

E-평가자: 순차적 가설 검정을 통한 신뢰할 수 있는 에이전트 검증기

Shuvom Sadhuka^{1,2}, Drew Prinster^{1,3}, Clara Fannjiang¹, Gabriele Scalia¹, Aviv Regev¹, Hanchen Wang^{1,4}

¹ Genentech

² MIT

³ 존스 홉킨스

⁴ 스탠퍼드

2025년 12월 4일

초록

행위자형 AI 시스템은 사용자 프롬프트에 반응하여 추론 단계나 도구 호출과 같은 일련의 행동을 실행합니다. 이러한 궤적의 성공 여부를 평가하기 위해 연구자들은 LLM 판정자나 프로세스-보상 모델과 같은 검증기를 개발하여 에이전트 궤적 내 각 행동의 품질을 점수화합니다. 이러한 휴리스틱 점수는 유용한 정보를 제공할 수 있지만, 에이전트가 성공적인 출력을 생성할지 여부를 결정하는 데 사용될 때 정확성을 보장하지는 않습니다. 여기서 우리는 블랙박스 검증기 점수를 거짓 경보율을 증명 가능하게 제어하는 결정 규칙으로 변환하는 방법인 *e-평가기*를 소개합니다. 우리는 성공적인 궤적(즉, 사용자의 프롬프트에 대한 올바른 응답으로 이어질 행동 시퀀스)과 실패한 궤적을 구별하는 문제를 순차적 가설 검정 문제로 정의합니다. *e-평가기*는 *e-프로세스*의 도구를 기반으로 에이전트 궤적의 모든 단계에서 통계적 유효성을 유지하는 순차적 가설 검정법을 개발하여, 임의로 긴 행동 시퀀스에서 에이전트를 온라인으로 모니터링할 수 있게 합니다. 경험적으로, 우리는 6개 데이터셋과 3개 에이전트에 걸쳐 *e-평가기*가 다른 전략보다 더 큰 통계적 검증력과 더 우수한 오경보율 제어를 제공함을 입증합니다. 또한 *e-평가기*를 활용하면 문제적 궤적을 신속히 종료하고 토큰을 절약할 수 있음을 추가로 입증합니다. 종합적으로 *e-평가기*는 검증자 휴리스틱을 통계적 보증을 갖춘 결정 규칙으로 변환하는 경량화되고 모델에 무관한 프레임워크를 제공하여, 보다 신뢰할 수 있는 에이전트 시스템의 배포를 가능하게 합니다.

1 서론

에이전트는 일련의 행동을 실행하여 자율적으로 작업을 수행하는 블랙박스 시스템으로, 이를 통칭하여 궤적(*trajectory*)이라 부릅니다. 이러한 궤적에는 도구 호출을 통한 외부 환경과의 상호작용, 코드 작성, 논리적 추론 단계 등 다양한 행동이 포함될 수 있습니다. 본 논문에서는 일반적으로 사용자 요청에 응답하는 대규모 언어 모델(LLM) 기반 에이전트를 지칭하지만, 에이전트 패러다임은 더 광범위하게 적용 가능합니다. 예를 들어, 일련의 기계적 동작을 통해 물리적 작업(예: 컵 집기)을 수행하는 로봇 에이전트[3], 게임 규칙에 부합하는 일련의 동작을 통해 게임(예: 포커 또는 체스)을 플레이하는 게임 에이전트[5, 51] 등이 포함됩니다. 에이전트는 폭넓게 응용될 수 있으며, 신약 개발[14, 52], 세포 생물학 및 유전체학[61, 19], 가설 검증[18] 등 다양한 분야에서 유망한 초기 결과를 보여주고 있습니다.

그럼에도 불구하고 에이전트는 실수를 저지르며, 이를 탐지할 수 있는 능력이 중요하다. 이를 위해 검증기 모델이 개발되어 에이전트의 궤적 내 각 행동에 수치 점수를 부여한다. 이러한 점수는 일반적으로 궤적이 올바른 최종 출력을 성공적으로 생성할 확률의 대리 지표로 사용된다. 검증기 예시로는 각 단계 후 점수(텍스트 출력)를 제공하는 판정형 대규모 언어 모델(LLM) [28]과, 궤적 내 각 단계가 "정확한"지 "부정확한"지에 대한 확률적 예측을 제공하도록 미세 조정된 프로세스-보상 모델 [29, 73, 31]이 있습니다. 이러한 검증기의 점수는 최종 출력이 부정확할 것으로 예상되는 실패한 궤적을 식별하는 데 활용될 수 있습니다.

현재 검증기의 주요 한계는 궤적의 오류를 표시할지 여부를 결정하는 데 점수를 활용할 때, 이 하류 결정의 오류 확률에 대한 보장이 전혀 없다는 점이다. 자율주행 연구실[27], 유전자 편집[42], 병원 운영[13] 등 현실 세계에 영향을 미치는 고위험 환경에 에이전트가 배치될 경우, 엄격한 보증은 특히 중요해질 수 있다.

특히, 우리는 오경보율, 즉 성공적인 궤적을 실패한 것으로 잘못 표시할 확률에 주목한다. 검증자 점수가 한계 보정과 같은 확률적 "정확성"에 대한 통념을 충족하더라도, 그러한 개념들은 오경보율에 대한 보장을 제공하지 않는다. 더욱이, 궤적 내 각 행동은 시간과 자원을 소모하며, 궤적은 길어질 수 있다. 따라서 우리는 궤적이 실패할 것임을 가능한 한 적은 행동 후에 조기에 탐지하고자 합니다. 전체 궤적의 비용을 감수한 후에야 결정을 내리는 대신 말이죠. 이러한 요구사항은 각 행동 후 검증기 점수를 평가할 것을 필요로 합니다. 이 경우, 특히 완전한 궤적의 길이가 사전에 알려지지 않기 때문에, 언제든 오경보를 발생시킬 확률에 대한 보증을 얻는 방법이 불분명합니다.

정밀 조정이나 더 나은 검증기 구축은 이러한 문제를 직접 해결하지 못합니다. 더욱이 배포 환경을 위한 검증기 정밀 조정은 상당한 공학적 노력을 수반하며, 충분한 컴퓨팅 자원과 검증기(그리고 아마도 에이전트) 가중치에 대한 화이트박스 접근이 필요합니다. 충분한 컴퓨팅 자원과 검증기 가중치 접근 권한이 있더라도, 검증기 미세 조정에 적합한 충분한 데이터를 확보하는 것은 비현실적일 수 있습니다. 배포 환경을 대표하는 궤적뿐만 아니라, 모든 궤적 내 모든 행동에 대한 "정확성" 레이블이 필요합니다[31].

이러한 과제를 해결하기 위해, 우리는 가벼운 통계적 래퍼인 e -평가기를 도입합니다. 이 래퍼는 블랙박스 검증기의 점수를 실패한 에이전트 궤적을 탐지하기 위한 결정 규칙으로 변환하며, 오탐률에 대한 보증을 제공합니다. 이를 위해 먼저 문제를 가설 검정으로 설정합니다. 검증기 점수 시퀀스가 성공적 궤적(즉, 올바른 최종 출력을 생성할 궤적)에 대해서는 분포 $p_{(1)}$ 에서, 실패한 궤적에 대해서는 분포 p_0 에서 추출된다고 가정합니다. 새로운 궤적에 대해, 문제는 가능한 한 적은 수의 동작 후 검증기 점수 시퀀스가 p_1 에서 추출되었는지 $p_{(0)}$ 에서 추출되었는지 결정하는 것으로 귀결됩니다. p_1 과 $p_{(0)}$ 를 평가할 수 있다면, 각 시간 t 마다 지금까지 수집된 점수가 p_1 에서 유래했는지 $p_{(0)}$ 에서 유래했는지 결정하는 잘 연구된 순차적 가설 검정인 순차적 가능성비 검정[60]을 적용할 수 있을 것이다. 그러나 p_1 과 p_0 은 일반적으로 알려져 있지 않다.

따라서 e -evaluator는 세 단계로 작동합니다: (1) 배포 환경에서 에이전트 궤적(즉, 행동 시퀀스와 중간 상태), 단계별 검증자 점수, 결과 레이블(즉, 최종 출력이 정답인지 오답인지)로 구성된 소규모 보정 세트를 수집합니다. (2) 각 단계 t 에서 각각 성공적 궤적과 실패적 궤적에 대한 검증자 점수 분포인 p_1 과 p_0 사이의 밀도 비율 모델을 학습합니다. 여기서 $p_{(1)}$ 은 각 단계 t 에서 성공적 궤적에 대한 검증기 점수의 분포를, $p_{(0)}$ 은 실패한 궤적에 대한 검증기 점수의 분포를 나타냅니다. (3) 보정 세트의 훌드아웃 분할을 사용하여 실패한 궤적을 표시하기 위한 결정 임계값을 찾습니다. 이때 사용자가 지정한 α 수준(즉, 성공적 궤적이 실수로 실패한 궤적으로 표시되는 비율)에서 오경보율(제1종 오류)을 제어합니다. 중요한 점은 e -평가기가 검증기의 현재 및 향후 개선 사항을 보완한다는 것입니다: 모든 에이전트와 검증기 조합에 대해 오경보율 제어를 보장하면서도, 더 우수한 검증기와 함께 배포할 경우 더 강력한 결정 규칙을 도출하는 경향이 있습니다. e -평가기는 최소한의 컴퓨팅 자원을 필요로 하며 표준 노트북에서 실행 가능하며, 코드는 PyPi(<https://pypi.org/project/e-evaluator/>)에서 이용 가능한 Python 패키지로 공개합니다.

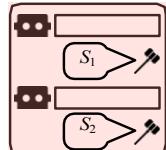
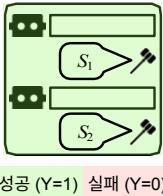
E -evaluator는 e -값과 순차적 가설 검정 접근법을 기반으로 합니다. e -값은 p -값의 대안으로, 일련의 가설 검정(예: "현재 진행 중인 궤적이 성공적인가?")을 수행하고자 하지만 사전에 결정 횟수를 알 수 없는 상황(궤적 길이가 가변적이기 때문)에서 특히 유용합니다.

실증적으로, 6개 데이터셋과 3개 에이전트에 걸쳐 e -평가기는 원시 검증기나 재교정 검증기 모델 단독과 같은 다른 기준 모델보다 더 우수한 오경보율(제1종 오류) 제어 및 통계적 결정력을 제공했습니다. e -평가기의 잠재적 활용 사례로 실패한 궤적의 조기 종료를 제시합니다. 이를 통해 토큰의 80%만으로 모델의 원래 정확도를 최대 90%까지 회복할 수 있습니다. 또한 비-LLM 환경(체스 엔진이 각 수 후 보드 상태를 수치화하여 점수화하는 경우)에서도 실험을 수행했습니다.

요약하면, 우리의 기여는 다음과 같습니다:

1. 에이전트 출력 검증 문제를 정식화하고, 기존 검증기 모델을 강화하기 위한 해결책으로 순차적 가설 검정을 제안한다.
2. 통제 가능한 오류율과 강력한 통계적 결정력을 통해 실패한 경로를 식별할 수 있는 통계적 래퍼인 e -평가기를 소개합니다.
3. e -평가기는 비-LLM 에이전트를 포함한 모든 블랙박스 에이전트-검증기 조합에 적용할 수 있습니다. e -평가기는 검증기에 대한 블랙박스 접근을 가정하므로, 검증기 모델을 직접 개선하는 방법(예: 미세 조정)이 이를 보완합니다.
4. e -평가기는 여러 데이터 세트, 검증기 및 에이전트에서 기준을 능가하는 성능을 보인다는 것을 경험적으로 입증합니다.

1. 궤적, 점수 및 결과 수집



성공 ($Y=1$) 실패 ($Y=0$)

2. 단계별 밀도 비율 학습 (e-process)

$$M_t = \frac{p(\mathbf{S}_{[1:t]} | Y=0)}{p(\mathbf{S}_{[1:t]} | Y=1)}$$

$$= \frac{p_0(\mathbf{S}_{[1:t]})}{p_1(\mathbf{S}_{[1:t]})}$$

3. 오경보율을 제어하기 위한 임계값 찾기

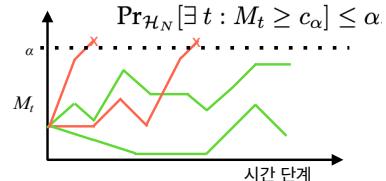


그림 1: e-평가기 개요. e-평가기는 세 단계로 작동한다. 첫째, 소규모 교정 세트인 궤적, 검증기 점수, 라벨을 수집한다. 둘째, 각 시간 단계 t 에서 밀도 비율 M_t 을 학습한다. 셋째, 빌의 부등식 또는 분위수 기반 경험적 접근법을 사용하여 오경보율을 제어하는 결정 임계값을 찾는다. 주어진 임계값에서, 실패한 궤적(빨간색)은 성공한 궤적(녹색)보다 더 높은 비율로 거부되어야 합니다.

2 방법

2.1 문제 설정

사용자 프롬프트 $o_{(0)}$ 이 주어지면, 에이전트는 $T \in \mathbb{N}_+$ 의 행동 시퀀스 (a_1, \dots, a_T) 를 실행한다. 여기서 $\mathbb{N}_+ := \{1, 2, \dots\}$ 이며, T 는 프롬프트와 에이전트의 내부 무작위성에 모두 의존하는 확률변수이다. 각 행동 $a_{(t)}$ 에는 행동 실행 후 환경 상태를 포착하는 관측값 o_t 가 연관됩니다(예: a_t 가 중간 산술 계산을 수행하는 경우, $o_{(t)}$ 에는 계산된 값이 포함될 수 있음). 행동과 관측은 함께 각 단계 t 에서 궤적 $H_t = (o_0, a_1, \dots, a_t, o_t)$ 를 형성합니다. 각 단계 t 이후, 블랙박스 검증기 모델 v 는 궤적 H_t 를 입력으로 받아 지금까지의 궤적 품질에 대한 휴리스틱 평가 역할을 하는 점수 $S_t = v(H_t)$ 를 제공합니다. 일반적으로 $S_t \in [0, 1]$ 이지만, e-evaluator는 모든 유형의 점수 값을 지원합니다. 검증자의 점수는 시퀀스 $\mathbf{S} = (S_1, \dots, S_T)$ 을 형성합니다.

