

Evaluating LLM Recommendation Safety under Hard Constraints

Tae-Yeon Eom*, Jae-Woong Kim**, Won-Young Jeon***

*Ph.D. Candidate, Dept. of Computer Engineering, Kongju National University, Cheonan, Korea

**Professor, Dept. of Software, Kongju National University, Cheonan, Korea

***Lecturer, Dept. of Software, Kongju National University, Cheonan, Korea

[Abstract]

This study compared final-action performance and rejection feedback conditions across four hard-constraint product recommendation procedures using a single local LLM (qwen3:8b). A fixed 40-monitor catalog, based on official manufacturer specifications and domestic prices, mapped 19 hard constraints to 46 product fields. With deterministic ground truth precomputed for 60 queries, we compared single pass (M1), self-refinement (M2), and Proposer-Critic with rejected-product-ID feedback (M3) or detailed rejection history containing violated constraint keys, required values, actual values, and natural-language reasons (M4). Under fixed evaluation conditions, final-action accuracy on REJECTION_STRESS was 40.0% (M1), 46.7% (M2), and 93.3% (M3 and M4). Both M3 and M4 recorded zero unsafe acceptances and correctly abstained on all 10 infeasible queries. Repeated same-constraint violation rates were 61.11% (M3) and 57.14% (M4), with a difference (M4-M3) of -3.97 percentage points (95% confidence interval: [-10.77, 0.00] percentage points). M4 showed a lower observed repeated-violation rate, although the confidence interval included zero, and slightly higher input-token use and response latency than M3.

▶ **Key words:** LLMs, hard-constraint recommendation, proposer-critic, abstention, deterministic ground truth

[요약]

본 연구는 단일 로컬 LLM(qwen3:8b) 기반 하드 제약 상품 추천의 네 절차와 기각 피드백 조건을 비교하였다. 제조사 공식 사양·국내 가격 기반의 모니터 40개 고정 카탈로그에 19개 하드 제약과 46개 상품 필드를 대응시키고, 60개 질의의 결정론적 정답을 사전 계산하였다. 비교 절차는 단일 통과(M1), 자체 수정(M2), 기각 상품 ID 피드백 기반 제안자-검토자(M3), 위반 제약 키·요구값·실제 값·자연어 이유의 상세 기각정보 이력 기반 제안자-검토자(M4)였다. 고정 평가 조건의 REJECTION_STRESS 층에서 최종 행동 정확도는 M1 40.0%, M2 46.7%, M3·M4 각각 93.3%였으며, M3·M4는 각각 위험한 승인 없이 실현 불가능 질의 10개 모두에서 올바르게 기권하였다. 동일 제약 반복 위반률은 M3 61.11%, M4 57.14%로 M4가 낮았으나, 차이(M4-M3) -3.97%p의 95% 신뢰구간 [-10.77, 0.00]%p는 0을 포함하였다. M4의 입력 토큰 사용량과 응답 지연은 M3보다 소폭 높았다.

▶ **주제어:** 대규모 언어 모델, 하드 제약 추천, Proposer-Critic, 기권, 결정론적 정답

- First Author: Tae-Yeon Eom, Corresponding Author: Jae-Woong Kim
- *Tae-Yeon Eom (utyv@nate.com), Dept. of Computer Engineering, Kongju National University
- **Jae-Woong Kim (jykim@kongju.ac.kr), Dept. of Software, Kongju National University
- ***Won-Young Jeon (jun1young@kongju.ac.kr), Dept. of Software, Kongju National University
- Received: 2026. 09. 08, Revised: 2026. 10. 08, Accepted: 2026. 10. 12.

I. Introduction

전자상거래와 대화형 검색 환경에서 대규모 언어 모델(LLM)은 복잡한 자연어 요구를 해석하고 상품 정보를 대화형으로 제시하는 도구로 활용되고 있다[1,2]. 이러한 활용에는 예산, 설치 공간, 해상도, 포트, 전력 공급처럼 명시적인 하드 제약의 준수 여부를 검토하는 절차가 필요하다. 후보 집합에 모든 조건을 만족하는 상품이 없는데도 모델이 임의의 상품을 선택하면, 사용자는 조건을 위반한 추천을 정상적인 대안으로 받아들일 수 있다. 따라서 하드 제약 상품 추천의 핵심은 선호에 가까운 후보를 제시하고, 최종 행동이 사용자 조건을 충족하도록 통제하는 데 있다.

