Clinical Reasoning-Driven Progress Evaluation of Medical Students Using Large Language Models

Heitor S. Mattosinho[1,*], Fernando Valente[2], Gabriel Leite[1], Ligia Maria Cayres Ribeiro[2], Marco A. de Carvalho Filho[2], and André Santanchè[1]

[1]Institute of Computing, University of Campinas, Brazil
[2]University Medical Center Groningen, University of Groningen, Netherlands
Email: heitor.mattosinho@ic.unicamp.br (H.S.M.); f.valente@hc.fm.usp.br (F.V.); gabriel.dfleite@gmail.com (G.L.); l.m.cayres.ribeiro@umcg.nl (L.M.C.R.); m.a.de.carvalho.filho@umcg.nl (M.A.C.F.); santanche@ic.unicamp.br (A.S.)
*Corresponding author

Manuscript received August 14, 2025; revised September 8, 2025; accepted November 6, 2025; published March 13, 2026

***Abstract*—Evaluating medical students' written answers to questions on a given topic can provide information about their mental representations of a disease—*i.e.*, illness script. However, limitations in methods for assessing how medical students develop clinical expertise hinder the advancement of educational practices. This study, therefore, proposes a technique that utilizes semantic annotations of the students' answers to trace a map of their knowledge concerning the topic of a question. Since manual text annotation is time- and effort-intensive, this study developed an innovative, illness-script-driven strategy using large language models. It identifies relevant medical information in the texts, creates a profile for each answer, clusters them, and categorizes the clusters. Practical experiments with Brazilian students demonstrate that the technique automatically traces consistent profiles from the responses, quantifying how knowledge of disease diagnosis evolves throughout the medical course.**

***Keywords*—illness script, medical education evaluation, large language model, clustering profiles**

I. Introduction

The continuous evaluation of a student's learning trajectory is a crucial component of education. It should inform to what extent the instructional activities are achieving their learning goals and provide students and teachers with cues to improve their processes. Evaluation, however, is a complex task. Some of its complexities lie in the fact that it requires a variety of strategies to accurately reflect different aspects of competency, such as knowledge, skills, and attitudes, as well as the workload they represent for evaluators and the challenges of scalability. In medical education, for example, researchers have developed different evaluation strategies [1]. These different methods also differ in their implementation. Observing a student taking care of real patients is more labor-intensive than creating multiple-choice questions for dozens of students. On the other hand, observation reveals aspects of competency that multiple-choice questions cannot access. One core practice of physicians that should be the focus of more evaluation, given its importance, is diagnostic reasoning.

Diagnostic reasoning is the complex process of gathering and interpreting clinical information, such as the history, physical examination, and laboratory test results, to determine the most likely diagnostic hypotheses for a specific patient. It is known that accumulating knowledge about diseases is necessary but is not enough for accurately diagnosing patients. It is also necessary for the knowledge to be organized [2]. As doctors develop expertise, they seem to store and structure knowledge in their long-term memory in the form of "*illness scripts*", which are mental representations of diseases [3]. Through repetitive practice with a large array of patients, physicians increase the number of illness scripts stored in their memory, and enrich them with new and nuanced data [4]. When a patient presents, for example, with chest pain, a doctor familiar with this symptom retrieves the illness scripts of different diseases that can cause chest pain and, collecting additional data, chooses the scripts that best fit that specific patient. Assessing such complex cognitive processes proves difficult, and no single method is sufficient for the task.

One way to evaluate illness scripts is through recall tasks. A question such as "tell me everything you know about chronic obstructive pulmonary disease" or "explain why you think this patient has chronic obstructive pulmonary disease", answered either verbally or in writing, can help elicit the student's illness script about this disease. It is commonly used by medical teachers when supervising students individually or in small groups while assisting patients, and has been explored in the medical education literature [5]. However, manually evaluating recall tasks are labor-intensive and time-consuming, which results in delayed feedback and limits scalability.

Therefore, an automated evaluation process is a desirable approach to improve the viability of use. In turn, this requires a structured evaluation process. The integration of computational tools, such as Large Language Models (LLMs), could expand the evaluation of illness scripts, offering students the opportunity to have evaluations on scripts of a larger number of diseases. These tools could also enable the evaluation of students' script progression over time.

The primary challenge of this work lies in the integration of an LLM into an evaluation process based on the illness script theory. This method is innovative because the LLMs do not operate as "black boxes", encompassing the entire evaluation process. Rather, they are fine-tuned or guided to facilitate the identification and annotation of illness script core ideas (components) in a text. The ideas and their sequence support an unsupervised, clustering-based method that categorizes students' responses into three distinct progression stages: novice, developing, and proficient.

Practical experiments show that aligning the illness script categories of annotations with the curricular components of Brazilian undergraduate medical courses reveals that the prominent categories at each stage reflect the curriculum's structure.

doi: 10.18178/ijiet.2026.16.3.2542

While related work has adopted LLMs mainly to annotate texts for evaluation purposes, this work innovates by structuring the process around the illness script theory and aligning it with an unsupervised method for categorizing students' answers. Besides being able to link student profiles to curriculum progression, this study also links components of clinical reasoning, made explicit through textual answers, to the knowledge and skills expected to be acquired in the Brazilian National Curricular Guidelines.

The remainder of the text is organized as follows. Section II presents the foundations, while Section III presents related work. Sections IV and V detail the approach, case study results, and contributions. Section VI presents conclusions and future work.

II. Foundations

A key element of this work involves exploring Transformer-based large language models for the Named Entity Recognition (NER) task, which supports a medical evaluation process.

NER is a natural language processing technique that focuses on identifying entities within a text. Traditionally, NER tasks are treated as a sequence labeling problem, where the text is segmented into a sequence of words and fed into the input. The output labels each sequence in order, detailing their respective categories and position [6]. This paper examines two strategies for this task—based on encoder- or decoder-only Transformer architectures—to investigate their potential and transparency.

Encoder-only strategies, like BERT—Bidirectional Encoder Representations from Transformers, are pre-trained masked language models [7], apt to infilling objectives based on the context. Besides the several applications built on top of BERT, NER techniques take advantage of its representation to recognize context-aware entities.