완전한 궤적 H_t 는 최종 출력 o_t 을 바른지 ($Y=1$) 아닌지 ($Y=0$)에 대한 이진 레이블 $Y \in \{0, 1\}$ 과 연관된다.¹ 우리는 $Y=1$ 인 궤적을 성공적인 궤적이라 부르고, $Y=0$ 인 궤적은

실패한 궤적. 우리는 보정 대/이/제/D_{cal} = $\{(\mathbf{S}^{(i)}, Y^{(i)})\}_{i=1}^n$ 에 대한 접근이 가능하다고 가정한다. 여기서 $(\mathbf{S}^{(i)}, Y^{(i)})$ 는 변수 길이 점수 시퀀스와 그 레이블에 대한 공동 분포인 P 에서 독립 동일 분포(i.i.d)로 추출된 것으로, 밀도 함수 p 를 가집니다.

2.2 가설 검정을 통한 평가

새로운 점수 시퀀스 \mathbf{S} 가 주어졌을 때, 우리의 목표는 에이전트의 궤적이 최종적으로 올바른 ($Y=1$) 출력인지 잘못된 ($Y=0$) 출력인지 판단하는 것이다. 이 목표를 가설 검정으로 형식화한다. P_1 과 $P_{(0)}$ 는 각각 올바른 최종 출력과 잘못된 최종 출력에 조건부로 주어진 점수 시퀀스의 분포를 나타내며, 각각의 밀도 함수는 p_1 과 p_0 이다. 즉, $p_1(\mathbf{S}) = p(\mathbf{S} | Y=1)$ 및 $p_0(\mathbf{S}) = p(\mathbf{S} | Y=0)$ 이다. 우리는 $P_1 \neq P_0$ 이라고 가정하며, 다음 두 가설 사이에서 검정을 수행하고자 한다:

$$H_N : \mathbf{S} \sim P_1 \quad (\text{최종 출력은 올바름})$$

$$H_A : \mathbf{S} \sim P_0 \quad (\text{최종 출력이 틀림})$$

P_1 과 $P_{(0)}$ 는 일반적으로 시간에 따른 점수 간의 복잡한 의존성을 인코딩합니다. 즉, 블랙박스 검증기 점수는 일반적으로 매 단계마다 고정된 분포에서 독립적으로 추출된 표본이 아니며, 편리한 가정을 적용하기 어렵습니다.

우리는 각 단계 t 에서 실행 가능한 H_N 과 H_A 사이의 순차적 검정을 구축합니다. 이 검정은 오직 $\mathbf{S}_{[1:t]}$ 만을 사용하며, 이는 에이전트의 t 단계까지의 궤적만을 기반으로 한 점수입니다. 구체적으로, 우리는 일련의 검정 통계량을 구성합니다.

¹ 본 프레임워크는 무한 길이의 궤적도 수용하는데, 이는 유한 시간 T 에 대해 최종 출력 o_t 를 생성하지 않으므로 $Y=0$ 으로 표기한다. 그러나 실용적 관련성을 위해, 본 논문에서는 항상 최종 출력을 생성하는 유한 길이의 궤적 설정에만 초점을 맞춘다.

$(M_t)_{t=1}^T$ 여기서 M_t 는 $S_{([1:t])}$ 의 실수 값 함수이며, c_α 는 사용자가 주어진 실수 값 결정 임계값입니다. 지정된 오류 수준, $\alpha \in (0, 1)$. 순차 검정은 다음과 같이 진행됩니다(알고리즘 1). $t = 1, 2, \dots$ 대해, $M_t \geq c_{(\alpha)}$ 이면 $H_{(N)}$ 을 기각합니다. 에이전트의 궤적 끝점인 $t = T$ 에 도달할 때까지 H_N 을 기각하지 않았다면 $H_{(N)}$ 을 수용합니다. 우리의 목표는 다음 두 기준을 충족하는 것입니다:

1. 오경보율 제어:

$$\Pr_{H_N} [\exists t \in [T] : M_t \geq c_\alpha] \leq \alpha, \quad (1)$$

여기서 T 는 완전한 궤적의 길이이며, 또한 임의의 값을 가진다. $[T] := \{1, \dots, T\}$ 이다. 즉, 오경보율, 즉 성공적인 궤적을 거부하는 비율은 최대 α 이하로 보장된다. 이 비율은 고전적으로 제1종 오류로도 알려져 있다. 고려 중인 확률은 다음과 같다.

$M_t \geq c_\alpha$ 를 초과하는 경우 — 즉, T_N 에 무엇이든 간에, 에이전트가 취하는 총 행동 횟수 T 에 관계없이, 우리가 귀무 가설 $H_{(N)}$ 을 기각하는 경우 — 이 유효성 개념은 언제든지 유효성(*anytime validity*)으로 알려져 있다.

2. 높은 검정력: 실패한 궤적은 종종 기각된다.

즉, 실패한 궤적에 대한 높은 기각률 $\Pr_{H_A} [\exists t \in [T] : M_t \geq c_\alpha]$ 를 원한다.

우리는 첫 번째 기준의 고학률 버전을 만족하는 방법을 개발하며, 경험적으로

두 번째 기준을 달성한다. 첫 번째 기준의 도전 과제는 통계량 시퀀스 $(M_t)_{t=1}^T$ 를 구성하는 것이다. 그리고 사전적으로 완전한 궤적의 길이 T 를 알지 못하더라도 순차 검정이 유효하도록 하는 결정 임계값 c_α 가 존재한다. 현대 통계적 가설 검정에서 표준 도구인 p -값은

이러한 설정에는 본질적으로 적합하지 않다. 직관적으로, p -값의 정의—모든 $\alpha \in [0, 1]$ 에 대해 $\Pr_{H_N} (q \leq \alpha) \leq \alpha$ 를 만족하는 확률 변수 q —는 그 자체로 특정 단계에서 생성된 p -값이 이전 단계의 p -값에 어떻게 의존하는지에 대해 아무것도 암시하지 않는다. 이러한 의존성은 데이터(여기서는 $S_{([1:t])}$)가

각 단계에서 관측된 값들은 상호 의존적일 뿐만 아니라, 각 단계의 데이터가 독립적이라 하더라도 단순히 이전 값들을 기반으로 다음 p -값을 생성할지 여부를 결정하는 과정 자체에 의존한다. 이러한 의존 구조에 대한 지식 없이 추가 가정과 이를 활용하는 기술[12, 22, 45, 53] 없이서는 어떤 기각의 확률을 평가해야 할지 불분명하다(자세한 논의는 부록 참조). 대신 우리는 귀무 가설에 대한 증거를 정량화하는 또 다른 대상인 *e-process*[44, 46]로 눈을 돌린다. 이 정의는 귀무 가설 하에서 검정 통계량의 현재 값과 과거 값 사이의 관계를 명시적으로 특성화하며, 이러한 설정에서 확률적 추론을 자연스럽게 가능하게 한다. $H(N)$ 에 대한 *e-process*는 $(E(t))_{t=1}^T$ 순서로, 각 단계에서 관측된 값들이 서로 의존적일 뿐만 아니라, 각 단계의 데이터가 독립적이라

현재 검정 통계량과 과거 값 사이의 관계를 자연스럽게 확률적 추론을 가능하게 하는 방식으로 명시적으로 특성화합니다. H_N 에 대한 *e*-과정은 $(E_t)_{t=1}^T$ 라는 시퀀스로, 각 E_t 는 H_N 에 대한 *e*-값입니다. 즉,

$E_{H_N}[E_t] \leq 1$ —그리고 H_N 에 대한 시험 마팅게일(M_t)이 존재하여 항상 $E_t \leq M_t$ 를 만족한다. 시험 마팅게일

$H_{(N)}$ 에 대한 시험 마팅게일은 두 조건을 만족하는 순열 $(M_t)_{t=1}^T$ 이다:

1. 비음성과 단위 평균: 모든 t 에 대해 M_t 는 비음수이며, $E_{H_N}[M_0] \leq 1$ 이다.

2. 마팅게일: $(M_t)_{t=1}^T$ 는 H_N 에 대한 마팅게일이다. 본 연구 설정에서 이는 다음을 의미한다: $E_{H_N}[M_t | M_0, M_1, \dots, M_{t-1}] = M_{t-1}$ 모든 t 에 대해 성립함을 의미한다.

모든 검정 마팅게일은 *e*-과정이지만, *e*-과정은 훨씬 더 광범위한 과정의 범주임을 유의하십시오.

테스트 통계량 시퀀스 $(M_t)_{t=1}^T$ 를 *e*-과정으로 구성함으로써 우리는 특정 특성을 활용할 수 있다.

각 단계 t 에서 전체 시퀀스에 대해 확률적으로 추론할 수 있는 속성들로, 에이전트가 최종적으로 취하는 총 단계 수와 무관합니다. 특히, 빌의 부등식은 모든 *e*-프로세스 $(M_t)_{t=1}^T$ 에 대해 다음을 나타냅니다.

$\Pr_{H_N} [\exists t : M_t \geq 1/\alpha] \leq \alpha$ 모든 $\alpha \in [0, 1]$ 에 대해 성립합니다. 즉, 최소한 $1 - \alpha$ 의 확률로 전체 시퀀스—

비록 무한히 진행된다 하더라도—모든 $\alpha \in [0, 1]$ 에 대해 $1/\alpha$ 미만이다. 이 부등식은 *e-process*가 왜 무효 가설에 대한 시간에 따른 증거의 정량화로 해석될 수 있는지 구체적으로 보여준다.

귀무가 참이라면, 높은 확률($1 - \alpha$)로 이 증거는 영원히 낮게 유지될 것이다(구체적으로 $1/\alpha$ 미만). 대립가설이 참이라면, 잘 설계된 *e*-과정은 잠시 후 설명하듯이 커진다.

² 보다 형식적으로, 마팅게일은 모든 t 에 대해 $E_{H_N}[M_t | F_{t-1}] = M_{t-1}$ 을 만족시키며, 여기서 $(F_t)_{t=1}^T$ 는 필터링이다. 설명의 명확성과 공간 제약으로 인해 여기서는 완전히 엄밀한 처리를 피하지만, 대략적으로 이는 F_{t-1} 에 이용 가능한 모든 정보에 조건을 거친다는 것을 의미하며, 이는 과거의 정확한 값에 조건을 거치는 것보다 더 많거나 적은 정보일 수 있다. 이는 F_t 가 소위 자연 여과(natural filtration)일 때 과거의 정확한 값에 조건을 거치는 것과 동등하다

우리 설정에서 e-프로세스를 구성하기 위해, 우리는 무효 밀도 p_1 과 대안 밀도 p_0 사이의 비율을 사용한다. 구체적으로, $M_0 = 1$ 로 설정하고, 각 단계 $t \in [T]$ 에 대해

$$M_t = \frac{p_0(\mathbf{S}_{[1:t]})}{p_1(\mathbf{S}_{[1:t]})}. \quad (2)$$

밀도 비율 과정은 시험 마팅게일이며 따라서 e-과정[44]이다(증명 참조). 이 선택된 M_t 및 $c_\alpha = 1/\alpha$ 를 사용하는 알고리즘 1은 언제나 유효한 오경보율 제어(정리 1)를 달성할 수 있게 한다.

알고리즘 1 에이전트 검증을 위한 순차적 가설 검정 프레임워크.

입력: 사용자 프롬프트, o_0 ; 각 단계 t 에 대해 점수를 입력으로 받아 검정 통계를 출력하는 함수, $M_t(\cdot)$, $t = 1, \dots$; 결정 임계값, $c_\alpha \in \mathbb{R}$.

출력: 결정, H_N 기각 또는 H_N 수용.

```

1:  $c_\alpha \leftarrow 1/\alpha$ 
2:  $H_0 \leftarrow (o_0)$ 
3:  $t = 1, \dots, T$ 에 대해
4:    $H_{t-1}$  인 경우, 에이전트는 행동  $a_{(t)}$  를 실행하여 환경 상태  $o_{(t)}$  를 생성한다.
5:   궤적 업데이트,  $H_t \leftarrow (o_0, a_1, o_1, \dots, a_t, o_t)$ .
6:   검증자 점수 획득:  $S_t \leftarrow v(H_t)$ .
7:   이 단계에 대한 검정통계량  $M_t(\mathbf{S}_{[1:t]})$ 를 평가한다
8:    $M_t(\mathbf{S}_{[1:t]}) \geq c_\alpha$ 인 경우  $\triangleright \mathbf{S}_{[1:t]} := (S_1, \dots, S_t)$ 
9:     반환:  $H_N$ 기각
10:    end if
11: for 종료
12:  $H_N$ 을 수락함

```

제안 1. 임의의 고정된 $\alpha \in (0, 1)$ 에 대해, 밀도 비율 과정 $M_t = p_0(\mathbf{S}_{[1:t]})/p_1(\mathbf{S}_{[1:t]})$, 그리고 결정 임계값 $c_\alpha = 1/\alpha$ 를 사용하는 알고리즘 1은 언제나 유효한 하위경보율 제어(식 1)를 달성한다.

증명. (개요) 밀도 비율 과정 $(M_t)^T$ 는 e-과정이다. 하위경보율에 대한 언제나 유효한 제어는

에 대한 언제나 유효한 제어는 빌의 부등식에 의해 따릅니다. 완전한 증명은 부록 8.1을 참조하십시오. \square

제1정리는 언제나 유효한 오경보율 제어를 확립하지만, 왜 다른 e-과정 대신 밀도비 과정을 선택해야 하는지는 설명하지 못한다. $H_{(A)}$ 하에서 밀도비 과정은 모든 e-과정 중 시간이 지남에 따라 (기대값 기준) 가장 빠르게 성장하는데, 이 개념을

제2정리에서 정밀하게 정의된 로그최적성(log-optimality)이라고 한다. 로그최적성은 비순차적 가설검정에서 "가장 강력한" 검정통계량과 유사하다: 직관적으로, $M_{(t)}$ 는 의사결정 임계값을 초과하는 경향이 있으며, 이에 따라 다른 e-과정보다 더 일찍 실패한 궤적을 팀지할 수 있게 한다.

