

Intelligence for Safe World

SURESOFT

VERIFAI-A 도구소개서



SURESOFT

TABLE OF CONTENTS

Background	3p
Solution	8p
Features	9p
ROI	23p
Options	24p
Specification	25p

VERIFAI-A

VERIFAI-A(베리파이-에이)는 비결정적인 AI 에이전트의 품질을 검증하는 테스트 · 평가 자동화 플랫폼입니다. 5종의 테스트 방법론과 LLM 기반 평가를 축으로, 유니버설 커넥터 통합 연동과 테스트 케이스 자동 생성 · 증강을 단일 환경에서 제공합니다.

핵심경쟁력
(차별점/강점)

테스트 방법론 5종



유니버설 커넥터

테스트 케이스
자동 생성 · 증강

평가자 스튜디오

AS-IS

- 1 고정 입출력 도구의 검증 한계
- 2 모델별 API · 프로토콜 개별 연동
- 3 수작업 테스트 데이터 생성
- 4 주관적 · 비일관 수동 평가
- 5 단발성 수동 실행 · 이력 관리 한계



To-BE

- 1 테스트 방법론 5종 · LLM 평가 기반 응답 검증
- 2 유니버설 커넥터로 인터페이스 통합 연동
- 3 시나리오·문서 기반 자동 생성 · 의미보존 증강
- 4 G-Eval · DAG · 평가단 기반 일관 평가 체계
- 5 응답 수집 · 평가 파이프라인 자동화

AI 에이전트 진화 5단계

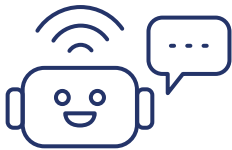


에이전트의 자율성이 높아질수록 기업의 관리 영역과 리스크도 비례하여 증가함

검증되지 않은 에이전트의 자율적 판단 : 기업의 법적 책임 인정 판례

에어캐나다 챗봇의 부정확한 정보 제공 및 배상 인정 판결 (2024)

사건 개요



고객이
에어캐나다 AI 챗봇으로부터
"항공권 구매 후 90일 이내에
경조사 할인을 소급 적용 받을 수
있다"는 잘못된 안내를 받고
예매를 진행

기업의 책임 회피



사후 환불을 거절당한
고객이 소송을 제기하자,
에어캐나다는 챗봇이 회사와
분리된 '별개의 법적 주체'라며
책임을 부정

재판소의 판결



민사해결재판소는 챗봇 역시
공식 웹사이트의 일부이므로
기업이 전적으로 책임져야 하며,
고객이 정보의 정확성을
일일이 교차 검증할 의무는
없다고 판시

글로벌 AI 법제도 정립

- AI 리스크 통제를 위한 글로벌 규제 강화와 고위험 시스템 사전 검증 법제화 추세

EU AI Act (` 24.1월 제정, ` 26.8월 전체 발효)

고위험 AI에 대한 사전 검증과
적합성 평가를 법적 의무로 규정

- 세계 최초의 포괄적 AI 규제법
- 법적 구속력을 기반으로 글로벌 AI 규범 표준 제시
- 2024년 제정 → 단계적 시행
- 2026년 8월 : 고위험 AI 규제 전면 적용**



EU Artificial
Intelligence Act

한국 AI 기본법

신뢰성 확보를 위한 전문 검증기관 지정
및 이용자 보호를 위한 사전 조치 의무화

- 아시아 태평양 지역 최초의 포괄적 AI 단행법
- 개발 · 이용 사업자 공동 책임 및 의무 분담
- 고영향 AI 검 · 인증 및 기본권 영향 평가 도입
- 고성능 AI 안전성 확보 및 결과 제출 의무**



AI Basic Act of the
Republic of Korea

AI 에이전트 테스트, 왜 어려운가?

환경 및 연동

복잡한 연동 환경과 개발 부담 가중



- 특화 에이전트 네트워크가 여러 앱과 비즈니스 기능 간 동적으로 협업
- UX의 1/3이 앱 조작 방식에서 에이전트 기반 프론트엔드로 전환

평가 및 기준

비결정적 응답과 주관적 검증



- 동일한 질문에도 매번 다른 답변
- 명확한 기댓값 정의 불가
- 멀티턴 문맥 유지력 측정 한계
- 일관성 없는 주관적 품질 평가

자원 및 관리

데이터셋 확보 및 이력 관리의 어려움



- 수작업 데이터 생성의 높은 비용/시간
- 엡지케이스 확보 난항
- 테스트 이력 추적 및 버전 관리 한계
- 일관된 주기의 자동화 검증 체계 부재

입출력이 고정된 기존의 테스트 도구로는 매번 달라지는 AI의 응답을 검증하기 어려움
AI 에이전트 고유의 비결정적 출력을 완벽히 통제하는 데 근본적인 한계가 존재