최종 행동의 안전성은 질의의 실현 가능성에 따라 달라진다. 유효한 상품이 존재하는 질의에서는 조건을 만족하는 상품을 선택해야 하고, 유효한 상품이 존재하지 않는 질의에서는 기권(ABSTAIN)해야 한다. 본 연구는 상품을 생성하거나 선택하는 기능과 함께 부적합 후보를 검토하고, 승인 가능한 후보가 없을 때 탐색을 중단하는 절차를 포함하여 추천 구조를 비교한다.

본 연구는 동일한 로컬 LLM에 서로 다른 역할과 피드백 조건을 부여하여 추천 구조를 통제된 환경에서 비교한다. 구체적으로 단일 통과 및 자체 수정 방식과 제안자(Proposer)-검토자(Critic)의 검토·반복 탐색·기권 종료로 결합한 방식을 비교하고, 상세 기각정보 이력을 전달하는 방식과 기각 상품 ID만 전달하는 방식 사이의 동일 제약 반복 위반률 차이를 확인한다. 평가의 객관성을 확보하기 위해 공식 사양에 기반한 고정 상품 카탈로그와 파이썬 기반 결정론적 검사기를 구축하였다. 각 질의의 유효 상품 집합과 최종 정답 행동은 실험 전에 계산하여 동결하였다.

연구 목적에 따라 세 가지 연구 질문을 설정하였다. RQ1은 네 절차 조건(M1-M4)의 최종 행동 정확도, 위험한 승인 및 실현 불가능 질의의 올바른 기권이 어떻게 나타나는지를 묻는다. RQ2는 상세 기각정보 이력을 제공하는 M4와 기각 상품 ID만 제공하는 M3 사이의 기각 이후 동일 하드 제약 반복 위반률 차이를 검토한다. RQ3은 네 절차 조건의 모델 호출 수, 입력·출력 토큰, 전체 토큰 및 응답 지연을 비교한다.

본 연구의 기여는 세 가지다. 첫째, 실제 상품 사양과 가격 자료를 바탕으로 40개 모니터의 고정 카탈로그와 19개 하드 제약 스키마를 구축하고, 결정론적 검사기로 60개 질의의 정답을 사전에 산출하였다. 둘째, 동일 모델의 네 절차 조건을 비교하고, M3와 M4가 첫 기각 이전에 동일한 호출 결과를 공유하도록 통제하여 상세 기각정보 제공

에 따른 차이를 평가하였다. 셋째, 최종 행동 정확도뿐 아니라 위험한 승인, 올바른 기권, 동일 제약 반복 위반 및 호출·토큰·지연 비용을 함께 보고하여 안전성과 효율성을 동시에 평가하였다.

II. Preliminaries

2.1 LLM-Based and Constraint-Based Recommendation

LLM 기반 추천 연구는 사용자와 상품에 관한 정보를 텍스트로 표현하고, 프롬프트와 미세조정을 활용해 다양한 추천 과업을 수행해 왔다[1,2]. 이러한 접근은 사용자의 복잡한 요구를 대화로 파악하고 추천 근거를 자연어로 설명하는 데 활용된다. 상품 추천에서는 이와 함께 예산, 크기, 호환성 등 사용자가 명시한 조건의 충족 여부를 확인해야 한다. 제약 기반 추천(constraint-based recommendation)은 이러한 요구와 상품 속성을 제약으로 표현하고, 이를 충족하는 상품과 서비스를 탐색하는 접근이다[3]. 사용자가 필수로 지정한 조건은 하드 제약으로 정의하여 후보 상품의 적합성을 판단하는 기준으로 사용한다. 본 연구는 이 관점을 고정 상품 카탈로그에 적용하고, LLM 기반 후보 탐색 과정에서 하드 제약의 충족 여부를 평가한다.

2.2 Self-Refinement, Critiquing, and Verbal Feedback Memory

LLM의 반복 수정은 피드백의 출처와 활용 방식에 따라 구분할 수 있다. 동일 모델의 자체 피드백을 활용하는 Self-Refine은 초기 출력을 생성한 뒤 피드백과 수정을 반복하는 구조를 제안하였다[4]. 외부 도구의 피드백을 활용하는 CRITIC은 검색 엔진이나 코드 실행기와의 상호작용을 통해 모델 출력을 검토하고 수정한다[5]. 이와 관련하여 모델 자체 판단에 기반한 내재적 자기수정에서는 일부 추론 과업의 정확도 하락이 보고되었으며, 피드백 조건에 따른 성능 평가의 필요성이 제기되었다[6]. 따라서 단순히 호출 횟수를 늘리는 자체 수정과, 후보 생성 및 제약 검토의 역할을 분리하고 검토 결과를 명시적으로 전달하는 구조는 구분하여 평가할 필요가 있다. 본 연구의 M2는 직전 출력을 참고하는 반복 판단을 사용하며, M3와 M4의 피드백은 동일 LLM의 검토 결과에서 얻는다.

