



HYBRID LEARNER PROFILING USING BEHAVIORAL TRACES AND LLM-BASED SEMANTIC ANALYSIS

Salma El Omari*	Intelligence, Data and Computing Team, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco	salma.elomari@usms.ma
Najlae Idrissi	Intelligence, Data and Computing Team, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco	N.IDRISSI@usms.ma

* Corresponding author

ABSTRACT

Aim/Purpose	This study proposes a hybrid learner profiling framework that combines behavioral data and semantic analysis based on Large Language Models (LLMs) to build interpretable and actionable learner representations.
Background	Existing learning analytics approaches rely primarily on quantitative interaction logs or static self-reports, failing to capture the underlying cognitive and emotional reasons behind learner behaviors. Most existing profiling approaches describe what the learner does, but not why they behave in a certain way.
Methodology	The framework relies on two complementary pillars. The first analyzes data from virtual learning environments (VLEs). The second uses an LLM as a semantic annotator to extract cognitive, self-regulatory, and emotional dimensions from learners' textual outputs. The approach was validated using real-world of-line datasets and synthetic pedagogical archetypes.
Contribution	This work contributes a multidimensional learner representation that bridges the gap between observable interaction logs and unobservable internal states.
Findings	The behavioral profiling module achieved a dropout prediction AUC of 0.871. Semantic annotation showed no correlation between cognitive complexity and emotional tone ($r = 0.00$). In controlled validation, planning cues were detected

Accepting Editor Natasha Boskic | Received: May 16, 2026 | Revised: August 3, 2026 |

Accepted: August 11, 2026.

Cite as: El Omari, S., & Idrissi, N. (2026). Hybrid learner profiling using behavioral traces and LLM-based semantic analysis. *Interdisciplinary Journal of Information, Knowledge, and Management*, 21, Article 17.

<https://doi.org/10.28945/5867>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

	in 92.3% of valid Strategist messages, compared with 0-16.7% of messages from the other archetypes.
Recommendations for Practitioners	Practitioners can use these multidimensional profiles as a Decision Support System (DSS) for the early identification of at-risk learners and for targeted, personalized interventions.
Recommendations for Researchers	Researchers should treat learner profiling as a multidimensional and partially latent construct requiring complementary data sources.
Impact on Society	This framework helps educational organizations reduce dropout rates and improve the overall quality and inclusivity of digital learning environments.
Future Research	Future developments will focus on deploying the framework in real-world settings, such as integrating it into a real Learning Management System (LMS) and evaluating its impact on adaptive learning and personalized support.
Keywords	learner profiling, learning analytics, large language models, explainable AI, decision support systems

INTRODUCTION

Digital learning, through platforms such as Massive Open Online Courses (MOOCs) and Virtual Learning Environments (VLEs), has significantly transformed educational practices (Capone & Lepore, 2022). However, these environments generate massive amounts of unstructured data, creating challenges for understanding learner behavior and supporting educational decision-making. This provides opportunities to use learning analytics approaches to better understand learners and predict their success (Kew & Tasir, 2022). In this context, learner profiling plays a central role as a knowledge representation tool, particularly for personalized information access and recommendation systems (Ezaldeen et al., 2022; Sassi Hidri, 2024).

Several studies have shown that behavioral traces are strongly linked to learner engagement and to the likelihood of success or failure (Winter et al., 2024; Ye & Pennisi, 2022). Understanding the multidimensional factors that influence learners' continuance intention is critical for addressing these dropout challenges (Jin et al., 2025). Many review studies indicate that approaches to learning analytics often rely on basic indicators such as clicks, time spent, or interactions, which do not capture the full complexity of a learner's profile (Khalil et al., 2023). However, a central research gap remains. Current log-based analytics primarily measure what learners do, but rarely explain why these behaviors occur. They provide limited access to cognitive, self-regulatory, and emotional states. More recent studies emphasize the adoption of a more holistic approach. This perspective views the learner profile as a complex, multidimensional system in which cognition, motivation, engagement, and emotions interact dynamically (Saqr et al., 2024).

Existing work can be divided into three main waves. The first wave relies on explicit questionnaires designed to measure dimensions such as learning styles, motivation, or certain personality traits (Ezzaim et al., 2025; Osei et al., 2022). These approaches provide direct access to learner characteristics but are limited by questionnaire and self-report biases and by their static nature. A second wave relies entirely on behavioral data. These approaches have demonstrated good predictive performance, but they primarily describe what the learner does rather than explaining why the learner adopts certain behaviors. Indeed, the same pattern of activity can correspond to very different educational situations (Ali & Hanna, 2022; Arizmendi et al., 2023).

More recently, certain studies have analyzed learners' written work, particularly through sentiment analysis or supervised language classification models (Dowell & Kovanović, 2022; L. Li et al., 2022; Yee et al., 2023). Although these approaches offer direct access to learners' discourse, they are rarely

integrated with behavioral data from learning platforms. Recent NLP research also shows that LLMs can annotate unstructured texts through task instructions and predefined criteria (He et al., 2024; Tan et al., 2024). However, these approaches mainly address general classification, dataset creation, or text mining. They do not use LLM annotations to build multidimensional pedagogical learner profiles.

To overcome these limitations, we propose a hybrid learner profiling framework that combines behavioral analysis with semantic analysis of learners’ textual output using generative language models. Within the semantic profiling module, the LLM is used for annotation rather than content generation. Synthetic content generation is used separately for controlled validation. The semantic annotation connects structured behavioral logs with pedagogical signals extracted from learner text. Unlike conventional sentiment analysis, the LLM identifies cognitive, self-regulatory, engagement, communication, and emotional dimensions. This use of LLM-based annotation represents the main semantic novelty of the proposed framework.

The framework contains two complementary modules. The behavioral module analyzes activity and assessment data from the OULAD dataset. The semantic module analyzes learner messages from the Stanford MOOC Posts dataset. OULAD supports behavioral validation, while Stanford MOOC Posts supports semantic validation. Controlled synthetic pedagogical archetypes are then examined for profile consistency across the proposed dimensions.

This paper makes the following main contributions:

- A hybrid learner profiling framework combining behavioral analytics and semantic dimensions to build interpretable representations.
- An explainable semantic annotation pipeline using instruction-based LLMs to extract unobservable internal processes from unstructured text.
- A multidimensional learner representation integrating cognitive, behavioral, and emotional signals to support educational decision-making.
- A controlled validation methodology using real-world data alongside synthetic pedagogical archetypes to test profile consistency.

To clarify this gap, the next section reviews questionnaire-based, behavioral, and textual profiling approaches and explains why each remains incomplete when used in isolation. The proposed methodological framework is then presented, detailing the two pillars of the hybrid profiling framework and the adopted validation strategy. Details of the experimental setup and results obtained on the different datasets are then provided, followed by a discussion of the framework’s practical implications, limitations, and future research prospects. Finally, the main contributions are summarized.

LITERATURE REVIEW

Learner profiling is an important step that aims to represent learners’ characteristics to support personalization and improve the learning experience. In general, learner profiling seeks to capture internal dimensions such as motivation, engagement, metacognition, self-regulation, and emotions, as these strongly influence how learners learn and progress (S. Li et al., 2022; M. Liu et al., 2022; Thomas et al., 2023).

EXPLICIT LEARNER PROFILING BASED ON PSYCHOLOGICAL QUESTIONNAIRES AND CLUSTERING

Early approaches to profiling generally rely on explicit questionnaires derived from psychology. These approaches aim to establish a link between psychological aspects and education and to measure dimensions of relative stability such as learning styles, motivation, and personality traits (Escolà-Gascón & Gallifa, 2022; Katsarou & Chatzipanagiotou, 2025; Subramaniam & Suhaimi, 2022). Several studies have shown that certain traits of internal character, such as motivation and metacognitive awareness, are strongly linked to academic success (Tuononen et al., 2023).

Similarly, other studies propose person-centered approaches to student modality profiling, grounded in data-driven methods such as clustering or latent profile characterization. These studies show that learners are not a homogeneous group and that distinct profiles are associated with different behaviors and performance levels (Bauer, 2022; Chen et al., 2026; Law & Liang, 2020; Ofori & Dake, 2026). These approaches are based on solid theoretical foundations and provide a detailed understanding of specific learner characteristics. They also have several limitations. The profiles generated by these approaches are generally static and rely on the information provided by the learners themselves. As well, many studies focus on a single set of profile dimensions, which provides only a partial view of the profile and fails to capture the complex nature of learning processes.

BEHAVIORAL TRACE-BASED PROFILING WITH LEARNING ANALYTICS

Other studies use learning analytics to leverage learner-generated data and better understand learner behavior (Barthakur et al., 2023; Fazil et al., 2024; Hernández-de Menéndez et al., 2022). In general, these studies focus on predicting academic outcomes based on learners' activity logs. They do not aim to define comprehensive learner profiles, which limits our in-depth understanding of learners (Arizmendi et al., 2023; Law & Liang, 2020). These studies use indicators such as access frequency, regularity, persistence, participation, and discussion activity to describe engagement and predict academic outcomes.

Another strand of research aims to provide a more pedagogical explanation, with studies focusing on self-regulated learning behaviors (Yu et al., 2025). Self-regulated learning describes how learners plan, monitor, and adjust their learning. It includes goal setting, time management, perseverance, reflection, and control of cognitive, motivational, affective, and behavioral processes. These studies show that self-regulated learning is an important factor in engagement and academic success and suggest using more structured indicators (Kara et al., 2024) within holistic and technology-assisted models (Michael et al., 2025).

