WP1: Gathering Learning Trace Data

Description:

In order to train our tutor Olé with Federated Reinforcement Learning, we need to first collect a dataset of user learning traces from a personalized learning environment. The collection of quality user preference data for training is a crucial factor in determining whether our model research advances are successful. Therefore, this work package will focus on the development of the learning environment (Scholé) and actively log users' interaction with the system including timestamped clickstream and text data (e.g., scrolling behavior, generated learning paths, requested modifications to learning paths, answers to the questions posed, engagement level on the current learning content, and questions asked to the tutor) in the form of implicit and explicit feedback.

Our initial focus is to test Scholé with adult learners in the workplace given that corporate learners have concrete use cases for data science learning aligned in their job contexts, as well as diverse skill levels and specific learning goals tied to their roles. Additionally, they have limited time for training, making them an ideal audience to evaluate the tutor's ability to deliver flexible durations of adaptive learning experiences with bite-sized blocks of content. We will start by exploring Responsible AI practices in the workplace.

The focus on Responsible AI practices (e.g. algorithmic bias, interpreting data visualizations, transparent AI) was chosen given its relevance for the European Union Artificial Intelligence Act published in the Official Journal of the European Union stating that all providers and deployers of AI systems need to ensure, to their best extent, a sufficient level of AI literacy among staff dealing with the operation and use of AI systems. We will design this system in a generalizable manner so that the training content and skill aptitude adaptation can be easily extensible to other domains and student populations.

Activities:

1. Develop the base learning platform that will host the tutor for extracting learning traces.
    a. We will run co-design workshops to iterate on the design and functionalities of the front-end using the existing wireframes.
    b. The frontend will be implemented using React, and the backend will use FastAPI, reusing some functionalities from our already implemented platform GELEX [5].
    c. We will host the application using credits from Microsoft Azure (we have been awarded 10k credits as a recipient of the Microsoft for Startup grants).
    d. The platform will have an interface to display bite-sized blocks of content and checkpoint questions extracted by Olé, our learning tutor. The content can be in the form of text (questions or reading), audio (podcasts) or video (lecture content).
2. Build the knowledge graph of learning materials

    a. For the initial scenario of “Responsible AI practices in the workplace”, we will build a knowledge graph of data science and artificial intelligence concepts where the nodes (e.g., “Model Bias”) will represent the topics and the links point to the prerequisite for the nodes (e.g., “Training Data”, “Analyzing Model Performance”). Each node will contain open-source data science training content (text, audio, video) at different levels of complexity. We will use the machine learning for behavioral data course materials (co-created by Paola and Vinitra) offered by the ML4ED lab at EPFL as a basis for our initial graph. We also have permission to use the University of Washington Intro to Machine Learning course materials and the UC Berkeley Foundations of Data Science tutoring materials for our graph.
    b. We will evaluate the knowledge graph with expert data science teaching professors in the field (e.g., Data Science professors from EPFL, Berkeley, University of Michigan, UCSD, and Monash University have already agreed to help us validate the knowledge graph). We will measure the node validity (verifying that every node represents a valid and meaningful entity in data science and AI) and the edge validity (verifying that the edges, prerequisite links, accurately represent logical and educational dependencies).
3. Integrate a pretrained baseline LLM to extract the curriculum path and edit examples in learning materials to be relevant to a users’ context. We will refer to this model as simple-Olé.
    a. Using an LLM (simple-Olé), extract the user’s context (e.g., work context and position). We intend to use a fine-tuned tutoring model LLama 3.1 70B (Ethel Tutor) based on TutorChat data, following the same training procedure as was conducted by our team in the Swiss AI Education vertical.
    b. Given the user’s input, we will recommend a learning path that is suitable for the user’s context. The learning path will be the recommended subgraph incorporating both the starting and the final module.
    c. An LLM will be used to adapt the content to the user’s context (e.g., tailor the exercises to the legal, finance, healthcare, or sports retail domain). It will also be used to edit follow-up MCQ or short response questions from existing learning materials into relevant examples to test a user’s understanding. As the examples are simply edited and not generated from scratch, we greatly minimize the risk of hallucinations.
    d. We will update the generated questions dynamically. Given a student data interaction stream, a simple knowledge tracing algorithm [6] utilizing our pre-written code from the PyKT library (our KT pipelines are available at https://github.com/epfl-ml4ed/mlbd-2023/tree/main/lectures/week-06) will identify the pace of the user's progress. If the progress is estimated to be slower than average (less than 0.5 for a node), the previous prerequisite node will be retrieved and added to the curriculum. If the user is advancing fast (higher than 0.75), the curriculum might skip some nodes and adapt the difficulty of the content and questions.