피드백을 후속 시도에 축적하여 활용하는 접근도 제시되었다. Reflexion은 시행착오에서 얻은 피드백을 언어적 성찰로 변환하여 일화 기억에 저장하고, 후속 시도의 의사

결정에 활용하는 접근을 제안하였다[7]. 본 연구의 M4는 이러한 이력 활용과 개념적으로 관련되며, 상품 ID와 검토자가 보고한 제약 키, 요구값, 실제값 및 위반 이유를 각각 이력으로 전달한다. M3-M4 비교에서는 이 상세 기각정보 묶음의 제공에 따른 차이를 평가한다.

2.3 Abstention and Structured Output Reliability

기권 연구는 질문의 답변 가능성에 따라 응답과 유보를 선택하는 능력을 다룬다. 이 과정에서는 모델이 자신의 지식 범위와 답변 가능성을 식별하는 자기지식이 평가 대상이 된다[8]. 이러한 판단을 학습에 반영하는 접근에서는 반복 생성한 답변의 정당 여부로 문항별 지식 점수를 추정하고, 이를 미세조정의 학습 가중치로 사용한다. 지식 점수가 0인 학습 사례에는 목표 응답 끝에 <IDK> 토큰을 추가하여 불확실성을 표현하도록 한다[9].

본 연구에서는 고정 후보 집합에서 유효 상품의 존재 여부를 결정론적으로 판정하여 정당 기권 여부를 정한다. 실행 중 후보 검토와 기권 결정은 LLM이 수행하고, 결정론적 검사기는 정당 산출과 결과 평가에 사용한다.

실제 시스템에서는 의미적 정당성뿐 아니라 출력 형식의 유효성도 중요하다. 문법 제약 디코딩은 형식 문법으로 생성 가능한 출력 공간을 제한하여 복잡한 구조의 생성을 안정화하는 접근을 제시한다[10]. 형식적으로 유효한 JSON에 담긴 상품 사양과 사용자 제약의 의미적 일치는 별도의 검증 대상이다. 따라서 구조화 출력의 형식 검증과 추천 결과의 의미 검증은 구분할 필요가 있다.

2.4 Synthesis of Prior Work and Positioning of the Present Study

본 연구는 LLM 추천, 제약 기반 추천, 자기수정, 언어적 피드백 기억 및 기권 연구를 하나의 고정 상품 환경에서 연결한다. M1은 단일 통과, M2는 동일 모델 자기수정 계열, M3는 역할 분리형 제안-검토와 기각 ID 피드백, M4는 상세 기각정보 이력을 포함한 조건으로 구성하여 네 절차의 성능과 M3-M4의 피드백 조건에 따른 차이를 비교한다. 또한 공식 상품 카탈로그와 결정론적 검사기를 정당 기준으로 사용하고, 실현 불가능 질의를 명시적으로 포함한다. 이에 따라 추천 내용의 정확성뿐 아니라 안전한 기권, 반복 위반 및 추론 비용을 함께 평가한다.

III. Methodology

3.1 Overview

본 연구는 단일 로컬 LLM 조건에서 네 가지 추천 절차의 성능을 비교하는 통제 실험으로 설계하였다. 각 실행에는 사용자 질의, 구조화된 하드 제약, 질의별 후보 상품 10개가 입력된다. 단일 통과와 자체 수정 방식은 제안자 역할만 사용하며, 역할 분리 방식은 제안자가 선택한 상품을 검토자가 활성 제약별로 다시 검사한다. 검토자가 상품을 기각하면 피드백 조건에 따라 다음 후보를 탐색하고, 최대 세 번의 반복을 마친 후에도 승인된 상품이 0개이면 ABSTAIN으로 종료한다.

실험은 입력 구성, 후보 제안, 제약 검토, 기각 정보 전달, 최종 행동 결정, 결정론적 정당(ground truth)과의 비교 순서로 진행하였다. 모든 방법의 최종 출력은 상품 승인 또는 기권으로 통일했으며, 최종 행동의 적합성뿐 아니라 반복 위반, 모델 호출 수, 토큰 사용량, 응답 지연을 함께 기록하였다. 전체 연구 절차는 Fig. 1과 같다.