제2정리. 밀도 비율 과정, $M = \frac{p_0(\mathbf{S}_{[1:t]})}{p_1(\mathbf{S}_{[1:t]})}$, 는 로그 최적이다. 즉, 다른 어떤 e-과정 $(M')^T$ $\triangleright t=0$

및 정지 시간 τ 에 대해, $E_{H_A} [\log M_\tau] \geq E_{H_A} [\log M'^\tau]$ 가 성립한다.

증명. (개요) 이는 Neyman-Pearson 정리[38]의 순차적 아날로그이다. [44]에도 제시된 이 증명은 먼저 특정 M_t 가 모든 e-변수 중 그 최적 e-변수임을 주목함으로써 시작된다. 이후 e-과정의 성질을 활용하여 증명을 완성할 수 있다. 완전한 증명은 부록 8.1을 참조하라. \square

2.3 밀도 비 추정

실제 적용 시 p_1 과 p_0 의 형태는 일반적으로 알려져 있지 않습니다. 따라서 제안된 방법인 e-평가기는 배포 전에 보정 단계(calibration phase)를 거치며, 이 단계에서 보정 데이터 $D_{(cal)}$ 를 활용하여

밀도 비율, $M^t(S_{[1:t]}) \approx \frac{p_0(S_{[1:t]})}{p_1(S_{[1:t]})}$ 의 모델을 학습합니다(알고리즘 2). 이를 위해 분류기 기반 밀도 비율

추정, 베이즈 법칙에 의한 다음 등식에 기반한 접근법 [4, 16]:

$$M_t = \frac{p_0(S_{[1:t]})}{p_1(S_{[1:t]})} = \frac{p(Y=0|S_{[1:t]}) p(Y=1)}{p(Y=1|S_{[1:t]}) p(Y=0)}. \quad (3)$$

구체적으로, 각 시간 단계 t 에 대해, 입력으로 $S_{[1:t]}$ 를 받아 $p(Y=1|S_{[1:t]})$ 의 추정값을 제공하는 분류기 f_t 를 훈련합니다. 또한 클래스 확률 $p(Y=1)$ 의 추정값 π^t_1 을 형성합니다. 이 두 추정값을 식 (3)에 대입하여 단계 t 에서의 추정 밀도 비율을 다음과 같이 구합니다:

$$M^t = \frac{1 - f_t(S_{[1:t]})}{f_t(S_{[1:t]})} \frac{\pi^t_1}{1 - \pi^t_1}.$$

이러한 추정된 밀도 비율 M^t 와 빌의 부등식에 기반한 동일한 결정 임계값 $c_\alpha = 1/\alpha$ 를 사용하여 알고리즘 1을 실행하는 것을 $1/\alpha$ 일계값을 가진 e -평가기라고 부를 것이다.

진술 1과 진술 2의 보장은 진정한 밀도 비율, 즉 M^t $= \frac{p_0(S_{[1:t]})}{p_1(S_{[1:t]})}$ 각 점 S 에 대해

와 각 단계 t 에 대해 학습할 때 적용된다.

실험에서 우리는 각 단계의 분류기 f_t 에 대해 단순 로지스틱 회귀를 사용했으며, 수백 개의 교정점에서 학습된 추정 밀도 비율이 경험적으로 오경보율과 오탐지를 모두 달성하고 다른 방법보다 우수한 성능을 보인다는 것을 발견했습니다(자세한 내용은 부록 8.2.2 참조). 경보율 제어와 우수한 성능을 모두 달성을 발견했습니다(자세한 내용은 부록 8.2.2 참조).

알고리즘 2 교정 데이터를 이용한 밀도 비율 추정.

입력: 보정 데이터, D_{cal} .

출력: 모든 단계에 대한 밀도 비율을 추정하는 함수, $\{M^t\}_{t \in \mathbb{N}_+}$.

1: D_{cal} 을 D_{DRE} 와 $D_{threshold}$ 로 무작위 분할하며, $D_{DRE} \cup D_{threshold} = D_{cal}$ 이 성립한다.

2: $\pi^t_1 \leftarrow \frac{1}{|D_{DRE}|} \sum_{(S, Y) \in D_{DRE}} Y \quad \triangleright$ 경험적 빈도로 클래스 확률을 추정합니다.

3: $T_{max} \leftarrow \max\{T : D_{DRE} \text{내 길이 } T \text{의 성공 및 실패 점수 시퀀스가 모두 존재하는 경우}\}$

4: **for** $t = 1, \dots, T_{max}$ **do**

5: $D_{DRE,t} \leftarrow \{(S_{[1:t]}, Y) : (S, Y) \in D_{DRE} : |S| \geq t\}$

6: $D_{DRE,t}$ 를 사용하여 $S_{[1:t]}$ 를 입력으로 받아 $p(Y=1 | S_{[1:t]})$ 를 예측하는 확률적 분류기 f_t 를 훈련한다.

7: $\hat{M}^t(\cdot) \leftarrow \frac{1 - f_t(\cdot)}{f_t(\cdot)} \frac{\pi^t_1}{1 - \pi^t_1}$

8: **for** 종료

9: 모든 $t > T_{max}$ 에 대해 $M^t(\cdot) \leftarrow \hat{M}^t(\cdot)$ 로 설정한다.

2.4 분위수 추정법을 통한 전력 증가

진밀도비 과정 M_t 가 주어졌을 때, M_t 가 결정 임계값 $c_\alpha = 1/\alpha$ 를 초과할 때마다 $H_{(N)}$ 을 기각하면 언제나 유효한 오경보율 제어가 달성된다(정리 1). 그러나 실제로는 추정된 밀도비를 사용하므로, 임계값 c_α 는 밀도비 추정 오차를 고려해야 한다. 또한, 언제나 유효성을 달성하기 위해 $c_{(0)}$ 를 설정하는 것은 지나치게 보수적일 수 있다. 이는 에이전트가 무한히 행동을 계속 실행하더라도 허위경보율의 상한을 보장하기 때문이다. 반면 실제 에이전트는 일반적으로

유한한 수의 행동 후 최종 출력을 생성하거나 종료됩니다.

따라서 우리는 언제든지 오경보 제어를 높은 확률로 유지하면서 이 두 요소를 모두 고려하기 위한 결정 임계값 c_α 설정의 대체 절차를 제안한다. 먼저 보정 집합 D_{cal} 을 D_{DRE} 와 $D_{threshold}$ 로 분할한다. 여기서 D_{DRE} 는 먼저 밀도 비율 추정(Alg. 2)에 사용되며, 이후 $D_{threshold}$ 를 사용하여 $c_{(0)}$ 를 다음과 같이 설정한다.

(알고리즘 2)에 사용되고, 이후 $D_{threshold}$ 는 다음과 같이 c_α 를 설정하는 데 사용됩니다.

먼저, 과정 M^t 가 $c_{(0)}$ 를 초과하는 경우 $H_{(N)}$ 을 기각하는 것은 $\max_t M^t \geq c_\alpha$ 일 때 $H_{(N)}$ 을 기각하는 것과 동등합니다.

c_α . 따라서 c_α 를 귀무가설 하에서 $\max_t M^t$ 분포의 $(1 - \alpha)$ 분위수로 설정하는 것으로 충분하다.

따라서 우리는 $D_{threshold}$ 내의 성공적 궤적에 주목하며, 이는 H_N 에 해당한다. 보정 집합 내 각 궤적 i 에 대해, 모든 단계에 걸쳐 추정된 최대 밀도 비율 $M^{(i)} = \max\{\hat{M}_1^{(i)}, \dots, \hat{M}_T^{(i)}\}$. 이 최대값 $M^{(i)}$ 는 귀무 분포의 표본이다. 이러한

표본을 바탕으로, 우리는 $q^{(1-\alpha)}$ 를 구성하는데, 이는 무효 분포의 $(1 - \alpha)$ -분위수에 대한 높은 확률 상한이다. (알고리즘 3). 추정된 밀도 비율 $\hat{M}_t^{(i)}$ 와 결정 임계값을 이 고확률 상한 $c_\alpha = q^{(1-\alpha)}$ 로 설정하여 알고리즘 1을 실행하는 것을, 다음 보증에 따라 '확률적으로 근사적으로 정확한(PAC)' 임계값을 가진 e-평가기라고 부른다. 즉, 사용자가 지정한 δ 에 대해 최소 $1 - \delta$ 의 확률로(즉, "확률적으로"), 이 절차는 사용자가 지정한 α

("대략적으로 정확함") 조건 하에서 언제나 오경보율을 제어할 수 있음을 의미한다.

제안 3. 알고리즘 2에 의해 출력된 추정 밀도 비율 함수를 $\{\hat{M}_t^{(i)}\}_{t \in \mathbb{N}_+}$ 로 표기한다. 고정된 오차 수준 $\delta \in (0, 1)$ 및 사분위수 수준 $\alpha \in (0, 1)$ 에 대해, c_α 를 알고리즘 3의 출력으로 정의한다. 그러면 $\{\hat{M}_t^{(i)}\}_{t \in \mathbb{N}_+}$ 와 결정 임계값 c_α 를 사용하는 알고리즘 1, 즉 PAC(probably-approximately-correct) 임계값을 가진 e-평가기는 다음을 만족한다.

$$Pr_{D_{cal}}(Pr_{H_N}(\exists t \in [T] : M_t \geq c_\alpha | D_{cal}) \leq \alpha) \geq 1 - \delta \quad (4)$$

증명. (개요) 높은 확률로 $q^{(1-\alpha)} \geq q_{1-\alpha}$ 를 만족하는 $q^{(1-\alpha)}$ 를 찾는 것은 교정 집합의 무효 표본들 중에서 순서 통계량 $M_{(k)}$ 가 $q_{1-\alpha}$ 보다 높을 확률이 높은 지표 k 를 찾는 문제로 귀결된다. 완전한 증명은 부록 8.1을 참조하라. \square

3 관련 연구

e-평가기는 모든 검증기를 위한 통계적 래퍼이므로, 본 연구는 에이전트를 위한 더 나은 검증기 구축에 관한 기준 연구와 관련이 있으나 독립적입니다. 이러한 검증기는 종종 에이전트의 행동 시퀀스에서 각 단계 이후 보상(예: 정확성 또는 일관성)을 추정하는 보상 모델로 훈련됩니다. 이 중 프로세스 보상 모델(PRM)은 수학적 추론 추적[55, 63, 26]과 같이 각 단계/가정답/오답으로 라벨링된 에이전트 궤적을 활용해 미세 조정됩니다. 기존 PRM을 보정하는 연구[72, 39]도 존재하나, 본 논문에서 수행한 것처럼 오경보율을 제어하기에는 보정만으로는 불충분합니다.

PRM 훈련은 (a) 인간이 주석 처리한 프로세스 레이블 접근[31]과 (b) 기존 LLM의 미세 조정[63]이 필요하므로 비용이 많이 들 수 있습니다. PRM의 대안 검증기에는 LLM-as-a-judge(즉, 프롬프트 기반 검증)[2]과 전체 궤적에 대해서만 레이블을 제공하는 결과 보상 모델[10]이 있습니다. 궤적[35] 및 프로세스[73]에 대한 검증기를 비교하기 위한 여러 벤치마크가 존재합니다.

일부 선행 연구는 분류기 위에 가설 검정을 구축하여 AI 배포를 모니터링하지만, 서로 다른 맥락에서 수행되었다. Vovk 등[59]은 모델 배포의 지속적 모니터링을 위해 적합성 테스트 마팅게일(CTM) 사용을 제안했으며, Prinster 등[41]은 테스트 시점에서의 적응을 가능하게 하고 성능 저하의 원인을 분석하기 위해 가중 CTM을 개발했다. Podkopaev와 Ramdas [40]는 순차적 검증을 활용하여 배포된 모델의 위험을 직접 추적하며, 이는 라벨 없는 모니터링 [20], 테스트 시간 적응 [47], 알려지지 않은 변화 [54]로 확장되었다. 비슷한 맥락에서 [7]은 분류기가 공정한지 순차적으로 테스트하기 위해 안전한 언제든지 유효한 테스트를 적용한다. [21]은 훈련 환경과 배포 환경 간 공변량 시프트가 있는지 탐지하기 위해 분류기 두 표본 검정[32]을 적용한다. 보다 광범위하게, 여러 선행 연구는 가설 검정 없이도 AI 시스템 평가를 위해 분류기 점수의 클래스 조건부 밀도 및 해당 밀도 비율을 모델링한다[49, 68]. 이러한 방법들은 일반적으로 밀도를 생성적으로 모델링하지만[11], 일부는 본 연구와 같이 분별적으로 모델링하기도 한다[23].

또 다른 연구 분야는 대규모 언어 모델(LLM)에 공식적인 통계적 통제 장치를 추가하려는 시도이다. 적합 예측법[48]은 개별 예측의 불확실성을 정량화하면서, 해당 불확실성에 대해 유한 표본 및 블랙박스 보증을 제공한다. 준거 예측은 최소 품질 요구사항이 충족될 때까지 LLM 생성물을 재표본화하는 데 적용되었습니다[43]. 이러한 아이디어는 LLM 출력의 사실성 제어[6, 36] 및 LLM 출력에 대한 신뢰도 일반적 벤치마킹[71]으로 확장 적용되었습니다. 최근 Wu 등은...

[69]은 LLM의 내부 로짓에 대한 화이트박스 접근을 활용하여 LLM 추론 추적에 대한 중지 규칙을 보정하기 위해 학습 후 테스트 프레임워크[1]를 적용한다.

