

인사이트

AI가 요구사항을 채우게 하면서, 우리는 무엇을 제한했나

방동걸 (주)웨이스 대표이사

2026년 9월 6일

저희는 회의 녹음과 문서에서 시스템 요구사항을 자동으로 채우는 기능을 만들고 있습니다. 담당자와 두세 시간 이야기한 녹음을 넣으면 표준 질문 트리가 채워지고, 그 결과가 진단서와 견적의 입력이 됩니다. 사람이 몇 주 걸려 정리하던 것을 자동으로 만드는 기능이니, 상식적으로는 "얼마나 많이 채우는가"가 성능일 것입니다. VEXPLOR에 그 기능을 넣는 웨이스가 상식을 그대로 따르지 않은 이유를 적은 글입니다.

저희가 이 기능에 제일 먼저 넣은 규칙은 반대쪽이었습니다. **근거가 없으면 채우지 않고 비운다**. 파일럿을 다시 돌린 회차에서 비운 항목이 미측정 6건, 확인이 안 끝난 채로 표시만 남긴 것이 22건이었고, **추측으로 넣은 값은 0건**이었습니다.

왜 자동으로 채우는 기능에서 안 채우는 규칙이 먼저인지, 그리고 그 규칙들을 왜 자동 개선 대상에서 빼 두었는지 적습니다.

01

자동으로 채우면 틀린 것도 자동으로 퍼집니다

이 기능의 출력은 문서 한 장에서 끝나지 않습니다. 채워진 값이 진단 판정으로 가고, 판정이 범위 산정으로 가고, 범위가 견적 금액이 됩니다. 사람이 손으로 정리할 때는 이상한 값이 중간에 걸렸지만, 자동으로 흐르면 틀린 값도 같은 속도로 끝까지 갑니다.

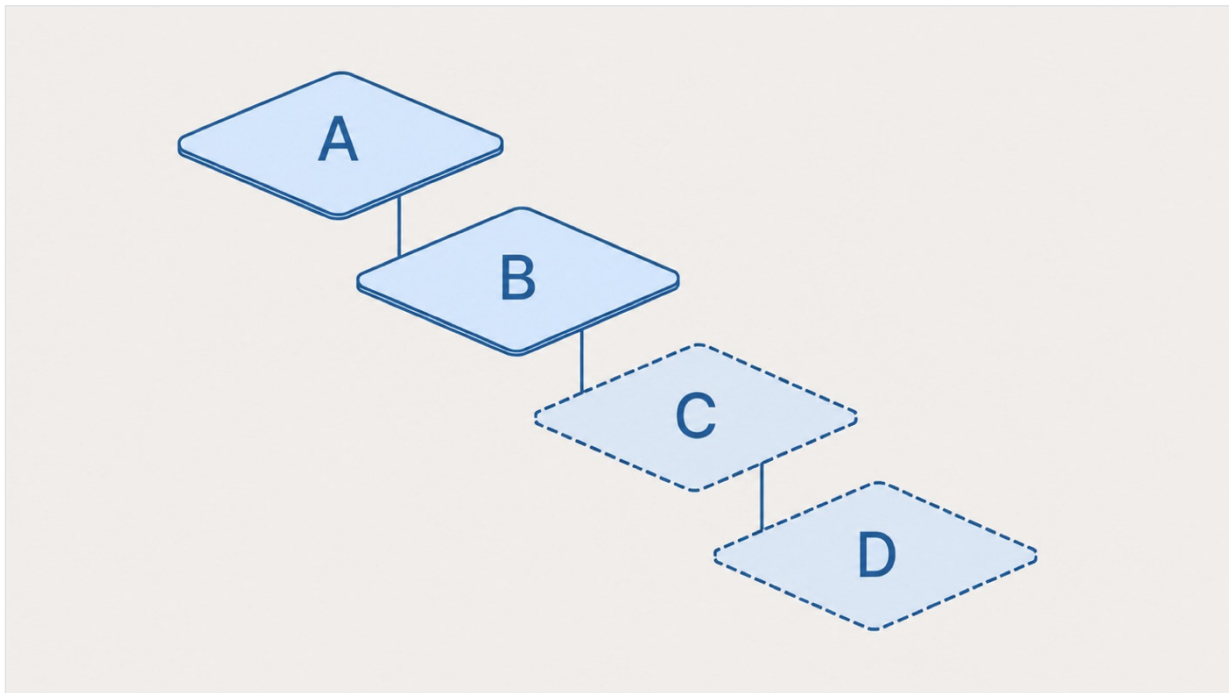
그래서 설계 원칙을 하나로 잡았습니다. **통과 근거가 없으면 멈춘다**. 전사에 없다는 것은 그런 사실이 없다는 뜻이 아니므로, 빈 항목은 추측으로 메우지 않고 미측정으로 표시해 현장 확인 목록으로 보냅니다. 비어 있는 진단서는 덜 그럴듯해 보이지만, 채워진 거짓말보다 훨씬 쌉니다.