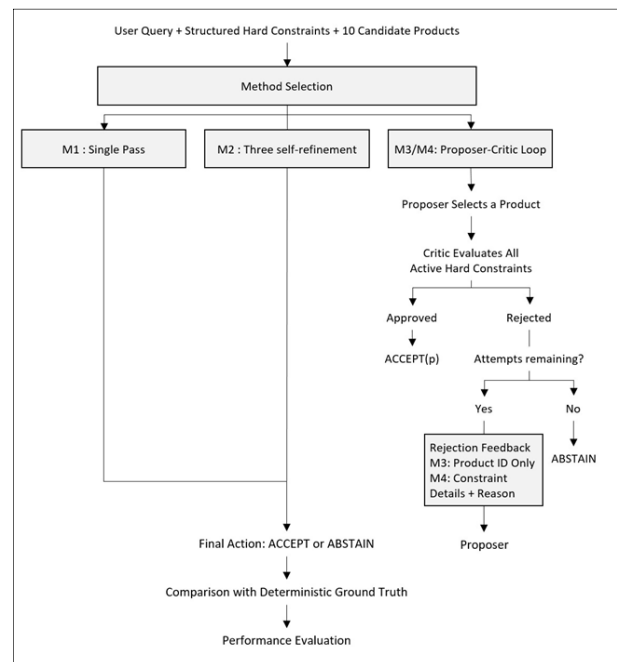


Fig. 1. Experimental workflows of the four methods and deterministic evaluation

3.2 Execution Environment and Evaluation Data

실험 조건의 재현성을 확보하고 외부 API 및 서비스 변경에 따른 영향을 줄이기 위해 모든 실행은 동일한 로컬 환경에서 수행하였다. 최종 평가에서는 모델, 추론 엔진, 생성 파라미터, 타임아웃 및 재시도 조건을 고정했으며, 구체적인 설정값은 Table 1에 제시하였다. 제안자는 상품

ID 또는 기권 여부를 간결하게 반환하는 반면, 검토자는 활성 제약별 판정 상태와 근거를 정해진 스키마에 따라 반환한다. 이러한 역할별 응답 길이의 차이를 반영하여 검토자에게 제안자보다 긴 최대 출력 범위를 적용하였다.

Table 1. Local Execution Environment and Model Configuration

Category	Configuration
OS & Hardware	Windows 11, RTX 3060 (12GB VRAM), 32GB RAM
Model & Engine	Ollama v0.33.2: Qwen3-8B (8.2B, Q4_K_M, ID: 500a1f067a9f)
Generation	Seed: 42, Temp: 0.0, Top-p: 1.0, Context: 8,192 tokens
Max Output	Proposer: 512 tokens Critic: 1,024 tokens
System	Timeout: 600s, Max HTTP retries: 2

상품 카탈로그는 제조사 공식 제품 페이지와 국내 가격 비교 자료를 기준으로 구축하였다. 초기 검토 대상 45개 모니터 중 가격 비교가 중단된 5개를 제외하여 최종 카탈로그를 40개 상품으로 구성하고, 상품별 사양·가격 출처와 수집일을 기록하였다. 상품 사양은 공식 자료에서 확인된 값을 사용하고, 결측값은 결측 상태로 유지하였다. 판정에 필요한 값에 결측 또는 형식 오류가 있으면 검사기가 DATA_ERROR를 반환하도록 하였다. 상품 데이터의 규모, 제약 수, 판정 논리 및 주요 제약의 계산 방식은 Table 2에 요약하였다.

Table 2. Summary of the Product Catalog and Hard-constraint Design

Item	Design
Catalog and constraint structure	
Candidates	45 monitors
Final Catalog	40 products
Catalog Fields	46 fields
Hard Constraints	19 constraints
Decision Logic	Logical AND across all active constraints
Error Handling	No imputation; the checker returns DATA_ERROR
Example decision rules	
Budget	$p_{price} \leq c_{budget}$
Resolution	$(p_w \geq c_w) \wedge (p_h \geq c_h)$
USB-C PD	$p_{video} = \text{TRUE} \wedge (p_w \geq c_h)$
Height	$p_{height, \min} \leq c_{height, \max}$

최종 평가 세트는 일반적인 추천 상황을 다루는 GENERAL 층과 기각 이후 동일 제약 반복 위반을 평가하는 REJECTION_STRESS 층에 각각 30개씩, 총 60개 질의로 구성하였다. 각 층에는 유효 상품이 존재하는 실현 가능 질의 20개와 기권을 정답으로 하는 실현 불가능 질의 10개를 배치하였다. 모든 질의에는 후보 상품 10개를 제공했으며, 활성 제약이 1개인 질의는 12개, 2개와 3개인 질의는 각각 24개로 구성하였다. REJECTION_STRESS에서는 후보 목록의 첫 두 상품이 동일한 제약을 위반하도록 구성하고, 실현 가능한 질의의 유효 상품은 뒤쪽에 배치하였다. 이를 통해 앞선 후보의 기각 이후 같은 제약을 다시 위반하는 후보를 제안하는지 평가하도록 하였다.

