Behavior, Difficulty & Grading: What Drives Retention on Lernnavi?

Omar Boudarka
EPFL
omar.boudarka@epfl.ch

Valentine Casalta
EPFL
valentine.casalta@epfl.ch

Nino Gerber
EPFL
nino.gerber@epfl.ch

ABSTRACT

In recent years, learner engagement and retention have emerged as critical challenges in the context of online education platforms. This study investigates factors influencing user retention on Lernnavi, an adaptive digital learning system offering exercises in mathematics and German for Swiss secondary students. Leveraging two and a half years of platform activity data, we explore the predictive power of behavioral patterns, perceived and actual task difficulty, and evaluation fairness, particularly in open-ended responses.

We model next-week engagement using both interpretable behavioral features and difficulty signals, applying time-aware machine learning approaches including Gradient Boosting and Random Forest classifiers on accumulated features, as well as a Long Short-Term Memory (LSTM) model on weekly sequences. Behavioral features, especially effort-based metrics like time spent per week, proved most effective in predicting sustained platform use—a finding confirmed by tree-based models. While difficulty-based features alone offered little predictive power, their integration with behavioral inputs modestly improved performance in math-related tasks. Regression models further showed that difficulty metrics had limited ability to explain variance in engagement intensity. Our best-performing model—a holiday-aware LSTM using only effort-based features—achieved an AUC of 0.81 and an F1 score of 0.72, significantly outperforming traditional baselines. This result highlights the power of temporal modeling in capturing short-term engagement dynamics and underscores the central role of user effort in predicting retention on digital learning platforms.

Finally, we conducted a fairness audit of the platform's grading system using a multilingual semantic similarity model (SBERT) as an external oracle. Discrepancies between human and semantic evaluations revealed both limitations in the embedding model and potential systemic bias in platform grading logic.

These findings provide insights for improving adaptive learning platforms through targeted behavioral modeling and evaluation auditing, and underline the importance of aligning perceived difficulty and feedback to foster long-term student retention.

1. INTRODUCTION

In recent years, the rise of MOOCs and online learning platforms has drawn growing research interest into how learners interact with these new educational environments. A consistent challenge identified in the literature is the difficulty of maintaining learner engagement over time. Notably, studies estimate that between 40% and 80% of students drop out of online courses [1]. In a 2019 systematic literature review, Muljana and Luo highlighted several contributing factors to student retention in online learning, including program difficulty, learning facilitation, course design, learner behavior, and various demographic and personal characteristics [4].

Lernnavi is an online learning support system developed in collaboration with the Department of Education of the Canton of St. Gallen. It offers adaptive, curriculum-aligned exercises in German and mathematics for secondary school students. Through automated feedback, structured learning modes, and a personalized recommendation engine developed with EPFL, Lernnavi promotes sustained skill development and autonomous learning. Its integration of performance tracking and adaptive task assignment makes it a valuable context for studying learner engagement and dropout behavior in an applied educational setting.

This project leverages a dataset of Lernnavi user interactions, including demographics and behavioral activity logs. A preliminary analysis reveals that the platform exhibits similar retention issues, with users typically remaining active for only two to four weeks over a two-and-a-half-year period Figure1.

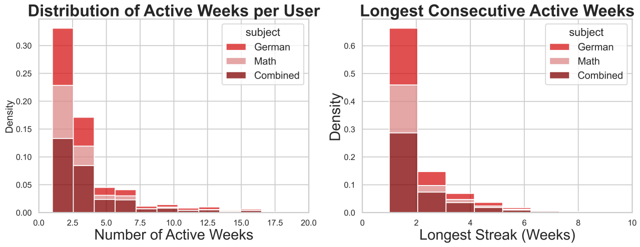


Figure 1: Distribution of user activity duration on Lernnavi

In light of these findings, this work investigates the impact of three core components on learner retention within the Lernnavi platform: (1) user behavior profiles, (2) the role of perceived and actual task difficulty, and (3) potential frustration linked to grading quality and bias, particularly in open-ended responses.

2. RELATED WORK

Our approach builds on prior work in learning analytics that emphasizes multi-dimensional modeling of student behavior. Mejia-Domenzain et al. [3] introduced an interpretable profiling framework across six dimensions of self-regulated learning, which we adapt to our weekly engagement context. Urrutia Cordero et al. [6] and Khalil et al. [2] have demonstrated the value of visualizing learner profiles and modeling future engagement, respectively. We extend these ideas by combining behavior profiling with next-week engagement forecasting.

3. METHODOLOGY

3.1 Models

3.1.1 How does user behavior influence retention rates on the platform?

Feature engineering To prepare the dataset for analysis, we aggregated event and transaction logs by calendar week, creating a weekly granularity for each user. Missing values were filled with zeros under the assumption that no recorded activity indicates genuine inactivity during that week.

