Enigma in Education
https://enigma.or.id/index.php/edu

e-ISSN: 3026-5827

ORIGINAL ARTICLE
Enigma in Education | Vol. 4, No. 2 (2026) | pp. 87–102

System Quality, Information Quality, and Technical Burden in Relation to Discontinuance Intention in GenAI-EdSent Digital Learning Applications: An Exploratory Rule-Based Review Analysis Guided by the DeLone–McLean Framework

Hesti Putri[1*], Matilda Munoz[2]

[1]*Department of Informatics Science, CMHC Research Center, Palembang, Indonesia*
[2]*Department of Physics, Camphece Science Center, Campeche, Mexico*

ARTICLE INFO

Article history
Received: August 20, 2026
Revised: September 15, 2026
Accepted: October 10, 2026
Available online: October 25, 2026
Published: November 1, 2026

Keywords
Digital learning; system quality; information quality; discontinuance intention; user reviews; GenAI-EdSent

Corresponding author
Hesti Putri
hesti.putri@gmail.com

ORCID
0009-0002-9568-3731

DOI
https://doi.org/10.61996/edu.v4i2.137

ABSTRACT

Background: Continued use of digital learning platforms is often explained by perceived experience quality. Public application store reviews, however, offer no individual-level retention data, so it is unclear how far stated intentions can stand in for actual retention.

Objective: To determine how often educational application reviews state explicit continuance or discontinuance intentions, and how these relate to complaints about system quality, information quality, and technical burden.

Methods: GenAI-EdSent was analyzed as a cross-sectional review corpus. Multi-stage filtering of 480,831 catalog rows from 22 educational applications left 204,294 unique English reviews (2021 to 13 August 2025). Lexical rules separated explicit discontinuance, explicit continuance, and unstated intentions before DeLone–McLean and Expectation-Confirmation coding, without human annotation. Exact conditional odds ratios (ORs), 95% exact confidence intervals (CIs), Fisher's exact tests, and Benjamini–Hochberg false discovery rate (q) adjustment were used.

Results: Only 220 reviews (0.108%) stated an explicit intention: 180 discontinuance and 40 continuance. System quality issues appeared in 10/180 (5.6%) versus 1/40 (2.5%) reviews (OR = 2.29; 95% exact CI: 0.31–102.08; p = 0.694), information quality issues in 13/180 (7.2%) versus 1/40 (2.5%) (OR = 3.02; 95% CI: 0.43–132.28; p = 0.474), and technical burden in 2/180 versus 0/40, an unbounded OR (95% CI: 0.04–∞; p = 1.000; all q = 1.00). Multivariable regression was withheld because only 34 continuance comparators existed across three eligible applications.

Conclusion: Large review catalogs do not guarantee enough explicit intention outcomes for adjusted modeling. The estimates are imprecise crude associations, not measures of retention.

Copyright: © 2026 The Author(s). Published by Enigma Institute on behalf of Enigma in Education. The authors retain copyright of their work.

Access license: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same license.

1. Introduction

Digital learning applications and artificial intelligence (AI)-assisted study companions have changed how secondary and higher education learners get homework support, conceptual explanations, and interactive quiz practice.[1–5] Mobile educational platforms now reach millions of downloads. Researchers and instructional designers now ask why learners keep or drop these tools.[6–10] Two theoretical traditions dominate the literature on persistence. The updated DeLone and McLean Information Systems (IS) Success Model holds that system quality, information quality, and service quality jointly shape user satisfaction and intention to use, and these in turn influence net educational benefits.[1] Bhattacherjee's Expectation-Confirmation Model (ECM) takes a different angle. It separates initial technology acceptance from post-adoption continuance intention and argues that continued use depends on how well actual performance confirms prior expectations and builds perceived usefulness and satisfaction.[2] Survey-based structural equation modeling studies in university and secondary school cohorts repeatedly find that stable system architecture, responsive interfaces, and accurate instructional content raise learner satisfaction and continuance intention.[11–18] The reverse holds too. Technical instability, hallucinated or misleading AI explanations, and intrusive resource burdens can wear down learner trust and lead to abandonment of the platform.[19–24]

The DeLone–McLean and ECM frameworks are theoretically clear. Moving their constructs from validated psychometric questionnaires to naturalistic user-generated text in public application marketplaces, though, raises serious methodological problems.[25,26] Public app store reviews, such as those compiled in the multi-application GenAI-EdSent repository, give an unobtrusive observational view of learner complaints and praise across commercial AI homework helpers, learning management systems, and study companions.[3,4] Unstructured store reviews still differ in kind from longitudinal cohort telemetry or standardized psychometric scales. A spontaneous complaint about

application crashes or incorrect math solutions in free text is not a calibrated latent score on a multi-item DeLone–McLean scale.[1,25,27,28] Public app store catalogs also supply no active user denominator, no individual session histories, and no behavioral follow-up after the review.[29,30] Some reviewers do write first-person statements that they intend to uninstall or stop using an application. A smaller fraction say they will keep using it. Most do neither. They offer general praise, brief frustration, or feature requests with no forward-looking continuance or discontinuance intention stated.[31–33]

Review mining invites two shortcuts. One treats overall catalog volume as statistical power for modeling behavioral outcomes. The other treats every review without an uninstall complaint as a continuing-user control group.[34–36] Both bring construct invalidity and selection bias. Hundreds of thousands of reviews with unstated intentions, counted as retained learners, merge silence about future use with active behavioral persistence. Comparator cell counts swell by orders of magnitude. Fitting multivariable regression or structural equation models to rule-coded thematic frequencies when explicit comparator outcomes are sparse also breaks minimum events-per-variable requirements and gives a false sense of adjusted precision.[37,38] Multi-application repositories add another problem. They combine platforms launched in different calendar years with very unequal review volumes. Pooled annual trends in discontinuance mentions can therefore shift simply because high-volume applications enter or dominate the catalog over time.[3,4]

This study takes up these methodological and empirical gaps through a rule-based exploratory analysis of the GenAI-EdSent digital learning review dataset.[3,4] The DeLone–McLean IS Success Model and Bhattacherjee's Expectation-Confirmation Model serve strictly as codebook organizing frameworks. We do not claim to test latent structural pathways. Three objectives guide the work.[1,2] The first is to quantify how many explicit first-person discontinuance and continuance intention statements remain after multi-stage corpus verification, temporal windowing, length filtering, language classification, and within-application text deduplication across the 480,831 initial catalog rows of GenAI-EdSent. The second is to estimate crude exact conditional associations of explicit discontinuance versus continuance intentions with three primary experiential complaint domains: system quality issues, information quality issues, and technical burden. Commercial paywall friction and positive quality mentions are handled separately as secondary descriptive constructs. The third objective is to judge whether the distribution of explicit intention outcomes across the 22 curated educational applications meets the pre-specified protocol feasibility thresholds for multivariable logistic regression. It also examines how sensitive annual discontinuance detection rates and crude odds ratios are to application composition and to alternative filtering specifications.

Public application store reviews call for a different analytical protocol from traditional psychometric surveys, and the reason becomes clear once we compare how constructs are measured in each paradigm.[6–15] In classical applications of the DeLone–McLean Information Systems Success Model and Bhattacherjee's Expectation-Confirmation Model, every sampled learner completes a structured multi-item questionnaire. It rates system reliability, information accuracy, perceived usefulness, expectation confirmation, satisfaction, and continuance intention on calibrated Likert scales.[1,2,10] Every participant answers every item, so the outcome denominator equals the full survey sample. Covariance-based or partial least squares structural equation modeling can then estimate simultaneous latent pathways across hundreds of complete observations.[16–21] A public review storefront such as Google Play or the Apple App Store works differently. It is an unprompted, self-selected channel. Learners write reviews at arbitrary points in their adoption lifecycle: right after downloading an application, after a frustrating software crash during homework, after meeting a sudden subscription paywall, or after preparing successfully for an examination.[27,31,32] A five-word review that omits any complaint about crashes or answer accuracy does not show that the learner met flawless system or information quality. It shows only that the learner did not mention those domains in that short text. The same logic applies when a learner writes a generic evaluative comment without saying whether they will keep or uninstall the application. Their future use is unobserved, not positive.[31,32,39]

Generative AI study companions spread rapidly between 2021 and 2025, and they changed what information quality and user burden mean in digital education.[3,4,27,40–46] In conventional learning management systems (such as Blackboard) or static flashcard platforms (such as Kahoot!), system quality concerns server uptime, login authentication, and page rendering. Information quality there depends on instructor-uploaded course materials.[47,48] Generative AI homework helpers and math solvers (such as Brainly: AI Homework Helper, Gauth: AI Study Companion, Question.AI, Answer.AI, Nerd AI, and Mathos AI) work differently. Learners rely on optical character recognition and large language models to parse photographed problem stems and produce step-by-step explanations in real time.[4,40,43,46] Suppose an AI solver hallucinates an algebraic derivation, misreads a chemical equation, or gives a confident but wrong final answer. The learner then experiences acute negative expectation disconfirmation about information quality.[2,27,43] Many commercial EdTech platforms also use aggressive freemium monetization, restricting daily AI explanations behind subscription paywalls, ad-viewing quotas, or virtual ticket systems.[32,49] Two questions follow. The first is whether stated discontinuance intentions in public reviews are associated with core DeLone–McLean technical domains (system quality, information quality, and technical resource burden) or with commercial paywall friction. The other is whether public review corpora hold enough explicit intention outcomes to support multivariable adjustment. Both must be settled for sound educational data science.[1–5]

2. Methods

2.1 Study design and data source

This study used a retrospective repeated cross-sectional design built on public application store reviews. It was not a longitudinal cohort. The data came from the publicly accessible GenAI-EdSent dataset, which compiles user reviews across 22 curator-selected educational and AI-assisted learning applications.[3,4] We analyzed sixteen primary release Parquet files holding 480,831 initial catalog rows across the 22 applications, after checking file integrity against the official repository checksums. Redundant application-level split copies were left out of the merge, so no review record was counted twice. In the repository structure, the primary release files make up the full consolidated review catalog and do not serve as a machine-learning partition in this study.

Five core variables were extracted from the raw records into the analytical matrix: application name, review identifier (ID), review body text, numerical star rating (1 to 5), and review timestamp date. Reviewer display names, profile photographs, developer contact metadata, and developer reply texts played no part in construct operationalization. They were dropped from the matrix at ingestion. Star ratings were screened for completeness and range validity (1–5). They were not used as a stand-in for explicit continuance or discontinuance intention, nor as an explanatory predictor, because treating a low rating as behavioral discontinuance would confound dissatisfaction with an actual intention to abandon the platform.[1,2] Review dates are calendar timestamps. Origin time zones of international storefronts are not verified in the source repository. Temporal analyses stayed at the calendar-year and date level. No hourly adjustment across countries was made.

2.2 Corpus selection, normalization, and deduplication

Corpus construction followed a pre-specified, multi-stage screening protocol (Table 1 and Figure 1). The observation window ran from 1 January 2021 through 31 December 2025. We began with 480,831 catalog rows spread over 16 Parquet files and 22 applications. A first structural check, which required non-empty text, a valid timestamp date, an integer star rating (1–5), and a review ID, removed 7 empty-text rows and left 480,824 valid records (99.999%). Deduplication used the application-and-review-ID pair and was set to keep the latest timestamp whenever an ID repeated; no repeated review IDs turned up anywhere in the catalog. Records outside the 1 January 2021–31 December 2025 window numbered 33,097 and were removed, which retained 447,727 reviews (93.115%). The final timestamps observed spanned 1 January 2021 to 13 August 2025.

Table 1. Multi-stage corpus formation, deduplication, and explicit intention eligibility flow of the GenAI-EdSent educational application review dataset.

