

AI/ML 연재 도구가 보고한 숫자를 실제와 대조하는 법 1/4

저희 검사 스크립트가 자기 성적을 잘못 보고하고 있었습니다

방동걸 (주)웨이스 대표이사

2026년 9월 2일

AI에게 일을 맡기기 시작하면 검사 장치가 따라 늘어납니다. 사람이 결과물을 일일이 다 볼 수 없으니 자동으로 훑는 스크립트를 만들어 붙이고, 그 스크립트가 통과를 내면 다음 단계로 넘깁니다. 코드 린터, 테스트, 배포 전 점검, 문서 검사 — 형태는 달라도 하는 일은 같습니다. **사람 대신 봐 주는 것입니다.** AI에게 일을 맡겨 VEXPLOR를 만드는 웨이스에서도 사내 검사 스크립트 하나가 그 일을 잘못하고 있었습니다.

그런데 그 검사 장치가 실제로 일하고 있는지는 누가 봅니까.

저희에게 그 질문이 사고로 돌아온 이야기를 적습니다. 대상은 **저희가 사내에서 쓰는 문체 검사 스크립트**입니다.

01

무엇을 검사하는 도구인가

저희는 제안서·보고서·블로그 글을 여러 사람이 나눠 작성합니다. 사람마다 말투가 다르면 회사가 두서없어 보여서, 쓰지 않기로 한 표현 목록을 만들어 두었습니다. 번역투와 영어 문장 구조를 그대로 옮긴 문장이 주 대상입니다.

그 목록을 사람이 매번 눈으로 대조할 수는 없어서 **파이썬 스크립트 하나를 만들었습니다.** 문서 경로를 넣고 돌리면 걸린 문장을 줄 번호와 고칠 말까지 짚어 줍니다. 문서를 낼 때마다 이걸 돌렸고, 매번 통과가 나왔고, 그 통과를 믿고 문서를 내보냈습니다.

그 스크립트가 얼마나 잘 잡는지는 저희 규약 문서에 "자가검증 15/15"라고 적혀 있었습니다. 15개 항목을 시험해 15개를 다 잡았다는 뜻입니다.

그런데 그 검증을 돌릴 시험 문제가 코드에 없었습니다. 15/15는 예전에 사람이 손으로 한 번 세어 보고 적어 둔 숫자였고, 그 뒤로 아무도 다시 재지 않았습니다. 검사기는 매일 돌면서 매번 통과를 냈지만, 그 통과가 무엇을 근거로 나오는지 확인한 적이 없었습니다.

어제 처음으로 실제로 켜졌습니다. 규칙 39종, 재현율 40/40, 자연문 47문장에 오답 0.

숫자 자체는 나쁘지 않았습니다. 문제는 그것을 아무도 모르고 있었다는 것입니다. 검사기가 잘 잡고 있었다는 것도, 못 잡고 있었다는 것도 확인할 방법이 없는 상태로 몇 주를 보냈습니다.

이 글은 그 사건을 그대로 적은 글입니다. 저희가 만든 검사 장치가 저희를 어떻게 속였는지, 그 첫 회차입니다.

02

검사기는 매일 돌았고 매번 통과했습니다

저희는 문서를 낼 때마다 금지 표현 검사기를 돌립니다. 번역투와 어색한 차용어를 잡는 도구입니다. 매번 [PASS]가 나왔습니다.

여기서 저희가 착각한 것이 하나 있습니다. 검사기가 통과를 냈다는 것은 "문제가 없다"가 아니라 "내가 문제를 못 찾았다"입니다. 두 문장은 화면에서 똑같이 초록색으로 보입니다.

잡아야 할 것을 하나도 못 잡아도 통과가 나옵니다. 오히려 아무것도 못 잡는 검사기가 가장 조용하고, 가장 자주 통과를 냅니다.

그래서 검사기에는 성적표가 따로 필요합니다. 이 검사기가 잡아야 할 것을 실제로 잡는가. 그것을 재는 별도의 장치입니다. 저희 문서에는 그 성적표의 숫자가 적혀 있었는데, 정작 그 숫자를 만드는 장치가 없었습니다.

