JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 8, pp. 41-49, August 2026
https://doi.org/10.9708/jksci.2026.31.08.041

Numerical Noise Detection in E-commerce Product Data Using a Category-Relative Ordinal AUM-Based LLM Pipeline

Sharon Kim*, Seok-Yong Yun**

*Ph.D. Student, Department of AI Information Science, Myongji University, Seoul, Korea
**Professor, Department of AI Information Science, Myongji University, Seoul, Korea

[Abstract]

E-commerce catalogs contain numerical noise that stays within the statistical range of a category and thus evades distribution-based detection. This study proposes REPAIR, a training-free LLM pipeline that detects semantic inconsistency between product names and weight values by mapping weights into category-relative ordinal bins and scoring the margin and ordinal distance between the bin predicted via product-contextual Chain-of-Thought reasoning and the observed bin (Category-Relative Ordinal AUM). On 798 Amazon Berkeley Objects samples with anomalies constructed by In-Distribution Semantic Swap over five seeds, REPAIR records an AUROC of 0.844±0.012 (+0.128 over Standard AUM) and 0.824–0.868 across three LLMs, confirming model-agnostic behavior.

▸ **Key words**: Anomaly Detection, Large Language Model, E-commerce, Area Under the Margin, Chain-of-Thought, Data Quality

[요 약]

이커머스 카탈로그에는 단위 혼동, 입력 오류, 유사 상품 복사로 수치 노이즈가 빈번하며, 분포 기반 방법은 정상 범위 내 문맥적 이상치의 탐지가 어렵다. 이에 본 연구는 상품명과 무게의 의미론적 불일치를 탐지하는 LLM 파이프라인 REPAIR를 제안한다. REPAIR는 서브카테고리 내 실제 무게 값을 교환하는 In-Distribution Semantic Swap으로 분포 내 평가 데이터를 구성한다. 중심 기제는 무게를 카테고리 상대 서열 구간에 매핑하고, 상품명 기반 CoT 추론의 예측 구간과 실제 구간의 마진·서열 거리를 점수화하는 Category-Relative Ordinal AUM이다. Amazon Berkeley Objects 798개 표본의 5개 시드 실험에서 AUROC 0.844±0.012로 Standard AUM 대비 +0.128 향상되었고, Qwen2.5-7B, Llama-3.1-8B, GPT-4o-mini의 영어 부분집합(N=724)에서 AUROC 0.824–0.868로 모델 독립성을 보였다.

▸ **주제어**: 이상치 탐지, 대형언어모델, 이커머스, Area Under the Margin, Chain-of-Thought, 데이터 품질

• First Author: Sharon Kim, Corresponding Author: Seok-Yong Yun
*Sharon Kim (shaaaron@mju.ac.kr), Department of AI Information Science, Myongji University
**Seok-Yong Yun (icanibe@mju.ac.kr), Department of AI Information Science, Myongji University
• Received: 2026. 07. 06, Revised: 2026. 08. 03, Accepted: 2026. 08. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

이커머스 플랫폼에서 무게, 가격, 용량과 같은 수치 속성은 검색, 추천, 배송비 산정, 재고 관리에 직접 활용되므로, 그 오류는 단일 상품 정보의 부정확성에 그치지 않고 후속 서비스 전반의 신뢰도에 영향을 미친다. 데이터 중심 인공지능(Data-Centric AI) 연구는 데이터 품질이 AI 시스템의 성능과 신뢰성에 중요한 영향을 미친다고 강조하지만[1], 판매자 직접 등록과 자동 수집이 혼재한 실제 환경에서는 단위 혼동, 입력 오류, 유사 상품 정보 복사에 따른 수치 노이즈가 빈번히 발생한다. 예를 들어 'ultra-slim'으로 소개되는 휴대폰 케이스의 무게가 단위 혼동으로 12kg로 기록되면, 소비자는 무게 기반 검색과 필터에서 해당 상품을 찾지 못하거나 과다 산정된 배송비를 안내받게 되고, 판매자는 반품 처리와 문의 대응 비용을 부담하게 된다.

수치 노이즈의 어려움은 값이 항상 극단값으로 나타나지 않는다는 데 있다. 문맥적 이상치(contextual anomaly)는 전체 분포에서는 정상처럼 보이더라도 객체와 속성의 조건부 관계 안에서는 비정상으로 판정될 수 있다는 점에서 점 이상치와 구별된다[2][3].

기존 통계 기반 방법인 Interquartile Range(IQR), Z-score, Isolation Forest는 수치 분포의 위치와 이탈 정도를 기준으로 이상치를 판단한다[4][5]. 따라서 동일 카테고리 내부에 실제 존재하는 값이 다른 상품으로 교환되면 해당 값은 분포 내부에 머무르므로, 상품명과 수치 속성 사이의 의미론적 모순을 포착하기 어렵다.

최근 대형언어모델(Large Language Model, LLM)은 상품명, 수식어, 재질, 용도와 같은 자연어 단서로부터 물리적 기대치를 추론할 수 있는 가능성을 보이나[6], 절대 무게 값을 직접 예측하게 하는 접근은 환각, 길이 편향, 도메인 지식 부족의 영향을 받을 수 있다[7].