Screening Stage	Remaining Reviews (n)	Proportion of Initial Catalog (%)	Exclusion Count and Substantive Selection Criterion
Initial catalog distribution rows	480,831	100.000%	16 primary data files covering 22 curated educational applications; separate application-specific splits were not merged to avoid double-counting.
Valid review text, date, rating (1–5), and review ID	480,824	99.999%	7 empty text records excluded; application–review ID deduplication verified with 0 repeated review IDs.
Within observation window (2021–2025)	447,727	93.115%	33,097 records outside the 1 January 2021 to 31 December 2025 observation window (observed final date 13 August 2025) excluded.
Minimum five alphabetic word tokens	230,084	47.851%	217,643 ultra-short reviews (<5 alphabetic tokens) excluded; operationalized as alphabetic word count rather than raw character length.
Detected as English language (langid 1.1.6 label en)	211,702	44.028%	18,382 non-en reviews excluded via automated language classification on up to the first 5,000 characters.
Unique normalized text within application (primary descriptive corpus)	204,294	42.488%	7,408 exact normalized text duplicates within the same application removed (NFKC, lowercase, whitespace unification, contraction expansion; earliest date retained).
Explicit discontinuance or continuance intention (primary association subset)	220	0.046% (0.108% of 204,294)	180 discontinuance intentions (0.088% of 204,294) and 40 continuance intentions (0.020% of 204,294); 204,074 unstated-intention reviews (99.892%) excluded from 2×2 association tests.
Candidate multivariable regression subset after application threshold	185	0.038% (0.091% of 204,294)	151 discontinuance and 34 continuance intentions across 3 applications with ≥5 reviews per outcome (Brainly, Gauth, Question.AI); below model feasibility threshold (minimum 100 and 110 for design rank 11).

Note: Counts represent individual public review records across filtering stages rather than unique longitudinal users or actual platform retention rates. The 204,294 deduplicated reviews form the primary descriptive denominator, whereas the 220 reviews with explicit first-person discontinuance or continuance statements form the denominator for crude 2×2 contingency associations. The last two rows give the proportion of the initial catalog and, in parentheses, of the 204,294-review primary corpus.

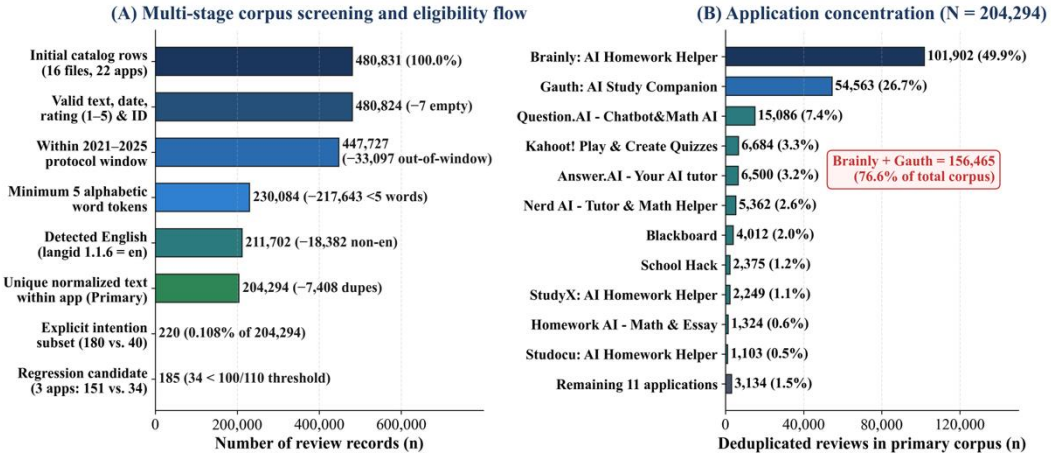


Figure 1. Multi-stage corpus screening cascade and application concentration across the GenAI-EdSent repository: (A) sequential filtering from 480,831 initial catalog rows across 16 Parquet files and 22 educational applications through valid text, date, star rating (1–5), and ID verification (480,824; 7 empty records removed), 2021–2025 protocol windowing (447,727), minimum five alphabetic words (230,084), automated English language detection via langid 1.1.6 (211,702), and within-application normalized text deduplication (204,294 unique reviews) to the 220 explicit intention reviews (180 discontinuance and 40 continuance) and 185 regression candidates (151 vs. 34 across 3 eligible applications); and (B) review concentration across the 22 applications showing that Brainly: AI Homework Helper (101,902; 49.9%) and Gauth: AI Study Companion (54,563; 26.7%) together comprise 156,465 reviews (76.6% of the primary corpus). Counts reflect individual reviews rather than unique longitudinal users or actual retention rates.

Rule-based syntactic parsing needs enough lexical context, so a review had to contain at least five alphabetic word tokens; a raw character count threshold was not used. This filter excluded 217,643 ultra-short reviews and retained 230,084 records (47.851%). Language identification came next, using langid version 1.1.6 on at most the first 5,000 characters of each review. Only English-labeled records (en) were kept. That step excluded 18,382 non-English records and left 211,702 English-

labeled reviews (44.028%). Automated detection is a machine heuristic. It can misclassify brief or code-switched strings. Its effect at the boundary was therefore documented explicitly. Text was then normalized through Unicode Normalization Form Compatibility Composition (NFKC), lowercase conversion, whitespace unification, and standardized expansion of common English contractions. When identical normalized texts appeared within one application, only the earliest observed timestamp was kept. This removed 7,408 within-application duplicate text copies and established the primary descriptive corpus of 204,294 unique reviews (42.488% of the initial catalog). Public review IDs cannot link separate reviews written by one person over time or across accounts, and exact normalization does not merge near-identical semantic paraphrases. Keeping the within-application duplicate texts (n = 211,702) was therefore planned in advance as a sensitivity analysis.

2.3 Operationalization of intention outcomes and experiential domains

The primary binary outcome contrasted two kinds of first-person statement. The first was an explicit plan to uninstall, delete, or stop using the application (explicit discontinuance intention). The second was an explicit statement of continuing to use, keeping, or refusing to uninstall it (explicit continuance intention; Table 2). Declarations of no longer using the application, or of being finished with it, counted as discontinuance intention. A narrower, prospective-only definition of discontinuance was evaluated in sensitivity testing. Conditional statements, such as an intention to uninstall if a defect persists, were kept as stated intentions, although stated intention does not guarantee later behavior. First-person reports of having already uninstalled the application were tracked as an auxiliary, non-exclusive label (n = 122) that could overlap with the primary categories. Reviews stating neither intention (n = 204,074) formed a separate unstated-intention category. They were never treated as continuing-user controls. Generic praise or recommendations to other learners did not automatically qualify as explicit first-person continuance intention.

Table 2. Operationalization of rule-based text proxies, target lexical evidence, and construct interpretation boundaries guided by the DeLone–McLean and Expectation-Confirmation frameworks.

Construct / Coding Domain	Role in Analytical Framework	Target Textual Evidence in Public Reviews	Construct Interpretation Boundary and Methodological Limitation
System quality issues	Primary explanatory proxy	Application crashes, software bugs, screen freezing, lag, login or access failures, and non-functioning system features.	Captures user-reported technical malfunctions in free text; does not measure objective server uptime or telemetry reliability.
Information quality issues	Primary explanatory proxy	Incorrect homework answers or solutions, misleading AI-generated outputs, and unclear or confusing explanations.	Captures perceived output inaccuracy or instructional unclarity; does not constitute expert pedagogical verification of solution correctness.
Technical burden	Primary explanatory proxy	Intrusive advertisements, excessive battery, mobile data, or memory consumption, and device overheating.	Strictly restricted to technical/resource and ad-interruption friction; monetary subscription costs and paywalls are separated.
Commercial burden	Secondary descriptive proxy	Paywalls, mandatory subscriptions, unexpected billing charges, refund disputes, high pricing, and ticket/coin access barriers.	Reported descriptively as a secondary monetization construct; not conflated with the primary technical burden domain.
Positive quality mentions (system & information)	Secondary descriptive proxy	Explicit praise for ease of use, system reliability, or accurate and clear instructional explanations.	Summarized descriptively; does not represent a validated psychometric latent satisfaction or quality scale score.
Explicit intention outcome	Primary binary comparison outcome	First-person statements of planning to uninstall or stop using (discontinuance) versus explicitly continuing or keeping the application (continuance).	Unstated intention (n = 204,074) remains a separate category and is never treated as a continuing-user control group; prior uninstall reports (n = 122) are tracked as an auxiliary tag.

Note: Contextual rule-based classification removes outcome-defining intention phrases from input strings prior to coding experience domains to prevent direct lexical leakage, and applies local negation checks (e.g., rejecting 'no bugs' or 'not too many ads'). Experience domains are non-mutually exclusive and are not summed into a composite score.

Classification relied on contextual lexical pattern-matching rules. These covered both the intention outcomes and the experiential domains, and they were guided by the DeLone–McLean IS Success Model and Bhattacherjee's ECM.[1,2] Outcome detection and domain coding could otherwise leak into each other. To prevent this, the exact phrase spans that triggered an intention classification were removed from the input string before it was scanned for experiential domain proxies. Three primary explanatory complaint domains were operationalized. (1) System quality issues covered crashes, software bugs, screen freezing, lag, login or access failures, and non-functioning system features. (2) Information quality issues covered incorrect homework answers or solutions, misleading AI outputs, and unclear or confusing explanations. (3) Technical burden covered only intrusive advertisements, excessive battery, mobile data, or memory consumption, and device overheating (Table 2). Commercial burden, meaning paywalls, mandatory subscriptions, billing charges, refund disputes, and high pricing, was kept as a separate secondary descriptive domain. Merging it into technical burden would have confounded the constructs of technical resource friction and monetization policy. Secondary positive mentions of system quality (ease of use, reliability) and information quality (accurate or clear answers) were also recorded descriptively. Local negation rules rejected simple negated phrases such as 'no bugs' or 'not too many ads'. Complex sarcasm, quotations, cross-app comparisons, and non-standard spelling remain beyond what automated lexical rules can fully resolve. A single review may voice complaints in several domains, so the domain indicators are non-mutually exclusive and were not summed into an arbitrary composite score.

2.4 Statistical analysis, feasibility stopping rules, and uncertainty

All descriptive proportions for the corpus used the 204,294 unique reviews as the denominator. Crude association analyses used a smaller denominator. This was the 220 reviews with an explicit discontinuance (n = 180) or continuance (n = 40) intention (Table 3). For each primary domain, a 2×2 contingency table cross-classified explicit intention status (discontinuance versus continuance) by domain mention status (present versus

absent). Exact conditional odds ratios (ORs) with 95% exact confidence intervals (CIs) came from the conditional maximum likelihood estimator in scipy.stats.contingency.odds_ratio(..., kind='conditional'). Two-sided exact p-values came from Fisher's exact test.[5] Zero cells were kept as observed. No artificial 0.5 Haldane–Anscombe continuity correction was added to force finite point estimates. The Benjamini–Hochberg false discovery rate (FDR) procedure (q-values) controlled multiplicity across the three primary hypothesis tests.

The initial protocol planned a multivariable logistic regression model to estimate adjusted associations. Predictors: the three primary domains, log-transformed review word length, application fixed effects, and calendar-year fixed effects. Small-sample sparse-data bias and overfitting were to be avoided through three mandatory feasibility criteria checked before execution: (1) each included application needed at least five reviews in each explicit intention outcome category (≥5 discontinuance and ≥5 continuance); (2) the eligible subset needed at least 100 reviews in each intention outcome overall; and (3) the sample needed at least 10 minority-outcome events per design-matrix rank parameter. Screening showed that only three applications met the ≥5 per-outcome criterion. They yielded 185 reviews (151 discontinuance and 34 continuance) across a design rank of 11. The 34 minority comparators fell far below both the absolute minimum of 100 and the rank-based minimum of 110. The stopping rule halted multivariable regression and cluster-bootstrap adjusted coefficient estimation automatically. No fabricated adjusted odds ratios or ad-hoc structural equation models were substituted for the halted model. Crude exact conditional estimates are reported instead. They are exploratory.

Five pre-specified sensitivity analyses tested whether the crude associations held under other corpus definitions: (1) retaining within-application duplicate normalized texts (n = 211,702); (2) applying a narrower prospective-only discontinuance intention definition (n = 204,264); (3) raising the minimum length filter to at least ten alphabetic words (n = 120,966); (4) restricting the observation window to complete calendar years 2021–2024 (n = 181,070); and (5) running a leave-one-application-out (LOAO) jackknife across all 22 applications. Temporal trajectories were compared separately of explicit discontinuance detection per 1,000 reviews from 2021 to 2025. The pooled catalog series was set against an equal-weighted standardized series. That series drew on the five applications with at least 20 reviews in every calendar year from 2021 through 2025 (Blackboard, Course Hero: AI Homework Help, Gauth: AI Study Companion, Kahoot! Play & Create Quizzes, and Studocu: AI Homework Helper). Data for 2025 are partial (1 January to 13 August 2025).

