Assessing LLMs as Cognitive Interpreters of Student Prompts: A Typological Framework

Tadeu da Ponte
Alma Mater Europaea University
Slovenska ulica 17, 2000 Maribor, Slovenia

Matevz Vremec
Alma Mater Europaea University
Slovenska ulica 17, 2000 Maribor, Slovenia

Matej Mertik
Alma Mater Europaea University
Slovenska ulica 17, 2000 Maribor, Slovenia

ABSTRACT

This paper introduces a typology of student cognitive actions in interactions with large language model (LLM)-based tutors. Drawing on the CoMTA dataset of 188 anonymized math tutoring dialogues from Khan Academy, student-generated questions were analyzed as evidence of reasoning processes. The methodology combines a natural language processing (NLP) pipeline for semantic clustering with a dual-stage human classification of communicative intent and cognitive action.

The resulting typology is synthesized into a partially ordered taxonomy that captures the complexity and multidimensionality of student thinking in AI-mediated learning contexts. Two research questions guide this investigation: (1) Can a typology be derived directly from unsupervised NLP clustering methods? and (2) To what extent can LLMs replicate expert-driven classification schemes?

Findings from RQ1 reveal that semantic clustering via PCA and KMeans offers only limited alignment with pedagogically meaningful distinctions. In contrast, results from RQ2 show that several LLMs–particularly Deepseek, Grok, and Gemini–can reliably

extend the typology to unseen data, demonstrating high accuracy in classification. These results suggest that scalable, cognitively informed AI tutoring may be supported by combining expert frameworks with strategically configured LLM architectures.

General Terms

Educational Technology, Artificial Intelligence

Keywords

LLM-based tutoring, cognitive typology, prompt classification, Khanmigo, educational interaction

1. INTRODUCTION

Traditional educational research methods rely on presenting learners with standardized prompts-either open-ended or multiple-choice questions-and analyzing their responses. These methods ensure that students begin from the same cognitive stimulus, making comparative analysis across individuals and groups feasible. As a result, the analytical task is typically centered on categorizing responses to the same question [24].

While this approach promotes consistency, replicability, and potential generalization, it also constrains the cognitive scope of students' responses. Once a question is posed, three immediate limitations emerge:

(1) **Reactive Cognition:** The student's thinking becomes predominantly responsive to the stimulus, rather than originating from spontaneous inquiry.
(2) **Constrained Thinking:** The structure of the question encourages convergent thinking-directed at identifying a single correct answer-while discouraging divergent exploration.
(3) **Recall Emphasis:** The working memory is directed toward retrieving information from long-term memory, which may inhibit the construction of new knowledge from exploratory engagement with external sources.

These cognitive dynamics have been documented across various studies. Multiple-choice formats tend to elicit recognition-based processing and recall [9], while open-ended questions foster elaboration and integration but still restrict spontaneity [12]. In contrast, student-generated questioning has been shown to activate higher levels of metacognition, curiosity, and ownership over the learning process [30, 3].

With the integration of large language models (LLMs) into education, new data sources for understanding student cognition are emerging. Systems like Khan Academy's Khanmigo now enable researchers to collect and analyze student-generated questions in authentic learning environments. This shift aligns closely with the growing focus on AI literacy-defined by Long and Magerko [14] as "a set of competencies that enable individuals to critically evaluate, interact with, and apply artificial intelligence."

AI literacy is increasingly seen as a core 21st-century skill. Its development, particularly in K-12 contexts, is supported through various pedagogical approaches, including digital storytelling [18, 5], project-based and game-based learning [10], and the use of accessible tools like Scratch or Teachable Machine [10].

Table 1. : Comparison of Student Cognitive Behaviors by Activity Type

Activity	Cognitive Characteristics	Implications
Answering Multiple-Choice Questions	Focuses on recall, recognition, and test-taking strategies [9]	Efficient assessment of content knowledge but limits creative engagement or process insight
Responding to Open-Ended Questions	Allows for elaboration and explanation [12]	Encourages expressive depth but constrained by the prompt's framing
Generating Questions	Involves metacognition, reflection, and transfer [30, 3]	Promotes learner agency, exploration, and diagnostic insight into cognitive processes

However, challenges remain in integrating ethical considerations, training teachers, and aligning curriculum frameworks [29, 15]. To address these, validated frameworks and assessment tools are being developed [6]. Recent contributions such as the OECD's AI Capability Indicators [19] and the AI Literacy Framework [7] offer structured models to define, benchmark, and assess AI-related competencies across cognitive, social, and ethical domains in education.
This study contributes to this emerging field by offering a fine-grained analysis of student behavior through question generation in LLM-mediated tutoring. Rather than assessing how students respond to questions, how they formulate them was analyzed. This lens provides insight into spontaneous reasoning patterns, enabling a more behavioral grounding of AI literacy competencies.
To support this inquiry, the CoMTA dataset introduced by Miller and Dicerbo [17], which contains 188 anonymized student-tutor dialogues captured from real use of Khanmigo, was used. These conversations include moments of clarification, confusion, exploration, and self-reflection-captured not through static responses, but through the dynamic generation of queries.
Given this context, new typologies and taxonomies are needed to model student thinking as expressed in self-formulated questions. While the science of educational assessment has long focused on how to code responses, it is now time to develop structured frameworks for analyzing student-initiated inquiry. Such models are essential to fully leverage the pedagogical possibilities offered by AI-supported learning environments.
One important application of such typologies and taxonomies is in refining the behavior of LLM-based tutors. Rather than merely providing correct answers, these systems can be guided to respond in ways that support the pedagogical intentions underlying the student's inquiry. As highlighted in Miller and Dicerbo's CoMTA report, "Effective tutoring involves helping students find the answers themselves, which can be difficult for LLMs that are optimized to provide direct answers" [17]. Understanding the kinds of questions students pose-and their cognitive implications-can help design AI-driven tutors that not only address students' requests but also steer the interaction toward higher-order thinking goals. In this sense, recognizing the structure and type of questions students formulate enables the LLM to prompt the student back, nudging them into a more intentional and cognitively productive line of inquiry.
This study is guided by two central research questions:

RQ1 Is it possible to create a typology of student questions directly through natural language processing models? What are the opportunities and limitations of this approach?