이에 본 연구는 상품명과 무게 속성 사이의 문맥적 불일치를 추가 분류기 학습 없이 점수화하는 LLM 파이프라인 REPAIR를 제안한다. 본 연구의 기여는 다음과 같다. 첫째, 동일 서브카테고리 내 실제 무게 값을 교환하는 In-Distribution Semantic Swap으로 분포 내부의 문맥적 이상치 평가 프로토콜을 구성한다. 둘째, 절대 무게 예측 대신 카테고리 상대 서열 구간을 예측하는 Category-Relative Ordinal Binning과 Product-Contextual CoT 프롬프팅[8]을 결합한다.

셋째, 예측 구간과 실제 구간의 AUM 마진[9]에 서열 거리를 반영하는 Ordinal AUM 점수화를 제안하고, ABO 데이터셋의 5개 시드 반복 실험과 세 종류의 LLM에서 효과를 검증한다.

II. Related Work

1. Contextual and Tabular Anomaly Detection

문맥적 이상치 탐지는 객체와 속성값 사이의 조건부 양립성 평가에 초점을 둔다. Foorthuis는 점·문맥적·집합 이상치를 구분하여 개념적 기반을 제시하였고[2], Song 등은 속성 간 조건부 결합 패턴으로 이상 여부를 판단하는 프레임워크를 제안하였다[3]. 이후 비즈니스 프로세스 및 객체-문맥 관계 연구는 의미적 문맥이 누락될 때 통계적으로 드문 사건과 실제 이상 사건이 혼동될 수 있음을 보였으나[10][11], 텍스트 의미와 수치 속성이 결합된 상품 메타데이터의 수치-의미 양립성은 직접 다루지 않았다.

통계 및 트리 기반의 표 데이터 이상치 탐지 방법은 상품명과 수치값 사이의 의미론적 양립성을 직접 모델링하지 않는다. IQR은 사분위 범위로 극단값을 식별하고[4], Isolation Forest는 관측값의 격리 용이성을 활용하며[5], Deep Isolation Forest는 비선형 구조에 대한 적응성을 높였으나[12], 세 방법 모두 판단 근거가 수치 분포 자체에 한정된다.

2. Training-Dynamics and LLM-Based Diagnostics

샘플 수준 데이터 품질 진단 연구는 학습 동역학을 활용한다. Example Forgetting은 반복적으로 잊히는 샘플을 어려운 사례로 해석하였고[13], Dataset Cartography는 신뢰도와 변동성으로 데이터셋을 진단하였으며[14], AUM은 정답 클래스와 경쟁 클래스 사이의 마진을 누적하여 라벨 오류 가능성이 높은 샘플을 식별한다[9]. 이들은 주로 분류 라벨 오류를 대상으로 하며, 연속형 수치 속성과 상품명 사이의 의미론적 불일치를 직접 점수화하지 않았다.

최근 LLM 기반 연구는 표 데이터 이상치 탐지에 의미론적 문맥을 도입할 수 있는 가능성을 제시하였다. AnoLLM은 텍스트 직렬화와 NLL을 활용해 표 데이터 이상치를 점수화하였고[15], AD-LLM은 LLM의 이상치 탐지 능력을 벤치마크하였다[16]. ReTabAD는 텍스트 메타데이터와 도메인 의미론의 복원이 탐지 성능과 해석 가능성에 기여함을 보고하였다[17]. 이들 접근은 수치 속성을 카테고리 상대 서열 구간으로 변환하거나 예측 구간과 실제 구간 사이의 마진과 서열 거리를 점수화하지 않으며, 생성 확률 단독 신호의 한계는 4.2절의 Target-PPL 결과로 확인된다.

LLM 기반 접근은 모델 내부 지식에 과도하게 의존할 때 환각이나 도메인 지식 부족의 영향을 받을 수 있다[7]. RAG는 외부 근거 검색으로 환각을 완화하는 접근이나[18], 본 연구는 상품명 내부의 물리 단서와 카테고리 상대 구간을 결합하는 경량 탐지 구조를 우선 설계한다.

본 연구가 채택한 LLM 파이프라인은 단일 모델 호출이 아니라 복수의 처리 단계를 연쇄해 하나의 과제를 수행한다. 복잡한 과제를 다단계 프롬프트 연쇄로 분해하면 단일 호출 대비 제어 가능성과 중간 결과의 투명성이 향상됨이 보고되었고[19], 각 단계를 선언적 모듈로 추상화하여 파이프라인을 체계적으로 설계하고 최적화하는 프로그래밍 모델도 제안되었다[20]. REPAIR는 이러한 설계 관점을 따라 구간화, 프롬프트 기반 구간 확률 추출, 마진 점수화를 검증 가능한 독립 단계로 분리한 학습-불필요 파이프라인이다.

선행 연구는 문맥적 판정, 통계 기반 탐지, 학습 동역학 진단, LLM 의미론의 가능성을 제시하였다. 그러나 텍스트 의미와 수치 속성이 결합된 상품 메타데이터에서 카테고리 상대 구간, 상품명 문맥 추론, 마진 기반 점수화를 하나의 탐지 파이프라인으로 통합한 연구는 제한적이다.

III. The Proposed Scheme

1. System Overview

REPAIR는 문맥적 수치 노이즈를 데이터 내부 교환, 카테고리 상대 구간화, LLM 문맥 추론, Ordinal AUM 점수화의 네 단계로 탐지한다. Fig. 1과 같이 Amazon Berkeley Objects(ABO) 메타데이터에서 유효한 무게 값을 가진 상품만 필터링하고[21], 서브카테고리 단위 분위수를 기준으로 가벼운 상품군과 무거운 상품군을 구성한다.

