# An Empirical Analysis of AI Tutor Usage, Skill-Based Clustering, Activity Metrics, and Essay Engagement on an Online Education Platform *

Alban Brüchig, Osman Örnek, Filip Mikovíny

## ABSTRACT

We investigate four complementary aspects of student interaction with GoGymi, an online educational platform enhanced by Gymitrainer (the application's AI tutor):

**RQ1**: How does usage of Gymitrainer/the AI model relate to students' mathematics quiz performance over time?

**RQ2**: Do clusters of students based on essay skill scores exhibit distinct performance patterns in mathematics and text quizzes?

**RQ3**: How do student clusters based on their average quiz scores in each subject and total activity count differ in their engagement patterns within the platform?

**RQ4**: Can metrics of student activity predict disengagement (churn) from the platform?

We leverage Bayesian Knowledge Tracing, Principal Component Analysis followed by K-Means Clustering, gradient-boosted Classification with SHAP explanations, and descriptive statistics on essay engagement. Our methods are exhaustively detailed to ensure reproducibility.

## Keywords

Bayesian Knowledge Tracing, Principal Component Analysis, K-Means Clustering, Gradient-Boosted Classification, SHAP

## 1. INTRODUCTION

Education is at once humanity's most enduring endeavor and its most urgent challenge: how do we kindle curiosity, nurture understanding, and cultivate the capacity for lifelong growth? In the digital age, online platforms promise to extend learning beyond the classroom's walls — allowing any student, anywhere, to learn at their own pace with nothing more than electricity and an internet connection. Yet this promise brings new questions: in a landscape of pixels and data, how do we ensure that every interaction advances genuine comprehension rather than superficial completion? How can we discern, from the tangled traces of clicks and keystrokes, the deeper arcs of progress, struggle, and engagement?

At the heart of these questions lies a fundamental belief: that each learner is a unique agent of their own growth, endowed with intrinsic curiosity and the capacity to push intellectual boundaries. Our mission is to honor that agency — to transform raw interaction logs into insights that illuminate how students truly learn, where they falter, and which supports propel them forward. By applying rigorously validated machine-learning techniques — Bayesian Knowledge Tracing to model evolving mastery, clustering to surface learning archetypes, and predictive analytics to flag early signs of disengagement — we seek not merely to measure outcomes but to reify the invisible processes of reflection, revision, and discovery.

Concretely, we investigate students' journeys on GoGymi, an adaptive platform augmented by the Gymitrainer AI tutor, designed to prepare middle-school learners in the Canton of Zurich for their high-school entrance exams. Rather than treating education as a pipeline of content delivery, we treat it as a tapestry woven from moments of insight, feedback, and renewed effort. Our questions are thus fourfold:

- How does the intensity of AI-driven support shape mathematical learning trajectories?
- What latent profiles emerge when we cluster users by essay-writing skills, and how do these profiles correlate with broader performance and engagement patterns?
- Which latent profiles emerge when we cluster users by subject, and how do these profiles correlate with broader performance and engagement patterns?
- And finally, can early activity signals predict which students risk falling away, so that timely interventions can be deployed?

By framing each sub-study as a step in the larger journey of understanding — and by rooting our methods in both

*(Does NOT produce the permission block, copyright information nor page numbering). For use with edm_article.cls.

educational theory and machine learning for education best practice — we aim to chart a path toward more humane, more personalized, and ultimately more effective learning experiences. In doing so, we contribute to the grander mission of education: empowering every student to become not just a consumer of knowledge, but an active architect of their own intellectual growth.

## 2. RELATED WORK

Baillifard et al. (2024) [2] present a controlled study in which embedding a personal AI tutor into a university-level neuroscience course led to significantly higher exam scores among students who actively used the system compared to a non-tutored control group. This case study underlines the promise of adaptive, AI-driven support for learning. Unlike Baillifard et al., who focus on university neuroscience, our work evaluates AI-tutor effectiveness in a middle-school math setting using Bayesian Knowledge Tracing to quantify learning dynamics over time.

We relied on the comprehensive framework laid out in Abdelrahman et al.'s survey of knowledge-tracing methods [1] to inform every aspect of our BKT implementation. In particular, we followed their recommendation for evaluation—using AUC as one of the primary fit metrics. Their discussion of data sparsity and per-skill parameter sharing also motivated our decision to fit separate cohort-specific BKT models (non-AI, light-AI, heavy-AI), ensuring robust parameter estimates even when interactions per student–skill pair are limited.