Following the framework introduced by Mejia-Domenzain et al. [3], we organized student behavior into six interpretable dimensions of self-regulated learning summarized in 1. The features were tailored to reflect indicators of skill proficiency and engagement patterns relevant to our research question. Users with insufficient data—fewer than 10 recorded events or activity in less than two weeks—were excluded from further analysis.

Dimension	**Feature**	**Description**
3*Effort	Weekly events	Total number of platform actions
	Weekly clicks	Number of interaction clicks
	Time spent	Total minutes spent online
2*Consistency	Mean session duration	Avg. length of user sessions
	Std of activity	Std. dev. of weekly events
Regularity	Days between sessions	Avg. gap between study days
2*Proactivity	Go to theory	Visits to theory pages
	Early sessions	Sessions before 8 AM
2*Control	Next freq.	Frequency of clicking “next”
	Skip freq.	Frequency of skipped tasks
2*Assessment	% Correct	Share of correct answers
	Challenges done	Completed challenges

Table 1: Behavioral features grouped by self-regulated learning dimensions.

Engagement labeling To capture user engagement, we defined a set of *meaningful actions* on the platform: `SUBMIT_ANSWER`, `REVIEW_TASK`, `NEXT`, `SKIP`, and `GO_TO_THEORY`. These events reflect both task completion and active navigation behavior. For each user-week pair, we assigned a binary label, *engaged_next_week*, indicating whether the user performed at least one meaningful action in the following week ($t + 1$).

The labeling process used ISO weeks as a time index and was implemented via a self-join on the activity table, shifting the week forward for engagement target generation.

Behavioral profiling To enhance the interpretability of behavioral dynamics, we clustered users within each self-regulated learning dimension: Effort, Consistency, Regularity, Proactivity, Control, and Assessment, using K-Means ($k = 3$) applied separately to the respective weekly features. Clustering was conducted offline on training weeks only (up to week `2022-09`) to prevent label leakage. Features were standardized using `StandardScaler`, and centroids were ordered post hoc to produce ordinal profile labels: *Low*, *Mid*, and *High*. This profiling procedure was stored using `joblib` and repeated once per semester.

During feature generation, each week’s behavioral vector was transformed using the saved scalers and assigned to its nearest cluster centroid for each dimension. The resulting profile labels were one-hot encoded and appended to the main feature matrix X.

Modeling We used a time-aware cross-validation strategy (`TimeSeriesSplit`) to ensure realistic evaluation that respects the temporal nature of user behavior. We evaluated several classification models, each offering strengths well-suited to our engagement prediction task:

- **Logistic Regression**: Serves as a strong baseline and is useful for identifying linear trends between behavioral features and future engagement.
- **Random Forest**: Captures complex, non-linear relationships in user behavior and is robust to noise—ideal for heterogeneous activity logs.
- **Gradient Boosting**: Builds on previous mistakes to improve predictions, which is effective for subtle patterns in weekly engagement signals.
- **LightGBM**: Efficient for large datasets and high-dimensional behavioral features, making it suitable for rapid experimentation and tuning.
- **CatBoost**: Especially effective with categorical data and works well even with limited preprocessing—helpful for behavioral labels like profile clusters.

Ablation studies compared performance with and without behavioral profile features. Feature importance was analyzed using gain-based metrics and SHAP values to interpret model decisions and understand which behaviors most strongly predict future engagement.

Time Series Modeling: LSTM To complement our traditional machine learning models, we implemented a Long Short-Term Memory (LSTM) model to predict weekly engagement based on time-series features. As input features, we selected only the Effort dimension, as our traditional machine learning models had performed best when focusing solely on them. Each user’s activity timeline was segmented into sequences of 3, 4, 7, and 10 weeks, and the model predicted a binary engagement label for each time step.

The model outputs a probability for each week, indicating whether the user is likely to be engaged or not. It was trained using a 2-layer LSTM architecture (64 units per layer), with a dropout rate of 0.3 between the layers. We used binary cross-entropy loss and the Adam optimizer. To address class imbalance, sample weights were applied. Early stopping based on validation AUC was employed to prevent overfitting.

3.1.2 *How does difficulty influence retention rates on the platform?*

Feature engineering To estimate perceived difficulty, we combined response time and evaluation outcome into a single metric. Response times were log-transformed and standardized, then weighted by the evaluation result: correct answers received lower weights, while incorrect ones were penalized more heavily. The final score, computed as the product of normalized time and evaluation weight, captures both effort and performance, reflecting how difficult a task likely felt to the learner. To represent weekly difficulty exposure, we computed aggregated features per user-week. These include the mean and standard deviation of both actual and perceived difficulty ratings, as well as the average discrepancy between perceived and actual difficulty. This set of five features captures not only the general challenge level but also the variability and alignment in users' task experience.