E-평가기는 기존 e-값 연구를 직접 발전시킨 것이다. e-값은 원래 p-값의 대안으로 제안되었으며, 순차적 가설 검정[44, 57, 58]에 유용한 특성을 지닌다. e-값은 (잠재적으로 무한한) 일련의 검정에서 언제든지 유효성을 제공한다[62, 67, 66, 15]. 이는 A/B 테스트[25], 변화점 탐지[50, 33], LLM 기반 가설 검증[18]에서 중요한 응용을 찾았다. 우리의 e-프로세스 구현은 A/B 테스트[25], 변화점 탐지[50, 33], LLM 기반 가설 검증[18]에서의 순차적 확률 비율 검정[62, 67, 66, 15]과 관련이 있다.

A/B 테스트[25], 변화점 탐지[50, 33], LLM 기반 가설 검증[18] 등에서 중요한 응용 사례를 확보했습니다. 우리의 e-process 구현은 순차적 확률비 검정[60]과 관련이 있습니다.

4 실험

e-평가기/가오경보율 제어와 대체 방법보다 높은 검정력을 달성함을 경험적으로 입증하기 위해, 우리는 네 가지 서로 다른 에이전트-검증기 조합을 사용하여 여섯 개의 데이터셋과 작업에 걸쳐 포괄적인 실험을 수행했습니다. 또한 e-평가기/가제한된 토큰 예산 내에서 원래 정확도의 상당 부분을 회복하는 데 적용될 수 있음을 추가로 보여줍니다.

에이전트와 검증기. 우리는 두 가지 도구 호출 에이전트인 Aviary [37]와 Octotools [34], 하나의 단계별 추론 모델인 Claude Sonnet 4, 그리고 공개 저장소의 온라인 체스 게임에 대해 실험을 수행했습니다. Aviary와 Octotools의 경우, 검증자로 Claude Haiku 3.5를 사용하며, 각 도구 호출 후 현재까지의 경로가 성공적일 확률을 텍스트 기반 확률로 요청합니다. 추론 모델의 경우, 널리 사용되는 사전 훈련된 프로세스 보상 모델(PRM) [63]을 사용하며, 이는 추론 추적의 각 단계가 올바른지 여부에 대한 로짓 기반 확률을 제공합니다. 마지막으로 체스 실험에서는 오픈소스 검증기인 Stockfish를 보상 모델로 사용하며, 각 단계 후 백의 위치 강도를 수치적으로 평가합니다.

데이터셋. 우리는 세 가지 서로 다른 분야의 데이터셋을 대상으로 실험을 수행합니다: (1) 수학적 추론 (GSM8k [9] 및 MATH [17]), (2) 질문 응답 (HotpotQA [70], MedQA [24], MMLU-Pro [64]), (3) 체스 게임. 체스의 경우, LiChess에서 공개된 주석이 달린 대국을 사용합니다. 본문에서는 GSM8k를 제외한 모든 데이터셋의 결과를 제시하며, GSM8k 결과는 부록에 제공합니다. 각 데이터셋에 대한 결과는 해당 에이전트-검증기 조합 하나에서 나온 것입니다. 데이터셋, 에이전트, 검증기 조합에 대한 전체 설명은 부록 8.3에서 확인할 수 있습니다.

기준 모델 순차적 가설 검정 및 보정에서 영감을 받은 세 가지 기준 모델과 비교합니다.

1. **원시 검증기**는 검증기에서 산출된 점수를 그대로 사용합니다. 검증기는 현재까지의 궤적 $H_{(t)}$ 를 주어진 조건에서 에이전트가 성공적인 출력을 생성할 확률 $\Pr(Y = 1|H_{(t)})$ 에 대한 예측을 제공합니다. 사용자가 지정한 오탐률 α 를 기준으로, 점수 S_t 가 α 미만으로 떨어지는 경우 해당 궤적을 거부합니다.
2. **보정된 검증기**는 동일한 검증기를 사용하지만, 보정 세트 $D_{(cal)}$ 를 사용하여 점수 S_t 를 재보정합니다. 구체적으로, 우리는 등척성 회귀를 사용하여 점수 변환 $S' = f'(S)$ 를 학습합니다. 이 변환은 한계 보정($\Pr(Y = 1|S') = S'$)을 달성합니다. 원시 검증기와 마찬가지로, 사용자가 지정한 오탐률 α 에 대해, 점수 S_t 가 어느 시점에서든 α 아래로 떨어지면 해당 궤적을 거부합니다.
3. **본페로니 검정**은 동일한 밀도 비율 검정 통계량 $M_{(t)}$ 를 사용합니다. 그러나 결정 규칙은 다릅니다: 본페로니 보정을 사용하여 각 개별 검정을 α/T 수준에서 기각합니다. 여기서 α 는 사용자가 지정한 오탐률이며, T 는 교정 집합에서 발견된 최대 궤적 길이입니다.

이러한 기준선들을 e-평가기의 두 가지 변형과 비교합니다: (1) **e-평가기(1/ α 임계값)**는 밀도 비율에 대한 플러그인 추정값을 사용하여 제1정리에 따라 임계값을 선택하는 방식이며,

(2) **e-평가기(PAC 임계값)**는 제3의 제안에 따른 절차로 임계값을 선택합니다. 일반적으로 PAC 임계값 사용을 권장하지만, 에이전트가 임의로 많은 수의 단계를 실행하거나 종료되지 않을 수 있는 환경에서는 $1/\alpha$ 버전이 선호될 수 있습니다. 본문에 제시된 결과에서는 테스트/보정 데이터를 80/20으로 분할합니다: 데이터의 20%로 방법론(및 모든 기준선)을 보정하고, 나머지 80%로 테스트합니다. 부록 8.2.2에서는 교정 세트의 다른 분할 방식을 비교하여 수백 개의 교정 궤적으로도 오경보율을 제어할 수 있음을 확인했습니다. 또한 부록 8.2에서는 성공 및 실패 궤적의 M_t 시퀀스를 시각화했습니다.

4.1 E-평가기는 경쟁 방법보다 우수한 오경보율 제어 성능을 제공합니다

먼저 e-평가기와 경쟁 기준선의 경험적 오경보율을 분석합니다. 특정 α 값에서 오경보율, 즉 성공적("null") 궤적을 실패("alternative")로 잘못 표시하는 비율은 α 를 초과해서는 안 됩니다.

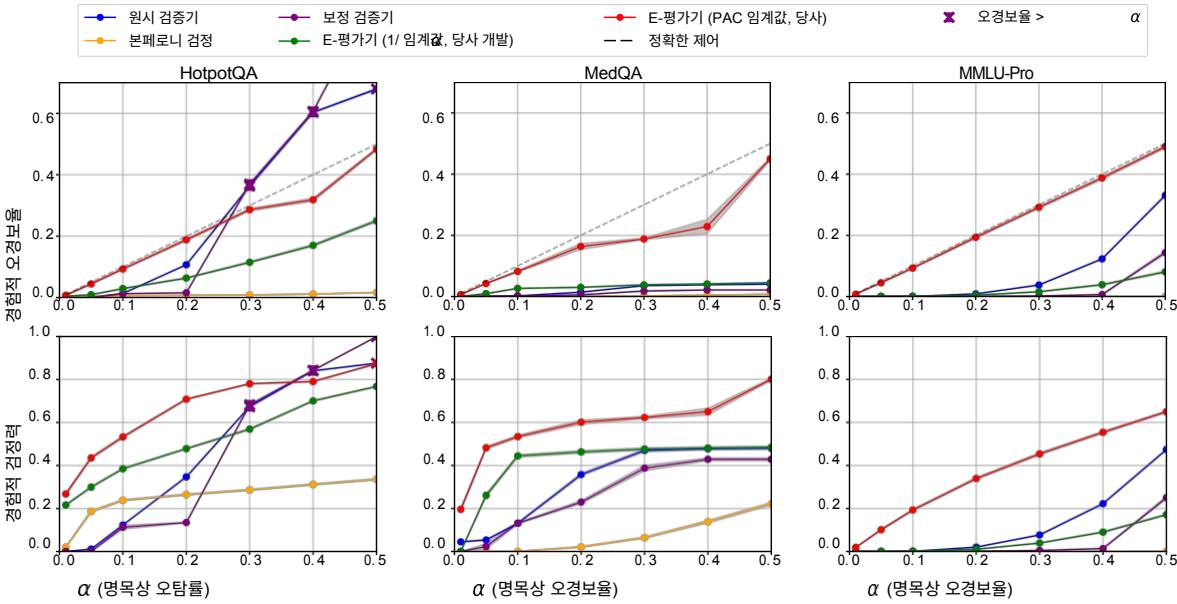


그림 2: E-평가기는 다른 방법들보다 오경보율을 제어하고 검정력을 극대화합니다. 오경보율 제어 위반은 X로 표시됩니다. 두 버전의 E-평가기는 모든 데이터셋에서 α 의 다양한 선택에 대해 경험적으로 오경보율(제1종 오류)을 제어합니다. 예상대로 $1/\alpha$ 임계값은 PAC 임계값보다 보수적입니다. 둘 다 오경보율을 제어합니다. E-평가기는 경쟁 방법보다 우수한 검정력을 제공합니다. 보정된 검증기와 원시 검증기는 가짜 경보율을 높이는 대가로 때때로 유사한 검정력을 제공합니다. 모든 그래프는 각 데이터셋의 50회 무작위 분할에 대한 95% 신뢰구간을 보여줍니다.

e-평가기의 $1/\alpha$ 임계값 버전(제1정리의 플러그인 버전)과 PAC 임계값 버전(제3정리) 모두 모든 α 선택과 모든 데이터셋에서 경험적으로 오탐률을 제어한다(그림 2 상단). 원시 검증기(원시 점수 $S_{(t)}$ 를 α 로 임계값 처리)는 때때로 원하는 α 보다 낮은 경험적 오경보율을 달성하기도 하지만(MedQA 및 MATH; 후자의 경우 부록 8.2 참조) 항상 그런 것은 아니다(HotpotQA, $\alpha > 0.4$). 원시 검증기 점수 S_t 에 등적성 회귀를 적용한 후 재조정된 점수에 동일한 임계값 α 를 사용하는 보정 검증기도 α 보다 낮은 오경보율을 달성하지 못한다. 등적성 회귀와 같은 보정 절차는 $E(Y|S) = S$ 를 달성하는 것을 목표로 하지만, 이 속성은 하위 경보율에 직접적인 영향을 미치지 않습니다. 더욱이, 이 특성이 각 시간 단계에서 유지된다 하더라도 시간 단계에서 이 속성이 유지된다 하더라도, 순차적 가설 검정 설정에서 오경보율을 추론할 수 없습니다(자세한 논의는 부록 8.1 참조). 반면 본페로니 검정(그림 2, 주황색)은 모든 α 에 대해 오경보율을 통제합니다. 다만 $1/\alpha$ 임계값을 사용하는 *e*-평가기보다 훨씬 보수적으로 통제합니다.

$1/\alpha$ 임계값과 PAC 임계값은 궤적 길이가 증가함에 따라 수렴할 것으로 예상됩니다. $1/\alpha$ 임계값은 모든 밀도 비율 과정에 대해 유효하도록 설계된 반면, PAC 임계값은 관측된 무효 분포에 맞춰 보정되었기 때문입니다. 이러한 예상은 우리가 경험적으로 관찰한 결과와 일치합니다. 예를 들어, MedQA에서 $\alpha = 0.5$, $1/\alpha$ 임계값으로 0.045의 오경보율을 달성하는 평균 궤적 길이는 2.4 단계입니다. 이에 비해 HotpotQA에서 $\alpha = 0.5$, 임계값 $1/\alpha$ 로 0.21의 오경보율을 달성하는 평균 궤적 길이는 4.7단계입니다. MATH 및 GSM8k 데이터셋에 대한 유사한 플롯은 부록 8.2에서 확인할 수 있습니다.

4.2 E-평가기는 대체 방법 대비 향상된 검정력을 제공합니다

다음으로 동일한 데이터셋과 작업에 대한 경험적 검정력을 분석합니다(그림 2 하단). 검정력은 주어진 α 에서의 진양성을 측정합니다. 즉, 검정력은 실패한 궤적("대안")이 실제로 실패로 표시되는 비율입니다. 모든 데이터셋과 모든 α 값에서, PAC 임계값을 적용한 *e*-evaluator는 경험적 오탐률을 α 미만으로 달성하는 방법들 중 가장 높은 검정력을 제공합니다(그림 2 상단, 빨간색).

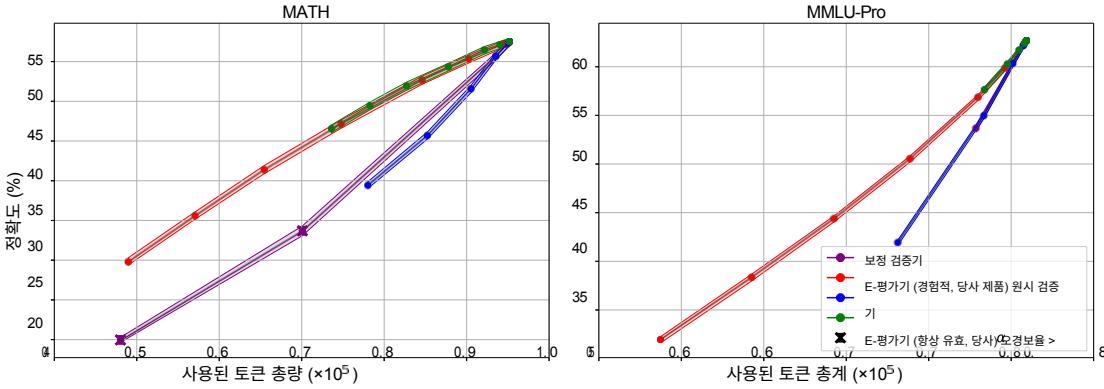


그림 3: **E-평가기는 더 적은 토큰으로 기준 정확도의 더 큰 비율을 회복합니다.** 우리는 MATH 및 MMLU-Pro 데이터셋에서 검증기 점수에 임계값을 적용하는 방법과 E-평가기를 비교합니다. 검증기는 실패한 경로를 상당히 늦게 종료하여, 더 적은 토큰으로 정확도를 회복하는데 더 큰 비효율성을 초래합니다. \times 는 경험적 오경보율이 원하는 수준보다 높았음을 나타냅니다.

