미국 소비자 신용 업무에서 에이전트의 반복 완수와 실패 양상

Flitto Agent Bench (Finance – US Consumer Credit)

플리토 에이전트 데이터 리서처 팀

2026년 11월 21일

초록

미국 카드 발급사는 거래·계정 기록을 정책과 대조해 소비자 신용 업무를 처리하고 그 결과를 보고한다. 본 연구는 *Flitto Agent Bench (Finance – US Consumer Credit)*로 같은 문항을 반복 실행했을 때 완수 판정이 얼마나 일관되는지 평가했다. 벤치마크는 합성 기록과 고정 정책을 사용하는 여덟 분야의 99문항으로 구성된다. 본 평가에서는 다섯 모델에 각 문항을 두 차례 맡겼다. 매 시도는 별도 세션에서 실행했으며, 채점 코드는 필수 제출값과 파일 형식, 실행 한도 내 정상 종료 여부를 검사했다. 990회 중 407회(41.1%)가 통과했다. 동일 모델·문항의 두 시도를 짝지은 495쌍에서 두 번 모두 통과한 경우는 165쌍, 한 번만 통과한 경우는 77쌍(15.6%), 두 번 모두 실패한 경우는 253쌍이었다. 비교 모델 두 개를 포함한 진단에서는 실패한 755회 중 392회(51.9%)의 제출값이 자료 해석·대조에 관한 사전에 정의한 오답 패턴과 일치했다. 두 회차의 전체 완수율은 비슷했지만, 쌍별 집계에서 같은 모델·문항의 판정이 엇갈렸다. 실행 한도에 도달한 시도 중 세 건은 필수 값 검사를 모두 통과한 파일을 남겼다.

1 서론

카드 명세서에 표시된 잔액과 고객이 납부해야 할 금액은 서로 다른 값이다. 원장의 거래를 정확히 합산하더라도 결제 정책을 잘못 적용하거나 계산한 값을 다른 항목에 기록하면 업무를 완수하지 못할 수 있다. 따라서 소비자 신용 업무를 평가할 때는 완수 기준을 먼저 정하고, 두 회차의 판정이 일치하는지 확인해야 한다. 실행 기록을 대조하면 자료 확인부터 답안 제출까지 어느 단계에서 문제가 드러났는지 살필 수 있다.

미국 카드 발급사의 기록 대조·분석 업무에 대해 두 가지를 살폈다. 첫째, 같은 모델에 동일 문항을 두 차례 맡겼을 때 완수 판정이 얼마나 일치하는가? 둘째, 제출값과 실행 종료 기록에서 어떤 실패 유형을 확인할 수 있는가?

여덟 분야의 발급사 업무를 읽기 전용 자료, 적용 정책, 요구 산출물이 연결된 문항으로 구성했다. 2026년 11월 18일 자 고정 보고서의 공식 판정으로 두 회차의 완수를 비교하고, 공개 실패 진단을 바탕으로 제출값과 종료 상태에서 확인할 수 있는 실패 유형을 살폈다 [1].

1.1 선행 연구와 본 연구의 범위

τ-bench는 같은 과제를 반복 수행해 모든 시도를 통과한 비율을 나타내는 pass^{k}를 제시했다 [2]. Bank-World는 은행 업무 과제를 반복 실행해 판단 기준을 충족하고도 요구된 후속 조치를 완료하지 못한 사례를 분석했다 [3]. APEX–Accounting은 전문가가 작성한 기준으로 월말 결산 산출물을 평가하고, 반복 실행의 완수율과 오류 양상을 보고했다 [4]. τ-Knowledge는 정책에 필요한 문서를 제공하는 조건을 바꿔 근거에 접근하는 과정과 근거를 업무에 적용하는 과정을 비교했다 [5].

금융 자료를 평가에 사용하는 방식도 다양하다. FinanceBench는 기업 공시에 근거한 질의응답을, DABstep은 결제 자료로부터 얻은 수치와 짧은 답을, FINCH는 기업 스프레드시트 중심의 업무

산출물을 평가한다 [6, 7, 8]. FinAgent-Bench는 카드 이용률과 상환을 포함한 소비자 대상 금융 안내를 다룬다 [9]. 본 연구의 평가 범위는 발급사 내부의 원장 · 계정 · 정책 대조 업무다. 완수 판정은 제출 파일의 필수 값과 형식, 실행 한도 내 정상 종료 여부에 근거한다.

GDPval은 여러 직종의 산출물을 전문가가 평가하는 설계를 제시한다 [10]. BenchCAD는 과제별 입력, 요구 산출물, 검사 기준의 관계를 구체화한 사례다 [11]. 이 비교는 평가 대상과 판정 근거의 차이를 보여 준다. 선행 평가와의 성능 우열을 추정하려면 같은 업무에서 별도 비교가 필요하다.