Modeling To evaluate the contribution of difficulty features to engagement prediction, we trained models using the five aggregated difficulty-based features described above. We compared these to a second setup that included both difficulty and the behavioral features described previously. For both configurations, we trained four classifiers: Dummy, Logistic Regression, Random Forest, and Gradient Boosting, using a time-aware cross-validation strategy. To further assess the effect of difficulty on engagement intensity, we modeled weekly time spent as a continuous target using difficulty features only. For this regression task, we used both linear models (Linear Regression, Ridge, Lasso) and a non-linear model (Random Forest Regressor) to balance interpretability and flexibility. All models included standardized preprocessing pipelines. The same modeling procedures were then repeated separately for mathematics and german to investigate whether the impact of difficulty and behavior varied across subject domains.

3.1.3 *Are there biases induced by the evaluation of open ended questions ?*

Data selection To evaluate the alignment between system-assigned evaluations and semantic similarity, we focused exclusively on open-ended responses in German. We filtered for valid `openInput` answers, ensuring clean JSON formatting and non-empty user input. Only active users were retained, based on prior engagement filtering, to ensure comparability with earlier analyses. We employ:

- **Sentence-BERT (SBERT)** [5]: We use the `distiluse-base-multilingual-cased-v2` model to generate 768-dimensional embeddings for each student answer and its canonical solution. SBERT is chosen for its strong performance on semantic similarity tasks and multilingual capability, which suits German text embeddings without extensive fine-tuning.
- **Cosine Similarity Oracle:** For each answer–solution pair, we compute $\mathrm{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$. We then discretize similarity into three labels:
  - `CORRECT` if $\mathrm{sim} \geq 0.80$,
  - `PARTIAL` if $0.50 \leq \mathrm{sim} < 0.80$,
  - `WRONG` if $\mathrm{sim} < 0.50$.

  These thresholds were chosen via initial ablation: varying $\tau_1 \in [0.75, 0.85]$ and $\tau_2 \in [0.45, 0.55]$ and selecting the pair that minimized overall oracle–system disagreement.
- **Clustering with KMeans:** We compute feature vectors by concatenating (1) the elementwise difference of SBERT embeddings ($\mathbf{e}_{\mathrm{ans}} - \mathbf{e}_{\mathrm{sol}}$), (2) the scalar cosine similarity, and (3) a binary `error_flag` (1 if oracle label $\neq$ system label, 0 otherwise). We then reduce these to 2D via PCA for visualization and apply KMeans ($k = 4$ via Elbow analysis) to identify clusters of high/low error rates.

3.2 Evaluation

3.2.1 *RQ1: Behavior and Retention*

Classification models predicting user engagement were evaluated using two standard metrics: ROC-AUC and F1 score. ROC-AUC measures the model's ability to distinguish between students who will remain active and those who will not, regardless of classification threshold, and is robust to class imbalance. The F1 score provides a threshold-dependent summary that balances precision and recall, offering practical insight into actionable retention predictions.

To ensure robust performance estimation, we used a 5-fold `TimeSeriesSplit`, preserving temporal order and avoiding data leakage. All features were standardized using `StandardScaler` within each pipeline. Comparisons were made between models trained on behavioral features alone, and models augmented with behavioral profile clusters derived from K-Means.

3.2.2 *RQ2: Difficulty and Retention / Time Spent*

We evaluated two tasks: (i) predicting next-week engagement as a binary classification task, and (ii) predicting weekly time spent as a regression task.

For classification, we again used ROC-AUC and F1 score, applied to models trained with difficulty-only features and those with combined difficulty plus behavioral inputs. Time-aware cross-validation ensured future weeks were predicted based on past activity, simulating a real deployment scenario.

For regression, we modeled weekly time spent using only difficulty features. We used R^2 to quantify variance explained, and RMSE to evaluate average error magnitude in seconds. Linear and non-linear models were evaluated using the same 5-fold temporal split. To explore domain-specific effects, both classification and regression analyses were repeated for Math and German subsets.

3.2.3 *RQ3: Bias in Grading of Open-ended Questions*

We evaluate based on:

1. **Error Flag Rate:** For each record, $\mathrm{error\_flag} = \big[\mathrm{oracle\_label} \neq \mathrm{sys\_label}\big]$. The cluster-level err_rate is the mean of those flags within each cluster.

2. **Gender Distribution:** We cross-tabulate error_flag and cluster assignments against self-reported gender (`FEMALE`, `MALE`, `STAR`). Any significant disparity suggests potential gender bias.