03

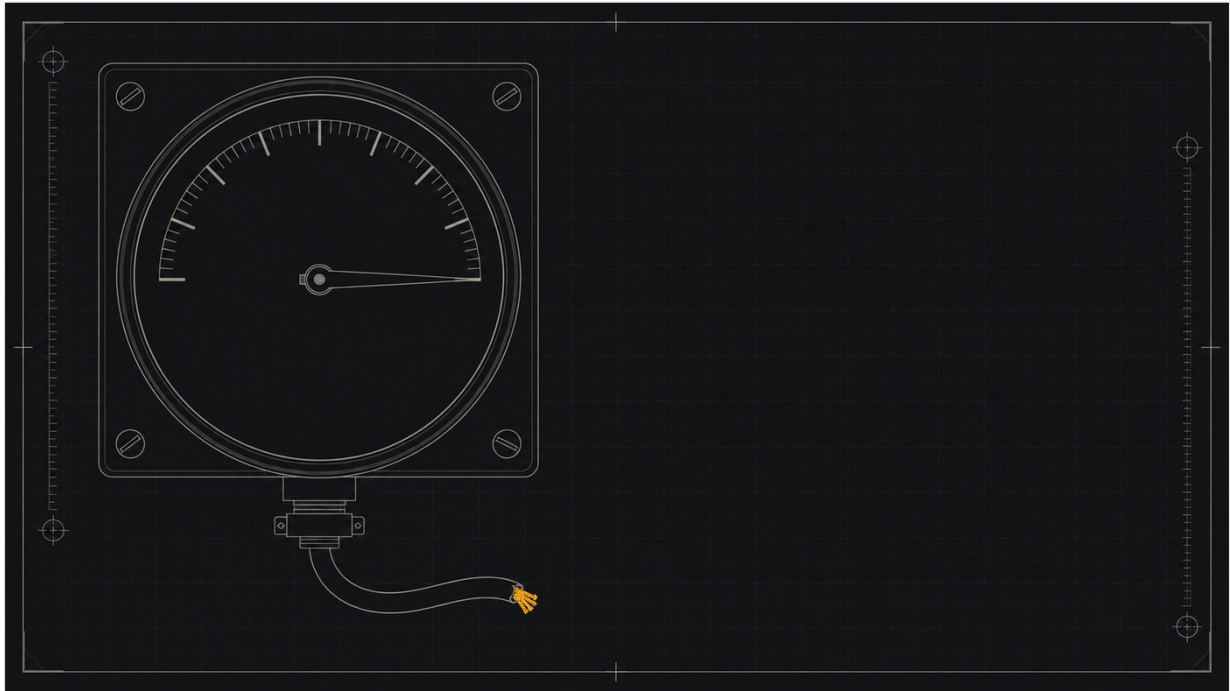
다른 일을 하다가 걸렸습니다

찾아낸 경위가 허무합니다. 저희를 잡으려 간 것이 아니었습니다.

외부 연구 하나를 저희 작업 방식에 적용할지 검토하던 중이었습니다. 그 검토에서 "우리 도구 중 자가검증이 있는 것"을 세어 보려고 명령어를 쳤습니다.

```
$ 금지어_스윙.py --selftest error: unrecognized arguments: --selftest
```

문서에 적혀 있는 기능이 코드에 없었습니다. 규약 문서는 그 명령어를 근거로 15/15를 말하고 있었고, 명령어 자체가 없었습니다.



숫자는 계기에 떠 있었고, 그 숫자를 만드는 것은 어디에도 붙어 있지 않았습니다

이런 결함은 스스로 신고하지 않습니다. 검사기가 매일 통과를 내고 있었으니 아무 경보도 울리지 않았습니다. 옆에서 다른 것을 세다가 우연히 손이 닿은 자리였습니다.