Wang (2025) [4] analyzes the learning behaviors of open-education learners by first reducing dozens of problem-interaction features via PCA, then applying K-means to effectively categorized students into different groups, based on engagement and performance. Each profile exhibited distinct course-completion and grade outcomes, demonstrating how unsupervised dimensionality reduction plus clustering can surface different learner archetypes. In a similar vein, our PCA–K-means clustering of essay-rubric scores reveals latent writing-skill cohorts that correlate with both math and text assignment performance.

Kim et al. (2023) [3] present a Student Dropout Prediction (SDP) system for a large university dataset. Their approach combined an XGBoost classifier (a gradient-boosted decision tree model) with SMOTE oversampling to handle class imbalance. Importantly, they applied SHAP (Shapley Additive Explanations) to identify the most influential features contributing to dropout risk. Building on Kim et al.'s use of XGBoost+SHAP for dropout, we adapt their interpretable pipeline to predict churn on GoGymi by incorporating detailed activity metrics.

## 3. METHODOLOGY

### 3.1 RQ1: AI-Tutor Engagement and Mathematical Learning

#### 3.1.1 Data Preprocessing

To prepare a dataset for modeling, we first linked each Gymitrainer-user interaction to each unique user ID using a mapping with a high confidence percentile, thereby enabling a per-student analysis of AI-tutor interactions alongside platform performance logs. Next, we standardized and inferred subject labels for each message based on keyword matching and the GoGymi curriculum taxonomy, ensuring that domain-specific effects could be isolated. To avoid contamination from non-student usage, all teacher and administrator accounts were removed.

For the mathematics question bank, we observed that raw question IDs yielded highly sparse and irregular learning traces. To mitigate this, we consolidated individual items into broader, curriculum-aligned skill categories (*Arithmetic*, *Geometry*, *Word Problems*) using a semi-automated procedure informed by both ChatGPT suggestions and the official GoGymi syllabus. This grouping enhances statistical stability and more faithfully represents students' conceptual progressions. All entries corresponding to empty or malformed questions were discarded to eliminate noise.

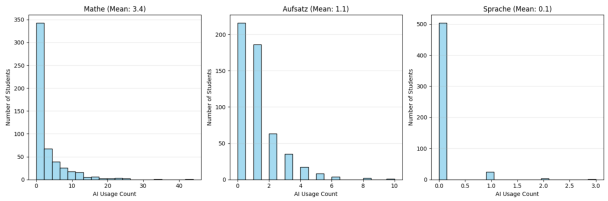


**Figure 1: Per-skill Distribution of Gymitrainer Message Counts across Users, illustrating Variability in AI-Tutor Engagement.**

We then quantified AI-tutor engagement by computing, for each user and each subject, the total number of messages exchanged with Gymitrainer. Users were stratified into three engagement cohorts:

- **Non-AI Users**: zero messages;
- **Light-AI Users**: 1 to the 75th percentile of nonzero message counts;
- **Heavy-AI Users**: above the 75th percentile.

Thresholds were determined empirically to balance cohort sizes and preserve interpretability. All event records were then chronologically ordered per user to respect the temporal dependencies intrinsic to learning processes.

Finally, we engineered a set of summary features capturing each student's practice history for each skill: total opportunities, cumulative success rate, time-between-attempts statistics, and rolling-window counts of recent practice. These features serve as covariates in our knowledge tracing models, allowing us to quantify how repeated exposure and past performance influence future mastery.

#### 3.1.2 Knowledge Tracing Model Implementation

We adopt the Bayesian Knowledge Tracing (BKT) framework to model each student's latent mastery over time for each mathematics skill. BKT represents knowledge as a hidden binary variable (learned vs. unlearned) and uses four interpretable parameters: initial mastery ($P_0$), learning transition probability ($T$), guess ($G$), and slip ($S$). We fit separate

BKT models for each AI-engagement cohort to assess how tutor usage modulates learning dynamics.

Model fitting proceeded as follows:

1. **Training–Test Split:** Interactions were partitioned by user using GroupShuffleSplit (70% train, 30% test, random_state=0).
2. **Parameter Estimation:** We employed the pyBKT library to maximize data likelihood via Expectation – Maximization, constraining parameters to $[0, 1]$.
3. **Evaluation:** On held-out data, we computed per-opportunity prediction accuracy, RMSE, and AUC for each skill.
4. **Learning Curves:** We aggregated predicted correct-response probabilities over repeated opportunities, plotting mean trajectories with 95% Wilson confidence intervals.

This approach yields both global metrics of model fit and fine-grained visualizations of how AI-tutor support influences the rate and stability of skill acquisition.

