외 2018)는 어떤 건을 사람에게 넘길지까지 모델이 함께 배우게 하면, 모델 혼자 답할 때보다 사람과 모델을 합친 전체 판단이 나아진다고 결론지었습니다. 이 연구는 제조 데이터로 실험하지 않았지만, 그 생각은 공장에서 AI와 검사원이 일을 나누는 방식에 그대로 옮겨 볼 수 있습니다.

05

공장이 AI의 불량 판정을 쓰기 전에 확인할 4가지

위 연구들을 공장의 질문으로 바꾸면 아래 4가지가 됩니다. AI 공급사를 고를 때도, 이미 쓰는 AI를 점검할 때도 같은 질문을 사용합니다.

확인할 것	공급사에 받을 것	없으면 생기는 일
유형별 성능	유형별 사례 수·정확도	약한 유형을 모름
확신 값 보정	확신 값별 실제 정답률	확신 값을 믿게 됨
처음 보는 입력	분포 변화 감지 방식	조용히 틀림
보류 출구	보류 기준·자동 처리 비율	애매한 판정도 자동 처리

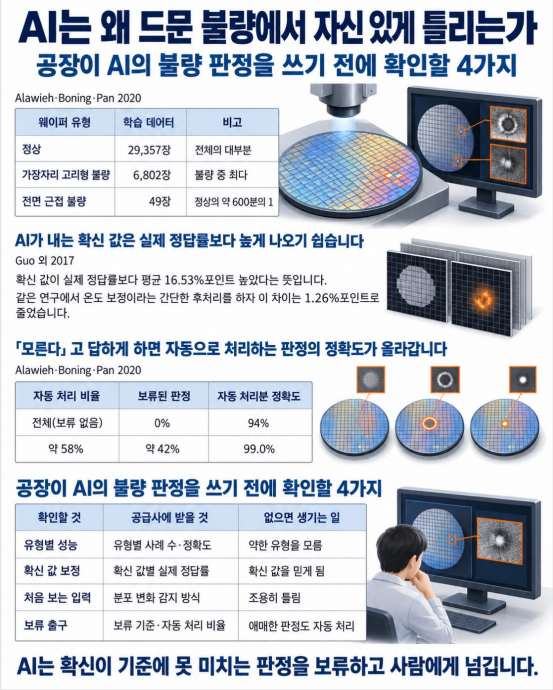
- **유형별 성능** — 전체 정확도 하나만 받지 말고 불량 유형별 학습 사례 수와 유형별 정확도를 함께 받습니다. 사례가 수십 건 이하인 유형이 어디인지 알면, 그 유형으로 판정이 나왔을 때 한 번 더 확인할 수 있습니다
- **확신 값 보정** — 「확신 90% 이상 판정의 실제 정답률은 얼마인가」를 묻습니다. 답이 90%와 크게 차이 난다면 그 확신 값으로 자동 처리 기준을 정하면 안 됩니다
- **처음 보는 입력** — 원자재·조명·설비 상태가 바뀌어 입력 분포가 달라졌을 때 이를 알아채는 장치가 있는지 묻습니다. 개별 판정의 확신 값만으로는 이 변화를 잡지 못합니다
- **보류 출구** — AI가 판정을 보류하고 사람에게 넘기는 기준이 있는지, 그 기준에서 자동 처리 비율이 얼마인지 묻습니다. 보류가 없는 AI는 처음 보는 불량도 반드시 이미 아는 유형으로 분류합니다

06

기록이 적은 공장일수록 4가지를 먼저 갖춰야 합니다

반도체 대기업은 수십 년 동안 쌓은 공정 기록이 있어도 드문 불량에서는 같은 문제를 겪습니다. 드문 불량은 정의상 기록이 적기 때문입니다. 중견·중소 공장은 흔한 불량의 기록조차 대기업보다 적어서, 같은 문제가 더 일찍 더 크게 나타납니다.

그래서 기록이 적은 공장이 AI를 들일 때 먼저 정할 것은 모델의 정확도 목표가 아니라 AI가 모를 때 어디로 빠지는지입니다. 유형별로 약한 곳을 알고, 확신 값을 보정하고, 입력이 바뀌면 알리고, 애매한 판정은 사람에게 넘기는 출구를 두면 드문 불량에서 나오는 오답이 현장에 그대로 흘러가지 않습니다.



이 글을 한 장으로 요약하면 이렇습니다 — 공장이 AI의 불량 판정을 쓰기 전에 확인할 4가지

AI는 드문 불량에서 틀리는 것보다 틀린 답을 자신 있게 내는 것이 더 위험하므로, 공장은 AI가 모를 때 사람에게 넘기는 출구부터 만들어야 합니다.

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

07

함께 읽어보세요

- AI 답의 오류, 내 불량 기록으로 10분 안에 잡기 (40분)
- 불량 기록 400행, AI로 상위 3원인 뽑기 (40분)
- 공장장이 AI 도입을 미루는 이유는 기술이 아니라, AI가 틀렸을 때 그 책임이 공장장에게 남기 때문입니다
- 상관은 알람을 만들고, 인과는 결정을 만든다

원문 blog.vexplor.com/post/why-ai-is-confidently-wrong-on-rare-defects
발행 2026년 10월 7일 · 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