3.3 Comparison Methods and Evaluation

Framework

단일 통과(M1), 자체 수정(M2), 기각 상품 ID 피드백(M3), 상세 기각정보 이력(M4)을 Table 3과 같이 비교하였다. M2는 직전 출력을 참고한 세 번째 판단을 최종 행동으로 사용한다. M3와 M4는 제안자와 검토자의 역할을 분리하며, M4의 상세 기각정보는 상품 ID, 위반 제약 키, 요구값, 실제값 및 자연어 이유로 구성된다.

Table 3. Role Configurations and Feedback Conditions for the Four Methods

Method	Role configuration	Number of attempts	Feedback used in the next attempt
M1	Proposer only	1	None
M2	Proposer only	3	Previous output
M3	Proposer-Critic	Up to 3	Rejected product ID
M4	Proposer-Critic	Up to 3	Full rejection details

기각 피드백에 따른 차이를 비교하기 위해 M3와 M4에는 동일한 모델·프롬프트·출력 스키마를 적용하고, 첫 제안과 첫 검토의 호출 결과를 공유하였다. 기각된 상품은 두 방법 모두 후속 후보에서 제외하였다. 실행 순서의 영향을 줄이기 위해 M1, M2, M3·M4 블록의 여섯 가지 순열을 60개 질의에 균형 있게 배치하였다.

정답은 파이썬 기반 결정론적 검사기로 산출하였다. 질의별 후보 상품 집합을 P , 활성 하드 제약 집합을 C , 상품 p 가 제약 c 를 만족하는지를 Satisfied(p, c)로 정의하였다. 검사기는 각 상품-제약 쌍에 대해 SATISFIED, VIOLATED, DATA_ERROR 중 하나를 반환하며, 검사 결과가 SATISFIED인 경우에만 Satisfied(p, c)를 참으로 처리하였다. 모든 활성 제약을 만족하는 유효 상품 집합

$V(P, C)$ 는 식 (1)과 같이 정의하였다. 데이터 오류 (DATA_ERROR)는 제약 위반(VIOLATED)과 구분하여 판정을 유보하며, 최종 평가 질의 60개에서 DATA_ERROR는 0개였다.

$$V(P, C) = \{p \in P \mid \forall c \in C, \text{Satisfied}(p, c)\} \quad (1)$$

질의의 실현 가능성은 후보 집합에 유효 상품이 하나 이상 존재하는지를 기준으로 판정하였다. 식 (2)에서 $V(P, C)$ 가 공집합이 아니면 실현 가능한 질의로, 공집합이면 실현 불가능한 질의로 분류하였다.

$$\text{Feasible}(P, C) \Leftrightarrow V(P, C) \neq \emptyset \quad (2)$$

복수의 상품이 모든 조건을 만족할 수 있으므로 결정론적 정답은 하나의 상품 ID가 아니라 허용 가능한 최종 행동의 집합으로 구성하였다. 식 (1)에서 정의한 유효 상품 집합을 V 라 할 때, $V \neq \emptyset$ 이면 V 에 속한 상품 중 하나를 승인하는 행동을 정답으로 인정하고, $V = \emptyset$ 이면 ABSTAIN만을 정답으로 인정하였다. 최종 정답 행동 집합 A^* 는 식 (3)과 같다.

$$A^* = \begin{cases} \{\text{ACCEPT}(p) \mid p \in V\}, & V \neq \emptyset, \\ \{\text{ABSTAIN}\}, & V = \emptyset. \end{cases} \quad (3)$$

모델이 반환한 최종 행동을 $\hat{a}$ 라고 할 때, $\hat{a} \in A^*$ 이면 올바른 행동으로 판정하였다. 모델이 $\text{ACCEPT}(p)$ 를 반환한 경우에는 선택 상품 p 가 식 (1)의 유효 상품 집합에 포함되어야 한다. 실현 불가능한 질의에서는 식 (2)에 따라 ABSTAIN을 반환한 경우에만 정답으로 처리하였다.