AI 에이전트 검증의 패러다임 전환

지능적 자율성 및 복잡도 증대

단순 기능을 넘어 자율적 판단을 수행함에 따라,
기존 도구로는 예측 불가능한 리스크 증대

비결정적 출력 특성

동일한 질문에도 매번 다른 형태의 응답을 하여,
정답이 고정된 기존 시스템으로 검증 불가

데이터 및 이력 관리의 한계

사용자의 예측 불가능한 입력에 대응하는
테스트 데이터 확보 및 검증 이력 관리의 한계

에이전트 특화 검증 체계

유니버설 커넥터 연동을 통한 검증 체계
구축 및 전용 방법론을 통한 검증 지원



LLM 기반 지능형 평가

LLM을 평가자 / 평가단을 통해
비결정적 응답에 대한 객관적인 평가



데이터 자동 생성 및 이력 통합 관리

LLM 기반 테스트케이스 자동 생성을 통한 커버리지
확보 및 스케줄러 및 변경 이력을 통한 체계적 관리

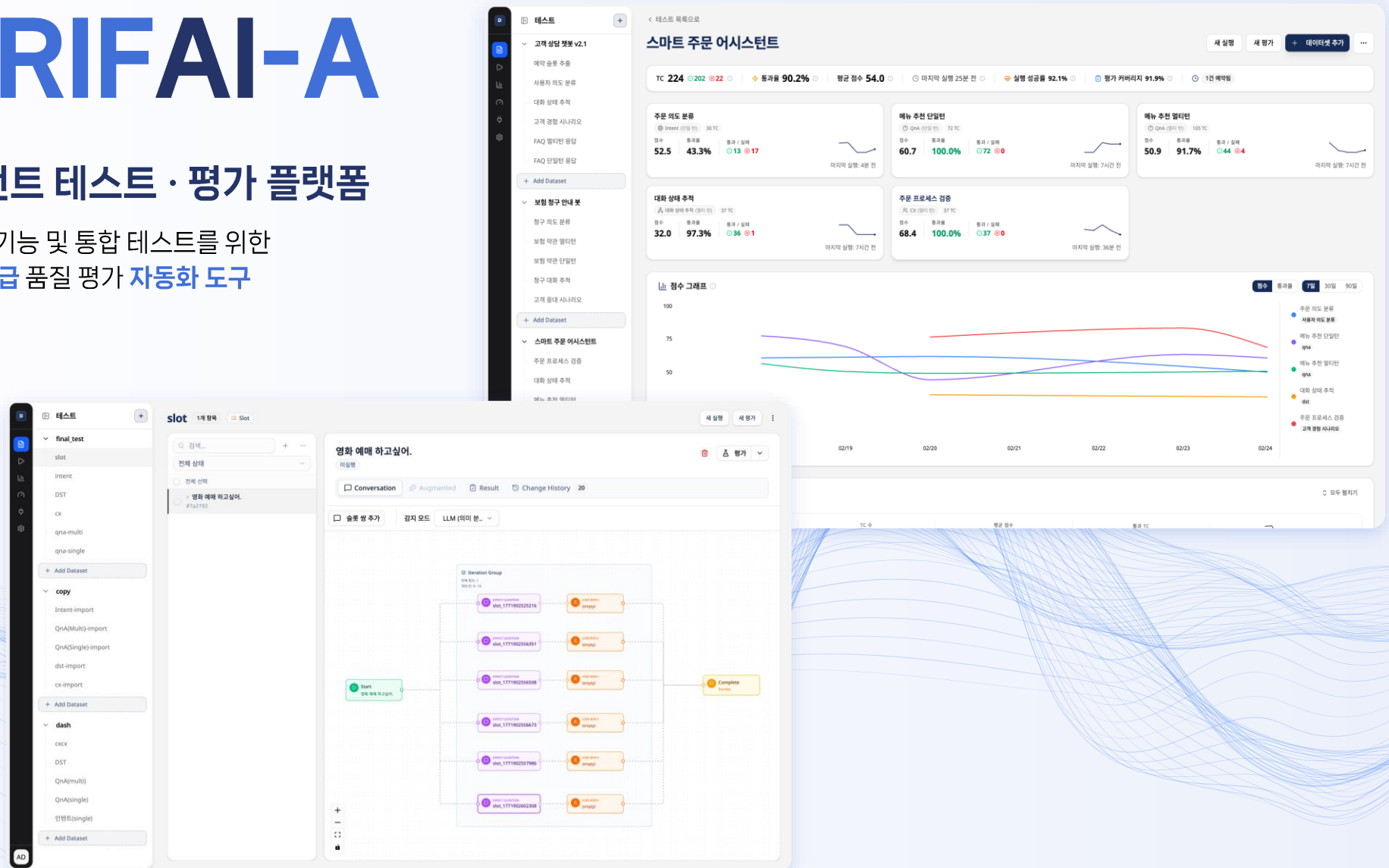


전문적인 5종 검증 방법론과 자동화된 시나리오 생성 기술을 결합하여,
다양한 평가 방식을 통해 AI 에이전트의 품질을 완벽히 보장하는 통합 검증 환경을 제공