## 3.2 RQ2: Essay - Skill Profiling and Cross-Domain Performance

### 3.2.1 Essay Submission Frequency Analysis

As an initial exploration of writing engagement, we counted the total number of essay submissions per student. To reveal broad participation patterns, we binned users into five frequency categories: 1, 2–3, 4–6, 7–10, and 11+ submissions. This categorization highlights the long-tail distribution of writing practice and informs subsequent clustering by indicating typical engagement levels.

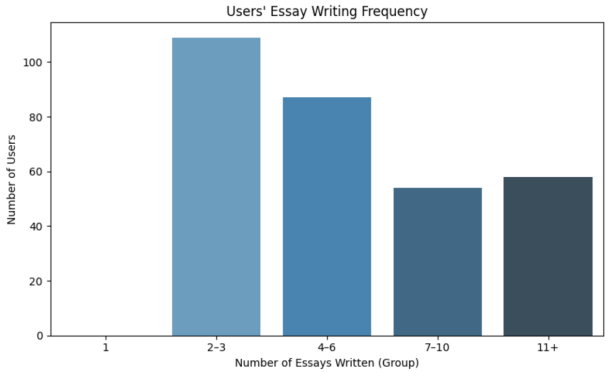


**Figure 2: Histogram of Essay Submission Counts per User, demonstrating the Heterogeneity of Writing Engagement.**

### 3.2.2 Dimensionality Reduction via Principal Component Analysis

Essay quality is assessed along fourteen rubric dimensions spanning Content (e.g. topical relevance, idea development), Structure (e.g. coherence, organization), and Language (e.g. grammar, style). To mitigate the curse of dimensionality and to facilitate clustering, we standardized each feature to zero mean and unit variance before applying PCA. We retained the first seven principal components, which together explain over 85% of the total variance, thereby preserving the majority of information while reducing noise.

### 3.2.3 K-Means Clustering and Cluster Validation

We performed K-means clustering in the seven-dimensional PCA space, systematically varying $k$ from 2 to 10. Optimal cluster count was determined by jointly considering:

- *Silhouette coefficient* (higher is better),
- *Davies–Bouldin index* (lower is better),
- *Calinski–Harabasz score* (higher is better).

All three metrics indicated that $k = 5$ achieved the best trade-off between cohesion and separation (Silhouette = 0.306; Davies–Bouldin = 0.929; Calinski–Harabasz = 211). We initialized K-means with 10 random restarts (random_state=42) to ensure replicable convergence.

Each student was thus assigned to one of five essay-skill clusters. To visualize cluster structure, we plotted PC1 versus PC2 with points colored by cluster label.

Finally, to examine cross-domain performance, we merged cluster labels with each student's mean quiz accuracy in Mathematics and Language domains, enabling a comparative analysis of how writing proficiency profiles relate to broader academic outcomes.

## 3.3 RQ3: Uncovering Learner Archetypes via Global Performance–Engagement Clustering

### 3.3.1 Preprocessing

To uncover latent patterns in how students allocate effort across subjects and engage with GoGymi, we constructed a feature space that combines both performance and usage metrics. Specifically:

- **Reused metrics:** We imported each student's AI-tutor engagement metrics from RQ1/BKT — namely, the per-subject AI Usage category based on users' Gymi-trainer message counts.
- **New metrics:** We computed each student's *mean quiz score* in each subject they attempted, and a *total activity count* aggregating all platform events (logins, problem attempts, essay submissions, studying a topic, etc.). These two axes capture, respectively, *what* students know (either have learned while using GoGymi or knew beforehand) and *how much* they have interacted with the platform.

### 3.3.2 Clustering, Validation and Confounders

Because not every student engages with every subject, direct comparisons are confounded by missing data: a zero in "mean score-Essay" may indicate either failure or non-participation. Rather than discarding partial profiles, we

opted for a higher cluster granularity — varying $k$ from 2 up to 10 — to allow the clustering algorithm to separate "non-takers" from "strivers." We applied K-means in this augmented space and evaluated cluster quality using the same familiar criteria from our earlier analyses (silhouette coefficient, Davies-Bouldin index and Calinski-Harabasz).

To mitigate the confounding effect of subject-coverage, we emphasize cluster interpretability over minimal inertia: our goal is not merely to compress the data, but to surface meaningful cohorts whose distinct mastery–engagement profiles suggest different pedagogical interventions.