2.5 Computational verification, reproducibility, and research ethics

Before final tabulation, ten adversarial codebook test cases and automated verification checks audited arithmetic identities, denominator consistency, review ID uniqueness, timestamp boundaries, and exact contingency reproducibility. A human audit framework with blank rater columns is ready. Independent human coding has not yet been done. No Cohen's κ, precision, recall, or human-validated accuracy metrics are therefore claimed, and every rule-based classification remains provisional. The GenAI-EdSent dataset is distributed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The MIT license governs the code scripts separately.[3] In derived feature tables, public review identifiers were replaced by local cryptographic hashes. Raw review texts were left out of shared summary outputs. No user text was sent to external commercial large language model APIs. The computational environment comprised Python 3.12, pandas 2.2.3, numpy 2.3.5, pyarrow 26.0.0, langid 1.1.6, scipy 1.18.1, and statsmodels 0.15.0.

2.6 Formal specification of exact conditional inference, feasibility gates, and direct standardization

The estimators implemented in the computational pipeline for the contingency and temporal analyses are defined formally below. Let the primary explicit-intention analytical subset of size $N_{int} = 220$ be partitioned into $n_1 = 180$ reviews expressing explicit discontinuance intention and $n_0 = 40$ reviews expressing explicit continuance intention. For each binary experiential domain indicator d, let a be the number of discontinuance reviews mentioning domain d. Let b be the number of continuance reviews mentioning it. Let $c = n_1 - a$ be the number of discontinuance reviews not mentioning domain d, and $f = n_0 - b$ the number of continuance reviews not mentioning it. Conditional on the fixed row and column margins ($n_1 = 180$, $n_0 = 40$, $m_1 = a + b$, and $m_0 = c + f$), the random variable A (the count in the upper-left cell) follows Fisher's non-central hypergeometric distribution. Its conditional probability mass function is $P(A = k \mid m_1, n_1, n_0, \psi)$ proportional to $C(n_1, k) \times C(n_0, m_1 - k) \times \psi^k$, where ψ is the true conditional odds ratio and C(n, k) is the binomial coefficient. The conditional maximum likelihood estimator (CMLE) in scipy.stats.contingency.odds_ratio(..., kind='conditional') solves for the value $\hat{\psi}$ at which the observed cell count a equals its conditional expectation $E[A \mid m_1, n_1, n_0, \hat{\psi}]$.[5] The unconditional sample cross-product ratio (a × f) / (b × c) depends on asymptotic normal approximations of log(OR), with standard error sqrt(1/a + 1/b + 1/c + 1/f). The exact conditional estimator differs. Its Clopper–Pearson-type exact 95% confidence bounds $[\psi_L, \psi_U]$ stay valid under extreme cell sparsity (such as b = 1 for system and information quality issues). When b = 0 (technical burden: a = 2, b = 0, c = 178, f = 40), the estimator returns $\hat{\psi} = \infty$ with a well-defined exact lower bound ($\psi_L = 0.0414$), and no artificial +0.5 Haldane–Anscombe continuity bias is introduced.

For the planned multivariable logistic regression model, let $Y_{ij} \in \{0, 1\}$ denote explicit discontinuance (1) versus continuance (0) intention for review i within eligible application j in calendar year t. The pre-specified linear predictor was $\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_{sys} X_{sys,ij} + \beta_{inf} X_{inf,ij} + \beta_{tech} X_{tech,ij} + \beta_{len} \log(W_{ij}) + \alpha_j + \gamma_t$. Here X_{sys}, X_{inf}, and X_{tech} are binary indicators for the three primary domains, and $\log(W_{ij})$ is the natural logarithm of alphabetic word count. The term α_j carries application fixed-effect contrasts across the $J_{elig} = 3$ applications that met the initial cell criterion of at least five reviews per intention outcome (Brainly, Gauth, and Question.AI), contributing $J_{elig} - 1 = 2$ degrees of freedom. The term γ_t carries calendar-year fixed-effect contrasts across the 5 observation years 2021–2025 (contributing 5 − 1 = 4 degrees of freedom). The design matrix rank is the sum of the intercept (1), three domain slopes (3), one length covariate (1), two application contrasts (2), and four year contrasts (4), giving K = 11. The protocol feasibility rules required (i) ≥5 outcomes per cell per included application, (ii) ≥100 total observations in each outcome category, and (iii) ≥10 minority outcome events per design rank parameter (at least 10 × 11 = 110 continuance comparators). The eligible subset of n = 185 reviews (151 discontinuance and 34 continuance) offered only 34 / 11 = 3.09 minority events per parameter. The model was terminated.

The temporal analysis spanned calendar years $t \in \{2021, 2022, 2023, 2024, 2025^*\}$. Let N_{jt} denote the number of deduplicated reviews for application j in year t, and D_{jt} the number of explicit discontinuance intentions detected in application j in year t. The pooled annual rate per 1,000 reviews is $R_{pooled}(t) = 1{,}000 \times (\sum_{j=1}^{22} D_{jt}) / (\sum_{j=1}^{22} N_{jt}) = \sum_{j=1}^{22} w_{jt} \times r_{jt}$. Here $r_{jt} = 1{,}000 \times (D_{jt} / N_{jt})$ is the application-specific rate, and $w_{jt} = N_{jt} / \sum_k N_{kt}$ is the time-varying review-volume weight of application j in year t. The weight w_{jt} shifts sharply when a high-volume application enters the catalog (e.g., Brainly contributing 101,902 reviews starting in January 2022). $R_{pooled}(t)$ then mixes within-application temporal change with shifts in cross-application composition. A composition-invariant benchmark was needed.

Let S_5 denote the fixed set of five applications with at least 20 deduplicated reviews in every calendar year from 2021 through 2025 (Blackboard, Course Hero: AI Homework Help, Gauth: AI Study Companion, Kahoot! Play & Create Quizzes, and Studocu: AI Homework Helper). The equal-weighted five-application standardized rate per 1,000 reviews is $R_{std5}(t) = (1 / 5) \times \sum_{j \in S5} [1{,}000 \times (D_{jt} / N_{jt})]$. Each continuously observed platform receives a constant weight of 0.20, whatever its annual review volume.

3. Results

3.1 Corpus curation flow, intention sparsity, and application concentration

The catalog held 480,831 initial rows, drawn from 16 Parquet files that covered 22 educational applications. Sequential quality and eligibility filters then narrowed the analytical sample (Table 1 and Figure 1A). Removing 7 empty text records left 480,824 valid reviews, each with a verified date, a star rating (1–5), and a unique application–ID pair. The 2021–2025 protocol window kept 447,727 reviews and excluded 33,097 records outside it. A minimum of five alphabetic word tokens then kept 230,084 reviews, with 217,643 short entries dropped. Automated English language filtering with langid 1.1.6 retained 211,702 English-labeled records and excluded 18,382 non-en records. Finally, exact deduplication of normalized text within each application removed 7,408 duplicate copies. This yielded the primary descriptive corpus of 204,294 unique reviews.

Explicit first-person intention statements were sparse in the primary descriptive corpus of 204,294 deduplicated reviews (Table 3 and Figure 2A). Only 220 reviews (0.108% of 204,294) stated an explicit intention eligible for binary comparative analysis. Of these, 180 were explicit discontinuance intentions (0.088% of 204,294; 81.8% of the 220 explicit intention reviews) and 40 were explicit continuance intentions (0.020% of 204,294; 18.2% of the 220 explicit intention reviews). The other 204,074 reviews (99.892%) held no explicit first-person continuance or discontinuance statement under the codebook rules. No review (0.000%) was classified as mixed intention. A separate group of 122 reviews (0.060% of 204,294) matched the auxiliary label for reviewers who reported having already uninstalled the application before or during review composition.

Table 3. Distribution of stated intention categories and auxiliary uninstall reports in the primary deduplicated corpus (N = 204,294).

Intention Classification Category	Review Count (n)	Percentage of Primary Corpus (% of 204,294)	Analytical Treatment and Substantive Interpretation
Explicit discontinuance intention	180	0.088%	Included in primary 2×2 association denominator (81.8% of the 220 explicit intention reviews).
Explicit continuance intention	40	0.020%	Included in primary 2×2 association denominator as the active explicit comparator group (18.2% of 220).
Mixed intention (both discontinuance and continuance)	0	0.000%	Zero reviews simultaneously triggered both opposing intention rules under the primary codebook.
Unstated intention	204,074	99.892%	Retained in descriptive corpus denominator; excluded from 2×2 association tests to avoid artificial control inflation.
Total primary mutually exclusive categories	204,294	100.000%	Complete deduplicated English review corpus (≥5 alphabetic words) across 22 educational applications.
Subtotal: Explicit intention analytical subset (discontinuance + continuance)	220	0.108%	Primary analytical sample for exact conditional odds ratios and Fisher's exact tests (180 vs. 40).
Auxiliary label: Past application uninstall report	122	0.060%	Non-exclusive auxiliary tag capturing statements of having already uninstalled the app; may overlap with primary categories.

Note: Percentages in the four primary mutually exclusive categories sum to exactly 100.000% (0.088% + 0.020% + 0.000% + 99.892% = 100.000%).

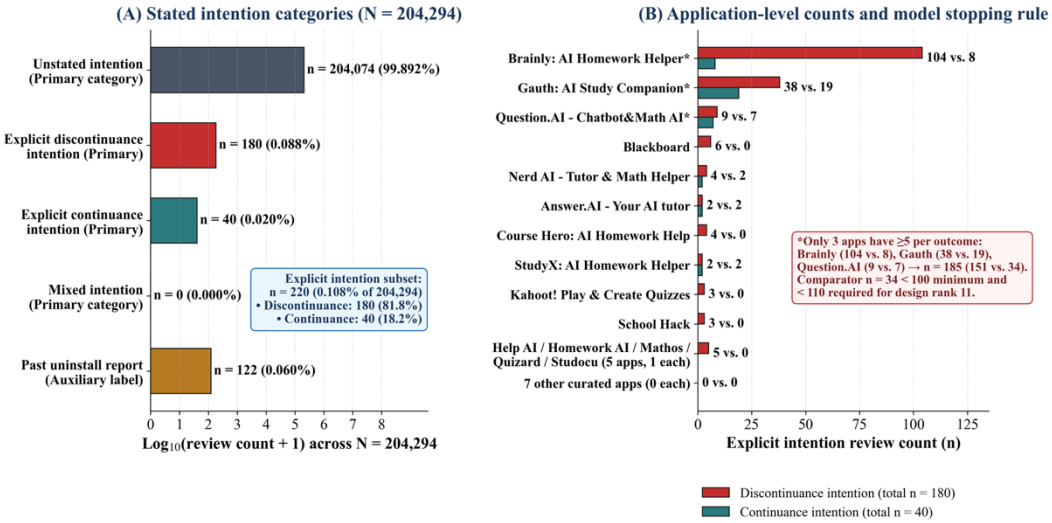


Figure 2. Stated intention category breakdown and application-level explicit outcome distribution: (A) distribution across the 204,294 unique English reviews showing 204,074 unstated-intention reviews (99.892%), 180 explicit discontinuance intentions (0.088%), 40 explicit continuance intentions (0.020%), and 0 mixed-intention reviews (0.000%), alongside 122 auxiliary past-uninstall reports (0.060%); and (B) application-level explicit discontinuance versus continuance counts across the 22 applications, illustrating why only three applications (Brainly: 104 vs. 8; Gauth: 38 vs. 19; Question.AI: 9 vs. 7; combined n = 185

with 151 discontinuance and 34 continuance intentions) met the ≥5 per-outcome threshold and why the 34 continuance comparators fell below the protocol regression feasibility thresholds (100 minimum and 110 for design rank 11).