최종 행동 정확도는 전체 평가 질의 가운데 모델의 행동이 정답 행동 집합에 포함된 비율로 계산하였다. 추천 안전성은 유효 상품 집합 밖의 상품을 승인한 위험한 승인 건수와 실현 불가능 질의에서 ABSTAIN을 반환한 올바른 기권률로 평가하였다. JSON 파싱 오류로 발생한 RUN_ERROR는 오답에 포함하여 각 방법의 정확도 계산에서 전체 60개 질의를 분모로 유지하였다.

기각 이후의 탐색 과정은 동일 제약 반복 위반률로 평가하였다. 이는 기각 이후 상품 ID를 반환한 후속 제안 가운데 앞선 기각에서 위반한 제약 키를 다시 위반한 제안의 비율이다. M3와 M4의 차이(M4-M3)에 대한 95% 신뢰구간은 REJECTION_STRESS의 30개 질의를 단위로 두 방법의 결과를 함께 200,000회 재표집하는 대응표본 부트스

트랩으로 산출하였다.

추론 비용은 정상 완료된 실행의 질의당 평균 모델 호출 수, 프롬프트·출력 토큰 수 및 응답 지연으로 측정하였다. 최종 행동 정확도, 위험한 승인, 올바른 기권 및 추론 비용은 고정된 평가 세트의 기술통계로 비교하였다. 최종 분석에는 검토자의 출력 한도와 기각 상품 재제안 통제를 보완한 v1.1의 전체 재실행 결과를 사용했으며, 카탈로그·질의·정답·프롬프트·검사 코드와 실행 로그를 버전 및 해시 기록과 함께 보관하였다.

IV. Results and Analysis

4.1 Run Completion and Final-Action Performance

전체 60개 질의에 네 가지 방법을 적용하여 240개 질의-방법 쌍을 실행하였다. 235개(97.92%)는 정상 완료되었고, 5개(2.08%)는 검토자 응답의 JSON 파싱 오류로 RUN_ERROR 처리하였다. 5건은 모두 GENERAL 층에서 발생했으며, REJECTION_STRESS 층의 120개 실행은 모두 완료되었다. RUN_ERROR는 오답으로 처리하고, 각 방법의 전체 60개 질의를 분모로 유지하였다.

방법별 결과는 Table 4에 제시하였다. M1과 M2는 REJECTION_STRESS의 실현 불가능 질의에서 올바른 기권 0/10을 기록했으며, 부적합 후보를 승인하는 사례가 발생하였다. 제안자와 검토자를 분리한 M3와 M4는 해당 평가층에서 위험한 승인 0건을 기록하고, 실현 불가능 질의 10개에 모두 기권하였다. 전체 60개 질의에서도 역할 분리 방식의 최종 행동 정확도가 단일 통과와 자체 수정 방식보다 높게 나타났다. 이 결과는 검토·반복 탐색·기권 종료를 결합한 절차의 관찰 성능을 나타낸다.

Table 4. Safety and Final-action Performance by Method

Method	REJECTION_STRESS subset				Full evaluation set		
	Correct	Acc. (%)	Unsafe	Infeasible	Correct	Acc. (%)	Errors
M1	12/30	40.0	18	0/10	28/60	46.7	0
M2	14/30	46.7	16	0/10	31/60	51.7	0
M3	28/30	93.3	0	10/10	55/60	91.7	3
M4	28/30	93.3	0	10/10	55/60	91.7	2

Note. REJECTION_STRESS results are based on 30 queries per method. Correct abstention is calculated over the 10 infeasible queries. Full-set accuracy is calculated over all 60 queries per method, with run errors counted as incorrect final actions.

4.2 Repeated Violations and Differences Associated with Detailed Rejection History

상세 기각정보 이력에 따른 동일 제약 반복 위반률 차이는 REJECTION_STRESS에서 M3와 M4의 기각 후 후속 제안을 기준으로 평가하였다. Table 5와 같이 M4의 반복 위반률은 M3보다 3.97%p 낮았다. 대응표본 부트스트랩으로 산출한 95% 신뢰구간은 $[-10.77, 0.00]\%$ 로 0을 포함하므로, 관찰된 차이의 크기와 추정치의 불확실성을 함께 고려한다. M3와 M4는 REJECTION_STRESS의 최종 행동 정확도, 위험한 승인 및 올바른 기권에서 같은 집계값을 보였다. 상세 기각정보 이력은 반복 위반률이 낮아지는 방향의 관찰 차이를 보였으며, 비교의 범위는 위반 제약 카요구값·실제값·자연어 이유를 함께 제공하는 정보 묶음이다. 질의별 행동의 일치와 방법 간 동등성은 별도의 대응 분석에서 평가할 항목이다.