VERIFAI-A

AI 에이전트 테스트 · 평가 플랫폼

Agentic AI 기능 및 통합 테스트를 위한
엔터프라이즈급 품질 평가 자동화 도구



AI 에이전트의 품질을 검증하는 올인원 테스트 플랫폼

- 에이전트 개발/품질 관리 팀이 테스트를 설계하고 실행 및 평가까지 할 수 있도록 설계된 직관적 UI



KEY FEATURE 01

5가지 테스트 방법론

QnA, Intent, Slot, DST, CX 방법론으로 에이전트를 체계적으로 검증



KEY FEATURE 02

유니버설 커넥터

다양한 에이전트를 표준화된 자체 인터페이스로 하나의 플랫폼에서 통합 관리



KEY FEATURE 03

LLM 기반 테스트 케이스 자동 생성

시나리오를 입력하여 다양한 테스트 케이스를 자동으로 생성



KEY FEATURE 04

문서 기반 테스트 케이스 자동 생성

제품 / 서비스 매뉴얼, FAQ 문서 등을 분석하여 테스트 케이스를 자동으로 생성



KEY FEATURE 05

퍼소나 기반 CX 자동 테스트

다양한 가상 고객 퍼소나를 통해 테스트 케이스를 자동으로 생성



KEY FEATURE 06

의미보존 데이터 증강

5가지 증강 기법을 통해 테스트 커버리지를 확장



KEY FEATURE 07