2 평가 범위와 과제 설계

이 트랙은 미국 카드 발급사의 소비자 신용 분석 업무를 다룬다. 담당자는 기록을 검토하고 정책을 적용해 요청된 결과를 제출한다. 카드 명세서에 남은 청구액을 확인하는 업무부터 월말 포트폴리오 지표를 산출하는 업무까지 포함한다. 에이전트는 읽기 전용 데이터룸을 조회하고 별도 작업 공간에서 필요한 값을 계산한다. 완수는 필수 제출물과 실행 종료 상태로 판정한다. 실제 고객과의 대화, 심사 승인, 자금 이동, 운영 시스템의 상태 변경은 평가 범위 밖이다 [1].

2.1 문항 구성과 업무 제약

각 문항에는 업무 요청, 기록과 정책을 담은 데이터룸, 제출 규약, 채점 기준이 있다. 거래 · 계정 · 신용 기록은 평가용으로 생성했다. 일부 생성기는 실제 금융 자료의 구조와 요약 통계를 참고해 보정했다. 미국의 공개 규정과 실무자 인터뷰 · 교육 자료 등을 참고해 업무 목적과 기준을 정했다. 문항에 따라 시점별 정책과 시스템별 금액 단위를 구분하고, 정정 · 반환 · 재시도 기록을 대조해야 한다 [1].

표 1은 분야별 문항 수와 기본 설계 수를 정리하고, 대표 업무의 자료 · 판단 · 제출 항목을 보여 준다. 부록 A에서는 분야별 사례 하나를 자세히 설명한다.

표 1: 공식 문항의 분야별 구성과 대표 업무. 문항 · 설계 수는 공식 평가에 해당한다. 각 업무 예시는 범위를 보여 주는 설명용 사례다.

업무 분야	문항	설계	요청자, 근거와 제출 항목
개인 금융	13	8	카드 회원이 특정 명세서의 남은 청구액을 묻는다. 명세서와 납부 이력에서 기준일까지 그 명세서에 반영된 납부를 찾아 합산한 뒤 남은 청구액과 함께 제출한다.
결제 · 청구	13	10	결제 운영 담당자가 월간 정산을 대조한다. 접수 · 재시도 · 정산 · 반환 기록을 원 결제 건별로 연결한다. 정산 건수와 반환액을 뺀 순정산액을 제출한다.
사기 · 분쟁	12	9	분쟁 담당자가 특정 기간에 접수해 처리가 끝난 사건을 검토한다. 접수 · 처리 시점을 기한 규칙과 대조하고 지급 · 취소 기록을 연결해 기한 초과 건수와 확정 보상액을 제출한다.
신용파일	11	7	신용정보 분석가가 신규 카드의 개설 시점 대표 신용점수를 요청한다. 계정 개설일과 신용조회 기록에 기간 · 기관 · 유효성 기준을 적용해 해당 계정의 대표 점수 평균을 제출한다.
심사	13	11	심사 책임자가 월별 승인 결과를 검토한다. 신청서의 월 소득 · 상환액과 최종 심사 결과를 연결해 승인 건수와 승인 건의 평균 부채상환비율을 제출한다.
이탈 · 유지	13	9	고객 유지 담당자가 분기 이탈률을 요청한다. 계정 원장과 해지 이력에서 분기 시작 시점에 열려 있던 계정을 정하고, 그중 분기 내 해지 건수와 이탈률을 제출한다.
추심	12	8	추심 담당자가 월말 장기 연체 추심 대상 계정 수를 요청한다. 계정 · 결제 주기 기록에 제공된 추심 기준을 적용해 대손상각되지 않은 표준카드 중 연속 연체 조건을 충족하는 계정 수를 제출한다.
포트폴리오	12	12	포트폴리오 분석가가 월말 신용한도 이용률을 요청한다. 월말 결제 주기와 계정 원장을 대조해 보고 대상과 잔액 단위를 맞춘 뒤, 미수채권 합계를 같은 계정의 한도 합계로 나눈다.
합계	99	74	데이터룸 63개

2.2 합성 예시의 자료, 계산과 판정

공개 과제 소개는 신용한도 이용률을 포트폴리오 업무의 지표로 든다 [14]. 아래에서는 세 계정의 월말 기록을 예로 들어 이용률 계산과 채점 기준을 설명한다. 이 예시는 공식 99문항과 성능 집계에 포함되지 않는다.

표 2: 설명용 월말 계정 기록. 신용한도와 미수채권은 달러 기준이다.