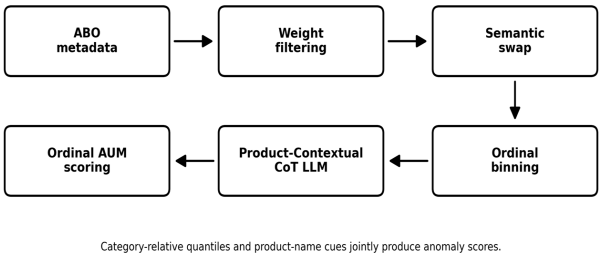

Fig. 1. Overview of the proposed REPAIR pipeline

해당 파이프라인의 핵심은 LLM을 절대 무게 생성기가 아니라 카테고리 내부 상대 위치 판단기로 사용하는 데 있다. LLM은 무게 값을 보지 못한 상태에서 상품명과 카테고리 정보만으로 예상 구간을 출력하며, 실제 구간과의 마진 및 서열 거리가 이상치 점수로 변환된다.

2. In-Distribution Semantic Swap

In-Distribution Semantic Swap은 분포 내부에 위치하지만 의미론적으로 불합리한 평가 데이터를 구성한다. 단순 극단값 주입은 IQR과 같은 통계 방법으로 쉽게 탐지되므로, 본 연구는 동일 서브카테고리 내부에서 실제 관측된 무게 값을 서로 교환하여 분포 내부성을 유지하면서 상품명과 수치값의 의미적 충돌을 생성한다. 주입되는 값이 동일 서브카테고리에 실재하는 다른 상품의 관측값이라는 점에서, 이 설계는 유사 상품 정보 복사처럼 다른 상품의 실제 값이 잘못 기입되는 오류 유형과 구조적으로 대응한다.

서브카테고리 C에 대해 분위수 Q_p^C를 산출하고, Q25와 Q75가 동일한 Zero-Variance 서브카테고리 및 표본 수가 10 미만인 서브카테고리를 제외한다. 이후 무게가 $1.05 \cdot Q10^C$ 이하($w_i \leq 1.05 \cdot Q10^C$)인 상품을 Group A, $0.95 \cdot Q90^C$ 이상($w_i \geq 0.95 \cdot Q90^C$)인 상품을 Group B로 정의하고 두 그룹의 무게 값을 1:1로 교환한다. 두 조건은 Q10에 +5%, Q90에 −5%의 허용 밴드를 각각 반영한 것이다. 교환 후 주입된 무게는 해당 카테고리에 실제 존재하는 값이므로 단순 분포 이탈 신호를 최소화한다.

카테고리 외부 값을 사용하면 탐지가 지나치게 쉬워지므로, 정상 표본과 이상 표본을 모두 동일 서브카테고리 내부에서 구성하여 평가 난이도를 현실적으로 유지한다. 다만 분포 폭이 좁은 서브카테고리에서는 두 그룹의 표본이 동일한 무게값을 가질 수 있어 교환 후에도 관측 무게가 변하지 않는 표본이 발생하며, 5개 시드 합계 7건(전체 이상치 1,989건의 0.35%)으로 이는 이상치 라벨의 소량 위양성으로 남는다.

3. Category-Relative Ordinal Binning

Category-Relative Ordinal Binning은 절대 무게 예측의 불안정성을 줄이기 위해 수치값을 카테고리 내부 상대 구간으로 변환한다. LLM에게 정확한 무게 값을 직접 예측하게 하면 수치 추정 오류와 환각의 영향을 받기 쉬우므로[7], 카테고리 내부 상대 위치를 나타내는 K=5개의 서열 구간을 예측 대상으로 설정한다. 카테고리 C의 p번째 분위수는 식 (1)과 같이 정의한다. 여기서 W_C^{pre}는 교환 적용 이전 카테고리 C에 속한 상품들의 원본 무게 집합 $\{w_i \mid item_i \in C\}$이며, 분위수 Q_p^C는 임의의 $p \in [0, 100]$에 대해 정의된다. 구간 경계 집합 B_C는 이 중 p

∈ {0, 20, 40, 60, 80, 100}의 여섯 분위수를 사용하며, 교환 범위 설정(3.2절)의 Q10, Q90 등도 동일한 정의를 따른다.

$$Q_p^C = Percentile_p\left(W_C^{pre}\right), p \in [0,100] \quad (1)$$
$$B_C = \left\{Q_p^C | p \in \{0,20,40,60,80,100\}\right\}$$

실제 무게 w_i를 카테고리 C 내부의 서열 위치로 매핑하는 함수 φ는 식 (2)와 같다. 구간 경계값은 하위 구간에 귀속되며(우폐 구간), 조건을 만족하는 k가 없는 상한 초과 값은 최대 구간(k=4)에 배정되므로 φ는 모든 w_i에 대해 정의된다.

$$\varphi(w_i, C) = \min\left\{k \in \{0,\ldots,4\} | w_i \le Q_{20(k+1)}^C\right\} \quad (2)$$

카테고리 상대 구간은 동일 값이라도 카테고리 내부 위치에 따라 다르게 해석하므로 상품 유형별 분포 차이를 탐지 과정에 반영한다. 구간 간 서열 거리는 식 (3)과 같이 정의된다.