Most important research in the NER field utilizes different Transformers encoder-only models [8, 9]. Such models require extensive curated and manually annotated data for training and validation, which can be a limiting factor in many areas of application.

Term extraction tasks can be even more challenging in clinical scenarios than in other areas of application. This is explained because medical entities can be highly context-sensitive, often requiring a nuanced comprehension of the text. Additionally, biomedical entities can have many synonyms and abbreviations [10]. All these differentiations can pose a challenge in the term extraction task.

Large Language Models (LLMs), such as ChatGPT and Llama, are decoder-only autoregressive models based on the Transformer architecture. Their task involves predicting (\textit{i.e.}, generating) the following tokens departing from the previous ones of the sequence. These cutting-edge models can perform complex assignments given a request in natural language. Their extensive pre-trained data and large number of parameters enable great results across many generative tasks. These LLMs allow detailed prompts for specific tasks. With In-Context Learning (ICL), users can provide examples to further improve results.

Even though LLMs are trained on a plethora of data, specific knowledge can be lacking or necessitate unique contextual awareness to complete the task. However, instead of necessitating an additional large set of annotated data for that specific task to perform fine-tuning—usually hard to acquire the data and can be expensive to train these large models—an alternative is to utilize ICL (*i.e.*, include similar training examples for the task in the prompt), which has shown great improvements [11].

Attempts to utilize generative language models for solving Information Extraction (IE) tasks began to be described after their potential to address a wide range of tasks was recognized. However, studies described limitations for generative LLMs when they attempt to mimic the output format of traditional encoder-only models [12]. The main difficulties are span positioning and unmentioned texts or entity types [10, 13, 14]. Because of their generative nature, these LLMs produce sentences that are not from the source text and label entities with types not specified in the instructions to the model. To that extent, many studies propose methods to augment the labeling process into the source text, reporting that it better suits the strengths of these generative models. Next, some of these approaches are further detailed.

The first studies to utilize generative pre-trained models were [15] and [16], which employed T5 and BART, respectively. These are transformer encoder-decoder language models. Both papers approached this generative sequence labeling task differently, with a focus on the strengths of text generation. Reference [15] proposes what they call a "*generative-style sequence labeling model*", where information is augmented into the target text. As an example, the sentence "*He is John Wethy from NBC News*" is transformed in the labeling process, resulting in "*He is [John Wethy | person] from [NBC News | org]*". On the other hand, [16] proposed an approach called "*Template-based NER*". First, all possible spans are enumerated in a sentence. Spans are fundamentally a consecutive sequence of tokens, and for each sentence they restrict the number of *n*-grams for a span. Later, the approach utilizes the prepared template for each span. Given a template "*⟨xi:j⟩is a⟨yk⟩entity*"—where *⟨xi:j⟩* denotes a candidate text span and *⟨yk⟩* denotes an entity type from the set of all possible types *y* —, the template is instantiated for the span *⟨xi:j⟩* across all entity types *⟨yk⟩* in *y*. Finally, the generative model scores each of these instantiated templates, and the one with the highest score is chosen for the span.

As examples of decoder-only LLMs, the NER approach in [13] augments the source text by surrounding the entities and their respective categories with special markers. However, it only identifies one category per prompt. Similarly, [10] utilizes text augmentation. The study uses an IE framework named TANL [17], which can solve various language tasks, such as relation extraction and nested named entity recognition. The differential is that they utilized a RAG-based (retrieval-augmented generation) method to improve the entity recognition and labeling process.

Compared to previous work, these studies that utilize text augmentation address the problem of the unmentioned text. Specifically, generative language models create entities not present in the text for approaches that involve listing discovered entities apart from the source text.

In summary, generative models [13, 14] can perform IE tasks competitively but tend to underperform traditional pre-trained encoder-only language models (*e.g.*, BERT, RoBERTa) when enough annotated data is available for fine-tuning. Additionally, the utilization of in-context learning considerably improves the performance of generative models but requires significantly less annotated data. This suggests that decoder-only LLMs can serve as an alternative in scenarios where annotated data is scarce, as they typically achieve good performance with a few examples via ICL [15, 18].

III. Related Work

Medical evaluations are typically conducted through standardized oral or written exams, which means that automatization of the evaluation process can be achieved through NLP techniques [19].

The extraction of information in student-written notes is a recurrent topic of interest. Two studies assessed students' encounters with standardized patients (*i.e.,* trained actors that simulate clinical conditions) [20, 21]. In the United States Medical Licensing Examination (USMLE), specialists score Patient Notes (PN)—written by examinees in encounters with standardized patients—according to predefined Key Essentials.

The annual exam results are reviewed by approximately 100 raters, and the total number of PNs exceeds 330,000 [20]. It is a given that these large numbers are cumbersome for the raters and that it is a complex task to guarantee a consistent scoring in such a scenario. To standardize the evaluation and ease the burden on human raters, [20] developed an automatic information extraction system to aid in the grading process. The objective is to determine if a specific standard concept, called Key Essential (KE) entry, is represented within a span of text in an examinee's patient note. The methodology consists of a semi-supervised approach, wherein the system is guided by a small set of human-annotated medical notes rather than being pre-trained on a large dataset. This approach sequentially implements a series of text comparison and similarity functions, ranging from exact string matching to fuzzy similarity with dynamic thresholding and set-overlap techniques.

In the same annual exam scenario, but with a different approach, [21] utilizes ChatGPT 3.5 to grade first-year medical students' notes according to a scoring rubric. In this instance, the comparison is between the scoring given by the standardized patient and the generative model. The study concluded that ChatGPT made fewer scoring errors than the standardised patient.

Unlike the proposed work, some of the aforementioned approaches operate as "black boxes", meaning the rationale behind the conclusions is obscured by a neural network. For those who rely on automated text annotation, the absence of theoretical guidelines, such as the illness script, makes it subjective which parameters an LLM should use to evaluate a given answer.