계정	원자료 잔액	미수채권	신용한도
A	$350	$350	$1,000
B	70,000센트	$700	$1,000
C	−$20	$0	$1,000

예시의 업무 규칙은 잔액 단위를 달러로 맞추고, 음수 잔액은 미수채권 합계에서 0으로 처리하는 것이다. 세 계정의 신용한도는 모두 분모에 포함한다. 미수채권 합계 $1,050을 신용한도 합계 $3,000으로 나누면 이용률은 0.35다. 답안 파일의 예는 `{"utilization": 0.35}`다.

이 예시에서는 신용한도가 모두 같으므로 계정별 이용률 0.35, 0.70, 0.00을 단순 평균해도 0.35가 나온다. 제출값 검사로는 두 계산 방법을 구별할 수 없다. 어떤 방식으로 계산했는지는 실행

기록에서 확인해야 한다.

3 평가 절차

3.1 문항 선정과 평가 집단

평가 범위는 2절의 미국 소비자 신용 · 카드 운영 업무다. 각 문항은 업무 요청, 합성 업무 기록, 정책 · 절차 문서, 필수 제출 항목으로 구성했다. 분야 구성과 본 평가 모델 5개의 개발 단계 응답을 고려해 후보 311개 중 99개를 선정했다. 이 문항은 기본 설계 74개와 데이터룸 63개에 걸쳐 있다 [1].

본 평가의 다섯 모델은 99문항을 각각 두 차례 수행했다. 동일 모델 · 문항의 두 시도를 짝지어 990회를 495쌍으로 묶었다. 문항 선별에 사용하지 않은 모델 2개도 같은 99문항을 두 번 수행했다. 실패 진단에는 이 비교 평가를 포함해 7개 모델의 1,386회 가운데 실패한 755회를 사용했다. 시도는 2026년 11월 5일부터 13일까지 수행했다. 공식 결과와 실패 진단은 2026년 11월 18일 자 공개 기술 리포트의 고정판을 따른다 [1]. 2026년 11월 20일 자 리더보드는 별도 집계이며 모델 14개를 포함한다.

3.2 실행 조건과 시도 채택

문항 · 프롬프트 · 데이터룸의 해시와 채점 코드 판본을 확인한 뒤 시도를 채택했다. 시도마다 작업 공간과 세션을 분리했고, 데이터룸에는 읽기 전용으로 접근할 수 있었다. 2회차에 1회차 답안이나 채점 피드백을 넘기지 않았다. 공통 실행 상한은 60턴과 3,600초였다. 도구 접근과 추론 설정은 모델 설정에 따라 달랐다 [1, 16].

모델 · 문항 · 회차별로 고정 조건에 맞는 첫 유효 시도를 채택했다. 사전에 정한 인프라 오류가 발생하면 재실행을 한 차례 허용했다. 채택한 시도의 오답 제출, 미제출, 실행 한도 소진은 실패로 판정했다. 같은 모델과 문항에서 두 회차 모두 유효한 시도를 확보한 경우 판정을 짝지었다 [1].

3.3 참조 답안과 코드 판정

연구진은 문항마다 업무 요구를 필수 결과값과 적용 규칙으로 구체화했다. 금융 전문가가 업무 논리와 참조 답안을 검토했고, 공개 규정과 데이터룸 문서에 근거가 있는지 대조했다. 참조 답안은 적용 범위, 기간, 단위와 자료 간 우선순위를 반영했다. 문항별 채점 기준표에는 필수 키와 자료형, 허용 표현, 수치 항목별 허용 오차를 명시했다. 두 가지 참조 답안을 허용하는 경우에는 그중 하나의 필수 값 조건을 모두 충족해야 했다 [1].

통과하려면 지정 경로에 읽을 수 있는 JSON 객체를 제출하고, 모든 필수 항목의 채점 기준을 충족하며, 실행 한도 내에 정상 종료해야 했다. 제출값이 모두 맞더라도 한도 소진으로 중단한 시도는 실패로 판정했다. 문항별 부분 점수는 부여하지 않았다. 검토자는 참조 답안과 채점 기준을 확인하고 실패 유형을 사후 분류했다. 공식 통과 · 실패 판정은 고정된 채점 코드가 내렸다. 저장한 답안과 종료 상태를 대조해 1,386회 시도의 집계가 공개 기술 리포트와 일치하는지 검산했다 [1].