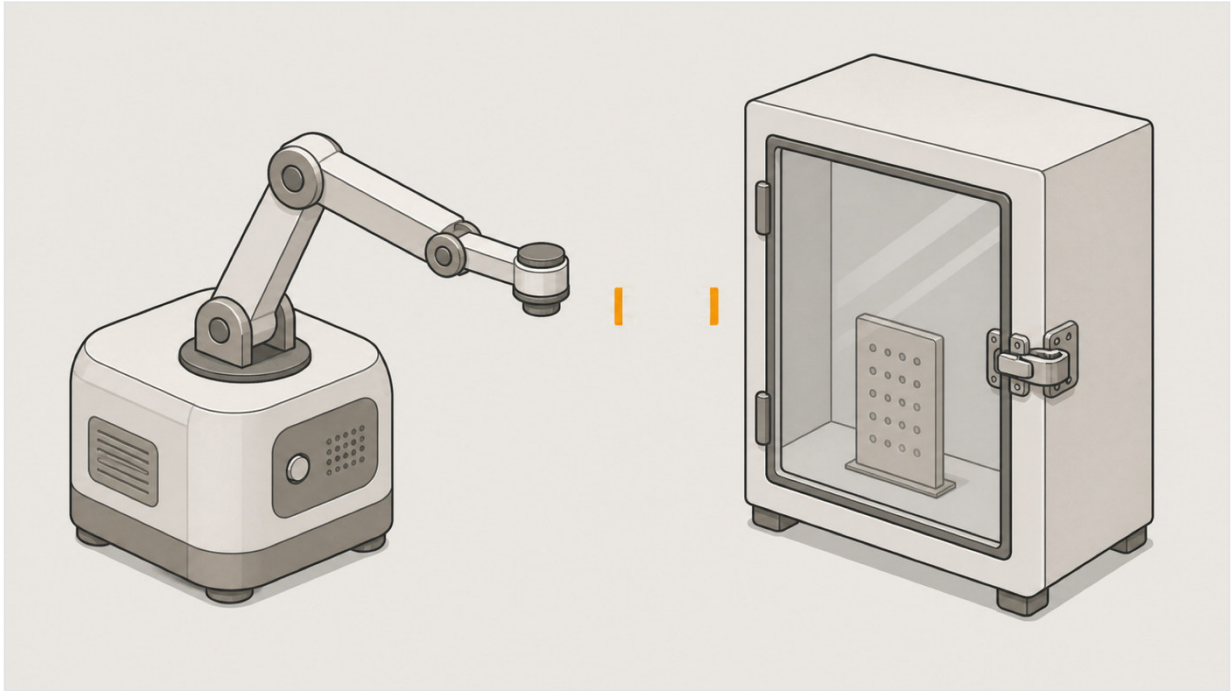
04

고친 방법 — 검사기가 자기 성적을 증명하게 했습니다

세 가지를 했습니다.

- **검증 세트를 실제 파일로 만들었습니다.** 걸려야 하는 문장 40개(양성)와 걸리면 안 되는 자연스러운 문장 47개(음성)를 규칙별로 채웠습니다. 규칙이 39종이니 종마다 자기를 걸리게 하는 문장이 최소 하나씩 있어야 합니다.
- **--selftest를 실제로 만들었습니다.** 이제 명령어 하나로 재현율과 오탐이 함께 나옵니다. 실측값은 재현율 40/40, 오탐 0/47, 규칙 39종 중 양성 보유 39종입니다.
- **그 세트를 자동 수정 대상에서 뺐습니다.** 검사기를 고치는 쪽이 시험 문제까지 고칠 수 있으면 성적은 언제든지 만점이 됩니다. 세트는 사람 승인 없이는 바뀌지 않는 자리에 두었습니다.

세 번째가 핵심입니다. 채점표를 만드는 것보다, 채점표를 시험 보는 쪽 손이 닿지 않는 곳에 두는 것이 어렵습니다.



채점표는 검사기 손이 닿지 않는 자리에 두었습니다

05

이 사건에서 남는 것

저희가 배운 것을 세 줄로 적으면 이렇습니다.

- 초록불은 상태가 아니라 그 검사기의 발언입니다. 발언의 신뢰도는 따로 재야 합니다.
- 문서에 적힌 숫자는 그 숫자를 만든 장치가 지금도 도는지까지 함께 적혀야 합니다. 한 번 손으로 확인한 값과 매번 자동으로 나오는 값은 문서에서 똑같이 생겼습니다.
- 자기 도구를 못 믿는 것이 기본값입니다. 저희는 자동화를 파는 회사인데, 저희 자동화가 저희에게 잘못 보고하고 있었습니다.

그리고 하나가 남았습니다. 이번에는 검사기를 검사하는 장치를 만들었는데, **그 장치가 제대로 도는지는 무엇이 보증합니까.**

그 자리에서 또 한 번 걸렸던 이야기를 다음 회차에 적겠습니다. 이번에는 저희 안에서 걸린 것이 아니라, 바깥에서 먼저 알려 주었습니다.

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



원문	blog.vexplor.com/post/the-checker-graded-itself
발행	2026년 9월 2일 · 방동걸 (주)웨이스 대표이사
회사	주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소	(18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1332·1333·1334·1335호 (영천동, 실리콘엘리동탄)
연락처	전화 070-4730-0629 · contact@wace.me
홈페이지	wace.me · linkedin.com/company/wace-inc



원문



홈페이지



링크드인