02

근거마다 등급을 붙입니다

무엇으로 채웠는지를 항목 단위로 기록합니다. 같은 문서라도 고객이 직접 말한 부분과 저희가 해석한 부분은 등급이 다릅니다.

등급	무엇으로 채웠나	어떻게 다루나
A	고객 작성·실데이터 실측	그대로 사용
B	고객 발화 회의 전사	시각 표시 필수
C	우리가 쓴 문서	미검증 표시 강제
D	표 한 줄 + 표준 보간	단독 채움 금지



실선과 점선의 경계가 기준입니다. 판정과 금액을 가르는 항목은 A·B 등급 근거가 없으면 아예 채우지 않습니다.

여기서 중요한 것은 C입니다. 제안서·사업계획서·분석 보고서는 우리가 쓴 문서라 읽기 편하고 이미 정리돼 있어서, 자동 채움의 재료로는 가장 유혹적입니다. 그런데 그것으로 채운 값을 다시 우리가 검증하면 **같은 편향을 두 번 통과시키는 것**이라 대조가 성립하지 않습니다.

D는 더 분명했습니다. 레퍼런스 표 한 줄과 도메인 상식만으로 채운 세트를 실제로 만들어 봤더니 항목 97개가 나왔습니다. 그럴듯한 97개였고, **검증이 물리적으로 불가능한 97개**였습니다. 같은 방식으로 만든

다른 세트들은 실제 인터뷰와 대조하자 예외 없이 판정이 여러 단계 내려갔습니다. 올라간 사례는 하나도 없었습니다.

그래서 판정과 금액을 가르는 항목 10개는 **A·B 등급 근거가 없으면 아예 채우지 않습니다.**

03

전사를 넣기 전에 여섯 가지를 봅니다

자동 채움의 입력은 대부분 회의 녹음을 받아 적은 전사입니다. 그런데 전사는 생각보다 자주 망가진 채로 도착합니다. 그래서 채움을 시작하기 전에 전사 자체를 먼저 검사합니다.

무엇을 보나	통과 조건
시각 표시	발화 10건당 8건 이상
화자 번호	판정 근거로 사용 금지
수치	사람 확인 후 반영
다른 회의 혼입	주제 교차 반복 없음
길이 절단	상한 부근 종료 아님
회의 종류	분류 + 근거 공백 명시

두 번째가 특히 값비싼 교훈이었습니다. 전사에는 화자1·화자2 같은 번호가 붙어 있어서 "고객이 말한 것"과 "우리가 말한 것"을 가르는 데 쓰고 싶어집니다. 그런데 **같은 녹음을 다시 받아 적으면 화자 수가 6명에서 9명으로 바뀌고, 같은 문장의 화자 번호도 달라졌습니다.** 번호는 파일마다 새로 매겨지는 임시 값이라 근거가 될 수 없습니다. 지금은 시각과 원문 문자열로 근거를 고정합니다.