3.2 Experiential domain mentions and crude exact conditional associations

Among the 220 reviews with an explicit intention statement (180 discontinuance versus 40 continuance), rule-coded mentions of the three primary experiential domains were modest in absolute frequency (Table 4 and Figure 3A). System quality issues, such as crashes, bugs, freezing, lag, or login failures, appeared in 10 of 180 discontinuance reviews (5.6%) and 1 of 40 continuance reviews (2.5%). That made 11 of 220 explicit intention reviews overall (5.0%). Information quality issues covered incorrect solutions, misleading outputs, or unclear explanations. They appeared in 13 of 180 discontinuance reviews (7.2%) and 1 of 40 continuance reviews (2.5%), or 14 of 220 reviews overall (6.4%). Technical burden was defined strictly as intrusive advertisements, excessive battery/data/memory drain, or device overheating. Only 2 of 180 discontinuance reviews (1.1%) and 0 of 40 continuance reviews (0.0%) mentioned it. Among the secondary descriptive proxies, commercial burden (paywalls, subscriptions, billing charges, or refund complaints) was more common than any primary technical domain. It occurred in 23 of 180 discontinuance reviews (12.8%) and 2 of 40 continuance reviews (5.0%; 25/220 [11.4%] overall). Secondary positive mentions of system quality numbered 1/180 (0.6%) among discontinuance reviews and 1/40 (2.5%) among continuance reviews. Positive mentions of information quality numbered 2/180 (1.1%) and 2/40 (5.0%), respectively (Table 4).

Table 5 and Figure 3B report crude exact conditional odds ratios, 95% exact confidence intervals, two-sided Fisher's exact p-values, and Benjamini–Hochberg FDR-adjusted q-values for the 220 explicit intention reviews. The odds were oriented as explicit discontinuance versus explicit continuance, comparing reviews that mentioned a domain with those that did not. System quality issues gave a conditional exact OR of 2.29 (95% exact CI: 0.31–102.08; two-sided Fisher's exact p = 0.694; Benjamini–Hochberg q = 1.00). Information quality issues gave a conditional exact OR of 3.02 (95% exact CI: 0.43–132.28; Fisher's exact p = 0.474; q = 1.00). Both 95% exact confidence intervals spanned more than two orders of magnitude and crossed the null value of 1.00. Point estimates above unity therefore offer no precise statistical evidence of an association. Technical burden (2/180 versus 0/40) had a zero cell in the continuance comparator group. The conditional exact OR was unbounded at infinity (∞), with a 95% exact confidence interval from an exact lower limit of 0.0414 (0.04) to infinity (Fisher's exact p = 1.000; q = 1.00). Such an unbounded estimate reflects sparse-cell degeneracy, not a large or certain effect. All three primary tests returned Benjamini–Hochberg FDR q-values of 1.00. These values show no statistically precise differentiation, yet they do not prove equivalence or the absence of a true underlying relationship.

Table 4. Frequency and proportion of primary and secondary text proxy mentions stratified by explicit intention category (n = 220).

Text Proxy Domain	Domain Tier	Discontinuance Intention (n = 180)	Continuance Intention (n = 40)	Total Mentions in Explicit Subset (n = 220)
System quality issues	Primary explanatory domain	10 / 180 (5.6%)	1 / 40 (2.5%)	11 / 220 (5.0%)
Information quality issues	Primary explanatory domain	13 / 180 (7.2%)	1 / 40 (2.5%)	14 / 220 (6.4%)
Technical burden	Primary explanatory domain	2 / 180 (1.1%)	0 / 40 (0.0%)	2 / 220 (0.9%)
Commercial burden (paywall / subscription / billing)	Secondary descriptive domain	23 / 180 (12.8%)	2 / 40 (5.0%)	25 / 220 (11.4%)
Positive system quality mentions	Secondary descriptive domain	1 / 180 (0.6%)	1 / 40 (2.5%)	2 / 220 (0.9%)
Positive information quality mentions	Secondary descriptive domain	2 / 180 (1.1%)	2 / 40 (5.0%)	4 / 220 (1.8%)

Note: Text proxy domains are not mutually exclusive because a single review can mention more than one experiential domain. Percentages are calculated within each explicit intention column (out of n = 180 discontinuance reviews and n = 40 continuance reviews).

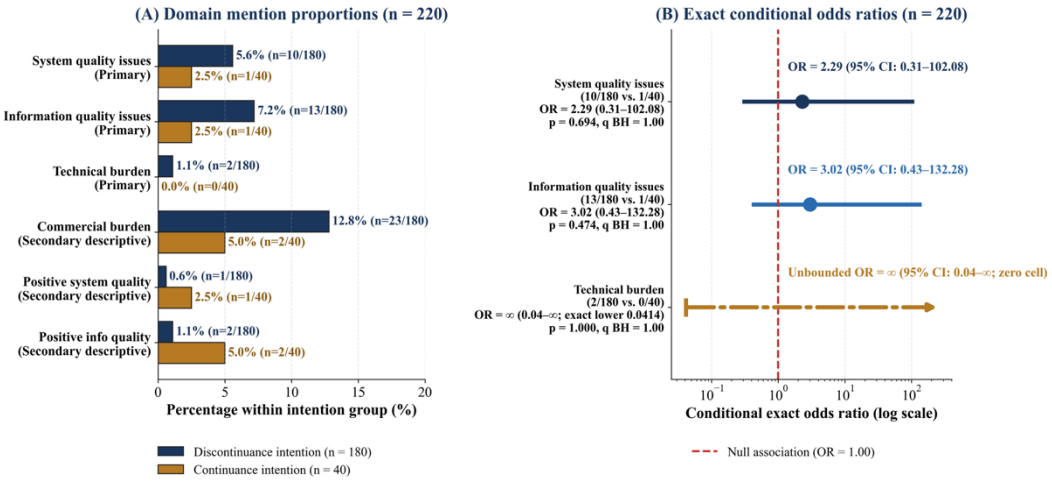


Figure 3. Experiential domain mentions and crude exact conditional odds ratios among the 220 reviews with explicit intention (180 discontinuance vs. 40 continuance): (A) proportion of reviews mentioning primary explanatory domains—system quality issues (10/180 [5.6%] vs. 1/40 [2.5%]), information quality issues (13/180 [7.2%] vs. 1/40 [2.5%]), and technical burden (2/180 [1.1%] vs. 0/40 [0.0%])—and secondary descriptive domains—commercial burden (23/180 [12.8%] vs. 2/40 [5.0%]), positive system quality (1/180 [0.6%] vs. 1/40 [2.5%]), and positive information quality (2/180 [1.1%] vs. 2/40 [5.0%]); and (B) log-scale forest plot of conditional exact odds ratios and 95% exact confidence intervals for system quality issues (OR = 2.29; 95% CI: 0.31–102.08; Fisher's exact p = 0.694; Benjamini–Hochberg q = 1.00), information quality issues (OR = 3.02; 95% CI: 0.43–132.28; p = 0.474; q = 1.00), and technical burden (unbounded OR = ∞; 95% CI: 0.04–∞ [exact lower limit 0.0414]; p = 1.000; q = 1.00).

Table 5. Crude exact conditional odds ratios, 95% exact confidence intervals, two-sided Fisher's exact test p-values, and Benjamini–Hochberg q-values among the 220 reviews with explicit intention.

Experiential Domain	2×2 Cell Counts (Discontinuance vs. Continuance: Present | Absent)	Conditional Exact OR	95% Exact Confidence Interval	Two-Sided Fisher's Exact p	Benjamini–Hochberg FDR q
System quality issues (Primary)	10 vs. 1 | 170 vs. 39	2.29	0.31–102.08	0.694	1.00
Information quality issues (Primary)	13 vs. 1 | 167 vs. 39	3.02	0.43–132.28	0.474	1.00
Technical burden (Primary)	2 vs. 0 | 178 vs. 40	Unbounded (∞)	0.04–∞ (exact lower: 0.0414)	1.000	1.00

Note: Odds ratios are oriented as the conditional odds of explicit discontinuance intention (n = 180) versus explicit continuance intention (n = 40) in reviews mentioning the domain relative to reviews not mentioning the domain, computed via conditional maximum likelihood estimation. Zero cells are preserved without adding an artificial 0.5 continuity correction. Benjamini–Hochberg FDR adjustment is applied across the three primary hypothesis tests. Planned multivariable logistic regression was not executed because only three applications met the ≥5 per-outcome threshold (151 discontinuance vs. 34 continuance; design rank = 11). OR = odds ratio; CI = confidence interval; q = Benjamini–Hochberg false discovery rate-adjusted p-value.

3.3 Annual trajectories and application composition sensitivity

Annual discontinuance intentions per 1,000 deduplicated reviews from 2021 to 2025 diverged between pooled catalog rates and composition-standardized rates (Table 6 and Figure 4). Across the full corpus, annual deduplicated review volumes were 8,655 in 2021 (13 discontinuance intentions), 50,173 in 2022 (52 discontinuance intentions), 48,383 in 2023 (54 discontinuance intentions), and 73,859 in 2024 (45 discontinuance intentions). The partial 2025 period, from 1 January to 13 August 2025, held 23,224 reviews (16 discontinuance intentions). Pooling all available applications within each year, the detected discontinuance rate fell by more than half across the complete calendar years. It was 1.50 per 1,000 reviews in 2021, 1.04 per 1,000 in 2022, 1.12 per 1,000 in 2023, and 0.61 per 1,000 in 2024. The partial year 2025 showed 0.69 per 1,000.

The picture changed when the annual trajectory was standardized with equal weights (1/5 each) across five applications. These were the applications with at least 20 deduplicated reviews in every calendar year from 2021 through 2025: Blackboard, Course Hero: AI Homework Help, Gauth: AI Study Companion, Kahoot! Play & Create Quizzes, and Studocu: AI Homework Helper. Between 2021 and 2024 the temporal direction reversed (Table 6 and Figure 4A). In this equal-weighted five-application series, discontinuance detections went from 2.13 per 1,000 reviews in 2021 to 1.59 per 1,000 in 2022. They then rose to 2.06 per 1,000 in 2023 and to 3.39 per 1,000 in 2024. The partial 2025 window recorded 1.49 per 1,000. The opposing trajectory suggests that the apparent drop in the pooled catalog rate, from 1.50 to 0.61 per 1,000 between 2021 and 2024, is consistent with a change in application composition. Brainly and newer AI tutor applications brought a large influx of reviews, which may have lowered the pooled per-review rate of explicit intention detection. The drop therefore need not reflect a true longitudinal reduction in user discontinuance propensity. These rates are descriptive.

Table 6. Annual trajectory of detected explicit discontinuance intentions per 1,000 reviews comparing the pooled available catalog against the equal-weighted five-application standardized series (2021–2025).

Calendar Year	All Deduplicated Reviews (n)	Explicit Discontinuance Intentions (n)	Pooled Rate per 1,000 Reviews	Equal-Weighted 5-App Standardized Rate per 1,000 Reviews	Temporal Coverage and Composition Note
2021	8,655	13	1.50	2.13	Full calendar year; baseline before high-volume Brainly entries (Brainly begins 16 January 2022).
2022	50,173	52	1.04	1.59	Full calendar year; rapid catalog expansion following Brainly inclusion.
2023	48,383	54	1.12	2.06	Full calendar year; multiple generative AI study companion apps enter the catalog.
2024	73,859	45	0.61	3.39	Full calendar year; pooled rate drops to 0.61 whereas 5-app standardized rate rises to 3.39 per 1,000.
2025 partial (1 Jan–13 Aug)	23,224	16	0.69	1.49	Partial calendar year through 13 August 2025; not directly comparable to full-year seasonal totals.

Note: Across 2021–2025, total deduplicated reviews sum to 204,294 (8,655 + 50,173 + 48,383 + 73,859 + 23,224 = 204,294) and explicit discontinuance intentions sum to 180 (13 + 52 + 54 + 45 + 16 = 180). The equal-weighted standardized series assigns equal weight (1/5 each) to the five applications with at least 20 reviews in every calendar year from 2021 to 2025: Blackboard, Course Hero: AI Homework Help, Gauth: AI Study Companion, Kahoot! Play & Create Quizzes, and Studocu: AI Homework Helper.