3.4 두 회차 결과와 실패 진단

동일 모델 · 문항의 판정 쌍을 두 번 통과, 1회차만 통과, 2회차만 통과, 두 번 실패로 나눴다. 평균 완수율은 두 회차 전체 유효 시도 중 통과한 시도의 비율이다. 반복 완수율(pass^2)과 한 번 이상

통과 비율(pass@2)은 각각 두 회차를 모두 통과한 쌍과 한 번 이상 통과한 쌍을 전체 쌍으로 나눈 값이다. 판정 일치율에는 두 회차 모두 실패한 쌍도 포함된다. 조건부 판정 변화율은 한 번 이상 통과한 쌍 중 한 회차에서만 통과한 쌍의 비율이다(4절) [1, 2].

고정판은 모델별 1회차 완수율의 구간을 문항 단위 부트스트랩으로 계산했다. 분야별 모델의 1회차 완수율에는 Wilson 구간을 썼다. 각 회차의 모델 비교는 같은 문항을 짝지은 정확 McNemar 검정과 Holm 보정을 사용했다 [1].

실패 진단은 제출값과 종료 상태를 사전에 정의한 오답 패턴과 대조했다. 한 시도에 여러 유형을 기록할 수 있었다. 오류가 발생한 단계를 확인하려면 업무 요청과 정책, 모델이 남긴 메시지, 도구 호출·응답, 제출값과 종료 상태를 시간순으로 대조한다. 공급자가 제공한 추론 기록이 남아 있으면 뒤따른 행동과 비교할 수 있다. 공개 감사에서는 평가 관련 표현과 평가 자료에 접근했는지를 별도로 확인했다(5절) [1, 12, 13].

4 공식 평가 결과

4.1 두 회차의 완수율과 판정 일치

본 평가의 다섯 모델이 각 문항을 두 차례 수행한 결과를 짝지었다(3절). 1회차에는 495회 중 200회(40.4%), 2회차에는 495회 중 207회(41.8%)가 통과했다. 두 회차를 합친 평균 완수율은 407/990(41.1%)이었다 [1].

표 3: 본 평가의 동일 모델 · 문항 $N = 495$쌍에서 두 회차의 통과 여부. 공개 고정판의 표 3 · 4에서 계산했다 [1].

1회차	2회차 통과	2회차 실패	합계
통과	165	35	200
실패	42	253	295
합계	207	288	495

2회차의 통과 건수는 1회차보다 일곱 건 많았다. 동일 모델 · 문항의 판정은 495쌍 중 77쌍(15.6%)에서 달랐다(표 3). 두 회차 모두 통과한 쌍은 165/495(33.3%), 두 회차 중 한 번 이상 통과한 쌍은 242/495(48.9%)였다. 한 번이라도 통과한 242쌍 가운데 77쌍(31.8%)은 한 회차에서만 통과했다. 판정 일치율 418/495(84.4%)에는 두 번 모두 실패한 253쌍이 포함된다. 이 수치는 3절의 별도 두 시도에서 관측한 결과다.

4.2 요청 범위에 따라 달라지는 제출값

판정 교차표는 두 회차의 결과가 얼마나 자주 달랐는지 보여 준다. 개별 판정이 갈린 지점을 확인하려면 업무 요청, 입력 자료, 제출값과 종료 기록을 대조해야 한다 [1]. 공개 결제 예시를 보면, 금요일에 시도한 $40 결제가 거절되고 월요일 재시도에서 정산됐다. 금요일 결제 시도액은 $40이지만 금요일 정산액은 $0이다. 같은 금액을 읽어도 보고 날짜와 결제 상태에 따라 제출값이 달라진다 [15].

5 실행에서 관측한 실패와 평가 검증

5.1 제출값에서 확인한 오답 유형

2026년 11월 18일 자 고정판의 본 평가와 비교 평가를 합친 1,386회 중 실패한 755회를 분석했다. 이 중 392회(51.9%)의 제출값에서 자료 해석 · 대조와 관련된 오답 패턴을 확인했다. 대상 · 잔액 정의에 관한 패턴은 158회, 정책 · 기준 시점에 관한 패턴은 44회에 나타났다. 한 시도에 여러 유형이 해당할 수 있다. 이 집계는 제출값이나 종료 상태에 나타난 유형의 빈도다. 개별 실패를 진단하려면 에이전트가 확인한 자료와 도구 응답, 제출값을 대조해야 한다 [1].

5.2 같은 오답에 이르는 여러 경로

공개 고정판의 실패 755회 중 36회는 결과 검토 또는 중복 집계와 관련된 유형에 해당했다 [1]. 공개 진단 글의 예시에서 금요일 마감 잔액이 $40만큼 크게 나온 이유는 월요일 거래를 포함했거나 금요일 거래 한 건을 두 번 셌기 때문일 수 있다. 조회 조건과 반환된 거래, 합계에 사용한 거래를 확인하면 두 가능성을 구별할 수 있다. 이 예시는 같은 제출값만으로 오류가 생긴 경로를 가를 수 없는 이유를 보여 준다 [15].