## 3.4 RQ4: User Engagement Analysis Through Behavioral Activity Metrics

### *3.4.1 Carving Nuanced Learner Activity Profiles through Innovative Feature Design*

To identify which activity signals have a connection to the student falling away from using GoGymi, we define multiple user-level features:

- **Churn.** We define a student as churned, i.e. $\text{churn}(i) = 1$, if the interval between their last recorded activity and the study's cutoff date exceeds 28 days. The cutoff date is set to the latest timestamp in our dataset—March 7, 2025—which coincides with the annual Gymnasium entrance-exam window (March 1–15). Consequently, any user whose most recent interaction occurred on or before February 7, 2025, is classified as churned; all others are considered active (churn = 0).
- **Active Weeks:** count of weeks in which the student completed at least one activity.
- **Longest Consecutive Weekly Streak:** the longest run of consecutive active weeks of the student.
- **Total Activities:** the total count of all activities done by the user.
- **Average Weekly Activity:** The average number of activities completed by the user per week in which they were active.
- **First Activity Timestamp:** The timestamp corresponding to the user's first recorded interaction with the platform, indicating when they first started using the app.
- **Standard Deviation of Activity Gaps:** The standard deviation of the time intervals (in days) between successive activities, reflecting the regularity or irregularity of the user's engagement behavior.

### *3.4.2 XGBoost Classification*

We then train an XGBoost classifier with the churn label as the target, and the remaining features as features to the model. We split the data 80/20 into a train and test set.

## 3.5 Evaluation Metrics

For BKT, we use learning curves to identify differences across different AI usage groups. Additionally, we use the Root Mean Squared Error (RMSE) to measure how closely the model's predicted probabilities align with the actual binary outcomes (correct or incorrect answers). Since BKT outputs probabilities and the ground truth is binary, a lower RMSE (ranging from 0 to 1) indicates better calibration.

The Area Under the ROC Curve (AUC-ROC) evaluates the model's ability to distinguish between correct and incorrect answers. Values range from 0.5 (no better than random guessing) to 1.0 (perfect discrimination), with higher values indicating better separability between correct and incorrect predictions.

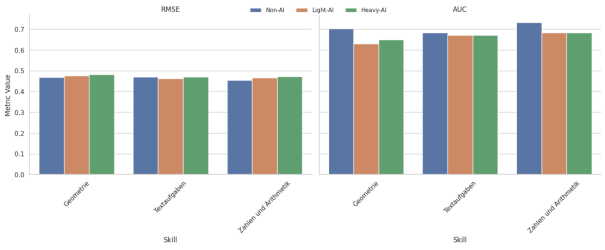


**Figure 3: BKT: RMSE & AUC per-AI Usage Category; per-Skill**

To evaluate the XGBoost classifier, after training, we predicted on the test set and computed a confusion matrix. The confusion matrix — plotted as a heatmap — revealed how many churned users the model correctly flagged versus how many remained undetected.

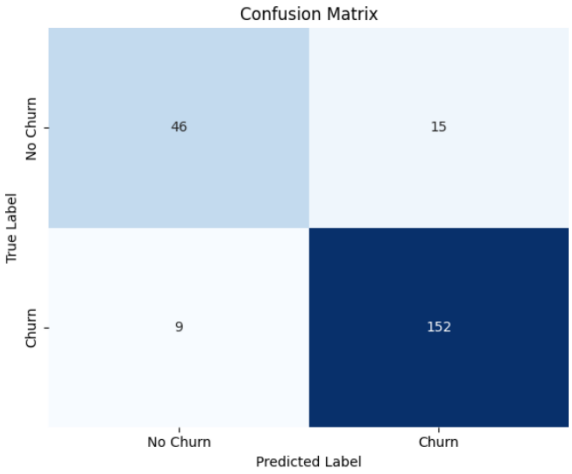


**Figure 4: Confusion Matrix of the XGBoost Classifier**

To choose a good number of clusters, we run KMeans for k from 2 to 14, recording both the inertia (total within-cluster sum of squared distances) and the average silhouette score. An "elbow plot" of inertia vs. k and a "silhouette plot" of silhouette score vs. k reveal a few promising candidate cluster counts where inertia flattens out and silhouette is relatively high.

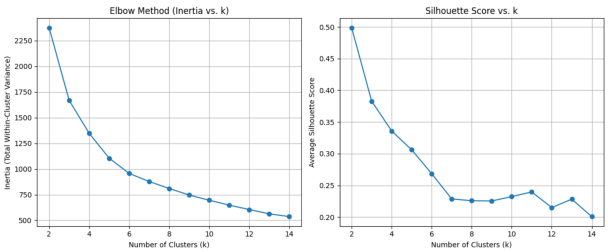


**Figure 5: Elbow Method and Silhouette Score (RQ2)**

We then narrow our focus to k = [3, 4, 5] and compute three cluster-quality metrics for each:

- Silhouette Score (higher is better)
- Davies-Bouldin Index (lower is better)
- Calinski-Harabasz Score (higher is better)

# 4. RESULTS

## 4.1 RQ1 (BKT)

Based on the BKT results, the AI chatbot does not appear to significantly support student learning in mathematics. Across all skills with sufficient data, the learning curves show minimal variation between AI usage groups. We observe that, after some time and when there are still enough observations, there is no particular improvement for this skill among non-AI and heavy-AI users, but we do observe on the contrary that light-AI users seem to progress. We also observe similar trends for other math skills.