However, several constraints remain. Profiles derived from behavioral traces mainly describe the “what,” i.e., what the learner is doing, without providing insight into the “why,” i.e., what they are thinking or understanding. Furthermore, learners may exhibit similar behaviors, but this does not necessarily indicate that they are in the same state, complicating pedagogical interpretation. Finally, cognitive, motivational, and emotional dimensions remain difficult to determine from behavioral cues alone reliably.

TEXTUAL AND SEMANTIC LEARNER PROFILING WITH LARGE LANGUAGE MODELS

To overcome the limitations of profiles that rely solely on behavioral traces, several studies have explored learners' textual data as a complementary source of information. Messages on forums or in comments offer direct access to the learner's cognitive, metacognitive, and emotional aspects (Ait-Adda et al., 2023; Zankadi et al., 2023).

Early approaches, based on classical semantic analysis, show that integrating textual content adds an interpretable dimension to the learner's profile. These studies highlight that textual data contain rich pedagogical signals (M. Li et al., 2022; Sharma et al., 2024). In most cases, these approaches are based on supervised models or predefined categories. This limits their ability to capture complex and constantly changing internal states. The use of generative Artificial Intelligence (AI) has revolutionized text analysis. Markowitz (2024) shows that generative models can infer reasoning styles from written text and that the form of the text provides insight into basic cognitive mechanisms. Similarly, Shahid et al. (2025) indicate that LLMs can transform unstructured text into actionable data, enabling the automatic identification of pedagogical signals. Some analyses, such as those by Zhang et al. (2024) and Kara et al. (2024), focus on specific pedagogical processes and show that language models can identify self-regulated learning (SRL) behaviors from learners' verbal expressions.

Recent studies use LLMs as text annotators. Given clear instructions and predefined criteria, these models can annotate text without task-specific training (He et al., 2024; Tan et al., 2024). This allows the extraction of multiple structured signals from a single learner message. However, annotation quality depends on prompt design, task definition, model choice, and domain context. Human annotations still outperform LLM annotations on complex language and emotion tasks (Nasution & Onan, 2024). LLM annotations may also reproduce biases, while learner messages can contain sensitive personal or educational information. These limitations require transparent prompts, fixed annotation criteria, reproducible settings, and cautious interpretation of the results.

Current semantic approaches have limitations. They often exclude behavioral and academic data. When they do extract features, they focus on restricted dimensions such as sentiment, cognition, or engagement. Consequently, they do not provide a complete pedagogical profile.

HYBRID AND MULTISOURCE LEARNER PROFILING APPROACHES

Several studies combine different data sources or analytical methods. Ezaldeen et al. (2022) integrate adaptive learner profiling with sentiment analysis for e-learning recommendation. This approach combines learner characteristics with affective information. However, it focuses primarily on recommendations and general sentiment. Jovanovic et al. (2024) combine trace-based learner profiles with learner self-reports. This combination improves the interpretation of observable behaviors, but it depends on information explicitly provided by the learners. Z. Liu et al. (2024) combine MOOC discussion behaviors with textual analysis. Their method uses BERT to identify cognitive and emotional engagement, and fsQCA to examine the relationship between engagement configurations and academic achievement. However, this method is limited to forum engagement and predefined dimensions.

Recent work explores implicit profiling using LLMs. David et al. (2025) propose ProfiLLM, a framework that builds dynamic profiles from chatbot conversations. It uses an LLM and a domain-specific taxonomy to infer technical proficiency. However, it does not model pedagogical, self-regulatory, or emotional states, and it does not use learning platform traces.

These studies show the value of combining multiple data sources. However, existing hybrid models focus on sentiment, self-reports, engagement configurations, or technical expertise. In this work, we connect behavioral and academic evidence with LLM-based pedagogical annotation to cover cognitive, self-regulatory, engagement, communication, and emotional dimensions.

RESEARCH PURPOSES

Regardless of the work carried out in this context, the literature still lacks a coherent framework that unifies behavioral data and semantic analyses. Our project proposes a hybrid profiling approach to address this issue. We combine the analysis of behavioral traces from learning environments with a semantic analysis of learners' textual outputs. Unlike existing approaches, our goal is not to optimize performance prediction. We aim to construct an interpretable pedagogical profile that explains observed behaviors in relation to underlying cognitive, self-regulatory, and emotional states.

This approach enables a transition from descriptive or predictive profiles to explanatory profiles. It paves the way for more interpretable, better-aligned personalization and recommendation mechanisms that address learners' real needs by extracting structured knowledge from interaction logs and text.

Based on this research gap, the present study is guided by the following research questions:

- RQ1.** To what extent does the integration of cognitive, self-regulation, and engagement dimensions improve dropout prediction compared with using each profile dimension independently?
- RQ2.** Can pedagogical inference from forum texts provide more direct and interpretable access to cognitive and emotional dimensions than behavioral proxies alone?

RQ3. Does the proposed hybrid profiling framework enable the identification of learner profiles that traditional profiling approaches cannot detect?

To address these research questions, the next section details the proposed methodology. It describes the hybrid framework architecture, the dual-pillar analysis process, and the evaluation strategy used to validate the generated profiles.

METHOD

The objective of this work is to propose a hybrid profiling framework based on several complementary dimensions. The framework represents the learner’s educational status in an interpretable manner. Unlike approaches that primarily focus on performance prediction, our framework constructs a multidimensional, evolving learner representation. Our approach relies on several types of data commonly available in online learning environments. We take into account behavioral information derived from platform interactions, pedagogical data from activities and assessments, and learner-generated text, such as messages and comments. These data sources are complementary because none is sufficient on its own. However, these online environments cannot contain all the necessary data. Therefore, although some information may be missing, we need to ensure our approach remains robust and effective.

Based on these data, we represent the learner’s profile through four families of pedagogical dimensions. The first family concerns the cognitive dimension. The second concerns SRL. The third dimension concerns engagement. Finally, the fourth family is about communicational and affective aspects. These dimensions can be instantiated in any environment, as they are defined at an abstract level. Our approach is thus independent of the platform or the dataset.

The proposed pipeline is illustrated in Figure 1. We first collect and organize the raw data according to its nature. From these traces, we extract behavioral and academic indicators. At the same time, reasoning, SRL, or emotional dimensions are annotated from learners’ textual content. A local Mistral-7B-Instruct model performs this semantic annotation. This information is then fused to create a holistic learner profile, which is updated over time to reflect the learner’s academic progression. The following subsections describe the annotation prompt, model settings, and synthetic generation parameters.

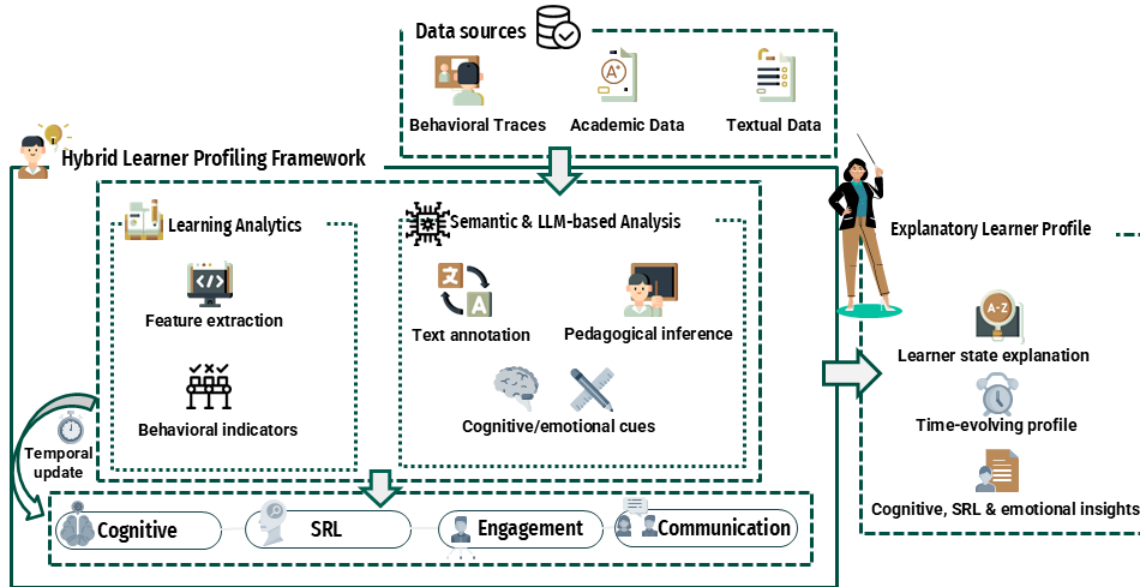


Figure 1. Overview of the proposed hybrid learner profiling framework

PEDAGOGICAL DIMENSIONS AND PROFILING SPACE

We have structured the learner profile around four categories of dimensions. Each category contains several sub-dimensions. Table 1 illustrates the multidimensional and hybrid nature of the profiling framework we propose.

Table 1. Synthesis of pedagogical dimensions and primary data sources

Dimension family	Sub-dimension	Primary data source
Cognitive	Knowledge level	Academic logs
	Mastery stability over time	Academic logs
	Cognitive complexity	Textual data
	Error type (none, conceptual, careless, or technical)	Logs + text
Self-regulation and motivation	Planning	Textual data
	Time management	Behavioral logs
	Perseverance	Logs + text
Engagement and behavior	Behavioral engagement	Behavioral logs
	Cognitive engagement	Logs + text
	Social engagement	Logs + text
Interaction and affect	Communication clarity	Textual data
	Emotional tone	Textual data

Cognitive dimension

The cognitive dimension represents the learner's level of understanding and quality of reasoning (Markowitz, 2024). To go beyond a vision limited to grades or scores, we have defined four sub-dimensions. The first concerns the level of knowledge, which reflects the degree of mastery of content. The second sub-dimension is the stability of mastery over time. It enables us to distinguish lasting mastery from one-off success. We also incorporate cognitive complexity, which corresponds to the depth of reasoning expressed in learner-generated text. Several studies show that language structure reflects underlying cognitive processes (Ait-Adda et al., 2023). Finally, the type of error allows us to differentiate between errors related to conceptual confusion and those due to misinterpretation of instructions or lack of attention.