평가자 스튜디오

LLM 기반 평가자 / 평가단으로 프로젝트 요구에 맞는 유연한 평가 체계



KEY FEATURE 08

파이프라인 자동화

응답 수집부터 평가까지 끊김 없는 자동화 흐름 구성

테스트 방법론 - QnA

- 사용자 질문에 대해 적절한 답변을 제공하는지 여부를 평가



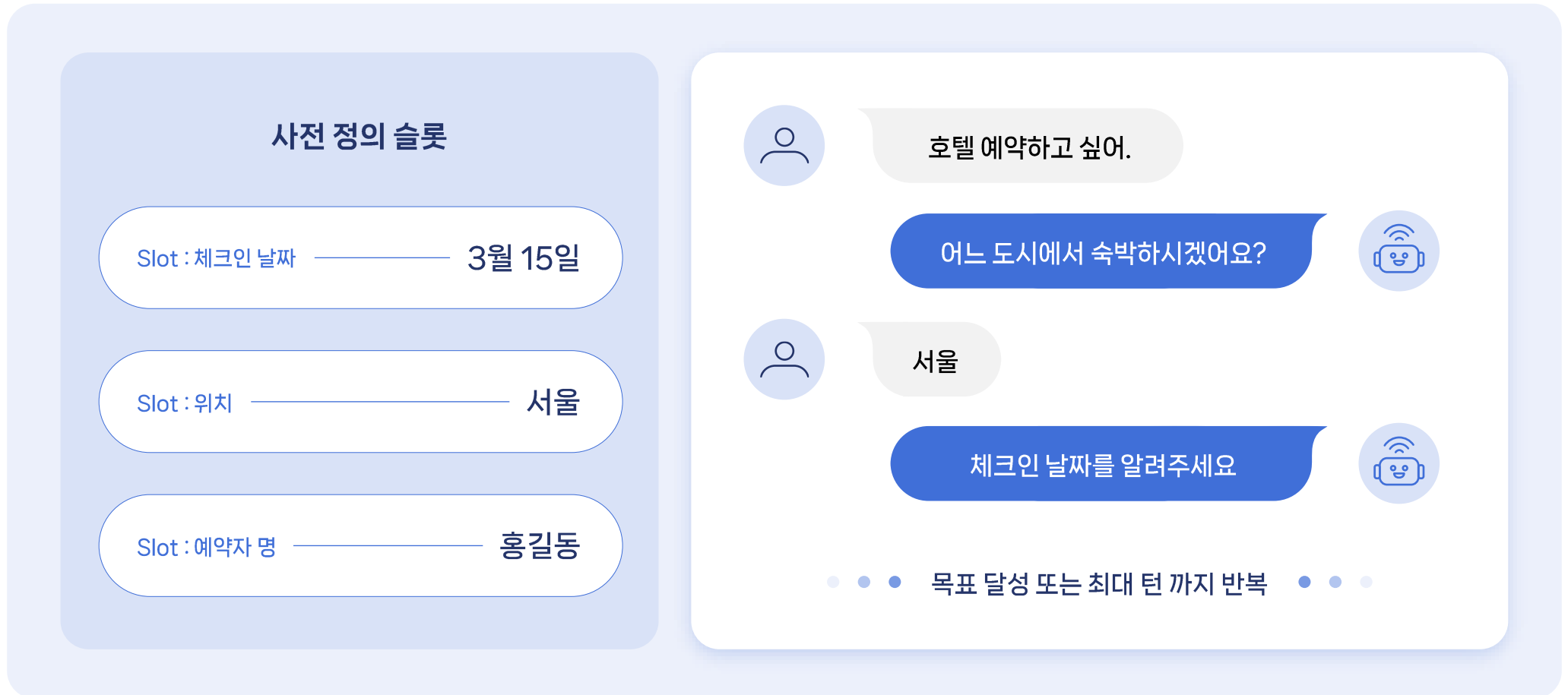
테스트 방법론 - Intent

- 사용자의 발화의 의도를 AI 에이전트가 정확하게 파악하는지 평가



테스트 방법론 - Slot

- 사전에 정의된 슬롯을 기반에 대한 사용자 응답을 유도하고 목적을 달성할 수 있는지 평가



평가 지표

슬롯 매칭도

목표 달성도

턴 별 문맥 관련성

테스트 방법론 - DST

- 에이전트가 사용자 발화에서 정보를 추출하여 대화 상태를 추적 하는지 평가

사용자 발화: 서울 호텔 예약하고 싶어요.

상태	값
숙박 위치	서울

에이전트 발화: 인원은 3명이에요.

상태	값
숙박 위치	서울
숙박 인원	3명

 **PASS** 모든 정보 추적 성공

사용자 발화: 서울 호텔 예약하고 싶어요.

상태	값
숙박 위치	서울

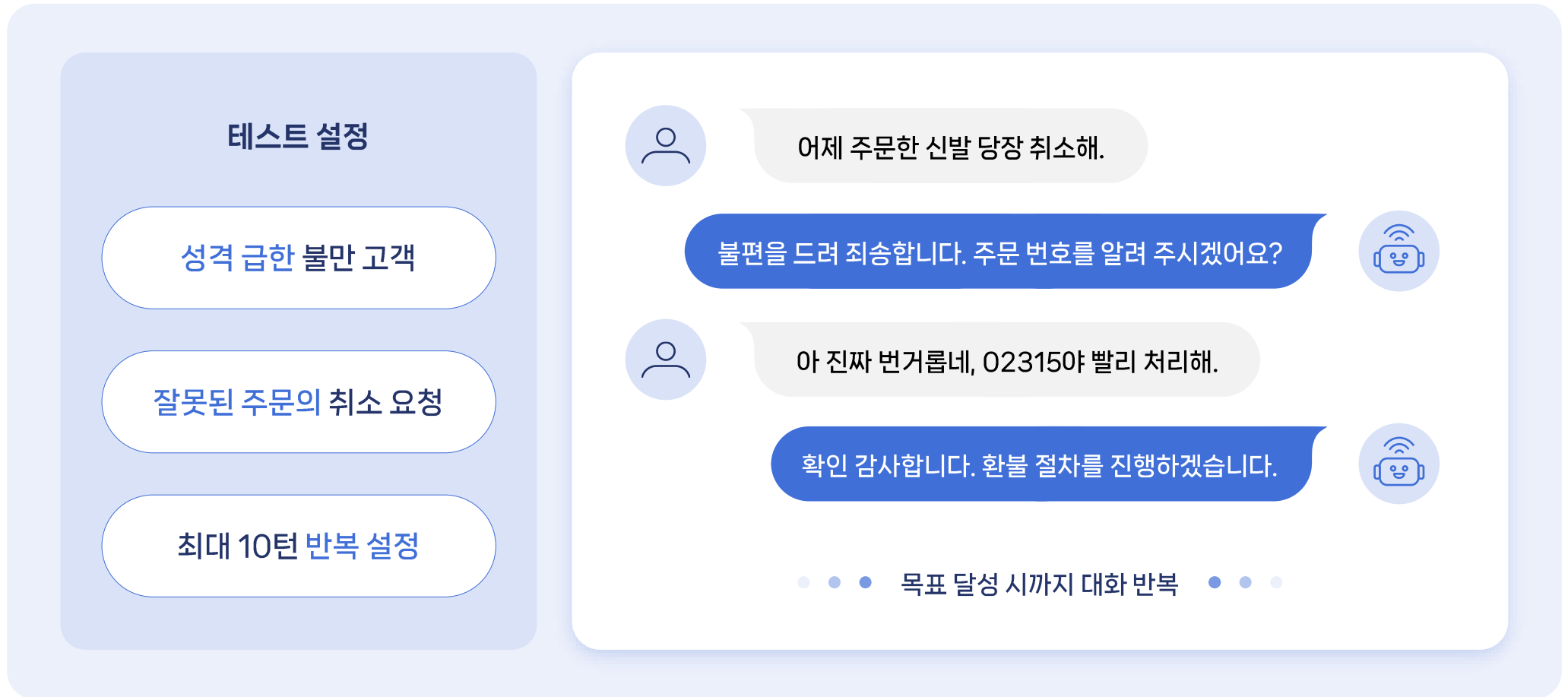
에이전트 발화: 인원은 3명이에요.