3. **Topic Error Rates:** We group by topic label (e.g., "Genus verbi/Active passive", "Interjections") to compute err_rate(t) $\frac{\sum[\text{error\_flag}=1]}{\text{count of topic } t}$. A high error rate per topic indicates potential oracle–rubric mismatch.

4. **Statistical Significance:** For topic vs. error flag contingency, we perform a chi-square test:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed count, E_{ij} the expected under independence. A $p < 0.05$ indicates that error rates vary significantly by topic.

4. RESULTS

4.1 How does user behavior influence retention rates on the platform?

Across all models, tree-based methods consistently outperformed baselines. As shown in Table 2, the best-performing model, Gradient Boosting without profile features, achieved an ROC-AUC of 0.752 and an F1 score of 0.55. Including profile features did not lead to a significant performance improvement, yielding an ROC-AUC of 0.751 and an F1 score of 0.54. However, the addition of profile features did not degrade performance and provided interpretable groupings of learner behaviors. Performance metrics remained stable across cross-validation folds, suggesting that engagement patterns are consistent over time and generalizable.

Model	Variant	AUC	F1 Score
Gradient Boosting	no_profile	**0.7520**	0.5506
Gradient Boosting	with_profile	0.7515	0.5452
Random Forest	no_profile	0.7498	0.5568
Random Forest	with_profile	0.7488	0.5558
CatBoost	no_profile	0.7473	0.5646
CatBoost	with_profile	0.7472	0.5627
LightGBM	with_profile	0.7387	**0.5786**
LightGBM	no_profile	0.7378	**0.5786**
Logistic Regression	no_profile	0.5793	0.4714
Logistic Regression	with_profile	0.5728	0.4661
Dummy	with_profile	0.5000	0.0000
Dummy	no_profile	0.5000	0.0000

Table 2: Performance comparison of models with and without profile features (best scores in bold).

From Figure 2 it appears that features from the *Effort* dimension emerged as dominant predictors. Notably, the feature `weekly_time_spent` alone accounted for approximately 80% of the model's total gain-based feature importance, indicating that it played a central role in decision-making across the ensemble of trees. Furthermore, the top eight features all originated from the Effort, Consistency, or Regularity dimensions.

These results support our hypothesis that student engagement can be effectively predicted using behavioral data, particularly features that capture effort and consistency over time. The stability of performance across time-based folds, and the dominance of effort-related predictors, suggest that retention is closely tied to regular and sustained platform usage. While the inclusion of profile labels did not significantly enhance predictive accuracy, it did offer interpretability that could support more personalized interventions.

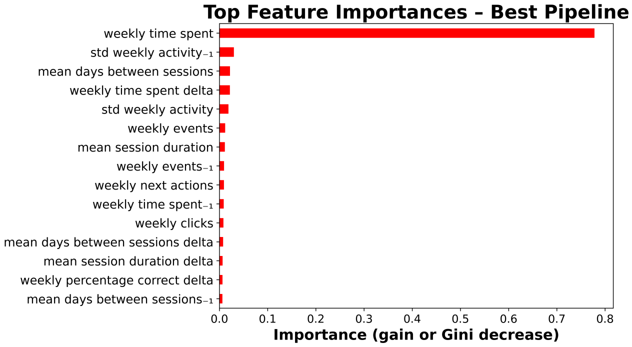


Figure 2: Top 15 features from the no-profile gradient boosting model.

Impact of Holidays on Engagement To investigate external factors influencing retention, we overlaid platform activity with a calendar of Swiss school holidays and the user activity (Figure 3). A clear pattern emerged: platform usage dropped significantly during official school holidays. This suggests that temporal externalities, particularly school calendars, must be accounted for in predictive modeling.

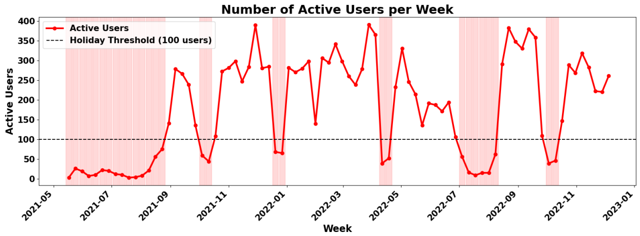


Figure 3: Weekly engagement patterns overlaid with holidays.

LSTM Sequence Model Results To complement tree-based models, we trained an LSTM sequence model to predict weekly engagement based on the last n weeks of user activity. Due to its capacity to capture temporal dependencies, the LSTM model outperforms our best gradient boosting baseline in terms of both F1 score and AUC, particularly when enhanced with holiday-aware features.

As shown in Table 3, shorter input sequences ($n = 3$ or 4) yielded the highest performance, with the holiday-aware LSTM achieving an F1 score of 0.72, compared to 0.55 for the baseline. Longer sequences, on the other hand, led to a noticeable decline in performance. This is likely due to label dilution and the inclusion of irrelevant time steps, particularly around school holiday periods, which introduce irregular usage patterns that may confuse the model. By explicitly incorporating holiday information, the model learns to adjust for these disruptions and better capture meaningful trends.