IV. Clinical Reasoning-driven Progress Evaluation with LLM

In this study, the goal is to explore the extracted structured information to build a map that reflects the progress of a medical student in comprehending a disease, its diagnosis, and treatment.

Fig. 1 presents a diagram illustrating the overall process pipeline. The process begins at the stage indicated by (1) in the figure. Students answer an open question.

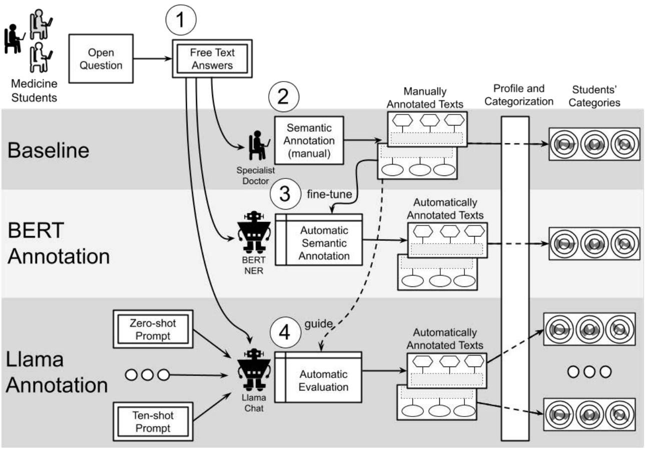

Fig. 1. Pipeline of the students' evaluation and categorization process.

The evaluation approach is based on illness script theory [6], which explains how clinical reasoning develops over time. Early on, students learn the causes of diseases. With practice in both simulated and real clinical settings, they begin to connect this knowledge to how diseases appear in patients. As students gain experience, these early ideas evolve into organized, structured mental models — called illness scripts. It typically includes risk factors (enabling conditions), symptoms (consequences), and mechanisms of diseases (fault). Fault mechanisms become increasingly simplified or encapsulated into easy-to-recall chunks, replacing detailed descriptions of the underlying pathophysiological process. Illness scripts are also disease-specific. As the exposure to clinical experience varies across students, its development is, therefore, individual.

To explicitly identify and evaluate illness scripts based on students' responses, the process begins with a specialist manually annotating the text to capture illness script idea units—thread (2) in Fig. 1. Next, a module builds profiles from the annotations and categorizes students' answers using an unsupervised clustering method.

The manual annotations form the basis for automation using Named Entity Recognition (NER), which is a valuable approach to bridging human natural language texts to structured information. The entities to be recognized have a model behind them, which is fundamental to supporting the processing of structured data.

The automation considers two methods: a more predictable and guided approach and a less predictable and autonomous one. The first is an encoder-only smaller BioBERTpt-based LLM, fine-tuned to detect and classify the target entities—thread (3). The second is a decoder-only bigger LLM, the Llama 3.1 70B, guided by few-shot prompts—thread (4). There is a tradeoff in this comparison. The first requires fine-tuning the model for acceptable outcomes, which results in the necessity of manually annotated data for each disease. While the second has no requirements for additional fine-tuning for each specific disease, it requires much more computational resources and can act less predictably, a side-effect of the autonomy of this language model, which in turn is also more generalist.

As Fig. 1 illustrates, both methods proceed to the profiling and categorization pipeline, which enables comparison and evaluation of the approaches.

This section details the proposed method. It begins by outlining the clinical reasoning-driven approach used to map student progress from written answers (Section III-A). Next, the dataset adopted for training and guidance is described (Section III-B). The automatization process for NER is then discussed, with separate subsections for BioBERTpt (Section III-C) and Llama (Section III-D). After comparing both approaches (Section III.E), the categorization and evaluation method derived from the NER is presented (Section IV).

A. Mapping the Progress from Students' Answers

Brazilian medical undergraduates are expected, under National Curricular Guidelines (NCG), to integrate knowledge, skills, and professional attitudes across five broad domains: health education, general scientific knowledge, health management, attention to public health needs, and attention to individual patient needs. The NCGs frame learning outcomes in terms of observable competencies such as gathering and interpreting clinical data, formulating diagnostic hypotheses, proposing evidence-based management plans, communicating with patients and teams, and acting within the health-care system. Progress throughout the six-year Brazilian curriculum is therefore judged not only by factual recall, but by how well students weave those facts into actions that benefit patients and society. This paper considers the Illness Script (IS) theory as the basis for exploring students' evolution throughout training in terms of their knowledge and structuring.

An illness script aggregates information about the "fault", the mechanism—or pathophysiology—by which a disease occurs; "enabling conditions", the patient-specific characteristics that facilitate or increase the likelihood of a disease manifestation and the "consequences", which are the observable manifestations of a disease [22]. The amount of correct and incorrect information, and the illness script components in the answers (fault, enabling conditions, and consequences), help estimate knowledge and reveal its structure. The order of this information helps identify how organized the student's illness script is. By detailing categories, it is also possible to understand student deficits in specific areas.

To probe script development, this study adopts a free-recall task, a staple of cognitive psychology research. Students were asked to write everything they knew about chronic obstructive pulmonary disease. Their responses were decomposed into small, semantically distinct idea units, each of which was judged for accuracy and then classified into the components of IS. To provide a more detailed comprehension of students' knowledge, some of these components were further stratified: faults into pathophysiology and etiology, and consequences into history, physical examination, laboratory tests, differential diagnosis, and therapeutic plan information. Knowledge about epidemiology was considered as enabling conditions. These eight categories, stratified from the illness script—pathophysiology, etiology, history, physical examination, laboratory tests, differential diagnosis, therapeutic plan, and epidemiology—were the ones utilized to label the idea units by the medical specialist and the automatic annotation approaches.

To evaluate students' knowledge of each of these components, their texts were decomposed into idea units. An idea unit can be a word or a few words that represent a distinct and meaningful information complex [23]. The idea units can also be classified into predefined categories, such as the aforementioned components of illness scripts. For instance, in the sentence "*the patient usually has a dry cough in the morning*", it is possible to identify three accurate idea units: cough, its dry quality, and its morning predominance, all of which belong to the consequences (history) component of the script. The text parts corresponding to each idea unit would be: "*cough*", "*dry cough*" and "*dry cough in the morning*". Even though all three contain the idea of cough, its dryness and its periodicity add information that can be used to differentiate between students with uneven levels of knowledge. The supplementary material provides details on how to annotate idea units and use the annotation tool.