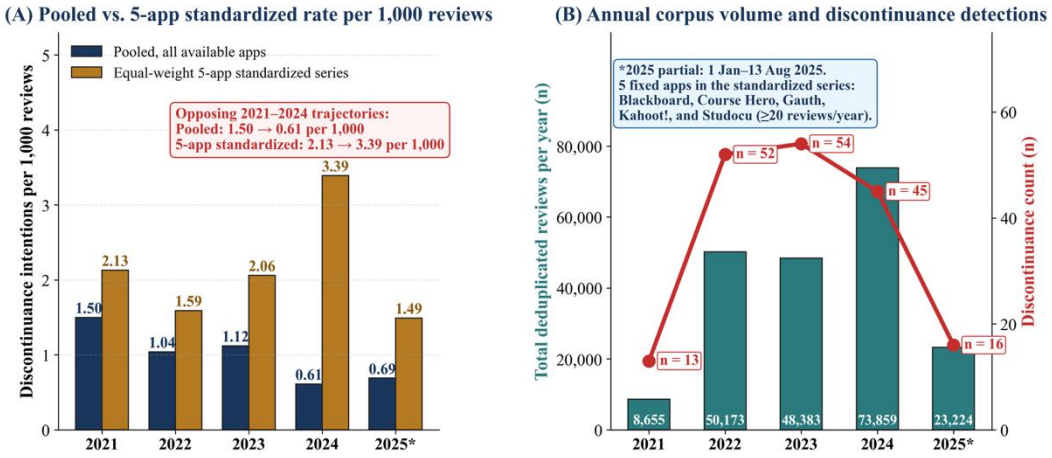


Figure 4. Annual trajectory of detected explicit discontinuance intentions per 1,000 reviews and catalog composition sensitivity from 2021 to 2025*: (A) comparison between the pooled rate across all available applications (declining from 1.50 per 1,000 in 2021 to 1.04 in 2022, 1.12 in 2023, and 0.61 in 2024, with 0.69 in partial year 2025*) and the equal-weighted five-application standardized rate across Blackboard, Course Hero, Gauth, Kahoot!, and Studocu (moving from 2.13 per 1,000 in 2021 to 1.59 in 2022, 2.06 in 2023, and rising to 3.39 per 1,000 in 2024, with 1.49 in partial year 2025*); and (B) annual deduplicated review volumes (8,655 in 2021; 50,173 in 2022; 48,383 in 2023; 73,859 in 2024; 23,224 in 2025*) and explicit discontinuance counts (13, 52, 54, 45, and 16; total n = 180). *Indicates partial year 1 January to 13 August 2025.

3.4 Sensitivity analyses and application-level coverage diagnostics

Sensitivity analyses across five alternative corpus and coding specifications confirmed that the imprecise crude odds ratios, and the shortage of comparator counts for multivariable modeling, were structural features of the dataset and not artifacts of one filtering choice (Table 7 and Figure 5). First, retaining within-application duplicate normalized texts raised the total corpus from 204,294 to 211,702 reviews (+7,408 duplicate texts). The explicit intention counts did not change, staying at 180 discontinuance and 40 continuance reviews, so the conditional exact ORs stayed identical at 2.29 for system quality issues and 3.02 for information quality issues. Second, a narrower prospective discontinuance definition for the outcome trimmed the corpus slightly to 204,264 reviews and cut explicit discontinuance intentions from 180 to 150, with continuance intentions held at 40. The conditional exact ORs stayed very imprecise at 2.19 for system quality issues and 3.07 for information quality issues. Third, a stricter minimum length of ten alphabetic words reduced the descriptive corpus by 40.8% to 120,966 reviews. It kept 177 discontinuance and 39 continuance reviews (98.2% of explicit intention cases), giving conditional exact ORs of 2.27 for system quality and 3.00 for information quality. Fourth, limiting the analysis to complete calendar years 2021–2024 (n = 181,070 reviews) left 164 discontinuance and 32 continuance intentions, with conditional exact ORs of 1.80 for system quality and 2.44 for information quality (Table 7 and Figure 5A). In every specification, technical burden kept a zero continuance comparator cell and an unbounded OR.

The leave-one-application-out (LOAO) jackknife across the 22 applications gave finite conditional exact OR point estimates from 1.19 to 2.64 for system quality issues and from 1.51 to 3.14 for information quality issues (Table 7 and Figure 5B). Only one continuance review (1/40) in the whole corpus mentioned a system quality issue, and only one (1/40) mentioned an information quality issue. Omitting the application that supplied that single review brought the comparator cell to zero. The corresponding domain OR then became unbounded (∞). Point estimates stayed above 1.00 in all finite LOAO iterations. Even so, the underlying contingency tables were extremely fragile, resting on single cells, and they did not support multivariable adjustment under any sensitivity specification.

Table 7. Sensitivity analysis of corpus size, explicit intention counts, and crude conditional odds ratios across alternative filtering and leave-one-application-out specifications.

Sensitivity Specification	Evaluated Corpus Size (n)	Explicit Discontinuance / Continuance (n / n)	Conditional OR: System Quality Issues	Conditional OR: Information Quality Issues	Conditional OR: Technical Burden & Model Feasibility
Primary analysis (unique normalized text within app)	204,294	180 / 40	2.29	3.02	Unbounded (2 vs. 0); multivariable threshold not met (34 comparators).
Within-application duplicate normalized texts retained	211,702	180 / 40	2.29	3.02	Unbounded (2 vs. 0); identical explicit intention counts and point estimates.
Narrower prospective discontinuance definition	204,264	150 / 40	2.19	3.07	Unbounded (zero comparator cell); wide confidence intervals persist.
Minimum ten alphabetic words threshold (≥10 words)	120,966	177 / 39	2.27	3.00	Unbounded (zero comparator cell); multivariable threshold still not met.
Complete calendar years only (2021–2024)	181,070	164 / 32	1.80	2.44	Unbounded (zero comparator cell); wide confidence intervals persist.
Leave-one-application-out jackknife (across 22 apps)	102,392–204,290	76–180 / 21–40	1.19–2.64 (finite range)	1.51–3.14 (finite range)	Becomes unbounded when the single comparator review mentioning the domain is omitted.

Note: None of the sensitivity specifications yielded sufficient explicit continuance comparators to satisfy the pre-specified multivariable logistic regression feasibility thresholds (≥5 per outcome per app, ≥100 per outcome overall, and ≥10 minority outcomes per design rank). OR = odds ratio; LOAO = leave-one-application-out.

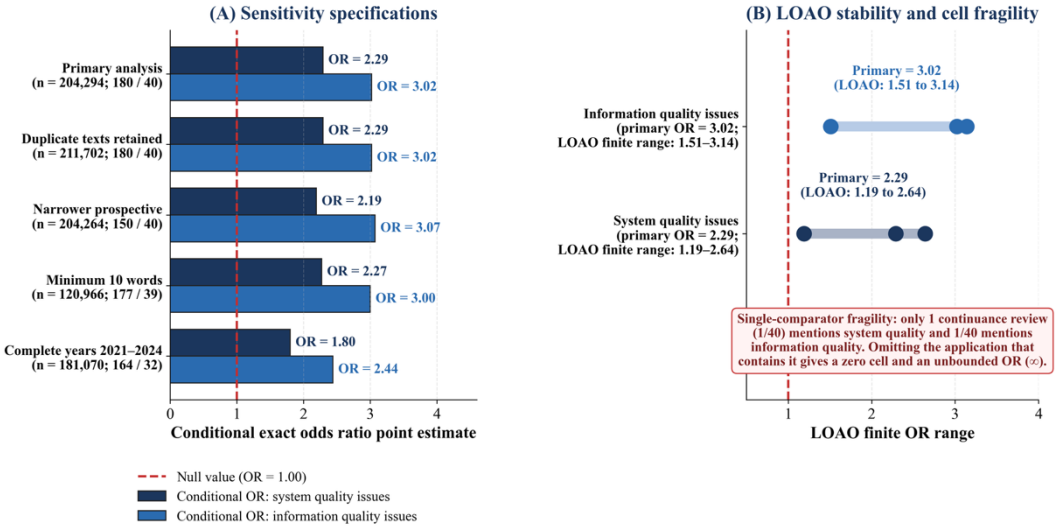


Figure 5. Sensitivity analysis across alternative corpus filtering rules and leave-one-application-out (LOAO) stability diagnostics: (A) conditional exact odds ratios for system quality issues and information quality issues across the primary corpus (n = 204,294; 180/40; OR = 2.29 and 3.02), retaining within-application duplicate texts (n = 211,702; 180/40; OR = 2.29 and 3.02), narrower prospective discontinuance definition (n = 204,264; 150/40; OR = 2.19 and 3.07), minimum ten alphabetic words (n = 120,966; 177/39; OR = 2.27 and 3.00), and complete calendar years 2021–2024 (n = 181,070; 164/32; OR = 1.80 and 2.44); and (B) LOAO finite point-estimate ranges across the 22 applications for system quality issues (1.19–2.64) and information quality issues (1.51–3.14), illustrating single-comparator cell fragility when the sole continuance review mentioning a domain (1/40) is omitted.

Review timestamps across the 204,294 deduplicated records ran from 1 January 2021 to 13 August 2025 (Table 8 and Figure 1B). Coverage was highly uneven in both review volume and observation duration. Several applications entered the observed corpus only in 2023 or 2024. Answer.AI - Your AI tutor first appeared on 30 March 2023 (n = 6,500), Question.AI - Chatbot&Math AI on 6 June 2023 (n = 15,086), Nerd AI - Tutor & Math Helper on 29 July 2023 (n = 5,362), and Homework AI - Math & Essay App on 9 August 2024 (n = 1,324). Two platforms dominated the catalog. Brainly: AI Homework Helper contributed 101,902 deduplicated reviews (49.880%, observed from 16 January 2022 to 13 August 2025). Gauth: AI Study Companion contributed 54,563 reviews (26.708%, observed from 2 January 2021 to 12 August 2025). Together, Brainly and Gauth accounted for 156,465 reviews, or 76.588% (76.6%) of the 204,294-review corpus. Across the 22 applications, the mean was 9,286.1 ± 23,743.9 reviews and the median was 1,008.5 [interquartile range, IQR: 143.0–5,024.5]. Seven of the 22 curated applications recorded zero explicit discontinuance and zero explicit continuance intentions (Table 8 and Figure 2B). These were AI Questions Generator (n = 166), AI Quiz Generator (n = 59), Academi AI: Study & Exam Tutor (n = 115), Edu AI - AI Homework Helper (n = 8), Quiz AI: AI Homework Helper (n = 140), Teacher AI - Language Practice (n = 50), and Tutor AI: Learning Assistants (n = 4). The pre-specified protocol feasibility gate required at least five reviews in both explicit intention categories, and only three platforms met it. Multivariable logistic regression was therefore not fitted. The three platforms were Brainly: AI Homework Helper (104 discontinuance versus 8 continuance), Gauth: AI Study Companion (38 discontinuance versus 19 continuance), and Question.AI - Chatbot&Math AI (9 discontinuance versus 7 continuance). Combined, these three eligible applications formed a candidate regression subset of 185 reviews, comprising 151 discontinuance intentions and 34 continuance intentions, with a model design matrix rank of 11. The 34 minority continuance comparators fell well below the pre-specified minimum of 100 outcomes per group and below the 110 minority outcomes required by the 10-events-per-design-parameter rule.

Table 8. Coverage, observation date range, and explicit intention counts across all 22 curated educational applications in the primary deduplicated corpus (N = 204,294).