4. Run pilot programs with organizations who wrote letters of intent (see attached), and on Prolific (a crowdsourcing platform) to obtain user learning traces from the platform while logging all interaction activity (clickstream and text).
    a. Allow modification of the learning path: Via a user drag-and-drop interface or via a conversation with the LLM, the user will be able to give feedback on the proposed learning path and modify it into a learning path they prefer (i.e. changing from video to audio or questions, deleting or reordering modules). These changes, and reflections on why they were made, will be tracked as explicit feedback.
    b. We commit to only saving data from users who consent to using our platform. The data will be logged in a Postgres database in the backend, and the interaction data will be stored in an anonymized way on protected servers at EPFL.
    c. We will collect two types of user feedback in the learning traces. Explicit feedback: A timestream of user ratings on module effectiveness, approval of suggested content changes, or reordering of course material by Scholé users. Implicit feedback: Metrics such as time spent on a module, skipped exercises, and post-training performance scores. These signals will provide a rich dataset for understanding what works in different learning contexts.

[1] Chui, K. T., Lee, L. K., Wang, F. L., Cheung, S. K., & Wong, L. P. (2023, November). A Review of Data Augmentation and Data Generation Using Artificial Intelligence in Education. In International Conference on Technology in Education (pp. 242-253). Singapore: Springer Nature Singapore.
[2] Zouleikha, B., Aida, C., & Samia, D. (2024, September). GAN-Based Data Augmentation for Learning Behavior Analysis in MOOCs. In Novel & Intelligent Digital Systems Conferences (pp. 632-638). Cham: Springer Nature Switzerland.
[3] Moreno, Y., Montero, A., Hidrobo, F., & Infante, S. (2023, June). Synthetic Data Generator for an E-Learning Platform in a Big Data Environment. In International Conference in Information Technology and Education (pp. 431-440). Singapore: Springer Nature Singapore.
[4] Radmehr, B., Singla, A., & Käser, T. (2024). “Towards Generalizable Agents in Text-Based Educational Environments: A Study of Integrating RL with LLMs” EDM 2024.
[5] Yazici, Aybars, et al. "GELEX: Generative AI-Hybrid System for Example-Based Learning" CHI Conference on Human Factors in Computing Systems. 2024.
[6] Pavlik, Philip I., Hao Cen, and Kenneth R. Koedinger. "Performance factors analysis–a new alternative to knowledge tracing." Artificial intelligence in education. Ios Press, 2009.
[7] Liu, Zitao, et al. "pyKT: a python library to benchmark deep learning based knowledge tracing models" Advances in Neural Information Processing Systems 35 (2022): 18542-18555.

Milestone 1 - Learning Traces from 100 Users Extracted

G1-1. A responsive front-end and working back-end. We will run usability tests and iterate on the platform until we achieve a System Usability Score from two HCI experts above 69 (higher than the average usability score). We will extract on average 10 traces from at least 100 users, leading to a rich dataset of at least 1000 traces. A trace is defined as a set of interactions between 30 seconds to 5 minutes that can be assigned to a purpose (i.e. conversing with Olé to set learning goals, modifying the proposed curriculum, consuming learning content, answering checkpoint questions).

G1-2. A verified knowledge graph (KG). We will verify the KG with human experts. We expect a substantial (0.61-0.80) inter-rater reliability measured by Cohen's kappa. We aim for a node validity (verifying that every node represents a valid and meaningful entity in data science and AI) higher than 90% and an edge validity (verifying that the edges, and prerequisite links, accurately represent logical and educational dependencies) higher than 90%.

G1-3. Olé will extract the relevant user context (e.g., job roles, existing skill levels, and learning objectives). Success will be evaluated through user satisfaction surveys conducted during pilot programs to assess the relevance and accuracy of the extracted context, aiming for an accuracy higher than 80%.

G1-4. Olé will recommend personalized knowledge paths based on user input and context. This will involve dynamically generating a subgraph from the data science knowledge graph. Success will be evaluated through user satisfaction surveys conducted during pilot programs, aiming for at least 80% positive feedback on the relevance and clarity of the recommended learning paths.