These results suggest that incorporating temporal context and calendar signals such as school holidays meaningfully improves sequence-based retention prediction.

Our best model, a holiday-aware LSTM using only effort-based features, achieved an AUC of 0.81 and an F1 score of 0.72, offering strong predictive performance despite noisy data. This provides compelling evidence that user effort alone is a highly informative signal for modeling retention behavior.

Model	**Seq Length (n)**	**AUC**	**F1 Score**
Gradient Boost (baseline)	–	0.752	0.55
LSTM	3	0.80	0.69
LSTM	4	0.79	0.64
LSTM	7	0.76	0.59
LSTM	10	0.74	0.52
LSTM (holiday-aware)	**3**	**0.81**	**0.72**
LSTM (holiday-aware)	4	0.81	0.65
LSTM (holiday-aware)	7	0.81	0.61
LSTM (holiday-aware)	10	0.82	0.58

Table 3: Comparison between traditional gradient boosting and LSTM models.

4.2 How does difficulty influence retention rates on the platform?

In order to better understand learners' perception of task difficulty, we compared perceived vs. actual difficulty ratings across subjects. As displayed in Figure 4, in both German and Math, the majority of tasks were perceived as easier than they were rated by the system. This effect was especially pronounced in German, where nearly 80% of tasks were perceived as easier. Very few tasks were perceived as harder than their computed difficulty, suggesting that learners rarely overestimate challenge. These patterns motivate the inclusion of perceived difficulty as a potential predictive signal for engagement modeling.

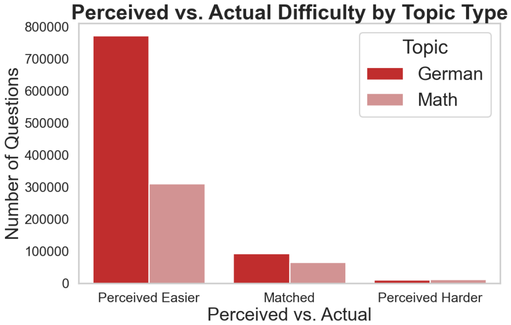


Figure 4: Comparison between actual and perceived difficulty of exercises per topic

To isolate the effect of difficulty on learner retention, we first trained models using only aggregated difficulty features. Across all classifiers, performance remained close to chance level, with the best model (Logistic Regression) achieving an AUC of 0.55 and an F1 score of 0.50. This suggests that difficulty signals alone are insufficient to reliably distinguish between retained and disengaged users.

Adding behavioral features described above led to a marked improvement in predictive performance. The Random Forest classifier reached an AUC of 0.73 and an F1 of 0.56, indicating that behavioral signals capture more actionable variance in user engagement. However, these performences remained lower than the ones obtained using behavioral features only as described in section 4.1. This difference suggests that including difficulty features might actually induce the model in error. Further analysis, such as feature importance analysis (Figure 5) revealed that weekly time spent was once again the dominant factor in both subjects. Perceived difficulty variability still proved to be a valuable indicator of retention regarding math topics.

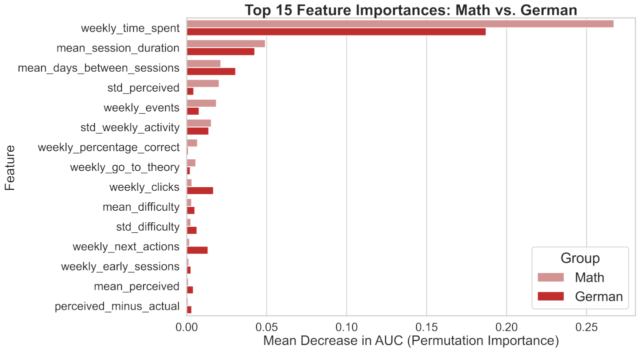


Figure 5: Enter Caption

To further examine the explanatory power of difficulty, we modeled weekly time spent as a continuous variable using only difficulty-related inputs. All regression models performed poorly, with near-zero R^2 values, confirming that difficulty features alone do not capture the factors driving time-on-platform.

When stratifying by subject, we observed comparable patterns. For both math and German, models using only difficulty features performed near baseline (AUCs $\simeq 0.54$). Including behavioral data led to strong improvements (Table 4), particularly in the math group, where the Random Forest reached an AUC of 0.77 and F1 of 0.62, this time improving predicting capabilities compared to section 4.1. Feature rankings were consistent across subjects, underscoring the central role of behavioral engagement in predicting retention.