Table 5. Repeated Same-constraint Violation Rates

Method	Repeated violations	Subsequent proposals	Violation rate (%)
M3: Rejected-ID feedback	22	36	61.11
M4: Detailed rejection history	20	35	57.14

Between-method difference: $M4 - M3 = -3.97$ percentage points

95% paired bootstrap confidence interval: $[-10.77, 0.00]$ percentage points

4.3 Operational Costs and Overall Interpretation

Table 6. Average Inference Costs by Method

Method	Model Calls	Prompt Tokens	Output Tokens	Latency (s)
M1	1.00	3,169.3	10.0	3.42
M2	3.00	9,667.8	31.1	11.30
M3	4.07	8,611.6	497.1	22.20
M4	4.03	8,708.2	494.8	22.47

Note. Values are averaged over completed runs: M1 and M2, $n = 60$ each; M3, $n = 57$; M4, $n = 58$.

방법별 평균 추론 비용은 Table 6에 제시된 바와 같다. M3와 M4는 M1보다 호출 수, 입력·출력 토큰 및 응답 지연이 모두 높았다. M2와 비교하면 호출 수, 출력 토큰 및 응답 지연은 높았으며, 입력 토큰과 전체 토큰은 낮았다. 입력과 출력을 합산한 전체 토큰은 M2 9,698.9개, M3 9,108.7개, M4 9,203.0개였다. M3와 M4의 평균 호출 수, 토큰 사용량 및 응답 지연은 유사한 수준이었다.

M4는 M3보다 입력 토큰이 평균 96.6개, 전체 토큰이 94.3개 많았고, 응답 지연은 0.27초 길었다. 평균 호출 수

는 0.04회, 출력 토큰은 2.3개 적었다.

결과를 종합하면, M3와 M4의 결합 절차는 M1과 M2보다 높은 최종 행동 정확도를 보였으며, REJECTION_STRESS에서는 더 높은 올바른 기권률을 기록하였다. 호출 수·출력 토큰·응답 지연도 높았으며, M2 대비 입력 토큰과 전체 토큰은 낮았다. 상세 기각정보 이력은 동일 제약 반복 위반률이 낮아지는 방향의 관찰 차이를 보였다. 이러한 결과는 각 절차의 성능과 비용을 지표별로 함께 평가할 필요성을 보여준다.

V. Conclusions

본 연구는 단일 로컬 LLM에서 네 가지 추천 구조(M1-M4)를 비교하였다. REJECTION_STRESS 평가에서 M3와 M4는 M1과 M2보다 높은 최종 행동 정확도를 기록했으며, 해당 평가층에서 위험한 승인 0건과 실현 불가능 질의 10개 모두의 올바른 기권을 기록하였다. 연구 결과는 후보 생성과 제약 검토, 반복 탐색 및 기권 종료로 결합한 절차의 관찰 성능을 보여준다. 상세 기각정보 이력을 제공한 M4의 동일 제약 반복 위반률은 M3보다 3.97%p 낮았으며, 95% 신뢰구간은 0을 포함하였다. M3와 M4는 REJECTION_STRESS의 최종 행동 지표에서 같은 집계값을 보였다. 이 결과는 상세 기각정보 묶음에 따른 관찰 차이로 해석하며, 차이의 크기와 신뢰구간을 함께 제시한다.

학술적 측면에서 본 연구는 고정 상품 카탈로그에서 동일 모델의 자체 수정과 역할 분리형 검토 절차를 비교하고, 상세 기각정보 제공에 따른 차이를 평가하였다. 정답 기준으로 결정론적 검사기를 사용했으며, 위험한 승인, 올바른 기권, 동일 제약 반복 위반 및 호출·토큰·지연 비용을 함께 측정하였다. 평가 범위를 최종 추천 정확도에서 안전 성과 추론 비용까지 확장했다는 점에서 의미가 있다.

실무적으로는 조건을 충족하는 상품이 없을 때 임의의 상품을 추천하는 것보다 명시적으로 기권하는 능력이 중요하다. LLM의 후보 탐색과 자연어 설명에 고정 상품 데이터와 결정론적 제약 검사 결과를 결합하여 최종 승인을 통제하는 방안을 후속 시스템 검증의 대상으로 제시한다. 역할 분리형 검토 절차에서는 M2보다 호출 수·출력 토큰·응답 지연이 높고 전체 토큰은 낮았으며, 구조화 출력의 실행 오류도 관찰되었다. 실제 시스템에 적용할 때는 안전 성과 추론 비용 사이의 상충관계를 고려해야 한다.