G1-5. The platform will generate tailored exercises, explanations, and follow-up questions using Olé. Adaptive mechanisms will dynamically adjust the difficulty and progression based on user performance and interaction data. Success will be evaluated through user satisfaction surveys conducted during pilot programs, targeting at least 80% user satisfaction with content personalization and adaptive difficulty.

G1-6. During the pilot programs, the platform will log all user interaction data. Success will be measured by achieving a data completeness rate of at least 95% across all user interactions.

WP2: Federated Reinforcement Learning from Human Feedback (FedRLHF) for Olé

Description:

Building on the dataset of learning traces we collected in WP1, WP2 focuses on the development, implementation, and validation of federated reinforcement learning with human feedback (FedRLHF) pipelines to power Olé [1, 2, 3]. The goal is to enable Olé to learn and adapt its teaching strategies real-time, continuously improving through reinforcement learning with LLM memories and using federated learning for scalability and privacy.

This work package has the bulk of the technical innovation of our proposal, and therefore the risk in our implementation. To our knowledge, FedRLHF approaches have never before been attempted for the education domain, and have not been implemented in a real-world environment. They have been proven theoretically and tested in benchmark settings (e.g. IMDB Movie Dataset or Movie Lens), but not at real-world scale nor with a data environment that does not provide direct preference feedback and instead a timestream of implicit and explicit signals [3, 4, 5]. Additionally, FedRLHF has never been integrated with LLM memory for short-term system adaptation in conjunction with periodic offline policy updates.

Activities:

1. Implement five Local RLHF pipelines by training one RL policy for each context (i.e. one for Decathlon, one for ETC, one for each Prolific setting) on top of simple-Olé.
    a. *Local Reward Model Training:* During local training, the reward model (RM) is fine-tuned on collected feedback from the organization's learners [1, 2]. As feedback may be explicit (e.g., module ratings or curriculum edits) or implicit (e.g., time-on-task) from WP1, we will need to extend the traditional RL setting to the education context to account for both of these types of signals [15]. This feedback is processed into labeled examples where higher engagement or performance improvements correspond to higher rewards. The reward model learns to associate particular training configurations (e.g., pacing, content depth, or interactivity) with these rewards. Importantly, the RM updates remain context-specific to each organization, ensuring that the scoring reflects local learning preferences and performance metrics.
    b. *Local Policy Optimization:* Once the reward model generates the reward signals, the Ole's LLM-based policy undergoes fine-tuning via reinforcement learning. Here, we use policy optimization techniques such as **Proximal Policy Optimization (PPO) [15]**, which ensures training stability by constrained policy updates, balancing exploration (e.g., trying new learning content or strategies) and exploitation (e.g., reinforcing previously successful curricula). The policy learns to generate learning curriculum paths that maximize learner engagement

while minimizing undesirable patterns, such as boredom or learner drop-off. Positive rewards might reinforce personalized curriculum structures that increase completion rates, while penalties discourage redundant or overly complex content. This RL phase occurs separately for each environment, ensuring that no raw feedback or learner data is communicated externally or across environments.

c. *LLM Memory Mechanism:* To enable real-time policy adjustment from the LLM without requiring a full policy update, we utilize a hierarchical LLM memory, as implemented by CLIN [17]. We integrate "memory" into the LLM context at two granularities: 1) **User-Level Memory,** which tracks user-specific learning history and preferences to personalize content recommendations to a specific learner. 2) **Organization-Level Memory:** Stores preferences, trends, and shared experiences across multiple users within an organization to guide broader content adaptation and curriculum design. The memory component is updated in real-time during each LLM interaction, and is implemented either as an external memory storage and update tool, or in the context-window added directly to the global prompt.