Model	**Math**		**German**	
	AUC	**F1**	**AUC**	**F1**
Random Forest	**0.773**	**0.625**	0.767	**0.660**
Gradient Boosting	0.772	0.614	**0.769**	0.655
Logistic Regression	0.620	0.574	0.611	0.631
Dummy Classifier	0.500	0.000	0.500	0.000

Table 4: Performance of different models on engagement prediction, split by subject (Math vs. German). Scores are averaged over 5-fold time-based cross-validation.

4.3 Are there biases induced by the evaluation of open ended questions ?

4.3.1 Clustering and Gender Skew

After K-Means ($k = 4$), we obtain the following summary:

Cluster	Count	Error Rate	Dominant Gender (F/M/S)
0	541	1.00	298 / 190 / 23
1	1587	0.00	867 / 583 / 54
2	217	0.00	125 / 82 / 5
3	1235	1.00	740 / 435 / 23

Interpretation: Clusters 1 and 2 exhibit *zero error rate* (system's labels fully align with the cosine oracle), whereas Clusters 0 and 3 have *100% error rate* (complete mismatch). A scatterplot in PCA space (Figure 6) reveals two well-separated error clusters (0,3) on the right, and two well-separated agreement clusters (1,2) on the left. Gender proportions within each cluster (see Table above) remain roughly consistent (F > M > S) across clusters, indicating **no significant gender skew** in misclassification patterns.

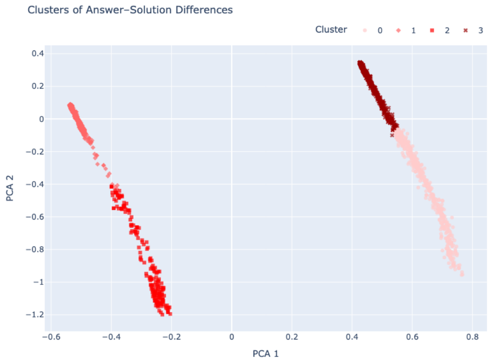


Figure 6: PCA Visualization of Answer–Solution Feature Vectors, colored by cluster. Clusters 0 and 3 correspond to 100% error, while 1 and 2 correspond to 0% error.

4.3.2 Topic-Based Error Rates

Table 5 ranks topics by their error rate (limited to those with ≥ 50 samples):

Topic	Count	Error Rate
Interjections	102	0.9412
Transferred meaning	357	0.7759
Main and subordinate clauses	94	0.5957
Determination of sentence elements	280	0.4893
Genus verbi/Active passive	2152	0.4675
Relationship clauses & subclauses	166	0.3313
Pronominal references	390	0.3103

Table 5: Top Topics by Error Rate (system $\neq$ cosine-oracle).

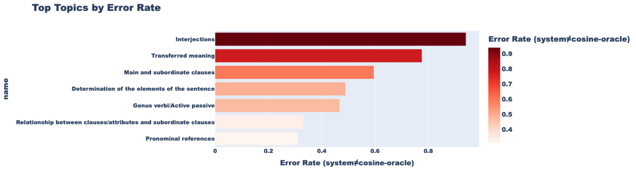


Figure 7: Top Topics by Error Rate (≥ 50 samples). 'Interjections' and 'Transferred meaning' show the highest mismatch.

Chi-Square Test: The test of independence between `topic` and `error_flag` yields $\chi^2 = 285.70,\ p < 10^{-50}$, indicating a highly significant association: certain topics are *more prone* to grading mismatches.

Topic vs. Cluster Proportions: Figure 8 (heatmap) reveals:

- *Cluster 0 (100% error)* is dominated by *Transferred meaning*, *Interjections*, and some *Active passive* items.
- *Cluster 1 (0% error)* contains nearly *all Effect of style levels* and *Rhetorical devices* (which were excluded above due to low sample count but have 100% agreement).
- *Cluster 3 (100% error)* is heavily *Genus verbi/Active passive*, suggesting that semantic similarity alone cannot capture German voice distinctions when the intended answer hinges on word order or morphological marking.
- *Cluster 2 (0% error)* successfully captures simpler topics like *Pronominal references.*

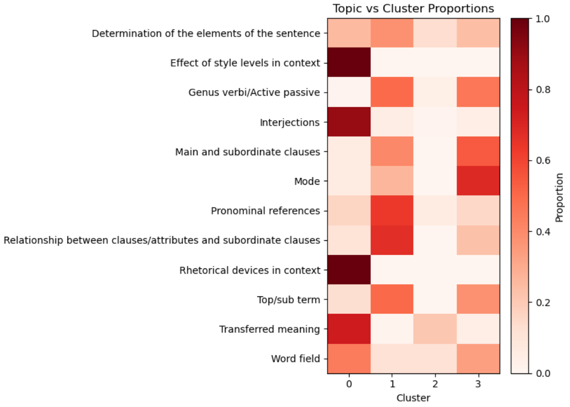


Figure 8: Normalized Proportion of Each Topic Across Clusters. Darker colors indicate higher representation.