Educational Application Name	Deduplicated Reviews (n)	Share of Corpus (%)	Earliest Observed Date	Latest Observed Date	Explicit Discontinuance (n)	Explicit Continuance (n)
AI Questions Generator	166	0.081%	2023-06-12	2025-08-04	0	0
AI Quiz Generator	59	0.029%	2023-06-15	2025-08-07	0	0
Academi AI: Study & Exam Tutor	115	0.056%	2023-11-19	2025-07-23	0	0
Answer.AI - Your AI tutor	6,500	3.182%	2023-03-30	2025-08-12	2	2
Blackboard	4,012	1.964%	2021-01-01	2025-08-11	6	0
Brainly: AI Homework Helper	101,902	49.880%	2022-01-16	2025-08-13	104	8
Course Hero: AI Homework Help	851	0.417%	2021-01-02	2025-08-09	4	0
Edu AI - AI Homework Helper	8	0.004%	2024-01-14	2024-08-26	0	0
Gauth: AI Study Companion	54,563	26.708%	2021-01-02	2025-08-12	38	19
Help AI: Your Homework With AI	914	0.447%	2023-09-09	2025-08-12	1	0
Homework AI - Math & Essay App	1,324	0.648%	2024-08-09	2025-08-11	1	0

Educational Application Name	Deduplicated Reviews (n)	Share of Corpus (%)	Earliest Observed Date	Latest Observed Date	Explicit Discontinuance (n)	Explicit Continuance (n)
Kahoot! Play & Create Quizzes	6,684	3.272%	2021-01-01	2025-08-12	3	0
Mathos AI: Math Helper & Tutor	152	0.074%	2024-04-04	2025-08-12	1	0
Nerd AI - Tutor & Math Helper	5,362	2.625%	2023-07-29	2025-08-12	4	2
Question.AI - Chatbot&Math AI	15,086	7.384%	2023-06-06	2025-08-12	9	7
Quiz AI: AI Homework Helper	140	0.069%	2022-09-27	2025-08-07	0	0
Quizard AI - Homework Helper	675	0.330%	2023-02-23	2025-08-11	1	0
School Hack	2,375	1.163%	2023-01-29	2025-08-02	3	0
Studocu: AI Homework Helper	1,103	0.540%	2021-01-05	2025-08-12	1	0
StudyX: AI Homework Helper	2,249	1.101%	2022-08-22	2025-08-12	2	2
Teacher AI - Language Practice	50	0.024%	2024-09-11	2025-08-09	0	0
Tutor AI: Learning Assistants	4	0.002%	2023-04-07	2024-01-01	0	0
Total across all 22 applications	**204,294**	**100.000%**	**2021-01-01**	**2025-08-13**	**180**	**40**

Note: Dates reflect the earliest and latest review timestamps observed within the filtered analytical corpus rather than official application release dates. Brainly: AI Homework Helper (101,902; 49.880%) and Gauth: AI Study Companion (54,563; 26.708%) together account for 156,465 reviews (76.588%, or 76.6% of the corpus). Only Brainly (104 vs. 8), Gauth (38 vs. 19), and Question.AI (9 vs. 7) met the initial threshold of ≥5 reviews in both explicit intention categories (totaling 151 discontinuance and 34 continuance reviews). IQR = interquartile range is reported in the text.

3.5 Cross-application stratification and lexical length

Table 8 lists the 22 curated educational applications, and a closer look shows how strongly review volume and explicit intention outcomes concentrate in a few platforms. Homework help and AI study companion platforms held the vast majority of records in the primary deduplicated corpus of 204,294 reviews. Brainly: AI Homework Helper contributed 101,902 reviews (49.880%), with 104 discontinuance and 8 continuance intentions. Gauth: AI Study Companion supplied 54,563 reviews (26.708%), with 38 discontinuance and 19 continuance intentions. Question.AI - Chatbot&Math AI added 15,086 reviews (7.384%), with 9 discontinuance and 7 continuance intentions. Answer.AI - Your AI tutor had 6,500 reviews (3.182%), with 2 discontinuance and 2 continuance intentions. Nerd AI - Tutor & Math Helper had 5,362 reviews (2.625%), with 4 discontinuance and 2 continuance intentions. School Hack contributed 2,375 reviews (1.163%), with 3 discontinuance and 0 continuance intentions. StudyX: AI Homework Helper supplied 2,249 reviews (1.101%), with 2 discontinuance and 2 continuance intentions. Homework AI - Math & Essay App added 1,324 reviews (0.648%), with 1 discontinuance and 0 continuance intentions. Studocu: AI Homework Helper had 1,103 reviews (0.540%), with 1 discontinuance and 0 continuance intentions. Help AI: Your Homework With AI contributed 914 reviews (0.447%), with 1 discontinuance and 0 continuance intentions. Course Hero: AI Homework Help supplied 851 reviews (0.417%), with 4 discontinuance and 0 continuance intentions. Quizard AI - Homework Helper added 675 reviews (0.330%), with 1 discontinuance and 0 continuance intentions. Mathos AI: Math Helper & Tutor had 152 reviews (0.074%), with 1 discontinuance and 0 continuance intentions. Institutional and gamified quiz platforms made up the next group. Kahoot! Play & Create Quizzes had 6,684 reviews (3.272%), with 3 discontinuance and 0 continuance intentions. Blackboard had 4,012 reviews (1.964%), with 6 discontinuance and 0 continuance intentions. The remaining seven smaller applications were AI Questions Generator (n = 166), Quiz AI: AI Homework Helper (n = 140), Academi AI: Study & Exam Tutor (n = 115), AI Quiz Generator (n = 59), Teacher AI - Language Practice (n = 50), Edu AI - AI Homework Helper (n = 8), and Tutor AI: Learning Assistants (n = 4). Together they contributed 542 reviews (0.265% of the corpus) and zero explicit intention statements.

Comparing the sensitivity specifications in Table 7 with the screening cascade in Table 1 reveals a difference in length between general storefront reviews and reviews that declare an explicit first-person intention. In initial screening, the five-word minimum eliminated 217,643 of 447,727 in-window reviews (48.6%). Nearly half of the raw app store submissions were thus one-to-four-word exclamations, such as 'good app', 'very helpful', or 'does not work'. Doubling the minimum from five to ten alphabetic words in sensitivity testing (Table 7) shrank the deduplicated English corpus by a further 40.8%, from 204,294 to 120,966 reviews (−83,328 reviews with between five and nine alphabetic words). Across that same reduction of 83,328 reviews, explicit discontinuance intentions fell by only 3 reviews (from 180 to 177; 98.3% retained). Explicit continuance intentions fell by only 1 review (from 40 to 39; 97.5% retained). The conditional exact ORs stayed nearly identical, at 2.27 for system quality issues and 3.00 for information quality issues. Likewise, a strictly prospective discontinuance phrasing rule removed only 30 retrospective discontinuance reviews (from 180 to 150 across n = 204,264 reviews, with OR = 2.19 for system quality and OR = 3.07 for information quality). Restricting the window to complete calendar years 2021–2024 kept 181,070 reviews with 164 discontinuance and 32 continuance intentions. This matches removing the 23,224 reviews and 16 discontinuance intentions of partial year 2025 shown in Table 6, along with 8 continuance intentions from 2025, and yields OR = 1.80 for system quality and OR = 2.44 for information quality.

4. Discussion

The main finding concerns a gap between catalog volume and the information usable for studying explicit behavioral intention in public educational app reviews. The GenAI-EdSent repository supplies 480,831 initial catalog rows and 204,294 deduplicated English reviews of at least five alphabetic words across 22 curated learning applications. Only 220 of these reviews (0.108%) state a first-person intention, either to discontinue (n = 180) or to continue (n = 40).[3,4] Large corpus size is often read as a sign of high statistical power in work on digital

learning and mobile application mining.[27,31] That reading does not hold in this corpus. When the question concerns post-adoption continuance or discontinuance intention, as conceptualized in Bhattacherjee's Expectation-Confirmation Model, the effective sample size is set by the 40 explicit continuance comparators and not by the 204,294 descriptive reviews (only 34 comparators fall within the three applications meeting minimum cell thresholds).[2] Counting the 204,074 unstated-intention reviews as continuing users would inflate the control group by more than five thousandfold. Such a choice would be a serious error, because it would equate the absence of a written uninstall threat with verified behavioral retention, and effect estimates would be distorted.

Within the 220 reviews that express an explicit intention, complaints about system quality (5.6% versus 2.5%; conditional OR = 2.29; 95% exact CI: 0.31–102.08) and information quality (7.2% versus 2.5%; conditional OR = 3.02; 95% exact CI: 0.43–132.28) were numerically more frequent among discontinuance reviews than among continuance reviews.[1,5] The direction of these point estimates matches survey-based DeLone–McLean and ECM studies, which report that software instability, interface lag, and inaccurate or hallucinated AI tutoring outputs weaken perceived usefulness and foster dissatisfaction.[6–10,16–22] The exact confidence intervals in public app store reviews, however, run from below 0.45 to over 100, and the Benjamini–Hochberg q-values equal 1.00. Neither domain can therefore be claimed as a statistically established determinant. Cross-sectional review text also cannot separate temporal sequence from reverse causality. The data cannot show whether meeting a system bug or a wrong homework solution led to the decision to uninstall. Learners who had already decided to leave over subscription paywalls may instead have been more inclined to list every technical or informational flaw. Unmeasured confounders such as application version updates, grade-level exam pressure, or device hardware constraints may also have driven both the complaints and the discontinuance statements.[29,50] Developers should not read a crude OR of 2.29 or 3.02 as a causal risk ratio that predicts how far user churn would fall if software bugs or answer accuracy were fixed.

The study also separates technical burden from commercial monetization friction (Table 2 and Table 4). Under a strict DeLone–McLean technical operationalization, which covers intrusive ad interruptions, battery or memory drain, data consumption, and overheating, technical burden appeared in only 2 of 180 discontinuance reviews (1.1%) and 0 of 40 continuance reviews (0.0%). The conditional OR was unbounded (95% exact CI: 0.04–∞).[1] The secondary descriptive proxy for commercial burden covers subscription paywalls, unanticipated billing charges, refund disputes, and high pricing. It appeared in 23 of 180 discontinuance reviews (12.8%) and 2 of 40 continuance reviews (5.0%). Corpus examples show that modern AI homework helpers often use hybrid access gates, such as daily question quotas, virtual coins, ticket systems, and blurred step-by-step explanations. The fixed primary codebook missed many of them. Recent studies of freemium mobile applications and generative AI tools (such as ChatGPT Plus and AI math solvers) report that perceived monetary cost and sudden paywall tightening trigger negative sentiment and discontinuance.[22,32,49,51] Had commercial paywall complaints been lumped into technical burden, monetization discontent would have been misattributed to technical system performance. Expanding the primary lexical dictionary after seeing sparse technical burden counts, on the other hand, would amount to post-hoc data dredging (HARKing). Refining lexicons for token-gated monetization and AI tutoring friction needs a separate development phase evaluated on held-out data.

The temporal analysis for 2021 to 2025 exposes a pitfall in longitudinal readings of curated multi-app repositories (Table 6 and Figure 4). Pooling reviews naively across all available applications, discontinuance intention detections seemed to fall steadily from 1.50 per 1,000 reviews in 2021 to 0.61 per 1,000 in 2024. Fixing the mix of apps changes the picture. Using equal weights across the five applications observed continuously with at least 20 reviews per year (Blackboard, Course Hero, Gauth, Kahoot!, and Studocu), the standardized rate rose from 2.13 per 1,000 in 2021 to 3.39 per 1,000 in 2024. This Simpson's-paradox-like reversal may reflect the fact that Brainly: AI Homework Helper (entering the corpus in January 2022 with 101,902 reviews) and several 2023–2024 AI study apps added very large numbers of brief reviews. Those reviews may have carried lower per-review rates of explicit intention detection and diluted the pooled denominator in later years. The presence of 'AI' in an application's storefront title when the dataset was curated in 2025 also does not show that reviews written in 2021 or 2022 evaluated the same generative AI features.[3,4] Valid inference over time needs a fixed panel of applications, synchronized release maturity windows, documented curator sampling frames, and version-level feature logs.

4.1 Methodological limitations and recommendations for future research

Several limitations apply (Figure 6). First, automated deterministic rules carried out all language filtering (langid 1.1.6), intention outcome classification, and experiential domain coding, with no independent human rater validation. Such rules are blunt. They can miss implicit discontinuance intentions expressed through metaphor or indirect phrasing. They can also misclassify temporary frustration, hypothetical conditions, sarcasm, or references to competing apps. The five-word minimum filter excluded 217,643 very short reviews. Exact normalized text deduplication within applications has limits too. It cannot link distinct reviews written by the same user or collapse semantic paraphrases. Human audit columns remain unrated. No Cohen's κ, sensitivity, or specificity adjustments could therefore be applied to correct for measurement error.