2. Design an adapted architecture for federated RLHF in an online tutoring setting, and implement this architecture as an extension of the Asynchronous-RLHF code across five local RLHF pipelines.
    a. *Secure Communication of Model Updates:* Once the local training process is complete, only the model weight updates with LoRA (not the local reward models) are sent to the central server. These updates represent changes to the model parameters, not the underlying raw data or feedback that informed the changes. We will implement secure communication protocols, such as **secure aggregation [7]** and **differential privacy [8, 9]**, to ensure that individual updates cannot be reconstructed or traced. Secure aggregation sums model updates from multiple clients, obscuring the contribution of any single client [7], while differential privacy can inject noise into the updates to further prevent re-identification [8, 9]. This ensures that client organizations contribute to improving the global model collaboratively without exposing sensitive learner information or proprietary training data [3, 4, 5].
    b. *Aggregation of Model Updates:* The central server aggregates the collected updates from multiple organizations using a federated optimization strategy such as **Federated Averaging (FedAvg) [10, 11]**. This approach computes a weighted average of the local model updates to produce an improved global model. The weight of each client's contribution may be proportional to its dataset size or learning performance to mitigate biases introduced by imbalanced data distributions. By aggregating across diverse training environments, the global Olé model generalizes to support a broader range of learner profiles and instructional contexts. Crucially, this aggregation step enables collective knowledge-sharing while maintaining local privacy, leveraging feedback from varied organizations to improve the global training policy.

3. Implement continuous and incremental global model improvement with periodic offline FedRLHF for Olé.
    a. *Distribution of Improved Global Model:* The aggregated global Olé model is redistributed to all participating organizations. Each organization receives a refined version of the model that integrates privacy-preserved insights from across the federation while preserving adaptability for local fine-tuning.
    b. *Incremental (Periodic) Offline RL updates*: The federated RLHF process is inherently iterative [12]. After each deployment cycle, Olé continues to observe learner engagement, collect feedback, and refine its policy locally before contributing new updates to the global Olé model [13,14]. The process cycles through repeated rounds of local training, secure aggregation [7], and global redistribution. Over time, this enables Olé to progressively align with dynamic learning objectives and emerging trends in professional training.

[1] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.
[2] Ivison, Hamish, et al. "Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback." arXiv preprint arXiv:2406.09279 (2024).
[3] Fan, Flint Xiaofeng, et al. "FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF." arXiv preprint arXiv:2412.15538 (2024).
[4] Ye, Rui, et al. "Openfedllm: Training large language models on decentralized private data via federated learning." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024.
[5] Noukhovitch, Michael, et al. "Asynchronous RLHF: Faster and More Efficient Off-Policy RL for Language Models." arXiv preprint arXiv:2410.18252 (2024).
[6] Cheng, Yujun, et al. "Towards Federated Large Language Models: Motivations, Methods, and Future Directions." IEEE Communications Surveys & Tutorials (2024).
[7] Bonawitz, Keith, et al. "Practical secure aggregation for federated learning on user-held data." arXiv preprint arXiv:1611.04482 (2016).
[8] Dwork, Cynthia. "Differential privacy." International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
[9] Abadi, Martin, et al. "Deep learning with differential privacy." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.
[10] Sun, Tao, Dongsheng Li, and Bao Wang. "Decentralized federated averaging." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.4 (2022): 4289-4301.
[11] Deng, Yuyang, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. "Distributionally robust federated averaging." Advances in neural information processing systems 33 (2020): 15111-15122.
[12] Yao, Liang, et al. "Efficient incremental offline reinforcement learning with sparse broad critic approximation." IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023).
[13] Song, Yuda, et al. "Hybrid rl: Using both offline and online data can make rl efficient." arXiv preprint arXiv:2210.06718 (2022).
[14] Krishna, Shambhavi, and Aishwarya Sahoo. "Solving the Inverse Alignment Problem for Efficient RLHF." arXiv preprint arXiv:2412.10529 (2024).

[15] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).
[16] Loftin, Robert, et al. "Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning." Autonomous agents and multi-agent systems 30 (2016): 30-59.
[17] Majumder, Bodhisattwa Prasad, et al. "Clin: A continually learning language agent for rapid task adaptation and generalization." arXiv preprint arXiv:2310.10134 (2023).
[18] Frick, Evan, et al. "How to Evaluate Reward Models for RLHF." arXiv preprint arXiv:2410.14872 (2024).

Milestone 2 - FedRLHF Implementation

G2-1. We will train local reward models for each organization (Decathlon, ETC, and each Prolific setting). Success will be measured by achieving a reward model accuracy across 12 metrics with an average score at +/- 5% of the range of the Llama 3.1 numbers on the Preference Proxy Evaluation (PPE) benchmark, which measures a reward model's aptitude for RLHF [18].