여섯 번째도 실측에서 나왔습니다. 자동으로 채워지는 비율은 회의 종류가 지배합니다. 현황 인터뷰는 70~80%대가 채워지는데, 착수·거버넌스 회의는 60% 아래로 떨어집니다. 영업 대화만 녹음된 건은 여러 모듈의 근거가 통째로 없었습니다. 그래서 채움률을 성능으로 보고하지 않고, **어느 영역이 근거 공백인지를 산출물에 함께 적습니다.**

04

검사를 통과하면서 틀리는 두 가지가 있습니다

등급과 게이트를 다 통과하고도 틀리는 유형이 둘 있었습니다. 둘 다 형식 검사로는 잡히지 않습니다.

첫째, 주어가 바뀝니다. 파일럿 하나에서 "메뉴 43개 중 29개 구현", "계정 1개", "일부 화면 미구현" 같은 서술이 고객 현황으로 들어가 있었습니다. 실제로는 전부 **저희가 만든 재현 앱**의 상태였습니다. 고객 시스템은 메뉴가 전부 있었고 계정도 여러 개였습니다. 이 오류는 판정 점수를 1도 움직이지 않아 어떤 검증에도 걸리지 않으면서, 진단서의 사실관계만 조용히 어긋나게 둡니다. 그래서 "구현·미구현" 같은 개발 관점 어휘나 저장소 경로가 근거 문장에 있으면 **주어를 다시 확인**하게 했습니다.

둘째, 목표가 현황으로 역류합니다. 같은 파일럿에서 "종이 수집률 85%"라는 값이 고객 현황으로 굳어 있었습니다. 출처를 따라가니 저희 제안서의 목표 표에 있던 기준선이었습니다. 고객 자료 어디에도 그 실측은 없었습니다. 우리가 그리려던 그림이 **고객의 현재로 되돌아온 것**입니다. 지금은 우리 목표 표에서 온 값에 별도 표시를 붙여 현황 항목에 들어가지 못하게 막습니다.

05

같은 시간이면 어떤 근거부터 확보하는 게 이득인지도 재 봤습니다

인터뷰를 길게 하는 것이 늘 최선은 아니었습니다. 재실행에서 근거 종류별로 얼마나 넓게 덮는지를 세어 보니 순서가 이렇게 나왔습니다.

우선	근거	무엇이 좋은가
1	모듈 사용 현황 실측	한 줄로 여러 영역 확정
2	운영 DB 집계	진술 없이 규모 확정
3	서면 문답	항목에 직결되나 어긋남 여지
4	현황 인터뷰 전사	전 영역을 고르게 덮음
5	착수·개선 회의 전사	의제 영역만 덮음

1번이 왜 맨 위인지가 뜻밖이었습니다. "이 모듈들은 사용하지 않습니다"라는 한 줄이 세 시간짜리 인터뷰보다 넓은 영역을 한 번에 확정합니다. 없는 것을 확인하는 문장이 있는 것을 설명하는 문장보다 정보가 많습니다.

3번에는 단서가 붙습니다. 담당자 답변과 실데이터가 어긋나는 일이 실제로 있었습니다. 한 파일럿에서 품목 수를 8,000여 개로 말씀하셨는데 운영 DB 집계는 15,000개를 넘었습니다. 누가 틀린 게 아니라 **세는**

기준이 서로 달랐던 것입니다(쓰지 않는 이력 품목이 포함되는가). 그래서 규모·금액·비율이 걸린 항목은 진술만으로 확정하지 않고 집계와 함께 봅니다.

06

비운 항목은 그냥 비워 두지 않습니다

fail-closed의 값어치는 비우는 데 있지 않고 **비운 자리를 추적하는 데** 있습니다. 채우지 못한 항목은 미측정으로 표시돼 현장 확인 목록으로 넘어가고, 그 목록이 다음 방문의 질문지가 됩니다. 산출물에는 어느 영역이 근거 공백인지가 함께 실립니다.