RQ2 To what extent can LLMs accurately classify student questions based on a human-derived taxonomy?

These questions guide the methodological design and inform the broader objective: to examine how student cognition is expressed through question generation and how this behavior can be modeled and enhanced in AI-mediated educational environments.

2. METHODOLOGY

This study employs a dual-method approach to examine how student cognition is expressed through question generation in LLM-mediated tutoring contexts. The methodological workflow unfolds in three main phases: (1) extraction and expert annotation of 212 student questions from the CoMTA dataset, resulting in a typology of 10 question types mapped to 6 cognitive actions; (2) evaluation of whether unsupervised NLP techniques-specifically BERT-based sentence embeddings, PCA reduction, and clustering-can recover cognitively meaningful groupings (RQ1); and (3) testing whether 10 commercial LLMs can generalize and apply the expert-designed typology to classify new student queries (RQ2). This design enables a comparative analysis between bottom-up computational discovery and top-down application of pedagogical structure, offering insight into how AI systems may support cognitive modeling in education.
To implement this workflow, 188 real-world math tutoring dialogues released by Khan Academy through the CoMTA benchmark [17] was analyzed. A custom natural language processing (NLP) pipeline was developed in Python to extract and preprocess student-generated questions from the JSON-formatted dataset. The following subsections detail each methodological component.

2.1 Data Extraction and Preprocessing

To begin the analysis, a Python-based pipeline to process the CoMTA dataset was developed, comprising structured dialogues between students and an LLM-based tutor. Each record contains a sequence of conversational turns, with each message tagged by the speaker's role (i.e., `"user"` for the student and `"assistant"` for the tutor).
The script traverses the dataset and extracts only utterances authored by students. A custom NLP routine was employed to detect question-like content, even in cases where canonical punctuation was absent. Specifically, a set of regular expression filters wa applied targeting lexical indicators of inquiry such as *how*, *why*, *what*, and related forms. These interrogative patterns are consistent with established methods in question classification [13], enabling the identification of epistemic intent regardless of syntactic formality.
As an illustrative case, Appendix A presents the full transcript of a sample session (test_id = 1) from the CoMTA dataset. From this session, the extraction method successfully identified multiple student inquiries, including:

> *"I'm struggling with multiplying decimal numbers, can you guide me through the process?"*
>
> *"What is the rule on where to put the decimal?"*
>
> *"So we multiply the whole numbers and then move the decimal to the left according to how many numbers are to the right of the decimal?"*

These examples highlight the method's ability to capture both direct questions and interrogative intent embedded in declarative syntax. This preprocessing step was essential for isolating semantically rich prompts and enabling subsequent cognitive classification.

2.2 NLP Embedding and Clustering Analysis

To explore semantic regularities in student queries and facilitate interpretability, a natural language processing pipeline was implemented using pre-trained sentence embeddings and clustering. Each student-authored utterance (role: "user") was embedded using the `all-MiniLM-L6-v2` model from the Sentence-BERT framework [25], which is optimized for semantic similarity tasks in short textual inputs.
The resulting 384-dimensional sentence vectors were projected onto a two-dimensional space using Principal Component Analysis (PCA) [11], enabling visual analysis and dimensionality reduction. Clustering was performed using the KMeans algorithm [16] with $k = 6$ to uncover underlying patterns in student behavior.

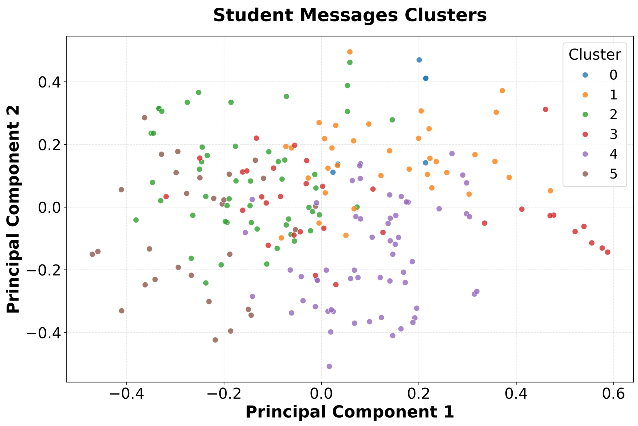

Fig. 1: PCA-based visualization of clustered student messages. Colors represent six cognitive behavior clusters.

The extracted questions were then subjected to a two-tiered human classification protocol, described below.

2.3 Cluster Inspection and Limitations of Semantic Groupings

Following the generation of semantic clusters, an exploratory inspection of the questions within each group to assess their potential for typological categorization was conducted. While some clusters revealed superficial thematic cohesion-such as questions referencing similar mathematical topics or phrased using structurally analogous sentence patterns-these groupings did not support consistent or meaningful distinctions in terms of cognitive function.
This limitation stems in part from the nature of the embedding and clustering algorithms used. Both Sentence-BERT embeddings and the PCA projection capture semantic proximity based on surface-level similarity in lexical and syntactic features. As a result, the clusters primarily reflect linguistic patterns or content overlap rather than the underlying reasoning strategies or cognitive processes driving student inquiry.