G2-2. Memory will be updated in real-time during each LLM interaction, capturing user-specific memories (contextual learning history) and organization-level memories (organization preferences and context). LLM Memory will be implemented following the CLIN method [17]. Success will be measured by having at least 50 user-specific memory drafts captured and 5 organization-level memory drafts captured. We will target at least 5% improvement in accuracy when using memory compared to not using memory.

G2-3. We will ensure privacy-preserving updates using secure aggregation and differential privacy. A global federated averaging model update will occur at least three times, aggregating 10 local updates per cycle. Success will be measured by maintaining an accuracy drift of no more than 5% between local and aggregated global models, with a differential privacy epsilon ≤ 3.0.

G2-4. We will redistribute the updated global model to all local environments every month, performing at least two redistributions. Success will be measured by achieving at least 5% accuracy increase after three federation rounds.

G2-5. We will compare the explicit feedback generated from the gold-standard user modified learning paths from WP1 with the generated learning paths from simple-Olé, Olé with LLM memory, RLHF Olé, FedRLHF Olé, and FedRLHF Olé with LLM memory. The model that performs best will be referred to as best-Olé. The learning paths generated by best-Olé will have at least 75% Top-K accuracy on the gold standard learning paths (extracted from human learners in WP1).

WP3: Learning Analytics

Description:

The goal of this WP is to extract insights of learning behavior across different learning profiles and different settings (i.e. Decathlon vs. ETC vs. Prolific) using WP2's best-Olé architecture. The purpose of collecting these insights is to warm-start the LLM Memory mechanism for new users or settings without having to learn each insight individually (and therefore slowly). This will be the secret sauce of Olé's ability to generalize to new environments and user profiles that it has never seen before, analogous to prior research in meta transfer learning that allowed us to create a generalized student behavior model across 26 EPFL MOOCs with diverse subject material [3].

We will begin by analyzing the differences in the interaction data with best-Olé tutor in comparison to the simple-Olé architecture with the sequential additions of RLHF, FedRLHF, and LLM Memory. We will test the best-Olé architecture from WP2 on the environment from WP1 and run usability studies to evaluate Olé variations as well as in-depth, semi-structured interviews with the users. This will enable us to understand if the FedRLHF architecture and LLM memory mechanisms work beyond quantitative evaluations, and are really preferred by users.

We will also deeply analyze collected learning behavior traces across different domain settings, focusing on what was learned as important by the reward functions in each setting vs. the global model, and what LLM memories were stored across different profiles of users, per organization, and on a global scale. We will encode these traces in an embedding model in a learning-trace latent space (i.e. using an autoencoder), which will allow us to generalize to new environments.

Activities:

1. We will conduct a between-subjects and within-subjects experiment comparing the two versions (simple-Olé model with LLM prompting vs. the best best-Olé architecture developed in WP2).
    a. We will teach two different learning concepts (C1 and C2) and adopt a counterbalanced design with four groups, where participants will learn one concept under one condition in different order:
        i. Group 1: C1 with simple Olé then C2 with best-Olé
        ii. Group 2: C2 with simple Olé then C1 with best-Olé
        iii. Group 3: C1 with best-Olé then C2 with simple Olé
        iv. Group 4: C2 with best-Olé then C1 with simple Olé
    b. After each session (learning a concept C1 or C2), participants will solve a quiz and questionnaire. Following both sessions, we will conduct semi-structured interviews to understand participants' perception and preferences.
2. We will analyze the results from the user experiment over three user axes.