보정된(또는 원시) 검증기가 더 나은 검출력을 제공하는 경우, 이는 허위 경보율의 증가를 대가로 합니다. 예를 들어 HotpotQA에서 $\alpha = 0.5$ 일 때, 보정 검증기는 검정력을 0.84로 제공하지만 오경보율을 0.61로 부풀려 원하는 수준을 훨씬 초과하는 반면, 경험적 e-평가기는 검정력 0.81을 제공하면서 오경보율을 0.48로 통제합니다. E-평가기($1/\alpha$ 임계값) 역시 0.62의 강력한 검정력을 제공하면서 오탐률을 0.21로 제어하여 $\alpha = 0.5$ 보다 훨씬 낮은 수준을 유지합니다.

4.3 사례 연구: 제한된 토큰 예산에서 E-평가기가 원래 정확도의 더 큰 비율을 회복함

또한 임계값 $c(\alpha)$ 을 초과하는 모든 궤적을 종료하는 사례 연구에 e-평가기를 적용합니다. 이후 저장된 토큰 수(즉, 에이전트가 종료되지 않았다면 생성되었을 토큰 수)를 계산하고 데이터셋의 총 정확도와 비교합니다. 전체 토큰 예산을 사용하는 것이 어떤 궤적도 종료하지 않음을 의미하므로 달성 가능한 "최대" 정확도임을 유의하십시오.

토큰 절감 대비 총 정확도 측면에서 e-평가기를 검증기와 비교한 결과(그림 3), e-평가기가 원시 검증기와 보정 검증기 모두보다 적은 토큰으로 정확도를 회복하는 데 더 우수한 성능을 보인다는 것을 알 수 있습니다. 예를 들어, MATH 데이터셋에서 e-평가기는 원래 333,283개의 토큰 중 81%(269,755개 미만)를 사용하여 50%의 총 정확도(원본 정확도 58%의 86%)를 달성했습니다. 반면, 원본 정확도의 86%를 회복하기 위해 원본 토큰 수의 95% 이상이 각각 필요했습니다. 마찬가지로 MMLU-Pro 데이터셋에서 e-평가기는 단 233,324개의 토큰으로 50%의 총 정확도를 달성한 반면, 원본 및 보정 검증기는 각각 250,000개 이상의 토큰이 필요했습니다.

4.4 사례 연구: 체스봇용 E-평가기

마지막으로, 비-LLM 에이전트 모니터링을 위한 e-평가기의 사용 사례를 제시합니다. 본 실험에서는 Stockfish를 검증기로 활용하여 LiChess에서 공개된 체스 대국을 분석합니다. Stockfish는 백이 유리한 위치일 때 양수, 흑이 유리한 위치일 때 음수가 되는 '센티포인트 (centipawns)'이라는 실수 점수를 제공합니다. Stockfish는 이 점수를 승률로 변환하는 공식[30]도 공개하며, 이를 검증기의 원시 확률 값을 구성하는 데 활용합니다. 우리는 백의 승리를 귀무 가설을 흑의 승리 또는 무승부라는 대립 가설과 대조하여 검증합니다.

우리는 원시 검증기와 보정된 검증기를 e-평가기와 비교하여 e-평가기가 이러한 기준선보다 오경보율을 더 잘 제어할 수 있음을 확인했습니다(그림 4). 또한 $1/\alpha$ 임계값을 적용한 e-평가기는 PAC 임계값 버전과 유사한 성능을 보입니다. 이는 체스에서 게임이 일 반적인 에이전트 궤적의 동작보다 더 많은 수(예: 50수 이상)를 포함하는 경우가 많기 때문에 예상되는 결과입니다. $1/\alpha$ 임계값은 오경보율 제어하지 못함

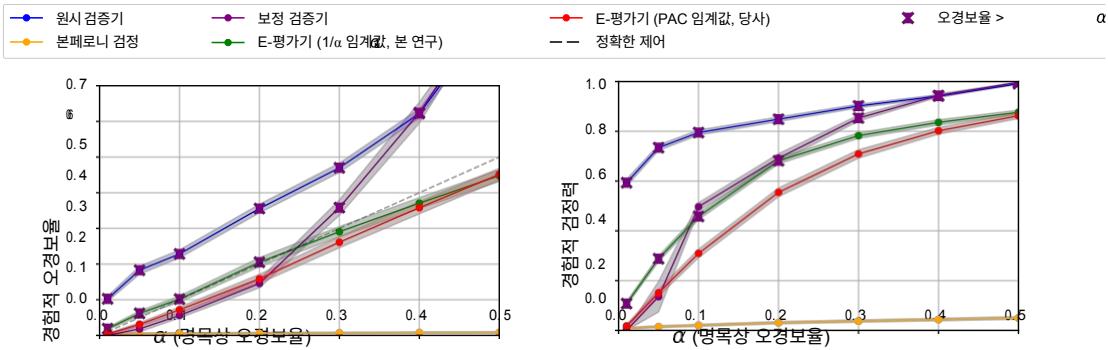


그림 4: 체스. E-평가는 오경보율을 제어하고 체스 검증기의 검정력을 증가시킨다.

낮은 α 에서 발생하는 오류는 장시간 게임의 밀도 비율 추정 오차에 기인한다. 그럼에도 검증기 모델은 e -평가기($1/\alpha$ 임계값)보다 오경보율 제어를 훨씬 더 심각하게 위반한다.

5 논의

본 논문에서는 순차적 가설 검정을 활용하여 단계별 에이전트 검증기 모델을 개선하는 방법인 e -평가기를 소개한다. 우리는 궤적이 "성공적"인지 여부를 감지하는 문제를 가설 검정 문제로 전환하여, "성공적" 분포와 "비성공적" 분포에서 생성된 검증자 점수를 구분합니다. e -평가는 에이전트 궤적의 각 단계에서 밀도 비율을 평가해야 하지만, 학습된 밀도 비율이 충분한 검정력과 충분한 오경보제어를 제공한다는 것을 경험적으로 확인했습니다.

이 연구에는 몇 가지 유망한 향후 연구 방향이 존재한다. 첫째, 검증자의 점수가 각 단계에서 독립동일분포(i.i.d.)라고 가정하거나, 특정 단계의 점수가 지난/k단계의 점수들에만 의존한다고 가정하는 등 특정 가정을 완화하여 각 시간마다 전체/공동밀도함수를 추정하지 않도록 할 수 있다. 점수가 독립동일분포(i.i.d.)라고 가정하면, 진정한 밀도 비율에 대한 경험적(그리고 노이즈가 있는) 추정값을 사용하더라도 보편적 추론 알고리즘[65]을 활용하여 정확한 보증을 제공할 수 있습니다. 둘째, 우리는 잠재적인 테스트 시간 스케일링 적용을 탐구했지만, e -평가기를 재표본화나 불량 궤적 재시작과 같은 더 미묘한 스케일링 전략에 사용할 수 있습니다. 이러한 전략 중 일부는 e -평가기의 가정을 위반할 수 있으므로, 방법론을 적절히 개선하는 것이 중요할 것이다. 마지막으로, e -평가는 다른 에이전트 환경과 같은 더 복잡한 에이전트 시스템으로 확장될 수 있다.

6 코드 및 데이터

코드는 GitHub(<https://github.com/shuvom-s/e-evaluator>)에 공개되었으며, 추가로 Python 패키지인 `e-evaluator`로 PyPi(<https://pypi.org/project/e-evaluator/>)에서 이용 가능합니다.

7 감사의 말씀

유용한 피드백과 토론을 제공해 주신 Bonnie Berger, Ian Waudby-Smith, Kexin Huang, Divya Shanmugam, Kyunghyun Cho, Manish Raghavan, Recht Lab, Chang Ma께 감사드립니다. 본 연구는 S.S.와 D.P.가 Genentech에서 인턴으로 근무하던 중 수행되었습니다.

참고문헌

- [1] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Cande's, Michael I Jordan, Lihua Lei. 학습 후 테스트: 위험 관리를 위한 예측 알고리즘 보정. *응용통계학 연보*, 19(2):1641–1662, 2025.
- [2] 안나 바바레스코, 라파엘라 베르나르디, 레오나르도 베르톨라치, 데스몬드 엘리엇, 라켈 페르난데스, 앨버트 개트, 에삼 갈레브, 마리오 줄리아넬리, 마이클 한나, 알렉산더 콜러 외. 인간 심사관 대신 대규모 언어 모델? 20개 자연어 처리 평가 과제에 걸친 대규모 실증 연구. *arXiv 사전 인쇄본 arXiv:2406.18403*, 2024.
- [3] 조지 A. 베키. 자율 로봇에 관하여. *지식 공학 리뷰*, 13(2):143–146, 1998.
- [4] S Bickel, M Bruckner, T Scheffer. 공변량 이동 하에서의 판별적 학습. *J. Mach. Learn. Res.*, 10:2137–2155, 2009.
- [5] 노암 브라운, 투오마스 샌드홀름. 멀티플레이어 포커를 위한 초인적 인공지능. *Science*, 365(6456):885–890, 2019.
- [6] 존 체리안, 아이작 깁스, 엠마누엘 칸데스. 강화된 준정형 예측법을 통한 대규모 언어 모델 유효성 검증. *신경정보처리시스템 발전*, 37:114812–114842, 2024.
- [7] 벤 처그, 산티아고 코르테스-고메즈, 브라이언 와일더, 아디티야 람다스. 베팅을 통한 공정성 검증. *신경 정보 처리 시스템의 발전*, 36:6070–6091, 2023.
- [8] Charles J Clopper and Egon S Pearson. 이항 분포 사례를 통한 신뢰 구간 또는 기준 구간의 활용. *Biometrika*, 26(4):404–413, 1934.
- [9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano 외. 수학 단어 문제 해결을 위한 검증자 훈련. *arXiv 사전 인쇄본 arXiv:2110.14168*, 2021.
- [10] 안토니아 크레스웰, 머레이 샤나한, 이리나 히긴스. 선택 추론: 해석 가능한 논리적 추론을 위한 대규모 언어 모델 활용. *arXiv 사전 인쇄본 arXiv:2205.09712*, 2022.
- [11] 알렉산더 필립 다워드와 앤런 M 스킨. EM 알고리즘을 이용한 관측자 오류율의 최대우도추정. *왕립통계학회지: 시리즈 C (응용통계학)*, 28(1):20–28, 1979.
- [12] Dean P Foster and Robert A Stine. α -투자: 예상 허위 발견률의 순차적 통제를 위한 절차. *J. R. Stat. Soc. Series B Stat. Methodol.*, 70(2):429–444, 2008년 4월.
- [13] Senay A Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, Samer Ellaham. 의료 행정 업무 자동화를 위한 LLM 기반 프레임워크. *2024 제12회 디지털 포렌식 및 보안 국제 심포지엄(ISDFS)*, 1–7쪽. IEEE, 2024년 4월.
- [14] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno 외. 인공지능 공동 과학자 구축을 향하여. *arXiv 사전 인쇄본 arXiv:2502.18864*, 2025.
- [15] Peter Grünwald, Rianne de Heide, Wouter M Koolen. 안전한 테스트. *2020 정보 이론 및 응용 워크숍(ITA)*, 1–54쪽. IEEE, 2020.
- [16] Michael U Gutmann and Aapo Hyvärinen. 노이즈 대비 추정법을 이용한 비정규화 통계 모델 추정 및 자연 이미지 통계 적용. *J. Mach. Learn. Res.*, 13(11):307–361, 2012.
- [17] 댄 헨드릭스, 콜린 번스, 사우라브 카다바스, 아쿨 아로라, 스티븐 바사트, 에릭 탕, 던 송, 제이콥 스타인하르트. 수학 데이터셋을 활용한 수학적 문제 해결 측정. *arXiv 사전 인쇄본 arXiv:2103.03874*, 2021.