For instance, certain clusters included multiple questions about trigonometric identities or derivatives, but these were posed from varied epistemic standpoints-ranging from rote answer verification to conceptual clarification or reflective engagement. Therefore, while the clustering step provided useful visualization and preliminary organization, it did not yield a basis robust enough to construct a typology of cognitive actions.
In light of this, it was decided not to proceed with interpreting or labeling clusters as distinct cognitive profiles. Instead, it was shifted to a general typological analysis of all filtered questions in the dataset, treating each as an independent unit for human coding of communicative intent and cognitive function.

2.4 Question Typing and Cognitive Classification

The elaboration of a typology from authentic student prompts involves a multi-step interpretive process. In this study, the first stage followed a grounded and iterative coding approach:

(1) Initially, a random sample comprising approximately 20% of the student questions was read and analyzed. Each question received a brief natural language description that aimed to capture the behavioral intent behind the utterance with high fidelity.
(2) Next, descriptions with similar meanings were grouped and their phrasing refined to ensure semantic consistency across items. These generalized labels were used to define the initial set of question types-an abstraction that maps diverse expressions to common communicative goals.
(3) The standardized types were then applied to the remaining 80% of questions. During this stage, one additional type emerged-*Questioning the tutor's behavior*-not previously identified in the initial sample.
(4) A final review was conducted to verify that no conceptual information was lost, distorted, or introduced inappropriately through the generalization. All types were re-evaluated in light of the full dataset.

The resulting typology consists of 10 descriptive categories that capture student communicative intent during LLM-tutored interactions:

Asking for guidance
Asking for problems to practice
Asking for an answer
Asking for assessment
Asking for alternative solution
Inviting to discuss concepts
Checking correctness
Checking understanding
Checking next step
Questioning tutor's behavior

Table 2 provides representative examples of each type. These illustrate the diversity of interaction intents and the level of abstraction used to define categories.
In a second phase of analysis, each question type was associated with a broader *cognitive action*, inferred from the underlying intention and epistemic role of the question. This stage moves beyond surface formulation to abstract the underlying cognitive strategies guiding the student's interaction. The cognitive action categories are:

Table 2. : Representative examples of question types used in classification

ID	Message	Type
122	so its -1/2(1)^2?	Checking correctness
122	so should i write it as -x/2(x^2)?	Checking correctness
126	Is this problem a sine or cosine function?	Asking for guidance
97	To determine the value of x?	Checking understanding
116	f(x) = sin x tan x can you provide a similar problem to this one?	Asking for problems to practice
135	cot(theta)(1-2)?	Checking correctness
143	well in my class we just use the established rule that sin^2(x) + cos^2(x) = 1. that means we can transform the left side into 1. i already know all the formulas, i want to try applying them. can you give me a hard problem?	Asking for problems to practice
126	How do you identify if you are dealing with a sine function and cosine function?	Asking for guidance
182	Given the functions f(x)= x^2-4 and g(x)= $\sqrt[3]{x}$-1. Can you tell me what (f-g)(x) and (f*g)(x) would be?	Asking for an answer
181	Can you help me find the limit as x approaches 1 for the equation (1-x+ln(x))/(1+cos(5pi x)?	Asking for guidance
179	Consider a function f where f'(x) = sin x^2 and f(0) = 0. Can you tell me the first three nonzero terms of the Maclaurin series for f?	Asking for an answer
132	I'm stuck on a problem in my homework. The question asks if the infinite series of ln(1+ 1/n^2) converges or diverges. I've hit a snag though. The instructions say to use the limit comparison test, but when I tried to find the limit of the original sequence divided by the series I'm comparing it to, I got -infinity. I'm not sure if I made a mistake or if that's the correct answer, and if it is, what does it imply?	Checking understanding
78	do i multiply it?	Checking next step
16	multiply by x?	Checking next step
109	is there a quicker way to solve this?	Asking for alternative solution
109	yes synthetic division, can we proceed step by step?	Checking next step
8	I'm 30 years old and have taken a break from education for some time. Now, I've decided to return and pursue a degree in computer science. However, I'm concerned that my math skills are equivalent to a 9th grader's. Before I delve into computer science, I want to enhance my math skills to a college level. Can you help me determine my current math level?	Asking for assessment
19	can you demonstrate the x-factor technique (not explain)?	Asking for alternative solution
77	Can you provide me with 10 basic arithmetic questions?	Asking for problems to practice
153	A farmer plans to enclose a rectangular plot of 17800 square feet in a pasture and then bisect the area with a fence running parallel to one side. What is the minimum amount of fencing that will be needed to accomplish this?	Asking for an answer
174	Hello there, Khanmigo! Could we discuss some equations related to economics?	Inviting to discuss concepts
19	why are you refusing to demonstrate?	Questioning tutors behavior

Confrontation

Extraction

Orientation

Confirmation

Exploration

Reflection

These actions represent high-level abstractions derived from the behavioral typology. Each type maps to a single cognitive action, as defined through iterative manual coding. For example:

(1) *"I'm struggling with multiplying decimal numbers, can you guide me through the process?"*
Type: Asking for guidance
Cognitive action: Orientation

(2) *"So we multiply the whole numbers and then move the decimal to the left according to how many numbers are to the right of the decimal?"*
Type: Checking understanding
Cognitive action: Confirmation

(3) *"What is the rule on where to put the decimal?"*
Type: Asking for guidance
Cognitive action: Orientation

This two-step approach-first typifying behavioral intent, then abstracting cognitive action-offers a structured lens to analyze student behavior in LLM-facilitated learning. These classifications will later serve as the foundation for evaluating trends, proposing taxonomic models, and informing AI tutor design.

2.5 Methodological Design for Typology Generation and Classification Testing

To address the two research questions, distinct but complementary methodological procedures were implemented. Each targets a different aspect of typology and taxonomy creation and application in AI-supported educational environments, but both were designed to investigate a specific research aim while ensuring clarity, interpretability, and replicability. The results of the analyses described below are presented in Section 4.