In Brazil, students who successfully completed high school are eligible to enter medical school. The medical undergraduate curriculum is organized in six years, with the first two years mostly dedicated to basic sciences, such as anatomy and physiology. In years three and four, activities are balanced between in-class discussions about diseases and patient care in real scenarios. Finally, in years five and six, the internships take place, during which students are dedicated full-time to patient care. It is expected, therefore, that in the earliest phases of training, Brazilian students would show isolated biomedical knowledge, with scant reference to contextual factors. As clinical experience increases, a growing proportion of etiology, clinical manifestations, and epidemiology knowledge should develop, signalling that learners are beginning to connect what they observe to the biomedical "fault" and to the conditions that enable diseases. In the senior years, students should have developed well-organized elements of the different categories of an illness script, while the fault component is expected to have been encapsulated. Annotations, therefore, do more than quantify what students know at a single moment; they trace the direction and pace of their evolution toward expert-level competence.

This process, however, is not only time-consuming but also requires attention regarding its reliability [24]. The analytic power of this method comes at a cost. Segmenting free text into idea units, assigning each to the correct category, and deciding on accuracy all introduce subjectivity. To mitigate this, a detailed manual of annotation procedures and patterns, assessor-training sessions, dual independent ratings, and inter-rater reliability checks were developed. However, the time, burden, and expertise required limit this type of idea unit analysis to research settings. Given the high resource intensity of the procedure, it would greatly benefit from automation strategies, particularly natural language processing pipelines capable of segmenting, coding, and scoring answers at scale.

Coding recall tasks at the granularity of idea units translates narrative student output into metrics that mirror the expectations of the Brazilian curriculum. This approach creates a bridge between the cognitive construct of illness scripts and educational practices that can support the curricular goals defined by the NCGs, opening a path for longitudinal tracking of script maturation. Ultimately, scaling the use of illness script analysis can transform educational practices, allowing teachers to offer feedback and remediation strategies tailored to individual students' needs.

B. Annotated Answers Dataset

The evaluation process, based on the annotation of idea units, was tested in real-world scenarios. Initially, specialists conducted manual annotations to further train/guide automatic processes.

Four hundred sixteen Brazilian medical students from seven schools, across all years of training, volunteered to answer the question, "*Tell me everything you know about chronic obstructive pulmonary disease (COPD).*" COPD was chosen because it is a prevalent disease that students are expected to engage with from their early years of training. The answers to this question were taken as a proxy of their illness script about COPD.

The number of answers per year of training ranged from 46 to 98. This allowed the analysis of how illness scripts about COPD are developed at the group level throughout undergraduate training.

Five medical doctors with expertise in education developed a structured analysis of students' answers according to the illness script theory. The process consisted of breaking the texts into idea units and classifying them into the aforementioned categories of illness scripts. After confirming the reliability of this analysis, two experts manually coded all the answers, blinded to students' year of training, using an annotation tool developed in a platform developed by the members of the research group.

This process of selecting and classifying idea units proved to be reproducible by different annotators. The baseline data was partly (20%) and independently annotated by two different human coders. The number of correct ideas and the number of disease dimensions were compared using a two-way random, absolute-agreement Intraclass Correlation Coefficient (ICC) to test for interrater reliability. COPD correct ideas ICC was 0.95, and COPD dimensions ICC was 0.91.

C. From Idea Units to Named Entities

This work treats the automatic annotation process as a Named Entity Recognition (NER) task. Idea units become named entities by identifying and delimiting entities in the text using an Inside-Out-Begin (IOB) format [25]. This approach enables multiple annotation layers. In the example "*The patient usually has a dry cough in the morning*", three entities are delimited as the "*history*" class in three layers: "*cough*", "*dry cough*", and "*dry cough in the morning*". The annotation system also allows for the association of more than one class per entity.

The annotation tool developed for this system allows specialists to identify idea units and relate them to one or more classes. Different annotations can occur in overlapping segments. These annotations are coded in the IOB+classes format, which is used to fine-tune BioBERTpt. However, the annotation format used in the Llama approach. It is either a list containing the entities and entity types or via text augmentation. Both formats are discussed in Section IV.E.

D. BERT NER

The first approach to automate the annotation process used a fine-tuned BioBERTpt model to recognize illness-script-related entities based on idea units. The BioBERTpt model [26] is a deep contextual encoder-only large language model for Brazilian Portuguese that was trained using clinical and biomedical narratives. To support the NER and categories required by this case study, the model was fine-tuned using the students' answers that were manually annotated by specialists.

To ensure an equal distribution of the training, validation, and test data, the students' year was used to balance the data. There were 249 training texts, 83 validation texts, and 84 test texts. The annotation follows the IOB format [25]. The 177M-parameter model is publicly available on Hugging Face Model Hub[1].

E. Llama Generative Model

As previously mentioned, unlike the first approach, the second approach used a decoder-only—and much larger—LLM. The Llama family of Large Language Models is open-source [27]. As explained in Section II, they are generative pre-trained transformer models. This means they excel at text generation tasks and follow instructions from prompts well. Using strictly open-source models ensures the reproducibility and transparency of the research. This research used the Llama 3.1 70B Instruct version.

The prompt was divided into three parts: context, task details and instructions, and output format. The context part was derived from [28], since they performed a similar task in a medical scenario. It provides a brief contextual description of the expertise required for the task ahead. In addition, the general instruction task, also derived from [16], was actively developed in collaboration with a medical specialist to fit the case study. This segment of the prompt describes the task's objective, including a list of possible entity types and their associated constraints. The constraints prevent unwanted token generation, such as the language of the texts (Portuguese) and emphasis on the task. The output format is described to be structured in JSON.