상태	값
숙박 위치	서울
숙박 인원	3명

 **FAIL** 이전 대화 정보 유실

테스트 방법론 - CX

- 가상 고객을 시뮬레이션 하여, 대상 에이전트의 종합적인 고객 경험을 평가



평가 지표

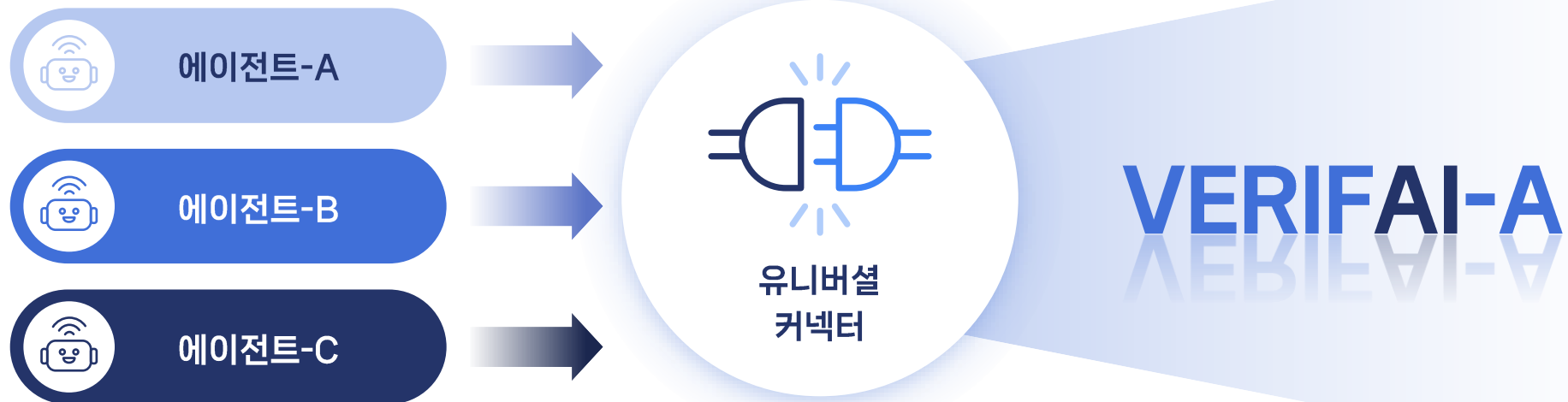
목표 달성도

턴 별 문맥 관련성

지식 유지

유니버설 커넥터

- 다양한 에이전트를 표준화된 자체 인터페이스로 하나의 플랫폼에서 통합 관리



LLM 기반 테스트 케이스 자동 생성

- 시나리오를 입력하면, LLM이 각 방법론에 맞는 테스트 케이스를 자동으로 생성

증강 예시

시나리오

쇼핑몰에서 구입한
운동화의 반품
규정과 방법에 대한
문의 및 반품 진행



운동화 반품할래



단순 변심에 의한 반품은
배송 완료 후 3일 이내에
가능합니다. 마이페이지 내
'주문/배송 조회' 메뉴에서
'반품 신청'을 클릭하여
접수해주시기 바랍니다.
보다 자세한 내용은..



어제 받은 운동화 반품
할텐데! 어떻게 해야
하냐?!?!?!?



주문번호 확인이 필요합니다.



아 귀찮게 뭘 자꾸 물어봐,
어제 주문한거 하나밖에
없잖아!!!!!!



어제 받은 운동화를 반품하고
싶은데 절차가 어떻게 됩니까?



단순 변심에 의한 반품은
배송 완료 후 3일 이내에
가능합니다. 마이페이지 내
'주문/배송 조회' 메뉴에서
'반품 신청'을 클릭하여
접수해주시기 바랍니다.
보다 자세한 내용은..

문서 기반 테스트 케이스 자동 생성

- 제품 / 고객 응대 매뉴얼 등 문서를 LLM이 분석하여 테스트 케이스를 자동으로 생성

지원 포맷

PDF, DOCX, DOC, PPT, HWP, HWPX, TXT, MD, CSV

증강 예시

Google Cloud Marketplace 제품의 요금제

Cloud Marketplace 제품을 제공하는 공급업체는 제품의 요금제를 설정하는 방법을 선택할 수 있습니다. 제품 유형에 따라 구독료, 사용량 요금 또는 이 두 가지의 조합이 청구될 수 있습니다.

이 페이지에는 각 유형의 Cloud Marketplace 제품에 사용할 수 있는 요금제 유형이 나열되어 있습니다.

SaaS 제품 요금제

Software as a service (SaaS) 제품의 요금제 유형은 다음과 같습니다.

월간 구독 또는 특정 기간 구독

공급업체는 베이직, 스탠더드, Pro 요금제와 같은 여러 정기 결제 요금제를 제공할 수 있습니다.

참고: 1년과 같은 특정 구독 기간을 선택한 경우 언제든지 서비스 수준을 높이거나 더 긴 구독 기간을 선택할 수 있습니다. 그러나 더 짧은 구독 기간을 선택하거나 다운그레이드하거나 서비스를 취소하려는 경우 구독 기간이 완료된 후에 변경사항이 적용됩니다. 예를 들어 고객이 현재 스탠더드 1년 요금제인데 스탠더드 월간 요금제 또는 베이직 요금제로 전환하려는 경우 스탠더드 요금제에서 1년이 완료된 후에 변경사항이 적용됩니다.