5.3 실행 종료와 참조 답안이 판정에 미친 영향

7개 설정의 실패 755회 가운데 47회는 실행 한도에 도달했다. 이 중 42회에는 답안 파일이 없었고, 2회는 파일이 있었으나 필수 값 검사에 실패했다. 나머지 3회의 답안 파일은 필수 값 검사를 모두 통과했다. 세 시도 모두 한도 내 정상 종료 조건을 충족하지 못해 공식 판정은 실패였다 [1, 15].

참조 답안은 98문항에서 하나였다. 한 문항에는 같은 요청에 대해 유효한 참조 답안 두 개를 적용했다. 그 문항에서는 두 답안 중 하나의 필수 값 조건을 모두 충족하면 제출값 검사를 통과했다 [1, 15].

5.4 공개 감사에서 확인한 평가 관련 표현

공개 고정판의 본 평가 990회에서 모델명과 판정을 가린 채 평가 관련 표현을 분류했다. 1회차 11/495회, 2회차 4/495회에서 해당 표현을 관측했다. 나머지 975회의 저장 기록에서는 해당 표현을 확인하지 못했다. 관측한 표현의 유무와 모델이 평가 상황을 실제로 인지했는지는 구분해야 한다. 저장된 도구 기록에서 참조 답안, 채점 코드, 저장소 또는 다른 실행의 자료에 접근한 사례는 확인되지 않았다 [1]. 평가 관련 표현, 평가 단서 탐색, 평가 자료 접근, 점수 영향은 각각 별도의 근거로 확인한다 [12, 13].

6 한계와 후속 검증

6.1 결과를 해석할 수 있는 범위

문항 선별과 표본 구조. 후보 311개 중 99개를 고를 때 본 평가 5개 모델의 개발 단계 응답을 사용했다. 개발 단계 완수율은 채택 문항에서 156/495(31.5%), 제외 문항에서 658/1,060(62.1%)이었다. 문항 선별에 사용하지 않은 두 모델도 같은 문항을 수행했다. 99문항 중 일부는 같은

기본 설계나 데이터룸을 쓴다. 이 때문에 문항 간 결과가 연관될 수 있다. 분야별 수치는 이 선별 집합에서 관측한 결과다. 공유 구조를 고려했을 때 통계 구간과 모델 간 차이가 얼마나 달라지는지는 별도로 확인해야 한다 [1].

실행과 측정의 범위. 모든 모델에 같은 문항과 명목상 실행 한도를 적용했다. 도구 접근과 추론 설정은 달랐으므로, 관측한 차이에는 모델과 실행 조건의 영향이 함께 반영됐을 수 있다. 두 회차는 이전 답안과 채점 피드백을 제공하지 않은 별도 세션에서 수행했다. 두 회차의 결과에는 공통 문항과 실행 조건에 따른 의존성이 남을 수 있다. 모델 · 문항별 관측값이 두 개뿐이라 장기 성공 확률을 안정적으로 추정하기 어렵다. 각 시도는 필수 값과 형식에 맞는 답안 파일을 제출하고 실행 한도 안에 정상 종료했을 때 통과했다. 평가 범위는 기록과 정책을 대조해 분석 결과를 제출하는 업무까지였다.

오답 집계와 실행 기록. 제출한 값이 사전에 정한 오답 패턴에 해당하면 그 유형으로 집계했다. 어떤 자료를 확인했고 어느 단계에서 오류가 드러났는지 알려면 실행 기록을 대조해야 한다 [1, 15]. 같은 문제가 전체 실행에서 얼마나 자주 발생했는지 알려면 분야와 모델 설정별로 기록을 나눠 추가로 검토해야 한다.

추론 기록과 감사 범위. 개별 실행에서는 모델이 남긴 메시지, 도구 호출과 응답, 제출값을 서로 대조해 해석한다. 모델 공급자가 반환한 추론 기록(CoT)이 남아 있으면 그 내용과 뒤따른 행동도 비교할 수 있다. 공개 감사 결과는 평가 관련 표현과 평가 자료 접근 여부처럼 저장 기록에서 확인한 항목을 보고한다. 특정 문항의 답을 외웠는지를 판단하려면 학습 자료나 문항 노출 이력에 관한 별도 근거가 필요하다 [1, 12].

공개 자료로 확인할 수 있는 범위. 공개 표로는 통과율 같은 집계 수치를 다시 계산할 수 있다. 개별 시도의 판정을 다시 확인하려면 답안 파일, 종료 기록, 참조 답안과 채점 코드가 필요하다. 같은 조건에서 평가를 다시 실행하려면 당시의 문항, 데이터룸, 실행기 판본과 모델 공급자 설정도 필요하다 [1].