This paper evaluated two Llama approaches for the NER output. The first approach generates a list of pairs (entity, entity type). These pairs are detached from the original text. This approach cannot explicitly extract the position of the entity in the source text, as explained in Section II.B. Another limitation is the problem of unmentioned text, which requires complex post-processing.

The second approach augments the entity tagging into the source text. That means the Llama model generates the source text with surrounding special marks to indicate a named entity, and within it, there is a separation between the entity and its type. This approach leverages the text generation capabilities of LLMs more effectively, and also mitigates errors such as entity positioning and unmentioned texts. For this particular implementation, the LLM did not always follow the format properly, which also required post-processing. Ultimately, both methods yielded similar F1 scores. This paper will focus on the results from the first approach.

[1] https://huggingface.co/harena-lab/bioberpt-dpoc-is-multiple

V. Comparing Llama to Fine-tuned BioBERTpt

This section discusses the results of the Named Entity Recognition (NER) task for Llama 3.1 70B and the two fine-tuned BioBERTpt versions. Since this work's primary goal is to map student progress, comparing automated annotations directly with human ones serves as a partial, rather than absolute, validation method. This is because manual annotation is subjective and sensitive to variations in scope and overlapping categories, which differ even among experts.

A key finding is the annotations' utility in consistently profiling and categorizing students. This consistency holds even when the automated annotations diverge from the human-generated baseline (*i.e.*, when F1 scores are lower), as will be detailed further. Table 1 shows the F1 scores of Llama variations (first five rows), the fine-tuned BioBERTpt, and the BioBERTpt fine-tuned on Llama-generated annotations (BioBERTpt-Llama). The comparison focuses on terms considered as idea units by both the human expert and the automated processes. Therefore, F1 scores are calculated by comparing the exact match of annotation spans, regardless of their categories.

Several task demonstrations were conducted to establish various scenarios and performance evaluations, as shown in Table 1. For clarity, these demonstrations will now be referred to as either "examples" or "shots".

Specifically, the Llama experiment included zero-shot, 1-shot, 2-shot, 3-shot, 4-shot, and 10-shot configurations. These are displayed in the tables' columns, with 0-shot listed in a separate row. For each target text that needs to be annotated, the prompt generation mechanism retrieves manually annotated texts that serve as examples (shots) for the Llama response. This mechanism ensures that the example texts will never be the same as the target text.

For the few-shot scenarios—defined as those with at least one example—four distinct example retrieval methods were applied: Random, Static, Text Similarity, and Labeled Entity Similarity. The Static method involves pre-selecting a fixed set of examples, chosen from the top 11 annotated texts—using a diversity metric, explained below—out of the entire dataset annotated by the medical specialist. The diversity metric considers both the variety of entity types and the total number of annotations: higher diversity reflects a broader usage of different entity types and more idea units. By contrast, the Text Similarity and Labeled Entity Similarity methods dynamically retrieve examples for each target text, using Term Frequency–Inverse Document Frequency (TF-IDF) to select the most similar examples—based on either the entire text or just the annotated entities, respectively. Finally, the Random method serves as a control, randomly retrieving examples from the whole dataset for dynamic shot retrieval.

The two dynamic approaches that utilize retrieval by similarity achieved the highest scores; however, compared to the static approach, they exploited all 416 manually annotated texts for the example retrieval. When considering a scenario with low availability of manually annotated data, it may be more reasonable to adopt the static approach. More research is needed to understand the acceptable size of a dataset for the utilization of dynamic example retrieval.

Table 1. The Llama separated labeling approach against BioBERTpt

Example Retrieval	1-shot	2-shot	3-shot	4-shot	10-shot
Zero-shot	0.7160	Zero-shot	0.7160	Zero-shot	0.7160
Static	0.731	0.771	0.776	0.790	0.792
Text Similarity	0.772	0.794	0.795	0.779	0.801
Labeled Entity Similarity	0.766	0.791	0.789	0.800	0.809
Random	0.742	0.772	0.756	0.773	0.791
Fine-tuned BioBERTpt			0.844		
BioBERTpt-Llama			0.834		

The BioBERTpt fine-tuned model achieved the best results. It can be interpreted as a consequence of a simpler model that has been highly trained for a specific scenario.

To balance the availability of annotated examples and computational cost, this study evaluated a version of BioBERTpt fine-tuned using Llama-generated annotations (BioBERTpt-Llama). The fine-tuning dataset was generated with the 10-shot Llama separated labeling approach, using the Labeled Entity Similarity method in the example retrieval process. While Llama has a high computational cost, this cost is restricted to the fine-tuning stage. Then, BioBERTpt-Llama can perform a recurrent task at a much more affordable computational cost.

The next section addresses the study's primary objective, which involves characterizing student profiles and categorizing them using an unsupervised approach.

VI. Illness-Script-driven Profile and Categorization

As previously described, the Illness Script theory (IS) provided the foundation for the model. The analysis began with eight specific categories: pathophysiology, epidemiology, etiology, history, physical examination, laboratory tests, differential diagnosis, and therapeutic plan. As detailed in Section III-A, these categories derive from the three core IS components: fault, enabling conditions, and consequences.

The hypothesis is that the evolution of a student's knowledge about a disease diagnosis can be evaluated by recognizing and analyzing the relations among these discrete entities. To evaluate the effectiveness of classifying students according to the eight entities, the analysis began with the manual annotation of 416 students' answers conducted by a physician. It is important to recall that the human annotator (and subsequently the automatic ones) was unaware of any students' data or demographics (*e.g.*, student year) during the annotation process.

The students' answers are then represented by an array of nine dimensions, which computes the occurrence of the eight categories, as well as the total number of entities. An unsupervised clustering method was adopted to categorize the students. Several clustering approaches were evaluated, and k-means achieved the best results. K-means initialization was set to k-means++ with the number of initializations set to ten. The maximum number of iterations was set to 300 and tolerance to 1×10^{-4}.

This study presents the approach to map students' annotation dimensions and the respective cluster-based categorization in four parts. First, it is applied to human annotations to show its validity and detail the insights it provides. Further, this result is contrasted with the automated processes in three contexts: Llama 3.1 70B with In-Context Learning (ICL) examples, BioBERTpt fine-tuned using human annotations, and BioBERTpt fine-tuned using Llama-generated annotations.