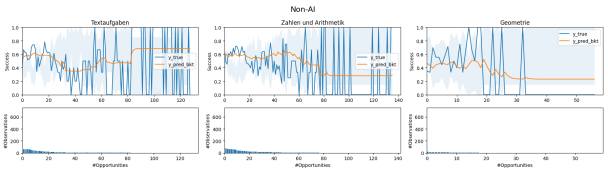


**Figure 6: Learning Curve of Non-AI Users by Math Skill**

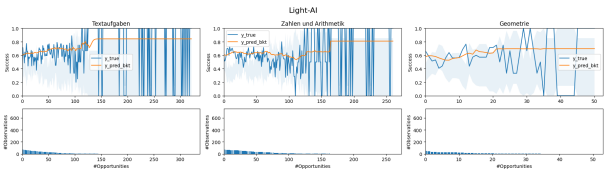


**Figure 7: Learning Curve of Light-AI Users by Math Skill**

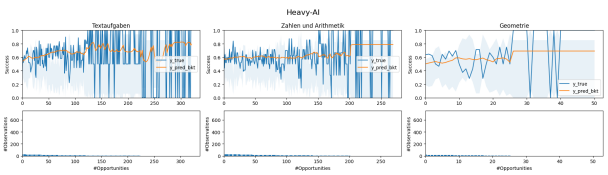


**Figure 8: Learning Curve of Heavy-AI Users by Math Skill**

## 4.2 RQ2 (Essay-Focused K-Means Clustering)

To visualize these fourteen-dimensional cluster centroids, we construct a radar plot.

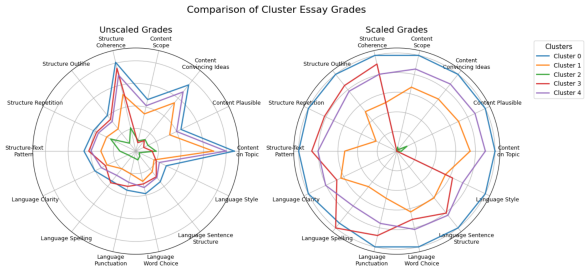


**Figure 9: Distribution of Chatbot Usage by Skill across Users**

We notice on Figure 9 from splitting in k=5 clusters that each cluster (but cluster 3 – which does not appear when k≤4) can be ordered to be better or worse than another, across all essay skills. This means that users are in general, i.e, when looking at a small number of clusters, distributed across their overall level at essays, rather than their level at an individual skill. This indicates low relevance of skills when looking at global essays performance.

| Cluster | Average math success | Average text success |
|---|---|---|
| 0 | 60.12% | 42.44% |
| 1 | 53.83% | 42.69% |
| 2 | 40.40% | 6.82% |
| 3 | 50.43% | 46.55% |
| 4 | 55.98% | 49.20% |

We also see that in most clusters students that perform better (respectively worse) at essays show better (respectively worse) average performance in math and text assignments. For k=5 we also observe the cluster 3 that scores very well except for content-related skills in essays as well as text questions, and is the first cluster to not respect the rule.

Students are well-divided by overall essay level, indicating that they learn(ed) these skills simultaneously.

## 4.3 RQ3 (K-Means Clustering Across All Subjects)

*Five-Cluster Solution.* With $k = 5$, the algorithm primarily segregates students by subject-coverage:

- *Cluster 3* consists almost exclusively of users who completed all three subjects.
- *Cluster 0*, *Cluster 2* and *Cluster 4* capture students who attempted only one or two subjects, respectively — their near-zero mean scores in the omitted domains reflect non-participation rather than poor performance.
- *Cluster 1* comprises students with virtually no recorded engagement — their total activity counts approach zero and their mean scores in all subjects are effectively zero. This "silent" cohort likely represents users

who signed up but never meaningfully interacted with the platform (e.g., brief account tests or immediate dropouts).

This coarse segmentation confirms the dominance of the "took-or-did-not-take" confounder and offers limited pedagogical insight beyond participation patterns.