6.2 후속 비교의 설계

기존 반복 기록의 층별 검수. 두 번 모두 실패한 253쌍과 판정이 엇갈린 77쌍을 분야 · 모델 설정 · 확인할 수 있는 기록 범위별로 나눠 검수할 수 있다. 두 검수자는 회차 순서와 상대 회차의 판정을 모르는 상태에서 각 기록을 독립적으로 분류한다. 두 번 실패한 쌍에서는 처음 관측됐고 종료까지 바로잡히지 않은 오류를 대조한다. 오류가 드러난 단계와 잘못 적용한 업무 기준이 각각 반복됐는지 기록한다. 판정이 엇갈린 쌍에서는 자료 선택, 계산, 제출, 종료의 차이를 기록한다. 판독할 수 없는 쌍의 수와 합의 전 검수자 간 일치도를 보고한다. 층별로 표본을 뽑으면 층의 크기로 가중하고, 판독 불가에 따른 반복률의 하한과 상한도 제시한다.

추가 관측과 개선 효과. 이전 답안과 피드백을 넘기지 않고 같은 설정에서 회차를 늘리면, 고정된 업무에서 반복 성공과 오류 재발을 더 정밀하게 추정할 수 있다. 새로 발견한 실패 유형의 수는 별도 지표로 보고한다. 시도를 늘리면 드문 실패를 더 많이 관찰할 수 있기 때문이다. 성능 개선을 평가하려면 공통 조건에서 비교할 대조군, 시도당 완수율, 첫 시도를 포함한 총시간과 비용을 사전에 정한다. 앞선 답안 또는 실무 검토 피드백을 제공하는 재시도는 별도 조건으로 측정한다.

새 문항과 입력 변경 대조. 선별이나 교정에 사용하지 않은 별도 기본 설계와 데이터룸에서 결과를 확인할 수 있다. 같은 업무 요청과 정책에서 정답을 바꾸는 원장 기록과 정답에 영향을 주지 않는 기록을 각각 하나씩 바꾼다. 각 변형의 참조 답을 독립적으로 계산하고, 같은 설정의 별도 세션에서 변경된 기준값을 맞힌 비율과 변하지 않은 기준값을 유지한 비율을 보고한다. 이 비교는 제출값이 판단에 필요한 기록의 변화에는 맞춰 바뀌고, 결과와 무관한 기록이 바뀌었을 때는 그대로 유지되는지 확인한다. 특정 문항의 사전 노출에는 별도 근거가 필요하다 [5].

7 결론

미국 소비자 신용 업무 99문항을 별도 세션에서 두 차례 평가했다. 본 평가의 전체 완수율은 41.1%였다. 두 회차의 평균 완수율은 비슷했지만 동일 모델 · 문항 495쌍 중 77쌍(15.6%)에서 판정이 바뀌었다. 비교 평가를 포함한 진단에서는 실패 755회 중 392회에서 자료 해석 · 대조와 관련된 오답 패턴을 확인했다. 실행 한도에 도달한 시도 중 세 건은 필수 값 검사를 모두 통과한 파일을 남겼다. 평균 완수율은 전체 성과를, 쌍별 집계는 회차 간 판정 변화를 보여 준다. 제출값과 종료 상태를 함께 확인하면 실패의 성격을 구분할 수 있다 [1, 15].

후속 연구에서는 일관된 분류 기준으로 판정이 엇갈린 쌍과 두 회차 모두 실패한 쌍에서 오류가 드러난 단계를 확인할 수 있다. 문항 선별과 수정에 쓰지 않은 과제를 여러 차례 실행하면 세 번째 이후 시도에서 처음 관측되는 실패 유형을 집계할 수 있다. 같은 조건의 새 세션에서 다시 실행하는 평가와 이전 답안이나 피드백을 제공해 회복을 측정하는 평가는 별도로 설계해야 한다.

참고문헌

[1] Flitto Agent Bench (Finance – US Consumer Credit). 2026년 11월 18일 고정 기술 리포트. 국문. 영문.

[2] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045, 2024.

[3] D. Rupawalla. Bank-World: A Benchmark for Evaluating Agents on Banking Workflows. arXiv:2610.39733v1, 2026.

[4] Julien Benchek et al. APEX-Accounting. arXiv:2607.27189v2, 2026.

[5] Q. Shi et al. τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. arXiv:2603.04370v1, 2026.

[6] P. Islam et al. FinanceBench: A New Benchmark for Financial Question Answering. arXiv:2311.11944, 2023.

[7] A. Egg et al. DABstep: Data Agent Benchmark for Multi-step Reasoning. 2025.