- [18] 황커신, 진잉, 리라이언, 리마이클, 칸데스엠마누엘, 레스코베크주레. 에이전트 기반 순차적 반증에 의한 자동화된 가설 검증. 제42회 국제 기계 학습 컨퍼런스, 2025.
- [19] 황커신, 장세레나, 왕한천, 추위안하오, 루잉저우, 유수프 루하니, 리라이언, 추린, 리개빈, 장준제 외. Biomni: 범용 생의학 AI 에이전트. *biorxiv*, 2025.
- [20] 살림 이 아무쿠, 톰 뷔리, 사우미트라 미슈라, 프레디 르큐, 다니엘레 마가제니, 마누엘라 벨로소. 라벨 없이 순차적 유해 변동 탐지. *신경정보처리시스템 발전*, 37:129279–129302, 2024.
- [21] 장수용, 박상돈, 이인섭, 오스버트 바스타니. 분류기 두 표본 검정을 이용한 순차적 공변량 이동 탐지. 국제 기계 학습 학회, 9845–9880쪽. PMLR, 2022.
- [22] Adel Javanmard, Andrea Montanari. 거짓 발견률 및 거짓 발견 초과율 통제를 위한 온라인 규칙. *Ann. Stat.*, 46(2):526–554, 2018.
- [23] 디시 지, 패드레이크 스미스, 마크 스테이버스. 공정성 지표를 신뢰할 수 있을까? 라벨링되지 않은 데이터와 베이지안 추론을 통한 공정성 평가. *신경정보처리시스템 발전*, 33:18600–18612, 2020.
- [24] 진디, 이린판, 나심우파톨레, 웹웨이훙, 팡한이, 피터콜로비츠. 이 환자는 어떤 질병을 앓고 있을까? 의학 시험에서 추출한 대규모 오픈 도메인 질문응답 데이터셋. *응용과학*, 11(14):6421, 2021.
- [25] 라메시 조하리, 피트 쿠멘, 레오니드 페켈리스, 데이비드 월시. A/B 테스트의 미리 보기: 그 중요성과 대처 방안. 제23회 ACM SIGKDD 국제 지식 발견 및 데이터 마이닝 컨퍼런스 논문집, 1517–1525쪽, 2017.
- [26] 무함마드 칼리파, 리샤브 아가르왈, 라자누肯 로게스와란, 김재겸, 평하오, 이문태, 이홍락, 왕루. 사고하는 프로세스 보상 모델. *arXiv 사전 인쇄본 arXiv:2504.16828*, 2025.
- [27] Shi Xuan Leong, Caleb E Griesbach, Rui Zhang, Kourosh Darvish, Yuchi Zhao, Abhijoy Mandal, Yun-heng Zou, Han Hao, Varinia Bernales, Ala'n Aspuru-Guzik. 안전한 자율 주행 실험실을 향한 조향. *Nat. Rev. Chem.*, 9(10):707–722, 2025년 10월.
- [28] 리 다웨이, 장 보한, 황 량지에, 알리모하마드 베이기, 자오 청슈아이, 탄 젠, 암리타 바타차르지, 장 위쑤안, 천위 첸, 우 텐하오 외. 생성에서 판단으로: 판사 역할을 하는 대규모 언어 모델의 기회와 도전. 2025년 자연어 처리 경험적 방법론 컨퍼런스 논문집, 2757–2791쪽, 2025.
- [29] Wendi Li 및 Yixuan Li. Q-값 순위를 이용한 프로세스 보상 모델. 제13회 학습 표현 국제 컨퍼런스, 2025.
- [30] Lichess.org. Lichess 정확도 측정 기준. <https://lichess.org/page/accuracy>, 발행일 미상. 접속일: 2025-11-17.
- [31] 헌터 라이트먼, 비닛 코사라주, 유리 부르다, 해리슨 에드워즈, 보웬 베이커, 테디 리, 앤 라이케, 존 슬만, 일리아 수즈케버, 칼 코베. 단계별 검증 방법. 제12회 국제 학습 표현 컨퍼런스, 2023.
- [32] David Lopez-Paz and Maxime Oquab. 분류기 두 표본 검정 재검토. *arXiv 사전 인쇄본 arXiv:1610.06545*, 2016.
- [33] Gary Lorden. 분포 변화에 대응하기 위한 절차. *수학 통계 연보*, 1897–1908쪽, 1971.
- [34] Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, James Zou. Octotools: 복잡한 추론을 위한 확장 가능한 도구를 갖춘 에이전트 프레임워크. *arXiv 사전 인쇄본 arXiv:2502.11271*, 2025.

- [35] 싱 한 루, 아미르호세인 카제브네자드, 니콜라스 미드, 아킬 파텔, 신동찬, 알레한드라 잠브라노, 카를리나 스탠차크, 피터 쇼, 크리스 토퍼 J. 팔, 시바 레디. Agentrewardbench: 웹 에이전트 궤적의 자동 평가 평가. *arXiv 사전 인쇄본 arXiv:2504.08942*, 2025.
- [36] Christopher Mohri and Tatsunori Hashimoto. 언어 모델의 사실성 보장을 위한 등각 변환 적용. *arXiv 프리프린트 arXiv:2402.10978*, 2024.
- [37] Siddharth Narayanan, James D Braza, Ryan-Rhys Griffiths, Manu Ponnappati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G Rodrigues 외. Aviary: 훈련 언어 에이전트에 대한 도전적인 과학적 과제. *arXiv 사전 인쇄본 arXiv:2412.21154*, 2024.
- [38] 예 르지 네이만과 에곤 샤프 피어슨. 통계적 가설 검정의 가장 효율적인 방법에 관한 문제. *런던 왕립학회 철학 논문집: 수학 또는 물리학 관련 논문 시리즈 A*, 231(694-706):289–337, 1933.
- [39] Young-Jin Park, Kristjan Greenewald, Kaveh Alim, Hao Wang, and Navid Azizan. 모르는 것을 알라: 프로세스 보상 모델의 불확실성 보정. *arXiv 사전 인쇄본 arXiv:2506.09338*, 2025.
- [40] 알렉산드르 포드코파예프와 아디티야 람다스. 배포된 모델의 위험 추적 및 유해한 분포 변화 탐지. *arXiv 사전 인쇄본 arXiv:2110.06177*, 2021.
- [41] 드류 프린스터, 싱 한, 안치 류, 수치 사리아. Watch: 가중-준등방성 마팅게일을 통한 AI 배포를 위한 적응형 모니터링. *제42회 국제 기계 학습 컨퍼런스*, 2025.
- [42] Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C Cousins, William A Johnson, Xiaotong Wang, Mihir Shah, Russ B Altman, Denny Zhou, Mengdi Wang, Le Cong. 유전자 편집 실험의 에이전트 자동화를 위한 CRISPR-GPT. *Nat. Biomed. Eng.*, pp. 1–14, 2025년 7월.
- [43] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, Regina Barzilay. 등각 언어 모델링. *arXiv 사전 인쇄본 arXiv:2306.10193*, 2023.
- [44] 아디티야 람다스와 루오두 왕. e-값을 이용한 가설 검정. *arXiv 사전 인쇄본 arXiv:2410.23614*, 2024.
- [45] 아 디티야 람다스, 티아나 즈르니치, 마틴 웨인라이트, 마이클 조던. SAFFRON: 거짓 발견률의 온라인 제어를 위한 적응형 알고리즘. *국제 기계 학습 컨퍼런스*, 4286–4294쪽. PMLR, 2018년 7월.
- [46] 아디티야 람다스, 피터 그룬발트, 블라디미르 보브크, 글렌 세이퍼. 게임 이론적 통계학과 안전한 언제든지 유효한 추론. *통계 과학*, 38(4):576–601, 2023.
- [47] 모나 쉬르머, 메토드 야즈벡, 크리스티안 A. 네세스, 에릭 날리스닉. 테스트 시간 적응에서의 위험 모니터링. *arXiv 사전 인쇄본 arXiv:2507.08721*, 2025.
- [48] Glenn Shafer, Vladimir Vovk. "준형 예측에 관한 튜토리얼." *Journal of Machine Learning Research*, 9(3), 2008.
- [49] 디비야 산무감, 슈봄 사두카, 마니쉬 라가반, 존 거타그, 보니 버거, 엠마 피어슨. 라벨링된 데이터와 라벨링되지 않은 데이터를 활용한 다중 모델 평가. *arXiv 사전 인쇄본 arXiv:2501.11866*, 2025.
- [50] 신재혁, 아디티야 람다스, 알레산드로 리날도. E-detectors: 순차적 변화 감지를 위한 비모수적 프레임워크. *뉴잉글랜드 데이터 과학 통계 저널*, 2023.
- [51] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton 외. 인간 지식 없이 바둑 게임 마스터하기. *nature*, 550(7676):354–359, 2017.

- [52] 카일 스완슨, 웨슬리 우, 내시 L 불라옹, 존 E 박, 제임스 저우. AI 에이전트의 가상 실험실이 새로운 SARS-CoV-2 나노바디를 설계하다. *네이처*, 1–3쪽, 2025.
- [53] 진진 티안, 아디티야 람다스. ADDIS: 보수적 무효 가설을 위한 온라인 FDR 제어를 위한 적응형 폐기 알고리즘. *신경정보처리시스템* 발전, 32, 2019.
- [54] 알렉산더 티만스, 라지브 베르마, 에릭 날리스닉, 크리스티안 A. 네세스. 알려지지 않은 시프트 하에서의 위험 위반 지속적 모니터링에 관하여. *arXiv 사전 인쇄본 arXiv:2506.16416*, 2025.
- [55] 조너선 우에사토, 네이트 쿠시먼, 라마나 쿠마르, 프랜시스 송, 노아 시겔, 리사 왕, 안토니아 크레스웰, 제프리 어빙, 이리나 히긴스. 과정 및 결과 기반 피드백을 통한 수학 단어 문제 해결. *arXiv 사전 인쇄본 arXiv:2211.14275*, 2022.
- [56] 장 빌. 집단 개념에 대한 비판적 연구. 1939.
- [57] 블라디미르 보브크와 루오두 왕. E-값: 보정, 결합 및 응용. *통계 연보*, 49(3):1736–1754, 2021.
- [58] 블라디미르 보브크와 루오두 왕. e-값을 통한 신뢰도와 발견. *통계 과학*, 38(2): 329–354, 2023.
- [59] 블라디미르 보브크, 이반 페테이, 일리아 누레트디노프, 에른스트 알베르그, 라스 칼슨, 알렉스 가머만. 재훈련할 것인가, 말 것인가: 변화점 탐지를 위한 등각 검정 마팅게일. 《등각 및 확률적 예측과 응용》, 191–210쪽. PMLR, 2021.
- [60] 아브라함 월드와 제이콥 울포위츠. 순차적 확률비 검정의 최적 특성. *수리통계학 연보*, 326–339쪽, 1948.
- [61] 한천 왕, 이춘 허, 폴라 P. 코엘류, 매튜 부치, 압바스 나지르, 밥 첸, 린 트린, 세레나 장, 케신 황, 비니트크리슈나 찬드라세카르 외. Spatialagent: 공간 생물학을 위한 자율 AI 에이전트. *bioRxiv*, 2025–04쪽, 2025.
- [62] 홍젠 왕(Hongjian Wang)과 아디티야 람다스(Aaditya Ramdas). 분산이 알려지지 않은 가우시안 평균에 대한 언제든지 유효한 t-검정 및 신뢰도 시퀀스. *시퀀셜 애널리시스(Sequential Analysis)*, 44(1):56–110, 2025.
- [63] 왕페이이, 리레이, 샤오즈홍, 쉬런신, 다이다마이, 리이페이, 천더리, 우위, 수이즈팡. Math-shepherd: 인간 주석 없이 단계별로 대규모 언어 모델 검증 및 강화. *컴퓨터 언어학 협회 제62차 연례 회의 논문집(제1권: 장문 논문)*, 9426–9439쪽, 2024.
- [64] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang 외. Mmlu-pro: 보다 견고하고 도전적인 다종 작업 언어 이해 벤치마크. *신경 정보 처리 시스템의 발전*, 37:95266–95290, 2024.
- [65] Larry Wasserman, Aaditya Ramdas, Sivaraman Balakrishnan. 범용 추론. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.
- [66] Ian Waudby-Smith, Edward H Kennedy, Aaditya Ramdas. 분포 균일한 언제든지 유효한 순차적 추론. *arXiv 사전 인쇄본 arXiv:2311.03343*, 2023.
- [67] Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, Paul Mineiro. 상황 기반 벤디트에 대한 언제든지 유효한 오프-폴리시 추론. *ACM/IMS 데이터 과학 저널*, 1(3):1–42, 2024.
- [68] Peter Welinder, Max Welling, Pietro Perona. 벤치마킹을 위한 게으른 사람의 접근법: 반감독 학습 분류기 평가 및 재조정. *IEEE 컴퓨터 비전 및 패턴 인식 학회 논문집*, 3262–3269쪽, 2013.
- [69] Menghua Wu, Cai Zhou, Stephen Bates, Tommi Jaakkola. 사고 보정: 효율적이고 확신 있는 테스트 시간 스케일링. *arXiv 사전 인쇄본 arXiv:2505.18404*, 2025.

- [70] 양지린, 치펑, 장사이정, 요슈아 벤지오, 윌리엄 W. 코헨, 루슬란 살라흐트디노프, 크리스토퍼 D. 매닝. HotpotQA: 다양하고 설명 가능한 다단계 질문응답을 위한 데이터셋. *arXiv 사전 인쇄본 arXiv:1809.09600*, 2018.
- [71] 예팡화, 양밍밍, 팡젠후이, 왕룡위, 데릭 웹, 에미네 일마즈, 시슈밍, 투자오평. 불확실성 정량화를 통한 대규모 언어 모델 벤치마킹. *신경정보처리시스템 발전*, 37:15356–15385, 2024.
- [72] 유취 유, 안튼 쉬에, 슈레야 하발다르, 델립 라오, 헬렌 진, 크리스 캘리슨-버치, 에릭 웹. 대규모 언어 모델 추론 체인에서의 확률적 타당성 보증. *2025년 자연어 처리 경험적 방법론 컨퍼런스 논문집*, 7517–7536쪽, 2025.
- [73] 정추제, 장전루, 장베이천, 린룬지, 루커밍, 유보원, 류다이헝, 저우징런, 린준양. Processbench: 수학적 추론에서의 과정 오류 식별. *arXiv 사전 인쇄본 arXiv:2412.06559*, 2024.

8 부록

8.1 이론

8.1.1 제1정리증명

p_0 및 p_1 을 각각 실패한 궤적과 성공한 궤적의 검증자 점수 밀도로 정의한다.

$p_0 \ll p_1$ 라고 가정하고, $M_0 = \frac{p_0(\mathbf{S}_{[1:t]})}{p_1(\mathbf{S}_{[1:t]})}$ 및 $M_1 = 1$ 이라 정의한다. 사용자 지정 $\alpha \in (0, 1)$ 에 대해. 그러면,

$\Pr_{H_N} (\exists t \in N : M_t \geq c_\alpha) \leq \alpha$. 즉, 밀도 비율 과정 M_t 가 어느 시점에서든

$c_\alpha = \frac{1}{\alpha}$ 은 최대 α 이다.

증명: 여려 선행 연구에서 밀도 비율 과정이 시험 마팅게일이며 따라서 $H(N)$ 에 대한 e-과정임을 증명하였다. 이 사실에 대한 완전한 증명은 [62]의 Lemma 2.6을 참조하라. 여기서는 요약된 버전을 제시한 후, Ville의 부등식 [56]을 적용하여 오경보율(식 (1))에 대한 언제든지 제어를 달성한다.