    a. Learning outcomes: We will see whether best-Olé improves learning outcomes and whether participants perform better in the quiz. For the learning outcome comparison, we will only use the results from the first session to remove the influence of learning gains. We will combine data from all groups and sessions for subjective and behavioral analysis.
    b. Behavioral analysis: We will examine the differences in behavior when interacting with the platform and run a cluster analysis to understand the differences in behavior. This analysis will further inform platform personalization. The cluster analysis will be done using the time series interaction data in alignment with [1].
    c. Perception: Using the interview transcripts and questionnaire answers, we will analyze user's preference, usability perception as well as what are the benefits and challenges perceived.
3. We will interpret trends across LLM memory (insights that could help generalize Olé to new domains).
    a. Conduct TFIDF word frequency analysis and embedding using Sentence Transformers for LLM memory edits across **user behavior clusters** to assess alignment with learning objectives. The user behavior clusters will be identified from the user experiment in Activity 1. We will determine which rewards are present over all user behavior clusters, and which are unique to certain user behavior clusters.
    b. **Domain-Specific Reward Trends:** Conduct TFIDF word frequency analysis and embedding using Sentence Transformers to identify which aspects of LLM memories are present across domains versus which are domain-specific.
4. We will interpret trends across Reward Models (insights that could help generalize Olé to new domains).
    a. **User Profile Behavioral Alignment:** For each user-behavior cluster, compare reward signals (e.g., engagement, task completion) with statistically observed user behaviors (e.g. average time between clicks, how often modules are completed) to assess alignment with learning objectives. The user behavior clusters will be identified from the user experiment in Activity 1. We will determine which rewards are present over all user behavior clusters, and which are unique to certain user behavior clusters.
    b. **Domain-Specific Reward Trends:** Identify which behaviors are consistently rewarded across domains versus those that are domain-specific. This can be done via saliency maps for reward model analysis [2].
5. We will encode learning behavior insights (average reward for each domain, features regarding memory edit behavior, average learning traces) for each of 5 settings (Decathlon, ETC, 3 Prolific Settings) in an embedding model, and examine Olé's ability to warm-start to a new domain.
    a. **Train Embedding Model**: Use an autoencoder with 128 latent dimensions trained on the collected learning insights encoded with SentenceTransformer. Validate with >90% reconstruction accuracy on a held-out dataset.
    b. **Test Warm Start**: Initialize LLM memory via the embedding model input for three new domains (Domain A: Biology, Domain B: Farming, Domain C: Pharmacy

Assistant). Measure cohesion of outputs through readability metrics (Flesch-Kincaid, SMOG) and a LLM-as-a-judge quantitative evaluation, analogous to the evaluation in iLLuMinaTE [4].

[1] Mejia-Domenzain, Paola, et al. "Identifying and comparing multi-dimensional student profiles across flipped classrooms." International Conference on Artificial Intelligence in Education. Cham: Springer International Publishing, 2022.
[2] [2012.05862] Understanding Learned Reward Functions
[3] Meta transfer learning for early success prediction in MOOCs
[4] From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback

Milestone 3 - Evaluation with more than 100 users

G3-1. We will conduct a pilot study on Prolific with 5 participants to test the comprehension of survey and interview questions and measure the duration of study tasks. Success will be defined as achieving at least 90% participant agreement on the clarity of questions and ensuring the study duration remains within the target range of 30–60 minutes. Insights from this pilot will be used to refine the study design.

G3-2. The large-scale study will engage at least 100 participants across 3-5 domains: teaching data visualization for sports retail, responsible AI practices for healthcare, and model bias for the legal domain, and Python for statistical analysis for recent graduates in finance. Participants will complete all tasks, quizzes, and interviews, with a minimum completion rate of 90% as a success metric. Engagement metrics, including average task duration, completion rates, and quiz performance, will be tracked to ensure robustness of the experimental design.

G3-3. We will compare learning outcomes using paired t-tests or ANOVA to analyze differences in quiz scores between simple-Olé and best-Olé architectures. A significant improvement in quiz scores for the best-Olé version over the simple-Olé version ($p < 0.05$) will indicate success.

G3-4. Using interaction data such as clickstream and response times, we will conduct clustering analyses to identifydistinct learner profiles. Successful clustering will be validated using metrics like silhouette scores, with values above 0.5 indicating well-defined clusters. These profiles will provide insights into behavioral patterns that can guide platform personalization.

G3-5. Memory utilization patterns will be analyzed by measuring edit frequency, diversity, and cross-domain reuse. We expect at least 30% of memory edits are reused across domains, indicating strong generalization.

G3-6. We will analyze reward signals to identify alignment with user actions and learning objectives. Success will be indicated by a correlation coefficient of at least 0.6 between rewards and quiz outcomes, as well as significant patterns of universal versus domain-specific reward trends. Additionally, the variance in reward signals explained by domain factors should exceed 30%, confirming the relevance of domain-specific customization.

G3-7. Achieve >90% reconstruction accuracy for the autoencoder on a held-out dataset of size 15% of the collected traces. Test warm-start performance by initializing LLM memory in three new domains (Biology, Farming, Pharmacy Assistant), with cohesion evaluated via readability metrics (Flesch-Kincaid ≤ 8.0, SMOG ≤ 7.0) and LLM-as-a-judge with decomposed questions scoring ≥ 80%.