*Eight-Cluster Solution.* Increasing to $k = 8$ reveals richer structure (Figure 10):

- **Math-only focused, low activity (Cluster 3):** These students submit comparatively few total events yet achieve high average scores in Mathematics, notably in light of other students. Two hypotheses arise: they may already possess strong math foundations and thus require less practice, or they are disproportionately motivated by mathematics and under-exposed to other subjects. A targeted recommendation is to encourage this cohort to diversify their practice: "You outperform 90% of your peers in Math-consider directing some effort to Language or Essay."
- **Essay-focused strivers (Cluster 4):** Analogously, a subset emerges whose activity is dominated by writing exercises, with minimal math engagement. While their essay scores improve steadily, their math and language scores lag. Adaptive prompts could nudge them toward a more balanced regimen: "Your persistence in writing is admirable–why not try a few math challenges today?"
- **Balanced high-achievers (Cluster 2) and mixed-ability generalists (Cluster 7):** Other clusters blend moderate activity with across-the-board performance, suggesting students who explore all subjects but vary in proficiency.
- **Math-focused very-high-activity strugglers (Cluster 5)**: This cohort exhibits an anomalously high total activity count — nearly three times the engagement of the math-only group — yet all interactions remain confined to the mathematics domain. Their disproportionate volume of practice, coupled with an absence of cross-subject exploration, suggests a subgroup of learners who are persistently grappling with mathematical concepts: they engage intensively, perhaps in search of mastery, but without diversifying their efforts. This pattern indicates the need for targeted scaffolding — such as adaptive hinting or reflective prompts — to transform their high-frequency attempts into deeper conceptual consolidation.

## 4.4 RQ4 (XGBoost Classification)

To interpret our XGBoost model, we used SHAP (SHapley Additive exPlanations). First, we built a SHAP explainer. We then computed SHAP values on the test set. A global SHAP summary bar plot ranked features by average absolute contribution, revealing which features the model relied on most. Next, a SHAP beeswarm (dot) plot illustrated

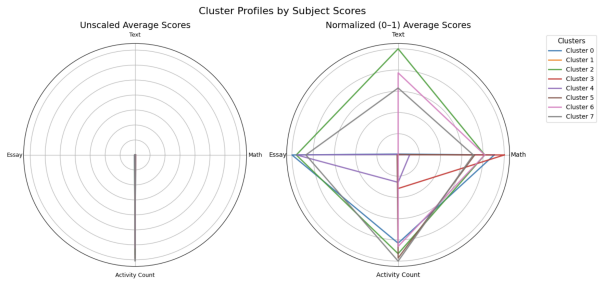


**Figure 10: Eight-Cluster Segmentation in the Space of per-Subject Mean Scores and Total Activity Count.**

how each feature's high or low value pushed individual predictions toward churn or non-churn; dots on the right side indicate "this feature drives churn," whereas dots on the left show "this feature pushes away from churn."

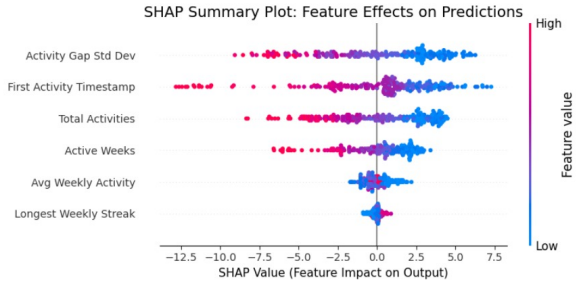


**Figure 11: SHAP Dot Plot**

We can also visualize the most important features, i.e. the ones that impact both churn and non-churn the most, by analyzing the mean absolute SHAP values.

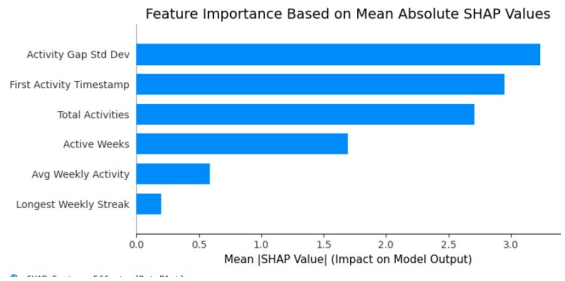


**Figure 12: Mean absolute SHAP values**

We deduce from Figure 5 that an early GoGymi first-use date is by far our best predictor for an engaged user, that is, a user that has used GoGymi at least once less than approx. a month before the exam date. Another feature that seem to predict user engagement is standard deviation of gaps between activities. Lower standard deviation in activity gaps indicates more regular usage, which is associated with higher engagement and a lower likelihood of churn.

# 5. DISCUSSION

Our integrated analysis of GoGymi's digital learning environment — spanning AI-tutor efficacy, writing-skill profil-

ing, churn-forecasting, and performance–engagement clustering — yields a constellation of insights, each tempered by data realities and methodological boundaries.