사용량 기반

공급업체가 설정한 측정 기준에 따른 요금입니다. 예를 들어 일부 공급업체는 앱에서 처리하는 데이터 양에 따라 요금을 청구할 수 있습니다.

구독 및 사용 요금의 조합

예를 들어 공급업체는 베이직 요금제에 대해 월 \$10의 요금을, 처리되는 모든 GiB 데이터에 대해 \$0.01의 추가 요금을 청구할 수 있습니다.

Cloud Marketplace는 구독 요금을 특정 Google Cloud 프로젝트와 연결하지 않습니다. Cloud Marketplace는 사용 요금을 자동으로 생성된 Google Cloud 프로젝트에 연결합니다. 이 프로젝트의 ID는 `pr-`로 시작하고 그 뒤에 고유한 일련의 문자와 숫자가 옵니다. 보고서의 정확성을 유지하려면 이 Google Cloud 프로젝트에서 Cloud Billing 계정으로 연결되는 링크를 삭제하지 마세요.



Google Cloud Marketplace의 SaaS 제품 요금제 유형에는 어떤 것들이 있습니까?

SaaS 제품 요금제 유형에는 월간 구독 또는 특정 기간 구독, 사용량 기반, 구독 및 사용 요금의 조합이 있습니다.



SaaS 제품의 특정 기간 구독을 선택한 경우 서비스 수준을 변경할 수 있는 조건은 무엇입니까?

특정 구독 기간을 선택한 경우 언제든지 서비스 수준을 높이거나 더 긴 구독 기간을 선택할 수 있지만, 더 짧은 기간을 선택하거나 다운그레이드하거나 서비스를 취소하려는 경우 구독 기간이 완료된 후에 변경사항이 적용됩니다.



사용량 기반 요금제는 어떻게 설정됩니까?



사용량 기반 요금제는 공급업체가 설정한 측정 기준에 따라 요금이 청구됩니다. 예를 들어, 앱에서 처리하는 데이터 양에 따라 요금을 청구할 수 있습니다.

의미보존 데이터 증강

- 기존 테스트 케이스를 LLM 기반 5가지 기법으로 자동 변형하여 테스트 커버리지를 확장

원본 문장

안녕하세요, 주문한 상품을 교환하려고 하는데 절차를 자세히 안내해 주실 수 있을까요?

증강 기법	증강 예시
오타	안녕하세요, 혹시 주문한 상품의 교환절차에 대해 자세히 알려주실 수 있을까요?
동의어	안녕하세요, 혹시 요청한 상품 의 교환 과정에 대해 상세히 안내해 주실 수 있나요?
대체	안녕하세요, 혹시 구매한 상품 교환 관련 절차를 상세히 알려주시면 감사하겠습니다.
어조	안녕! 주문한 물건 바꾸려고 하는데, 어떻게 해야 할지 좀 가르쳐 줄래?
재구성	안녕하세요. 주문한 상품의 교환 절차를 자세히 알고 싶습니다. 혹시 설명해 주실 수 있나요?

평가자 스튜디오

- LLM-as-a-Judge 평가자부터 여러 평가자를 조합한 평가단을 구성으로 유연한 평가 환경 제공



G-Eval 평가자

- 평가 기준과 단계를 자연어로 정의하면, LLM이 해당 기준에 따라 평가

평가 설계

평가 기준

"무엇을 평가할 것인지" 명확한 정의

예

사용자의 의도를 파악하고, 사실에 기반한 답변을 제공하는지 평가함

평가 단계

평가자가 응답을 평가할 때 따라야 할 단계 정의 단계는 순차적으로 처리

예

- ① 사용자 질문의 핵심 의도를 파악한다.
- ② 근거 없거나 사실과 다른 내용이 없다
- ③ 위 분석을 종합하여 100점 만점으로 채점



평가 커넥터

결과 판정

통과 임계값

점수 기준값과 비교 연산자 설정

예

> 80

결과 도출

평가 단계에 따라 도출된 점수를 통과 임계값 기준으로 PASS / FAIL 도출

예

85 / 100점 **PASS**

채점 근거: 핵심 의도는 파악하였으나, 사실 확인 시 일부 정보가 부정확함

DAG 평가자

- 작업, 판단, 판정 노드를 그래프로 연결하여 분기 조건이 있는 복합 평가

노드 구성

작업 (Task)

LLM에 지침을 전달하여
입력 처리 및 결과 전달

이진 판단 (Binary)