Self-regulated learning and motivation

Several studies indicate that self-regulated learning is directly linked to the learner's final performance. This family aims to characterize how the learner organizes, controls, and maintains their learning (Kara et al., 2024; Teng, 2024). We model this dimension through three main sub-dimensions: (1) Planning, which reflects the learner's ability to structure their work and set goals; (2) time management, which reflects the temporal organization of activities and the regularity of learning; and (3) perseverance, which describes the learner's continuity of effort in the face of difficulties and failures.

Engagement

Engagement describes the intensity and nature of learners' involvement in a learning environment. Research in learning analytics often focuses on engagement as a key dimension, typically measured by the frequency or duration of interactions (Arizmendi et al., 2023; Z. Liu et al., 2024). However, these quantitative indicators are insufficient to understand the true nature of engagement. That is why we distinguish three complementary forms. Behavioral engagement corresponds to observable activities on the platform. Cognitive engagement reflects visible mental effort and investment in learning tasks. Finally, social engagement reflects participation in exchanges and interactions with peers.

Interaction and communication

This fourth family aims to capture pedagogical signals that are difficult to access from logs alone. It is based on learners’ unstructured writing (Jovanovic et al., 2024). We identify two main sub-dimensions within this category. One is the clarity and structure of communication, which corresponds to the quality of the learner’s writing and the structure of their text. The other is emotional tone, which refers to the emotional cues expressed in language, such as frustration, confidence, anxiety, or satisfaction. These dimensions provide information about expressed emotional cues and communication patterns that may influence the learning process.

BEHAVIORAL PROFILING MODULE

In this module, we convert raw data from platform interactions into structured metadata to assess the intensity and quality of engagement. In our methodology, the learner is not viewed as a static entity. We follow three key steps to process behavioral traces:

- *Behavioral representation:* To characterize learner behavior, we calculate a set of attributes for each period based on interaction logs and assessments. We group them into three behavioral dimensions. Activity intensity is measured using metrics such as the number of clicks, time spent on educational resources, and the variety of content viewed. Work consistency is calculated based on the variety of activities per week, the number of active days, and the duration of inactive periods. These measures distinguish consistent work from last-minute work. Academic performance is assessed based on assessments, which include the average grade, grade variability, and the number of assessments submitted, providing insights into the learner’s level of mastery and persistence. The self-regulation dimension is derived from these indicators related to behavior and academic performance. This dimension consists of indirect measures of planning based on anticipated engagement, time management based on the regularity of activity, perseverance based on participation in assessments, and self-control based on the combination of performance level and consistent participation in assessments.
- *Temporal Aggregation:* Raw data extracted from learning environments is typically fragmented. To address this, we organize these events along a fixed timeline, which allows us to transform isolated actions into a time series and analyze changes in behavior over time.
- *Decay function:* For the profile to remain relevant, it must describe the learner’s current situation. A click from two months ago provides less information than a recent interaction. We therefore apply a time decay function to our metrics. This allows our framework to place greater weight on recent actions. As a result, the profiling system is highly responsive to the first signs of dropout or disengagement, supporting proactive educational decision-making. The cumulative engagement score E_t at time t is calculated as follows:

$$E_t = \sum_{i=0}^t A_i \cdot e^{-\lambda(t-t_i)}$$

where A_i represents the amplitude of activity at time t_i , and λ is the decay parameter that controls the rate at which past interactions become obsolete.

The engagement score is updated at each observation period. The decay parameter λ remains fixed during the analysis.

TEXTUAL SEMANTIC PROFILING VIA LLM-BASED ANNOTATION

To enable semantic profiling, we use learners’ written output as an additional source of information. The textual data considered in this module originates primarily from messages posted on discussion forums. They include comments on activities and other written expressions from the learning environment.

Unfortunately, these sources are generally unstructured and heterogeneous, making them difficult to exploit with conventional methods. For this reason, our approach relies on LLMs for semantic inference and pedagogical annotation. In the semantic profiling module, the LLM analyzes only learner-generated text. Synthetic generation is used separately for controlled validation. The role of the semantic module is to transform raw text into structured and pedagogically interpretable signals.

We create a structured prompt for annotation. The prompt specifies the required dimensions, their data types, and their accepted ranges. Our model is instructed to return numerical values in a fixed and structured format. We directly instruct the model as follows: Cognitive complexity and communication clarity are evaluated on a discrete scale from 1 to 5. Signals related to engagement also follow the same discrete scale. Self-regulation planning is represented as a binary indicator in the interval $[0,1]$. Emotional tone is expressed on a continuous scale from -1 to 1 . Error type uses four categories: none, conceptual, careless, and technical. The complete annotation prompt appears in the Appendix.

In semantic processing, the LLM links each text to the pedagogical benchmarks described in the previous section. These benchmarks correspond to cognitive complexity, linguistic clarity, self-regulated learning, and emotional tone. The textual content is first analyzed using several predefined dimensions derived from the text, such as cognitive elaboration, planning indicators, engagement signals, clarity of communication, and affective tone.

We selected these dimensions because they can be estimated from explicit textual cues. They also require no additional manual input from learners. The prompt instructs the model to act as a structured semantic annotator. It produces quantifiable indicators rather than an open textual response. This fixed output structure supports comparison across messages and learners.

The model output follows a structured schema to ensure consistency across annotations. For each message, the LLM returns a set of key-value pairs corresponding to the defined dimensions, for example: `{ "cognitive_complexity": 4, "error_type": "Conceptual", "autoreg_planning": 1, "engagement_cognitive": 3, "comm_clarity": 3, "comm_tone": 0.2 }`.

The annotation output is parsed as a JSON object. Markdown formatting is removed before parsing. Only successfully parsed outputs are retained for semantic analysis. Malformed field names are harmonized during the standardization stage.

We then aggregate the annotations generated by the semantic module at the learner level by averaging the available observations. This aggregation provides a coherent representation of the dimensions. It also allows us to analyze their progression over time. Before aggregation, these raw scores are normalized to the interval $[0,1]$. The categorical error type is retained separately. This process places the semantic dimensions on a common scale. It also allows their combination with the behavioral profile.

The semantic profile is updated whenever new learner-generated text becomes available. Between two textual observations, the latest semantic representation is retained. Detailed model settings and normalization procedures are reported in the Experimental Setup section.

DYNAMIC INTEGRATION OF HETEROGENEOUS DATA

After quantifying the behavioral and semantic features, we combine the information to obtain a comprehensive learner representation. The learner profile is represented as a unified state vector. However, another problem remains: the two data sources considered in this framework exhibit different temporal dynamics. Behavioral traces from the platform logs are generally continuous and dense. In contrast, text outputs are sparse and irregular. We adopt an asynchronous updating strategy to manage this temporal heterogeneity.

For dimensions derived from behavioral data, we propose updating them at regular intervals. Their evolution follows a smoothed update rule. This rule combines newly observed information with the previous learner state:

$$D_t = \alpha \cdot D_t^{\text{new}} + (1 - \alpha) \cdot D_{t-1}$$

where D_t^{new} denotes the value computed from recent behavioral traces. D_{t-1} represents the previous state of the dimension. The parameter $\alpha \in (0,1)$ controls the adaptation speed. A higher value gives more weight to recent observations.

In contrast, semantic dimensions extracted from textual data follow a memory-based update mechanism. Due to the irregular nature of learner writings, their values are updated only when new textual content is available:

$$D_t = \begin{cases} D_t^{\text{new}}, & \text{if a new textual contribution is observed at time } t \\ D_{t-1}, & \text{otherwise} \end{cases}$$

This strategy maintains a complete and stable learner profile, even when no new messages are available. It also respects the irregular nature of textual interactions in online learning environments. Therefore, behavioral dimensions evolve regularly, while semantic dimensions evolve only after a new textual observation.

INSTANTIATION UNDER PARTIAL DATA AVAILABILITY

Our proposed framework is designed to operate in real-world learning environments where available information is often partial and heterogeneous. We consider two real datasets widely used in the literature on learning analytics and online interaction analysis. These datasets have complementary characteristics and illustrate realistic situations in which the available information is partial and heterogeneous, reflecting the majority of real learning environments. In both datasets, the framework is not modified. Only the dimensions that can be instantiated from the available data are activated. Other dimensions remain unobserved. Unobserved dimensions are not interpreted as zero values. This flexibility produces partial but consistent pedagogical profiles. It also allows the framework to use available information without requiring every data source.

SYNTHETIC VALIDATION

Validating a hybrid system such as ours raises a methodological challenge. Most public datasets are either rich in behavioral logs but poor in textual data or vice versa. To address this limitation, we have developed a validation protocol based on synthetic data. A pedagogical archetype represents a predefined combination of behavioral patterns, academic outcomes, and textual cues. These archetypes test whether the framework produces coherent profiles under controlled conditions. The use of LLM-generated synthetic data has recently proven effective in methodological validations where empirical data is sparse or imbalanced (Nuryani et al., 2024).