$$d(j,k) = \frac{|j-k|}{K-1}, j,k \in \{0,\ldots,K-1\} \quad (3)$$

4. Product-Contextual CoT Prompting

Product-Contextual CoT Prompting은 LLM이 보유한 카테고리 수준의 물리 상식을 구간 판단으로 유도하는 구조화된 추론 인터페이스로서, 상품명에 포함된 재질·크기·용도 등 물리적 단서를 보조 근거로 명시화한다. CoT 프롬프팅은 중간 추론 단계를 명시하여 복합 추론 성능을 향상시키며[8], LLM은 일상 사물의 물리적 속성을 일부 추론할 수 있으나 객체-속성 수준에서는 한계를 가진다[6]. 4.4절의 상품명 셔플 실험은 상품명 문맥의 기여가 통계적으로 유의하되, 성능의 상당 부분이 카테고리 수준 상식에서 비롯됨을 보인다.

프롬프트는 Step 1에서 ultra-slim, heavy-duty steel, compact와 같은 물리적 수식어를 추출하고, Step 2에서 해당 단서가 카테고리 내부에서 상대적으로 가벼운지 무거운지를 판단하며, Step 3에서 0부터 4까지의 구간 중 하나를 출력하도록 제한한다. 모델의 마지막 토큰 위치에서 구간 k에 해당하는 숫자 토큰의 로짓 z_k를 추출하고, 소프트맥스를 적용하여 식 (4)와 같이 구간 확률 p_θ(k|x_i)를 계산한다.

$$p_\theta(k|x_i) = \frac{\exp(z_k)}{\sum_{j=0}^{4}\exp(z_j)}, k \in \{0,\ldots,4\} \quad (4)$$

프롬프트는 (i) 과제와 역할을 지정하는 시스템 지시, (ii) 카테고리명과 상품명을 제공하되 무게 값은 마스킹하는 입력부, (iii) Step 1-3의 추론 절차 지시, (iv) 마지막 행에 0-4 중 단일 숫자만 출력하도록 강제하는 출력 형식 제약으로 구성된다(Table 1의 Step 5). Qwen과 Llama는 동일한 프롬프트를 사용하고, GPT-4o-mini는 (iv)의 출력 지시 행을 2단계 파이프라인의 별도 사용자 턴으로 분리한다. 본 연구는 zero-shot 설정으로 few-shot demonstration을 포함하지 않으며, Step 1의 어휘 항목은 물리적 수식어의 의미 방향을 안내하는 힌트로서 입력-정답 쌍 형태의 시연이 아니다. 전체 프롬프트 전문은 Fig. 2와 같으며, 주 평가와 상품명 셔플 실험(4.4절)은 동일한 프롬프트를 사용한다. GPT-4o-mini의 2단계 추출은 temperature=0으로 수행되며, Stage 2에서 로그확률을 정규화하고 다섯 구간 토큰이 모두 관측되지 않으면 균등 분포로 폴백한다.

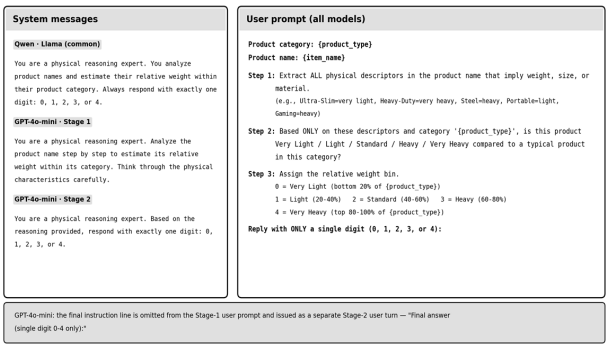

Fig. 2. Prompt templates of REPAIR (zero-shot; placeholders {product_type} and {item_name} follow the implementation)

5. Ordinal AUM Scoring

본 연구는 AUM의 마진 누적 관점[9]을 상품명 기반 예상 구간과 실제 주입 구간 사이의 불일치를 측정하는 점수로 확장하며, 각 표본의 마진은 식 (5)와 같이 계산한다.

$$m_i = p_\theta\left(b_i^{true}|x_i\right) - \max_{k \ne b_i^{true}} p_\theta(k|x_i) \quad (5)$$

여기서 b_i^pred = argmax_k p_θ(k|x_i)이다. m_i가 양수이면 모델이 실제 구간을 가장 높은 확률로 예측한 것이며, 음수이면 상품명으로부터 추론한 예상 구간이 주입된 무게 구간과 충돌함을 의미한다. 각 표본의 이상치 점

수 s_i는 식 (6)과 같이 정의된다. 음수 마진 표본 집합 S^-에서만 양의 점수가 발생하며, 정상 예측 표본은 max(0, -m_i) = 0이 되어 자동으로 0점을 갖는다.

$$s_i = d\left(b_i^{true}, b_i^{pred}\right) \cdot \max(0, -m_i) \quad (6)$$

Standard AUM(max(0, -m_i)로 정의)이 실제 구간과 경쟁 구간의 확률 차이만 반영하는 것과 달리[9], Ordinal AUM은 예측 구간과 실제 구간의 서열 거리를 함께 반영하여[22] 구간 차이가 클수록 더 큰 이상치 점수를 부여한다. b_i^true는 교환 후 주입된 무게를 기준으로 산출한다.