기준에 따라 예 / 아니오로
평가 흐름 분기

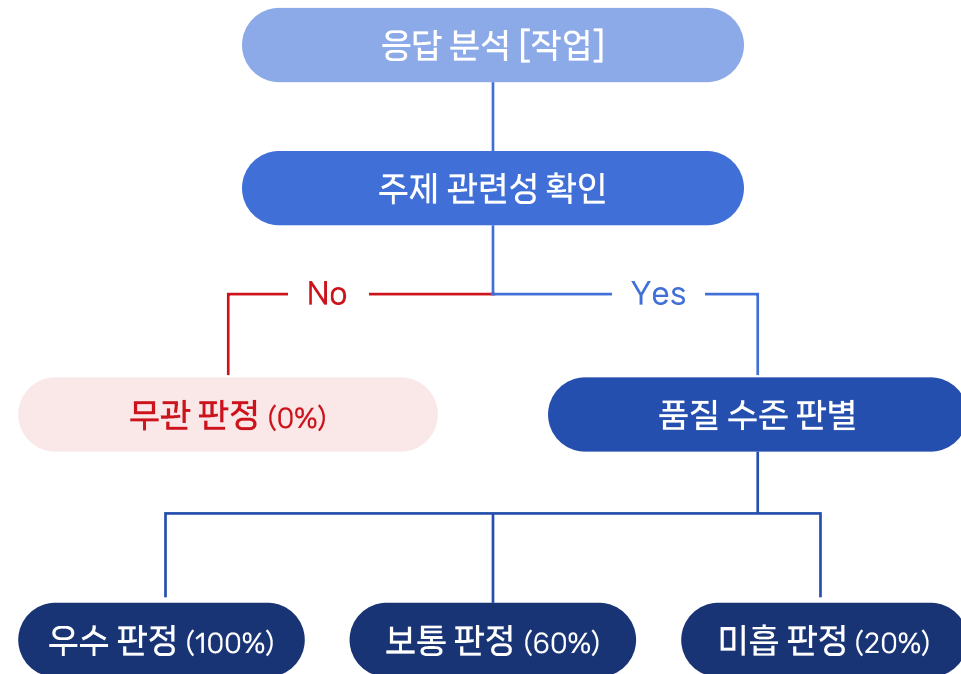
다중 판단 (Multi)

여러 선택지 중 하나를
선택해 세분화된 분기 처리

판정 (Result)