We define five representative pedagogical archetypes. Each archetype includes three learners observed across twelve weeks. This process produces 180 behavioral records. The initial archetype corresponds to the frustrated expert. It unites high scores and also high activity, and the discourse tone is negative or neutral-critical. The second archetype is the social novice, whose performance ranges from poor to moderate; however, participation in forums is frequent, with a tone that is globally positive. The third archetype is the dropout, marked by an exponential decline in activity, accompanied by messages expressing confusion and a negative emotional tone. The fourth archetype is the Strategist. It shows moderate and regular activity. Its messages contain explicit cues for planning and learning strategies. The final archetype concerns anxious but high-performing individuals. This archetype is characterized by a high click-through rate and decent satisfaction scores, despite an unstable emotional tone and messages expressing a degree of concern. The selection of these archetypes was not arbitrary. We based our selection on learner patterns reported in the learning analytics and MOOC literature (Arizmendi et al., 2023; Z. Liu et al., 2024). The Mistral-7B-Instruct model generates forum

messages from rule-based descriptions of each archetype. The exact generation parameters appear in the Experimental Setup section. The complete generation prompt appears in the Appendix.

Even though the same family of LLMs is used for text generation and annotation, a strict separation of functions mitigates the risk of circularity. The generation model is conditioned by rule-based archetype descriptions. The annotation model remains independent of archetype labels. This separation reduces direct label leakage, although model-specific consistency effects may remain. Synthetic data is not used for performance evaluation. It is used to verify pedagogical consistency and interpretability under controlled conditions, as a complementary analysis that does not replace empirical evaluation using real-world data.

EXPERIMENTS

DATASET DESCRIPTION

We conducted the experiment on two real-world offline datasets: The Open University Learning Analytics Dataset (OULAD) and the Stanford MOOC Posts Dataset. Table 2 provides a concise descriptive overview of these two datasets. The datasets support two complementary case studies and are not linked at the learner level. OULAD supports behavioral profiling, while Stanford MOOC Posts supports semantic profiling.

Table 2. Core characteristics of the datasets used in this study

Characteristic	OULAD	Stanford MOOC posts
Learners	28,785	2,882
Courses/modules	7	7
Total interactions/posts	10.6M	3,503
Avg. words per post	-	96.2
Dropout rate	31.2%	-

OPEN UNIVERSITY LEARNING ANALYTICS DATASET

The first dataset used is OULAD. It was collected from the online learning platform of the Open University, a British university that specializes in distance learning. The Open University offers seven training modules to its students over several academic sessions. The data collected from these sessions includes information on the learners’ interactions with the VLEs, as well as their academic performance and demographics.

In this case study, there is no textual data from discussion forums, but there is behavioral data regarding the learners’ interactions with the VLEs.

Behavioral validation used all seven OULAD modules and all available presentations. The initial dataset included 28,785 unique learners. Interactions from week 0 to week 40 were analyzed. Logs recorded before the course start were excluded. Learners without interaction or assessment traces were also removed. The final analytical sample included 29,144 active enrollments. A learner could appear in more than one module or presentation.

STANFORD MOOC POSTS DATASET

The second dataset is the Stanford MOOC Posts Dataset, a collection of posts from learners on the forums of several MOOCs offered by Stanford University. The posts have been organized by the courses for which they were made, as well as an aggregated dataset containing all the posts used in previous training and cross-validation work. Unlike OULAD, this dataset focuses solely on learners’

text production. It does not provide information on platform interactions or behavioral logs. Therefore, it is used to instantiate and examine the semantic dimensions of the framework. The source dataset contained 3,503 posts. After removing posts with fewer than five words, 3,404 posts remained. One annotation could not be parsed. The final semantic dataset, therefore, contained 3,403 saved annotations.

EXPERIMENTAL SETUP

Before instantiating the proposed model, we applied a standard preprocessing pipeline. We grouped behavioral interactions every week. Incomplete or noisy records were filtered out, and all numerical dimensions were normalized to the interval $[0, 1]$. Categorical fields were processed separately. This ensured comparability within the learner profile vector. When a given dimension could not be instantiated from the available data, it was treated as unobserved. This did not affect the internal consistency of the learner’s representation.

We conducted a preliminary sensitivity analysis for the behavioral parameters. Candidate values were $\lambda \in \{0.05, 0.1, 0.2, 0.3\}$ and $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$. We considered Macro-F1, AUC, and temporal profile stability. The selected values were $\lambda = 0.1$ and $\alpha = 0.3$.

The prediction target was binary. Withdrawn learners formed the positive class. Distinction, Pass, and Fail formed the non-withdrawn class. We used five-fold GroupKFold cross-validation. Groups were defined by the module and presentation combination. This design separated course offerings across the validation folds.

For dropout prediction, we used a Random Forest classifier. The model included 300 estimators and a maximum depth of 14. The `min_samples_split` was 20. The `min_samples_leaf` was 10. Class weights were balanced to address class imbalance. Missing values were median-imputed. Features were standardized before classification. All preprocessing steps were fitted only on each training fold. A random state of 42 was used to ensure reproducibility. Unspecified parameters retained their default scikit-learn values.

For the semantic annotation module, we used an instruction-tuned large language model as an inference engine. In our experiments, we used the mistralai/Mistral-7B-Instruct-v0.2 model as the inference engine. In this module, the model analyzed only texts produced by learners. The model operated in a zero-shot configuration. The model ran locally with 4-bit NF4 quantization. Double quantization was enabled, and computations used float16. Messages were processed sequentially with a batch size of one. Annotation used a temperature of 0.1, sampling, and a maximum of 200 new tokens. One annotation was produced per message.

We used a structured prompt to extract interpretable indicators from the pedagogical framework. The prompt requested cognitive complexity, error type, planning, cognitive engagement, communication clarity, and emotional tone. The model returned these indicators as a strict JSON object. Markdown markers were removed before parsing. The parser extracted the content between the first and last braces. Only successfully parsed outputs were retained. The complete annotation prompt appears in the Appendix.

Other instruction-oriented language models could be used in the same way without modifying the core methodology.

Finally, we generated synthetic pedagogical archetypes using the same Mistral model under rule-based pedagogical conditions. Each of the five pedagogical archetypes represented three learners observed over twelve weeks. This process produced 180 behavioral records. Synthetic text generation used 4-bit quantization, a temperature of 0.85, sampling, and a maximum of 80 new tokens. The complete generation prompt appears in the Appendix. Non-empty synthetic messages were then analyzed using a separate annotation prompt. This prompt extracted cognitive complexity, planning, and emotional tone. Archetype labels were excluded from the annotation input.

RESULTS

Figure 2 summarizes the experimental workflow through three validation paths. Panel A presents behavioral validation on OULAD. Panel B presents semantic profiling on Stanford MOOC posts. Panel C presents controlled hybrid validation using pedagogical archetypes.

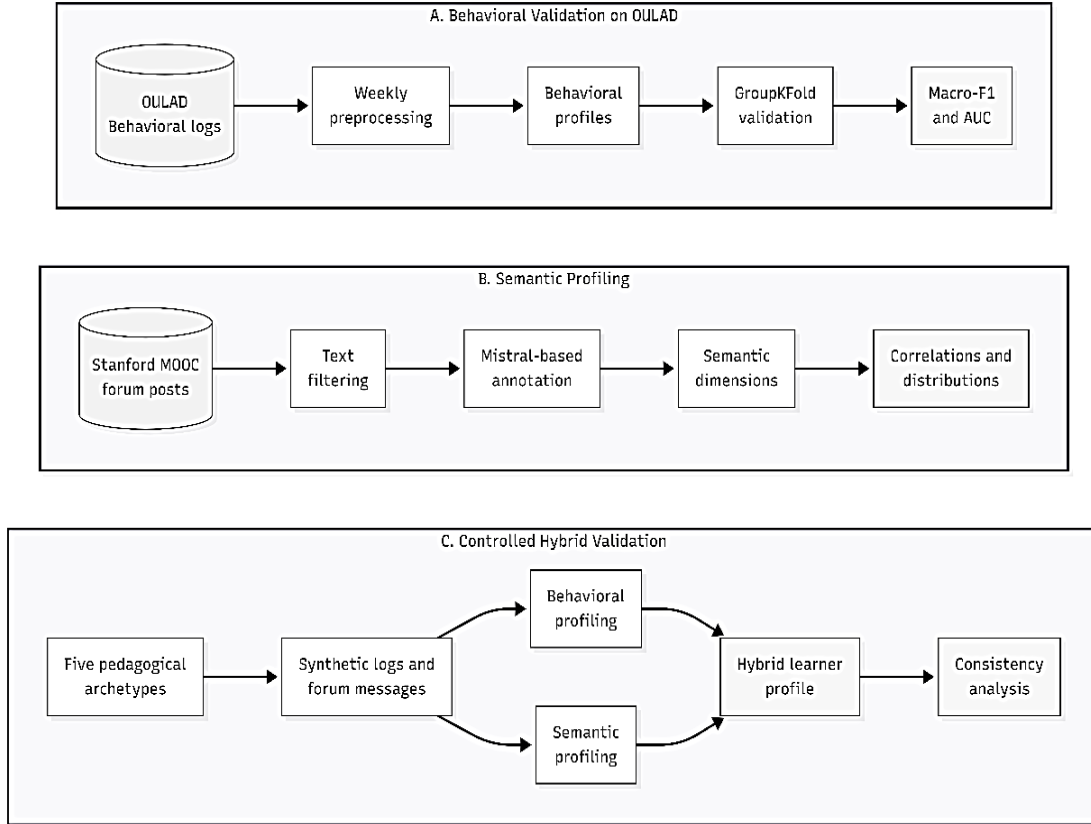


Figure 2. Experimental workflows for behavioral, semantic, and controlled hybrid validation

BEHAVIORAL VALIDATION ON OULAD

Figure 3 shows the evolution of Macro-F1 over the weeks of the course for the different dimensions of the learner profile. The F1-score is the harmonic mean of precision and recall. Macro-F1 calculates one F1-score for each class. It then averages these scores without weighting class frequency. Here, the classes are withdrawn and non-withdrawn learners. For each dimension family, sub-dimensions are aggregated using normalized averaging to yield a stable, interpretable macro-level comparison. Performance increases gradually for all dimensions over time. The hybrid profile consistently achieves the highest scores over the entire observation period. The cognitive dimension shows intermediate performance, followed by self-regulation, while behavioral engagement shows the lowest values. No abrupt fluctuations are observed.