*RQ1: AI-Tutor Engagement and Mathematical Mastery.* We observe that the density of student–skill interactions is the dominant driver of Bayesian Knowledge Tracing accuracy: the richest pool of data, be it in the first five to ten opportunities for some (skill, AI usage group pairs), in the first 10 through 60 for some others, yields the most reliable mastery estimates; whereas beyond that window data become sparse and model confidence wanes. Within this dense region, only the *light-AI* cohort exhibits a continued, modest increase in predicted mastery for Arithmetic, Geometry, and Word Problems. Both *non-AI* and *heavy-AI* users show initial gains but plateau early, suggesting that moderate hinting provides sufficient error correction without inducing dependence. In contrast, heavy-AI students may over-rely on scaffolds, and non-AI learners may leave misconceptions unaddressed.

**Implications and Refinements:**

- Since observational volume is key, enriching log data — e.g. detailed hint types, time-to-response, or partial-credit attempts — could extend the reliable modeling horizon beyond what we have now, e.g. for Geometry.
- Alternative tracing frameworks (e.g. Deep Knowledge Tracing, Performance Factor Analysis) might better leverage sparse later-stage data and capture non-binary knowledge transitions.
- Experimentally varying the granularity and timing of AI feedback (for instance, delaying hints by one step to encourage reflection) could test the "Goldilocks" hypothesis of optimal scaffolding.

*RQ2: Essay-Skill Clustering and Cross-Domain Correlates.* Our PCA–K-means segmentation of fourteen rubric scores surfaces five coherent writing profiles out of which four central clusters (the ones with a relatively large number of users) whose Content, Structure, and Language dimensions rise and fall in concert. These clusters map strongly onto mathematics and reading quiz accuracy, affirming writing's cross-curricular value.

**Metadata Enrichment:** To deepen our understanding, we propose augmenting the clustering feature set with contextual metadata drawn directly from the essay logs, such as:

- *Prompt Characteristics*: thematic category (e.g. narrative vs. argumentative), estimated complexity or word-count target.
- *Submission Dynamics*: time of day or day of week when the essay was written, elapsed time between prompt presentation and submission.
- *Revision Metrics*: number of drafts saved, edits per session, total word-count change between first and final draft.
- *Engagement Signals*: count of per-essay specific Gymitrainer interactions (instead of global user_id-tied number of AI-Tutor interactions), scroll-depth or time spent on feedback screens.

By integrating these metadata, we can distinguish, for example, clusters that excel under high-stakes prompts from those that perform best in low-pressure contexts, or identify learners whose revision habits predict writing gains.

*RQ3: Clustering by Per-Subject Performance and Activity.* Our eight-cluster model transcends mere participation splits, revealing cohorts such as:

- **Math-only high-activity strugglers (Cluster 5):** logging nearly three times the normalized activity of typical math-only peers yet achieving only modest gains, indicating persistent conceptual hurdles.
- **Essay-focused practitioners (Cluster 7):** heavy writers whose domain-specific zeal leaves other subjects untended.
- **Balanced high-achievers and generalists:** learners with moderate activity and strong, cross-subject proficiency.
- **Silent registrations (Cluster 1):** near-zero activity and scores, highlighting accounts that never converted into active learning.

**Possible Enhancements:**

- Employ model-based clustering that explicitly handles missing subject scores to reduce bias from unattempted domains.
- Track cluster membership over time to detect transitions — do strugglers diversify, or remain siloed?
- Incorporate sequence-similarity measures (e.g. dynamic time warping on skill trajectories) to cluster by learning process rather than summary statistics alone.

*RQ4: Early Predictors of Disengagement.* Our XGBoost – SHAP analysis highlights two dominant churn predictors: early registration date and low variability in inter-action intervals. While these features signal trust and regularity, they may mask external confounders such as school schedules or concurrent tutoring.

**Next Steps:**

- Conduct A/B trials of gamified streak mechanics versus static calendars to causally assess their effect on usage consistency.
- Tailor reminder cadences — comparing just-in-time nudges to weekly digests — to identify the most effective re-engagement strategy.

- Complement quantitative signals with short surveys probing motivational drivers, thereby linking model predictions to student intentions.

# 6. LIMITATIONS

Our study is constrained by two central limitations. First, we lack access to students' actual Gymnasium entrance exam scores, preventing direct validation of our AI-usage, writing-cluster, and churn-prediction insights against real-world academic outcomes. Consequently, we cannot ascertain whether high-performing essay clusters truly translate into superior exam performance, or how early disengagement forecasts final results.