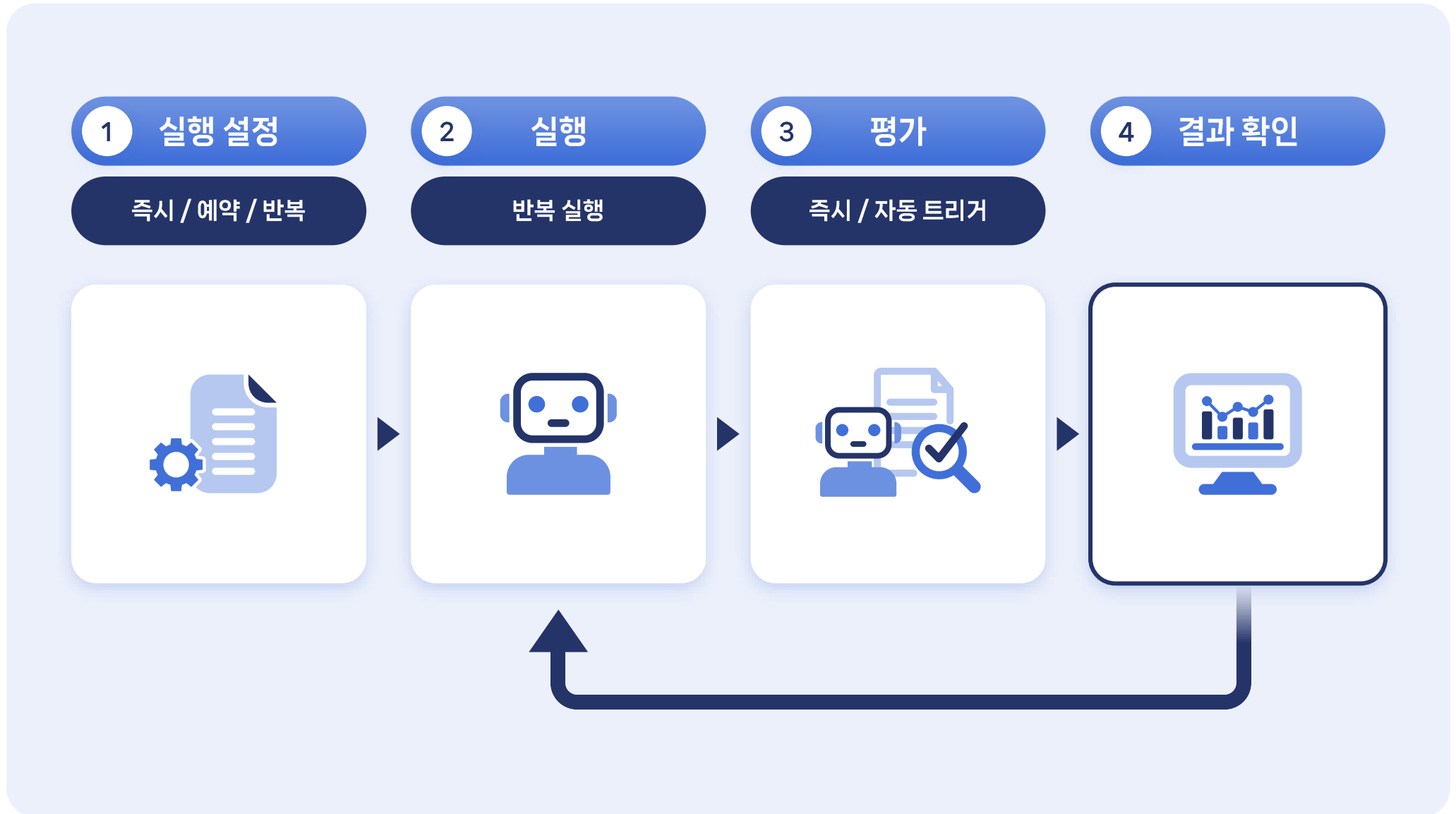
최종 점수를 부여하여
평가 종료 및 판정 완료

평가 그래프 설계 (예시)



파이프라인 자동화

- 응답 수집부터 평가까지 끊임 없는 자동화 흐름 구성

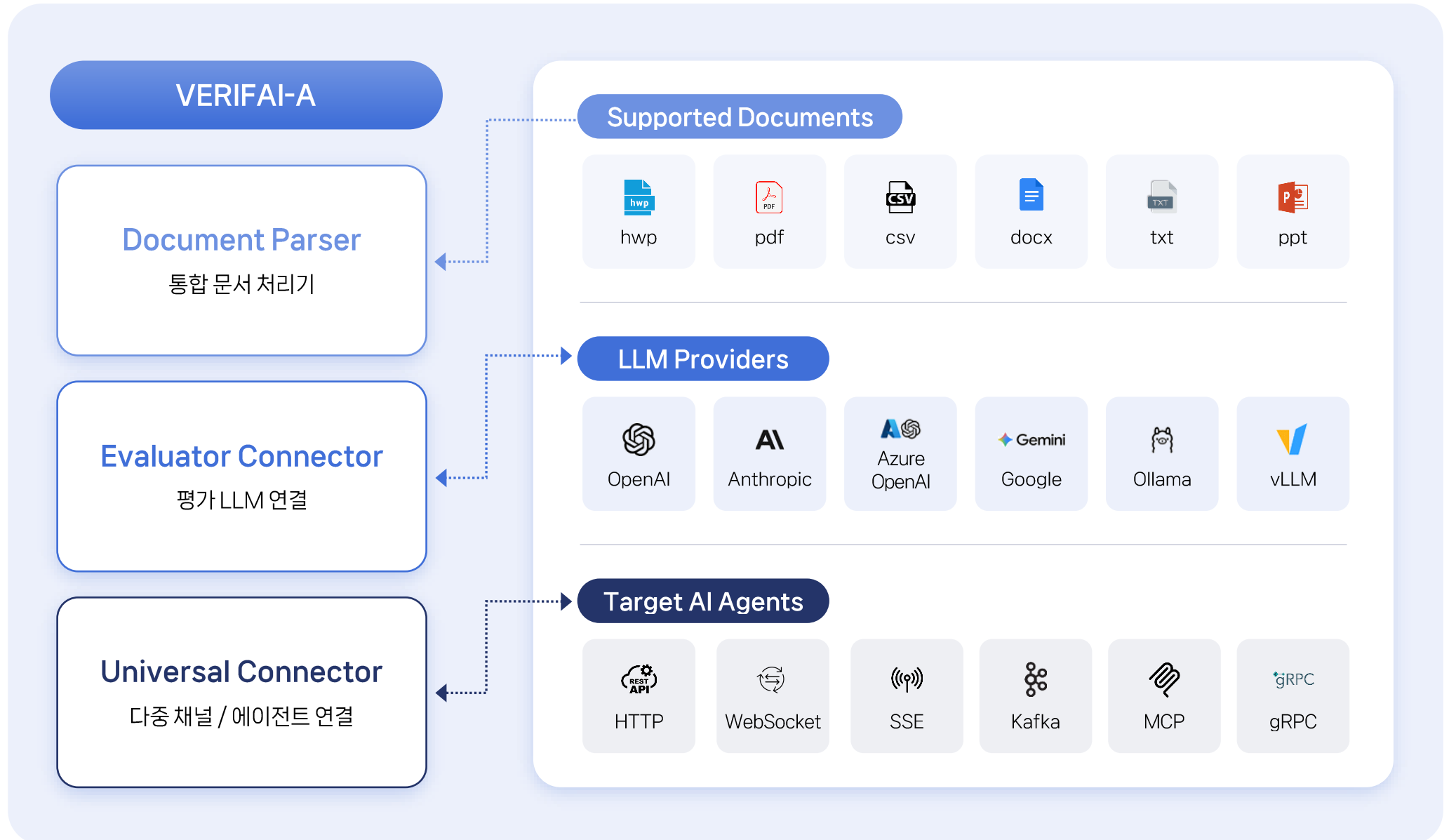


| 지원 환경

구분	세부 사항
지원 브라우저	Chrome
배포 방식	Azure, On-Premises SaaS
인증 방식	SSO
다국어	한국어



환경구성





Thank You