6. Algorithmic Procedure

Table 1은 REPAIR가 추가 분류기 없이 이상치 점수를 계산하는 절차를 요약한다. 입력은 상품명, 카테고리, 관측 무게이며, 출력은 각 표본의 이상치 점수이다.

Table 1. Pseudocode of the proposed REPAIR pipeline

Step	Procedure
1	Filter ABO items with valid product names, categories, and weight values.
2	For each subcategory C, compute quantiles and remove low-support or zero-variance categories.
3	Swap Q10 and Q90 weight groups to construct in-distribution semantic anomalies.
4	Map the injected weight into a category-relative ordinal bin using Eq. (1)-(3).
5	Prompt the LLM with masked weight information and extract bin probabilities using Eq. (4).
6	Compute margin and Ordinal AUM score using Eq. (5)-(6).

IV. Experimental Results

1. Experimental Setup

실험에는 Amazon이 공개한 실제 이커머스 상품 메타데이터인 Amazon Berkeley Objects(ABO) 데이터셋을 사용하였다[21]. 유효한 무게 값을 가진 상품을 필터링하여 102,141개의 정상 후보를 확보하고, In-Distribution Semantic Swap을 적용하여 검증 집합 798개와 시험 집합 798개를 구성하였다.

Table 2. Experimental configuration

Component	Configuration
Base LLM	Qwen2.5-7B-Instruct
Comparison LLMs	Llama-3.1-8B-Instruct, GPT-4o-mini
GPU	NVIDIA TITAN RTX (24GB)
Precision	float16
Number of bins (K)	5
Swap range	Q10-Q90
Distance matrix	5 × 5 ordinal (d = |j-k|/4)
Evaluation metrics	AUROC, AUC-PR (primary), F1 (reference)

평가 지표로는 AUROC와 AUC-PR을 주 지표로, F1을 참고 지표로 사용하였다. F1은 각 방법의 정규화 점수에 대해 [0, 1] 구간을 101개 지점으로 탐색하여 해당 평가 집합에서 F1이 최대가 되는 임계값에서의 값으로, 임계값 선택의 영향을 배제하고 방법 간 상한 성능을 동일 조건에서 비교하기 위한 것이다. 주 지표인 AUROC와 AUC-PR은 임계값 선택과 무관하다. 비교 방법에는 IQR, Z-score, Isolation Forest, Target-PPL, Standard AUM을 포함하였다[4][5][9]. Target-PPL은 NLL 기반 이상치 탐지[15]와 동일한 원리를 따르는 생성 확률 기반 참조선으로, 동일한 CoT 프롬프트와 상품명 입력 조건에서 산출하였다. 주 평가는 시드당 시험 집합 전체(N≈798)를 사용하였으며, 상품명 셔플 실험과 모델 독립성 비교는 각각 영어 부분집합 665개와 724개를 사용하였다.

2. Performance Comparison

제안 방법은 모든 기준선보다 높은 탐지 성능을 보였다. 다만 이상치 생성과 탐지가 카테고리 상대 좌표계를 공유하므로 통계 기반 기준선의 낮은 성능은 부분적으로 설계에 내재된 결과이며, 핵심 근거는 동일 좌표계에서 산출되는 Standard AUM 대비 향상에 둔다. 아울러 상품명 셔플에서의 유의한 성능 하락(4.4절), 세 이종 LLM에서의 일관된 성능(4.5절), 검증 집합에 없던 서브카테고리를 포함한 시험 평가(4.5절)는 이 향상이 데이터 생성 가정에 의존하지 않음을 보이는 보조 증거이다. Table 3은 5개 스왑 시드(seed 42-46)로 데이터 생성부터 평가까지 전체 파이프라인을 반복한 평균±표준편차이며, Ordinal AUM은 AUROC 0.844±0.012, AUC-PR 0.816±0.014를 기록하였다(대표 시드 42 기준 AUROC 0.855). IQR의 AUROC는 0.490으로 무작위 수준에 근접하여, 생성된 이상치가 통계적 극단값이 아니라 분포 내부의 문맥적 이상치임을 보여준다.

Table 3. Performance comparison across 5 swap seeds (mean±std, seeds 42-46; single-seed N≈798)

Method	AUROC	AUC-PR	F1
IQR	0.490±0.010	0.561±0.006	0.666±0.001
Z-score	0.606±0.019	0.604±0.014	0.678±0.012
Isolation Forest	0.527±0.012	0.578±0.009	0.666±0.001
Target-PPL	0.513±0.017	0.505±0.019	0.666±0.001
Standard AUM	0.716±0.021	0.699±0.022	0.707±0.016
Ordinal AUM (Proposed)	0.844±0.012	0.816±0.014	0.822±0.013

IQR, Isolation Forest, Target-PPL의 F1 0.666은 균형 데이터에서 모든 표본을 단일 클래스로 예측할 때 나타나는 값으로(Precision 0.5, Recall 1.0), 이들 방법이 문맥적 이상치를 실질적으로 변별하지 못함을 의미한다.

Standard AUM(0.716) 대비 Ordinal AUM(0.844)의 향상은 예측 구간과 실제 구간 사이의 거리 정보가 탐지 점수에 실질적으로 기여함을 보여준다[22].