Table 3 shows the dropout prediction performance on the OULAD dataset, evaluated using five-fold GroupKFold cross-validation, for different learner profile dimensions. The hybrid profile reaches the highest AUC, at 0.871. It also achieves the highest F1-score, at 0.689. The cognitive-only profile remains close, with an AUC of 0.857 and an F1-score of 0.685. The self-regulation and engagement profiles show lower recall values. This reduces their F1-scores despite relatively high precision.

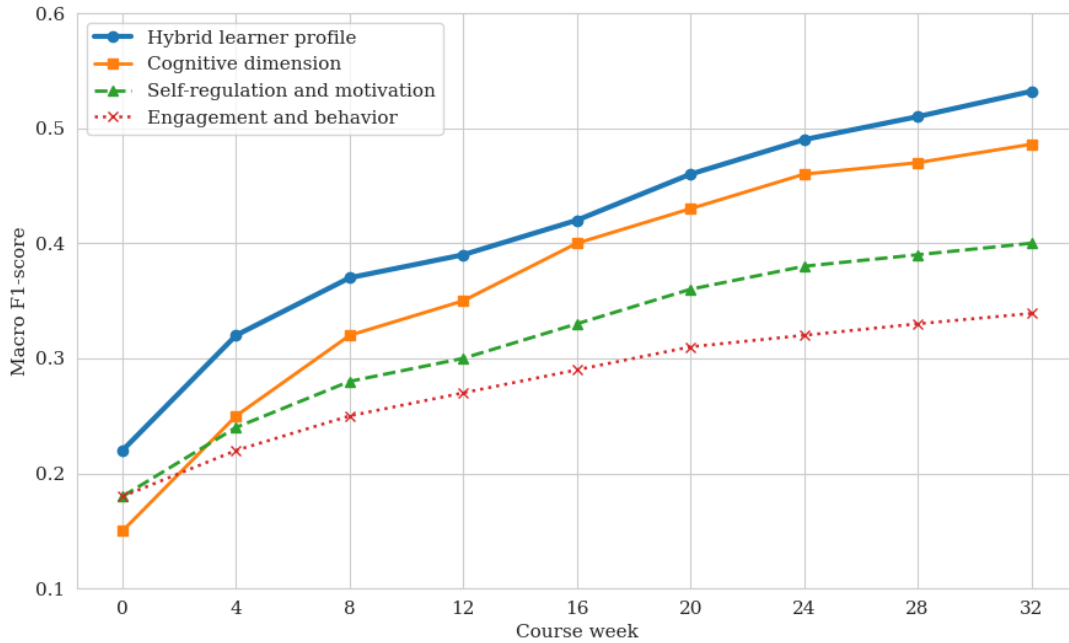


Figure 3. Temporal evolution of Macro-F1 scores across course weeks for different learner profile dimensions

Table 3. Dropout prediction performance across learner profile dimensions on OULAD (target: Withdrawn, 5-fold GroupKFold validation)

Profile/model	Precision	Recall	F1-score	AUC
Cognitive only (D1)	0.698	0.672	0.685	0.857
Self-regulation only (D2)	0.718	0.461	0.562	0.796
Engagement only (D3)	0.720	0.444	0.549	0.776
Hybrid profile (D1+D2+D3)	0.719	0.662	0.689	0.871

SEMANTIC PROFILING RESULTS

Figure 4 represents the obtained correlation matrix between the dimensions of the learner profile. The results indicate a weak to moderate correlation among the cognitive, self-regulation, behavioral, and emotional axes. The results show weak overlap between the dimensions. The correlation between cognitive complexity and cognitive engagement remains moderate, at $r = 0.65$. The results also show that correlations involving emotional tone remain low across the other semantic dimensions. The results highlight differences between the dimensions.

Figure 5 presents the distribution of message-level semantic annotation scores. Cognitive complexity is shown after normalization to $[0,1]$, while the remaining dimensions are shown on their annotation scales. The cognitive score has a very large range. In contrast, the planning indicator is less frequent. Emotional tone scores are centered around more neutral values, unlike the clarity of communication, which varies gradually.

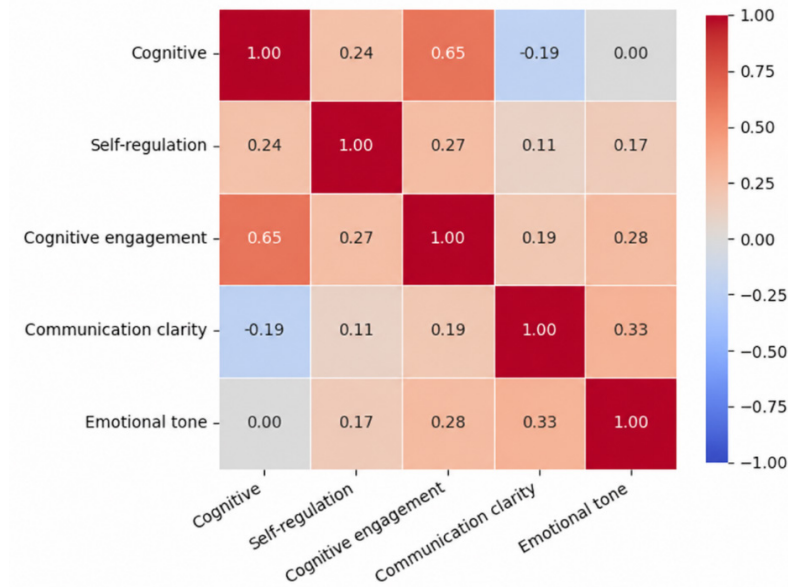


Figure 4. Correlation matrix between learner profile dimensions

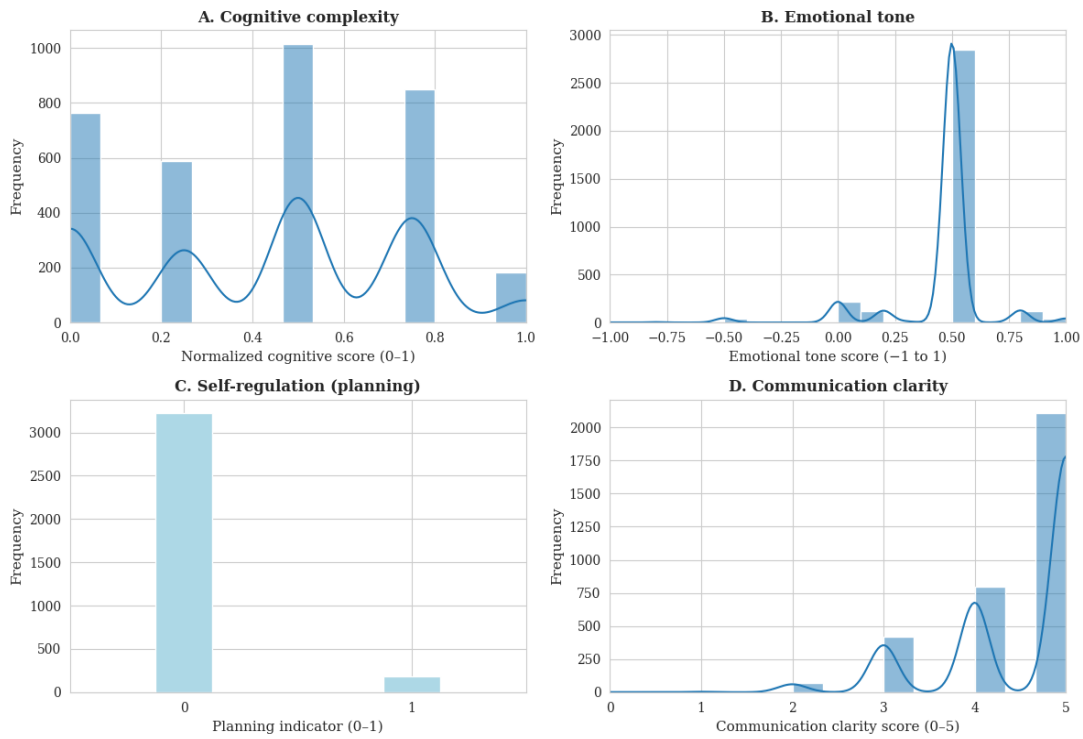


Figure 5. Distribution of scores for the semantic dimensions of the learner profile

HYBRID VALIDATION ON SYNTHETIC PEDAGOGICAL ARCHETYPES

Figure 6 shows normalized cognitive complexity and emotional tone across the five pedagogical archetypes. The Frustrated Expert shows moderate-to-high cognitive complexity and variable emotional tone. The Social Novice shows low cognitive complexity and a consistently positive tone. The Anxious Achiever presents wider dispersion across both dimensions. The Strategist shows limited dispersion. The Dropout shows moderate cognitive complexity and variable emotional tone.