4.3.3 Violin Plot Analysis by Topic

Figure 12 (see Apendix) displays *cosine_sim* distributions stratified by `sys_label` for three exemplar topics: *Transferred meaning*, *Interjections*, *Genus verbi/Active passive.*

Key Observations:

- *Transferred meaning*: The `WRONG` label cluster has a non-negligible density at high similarity ($\sim$ 0.4–0.6), indicating semantic synonyms that the system treated as incorrect.
- *Interjections*: `WRONG`-labeled samples show extremely low *_sim* (near 0), yet some `CORRECT` samples also appear at low similarity—implying the system's rubric deems certain emotional statements correct even if embeddings diverge semantically.
- *Genus verbi/Active passive*: Cosine similarity is uniformly high (> 0.8) across `CORRECT`, `PARTIAL`, and even some `WRONG` samples, highlighting that sentence-level semantic embeddings cannot capture German voice distinctions.

5. DISCUSSION

Our findings offer several practical takeaways for educational platforms aiming to improve student retention and learning outcomes.

First, our results confirm that effort and regularity are the most influential behavioral predictors of future engagement. In particular, the feature weekly_time_spent consistently dominated model importance scores, accounting for the majority of gain-based importance across all model variants. This highlights the central role of sustained time investment and repeated platform usage in fostering learner retention.

Given that Lernnavi is embedded in the school context, these findings suggest that promoting regular in-class sessions or structured homework using the platform could be a valuable lever. Such scaffolding might help cultivate habits that extend beyond class, encouraging students to revisit content independently—especially for revision or exam preparation.

This hypothesis is supported by our observation of sharp drops in activity during school holidays, suggesting that absent external reinforcement, learners are unlikely to engage voluntarily. From a product and policy standpoint, integrating Lernnavi more formally into classroom routines could serve as a behavioral anchor, nudging students toward more autonomous learning outside school hours.

While difficulty-based features alone yielded poor predictive performance, our exploratory analysis revealed that most students perceived tasks as easier than their assigned difficulty levels. This could indicate that learners are completing exercises too quickly or effortlessly, potentially reducing their perceived value. Introducing targeted challenge—via adaptive item selection or spaced retrieval of harder questions—might help increase engagement duration and depth.

When we stratified results by subject, important differences emerged between math and German. In math, perceived difficulty variation (std_perceived) ranked among the top predictive features, while in German, behavioral dimensions such as effort, consistency, and regularity fully dominated. These patterns support the idea that retention mechanisms are domain-dependent, with cognitive challenge playing a greater role in structured subjects like math, and behavioral consistency mattering more in linguistically complex fields like German.

Finally, our analysis of automated grading in open-ended questions revealed nuanced failure modes in Lernnavi's evaluation pipeline. Although we found no evidence of gender bias, several linguistic phenomena caused consistent misclassifications:

Passive voice constructions, although semantically equivalent to their active counterparts, were often marked as incorrect.

Interjections and idiomatic expressions led to wide semantic variation despite meeting the rubric's intent, causing misalignment between semantic similarity and correctness labels.

These issues underscore the limitations of relying solely on semantic embeddings, particularly for morphosyntactic tasks. To improve grading accuracy, especially in language-based topics, future versions of the evaluation system should integrate syntactic cues—such as part-of-speech tags or dependency parses—into their decision-making. Moreover, our results lend support to the Lernnavi development team's ongoing efforts to restructure the grading module with LLM-based architectures, which may better capture the nuanced interplay between meaning and form in open-ended responses.

Together, these findings provide a roadmap for improving adaptive educational systems: emphasize behavioral engagement, tailor difficulty by subject, and modernize evaluation techniques for open-ended tasks.

6. LIMITATIONS

While our findings provide valuable insights into learner engagement and evaluation quality on Lernnavi, several limitations must be acknowledged.

First, the scope of the dataset is relatively constrained, both in size and temporal coverage, which limits the generalizability of our conclusions to broader populations or longer-term usage. Although our analysis clearly shows that engagement patterns are influenced by school holidays, these external factors were not formally modeled beyond binary holiday indicators. Future work could benefit from more granular integration of academic calendars to differentiate between structural inactivity and genuine disengagement. Additionally, we do not account for how Lernnavi is embedded in classroom instruction—whether it is used sporadically in class or systematically encouraged for independent study—which could introduce substantial variation in user behavior and retention outcomes.