고객 앞에 비어 있는 진단서를 내놓기는 불편합니다. 다 채워진 문서가 일을 더 많이 한 것처럼 보입니다. 그런데 저희가 파일럿에서 확인한 것은 그 반대였습니다 — **채워진 값 중 검증이 불가능한 것들이 나중에 가장 비쌌습니다**. 진단이 틀리면 범위가 틀리고, 범위가 틀리면 견적이 틀립니다. 비어 있는 항목은 다음 회의에서 30초면 채워집니다.

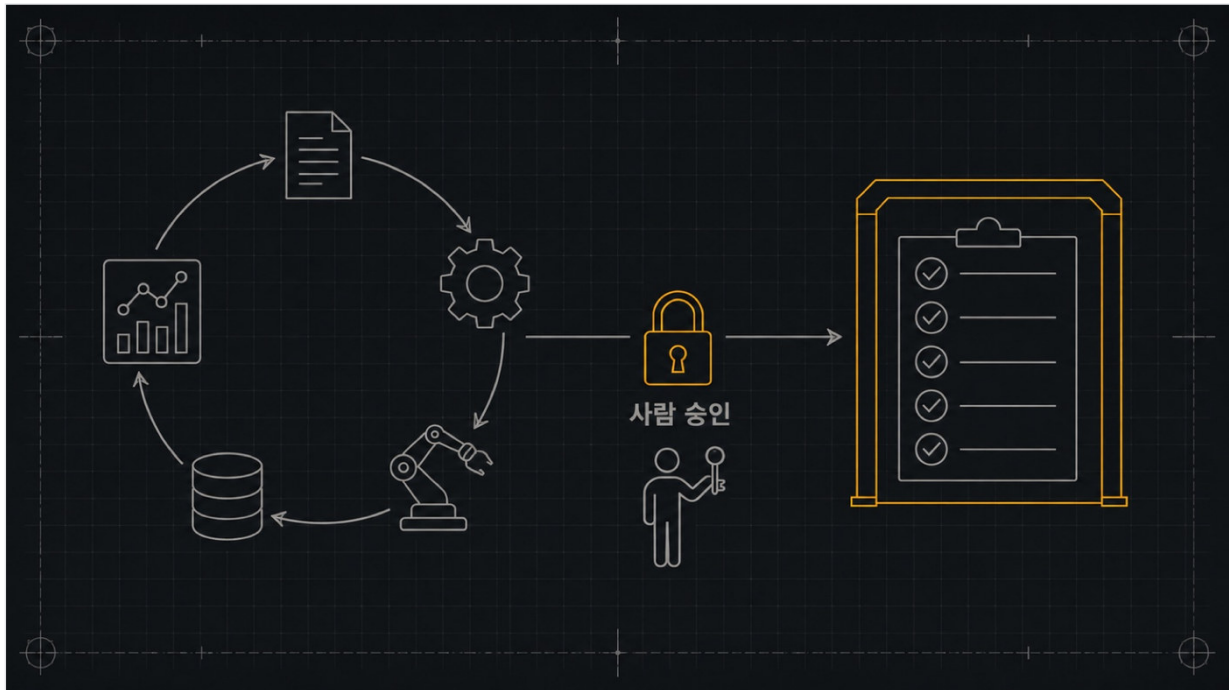
07

이 규칙들은 시스템이 스스로 못 고칩니다

여기가 이 글에서 제일 하고 싶은 이야기입니다.

저희 작업 환경은 스스로 규칙을 늘리고 고칩니다. 사고가 나면 그 유형을 막는 조항이 붙고, 반복되는 실패는 검사 도구로 승격됩니다. 그렇게 하는 편이 사람이 매번 기억하는 것보다 안전합니다.