한편 생성 확률만 사용하는 Target-PPL이 무작위 수준에 머문 결과(Table 3)는 NLL 계열 접근[15]과 마찬가지로 LLM 기반 탐지에서도 구간화와 마진 기반 점수화 같은 구조적 신호 설계가 필요함을 보여준다.

오탐 분석 결과, 오탐의 주원천은 카테고리 전형 대비 이례적으로 가벼운 정상 상품으로 확인되었다(예: 벽걸이형 디스플레이 마운트 중 경량 제품). 상위 100건 검수 큐 시나리오에서 precision@100은 5개 시드 평균 0.890±0.021이었으며(시드당 평균 오탐 11.0건), 오탐된 정상 표본 55건을 실제 구간별로 층화하면 최하위 구간(bin 0)의 오탐률이 0.143으로 가장 높고 최상위 구간(bin 4)이 0.037, 중간 구간(bin 1-3)은 0.000-0.014에 그쳤다. 이는 식 (6)의 점수가 실제 구간과 예측 구간의 서열 거리에 비례해 커지는 구조에서 비롯되며, 검수 큐 운용 시 극단 구간의 정상 상품에 대한 보조 확인 규칙을 결합하면 오탐 부담을 줄일 수 있다.

3. Statistical Significance

Ordinal AUM의 성능 향상은 모든 기준선 대비 통계적으로 유의하였다. 5개 스왑 시드(seed 42-46) 각각의 시험 집합에서 Ordinal AUM과 각 기준선의 AUROC 차이에 대해 쌍체 Bootstrap Resampling(N=10,000, one-sided)을 수행한 결과[23], Table 4와 같이 다섯 가지 비교 모두 5개 시드 전부(5/5)의 95% 신뢰구간 하한이 0을 초과하였고 최대 p-value가 0.001 미만으로, 성능 향상이 시드 구성과 표본 재추출 양쪽에서 안정적으로 유의하다.

Table 4. Statistical significance via paired Bootstrap Resampling across 5 swap seeds (N=10,000, one-sided; ΔAUROC = Ordinal AUM − baseline)

Comparison	Δ AUROC (mean ±std)	Range	Conservative lower CI	Max p	Sig. seeds
IQR	+0.354 ±0.016	[+0.330, +0.372]	+0.278	<0.001	5/5
Z-score	+0.237 ±0.019	[+0.223, +0.269]	+0.174	<0.001	5/5
Isolation Forest	+0.317 ±0.013	[+0.300, +0.329]	+0.250	<0.001	5/5
Target-PPL	+0.331 ±0.023	[+0.304, +0.365]	+0.254	<0.001	5/5
Standard AUM	+0.128 ±0.011	[+0.117, +0.144]	+0.084	<0.001	5/5

4. Ablation Study

절제 연구는 교환 범위, 서열 거리 페널티, 상품명 정보의 세 요인을 분리하여 검증한다. 첫째, Fig. 3와 같이 Cat.-IQR Rate는 교환 범위가 넓어질수록 18.0%에서 42.1%까지 증가하여 통계적 노출도가 커진다.

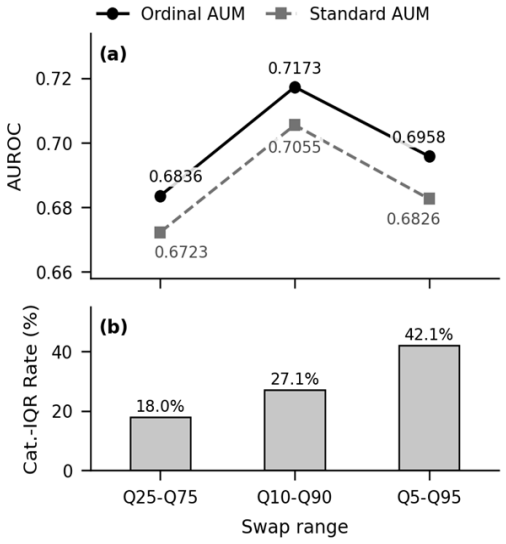

Fig. 3. Ablation on the swap range (validation set): (a) AUROC of Standard and Ordinal AUM, (b) Cat.-IQR Rate

Fig. 3의 AUROC는 검증 집합 기준으로 산출되어 시험 집합 기준인 4.2절의 수치와 절대값을 직접 비교하지 않는다. Ordinal AUM은 Q10-Q90에서 가장 높은 0.7173을 기록한 반면, 통계적 노출도가 가장 높은 Q5-Q95(Cat.-IQR Rate 42.1%)에서는 오히려 0.6958로 감소하여 통계적 노출도와 의미론적 탐지력이 단조 관계를 이루지 않음을 보인다.

이에 따라 두 요인이 균형을 이루는 Q10-Q90을 기본 스왑 범위로 설정한다. 둘째, 세 교환 범위 모두에서 Ordinal AUM이 Standard AUM보다 일관되게 높아 서열 거리 페널티가 탐지 성능에 기여함을 확인하였다.