Second, sampling is non-probabilistic for both applications and reviewers. Two applications (Brainly and Gauth) generate 76.6% of the entire deduplicated corpus. Reviewer country is unverified, and English review text does not imply that learners live in anglophone education systems. Active user denominators, download volumes, and silent non-reviewing dropouts are unobserved. The GenAI-EdSent repository is shared under a CC BY-NC 4.0 license, yet public availability of storefront text does not exempt researchers from institutional ethical oversight or from privacy protection obligations for traceable verbatim quotes.[3]

Third, the crude exact conditional odds ratios were run as an exploratory analysis. The pre-specified multivariable logistic regression had failed its sample-size feasibility gate (34 continuance comparators across three eligible apps versus the required 100 overall and 110 for design rank 11). Exact confidence intervals and Benjamini–Hochberg adjustments do not correct for selection bias, within-app clustering, or unvalidated coding errors. Future work should follow a two-stage protocol if it aims to identify valid determinants of digital learning continuance. The first stage is blinded two-rater human annotation across a stratified sample of applications, calendar years, and candidate intention strata, with formal conflict adjudication and held-out precision, recall, and Cohen's κ benchmarking. The second links the validated textual constructs to consented, anonymized longitudinal usage telemetry with prespecified inactivity churn definitions.

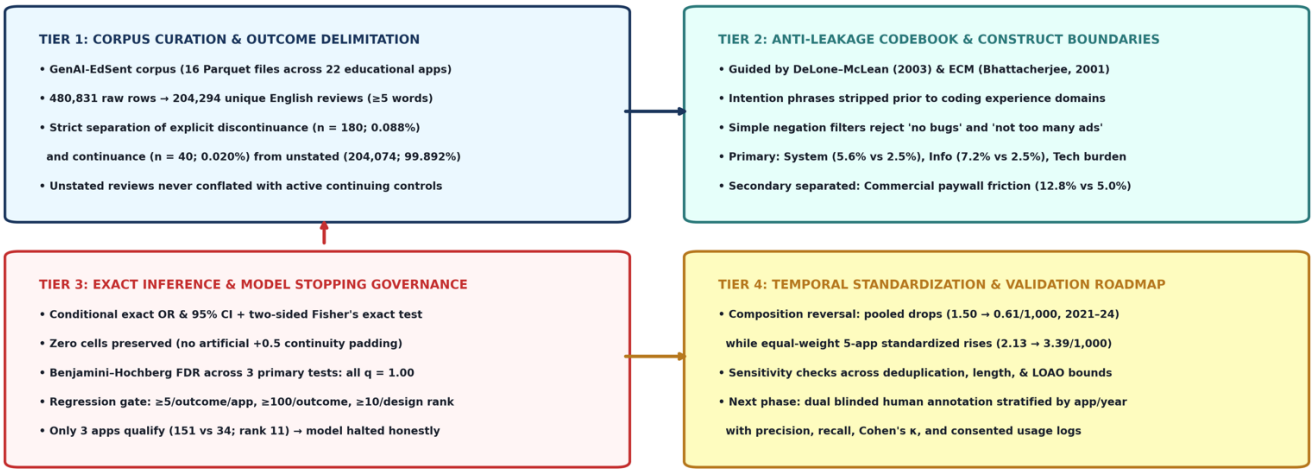


Figure 6. Four-tier methodological and theoretical governance architecture for public review continuance analysis: integrating (Tier 1) multi-stage corpus curation and strict separation of explicit discontinuance (n = 180) and continuance (n = 40) intentions from unstated-intention reviews (n = 204,074), (Tier 2) anti-leakage lexical coding and construct boundary separation between DeLone–McLean technical domains and secondary commercial paywall friction, (Tier 3) exact conditional hypergeometric inference and pre-specified multivariable regression stopping rules (halting adjusted modeling at 34 comparators vs. 100/110 required for design rank 11), and (Tier 4) equal-weighted temporal standardization and a next-phase dual-rater blinded human validation and telemetry linkage protocol.

4.2 Comparative synthesis with continuance and review-mining literature

Placing our empirical results within the 2021–2026 international literature helps explain why findings from open application store corpora differ so sharply from those obtained with cross-sectional student questionnaires.[6–24,47,48,52–55] Psychometric survey studies grounded in the DeLone–McLean IS Success Model and Bhattacherjee's ECM have been run across higher education and K–12 settings in China, India, Indonesia, Malaysia, South Korea, Vietnam, Europe, and North America. They report statistically significant positive standardized path coefficients (typically $\beta = 0.20$ to 0.55, $p < 0.001$) linking system quality and information quality to user satisfaction and continuance intention.[6–19] Recent structural equation and hybrid SEM–artificial neural network studies of mobile learning and generative AI tutoring platforms are one example: Qiu et al.,[6] Begum and Venkatesan,[7] Dai et al.,[10] Jung and Jo,[16] Ma and Zhang,[17] and Li et al.[18] show that learners prompted to rate every quality dimension and intention item on multi-point scales let perceived accuracy, responsiveness, and expectation confirmation explain much of the variance in stated continuance intention. Studies built on stimulus–organism–response and extended ECM frameworks point the same way. Pasupuleti and Thiyyagura,[14] Kala and Chaubey,[19] Silalahi et al.,[20] Duong,[21] and Nguyen[22] confirm that epistemic reliability and output clarity are central cognitive appraisals during AI-assisted study.

Why do psychometric surveys with N = 250 to 800 detect precise associations, while a public app store corpus of N = 204,294 deduplicated reviews yields wide, statistically non-significant exact confidence intervals (conditional OR = 2.29 [95% CI: 0.31–102.08] for system quality and OR = 3.02 [95% CI: 0.43–132.28] for information quality)? Three structural mechanisms explain it. First, every participant in a structured survey (100%) gives an explicit score on both the explanatory quality scales and the continuance intention outcome, so outcome missingness is eliminated. In the GenAI-EdSent storefront corpus, 99.892% of reviewers (204,074 of 204,294) do not state an explicit first-person continuance or discontinuance intention. The analyzable outcome sample shrinks to n = 220 (0.108%).[3,4] Second, who writes a storefront review is shaped by an asymmetry in vocalization. Learners who decide to abandon an application use the public review box as an exit protest, warning peers or pressuring developers (180 explicit discontinuance declarations). Satisfied learners who plan to keep using the application usually leave brief praise (illustrative wording, not a corpus quotation: 'great math helper, saved my homework') without any explicit first-person continuance clause (only 40 explicit continuance comparators).[27,31] Third, free-text reviewers mention only one salient thing. Even among the 220 explicit intention reviews, they name the single most salient source of their frustration or praise instead of working through a checklist covering system quality, information quality, and technical burden. Domain mention prevalences in brief storefront text therefore stay low (5.6% vs. 2.5% for system quality; 7.2% vs. 2.5% for information quality; 1.1% vs. 0.0% for technical burden). The continuance comparator column ends up with single-confirmation cells ($b = 1$ or $b = 0$), and the exact hypergeometric confidence intervals become wide.

Our findings can also be set beside recent computational app-review studies by Alabduljabbar,[27] Zhou and Zhou,[31] and Celik et al.[32] Those studies show the value of separating commercial paywall friction from technical burden and of enforcing anti-leakage coding rules. Many sentiment and topic-modeling analyses of generative AI and freemium mobile apps pool all negative review topics, including subscription pricing, token limits, software bugs, and answer errors, into undifferentiated negative sentiment clusters and correlate them with 1-star ratings.[27,31,32] Disaggregation gives a different order among learners who explicitly declare an intention to stop using digital learning apps. Commercial burden complaints (23/180; 12.8%) come first, outnumbering information quality complaints (13/180; 7.2%), system quality complaints (10/180; 5.6%), and technical resource burden (2/180; 1.1%). In modern AI-enabled homework companions, learners often download an application expecting free step-by-step tutoring and meet paywalls, mandatory credit-card trials, or daily ticket quotas after scanning their first problem.[32,49] This barrier causes immediate expectation disconfirmation that differs from software bugs or LLM hallucination. Future DeLone–McLean and ECM extensions for commercial EdTech should therefore model perceived financial value and paywall transparency next to system and information quality.[1,2,22,32,49]

4.3 Blueprint for dual-rater human calibration and misclassification-adjusted inference

Confirmatory inference needs more than provisional rule-based exploration. In educational review mining it requires the four-tier governance architecture shown in Figure 6. In the validation phase, two independent human annotators, blinded to the automated lexical labels, should code a stratified random sample drawn across three orthogonal dimensions. The first is application stratum, which should represent dominant platforms such as Brainly and Gauth alongside mid-sized and institutional apps. The second is year (2021 through 2025). The third is candidate intention stratum, oversampling candidate

discontinuance, candidate continuance, and past-uninstall reviews together with a random sample of unstated-intention reviews. Explicit intention is rare in the unstratified corpus (0.108%). Inter-rater reliability reporting should thus include both unweighted Cohen's κ and prevalence-adjusted bias-adjusted kappa (PABAK). Class-specific positive predictive value (precision), sensitivity (recall), and F1 scores should accompany them, evaluated on a strictly held-out test partition before any lexical dictionary refinement. Once human-validated sensitivity and specificity matrices exist, researchers can apply misclassification-adjusted contingency estimators. They can also link anonymized, consented platform usage telemetry (measuring actual session decay and 30-day or 90-day inactivity churn) to test whether stated storefront intentions and quality complaints predict objective behavioral discontinuance in digital education.[56,57]

5. Conclusion

This study analyzed 480,831 initial catalog rows and 204,294 deduplicated English reviews across 22 educational applications in the GenAI-EdSent repository, using a reproducible multi-stage procedure. It illustrates that a large review catalog does not ensure enough explicit intention outcomes for multivariable determinant modeling. Only 220 reviews (0.108%) held an explicit first-person discontinuance (n = 180; 0.088%) or continuance (n = 40; 0.020%) intention. Just 34 continuance comparators sat within the three applications that met initial cell-size criteria, well below the protocol feasibility thresholds for adjusted regression. Under the rule-based codebook guided by the DeLone–McLean and Expectation-Confirmation frameworks, neither system quality issues (conditional exact OR = 2.29; 95% exact CI: 0.31–102.08; p = 0.694; q = 1.00) nor information quality issues (conditional exact OR = 3.02; 95% exact CI: 0.43–132.28; p = 0.474; q = 1.00) showed a statistically precise association with explicit discontinuance intention. Technical burden (2/180 versus 0/40) gave an unbounded estimate because one comparator cell was zero (95% exact CI: 0.04–∞; p = 1.000; q = 1.00). The 2021–2024 trajectories also diverged. Pooled catalog rates declined from 1.50 to 0.61 per 1,000, while equal-weighted five-application standardized rates rose from 2.13 to 3.39 per 1,000. Unstandardized catalog trends are therefore easily distorted by shifts in application composition. These findings do not identify determinants of learner retention. They support the case for construct validation, two-rater human calibration, and linkage to longitudinal telemetry in educational technology continuance research.

Declaration

Artificial Intelligence Use

The authors did not use generative artificial intelligence or AI-assisted tools to write, edit, or analyze any part of this work. No review text from the dataset was passed to an external language model service. The authors reviewed all content and take full responsibility for it.

Author Contributions

Conceptualization, H.P.; Methodology, H.P.; Software, H.P.; Validation, H.P. and M.M.; Formal analysis, H.P.; Data curation, H.P.; Writing – original draft, H.P.; Writing – review & editing, H.P. and M.M.; Visualization, H.P.; Project administration, H.P. All authors have read and agreed to the final version of the manuscript.

Funding

This study received no funding from any public, commercial, or not-for-profit agency. The work was carried out without external financial support, and no funder had any role in its design, analysis, or reporting.

Conflict of Interest

The authors declare no competing financial or non-financial interests related to this work. The 22 educational applications in the GenAI-EdSent corpus are analyzed only to document corpus coverage and to evaluate the methodological and empirical limits of digital learning review research, without commercial endorsement or product ranking.

Ethics Approval and Consent

This retrospective secondary analysis used publicly available application store review records from the GenAI-EdSent repository (CC BY-NC 4.0 license). Reviewer account display names, profile photographs, and developer reply metadata were excluded from the analytical matrix, public review IDs were replaced by local cryptographic hashes in derived feature files, and no raw review text was sent to external language model services.

Consent for Publication

Not applicable. No identifiable individual user account metadata, profile images, or private learner telemetry records are reproduced in this manuscript.

Data Availability

The GenAI-EdSent dataset is publicly available from the Hugging Face repository (https://huggingface.co/datasets/Erfan-Nourbakhsh/GenAI-EdSent) under the CC BY-NC 4.0 license. The coding rules, analysis code, and derived tables can be obtained from the corresponding author on reasonable request.