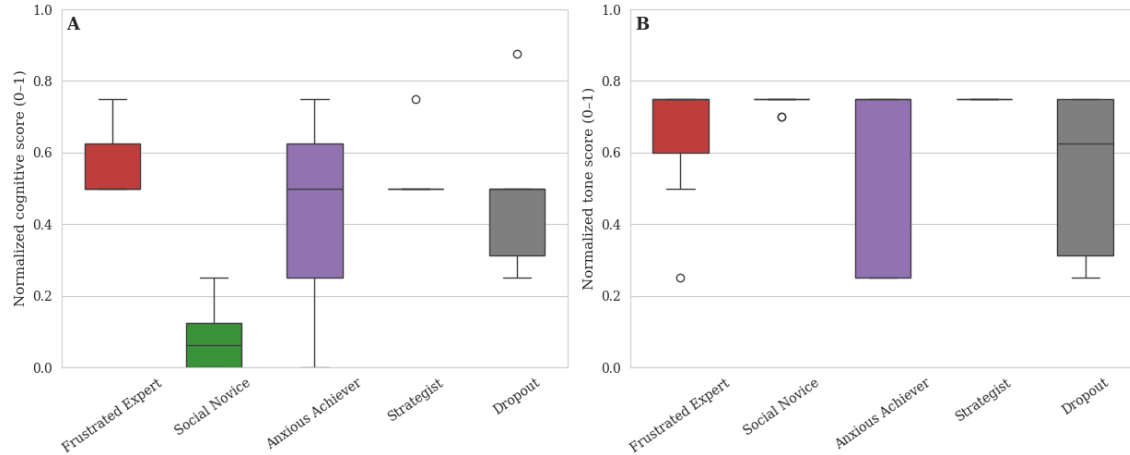


Figure 6. Distribution of normalized scores by pedagogical archetype
(A) Cognitive discrimination, (B) Emotional tone discrimination

Table 4 reports planning detection across the five pedagogical archetypes. Planning is mainly detected for the Strategist archetype, at 92.3%. Lower rates appear for the Dropout and Frustrated Expert. No planning cue appears for the Social Novice or Anxious Achiever.

Table 4. Detection rate of the metacognitive planning strategy by pedagogical archetype

Pedagogical archetype	Planning detection rate (%)
Frustrated Expert	6.7
Social Novice	0.0
Anxious Achiever	0.0
Strategist	92.3
Dropout	16.7

DISCUSSION

The primary goal of our approach was not to optimize predictive performance, but to demonstrate that behavioral data alone is insufficient to capture the complexity of a learner’s profile. Cognitive, metacognitive, and emotional signals must supplement it. This supports earlier findings that learner profiles based solely on traces are incomplete and require additional sources, such as self-reports or insights into a learner’s internal states, to be more comprehensive (Jovanovic et al., 2024). Beyond this limitation, our results suggest that learner profiles should draw on multiple data sources that combine behavioral signals with latent cognitive and affective information.

The temporal analyses we conducted on OULAD indicate that classification performance evolves gradually over the course of the weeks. This reflects the nature of learning, as at the beginning of the course, the available traces are limited, but they gradually become richer, allowing for more robust and better-defined profiles. Figure 3 shows that the hybrid profile outperforms the isolated dimensions. This highlights the complementary nature of the signals employed. Similar patterns have been reported in studies that combine self-regulated learning indicators, engagement measures, and log data, in which models that combine constructs outperform unidimensional models (Ali & Hanna,

2022). We can therefore confirm that the hybrid integration of cognitive, self-regulation, and engagement dimensions reduces decision-making ambiguity. It provides a better understanding of the heterogeneity of learning trajectories. This directly addresses RQ1.

Among the strongest predictors is the cognitive dimension, but it alone is not sufficient to explain the risk of dropping out. The self-regulation and engagement dimensions show weaker predictive performance in isolation. However, they provide complementary information within the hybrid profile. This pattern is plausible in education. Inactivity often relates to failure, but activity alone does not guarantee success. In reality, many learners continue to behave actively right up until they drop out. This behavior has been described as gaming the system or shallow engagement (Khanna et al., 2025). This result suggests that learner profiling should be viewed as a multidimensional representation problem rather than as a simple prediction task.

Methodologically, we opted for validation using GroupKFold, unlike in previous studies, which limits data leakage between the training and test sets. This choice enhances the reliability of the results presented. However, GroupKFold does not remove bias from the dataset itself. The results still reflect one educational context. Their transfer to other settings remains uncertain.

Our semantic analysis highlights that the dimensions are weakly correlated with one another. These results suggest that they represent different types of learner behavior. This supports the use of a multidimensional profiling model. It also supports RQ2 by demonstrating that semantic inference facilitates our interpretation of cognitive and emotional dimensions that cannot be reliably inferred from behavioral data alone. In this sense, the semantic module does not simply improve prediction but changes the nature of the learner model by making previously unobservable dimensions accessible through textual analysis. However, weak correlations do not prove annotation accuracy. They only show limited statistical overlap between the generated scores. LLM annotations may depend on prompt wording, language, writing style, and course context. The model may also reproduce linguistic or cultural biases. Short or multilingual messages may receive less stable scores. Another model may produce different annotations for the same text. The study does not compare the annotations with a human-coded reference set. Therefore, the semantic scores should not be treated as ground truth. Human validation remains necessary before real educational use.

This finding differs from the standard approaches presented in previous literature on forum analysis, where forums are generally studied from the perspective of sentiment classification or certain predefined categories, without delving into specific pedagogical concepts such as cognitive complexity or self-regulation (Yee et al., 2023). We also found almost no correlation between cognitive complexity and emotional tone, suggesting that the LLM extractor we use effectively separates cognitive ability from emotional expression. This approach contrasts with traditional approaches in the field of educational technology, where sentiment analysis is used as an indicator of satisfaction or success (Jovanovic et al., 2024). Also, previous research on MOOCs shows that sentiment analysis is primarily used to identify whether learner sentiments are positive or negative, which does not provide much insight into their deeper learning processes (L. Li et al., 2022). Our results show that a learner can be cognitively competent yet frustrated, or socially positive despite experiencing cognitive difficulties. Similarly, the distributions shown in Figure 5 indicate that certain competencies, such as self-regulation and planning, are rare yet highly informative.

In this paper, a learner profile refers to the representation produced by the framework. A pedagogical archetype refers to a predefined synthetic pattern. These terms describe different elements. Validation using synthetic pedagogical archetypes supports profile analysis under controlled conditions. The fact that the metacognitive planning strategy is detected primarily in the Strategist archetype suggests that the model does not conflate this competency with more common factors related to activity or performance. Although the synthetic pedagogical archetypes are designed to capture specific learner archetypes, not all dimensions are presented explicitly in each generated message. Planning is only signaled when explicitly verbalized, which explains the absence of planning cues in some archetypes.

This behavior is consistent with the model extracting only explicit signals rather than inferring latent traits. Detected dimensions reflect actual evidence in the text rather than assumed traits. The very low variability in some dimensions for certain archetypes is expected due to the nature of the synthetic pedagogical archetypes, in which each archetype reflects a specific set of dominant characteristics, resulting in stable values for some dimensions within each profile. This behavior indicates that the generated profiles are not artificially diversified, but rather consistent, supporting the internal consistency of the archetype design. This reinforces our hybrid methodology and directly addresses RQ3. The controlled experiment suggests that textual indicators can distinguish archetypes that would not be fully represented by behavioral variables alone. These results highlight that learner profiling should be viewed as a multifaceted, partially latent construct rather than a straightforward function of observable interactions. The proposed framework supports this view by explicitly considering both observable behaviors and textual indicators related to cognitive and affective processes.

Synthetic validation provides a controlled consistency check. It does not measure real-world predictive performance. The five archetypes simplify the diversity of real learners. Their rule-based design may also reduce score variability. The same Mistral model family generates and annotates the synthetic messages. Separate prompts reduce direct label leakage. However, model-specific language patterns may remain. This may increase the apparent consistency of the results.

Profiling results must be interpreted in terms of personalization and decision-making, rather than simple predictions. From an organizational perspective, educational institutions and LMS administrators can utilize this hybrid framework as a Decision Support System. When a dimension does not provide significant predictive value in a given dataset, it still offers actionable knowledge for personalizing learning paths and allocating institutional resources, such as human tutoring. This work can serve as a foundation for explainable personalization systems. For us, learner profiles are not merely opaque scores; they are structured around interpretable dimensions. This view aligns with recent trends in information management and learning analytics, emphasizing the need for interpretable and explainable learner models to support educational decision-making and adaptive systems (Dowell & Kovanović, 2022).

Computational cost remains an important limitation. Four-bit quantization reduces the required memory. However, the model processes messages sequentially with a batch size of one. This configuration may increase processing time. Scalability, therefore, remains unverified for large courses. Larger deployments may require batching, caching, or smaller language models. The asynchronous update strategy can reduce repeated processing when no new text appears.

Integration into a real LMS also creates technical challenges. Logs and messages need consistent learner identifiers. Their timestamps and course structures must also align. Platform schemas may differ across institutions. OULAD and Stanford MOOC Posts do not contain the same learners. Therefore, real end-to-end fusion is not evaluated on linked data. The complete hybrid profile is examined through synthetic archetypes.

This work addresses several challenges in current learning analytics research, as educational systems increasingly rely on automated models to support decision-making. In this context, the need for an explainable pedagogical learner profile has become essential. The hybrid profiling approach proposed in this study is situated within the field of explainable AI in education and aligns with recent calls for responsible and transparent personalization in digital learning environments.

CONCLUSION

We introduced a hybrid profiling framework that integrates behavioral traces with semantic inference. The framework differs from static questionnaire profiles and behavior-only prediction models. It combines observable traces with structured semantic indicators extracted from learner text. This

integration connects learner activity with expressed reasoning, planning, communication, and emotional cues. The framework also supports temporal updates and partial profiles when some data are unavailable.

The empirical results show three main strengths. First, the hybrid behavioral profile reaches the highest AUC, at 0.871. Second, cognitive complexity and emotional tone show no correlation ($r = 0.00$). This result supports their complementary role within the semantic profile. Third, planning cues were detected in 92.3% of the valid Strategist messages. These findings reveal distinctions that remain difficult to capture through single-source profiles. Unlike opaque prediction scores, each profile dimension has a clear pedagogical meaning. This structure supports more transparent educational decisions. The main contribution, therefore, lies in combining prediction, interpretation, and temporal profiling into a single coherent framework.