셋째, 상품명 정보의 기여도를 검증하기 위해 각 상품명을 동일 서브카테고리 내 다른 상품명으로 교체하는 within-category title shuffle 실험을 셔플 가능한 영어 부분집합 665개에서 수행하고, 원본과 셔플 조건의 AUROC를 쌍체 부트스트랩(N=10,000)으로 비교하였다. 하락은 세 LLM 모두 유의하였으나(Qwen 0.857→0.807, $p<0.001$; Llama 0.871→0.840, $p<0.01$; GPT 0.824→0.683, $p<0.001$), Qwen과 Llama의 하락폭은 작아 성능의 상당 부분이 카테고리 수준의 물리 상식에서 비롯됨을 확인하였다[6]. GPT의 큰 하락은 2단계 CoT 구조에서 상품명 기반 추론이 이후 단계로 전파되어 개별 상품명 해석에 상대적으로 더 의존하기 때문으로 해석된다.

5. Model-Agnostic Evaluation

모델 독립성 실험은 REPAIR의 핵심 신호가 특정 LLM에 종속되지 않음을 보여준다. Llama와 GPT의 비라틴 문자 처리 한계를 통제하기 위해 비라틴 스크립트 상품명 74개를 제외한 영어 부분집합 724개를 사용하였으며, Table 5에서 세 모델 모두 AUROC 0.824 이상을 기록하였고 최고(Llama-3.1-8B, 0.868)와 최저(GPT-4o-mini, 0.824)의 차이는 0.044에 그쳤다.

Table 5. Model-agnostic evaluation (English subset, N=724)

Model	AUROC	AUC-PR	F1
Qwen2.5-7B	0.853	0.823	0.841
Llama-3.1-8B	0.868	0.837	0.841
GPT-4o-mini	0.824	0.796	0.769

Llama-3.1-8B는 채팅 템플릿과 시스템 프롬프트 적용 후 숫자 토큰 예측 편향이 완화되어 AUROC가 0.47 수준에서 0.87 수준으로 개선되었다. 이는 REPAIR가 모델의 원시 생성 능력보다 구조화된 추론 인터페이스에 의존함을 시사한다.

시험 집합에는 검증 집합에서 관측되지 않은 서브카테고리 29개(17.8%)가 포함되었으며, 이를 포함한 전체 평가에서도 대표 시드(seed 42) 기준 AUROC 0.855(영어 부분집합 0.853)로 성능이 유지되어, 검증 집합에서 선택한 설계 요소(스왑 범위, K=5)가 미관측 서브카테고리로 일반화됨을 확인하였다.

V. Conclusions

본 연구는 상품명과 수치 속성 사이의 문맥적 불일치를 탐지하기 위해 카테고리 상대 Ordinal AUM 기반 LLM 파이프라인인 REPAIR를 제안하였다. 제안 방법은 In-Distribution Semantic Swap, Category-Relative Ordinal Binning, Product-Contextual CoT, Ordinal AUM Scoring을 결합하여 분포 내부의 문맥적 수치 노이즈를 점수화한다.

ABO 데이터셋에 대한 5개 스왑 시드(seed 42-46) 반복 실험에서[21] 제안 방법은 AUROC 0.844±0.012를 달성하였고, IQR 대비 +0.354, Standard AUM 대비 +0.128의 AUROC 향상을 보였다. 쌍체 Bootstrap Resampling에서 다섯 가지 비교 모두 모든 시드의 95% 신뢰구간 하한이 0을 초과하였으며[23], 영어 부분집합의 세 종류 LLM에서 AUROC 0.824-0.868을 기록하였다. 실무 적용 관점에서 제안 점수는 고정 임계값 분류기가 아니라 검토 우선순위를 부여하는 순위 신호로 설계되었으며, 인력 검수 용량에 맞춰 상위 k건을 선별하는 운용이 가능하다.

본 연구는 단일 공개 데이터셋 중심 검증, 물리적 수식어가 부족한 상품명에서의 추론 신호 약화, API 기반 LLM의 추론 비용이라는 한계를 가진다. 평가 데이터가 In-Distribution Semantic Swap으로 합성되었다는 점 또한 적용 범위를 규정한다. 실제 운영 환경의 수치 노이즈는 단위 입력 오류, 오타, OCR 오류, 소수점 위치 오류, 유사 상품 정보 복사 등 다양한 기제로 발생하며, 그 결과값은 분포의 극단으로 이동하는 경우와 정상 범위 내부에 남는 경우로 나뉜다.

전자는 기존 분포 기반 기법으로 탐지될 여지가 있는 반면, 본 벤치마크는 후자, 즉 오류 값이 분포 내부에 남아 통계적 신호가 사라지는 가장 어려운 부분집합을 겨냥한다. 스왑은 복사형 오류와 구조적으로 대응하며 소수점·단위 오류 중 결과값이 정상 범위에 남는 사례까지 포괄하는 이상화된 모델이나, 오류 기제별 발생 빈도와 값 분포를 정량적으로 재현하지는 않는다. 따라서 본 결과는 분포 내부에 은닉되는 문맥적 오류에 대한 탐지 성능 평가로 해석되어야 한다. 또한 이상치 생성과 탐지가 좌표계를 공유하는 설계 의존성과 무관하게, Standard AUM 대비 Ordinal AUM의 향상은 유의하였다(Table 4). 이를 보완하기 위해 향후 연구에서는 검색 증강 기반 자동 교정 모듈을 결합하고, 가격·용량 등 다른 수치 속성으로 확장하여 자연 발생 오류에 대한 검증을 수행할 계획이다.