RQ1: Assessing the Interpretability of NLP + PCA Clustering

To address the first research question-whether a typology of student questions can emerge from unsupervised NLP methods-dimensionality reduction and clustering techniques to semantically embed and organize the dataset was applied. Specifically, the `all-MiniLM-L6-v2` model from Sentence-BERT to generate 384-dimensional sentence embeddings for each student-authored question was used. These high-dimensional vectors were then projected into two dimensions using Principal Component Analysis (PCA), enabling both visualization and further analysis.
The reduced vectors were clustered using the KMeans algorithm with $k = 6$, a configuration chosen to allow for manageable interpretability while preserving variability in student behavior. Each student question was thus assigned to one of six PCA-derived clusters. These cluster assignments were then compared to the `Type` and `Cognitive Action` categories assigned through expert manual coding.
To evaluate how well the unsupervised clustering aligned with the human-defined labels, a contingency matrix (also known as a confusion matrix) was constructed. This matrix tabulates the frequency with which each cluster overlaps with each expert-assigned category, thereby visualizing the degree of agreement or divergence.
To quantify this alignment, the F1 Score for each cluster-category pairing was computed. The F1 Score is defined as the harmonic mean of two metrics: precision and recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Here, precision is the proportion of correct predictions among all instances assigned to a particular category by the clustering

algorithm, while recall is the proportion of correct predictions among all instances that truly belong to that category.
The F1 Score provides an intuitive and balanced measure of performance. High scores indicate strong alignment between the PCA-derived clusters and the manually constructed typology, suggesting that the unsupervised method may hold interpretive value. Conversely, lower scores highlight limitations in using clustering alone to capture cognitively meaningful groupings. This analysis enables a systematic exploration of the potential and boundaries of using PCA and NLP embeddings for typological generation.

RQ2: Evaluating LLMs' Ability to Generalize a Human-Derived Typology

The second research question evaluates the capacity of various commercially available large language models (LLMs) to apply a predefined pedagogical typology at scale. The approach simulated a real-world scenario in which an educational specialist manually annotates a small portion of student questions-20% of the dataset (42 out of 212 questions)-while an LLM is tasked with classifying the remaining 80% based on the provided labels and instructions.
The goal was not generativity but rather interpretation: can the LLMs faithfully extend a pedagogically meaningful classification scheme to unseen examples? Such capacity is foundational for future intelligent tutors that not only respond to student queries but also interpret them in light of structured educational goals.
The LLMs received the following prompt:

Instruction Prompt for LLMs

”You are an educational-NLP classifier, and you are an expert in cognitive science with teaching and learning experience.
A. I will give you an Excel file that contains 212 student questions.
B. Rows 1–42 are already labeled in column `"Type"` with the ten categories of my custom typology.
C. Rows 43–212 have an empty `"Type"` cell.
Task:

(1) Learn each category purely from the 42 labeled examples (do not invent new categories).
(2) For every unlabeled question (rows 43–212), assign exactly one of those same category labels.
(3) Write the completed dataset back to me as a CSV with the original columns plus the filled-in `"Type"` column.
(4) Preserve all text exactly; only add the new labels.

Log meta process information such as your reasoning process and store the time that you took to do the task. By the end, give me the CSV file as required above and a .txt file with a summary in English of this meta process information, including your complete credit of your model information to insert in my paper.”

Ten LLMs were tested in this task. Their methodologies and system descriptions are summarized below for comparison:

ChatGPT-4o - Applied a supervised classification approach using a k-Nearest Neighbors (k=3) algorithm on TF-IDF vectorized features (1-2 n-grams). Preprocessing included stopword removal. Executed using Python (scikit-learn and pandas) under guidance of the ChatGPT-4o model [20].

ChatGPT-o3 - Implemented a TF-IDF + Logistic Regression pipeline. The model was responsible for analyzing the initial 42 labeled samples and scripting the classification code. The execution was done locally with Python and scikit-learn [21].

ChatGPT-o4-mini - Followed the same logistic regression pipeline as o3. TF-IDF was used for vectorization. Model credited: ChatGPT o4-mini [22].

ChatGPT-o4-miniHigh - Variant of o4-mini, using the same TF-IDF + Logistic Regression setup for classification. Model credit: ChatGPT o4-miniHigh [23].

Claude (Anthropic) - Claude 3.7 Sonnet applied few-shot learning, identifying linguistic and semantic features from the 42 labeled examples. Classification was guided by structural patterns in student queries and interpreted intent [2].

Deepseek (LLM) - The DeepSeek-V3 model was used. The approach was zero-shot learning, classifying 170 questions based solely on the 42 labeled examples through semantic comparison and pattern matching. There was no model retraining, use of external datasets, or modification of the original 10 categories [26].

Gemini 25Pro (Google) - Used a few-shot classification strategy leveraging patterns in phrasing and inferred student intent. Category learning was grounded in the 42 labeled examples with no external data [8].

Grok 3 (xAI) - Conducted manual interpretation and rule-based classification using labeled examples. Each question was analyzed for semantic intent, preserving original wording. The process focused on ensuring consistent application of categories [28].

Manus (Ensemble ML) - Used an ensemble of Naive Bayes, SVM, and Random Forest classifiers. Classification was based on TF-IDF vectorized questions. Evaluation included 5-fold cross-validation to measure performance [27].

Perplexity AI Assistant - Zero-/few-shot classifier that relied on semantic similarity and keyword patterns from the labeled examples. No retraining was performed. Classification adhered to given categories [1].

These configurations reflect a diverse range of strategies from fully automated pipelines to pattern-based manual inference, enabling a comparative assessment of interpretive alignment with a human-defined educational taxonomy.