[8] Haoyu Dong et al. Finch: Benchmarking Finance & Accounting across Spreadsheet-Centric Enterprise Workflows. Findings of the Association for Computational Linguistics: ACL 2026.

[9] t54 Labs. FinAgent-Bench. 제작진 문서.

[10] OpenAI. Measuring the Performance of Our Models on Real-World Tasks. 2025.

[11] H. Zhang et al. BenchCAD: A Comprehensive, Industry-Standard Benchmark for Programmatic CAD. arXiv:2605.10865v2, 2026.

[12] Anthropic. Eval awareness in Claude Opus 4.6's BrowseComp performance. 2026.

[13] B. Schoen and J. Nitishinskaya. Metagaming matters for training, evaluation, and oversight. 2026.

[14] Flitto Agent Bench. 플리토 공개 리더보드. US Consumer Credit 과제 분야와 채점 기준.

[15] 플리토 에이전트 데이터 리서처 팀. *Credit Bench Stories*. 정답 파일이 있었던 실패 3회. 결제 정산일 예시. 실패 유형 진단.

[16] Flitto Agent Bench. 공개 리더보드. 성능 · 효율. 공개 집계 JSON. 2026년 11월 20일자 스냅샷.

A 분야별 대표 업무의 판정 기준

공개 과제 소개의 여덟 업무 분야를 바탕으로 대표 사례를 구성했다 [14]. 각 사례는 결과를 산출하는 데 쓰는 자료와 판단 기준을 설명한다. 사례에는 공식 문항이나 공식 실행 결과가 포함되지 않는다.

개인 금융. 카드 회원이 7월 5일 현재 6월 명세서에 남은 청구액을 묻는다. 명세서와 납부 이력에서 해당 명세서에 반영된 납부액을 합산한다. 아직 반영되지 않은 납부와 다음 청구 주기의 구매는 6월 명세서에 반영된 납부액에 포함하지 않는다. 필수 결과는 반영된 납부 합계와 명세서에 남은 청구액이다.

결제 · 청구. 결제 운영 담당자가 7월 정산 건수와 순정산액을 요청한다. 접수 · 재시도 · 정산 이력을 원 결제 건별로 연결한다. 7월에 정산된 한 거래는 시도가 여러 번이어도 한 건으로 센다. 반환 기록에서 7월 말까지 해당 거래에 반영된 금액을 합산해 총정산액에서 뺀다. 정산 건수와 순정산액을 제출한다.

사기 · 분쟁. 분쟁 담당자가 9월에 접수돼 기준일까지 처리가 끝난 사건을 검토한다. 사건별 접수일과 완료일을 업무에 주어진 기한과 대조한다. 사건 번호로 보상, 취소, 확정 기록을 연결해 최종 보상액을 정한다. 분쟁 처리 기한을 넘긴 사건 수와 확정 보상액을 제출한다.

신용파일. 신용정보 분석가가 개설 시점의 신용점수 평균을 요청한다. 계정 개설일과 조회 기록을 연결한다. 예시의 규칙에 따라 개설일 기준으로 유효한 점수 중 가장 최근 값을 계정별로 고른다. 점수가 확인된 계정의 평균을 제출한다.

심사. 심사 책임자가 6월에 심사를 마친 신청의 승인 결과를 점검한다. 신청서와 최종 심사 기록을 연결해 승인 건을 확인한다. 각 승인 건의 월 상환 부담을 해당 신청의 월 소득으로 나눈다. 승인 건수와 승인 건의 평균 부채상환비율을 제출한다.

이탈 · 유지. 고객 유지 담당자가 3분기 시작 시점에 열려 있던 계정의 분기 이탈률을 요청한다. 계정 개설 · 해지 이력을 대조해 기초 계정 수를 정한다. 그 계정 중 분기 내에 해지된 계정만 이탈 건수에 포함한다. 분기 시작 계정 수, 해당 해지 건수, 이탈률을 제출한다.

추심. 추심 담당자가 월말 장기 연체 추심 대상 계정 수를 요청한다. 계정 원장과 월별 연체 이력을 연결한다. 예시의 추심 기준이 정한 연속 연체 기간을 충족하고 아직 대손상각되지 않은 계정의 수를 제출한다.

포트폴리오. 포트폴리오 분석가가 월말 신용한도 이용률을 요청한다. 월말에 보고할 계정을 정한 뒤 잔액 단위를 달러로 맞춘다. 예시의 규칙에 따라 음수 잔액은 미수채권 합계에서 0으로 처리한다. 같은 계정의 미수채권 합계를 신용한도 합계로 나눈다. 2.2절에서 세 계정으로 계산을 재현할 수 있다.