A further limitation lies in our analysis of grading consistency for open-ended questions. Our approach uses SBERT-based cosine similarity as a semantic oracle to evaluate the alignment between student answers and canonical solutions. While this offers a scalable and language-agnostic proxy for correctness, it falls short of capturing grammatical accuracy and morphosyntactic detail—particularly in German. For instance, semantically equivalent responses that differ in voice or sentence structure may receive high similarity scores despite being grammatically incorrect. Conversely, creative or idiomatic responses may be penalized due to low surface-level similarity. Moreover, the thresholds used to assign correctness labels based on similarity (≥ 0.80 for correct, < 0.50 for wrong) are fixed and heuristic, potentially misclassifying edge cases. Several topics also had insufficient sample sizes to support detailed bias analysis, leaving possible gaps in coverage. Future work should consider extending our semantic evaluation pipeline by incorporating syntactic features, fine-tuning embedding models on domain-specific data, and dynamically calibrating similarity thresholds using gold-standard annotations. Expanding data collection to underrepresented topics would also enable a more robust assessment of grading fairness across all content areas.

7. CONCLUSION

This study presents an exploration of students retention dynamics on Lernnavi, integrating behavioral profiling, difficulty modeling, and grading fairness analysis. Our results consistently highlight user's efforts, particularly weekly time spent, as the strongest predictor of future engagement, with temporal models like LSTM further improving performance when accounting for holiday disruptions. While difficulty features alone had limited explanatory power, their variance proved informative for math-related retention when combined with behavioral signals. Importantly, our audit of open-ended grading revealed that semantic embeddings can misalign with rubric-based correctness, especially in morphosyntactically sensitive contexts like German. These findings underscore the value of combining interpretable behavioral models with domain-specific refinements in both difficulty tracking and evaluation logic to support more effective and equitable learning platforms.

8. REFERENCES

[1] P. Bawa. Retention in online courses: Exploring issues and solutions—a literature review. *SAGE Open*, 6(1):2158244015621777, 2016.

[2] M. Khalil, D. Urrutia Cordero, and M. A. Chatti. Student answer forecasting: Transformer-driven answer choice prediction for next-week engagement. *arXiv preprint arXiv:2405.20079*, 2024.

[3] P. Mejia-Domenzain, M. Marras, C. Giang, and T. Käser. Identifying and comparing multi-dimensional student profiles across flipped classrooms. *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*, 2022.

[4] P. S. Muljana and T. Luo. Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research*, 18:19–57, 2019.

[5] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[6] D. Urrutia Cordero, M. Scheffel, E. Ternieden, R. Pursian, and M. A. Chatti. Visualizing self-regulated learner profiles in dashboards: Design insights from teachers. *arXiv preprint arXiv:2305.16851*, 2023.

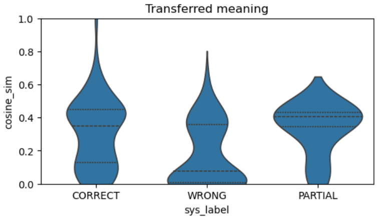
APPENDIX

A. SUGGESTIONS

To better support future research and improve modeling of user retention, we recommend collecting contextual metadata about how Lernnavi is used in school environments. In particular, it would be valuable to know whether usage is teacher-driven or voluntary, and if the tool is assigned for a one-off session versus integrated consistently throughout the school year. This distinction could shed light on how pedagogical framing affects engagement: for example, students may be more likely to return to Lernnavi on their own if they perceive it as aligned with classroom instruction or exam preparation. Logging this usage context, either through teacher-facing assignment settings or optional metadata tags, would significantly improve the interpretability of retention patterns.

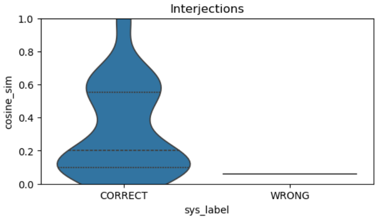
Additionally, integrating lightweight student feedback mechanisms would help quantify subjective experiences of the platform. After completing a session, students could briefly rate how difficult they found the material and whether they felt the grading was fair, especially for open-ended questions. This would provide insight into learners' frustration thresholds and perceived relevance. Over time, such feedback could help calibrate both adaptive difficulty settings and evaluation algorithms, reinforcing trust and fostering a more personalized learning experience.

B. ADDITIONAL FIGURES



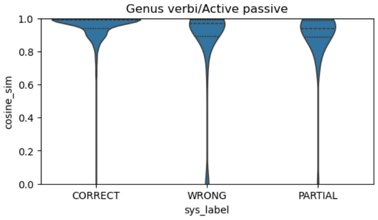
[b]

Figure 9: Transferred meaning



[b]

Figure 10: Interjections



[b]

Figure 11: Genus verbi/Active passive

Figure 12: Violin plots of cosine similarity by system label. Dashed lines denote quartiles.