3. TYPOLOGY AND TAXONOMY OVERVIEW

This section synthesizes the typology and corresponding taxonomy constructed through a human-expert annotation process, as detailed previously. The ten communicative question types identified were systematically mapped to six broader cognitive actions, allowing for higher-level abstraction of student intent during AI-mediated tutoring.
Rather than reiterating the full process, it is important to note that each question was assigned both a surface-level `Type` and a deeper `Cognitive Action`, capturing not only what students asked but how they engaged cognitively. The table below presents the resulting correspondence between these two dimensions, which underpins the evaluations in the subsequent Results section.

Table 3. : Mapping of question types to cognitive actions

Question Type	Cognitive Action
Questioning tutor's behavior	Confrontation
Asking for an answer	Extraction
Checking correctness	Confirmation
Checking understanding	Confirmation
Checking next step	Orientation
Asking for guidance	Orientation
Asking for alternative solution	Exploration
Asking for problems to practice	Exploration
Asking for assessment	Reflection
Inviting to discuss concepts	Reflection

3.1 Descriptive Statistics

To evaluate the distribution of question intents and cognitive actions, the frequency of each label based on the annotated dataset was computed. Although the dataset does not constitute a fully representative or randomized sample of all student behavior, it serves as a rich and illustrative corpus of authentic tutoring interactions with an LLM-based assistant. The extracted prompts come from dynamic dialogues where students attempted to reason, explore, and validate their understanding of mathematical concepts. Approximately 80% of all questions were classified under the **Orientation** and **Confirmation** categories, which aligns with the stated pedagogical goals of the Khanmigo assistant—that is, to guide rather than answer directly [17]. Only about 12% of the queries were direct requests for answers (*Extraction*), further reinforcing that students largely engage with the tool as intended.

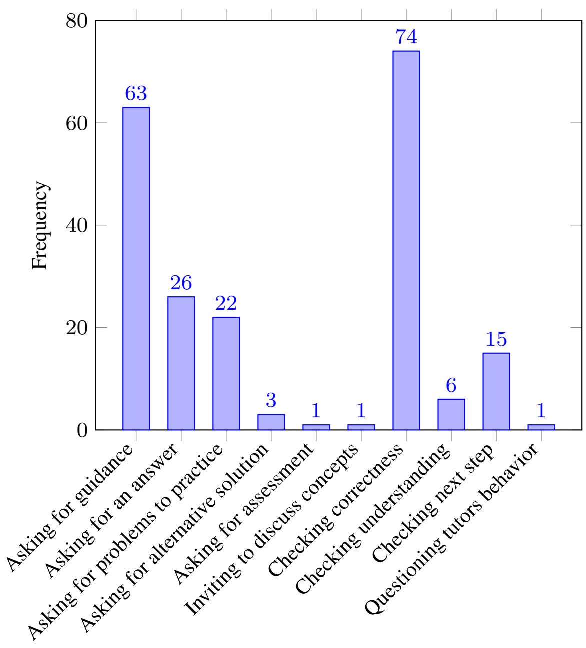

Fig. 2: Distribution of question types across the dataset

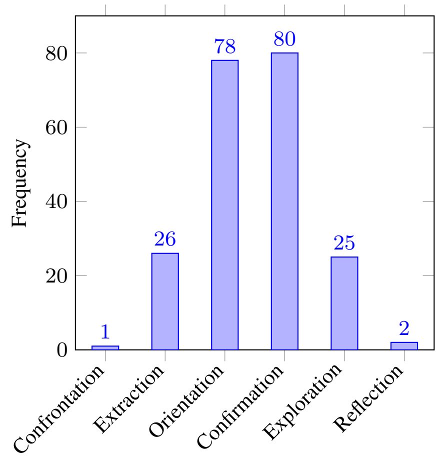

Fig. 3: Distribution of cognitive actions derived from manual classification

3.2 Partially Ordered Taxonomy

To structure these categories beyond flat labeling, it was conducted pairwise comparisons of cognitive actions to determine whether a hierarchy of cognitive complexity could be established between them. In several cases-such as between *Extraction* and *Confirmation*, or between *Orientation* and *Exploration*-a clear developmental or epistemic progression was inferred. However, in other pairs, no consistent ordering could be justified based on context, frequency, or underlying student intent.

As a result, it was adopted a *partially ordered taxonomy* model. In contrast to strictly hierarchical frameworks like Bloom's taxonomy [4], a partially ordered taxonomy accommodates categories that are not all mutually comparable in terms of complexity or sequence. Some cognitive actions occupy parallel levels-operating with similar epistemic sophistication but toward different ends-while others form vertical relationships reflecting ascending levels of cognitive engagement.

Figure 4 depicts this structure, illustrating both converging and diverging paths among the six cognitive actions. It emphasizes that student reasoning in LLM-mediated interactions is neither strictly linear nor reducible to a single trajectory, but instead reflects multidirectional pathways shaped by pedagogical context and learner agency.

To support future applications of this framework, Table 3 outlines the relationship between the question types identified and their corresponding cognitive actions, clarifying how surface-level communicative forms align with deeper epistemic functions.

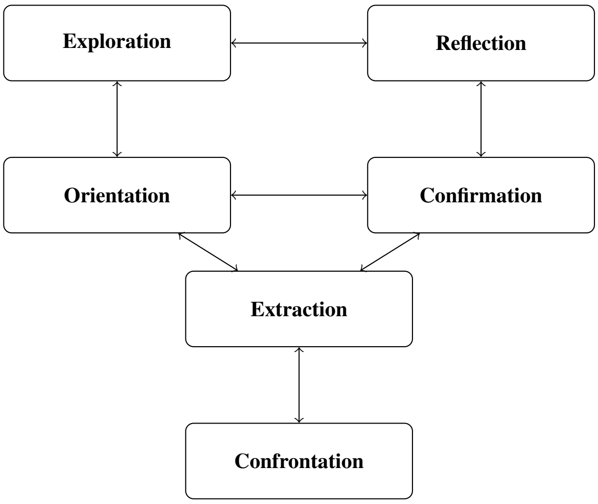

Fig. 4: Partially ordered taxonomy of student cognitive actions derived from manual classification.