A. Profile and Categorization from Human Annotations

A qualitative analysis followed a comparison of silhouette values. The analysis revealed that the cluster division yielding the best silhouette—i.e., the one for three clusters—exhibits a high correlation with the stage of students in the course.

Figs. 2 and 3 illustrate the distribution of students' profiles across the clusters.

An analysis of the individual profiles within each cluster revealed a strong correlation between cluster membership and student progress in the course. Based on this correlation, each cluster was named according to the dominant years of study represented within it, reflecting students' stages in the medical curriculum as follows:

Novice cluster—Students are predominantly from the first and second years.

Developing cluster—Students are predominantly from the third and fourth years.

Proficient cluster—Students are predominantly from the last two years (fifth and sixth years).

Figs. 2 and 3 ground the evaluation methodology in concrete examples, aligning with the expected progression of medical students through their training. The literature on illness script development shows that students' knowledge organization shifts over time. They transition from pathophysiology-heavy mental models to illness scripts enriched by clinical experience, focusing on the patient's clinical characteristics. This analysis examines idea units in the therapeutic category separately, since the literature on treatment scripts is limited.

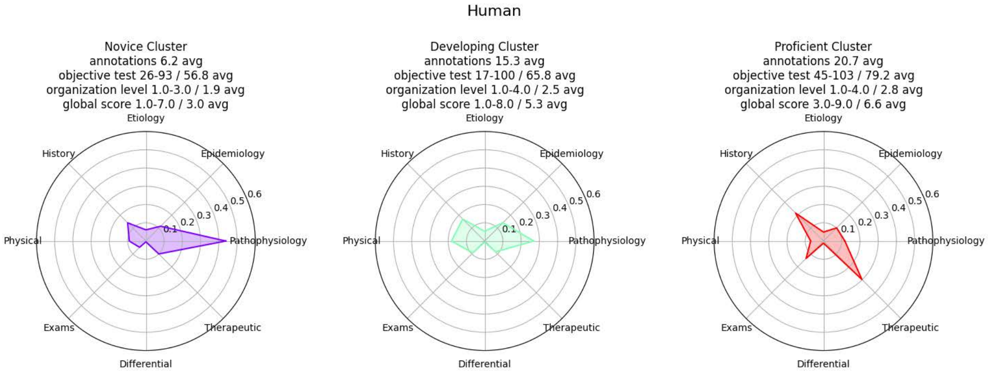

Fig. 2. Categories' distribution inside each cluster, plus the average number of annotations in each cluster, the respective smallest, largest, and average of three scores: objective test, organization level, and global score. Based on human annotations.

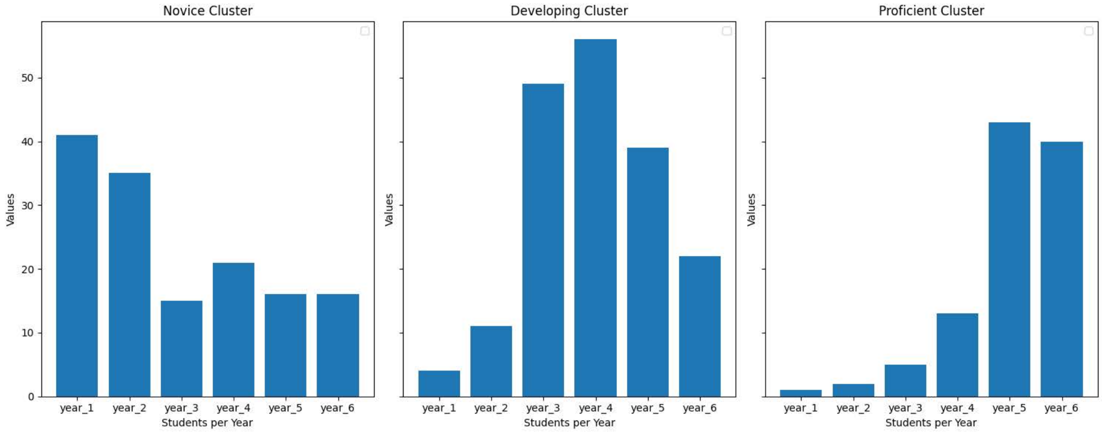

Fig. 3. Distribution of the student's year in each cluster. Based on human annotations.

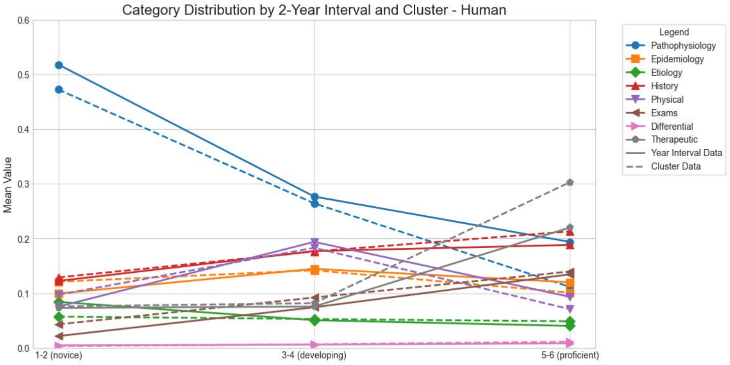

Fig. 4. Dimension distribution of human annotations along the years and clusters.

Fig. 4 illustrates the distribution of the relative presence of the eight dimensions in the text, aggregated by students' year in the course (continuous line) and by the discovered clusters (dashed lines). The vertical scale reflects the relative participation of the category in the annotations. The figure indicates a strong correlation between clustering distribution and course year. In years 1 and 2 and the Novice cluster, "*pathophysiology*" accounts for nearly 50% of annotations. This declines to about 30% in years 3 and 4 (Developing cluster) and falls to nearly 20% in the last years (Proficient cluster). This decline matches the academic course's structure, since the first two years focus on basic sciences, such as anatomy and physiology, as outlined in Section III.A.