The study also has important limitations. OULAD and Stanford MOOC Posts do not include the same learners. Therefore, real end-to-end fusion remains untested on linked data. Synthetic archetypes also simplify the diversity of real learners. LLM annotations depend on text quality, prompt design, language, and model bias. Future work will validate these annotations with human experts and linked real-world data. It will also assess computational cost, scalability, and integration within a real LMS. These steps will determine the framework's value for adaptive support and educational decision-making.

REFERENCES

-
- Ait-Adda, S., Bousbia, N., & Balla, A. (2023). Enriching the learner's model through the semantic analysis of learning traces. *E-Learning and Digital Media*, 20(1), 1–24. <https://doi.org/10.1177/20427530221102993>
- Ali, A. D., & Hanna, W. K. (2022). Predicting students' achievement in a hybrid environment through self-regulated learning, log data, and course engagement: A data mining approach. *Journal of Educational Computing Research*, 60(4), 960–985. <https://doi.org/10.1177/07356331211056178>
- Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2023). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, 55(6), 3026–3054. <https://doi.org/10.3758/s13428-022-01939-9>
- Barthakur, A., Dawson, S., & Kovanovic, V. (2023). Advancing learner profiles with learning analytics: A scoping review of current trends and challenges. *Proceedings of the 13th International Learning Analytics and Knowledge Conference* (pp. 606–612). Association for Computing Machinery. <https://doi.org/10.1145/3576050.3576083>
- Bauer, J. (2022). A primer to latent profile and latent class analysis. In M. Goller, E. Kyndt, S. Paloniemi, & C. Damşa (Eds.), *Methods for researching professional learning and development: Challenges, applications and empirical illustrations* (pp. 243–268). Springer. https://doi.org/10.1007/978-3-031-08518-5_11
- Capone, R., & Lepore, M. (2022). From distance learning to integrated digital learning: A fuzzy cognitive analysis focused on engagement, motivation, and participation during the COVID-19 pandemic. *Technology, Knowledge and Learning*, 27(4), 1259–1289. <https://doi.org/10.1007/s10758-021-09571-w>
- Chen, J., Zhang, L. J., & Chen, X. (2026). L2 learners' self-regulated learning strategies and self-efficacy for writing achievement: A latent profile analysis. *Language Teaching Research*, 30(1), 14–35. <https://doi.org/10.1177/13621688221134967>
- David, S., Meidan, Y., Hersko, I., Varnovitzky, D., Mimran, D., Elovici, Y., & Shabtai, A. (2025). *ProfilLLM: An LLM-based framework for implicit profiling of chatbot users*. arXiv. <https://doi.org/10.48550/arXiv.2506.13980>
- Dowell, N., & Kovanović, V. (2022). Modeling educational discourse with natural language processing. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *The handbook of learning analytics* (2nd ed., pp. 105–119). Society for Learning Analytics Research. <https://doi.org/10.18608/hla22.011>

- Escolà-Gascón, Á., & Gallifa, J. (2022). How to measure soft skills in the educational context: Psychometric properties of the skills-in-one questionnaire. *Studies in Educational Evaluation*, 74, 101155. <https://doi.org/10.1016/j.stueduc.2022.101155>
- Ezaldeen, H., Misra, R., Bisoy, S. K., Alatrash, R., & Priyadarshini, R. (2022). A hybrid e-learning recommendation integrating adaptive profiling and sentiment analysis. *Journal of Web Semantics*, 72, 100700. <https://doi.org/10.1016/j.websem.2021.100700>
- Ezzaim, A., Dahbi, A., Aqqal, A., & Haidine, A. (2025). AI-based learning style detection in adaptive learning systems: A systematic literature review. *Journal of Computers in Education*, 12(3), 731–769. <https://doi.org/10.1007/s40692-024-00328-9>
- Fazil, M., Rísquez, A., & Halpin, C. (2024). A novel deep learning model for student performance prediction using engagement data. *Journal of Learning Analytics*, 11(2), 23–41. <https://doi.org/10.18608/jla.2024.7985>
- He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., & Chen, W. (2024). AnnoLLM: Making large language models to be better crowdsourced annotators. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 6, pp. 165–190). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-industry.15>
- Hernández-de Menéndez, M., Morales-Menendez, R., Escobar, C. A., & Ramírez Mendoza, R. A. (2022). Learning analytics: State of the art. *International Journal on Interactive Design and Manufacturing*, 16(3), 1209–1230. <https://doi.org/10.1007/s12008-022-00930-0>
- Jin, K., Xu, W., & Zhong, X. (2025). A multi-dimensional synthesis of factors influencing e-learning continuance intention: A systematic review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 20, Article 28. <https://doi.org/10.28945/5614>
- Jovanovic, J., Gašević, D., Yan, L., Baker, G., Murray, A., & Gasevic, D. (2024). Explaining trace-based learner profiles with self-reports: The added value of psychological networks. *Journal of Computer Assisted Learning*, 40(4), 1481–1499. <https://doi.org/10.1111/jcal.12968>
- Kara, A., Ergulec, F., & Eren, E. (2024). The mediating role of self-regulated online learning behaviors: Exploring the impact of personality traits on student engagement. *Education and Information Technologies*, 29(17), 23517–23546. <https://doi.org/10.1007/s10639-024-12755-3>
- Katsarou, E., & Chatzipanagiotou, P. (2025). Examining the association of personality traits and grit on Greek students' wellbeing in higher education. *Education Sciences*, 15(1), 57. <https://doi.org/10.3390/educsci15010057>
- Kew, S. N., & Tasir, Z. (2022). Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning*, 27(2), 405–427. <https://doi.org/10.1007/s10758-021-09541-2>
- Khalil, M., Prinsloo, P., & Slade, S. (2023). The use and application of learning theory in learning analytics: A scoping review. *Journal of Computing in Higher Education*, 35(3), 573–594. <https://doi.org/10.1007/s12528-022-09340-3>
- Khanna, P., Mathieu, K., Norberg, K., Almoubayyed, H., & Fancsali, S. E. (2025). Math content readability, student reading ability, and behavior associated with gaming the system in adaptive learning software. In C. Mills, G. Alexandron, D. Taibi, G. Lo Bosco, & L. Paquette (Eds.), *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 465–472). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.15870268>
- Law, N., & Liang, L. (2020). A multilevel framework and method for learning analytics integrated learning design. *Journal of Learning Analytics*, 7(3), 98–117. <https://doi.org/10.18608/jla.2020.73.8>
- Li, L., Johnson, J., Aarhus, W., & Shah, D. (2022). Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. *Computers & Education*, 176, 104354. <https://doi.org/10.1016/j.compedu.2021.104354>

- Li, M., Ge, M., Zhao, H., & An, Z. (2022). Modeling and analysis of learners' emotions and behaviors based on online forum texts. *Computational Intelligence and Neuroscience*, 2022, 9696422. <https://doi.org/10.1155/2022/9696422>
- Li, S., Huang, X., Wang, T., Pan, Z., & Lajoie, S. P. (2022). Examining the interplay between self-regulated learning activities and types of knowledge within a computer-simulated environment. *Journal of Learning Analytics*, 9(3), 152–168. <https://doi.org/10.18608/jla.2022.7571>
- Liu, M., Cai, Y., Han, S., & Shao, P. (2022). Understanding student navigation patterns in game-based learning. *Journal of Learning Analytics*, 9(3), 50–74. <https://doi.org/10.18608/jla.2022.7637>
- Liu, Z., Tang, Q., Ouyang, F., Long, T., & Liu, S. (2024). Profiling students' learning engagement in MOOC discussions to identify learning achievement: An automated configurational approach. *Computers & Education*, 219, 105109. <https://doi.org/10.1016/j.compedu.2024.105109>
- Markowitz, D. M. (2024). Can generative AI infer thinking style from language? Evaluating the utility of AI as a psychological text analysis tool. *Behavior Research Methods*, 56(4), 3548–3559. <https://doi.org/10.3758/s13428-024-02344-0>
- Michael, J., Santoso, H. B., Junus, K., & Sandiwarno, S. (2025). Systematic review of self-regulated learning intervention and measurement in higher education: Towards a holistic, integrated, and technology-assisted model. *Journal of Information Technology Education: Research*, 24, Article 33. <https://doi.org/10.28945/5636>
- Nasution, A. H., & Onan, A. (2024). ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. *IEEE Access*, 12, 71876–71900. <https://doi.org/10.1109/ACCESS.2024.3402809>
- Nuryani, Y., Budi, I., & Hasibuan, Z. A. (2024). BERT-based model and LLMs-generated synthetic data for conflict sentiment identification in aspect-based sentiment analysis. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19, 511–537. <https://doi.org/10.28945/5439>
- Ofori, E., & Dake, D. K. (2026). Deep embedded clustering and statistical validation for student modality profiling. *Interdisciplinary Journal of Information, Knowledge, and Management*, 21. <https://doi.org/10.28945/5724>
- Osei, H. V., Kwateng, K. O., & Boateng, K. A. (2022). Integration of personality trait, motivation and UTAUT 2 to understand e-learning adoption in the era of the COVID-19 pandemic. *Education and Information Technologies*, 27, 10705–10730. <https://doi.org/10.1007/s10639-022-11047-y>
- Saqr, M., Schreuder, M. J., & López-Pernas, S. (2024). Why educational research needs a complex system revolution that embraces individual differences, heterogeneity, and uncertainty. In M. Saqr & S. López-Pernas (Eds.), *Learning analytics methods and tutorials* (pp. 723–734). Springer. https://doi.org/10.1007/978-3-031-54464-4_22
- Sassi Hidri, M. (2024). Learning-based models for building user profiles for personalized information access. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19, Article 10. <https://doi.org/10.28945/5275>
- Shahid, F., Hsu, M.-H., Chang, Y.-C., & Jian, W.-S. (2025). Using generative AI to extract structured information from free text pathology reports. *Journal of Medical Systems*, 49, Article 36. <https://doi.org/10.1007/s10916-025-02167-2>
- Sharma, P., Akgun, M., & Li, Q. (2024). Understanding student interaction and cognitive engagement in online discussions using social network and discourse analyses. *Educational Technology Research and Development*, 72(5), 2631–2654. <https://doi.org/10.1007/s11423-023-10261-w>
- Subramaniam, T. T., & Suhaimi, N. A. D. (2022). Learner profiling: A study on big five personality traits and lifelong learning skills. *Journal of E-Learning and Knowledge Society*, 18(2), 50–57.
- Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large language models for data annotation and synthesis: A survey. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 930–957). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.54>