4. RESULTS

This section presents the results corresponding to the two research questions. RQ1 investigates whether NLP-based embeddings combined with unsupervised clustering (via PCA) can yield clusters that align meaningfully with a human-derived typology. RQ2 evaluates the extent to which commercial LLMs can accurately classify student-generated questions according to this expert-designed taxonomy.
The categories and their corresponding codes are listed in Table 4 and Table 5. This mapping served as the reference for both the PCA cluster comparison (RQ1) and LLM classification (RQ2).

Table 4. : Codes used for question types

Type	**Code**
Checking correctness	RT01
Asking for guidance	RT02
Checking understanding	RT03
Asking for problems to practice	RT04
Checking next step	RT05
Asking for alternative solution	RT06
Asking for assessment	RT07
Asking for an answer	RT08
Inviting to discuss concepts	RT09
Questioning tutors behavior	RT10

Table 5. : Codes used for cognitive actions

Cognitive Action	**Code**
Confirmation	CA1
Orientation	CA2
Exploration	CA3
Reflection	CA4
Extraction	CA5
Confrontation	CA6

4.1 RQ1: Alignment Between NLP+PCA Clusters and Human Typology

To assess whether PCA-based semantic clustering recovers the categories defined by experts, each of the six KMeans clusters was cross-tabulated against both question `Types` (RT01-RT10) and broader `Cognitive Actions` (CA1-CA6). Tables 6 and 7 present the raw co-occurrence counts, while Tables 8 and 9 summarize the corresponding F1 scores, which combine precision and recall into a single interpretability metric.

Table 6. : Contingency Matrix: Clusters vs. Type Codes

Cluster	RT01	RT02	RT03	RT04	RT06	RT07	RT08	RT09	RT10
0	20	20	3	3	0	0	3	0	0
1	5	14	2	4	3	1	0	0	0
2	11	6	0	2	0	0	0	0	0
3	4	14	0	13	0	0	9	1	1
4	29	4	1	0	0	0	0	0	0
5	5	5	0	0	0	0	14	0	0

Table 7. : Contingency Matrix: Clusters vs. Cognitive Actions

Cluster	CA1	CA2	CA3	CA4	CA5	CA6
0	23	20	3	0	3	0
1	7	23	7	1	0	0
2	11	6	2	0	0	0
3	4	16	13	1	9	1
4	30	8	0	0	0	0
5	5	5	0	0	14	0

Table 8. : F1 Scores: PCA Clusters vs. Human Types

Cluster	RT01	RT02	RT03	RT04	RT06	RT07	RT08	RT09	RT10
0	0.33	0.36	0.11	0.08	–	–	0.08	–	–
1	0.10	0.30	0.11	0.16	0.19	0.07	–	–	–
2	0.24	0.15	–	0.10	–	–	–	–	–
3	0.07	0.27	–	0.41	–	–	0.26	0.05	0.05
4	0.54	0.08	0.05	–	–	–	–	–	–
5	0.10	0.11	–	–	–	–	0.56	–	–

Table 9. : F1 Scores: PCA Clusters vs. Cognitive Actions

Cluster	CA1	CA2	CA3	CA4	CA5	CA6
0	0.36	0.31	0.08	–	0.08	–
1	0.12	0.40	0.22	0.05	–	–
2	0.22	0.12	0.09	–	–	–
3	0.06	0.26	0.38	0.04	0.26	0.04
4	0.51	0.14	–	–	–	–
5	0.10	0.10	–	–	0.56	–

Interpretation. As shown in Tables 8 and 9, Cluster 4 achieves the highest F1 score for `Confirmation` (CA1) and Cluster 5 for `Extraction` (CA5). The remaining clusters exhibit predominantly low-to-moderate F1 values (< 0.40), indicating that PCA + KMeans alone only partially captures the nuanced, pedagogy-driven distinctions defined by human experts.

Table 10. : Accuracy of LLMs in Classifying Questions by Type

Model	Accuracy (%)	Classification Strategy Keywords
Deepseek-V3	87%	Zero-shot; DeepSeek-V3; pattern matching; in-context examples; no retraining
Grok 3 (xAI)	81%	Manual classification; semantic interpretation; rule-based
Gemini 25Pro (Google)	77%	Few-shot; phrase patterns; no external data
Perplexity AI Assistant	71%	Few-shot; semantic similarity; keyword patterns
Claude (Anthropic)	60%	Few-shot; linguistic features; structural inference
ChatGPT-o4-mini	56%	TF-IDF; logistic regression; scikit-learn
ChatGPT-o4-miniHigh	56%	TF-IDF; logistic regression; high memory variant
ChatGPT-o3	48%	TF-IDF; logistic regression; manual scripting
Manus (Ensemble ML)	43%	Ensemble; Naive Bayes; SVM; Random Forest; TF-IDF
ChatGPT-4o	36%	kNN (k=3); TF-IDF; 1–2 n-grams; stopword removal

4.2 RQ2: LLM Generalization of Human Typology

To evaluate whether large language models (LLMs) can generalize a human-derived typology, ten commercial models were prompted with a seed dataset of 42 labeled questions and asked to classify the remaining 170 unlabeled questions.

Table 10 shows overall classification accuracy per model, ordered from highest to lowest. These results reveal varying capabilities across models.