Conversely, dimensions such as "*history*" and "*therapeutic*", which are more relevant to later course stages, begin at under 10% in the first year and then show greater increases in subsequent years. While the "*epidemiology*" dimension rises from the early years to the middle years as expected, it declines in the later years, reflecting a relative decrease due to the notable growth of "*therapeutic*" annotations. The low presence of certain dimensions, such as etiology, can be attributed to the nature of the question, which does not elicit etiological aspects.

B. Profile and Categorization from Llama

Fig. 5 compares clusters derived from manual and Llama annotations in a heatmap. Figs. 6 and 7 display the profiles of the clusters. Even though there are differences in human and Llama annotations, as previously detailed, both approaches grouped students in equivalent groups. Devising the Developing clusters from the neighbors is more difficult since the boundaries are fuzzy. Nevertheless, students who are differently classified are always in a neighboring cluster, *i.e.*, there is no Novice student—from human annotations—classified as Proficient and vice-versa.

A comparison of Figs. 6 and 7 reveals that, as with the charts generated from human annotations, each cluster is characterized by the predominance of one or a few dimensions. An analysis of the individual clusters allows for conclusions to be drawn regarding how well these groups reflect reality.

In the Novice cluster, there is a predominance of students from the first years of medical undergraduate training and a predominance of idea units classified as pathophysiology in the recall protocols. As previously mentioned, this is a faithful representation of illness script theory, by which novice students represent diseases in long-term memory primarily through networks of pathophysiology ideas.

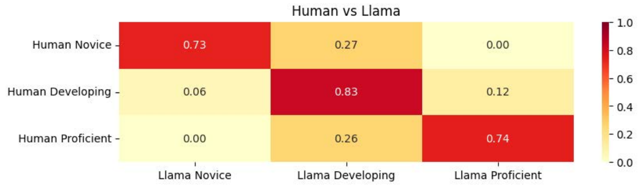

Fig. 5. Comparison between human and Llama clusters. Each grid's cell informs the percentage of students' answers in human clusters that also appear in the Llama cluster.

The Developing cluster, on the other hand, mostly represents students from the middle of the undergraduate period, and their texts have a less clear predominance of one dimension. The radar chart indicates that there are still many pathophysiology ideas, but they are not clearly predominant, with other dimensions increasing in importance. The less defined nature of this cluster may possibly mirror a period in medical training where Brazilian students are still consolidating their knowledge and discovering what is truly important through initial clinical contact with patients. Furthermore, as the categorization process is a continuous process, it is expected that there may be some spillover of characteristics from the edge of one cluster to the other, and this separation will never be perfect.

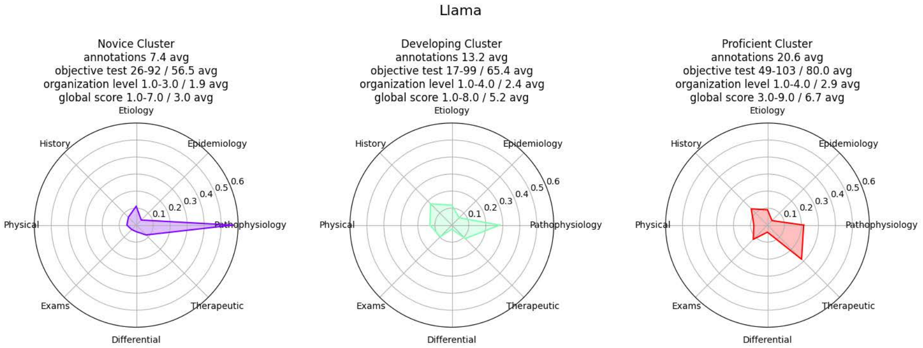

Fig. 6. Categories' distribution inside each cluster. Based on Llama annotations.

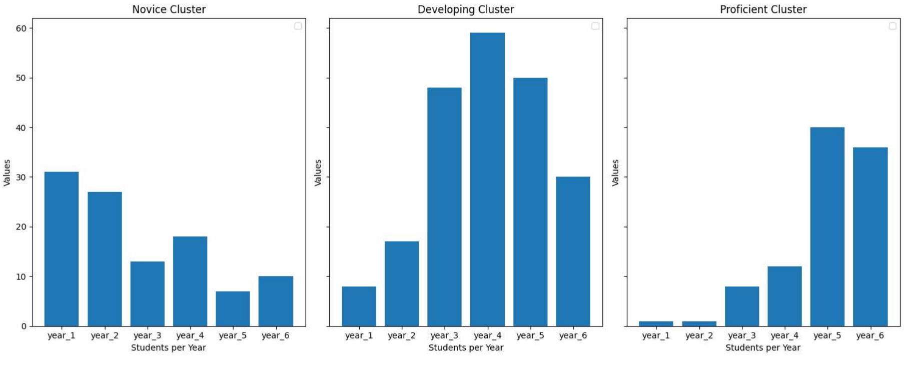

Fig. 7. Distribution of the student's year in each cluster. Based on Llama annotations.

Finally, the Proficient cluster aggregates senior students. There are some distinctions between the human and Llama profiles. Both identify the therapeutic dimension as predominant. As discussed in the previous section, this is understandable, since knowledge of treatments is typically acquired toward the end of medical training. Humans and Llama profiles differ in the second predominant dimension: history and pathophysiology, respectively. As students accumulate clinical experience, they become increasingly aware—and rightfully so—that history is important. The human evaluation is consistent with this tendency, and it may account for the observed differences in categorization.

C. Profile and Categorization from BioBERTpt

Fig. 8 presents the same comparison between clusters based on human and BioBERTpt annotations. Figs. 9 and 10 display the respective profiles of the clusters. This comparison adopts a smaller group of texts—20% of the previous comparison—since this solution required the remaining data for training and validation.

The BioBERTpt's Novice cluster demonstrated better human similarity than Llama, while it had worse similarity in the Developing cluster due to the fuzzy boundaries. Compared to humans, a very low rate of students grouped in the Novice cluster are classified as Proficient, and vice versa.