그런데 **판정하는 쪽까지 같은 방식으로 두면 안 됩니다**. 자기 성적을 매기는 규칙을 자기가 고칠 수 있으면, 좋아지는 가장 쉬운 길은 채점 기준을 낮추는 것입니다. 사람이 나쁜 뜻으로 그러는 게 아니라, 개선 루프의 자연스러운 최적화 방향이 그쪽입니다.



사실 확인, 불변 조건 검사, 완주 점검, 자동 채움 게이트는 자동 개선 대상에서 빼 두었고 이 계층의 변경은 사람 승인 경로로만 갑니다.

그래서 사실 확인 규칙, 불변 조건 검사, 완주 점검, 그리고 위에 적은 자동 채움 게이트는 **자동 개선 대상에서 뺐습니다**. 이 계층의 변경은 사람 승인 경로로만 갑니다. 개선 속도를 일부러 늦춘 자리입니다.

바깥 근거도 마침 같은 방향입니다. 모델에게 자기가 쓸 실행 환경을 스스로 만들고 키우게 한 뒤 2,207개 비공개 과제로 채점한 최근 연구(arXiv 2609.01437, 아직 심사 전 프리프린트입니다)에서, 스스로 만든 규칙 묶음은 코드·검색 영역에서 사람이 만든 기준선에 크게 못 미쳤고 반복 개선의 이득도 불안정했습니다. 저희가 규칙 문서를 자동화하면서도 판정 계층만은 손대지 못하게 둔 이유와 같은 자리를 가리킵니다.

08

아직 못 한 것

이 구조는 아직 완성이 아닙니다.

- 게이트는 **입력에만** 걸려 있습니다. 채워진 값이 진단·견적으로 흘러간 뒤 뒤늦게 틀렸다는 것이 밝혀졌을 때, 그것을 되짚어 회수하는 절차는 아직 문서로만 있습니다.
- 확인 목록으로 보낸 항목이 실제로 현장에서 확인돼 돌아오는 비율을 **아직 세지 않았습니**다. 비우는 규칙은 확인이 돌아와야 완성됩니다.
- 자동 채움은 파일럿 단계입니다. 고객 환경에서 상시로 도는 기능이 아닙니다.
- 등급 판정 자체가 아직 사람 손을 탐니다. 무엇이 A이고 무엇이 C인지는 규칙으로 적혀 있지만, 경계에 있는 근거(고객이 보내 준 우리 양식 문서 같은 것)는 매번 사람이 정합니다.

한 가지 더 적어 둡니다. 이 게이트들은 전부 사고가 난 뒤에 생겼습니다. 표 한 줄로 97개 항목을 만들어 본 것도, 우리 재현 앱 상태가 고객 현황으로 들어간 것도, 우리 목표가 고객의 현재로 되돌아온 것도 저희가 실제로 겪은 일입니다. 설계가 좋아서 미리 막은 게 아니라, 파일럿을 다시 돌려 대조했기 때문에 보였습니다. 그래서 다음 게이트도 아마 다음 사고에서 나올 것이라고 봅니다.

그래서 저희가 지금 말할 수 있는 것은 "자동으로 다 됩니다"가 아니라 이 정도입니다 — 자동으로 채우되, 못 채운 자리를 못 채웠다고 표시하고, 그 판정 기준만은 자동으로 바꾸지 않습니다.

시스템에 자동 개선을 붙이신다면, 무엇을 자동 개선 대상에서 빼 두셨나요.

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



원문 blog.vexplor.com/post/what-we-locked-from-self-improvement

발행 2026년 9월 6일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 ·
사업자등록번호 439-87-02852

주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동)
1702·1706호 (영천동, 실리콘밸리동탄)

연락처 전화 070-4730-0629 · contact@wace.me

홈페이지 wace.me · linkedin.com/company/wace-inc



원문



홈페이지



링크드인