본 연구는 qwen3:8b 단일 모델, 단일 seed, 모니터 단일 도메인, 한국어 질의, 고정된 40개 오프라인 카탈로그

를 사용하였다. 결과의 해석 범위는 이러한 조건에서 수행한 60개 질의의 관찰 결과이다. 제안자와 검토자는 동일 LLM의 역할로 구성되며, 정답 산출과 결과 평가는 별도의 결정론적 검사기가 담당한다.

후속 연구에서는 추천 결과의 안전성을 한층 강화하기 위해 결정론적 검증기, 실패 시 차단 상태 제어, 감사 로그를 결합한 시스템 보증 계층을 구축하고, 다양한 예외 상황에서 승인 차단과 기권이 적절히 작동하는지 오프라인 환경에서 평가할 예정이다. 호출 예산과 종료 규칙을 갖춘 조건에서 검토 절차와 기권 규칙을 각각 비교하고, 기각 상품 ID·구조화 기각정보·구조화 정보와 자연어 이유의 조합을 비교하여 구성요소별 기여를 평가할 예정이다. 복수 모델·seed·도메인과 독립 질의 세트를 대상으로 한 외적 타당성 검증은 장기 연구 과제로 확장할 필요가 있다.

REFERENCES

- [1] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)," Proceedings of the 16th ACM Conference on Recommender Systems, pp. 299–315, 2022. DOI: 10.1145/3523227.3546767.
- [2] Z. He, et al., "Large Language Models as Zero-Shot Conversational Recommenders," Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 720–730, 2023. DOI: 10.1145/3583780.3614949.
- [3] A. Felfernig and R. Burke, "Constraint-Based Recommender Systems: Technologies and Research Issues," Proceedings of the 10th International Conference on Electronic Commerce, Article 3, 10 pages, 2008. DOI: 10.1145/1409540.1409544.
- [4] A. Madaan, et al., "Self-Refine: Iterative Refinement with Self-Feedback," Advances in Neural Information Processing Systems, Vol. 36, pp. 46534–46594, 2023. DOI: 10.52202/075280-2019.
- [5] Z. Gou, et al., "CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing," International Conference on Learning Representations, 2024.
- [6] J. Huang, et al., "Large Language Models Cannot Self-Correct Reasoning Yet," International Conference on Learning Representations, 2024.
- [7] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," Advances in Neural Information Processing Systems, Vol. 36, pp. 8634–8652, 2023. DOI: 10.52202/075280-0377.
- [8] Z. Yin, Q. Sun, Q. Guo, J. Wu, X. Qiu, and X. Huang, "Do Large Language Models Know What They Don't Know?," Findings of the Association for Computational Linguistics: ACL 2023, pp. 8653–8665, 2023. DOI: 10.18653/v1/2023.findings-acl.551.
- [9] J. Lee, H. Jo, D. Ko, K. Chae, C. Park, and J. Kim, "What Models Know, How Well They Know It: Knowledge-Weighted Fine-Tuning for Learning When to Say 'I Don't Know'," arXiv preprint arXiv:2604.05779v2, 2026. DOI: 10.48550/arXiv.2604.05779.
- [10] S. Geng, M. Josifoski, M. Peyrard, and R. West, "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning," Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 10932–10952, 2023. DOI: 10.18653/v1/2023.emnlp-main.674.

Authors



Tae-Yeon Eom received the B.S. degree in Electrical, Electronic and Information Engineering from Kongju National University, Korea, and the M.S. degree in Library and Information Science from Chungnam National

University, Korea. He has completed the coursework for the Ph.D. program in Computer Engineering at Kongju National University, Korea. He is currently the CEO of Multiplaza. He is interested in e-commerce and related fields.



Jae-Woong Kim received the B.S. degree in Electronic Computer Science from Chung-Ang University, Korea, in 1983, the M.S. degree in Computer Engineering from Chung-Ang University, Korea, in 1988, and the Ph.D.

degree in Computer Engineering from Daejeon University, Korea, in 2000. Dr. Kim joined the faculty of the Department of Software at Kongju National University, Korea, in February 1992. He is currently a Professor in the Department of Software at Kongju National University. He is interested in software engineering, big data, and multimedia engineering.



Won-Young Jeon received the B.S. degree in Health Science from Dongnam Health University, Korea, in 2019, and the Ph.D. degree from Kongju National University, Korea, in 2026. Dr. Jeon joined the faculty

of the Department of Software at Kongju National University, Korea, in March 2026. She is currently a faculty member in the Department of Software at Kongju National University. She is interested in artificial intelligence, big data, LLM, and image processing.