- Teng, L. S. (2024). Individual differences in self-regulated learning: Exploring the nexus of motivational beliefs, self-efficacy, and SRL strategies in EFL writing. *Language Teaching Research*, 28(2), 366–388. <https://doi.org/10.1177/13621688211006881>
- Thomas, R. K., Strekalova-Hughes, E., Nash, K. T., Holley, M., Warner, C. K., Enochs, B., Prendergast, P., & Ricklefs, M. (2023). The learner profile: Piloting a tool to support contextualized understanding of the learner. *Journal of Early Childhood Teacher Education*, 44(3), 349–372. <https://doi.org/10.1080/10901027.2022.2047840>
- Tuononen, T., Hyytinen, H., Räisänen, M., Hailikari, T., & Parpala, A. (2023). Metacognitive awareness in relation to university students' learning profiles. *Metacognition and Learning*, 18, 37–54. <https://doi.org/10.1007/s11409-022-09314-x>
- Winter, M., Mordel, J., Mendzheritskaya, J., Biedermann, D., Ciordas-Hertel, G.-P., Hahnel, C., Bengs, D., Wolter, I., Goldhammer, F., Drachsler, H., Artelt, C., & Horz, H. (2024). Behavioral trace data in an online learning environment as indicators of learning engagement in university students. *Frontiers in Psychology*, 15, Article 1396881. <https://doi.org/10.3389/fpsyg.2024.1396881>
- Ye, D., & Pennisi, S. (2022). Using trace data to enhance students' self-regulation: A learning analytics perspective. *The Internet and Higher Education*, 54, 100855. <https://doi.org/10.1016/j.iheduc.2022.100855>
- Yee, M., Roy, A., Perdue, M., Cuevas, C., Quigley, K., Bell, A., Rungta, A., & Miyagawa, S. (2023). AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums. *Frontiers in Education*, 8, Article 1250846. <https://doi.org/10.3389/educ.2023.1250846>
- Yu, L., Halpin, P. F., Bernacki, M. L., Ren, S., Plumley, R. D., & Greene, J. A. (2025). Interpreting predictive learning sequences in a college math course through a self-regulated learning framework. *Journal of Learning Analytics*, 12(3), 66–86. <https://doi.org/10.18608/jla.2025.8865>
- Zankadi, H., Idrissi, A., Daoudi, N., & Hilal, I. (2023). Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. *Education and Information Technologies*, 28(5), 5567–5584. <https://doi.org/10.1007/s10639-022-11373-1>
- Zhang, J., Borchers, C., Aleven, V., & Baker, R. S. (2024). Using large language models to detect self-regulated learning in think-aloud protocols. In B. Paaßen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 157–168). International Educational Data Mining Society. <https://doi.org/10.35542/osf.io/hrtz6>

APPENDIX

This Appendix reports the prompts used in all modules. The prompts are reproduced in their original form.

STANFORD SEMANTIC ANNOTATION PROMPT

The following prompt is used to annotate learner-generated forum messages:

Instruction prompt

You are an expert annotator in Learning Analytics.

Analyze only the information explicitly expressed in the student message.

Do not infer hidden traits or intentions.

Use null when the text provides insufficient evidence.

Return one valid JSON object with the following fields:

1. "cognitive_complexity":

Integer from 1 to 5, or null.

1 = simple statement or basic question.

2 = limited explanation.

- 3 = clear reasoning with some detail.
- 4 = detailed reasoning connecting several ideas.
- 5 = complex reasoning, evaluation, or abstraction.
- 2. "error_type":
 - "None", "Conceptual", "Careless", "Technical", or null.
 - Conceptual = misunderstanding of a concept.
 - Careless = minor mistake despite apparent understanding.
 - Technical = platform, software, or access problem.
 - None = no error is expressed.
- 3. "autoreg_planning":
 - 1 only when explicit planning, monitoring, goal setting, or strategy adjustment appears.
 - 0 when such evidence is absent.
- 4. "engagement_cognitive":
 - Integer from 1 to 5, or null.
 - 1 = minimal intellectual effort.
 - 3 = active explanation or problem solving.
 - 5 = sustained analysis, reflection, or strategy use.
- 5. "comm_clarity":
 - Integer from 1 to 5.
 - 1 = difficult to understand.
 - 3 = generally understandable.
 - 5 = clear and well structured.
- 6. "comm_tone":
 - Float from -1.0 to 1.0.
 - 1.0 = clearly negative.
 - 0.0 = neutral.
 - 1.0 = clearly positive.

Output only the JSON object.

No explanation.

Input wrapper

<s>[INST] {annotation_prompt}

Student Message: "{student_message}" [/INST]

SYNTHETIC FORUM-POST GENERATION

The following system prompt generated short forum messages for the controlled validation.

System prompt

You are a student participating in an online course forum.

Write one natural student message of one or two sentences.

Use the following situation:

Knowledge level: {knowledge_level}

Emotional state: {emotional_state}

Planning behavior: {planning_behavior}

Social behavior: {social_behavior}

Current difficulty: {difficulty}

Do not mention a profile, persona, archetype, or simulation.

Do not list the attributes.

Do not explain the situation.

Express only signals that a real student could naturally communicate.

Output only the student message.

Input wrapper

<s>[INST] {synthetic_generation_prompt} [/INST]

SYNTHETIC SEMANTIC VALIDATION PROMPT

The following prompt analyzed each generated forum message. The annotation input contained only the generated message. It did not contain the pedagogical archetype label.

Instruction prompt

You are an expert annotator in Learning Analytics.

Analyze only evidence explicitly expressed in the student message.

Do not infer stable personality traits.

Do not infer facts that are absent from the message.

Return exactly one valid JSON object with these six keys.

1. "cognitive_complexity": integer from 1 to 5.
 - 1 = simple statement or basic question without reasoning.
 - 2 = limited explanation or one simple relation.
 - 3 = clear reasoning using relevant concepts.
 - 4 = detailed reasoning connecting several concepts.
 - 5 = deep analysis, evaluation, comparison, or abstraction.
2. "error_type": one of:
 - "None", "Conceptual", "Careless", or "Technical".
 - "None" = no explicit error is present.
 - "Conceptual" = misunderstanding of a learning concept.
 - "Careless" = minor mistake despite apparent understanding.
 - "Technical" = platform, software, access, or tool problem.
3. "autoreg_planning": integer 0 or 1.
 - 1 = explicit planning, monitoring, goal setting, time management, strategy selection, or strategy adjustment.
 - 0 = no explicit evidence of these processes.
4. "engagement_cognitive": integer from 1 to 5.
 - 1 = minimal or passive intellectual involvement.
 - 2 = basic interest or limited effort.
 - 3 = active explanation, questioning, or problem solving.
 - 4 = sustained reasoning or visible effort.
 - 5 = reflection, evaluation, or explicit strategy use.

5. "comm_clarity": integer from 1 to 5.
 1 = very difficult to understand.
 2 = partly unclear.
 3 = generally understandable.
 4 = clear.
 5 = very clear and well structured.
6. "comm_tone": float from -1.0 to 1.0.
 -1.0 = strongly negative, angry, or frustrated.
 -0.5 = moderately negative.
 0.0 = neutral or factual.
 0.5 = moderately positive.
 1.0 = strongly positive or enthusiastic.

Output only the JSON object.

Do not add explanations.

Do not add markdown.

Do not add extra keys.

Input wrapper

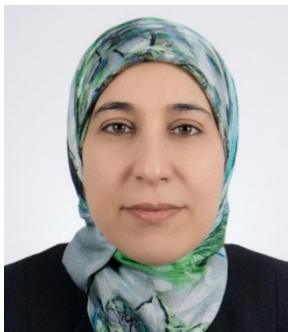
<s>[INST] {prompt}

Message: '{text}' [/INST]

AUTHORS



Salma El Omari received her Master's degree in Cloud and High-Performance Computing from the National School of Computer Science and Systems Analysis (ENSIAS) in Morocco in 2022. She is currently a PhD candidate in the Department of Computer Science at Sultan Moulay Slimane University, Morocco. Her primary research interests lie at the intersection of artificial intelligence and education, with a specific focus on the design of recommender systems, the application of reinforcement learning algorithms, and the development of personalized learning technologies.



Najlae Idrissi is a Full Professor in the Computer Science Department at the Faculty of Sciences and Techniques, Sultan Moulay Slimane University in Beni Mellal, Morocco. She is the head of the Intelligence, Data, and Computing Team within the same institution. Professor Idrissi has directed and continues to supervise numerous doctoral theses across various domains of computer science. Her extensive research interests encompass the application of artificial intelligence in education, medical image processing, wireless sensor networks, driver fatigue detection using deep learning algorithms, and the development of adaptive intelligent systems. She is an academic editor for several renowned journals and also the founder and general chair of ICAIGC and ISIVC 2026 symposium.