Acknowledgments

The authors thank Erfan Nourbakhsh and A. Mazaherian for curating the GenAI-EdSent educational application review repository and for releasing it publicly for non-commercial academic research. No other individual or organization contributed to this work in a way that requires acknowledgment.

References

1. DeLone WH, McLean ER. The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst. 2003;19(4):9-30. doi:10.1080/07421222.2003.11045748
2. Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Q. 2001;25(3):351-370. doi:10.2307/3250921
3. Nourbakhsh E. GenAI-EdSent dataset [dataset]. CC BY-NC 4.0 license. Hugging Face; 2025. Available from: https://huggingface.co/datasets/Erfan-Nourbakhsh/GenAI-EdSent
4. Mazaherian A, Nourbakhsh E. Unveiling user perceptions in the generative AI era: a sentiment-driven evaluation of AI educational apps role in digital transformation of e-teaching. arXiv [Preprint]. 2025:arXiv:2512.11934. Available from: https://arxiv.org/abs/2512.11934
5. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2
6. Qiu X, Wu J, Li H. Continuance intention to use mobile learning among college students: integrating quality factors and the expectation confirmation model. Sci Rep. 2026;16(1):5691. doi:10.1038/s41598-026-35949-w
7. Begum NF, Venkatesan D. Understanding continuance intention towards e-learning platforms: a structural model based on expectation confirmation and individual innovativeness. Front Educ. 2026;11:1892980. doi:10.3389/feduc.2026.1892980
8. Wang T, Lin C, Su Y. Continuance intention of university students and online learning during the COVID-19 pandemic: a modified expectation confirmation model perspective. Sustainability. 2021;13(8):4586. doi:10.3390/su13084586
9. Faozi FH, Handayani PW. The antecedents of mobile-assisted language learning applications continuance intention. Electron J e-Learn. 2023;21(4):299-313. doi:10.34190/ejel.21.4.2744
10. Dai J, Zhang X, Wang C. A meta-analysis of learners' continuance intention toward online education platforms. Educ Inf Technol. 2024;29(16):21833-21868. doi:10.1007/s10639-024-12654-7

11. Chandra A, Sengupta S, Vaish A. An investigative study among higher education students' continuance intention towards e-learning in India. Educ Inf Technol. 2025;30(4):4933-4957. doi:10.1007/s10639-024-12997-1
12. Ni A, Cheung A. Understanding secondary students' continuance intention to adopt AI-powered intelligent tutoring system for English learning. Educ Inf Technol. 2023;28(3):3191-3216. doi:10.1007/s10639-022-11305-z
13. Jang Y, Chiang I. Incorporating desire and persistence into understanding Gen Z learners' continuance intention toward using YouTube for learning in digital learning context. Educ Inf Technol. 2024;29(8):10043-10068. doi:10.1007/s10639-023-12202-9
14. Pasupuleti RS, Thiyyagura D. An empirical evidence on the continuance and recommendation intention of ChatGPT among higher education students in India: an extended technology continuance theory. Educ Inf Technol. 2024;29(14):17965-17985. doi:10.1007/s10639-024-12573-7
15. Zhao X, Zhang Y, Yue S, et al. Understanding continuance learning intention in online educational videos: the role of danmaku (real-time interactive comments) in informal learning. Educ Inf Technol. 2026;31(14):5555-5594. doi:10.1007/s10639-026-14008-x
16. Jung YM, Jo H. Understanding continuance intention of generative AI in education: an ECM-based study for sustainable learning engagement. Sustainability. 2025;17(13):6082. doi:10.3390/su17136082
17. Ma T, Zhang W. Modeling K-12 teachers' continuance intention to use generative AI: influencing factors and a hybrid SEM-ANN approach. Front Educ. 2026;11:1755301. doi:10.3389/feduc.2026.1755301
18. Li H, Xu Y, Brown C, et al. Sustainable AI integration in education: factors influencing pre-service teachers' continuance intention to use generative AI. Sustainability. 2026;18(7):3291. doi:10.3390/su18073291
19. Kala D, Chaubey DS. Learning by the prompt: impact of generative AI on student engagement, satisfaction, and continuance intention using stimulus-organism-response model. Asian Educ Dev Stud. 2026;15(2):1-19. doi:10.1108/aeds-08-2025-0373
20. Silalahi ADK, Riantama D, Phuong DTT. Redefining socio-technical systems for generative AI in education: evaluating its impact on continuance intention. J Sci Technol Policy Manag. 2026;17(1):1-30. doi:10.1108/jstpm-07-2025-0305
21. Duong CD. Modeling the determinants of HEI students' continuance intention to use ChatGPT for learning: a stimulus-organism-response approach. J Res Innov Teach Learn. 2024;17(2):391-407. doi:10.1108/jrit-01-2024-0006
22. Nguyen BN. Explaining ChatGPT Plus users' continuance intention: comparing TAM, ECM, and extended ECM. Electron Libr. 2026;44(6):1195-1216. doi:10.1108/el-12-2025-0478
23. Alshammari T. Instrument development for assessing continuance intention toward generative AI tools in education. Int J Inf Educ Technol. 2026;16(9):2541-2550. doi:10.18178/ijiet.2026.16.9.2711
24. Du K. Continuance intention to use gamified mobile apps for English speaking skills learning in Chinese English as Foreign Language (EFL) higher education. Int J Inf Educ Technol. 2024;14(9):1217-1226. doi:10.18178/ijiet.2024.14.9.2151
25. Baden C, Pipal C, Schoonvelde M, et al. Three gaps in computational text analysis methods for social sciences: a research agenda. Commun Methods Meas. 2022;16(1):1-18. doi:10.1080/19312458.2021.2015574
26. Martens D, Maalej W. Towards understanding and detecting fake reviews in app stores. Empir Softw Eng. 2019;24(6):3316-3355. doi:10.1007/s10664-019-09706-9
27. Alabduljabbar R. User-centric AI: evaluating the usability of generative AI applications through user reviews on app stores. PeerJ Comput Sci. 2024;10:e2421. doi:10.7717/peerj-cs.2421
28. Korfiatis N, Stamolampros P, Kourouthanassis P, et al. Measuring service quality from unstructured data: a topic modeling application on airline passengers' online reviews. Expert Syst Appl. 2019;116:472-486. doi:10.1016/j.eswa.2018.09.037
29. Xie Y, Yeoh W, Wang J. How self-selection bias in online reviews affects buyer satisfaction: a product type perspective. Decis Support Syst. 2024;181:114199. doi:10.1016/j.dss.2024.114199
30. Smironva E, Kiatkawsin K, Lee SK, et al. Self-selection and non-response biases in customers' hotel ratings: a comparison of online and offline ratings. Curr Issues Tour. 2020;23(10):1191-1204. doi:10.1080/13683500.2019.1599828
31. Zhou C, Zhou L. Understanding user experience in generative AI application: evidence from topic modelling and sentiment analysis of user reviews. Inf Res. 2026;31(iConf):235-253. doi:10.47989/ir31iconf64160
32. Celik AA, Balcioglu YS, Altindag E. Analyzing global user sentiments on freemium mobile app subscriptions. Int J E-Serv Mob Appl. 2025;17(1):1-38. doi:10.4018/ijesma.396957
33. Mcilroy S, Shang W, Ali N, Hassan AE. User reviews of top mobile apps in Apple and Google app stores. Commun ACM. 2017;60(11):62-67. doi:10.1145/3141771
34. Han S, Anderson CK. The platform matters: selection and measurement bias in online reviews. Cornell Hosp Q. 2025;67(1):42-53. doi:10.1177/19389655251327536
35. Meng XL. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. Ann Appl Stat. 2018;12(2):685-726. doi:10.1214/18-AOAS1161SF
36. Bradley VC, Kuriwaki S, Isakov M, et al. Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature. 2021;600(7890):695-700. doi:10.1038/s41586-021-04198-4
37. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163. doi:10.1186/s12874-016-0267-3
38. Gosho M, Ohigashi T, Nagashima K, et al. Bias in odds ratios from logistic regression methods with sparse data sets. J Epidemiol. 2023;33(6):265-275. doi:10.2188/jea.je20210089
39. Henderikx MA, Kreijns K, Kalz M. Refining success and dropout in massive open online courses based on the intention-behavior gap. Distance Educ. 2017;38(3):353-368. doi:10.1080/01587919.2017.1369006
40. Dabbagh N, Fake H. Generative AI (GenAI) as a mindtool that supports generative learning (GL). Comput Educ Artif Intell. 2026;11:100626. doi:10.1016/j.caeai.2026.100626
41. Jin S. Measures of learner-generative AI relationships. Comput Educ Open. 2025;8:100258. doi:10.1016/j.caeo.2025.100258
42. Velander J. Beyond operational skills: teachers' AI knowledge and interactions with generative AI in lesson planning. Comput Educ Open. 2026;10:100371. doi:10.1016/j.caeo.2026.100371
43. Sagynbayeva A, Pyo A, Yoon S, et al. Evaluating user performance with RAG-based generative AI: a scenario-based experiment on AI-assisted information retrieval. Comput Hum Behav. 2026;180:108952. doi:10.1016/j.chb.2026.108952
44. Tafazoli D. Exploring the potential of generative AI in democratizing English language education. Comput Educ Artif Intell. 2024;7:100275. doi:10.1016/j.caeai.2024.100275
45. Chiu TKF. Future research recommendations for transforming higher education with generative AI. Comput Educ Artif Intell. 2024;6:100197. doi:10.1016/j.caeai.2023.100197
46. Pecuchova J, Benko L, Drlik M. Reimagining feedback through generative AI in engineering education. Comput Educ Artif Intell. 2026;10:100574. doi:10.1016/j.caeai.2026.100574
47. Satiman LH, Zulkifli N. Determinants of teachers' continuance intention to adopt virtual learning environment within Malaysian schools. Int J Interact Mob Technol. 2022;16(14):31-42. doi:10.3991/ijim.v16i14.30597
48. Zhu M, Zhu J, Cao Y, et al. Understanding the impact of MOOC platform affordances on learners' continuance intention: an integrative perspective of platform identity and social identity. Interact Learn Environ. 2026;34(3):1-19. doi:10.1080/10494820.2026.2731475
49. Deng Y, Lambrecht A, Liu Y. Spillover effects and freemium strategy in the mobile app market. Manag Sci. 2023;69(9):5018-5041. doi:10.1287/mnsc.2022.4619
50. Li X, Zhang B, Zhang Z, et al. A sentiment-statistical approach for identifying problematic mobile app updates based on user reviews. Information. 2020;11(3):152. doi:10.3390/info11030152
51. Koghut M, Al-Tabbaa O. Exploring consumers' discontinuance intention of remote mobile payments during post-adoption usage: an empirical study. Adm Sci. 2021;11(1):18. doi:10.3390/admsci11010018
52. Jin L, Shang D. What drives art education massive open online courses continuance intention? Exploring determinants from

value-software-hardware-design perspective. Interact Learn Environ. 2022;32(4):1-17. doi:10.1080/10494820.2022.2153146

53. Rong Y, Bao Z, Huang Y, et al. Driving mechanisms of students' continuance intention to use AIDTs in co-creation contexts: a multi-method analysis using SEM-ANN-NCA. Interact Learn Environ. 2026;34(2):1-30. doi:10.1080/10494820.2026.2702022
54. Yu T, Dai J, Chen X, et al. Factors influencing continuance intention in blended learning among business school students in China: based on grounded theory and FsQCA. Interact Learn Environ. 2024;32(10):1-28. doi:10.1080/10494820.2024.2370477
55. Choi S, Jang Y, Kim H. Exploring factors influencing students' intention to use intelligent personal assistants for learning. Interact Learn Environ. 2024;32(8):4049-4062. doi:10.1080/10494820.2023.2194927
56. Moreno-Marcos PM, Muñoz-Merino PJ, Maldonado-Mahauad J, et al. Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. Comput Educ. 2020;145:103728. doi:10.1016/j.compedu.2019.103728
57. Kwon H, Kim HH, An J, et al. Lifelog data-based prediction model of digital health care app customer churn: retrospective observational study. J Med Internet Res. 2021;23(1):e22184. doi:10.2196/22184