The top-performing models – Deepseek, Grok, and Gemini – demonstrated strong alignment with human annotation, suggesting their potential to scale expert-designed typologies in educational tools. Their success appears closely linked to methodological fidelity: Deepseek employed zero-shot learning with strict pattern matching and no external data [26]; Grok conducted manual semantic classification grounded in interpretive consistency [28]; and Gemini applied few-shot learning guided by intent inference from representative samples [8]. These approaches preserved the communicative and contextual features present in the labeled data, crucial for pedagogical accuracy.

Conversely, the three least accurate models – ChatGPT-4o, Manus, and ChatGPT-o3 – showed lower alignment with expert labels. ChatGPT-4o relied on k-Nearest Neighbors with TF-IDF and minimal context modeling [20], likely limiting its ability to generalize beyond lexical similarity. Manus used a traditional ensemble (Naive Bayes, SVM, Random Forest) [27], which, despite leveraging TF-IDF, struggled with semantic nuance and overfitted frequent patterns. ChatGPT-o3, though following a logistic regression pipeline [21], demonstrated difficulty in adapting category definitions due to its rigid vector-based strategy. These results suggest that performance on classification tasks demanding human-like interpretive judgment is sensitive to architectural features such as contextual memory, semantic abstraction, and prompt alignment with educational goals.

5. CONCLUSION

This study introduced and tested a typology of student cognitive actions during interactions with an LLM-based tutor. Drawing on the CoMTA dataset of 188 real tutoring dialogues from Khan Academy, a two-stage classification process to map student-generated questions to communicative types and cognitive actions was designed.

The contributions are twofold. First, it was developed a grounded typology of question intents linked to a six-part taxonomy of cognitive actions, providing a new framework for analyzing student reasoning in emergent AI-mediated contexts. Second, it was tested two strategies for scaling this classification: (1) unsupervised clustering via NLP embeddings and PCA, and (2) interpretation by large language models (LLMs).

The clustering-based approach (RQ1) revealed only modest alignment between PCA-derived clusters and human labels. While some clusters (e.g., Cluster 4 with *Confirmation*, Cluster 5 with *Extraction*) showed moderate F1 scores, the majority lacked interpretive coherence. This suggests that unsupervised clustering alone is insufficient for constructing typologies that reflect cognitive intent.

In contrast, the evaluation of ten LLMs (RQ2) showed that several models – particularly Deepseek, Grok, and Gemini – achieved high classification accuracy (87%, 81%, and 77%, respectively), approaching human-level labeling. These models shared architectural and methodological traits that likely contributed to their success: Deepseek relied on strict zero-shot inference grounded in semantic pattern matching, Grok used rule-based human-designed logic, and Gemini leveraged few-shot learning with clear intent alignment. Together, they preserved the contextual and communicative signals embedded in the labeled examples – an essential trait for educational classification tasks.

Conversely, models such as ChatGPT-4o (36%) and Manus (43%) struggled. ChatGPT-4o used a shallow kNN-based classifier on TF-IDF features, limiting its semantic abstraction. Manus, although ensemble-based, suffered from overfitting frequent patterns and lacked interpretive generalization. These findings underscore that architectural features – such as memory depth, context modeling, and training objectives – have a direct impact on the model's ability to replicate pedagogically meaningful human judgments.

While this study is limited to a single dataset and a narrow domain (math tutoring), its implications are broader. As AI continues to mediate student learning, typologies like the one developed here can inform the design of more responsive, interpretable, and pedagogically aligned tutoring systems. Future work should expand the typology to other subjects, assess its cross-cultural validity, and explore real-time feedback applications in AI tutors.

Ultimately, the value of this research lies in bridging cognitive theory with machine understanding. Typological frameworks allow not only the analysis of student reasoning but also the development of AI systems that respect, reflect, and respond to the nuance of human thought in educational settings.

6. REFERENCES

[1] Perplexity AI. Perplexity assistant semantic classifier, 2025. Zero-/few-shot classification based on semantic alignment with labeled examples.

[2] Anthropic. Claude 3.7 sonnet classification process, 2025. Few-shot reasoning using semantic patterns from labeled student questions.

[3] Simon P. Bates, Ross K. Galloway, Jonathan Riise, and Danny Homer. Assessing the quality of a student-generated

question repository. *Physical Review Special Topics - Physics Education Research*, 10(2):020105, 2014.

[4] Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company, New York, 1956.

[5] Ng D., Luo Wanying, Chan H., and Chu S. Using digital story writing as a pedagogy to develop AI literacy among primary students. *Computers and Education: Artificial Intelligence*, 2022.

[6] Ng D., Wu Wenjie, Leung J., Chiu T., and K. Chu S. Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology*, 2023.

[7] Jennifer Daries and colleagues. Ai literacy framework: Review draft. Unpublished draft, 2024.

[8] Google DeepMind. Gemini 25pro question classification process, 2025. Few-shot classification based on phrasing and inferred student intent.

[9] Michael Hansen, Claire Scoular, and Patrick Griffin. Evidence for assessment of 21st century competencies: A review of 19 instruments. *Assessment in Education: Principles, Policy & Practice*, 30(1):43–68, 2023.

[10] H. Y. Yim I. and Su Jiahong. Artificial intelligence (AI) learning tools in K-12 education: A scoping review. *Journal of Computers in Education*, 2024.

[11] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[12] Stacy M. Kula and Ting Chang. Student-generated questions: Encouraging academic engagement, critical thinking, and student voice. *College Teaching*, 65(3):126–135, 2017.

[13] Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, 2002.

[14] David Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, pages 1–16, New York, NY, USA, 2020. ACM.

[15] Casal-Otero Lorena, Catalá Alejandro, Fernández-Morante Carmen, Taboada M., Cebreiro Beatriz, and Barro S. Ai literacy in K-12: a systematic literature review. *International Journal of STEM Education*, 2023.

[16] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[17] Pepper Miller and Kristen Dicerbo. Llm based math tutoring: Challenges and dataset. Technical report, Khan Academy, 2024. Available at: https://github.com/Khan/tutoring-accuracy-dataset/blob/main/LLM_Based_Math_Tutoring.pdf.