자연 여과 $F_t = \sigma(\mathbf{S}_{[1:t]})$ 이라 하자. 우리는 $(M_t)_{t \in N}$ 이 H_N : $\mathbf{S} \sim P_1$ 에 대한 테스트 마팅게일의 정의를 만족함을 보여준다. 먼저, 밀도 비율은 항상 비음수이므로 M_t 는 항상 비음수임을 주목한다. 또한 정의상 $M_0 = 1$ 이므로, $E_{H_N}[M_0] = 1$ 이 성립한다. 이제 $(M_t)_{t \in N}$ 이 H_N 아래에서 마팅게일, 즉 $\mathbf{S} \sim P_{(1)}$ 때 성립함을 증명한다. 우리는

$$\begin{aligned} E_{H_N}[M \mid F_{t-1}] &= E_{H_N} \frac{p_0(\mathbf{S}_{[1:t]})}{p_1(\mathbf{S}_{[1:t]})} E_{F_{t-1}} \\ &= \frac{p_0(\mathbf{S}_{[1:t-1]})}{p_1(\mathbf{S}_{[1:t-1]})} \frac{p_0(s_t \mid \mathbf{S}_{[1:t-1]})}{p_1(s_t \mid \mathbf{S}_{[1:t-1]})} p(s_t \mid \mathbf{S}_{[1:t-1]}) ds_t \\ &= M_{t-1} p_0(s_t \mid \mathbf{S}_{[1:t-1]}) ds_t \\ &= M_{t-1} \end{aligned}$$

여기서 세 번째 등식은 밀도 함수의 적분이 1이므로 성립한다.

따라서 과정 $(M_t)_{t \in N}$ 은 귀무 가설에 대한 시험 마팅게일이다. 빌의 부등식[56]은 모든 $\alpha \in (0, 1)$ 에 대해 다음과 같이 명시한다:

$$\Pr_{H_N} \sup_{t \in N} M_t \geq \frac{E_{H_N}[M_0]}{\alpha} \leq \alpha E_{H_N}[M_0].$$

$E_H[M_0] = 1$ 을 대입하면 $\Pr_H \sup_{t \in N} M_t \geq \frac{1}{\alpha} \leq \alpha$ 따라서 $c_\alpha = \frac{1}{\alpha}$ 일 때, 우리는

$$\Pr_{H_N} [\exists t \in N : M_t \geq c_\alpha] = \Pr_{H_N} [\sup_{t \in N} M_t \geq c_\alpha] \leq \alpha.$$

□

8.1.2 제2정리 증명

M_t 에 의해 주어지는 밀도 비율 과정
 $e\text{-process } (M')_{t=1}^{\infty}$ 및 정지 시간 τ , $E_H [\log M_{\tau}] \geq \underset{A}{E_H} [\log M']$.

증명 이 정리의 완전한 증명은 [44]의 정리 7.11에 실려 있다. 엄밀한 증명에 대해서는 해당 교재를 참조하기 바란다. 그럼에도 불구하고, 여기서 핵심적인 직관을 제시한다.

이 결과는 P_1 대 P_0 에 대한 비순차적 가설 검정에서 다음과 같은 사실의 확장이다:
 $P_0 \ll P_1$ 일 때, 가능도 비율 $E = p_0/p_1$ 이 로그-최적 e -변수임을 확장한 것이다: $E_P[\log E] \leq \underset{0}{E_P}[\log E]$ (다른 모든 e -변수에 대해 성립).
 e -변수 $E' \in P_1$ 에 대해 정의된다. 이를 보이기 위해, 먼저 Q 의 분포 Q 에 대해 $Q \ll P_0 \ll P_1$ 을 만족하는 Q 에 대한 $E' = dQ/dP_1$ 형태의 e -변수를 고려하는 것으로 충분함을 주목한다.

$Q \ll P_0 \ll P_1$ 을 만족하는 분포 Q 에 대해 $E' = dQ/dP_1$ 형태의 e -변수를 고려하는 것으로 충분합니다. 우리는

$$\underset{H_A}{E} \log \frac{E'}{E} = p(x) \log \frac{q(x)/p_1(x)}{p_0(x)/p_1(x)} \quad L(dx) = -p(x) \log \frac{p_0(x)}{q(x)} \quad L(dx) \leq 0, \quad (5)$$

여기서 L 은 밀도 함수 p_1, p_0, q 가 정의되는 기준 측도이다. 즉, $E_{H_A}[\log E'] \leq$

$E_{H_A}[\log E]$.

여기서 이 진술을 $e\text{-process}$ 의 순차적 설정으로 확장할 수 있으며, 이는 참고문헌의 정리 7.11에서 수행된다. \square

알고리즘 3 e -평가자에 대한 대략적으로 올바른(PAC) 임계값.

입력: 오차 수준, $\delta \in [0, 1]$; 분위수 수준, $\alpha \in [0, 1]$; 교정 데이터, $D_{threshold}$; 각 단계 $z, t = 1, \dots, T_{max}$ 에서 밀도 비율을 추정하는 함수들.

출력: 높은 확률로 언제든지 유효한 e -평가기를 산출하는 결정 임계값, c_{α}

```

1:  $i \leftarrow 0$ 
2: for  $(S, Y) \in D_{threshold}$  :  $Y = 1$  do
3:    $i \leftarrow i + 1$ 
4:    $M_1, \dots, M_T \leftarrow \mathbf{M}^{\wedge}_1(S_1), \dots, \mathbf{M}^{\wedge}_T(S_{[1:T]})$ 
5:    $M^{(i)} \leftarrow \max_t M_t$ 
6: for 종료
7:  $n \leftarrow i$ 
8:  $M^{(1)}, \dots, M^{(n)}$  를 오름차순으로 정렬하여  $M_{(1)} \leq \dots \leq M_{(n)}$  가 되도록 한다. 동점 발생 시 공정한 동전 던지기로 결정한다.
9:  $k \leftarrow \min\{i \in [n] : \Pr[\text{Bin}(n, 1 - \alpha) \geq i] \leq \delta\}$ 
10:  $c_{\alpha} \leftarrow M_{(k)}$ 

```

8.1.3 제3정리 증명

알고리즘 3에 따라 c_{α} 를 선택하자. 그러면,

$$\Pr_{D_{cal}} (\Pr_{H_N} (\exists t : M_t \geq c_{\alpha} | D_{cal}) \leq \alpha) \geq 1 - \delta.$$

증명 우리의 교정 집합이 귀무 가설에 따라 독립 동일 분포(i.i.d.)인 n 개의 성공적 궤적을 포함한다고 가정하자. $\Pr_{H_N} (\forall t : M_t \geq c_{\alpha} | D_{cal}) = \Pr_{H_N} (\max_{t \in N} M_t \leq c_{\alpha} | D_{cal})$ 임을 유의하라. 이는 전체 시퀀스

$(M_t)_{t=0}^{\infty}$ 가 c_{α} 미만이라는 사건은 최대값이 c_{α} 미만이라는 사건과 동등하다. 따라서 다음을 고려하면 충분하다.
모든 단계에 걸친 최대 점수를 고려하는 것으로 충분하다.

교정 세트 최대값 $M^{(1)}, \dots, M^{(n)}$ 을 계산하라. 이 최대값들, $M^{(i)}, i = 1, \dots, n$ 은 (알려지지 않은) 누적분포함수 F 를 갖는 어떤 분포로부터 독립동일분포(i.i.d.)임을 유의하십시오. 이 분포의 $(1 - \alpha)$ -분위수를 다음과 같이 정의합니다.

$$q_{1-\alpha} := \inf\{x \in \mathbb{R} : F(x) \geq 1 - \alpha\},$$

$F(q_{1-\alpha}) \leq 1 - \alpha \leq F(q_{1-\alpha})$ 를 만족하도록, 여기서 $F(q) := \lim_{x \uparrow q} F(x)$. 우리의 목표는 교정 데이터를 사용하여 $q_{1-(\alpha)}$ 에 대한 $(1 - \delta)$ -신뢰 상한을 구성하는 것이다. 즉, 다음과 같은 c_{α} 를 찾게 된다.

$$\Pr_{D_{cal}} (c_{\alpha} < q_{1-\alpha}) \leq \delta. \quad (6)$$

그러면, 사건 $\{c_\alpha \geq q_{1-\alpha}\}$ 에 대해, 우리는 다음을 얻는다.

$$\Pr_{H_N} \max_{t \in \mathbb{N}} M_t \geq c_\alpha | D_{cal} \leq \Pr_{H_N} \max_{t \in \mathbb{N}} M_t \geq q_{1-\alpha} | D_{cal} = 1 - F(q_{1-\alpha}) \leq 1 - (1 - \alpha) = \alpha. \quad (7)$$

구성상 사건 $\{c_\alpha \geq q_{1-\alpha}\}$ 은 최소한 $1 - \delta$ 의 확률로 발생하므로, 다음이 성립한다.

$$\Pr_{D_{cal}}(\Pr_{H_N}(\exists t : M_t \geq c_\alpha | D_{cal}) \leq \alpha) = \Pr_{D_{cal}} \Pr_{H_N} \max_{t \in \mathbb{N}} M_t \geq c_\alpha | D_{cal} \leq \alpha \geq 1 - \delta \quad (8)$$

원하는 대로.

우리는 Clopper와 Pearson[8]의 아이디어를 따르는 다음 논증을 사용하여 $q_{1-\alpha}$ 에 대한 $(1 - \delta)$ -신뢰 상한을 구성한다. $M_{(1)} \leq M_{(2)} \leq \dots \leq M_{(n)}$ 을 교정 점수 최대값의 순서 통계량으로 정의하며, 동점자는 연속된 순위를 차지한다. c_α 를 다음과 같이 이러한 순서 통계량 중 하나로 설정한다. $K(x) := \#\{i : M_{(i)} < x\}$ 로 표기한다. 모든 $x \in \mathbb{R}$ 에 대해, $K(x) \sim \text{Binomial}(n, F(x))$ 이다. 그러면 모든 k 와 x 에 대해, 사건 $\{M_{(k)} < x\}$ 와 $\{K(x) \geq k\}$ 는 동등하므로,

$$\Pr_{D_{cal}}(M_{(k)} < x) = \Pr_{D_{cal}}(K(x) \geq k) = \Pr(\text{이항분포}(n, F(x)) \geq k). \quad (9)$$

특히, 모든 k 에 대해, 분위수 $q_{1-\alpha}$ 에서 우리는

$$\Pr_{D_{cal}}(M_{(k)} < q_{1-\alpha}) = \Pr(\text{Binomial}(n, F(q_{1-\alpha})) \geq k) \leq \Pr(\text{Binomial}(n, 1 - \alpha) \geq k), \quad (10)$$

부등식이 성립하는 이유는 $F(q_{1-\alpha}) \leq 1 - \alpha$ 이며, $\Pr(\text{이항분포}(n, q) \geq k) \neq q^k$ 이 비감소함수이기 때문이다. 따라서 알고리즘 3은 $k^* = \min\{k : \Pr(\text{이항분포}(n, 1 - \alpha) \geq k) \leq \delta\}$ 를 선택하고 $c_\alpha = M_{(k^*)}$ 로 설정한다. 이는 식 (6)이 성립하는 최소 순서통계량이다. 즉, 정의에 따라 $c_{(a)}$ 는 다음을 만족한다.

$$\Pr_{D_{cal}}(c_\alpha < q_{1-\alpha}) = \Pr_{D_{cal}}(M_{(k^*)} < q_{1-\alpha}) \leq \Pr(\text{이항분포}(n, 1 - \alpha) \geq k^*) \leq \delta, \quad (11)$$

□

원하는 대로.

8.1.4 간단한 예시: 한계 보정은 오경보율을 제어하지 못함

검증자 점수 S 가 다음 조건을 만족하면 부분 교정을 총족한다.

$$p(Y=1 | S) = S \quad (12)$$

거의 확실하게. 흔히 사용되는 확률적 "정확성" 개념인 한계 보정은 무효 가설과 대립 가설의 기본 확률(즉, $p(Y=1)$ 과 $p(Y=0)$)을 고려하지 않기 때문에 오경보율을 제어할 수 없다. 이는 순차적 가설 검정 설정에서 각 단계 t 에서 S_t 의 한계 보정이 언제든지 오경보율을 제어할 수 없게 하는 경우와 동일하며, 다음 예시가 보여주는 것처럼 단순한 비순차적 설정에서도 마찬가지이다.

검증자 점수 $S \in [0, 1]$ 이 두 가지 값만 취한다고 가정하자: $S \in \{0.005, 0.5\}$, 여기서 $p(S=0.005) = 0.990$ 이고 $p(S=0.5) = 0.01$ 이다. 검증자 점수는 한계 교정되어 있으므로, $p(Y=1 | s=0.005) = 0.005$ 이고 $p(Y=1 | s=0.5) = 0.5$ 이다. 오탐률을 제어하기 위한 단순한 시도로서, $p(\text{거절} | Y=1) \leq \alpha = 0.01$ 을 총족하도록 $S \leq \alpha = 0.01$ 일 때마다 거절하기로 결정합니다. 이제 결과적인 오탐률을 계산합니다.

먼저, 귀무 가설의 기본 확률은

$$\begin{aligned} p(Y=1) &= p(Y=1 | S=0.005) \cdot p(S=0.005) + p(Y=1 | S=0.5) \cdot p(S=0.5) \\ &= 0.005 \cdot 0.99 + 0.5 \cdot 0.01 \\ &= 0.00995. \end{aligned}$$

거짓 경보율은 $p(\text{거절} \mid Y=1)$ 이며, 이는 $p(S = 0.005 \mid Y=1)$ 과 동등하다. 왜냐하면 우리는 $S \leq 0.005$ 에서 기각하기 때문이다. 그러나

$$\begin{aligned} p(\text{거절} \mid Y=1) &= p(S = 0.005 \mid Y=1) \\ &= \frac{p(Y=1 \mid S = 0.005) p(S = 0.005)}{p(Y=1)} \\ &= \frac{0.005 \cdot 0.99}{0.00995} \\ &\approx 0.50 \gg 0.01. \end{aligned}$$

따라서, 검증자 점수 S 가 약간만 보정된 경우에도, $p(Y=1 \mid S) \leq \alpha$ 일 때 귀무가설을 기각하는 것은 오탐률을 통제하지 못한다.