Second, sample size and data sparsity limit statistical power and generalizability. Many skills record few repeated opportunities per student, and the absence of granular logs (e.g. keystroke dynamics, partial solution attempts) restricts the depth of process analysis. Furthermore, our cohort — middle-school learners in Canton Zurich — may differ in motivation and curriculum focus from other populations.

To address these gaps, we envision:

- *Exam Linkage:* Secure partnerships to obtain anonymized exam score data, enabling end-to-end validation of model-driven interventions.
- *Data Enrichment:* Instrument the platform to capture finer-grained interaction traces (hint types, response latencies, draft revisions) and to log affective signals (self-reported confidence, frustration alerts).
- *Expanded Cohorts:* Deploy the analytical framework in diverse educational contexts — different regions, age groups, subject areas — to test robustness and uncover new profile archetypes.
- *Iterative Field Trials:* Integrate cluster-specific micro-interventions (e.g. targeted content pathways, gamified streak incentives, cross-domain prompts) and evaluate their efficacy through controlled experiments.

By confronting these limitations head-on and pursuing these future directions, we aim to refine GoGymi into a truly adaptive, evidence-based learning companion — one that not only measures progress, but actively stewards each learner toward enduring mastery and self-directed growth.

# 7. CONCLUSION

In weaving together quantitative tracings of AI-tutor engagement, nuanced profiles of writing competence, early signals of disengagement, and rich performance – activity typologies, this study illuminates the manifold pathways by which learners chart their own growth within GoGymi's digital milieu. We find that moderate guidance kindles sustained mathematical mastery, that writing subskills blossom in concert and echo across domains, that consistent rhythms of practice anchor commitment, and that finely grained clusters reveal both hidden strugglers and silent registrations. Though bounded by data sparsity and the absence of formal exam scores, our work nevertheless affirms a central tenet of education: learners flourish most when feedback is calibrated just so, when practice is both deliberate and diversified, and when insights from data science are woven seamlessly with pedagogical wisdom. By marrying analytical rigor with a commitment to each student's intrinsic agency, we lay a foundation for learning environments that honor the complexity of human growth and guide every individual along a personalized odyssey of discovery.

# 8. SUGGESTIONS TO GOGYMI

1. **Calibrated Gymitrainer Usage and Enhanced Instrumentation.** To encourage deeper cognitive engagement, GoGymi should introduce a controlled quota on Gymitrainer interactions — either as a daily cap or per-problem budget — thereby prompting students to formulate more deliberate, hypothesis-driven queries rather than relying on unlimited hints. Concurrently, the platform ought to enrich its telemetry: log fine-grained AI-interactions metadata (discussion content, precise start and end timing, problem resolution i.e. whether the problem was correctly solved after an obtained answer from Gymitrainer – rather than having to manually assign the content of the chat for example (which we did not do because of how time-intensive it would be), response latencies, keystroke dynamics, draft-revision histories to observe how students go about improving their own essays, and in-essay feedback interactions. This dual strategy both preserves the pedagogical benefits of AI scaffolding and supplies researchers with the contextual features needed to model learning processes more accurately and extend reliable knowledge-tracing beyond early practice windows.

2. **Automated Cohort Assignment and Adaptive Micro-Lessons.** Building on our clustering analyses, GoGymi should implement real-time cluster labeling for each student — dynamically assigning them to one of the empirically derived learner profiles — and surface tailored micro-lessons or prompts accordingly (e.g. grammar drills for writing-focused clusters, reflective math puzzles for high-activity strugglers). Embedding A/B-testing and feature-flagging capabilities will allow the team to evaluate the causal impact of these interventions, while a centralized analytics dashboard — complete with cluster-level and student-level visualizations — will accelerate iteration. Finally, forging partnerships to secure anonymized Gymnasium entrance exam scores will close the validation loop, ensuring that platform innovations translate into measurable academic success.

# 9. REFERENCES


[1] G. Abdelrahman, Q. Wang, and B. Nunes. Knowledge tracing: A survey. *ACM Comput. Surv.*, 55(11), Feb. 2023.

[2] A. Baillifard, M. Gabella, P. Banta Lavenex, and C. Martarelli. Effective learning with a personal ai tutor: A case study. *Education and Information Technologies*, 30:297–312, 07 2024.

[3] S. Kim, E. Choi, Y.-K. Jun, and S. Lee. Student dropout prediction for university with high precision and recall. *Applied Sciences*, 13(10), 2023.

[4] S. Wang. Analysis of learning behaviour of open education learners based on principal component analysis and k-means clustering algorithm. In *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, pages 1–7, 2025.