[18] Davy Tsz Kit N.G., Luo Wanru, Chan H., and Chu S. An examination on primary students' development in AI literacy through digital story writing. *Computers and Education: Artificial Intelligence*, 2022.

[19] OECD. *Introducing the OECD AI Capability Indicators*. OECD Publishing, Paris, 2025.

[20] OpenAI. Chatgpt-4o nlp classification meta-process, 2025. Performed classification of student questions using TF-IDF + kNN pipeline in Python.

[21] OpenAI. Chatgpt-o3 classification meta-process, 2025. Executed TF-IDF + logistic regression on labeled student questions using scikit-learn.

[22] OpenAI. Chatgpt-o4-mini classification meta-process, 2025. TF-IDF + logistic regression classification from 42 labeled examples.

[23] OpenAI. Chatgpt-o4-minihigh classification meta-process, 2025. Same approach as o4-mini, applied to the same task using TF-IDF + logistic regression.

[24] Stacey Pylman. Student-generated questions as formative assessment: A review of the literature. *Educational Practice and Theory*, 42(2):5–25, 2020.

[25] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[26] Deepseek Team. Gpt-4 based semantic classification, 2025. Used semantic matching to assign question categories without retraining.

[27] Manus AI Team. Manus ensemble classification engine, 2025. Used Naive Bayes, SVM, and Random Forest in soft-voting mode with TF-IDF inputs.

[28] xAI. Grok 3 student query classification, 2025. Manual category assignment with semantic and intent-based rule application.

[29] Zhou Xiaofei, Van Brummelen Jessica, and Lin Phoebe. Designing AI Learning Experiences for K-12: Emerging Works, Future Opportunities and a Design Framework. *arXiv.org*, 2020.

[30] Fu-Yun Yu. Student question-generation: The learning processes involved and their relationships with students' perceptions of a learning environment. *Instructional Science*, 39(3):325–346, 2011.

APPENDIX

Appendix A: Sample Tutoring Session from CoMTA Dataset

Below is a complete dialogue between a student and the LLM-based tutor, extracted from the CoMTA dataset (`test_id = 1`). The exchange exemplifies various question types and cognitive actions addressed in this study.

> **User:** I'm struggling with multiplying decimal numbers, can you guide me through the process?
> **Assistant:** Absolutely! Let's begin with a basic problem. How about 0.2×0.3? Try to solve it and let me know your answer. Remember, you can initially treat the decimals as if they were whole numbers!
> **User:** 0.6
> **Assistant:** Let's reconsider this. If we treat 0.2 and 0.3 as 2 and 3, we get $2 \times 3 = 6$. But these are not whole numbers, they're tenths! So, what should we do with the decimal point in our result?
> **User:** what is the rule on where to put the decimal?
> **Assistant:** Excellent question! Here's a tip: count the total number of digits after the decimal in the original numbers. How many are there in total?
> **User:** 2
> **Assistant:** That's correct! Now, let's place the decimal point in our result. Where should it go if we have 2 digits after the decimal in total?
> **User:** 2 places to the right
> **Assistant:** Almost there, but not quite. We want to move the decimal point to the left, not the right. So, if we have 2 digits after the decimal in total, where should the decimal point go in our result?
> **User:** 2 places to the left
> **Assistant:** Exactly! So, if we move the decimal point 2 places to the left in our result of 6, what do we get?
> **User:** .006
> **Assistant:** Not quite. Remember, we're starting with the number 6. If we move the decimal point 2 places to the left, we get a different number. Try again!
> **User:** .06
> **Assistant:** Yes, that's correct! So, $0.2 \times 0.3 = 0.06$. Well done! Would you like to try another problem, or is there something else you'd like to learn about?
> **User:** so we multiply the whole numbers and then move the decimal to the left according to how many numbers are to the right of the decimal?
> **Assistant:** Yes, you've got it! When multiplying decimals, we first multiply as if they're whole numbers. Then, we count the total number of digits to the right of the decimal in the original numbers. We move the decimal point in our result that many places to the left. This gives us the correct decimal placement in our result. Great job! Would you like to try another problem to practice this concept?
> **User:** give me a more challenging example
> **Assistant:** Sure thing! Let's try 0.25×0.4. Remember, you can start by multiplying as if they're whole numbers. What do you get?
> **User:** 100

This sample showcases multiple levels of cognitive engagement, including conceptual inquiry, procedural clarification, and reflective reasoning. It serves as an illustrative example of the annotation approach adopted in this research.

Appendix B: Detailed LLM Classification by Type

Table 11. : Number of Correct Predictions per Question Type by Model

Type	ChatGPT-4o	ChatGPT-o3	ChatGPT-o4-mini	Claude	Deepseek	Gemini	Manus	ChatGPT-o4-miniHigh	Perplexity	Grok
Checking correctness	61	57	56	48	53	46	59	56	42	48
Asking for guidance	0	17	35	38	50	36	9	35	44	41
Checking understanding	0	0	0	3	1	2	0	0	2	2
Asking for problems to practice	1	8	5	15	17	17	4	5	16	17
Asking for an answer	0	0	0	16	18	22	1	0	8	22
Checking next step	0	0	0	7	8	7	0	0	8	7
Asking for alternative solution	0	0	0	1	1	1	0	0	1	1
Asking for assessment	0	0	0	0	0	0	0	0	0	0
Inviting to discuss concepts	0	0	0	0	0	0	0	0	0	0
Questioning tutors behavior	0	0	0	0	0	0	0	0	0	0
Total Correct	62	82	96	128	148	131	73	96	121	138
Accuracy	36%	48%	56%	60%	87%	77%	43%	56%	71%	81%