B 분야별 평가 결과

다섯 본 평가 설정을 대상으로 집계했다. 같은 분야의 문항은 자료룸이나 기본 설계를 공유할 수 있으므로 분야별 비율은 이 문항 세트에서의 관측값으로 해석한다. 공개 집계의 모델 · 분야별 결과로 계산했다 [16].

표 4: 두 회차의 분야별 완수 결과.

분야	문항	1회차	2회차	두 회차 통과	한 번 이상	평균
개인 금융	13	23/65	25/65	17/65	31/65	36.9%
결제 · 청구	13	28/65	30/65	25/65	33/65	44.6%
사기 · 분쟁	12	21/60	20/60	17/60	24/60	34.2%
신용파일	11	26/55	28/55	23/55	31/55	49.1%
심사	13	29/65	28/65	22/65	35/65	43.8%
이탈 · 유지	13	30/65	31/65	25/65	36/65	46.9%
추심	12	19/60	20/60	17/60	22/60	32.5%
포트폴리오	12	24/60	25/60	19/60	30/60	40.8%
합계	99	200/495	207/495	165/495	242/495	41.1%

C 추가 모델과 자원 사용

2026년 11월 20일자 리더보드 스냅샷은 설정 14개에서 설정당 99문항을 두 차례 실행한 2,772회 시도를 집계한다. 본문의 고정판 집계와 별도다 [16].

표 5: 2026년 11월 20일자 모델 설정별 완수 결과.

모델	1회차	2회차	평균	두 회차 통과	한 번 이상
Claude Fable 5.2	85	85	85.9%	84	86
Claude Opus 5.5	81	76	79.3%	75	82
Gemini 4 Carbon	76	74	75.8%	70	80
GPT-6.2 Astra	58	49	54.0%	44	63
Gemini 4 Argon	50	56	53.5%	44	62
Qwen4 Max	43	44	43.9%	36	51
GPT-6.2 Sol	43	39	41.4%	30	52
Hunyuan 4	32	42	37.4%	24	50
Muse Spark 1.3	34	32	33.3%	22	44
GPT-6 Astra	29	28	28.8%	19	38
Qwen3.8 Max 0902	26	29	27.8%	17	38
GLM-5.3 Prime	29	25	27.3%	16	38
Kimi K3.1 Pro	14	18	16.2%	10	22
Solar Pro 4	8	11	9.6%	7	12

문항당 경과 시간은 모델 설정별 두 회차의 경과 시간 합을 99문항으로 나눈 값이다. 그림은 이를 한 번 이상 통과한 문항의 비율과 함께 보여 준다 [16].

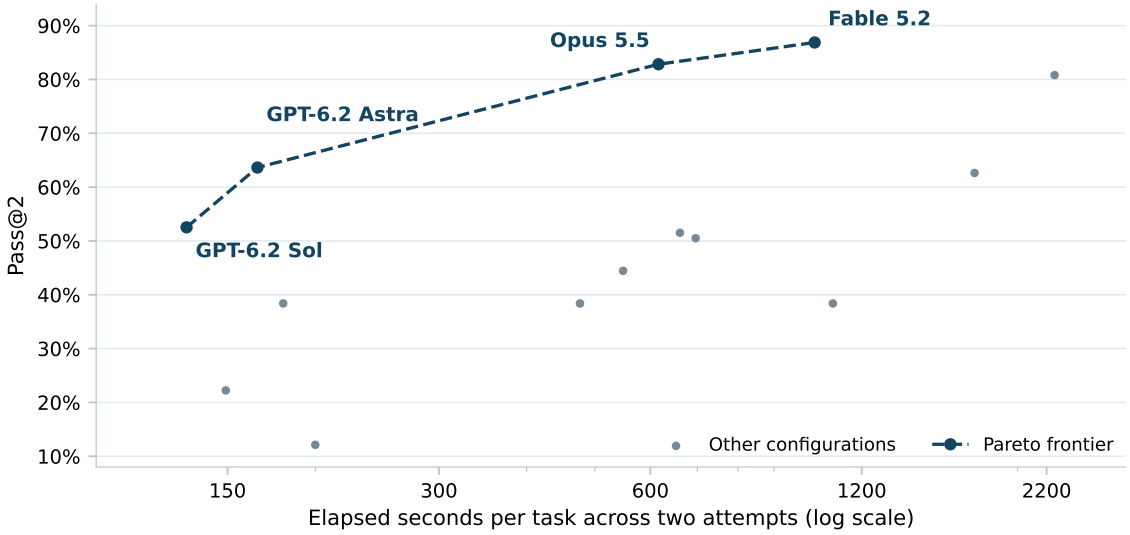

그림 1: 두 회차 합산 시간과 한 번 이상 통과 비율의 파레토 경계.