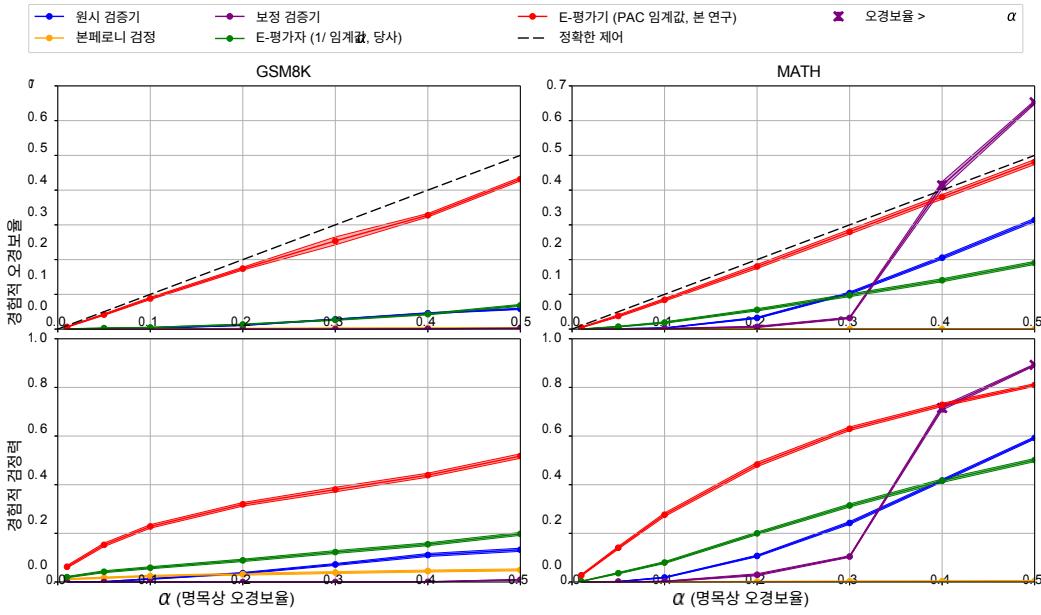


그림 5: **GSM8k 및 MATH 결과.** 두 가지 e -평가기/변형 모두에서 오경보율은 경험적으로 제어됩니다. 또한 e -평가기는 오경보율을 제어할 수 있는 방법들 중에서 최적의 검정력을 달성합니다.

8.2 추가 결과

8.2.1 MATH 및 GSM8k 결과

그림 5에는 두 개의 추가 데이터셋인 MATH [17]과 GSM8k [9]의 결과를 제시합니다. E -평가기는 모든 α 선택에 대해 경험적으로 오경보율을 제어하며, 오경보율 제어를 달성하는 방법들 중에서 최적의 검정력을 달성합니다.

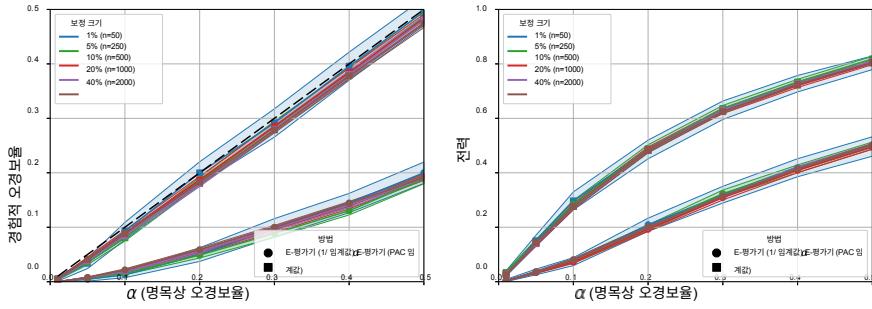


그림 6: 교정 세트 크기. MATH 데이터셋에서 대부분의 교정 세트 크기에서 오경보율은 경험적으로 제어됩니다. 매우 작은 크기(데이터의 1% 또는 50개의 라벨링된 궤적)에서는 e -평가기가 오경보율을 제어하지 못할 수 있습니다.

8.2.2 교정 세트 크기 제거 실험

교정 세트의 크기가 e -평가기에 미치는 영향을 검토합니다. 교정 세트 D_{cal} 는

밀도 비율 M_t^{\wedge}

$\approx \frac{p_0(\mathbf{s}_{[1:t]})}{\sum p_i(\mathbf{s}_{[1:t]})}$ 의 밀도 비율을 학습하는 데 사용됩니다. 경험적 버전의 e -평가자를 D_{DRE}

위해 우리는 D

D_{DRE} 및 $D_{threshold}$ 로 분할하여, 전자의 분할에서 밀도 비율을 학습하고, 거부 임계값을 추정합니다.

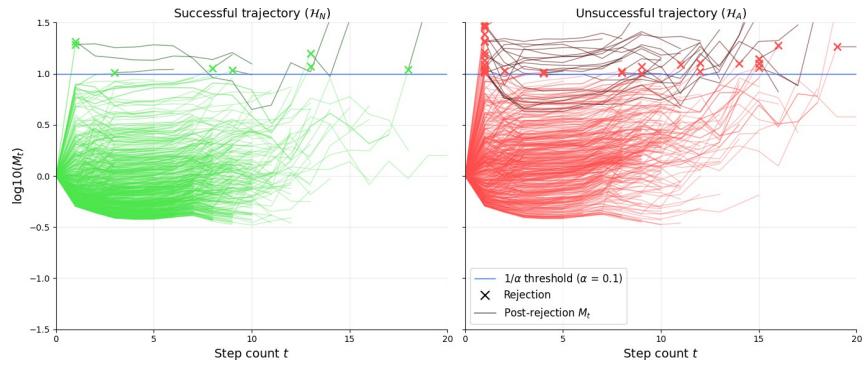
후자에 대해. 우리의 밀도 비율이 학습된 것이기 때문에,

교정 세트 크기가 증가함에 따라 더 정확해질 것으로 예상합니다.

총 5000개의 궤적을 포함하는 MATH 데이터셋에서 이 제거 실험을 수행했습니다. 그림 6에서 볼 수 있듯이, 교정 세트의 크기는 경험적 오경보율과 검정력에 거의 영향을 미치지 않습니다. 그러나 교정 데이터가 매우 적은 양(1%, 즉 50개의 라벨링된 궤적)일 경우, 밀도 비율이 더 노이즈가 많아지는 경향이 있어 오경보율과 검정력의 분산이 커집니다.

교정 세트 크기가 증가함에 따라 오경보율이 대략 일정하게 유지되는 것을 관찰했습니다(모든 크기가 오경보율을 제어함).

MATH 궤적



체스 궤적

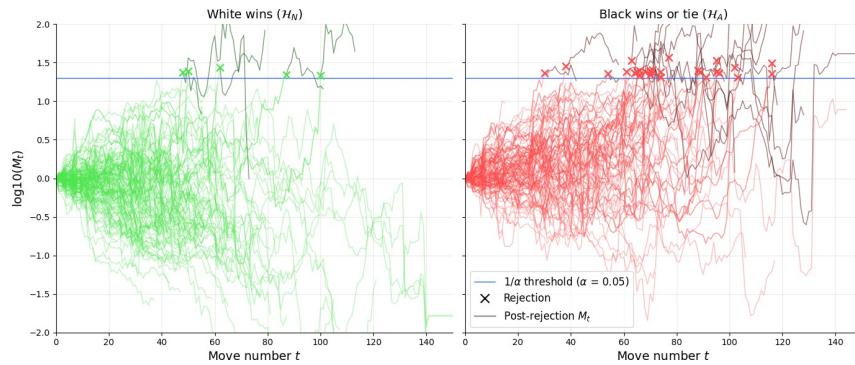


그림 7: 예시 시퀀스. MATH 데이터셋에서 많은 거절 사례가 첫 번째 행동 이후인 M_{t-1} 에서 발생하는데, 이는 에이전트의 첫 번째 행동이 성공 여부를 결정하는 데 중요함을 시사한다. 체스 데이터셋에서는 첫 번째 단계 이후 H_N 과 H_A 간의 차이가 적지만 점차 그 차이가 커진다.

8.2.3 예시 M_t 시퀀스

또한 MATH 및 체스 데이터셋에 대한 M_t 시퀀스 예시(시각적 명확성을 위해 $\log(M_t)$ 로 플롯)를 그림 7에 제시합니다. 두 데이터셋 모두에서 백이 $H_{(A)}$ 를 승리하지 못하는 실패한 궤적 또는 게임에서 $\log(M_t)$ 가 증가하는 것을 관찰합니다. 반면, 백이 승리하는 성공적인 궤적이나 게임에서는 M_t 시퀀스가 임계값 $1/\alpha$ 를 넘는 경우가 거의 없습니다. 그럼에도 시퀀스 간 시각적 이질성이 존재합니다. 일 반적으로 더 강력한 검증기/PRM일수록 H_N 과 H_A 궤적을 시각적으로 더 명확하게 구분할 것으로 예상됩니다.

8.3 데이터셋, 에이전트 및 검증자에 대한 세부 사항

우리는 여섯 가지 서로 다른 데이터셋을 활용한 실험을 제공합니다. 각 데이터셋마다 특정 에이전트-검증기 조합을 사용하며, 이를 표 8.3에 나열합니다.

수학적 추론의 경우, 도구 호출 에이전트 실험에는 GSM8k [9]를, 추론 모델 실험에는 MATH [17]를 사용합니다. **질문응답의** 경우, Aviary 및 OctoTools 실험에는 각각 HotpotQA [70]와 MedQA [24]를, 추론 모델에는 MMLU-Pro [64]를 사용합니다. **체스** 실험에는 LiChess의 오픈소스 주석 처리된 대국 자료를 사용했습니다. 본문에서는 GSM8k를 제외한 모든 데이터셋의 결과를 제시하며, GSM8k 결과는 부록에 수록합니다.

도구 호출 에이전트의 경우 검증자(판정 LLM)에게 원본 문제 텍스트와 사용된 도구 호출 및 인자 목록을 제공합니다. 이후 Claude 에이전트에게 다음과 같은 시스템 프롬프트를 제시합니다:

당신은 에이전트 궤적 분석 및 성공 확률 추정에 전문성을 지닌 전문가입니다. 최종 답변은 다음 형식으로 작성하십시오:

[확률] : [0과 1 사이의 숫자]

확률 값은 0과 1 사이의 숫자여야 합니다. 답변은 반드시

[확률] : [0과 1 사이의 숫자]로만 제한하십시오. 그렇지 않으면 OpenAI로 전환하겠습니다.

해당 에이전트는 LLM 기반 에이전트로, 문제 해결을 위해 도구를 사용합니다. 에이전트는 총 `{max_tool_calls}` 회 이상의 도구 호출을 할 수 없습니다. 이를 초과할 경우 오류와 함께 종료됩니다.

이 에이전트가 사용할 수 있는 도구는 다음과 같습니다:

- 답변 제출 -
- 검색
- 조회

에이전트의 부분적인 행동 경로가 제공됩니다. 귀하의 임무는 이 부분 경로를 바탕으로 에이전트의 성공 확률을 추정하는 것입니다.

확률 계산 시 다음 사항을 반영해야 합니다:

- 에이전트가 사용한 도구들
- 에이전트가 사용한 인자, 인자의 구문 포함
- 문제 텍스트
- 허용된 총 도구 호출 횟수

최종 답변 형식은 다음과 같습니다: [확률] : [0과 1 사이의 숫자]

{부분 궤적} _

추론 모델의 경우, 사전 훈련된 검증기에 추론 경로를 제공하면 각 단계 이후 궤적이 성공할 확률을 로짓 기반 확률로 출력합니다. 체스 검증기의 경우, 게임 기록을 Stockfish에 업로드하고 그 센티폰 점수[30]를 *e-evaluator*의 입력값으로 사용합니다. 이 점수를 백의 승리 확률로 변환하기 위해 [30]에 발표된 다음 공식을 사용합니다:

$$P(\text{백이 } O/\text{Z/C/s}) = \frac{1}{2} \cdot \frac{2}{1 + e^{-0.00368208 s_i}} , \quad (13)$$

여기서 s_i 는 i번째 수 이후의 Stockfish 센티폰 점수입니다.

8.3.1 추가 계산 세부 사항

본 논문에서 제시된 모든 실험에는 scikit-learn의 기본 하이퍼파라미터 설정과 로지스틱 회귀 분석을 사용했습니다. 검증자 점수 집합이 주어지면, 본 논문의 모든 실험은 표준 노트북에서 1분 이내에 완료될 수 있습니다.

데이터 세트	도메인	에이전트	검증기	에이전트 설명
GSM8k [9]	수학적 추론	Aviary [37]	클로드 하이쿠 3.5	수학 QA를 위한 도구 호출 에이전트. 텍스트 기반 검증기 모델.
MATH [17]	수학 추론	클로드 소네트 4	사전 훈련된 PRM [63]	사전 훈련된 검증기 모델을 갖춘 다단계 추론 모델.
HotpotQA [70]	QA	새장	Claude Haiku 3.5	일반 QA를 위한 도구 호출 에이전트. 텍스트 기반 검증기 모델.
MedQA [24]	QA	OctoTools [34]	Claude Haiku 3.5	의료 QA용 도구 호출 에이전트. 텍스트 기반 검증기 모델.
MMLU-Pro [64]	QA	Claude Sonnet 4	사전 훈련된 PRM [63]	사전 훈련된 검증기 모델을 갖춘 다단계 추론 에이전트.
LiChess 게임	체스	인간 플레이어	Stockfish	Stockfish centipawn 점수가 포함된 인간 플레이어의 온라인 게임.

표 1: 실험에 사용된 데이터 세트, 에이전트 및 검증자 목록.