REFERENCES

[1] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric Artificial Intelligence: A Survey," ACM Computing Surveys, Vol. 57, No. 5, Article No. 129, pp. 1-42, May 2025. DOI: 10.1145/3711118

[2] R. Foorthuis, "On the Nature and Types of Anomalies: A Review of Deviations in Data," International Journal of Data Science and Analytics, Vol. 12, No. 4, pp. 297-331, Dec. 2021. DOI: 10.1007/s41060-021-00265-1

[3] X. Song, M. Wu, C. M. Jermaine, and S. Ranka, "Conditional Anomaly Detection," IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 5, pp. 631-645, May 2007. DOI: 10.1109/TKDE.2007.1009

[4] J. W. Tukey, "Exploratory Data Analysis," Addison-Wesley, pp. 1-688, 1977.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM), pp. 413-422, Pisa, Italy, Dec. 2008. DOI: 10.1109/ICDM.2008.17

[6] Y. R. Wang, J. Duan, D. Fox, and S. Srinivasa, "NEWTON: Are Large Language Models Capable of Physical Reasoning?," Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 9743-9758, Singapore, Dec. 2023. DOI: 10.18653/v1/2023.findings-emnlp.652

[7] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Transactions on Information Systems, Vol. 43, No. 2, Article No. 42, pp. 1-55, Jan. 2025. DOI: 10.1145/3703155

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, pp. 24824-24837, New Orleans, LA, USA, Dec. 2022.

[9] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger, "Identifying Mislabeled Data using the Area Under the Margin Ranking," Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 17044-17056, Virtual Conference, Dec. 2020.

[10] W. Guan, J. Cao, J. Gao, H. Zhao, and S. Qian, "DABL: Detecting Semantic Anomalies in Business Processes Using Large Language Models," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, No. 11, pp. 11735-11744, Philadelphia, PA, USA, April 2025. DOI: 10.1609/aaai.v39i11.33277

[11] S. Mishra, D. Stricker, and J. Rambach, "When Anomalies Depend on Context: Learning Conditional Compatibility for Anomaly Detection," arXiv preprint arXiv:2601.22868, Jan. 2026. URL: https://arxiv.org/abs/2601.22868

[12] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 12, pp. 12591-12604, Dec. 2023. DOI: 10.1109/TKDE.2023.3270293

[13] M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An Empirical Study of Example Forgetting during Deep Neural Network Learning," Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, May 2019.

[14] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9275-9293, Online, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.746

[15] C.-P. Tsai, G. Teng, P. Wallis, and W. Ding, "AnoLLM: Large Language Models for Tabular Anomaly Detection," Proceedings of the International Conference on Learning Representations (ICLR), Singapore, April 2025.

[16] T. Yang, Y. Nian, S. Li, R. Xu, Y. Li, J. Li, Z. Xiao, X. Hu, R. A. Rossi, K. Ding, X. Hu, and Y. Zhao, "AD-LLM: Benchmarking Large Language Models for Anomaly Detection," Findings of the Association for Computational Linguistics: ACL 2025, pp. 1524-1547, Vienna, Austria, July 2025. DOI: 10.18653/v1/2025.findings-acl.79

[17] S. Yoon, D. Kim, S. Yoon, Y. S. Sim, S. Yoa, H.-S. Cho, S. Lee, H. Lee, and W. Lim, "ReTabAD: A Benchmark for Restoring Semantic Context in Tabular Anomaly Detection," arXiv preprint arXiv:2510.02060, Oct. 2025. URL: https://arxiv.org/abs/2510.02060

[18] H. Thimonier, F. Popineau, A. Rimmel, and B.-L. Doan, "Retrieval Augmented Deep Anomaly Detection for Tabular Data," Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), pp. 2250-2259, Boise, ID, USA, Oct. 2024. DOI: 10.1145/3627673.3679559

[19] T. Wu, M. Terry, and C. J. Cai, "AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts," Proc. ACM CHI Conf. on Human Factors in Computing Systems, pp. 1-22, April 2022. DOI: 10.1145/3491102.3517582

[20] O. Khattab et al., "DSPy: Compiling declarative language model calls into self-improving pipelines," Proc. ICLR, May 2024. DOI: 10.48550/arXiv.2310.03714

[21] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik, "ABO: Dataset and Benchmarks for Real-World 3D Object Understanding," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), pp. 21126-21136, New Orleans, LA, USA, June 2022. DOI: 10.1109/CVPR52688.2022.02045

[22] R. Diaz and A. Marathe, "Soft labels for ordinal regression," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4738-4747, June 2019. DOI: 10.1109/CVPR.2019.00488

[23] B. Efron and R. J. Tibshirani, "An Introduction to the Bootstrap," Chapman & Hall/CRC, 1993.

Authors

Sharon Kim received the M.S. degree in AI Information Science from Myongji University, Korea, in 2025. She is currently pursuing the Ph.D. degree in AI Information Science at Myongji University, Seoul, Korea.

Her research interests include retrieval-augmented generation (RAG), synthetic data, ontology, and data quality.

Seok-Yong Yun received the M.S. degree in Information Industry from Soongsil University, Korea, in 2000, and the Ph.D. degree in IT Policy and Management from Soongsil University, Korea, in 2018.

He is currently a Professor in the Department of AI Information Science, Myongji University, Seoul, Korea. His research interests include data analysis, machine learning, artificial intelligence, natural language processing, LLM/sLLM, databases, data governance, and clinical decision support systems.