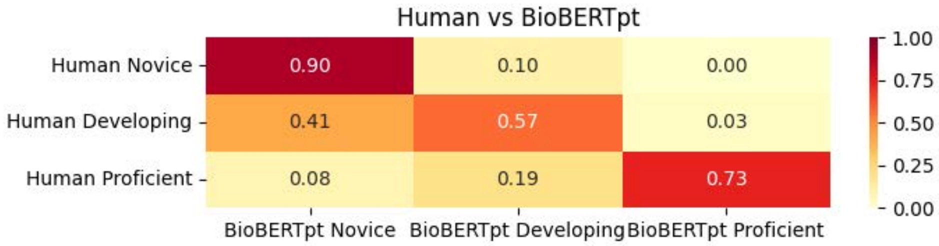

Fig. 8. Comparison between human and BioBERTpt clusters.

As Fig. 8 shows, the BioBERTpt produced similar profiles to the human version. This is a positive and expected result, as mentioned earlier, its simpler architecture can be tuned to enclose the human annotation.

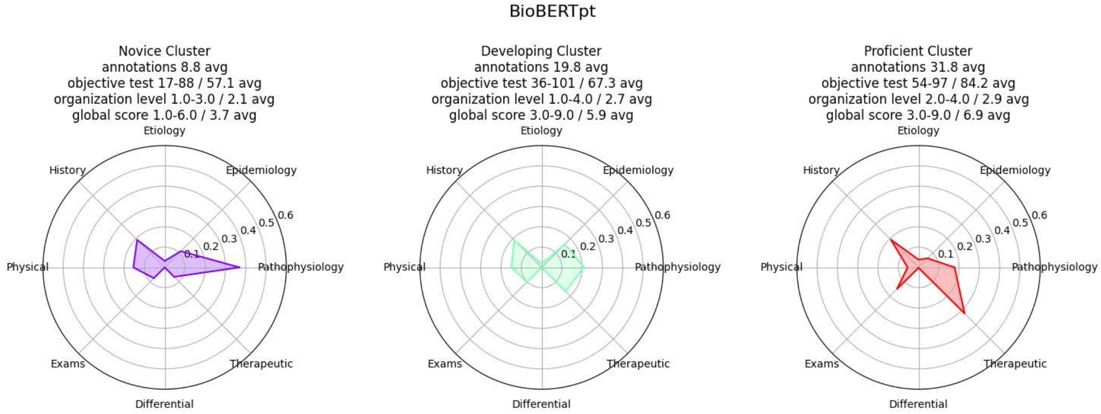

Fig. 9. Categories' distribution inside each cluster. Based on BioBERTpt annotations.

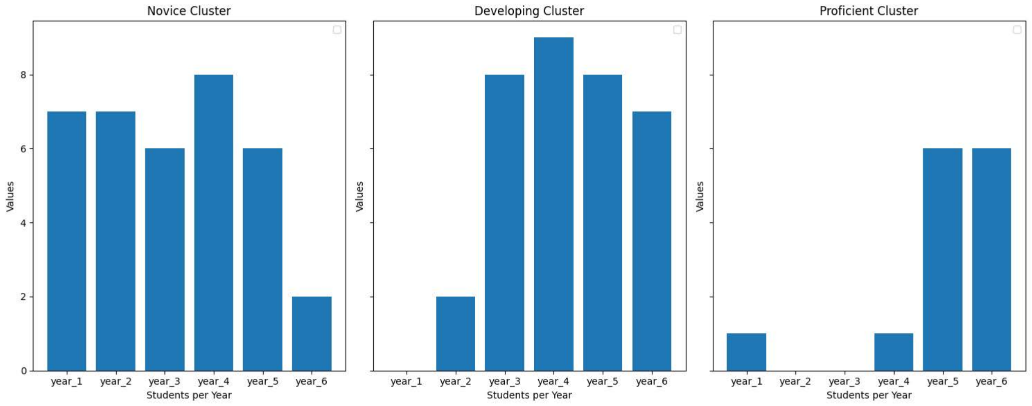

Fig. 10. Distribution of the student's year in each cluster. Based on BioBERTpt annotations.

The two automatic approaches achieved results close to human annotations. Some classification differences stem from the unsupervised discovery mechanism, which uses a non-deterministic method to identify classes with fuzzy boundaries. Still, they applied a consistent distinction policy, with differences only in neighboring classes.

There is a tradeoff of human/machine cost in the automatic mechanisms adopted. While Llama requires much less effort to guide via in-context learning, it has a high computational cost for daily use, necessitating expensive computational infrastructure and consuming significantly more energy. On the other hand, BioBERTpt, a much simpler and cheaper solution that runs even on personal computers and some mobile devices, required a significant amount of human effort to train.

D. Profile and Categorization from BioBERTpt-Llama

To address this tradeoff, a BioBERTpt model was trained using Llama annotations instead of human annotations. This solution can lead to a good balance between the low cost of human annotations and the low cost of computational infrastructure and energy consumption. The results are presented in Figs. 11, 12, and 13.

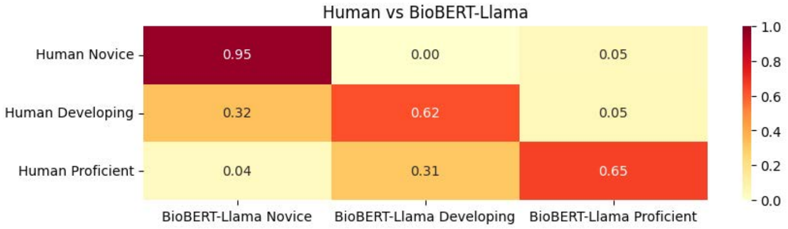

Fig. 11. Comparison between human and BioBERTpt clusters fine tuned with Llama (BioBERTpt-Llama).

As shown in Fig. 11, the model demonstrated performance comparable to that of the previous approaches. While Novice students' categorization achieved the best results, Development and Proficient students differ by one-third of the students classified in the previous cluster. However, this model maintains a consistent distinction policy, as differences in classification are always within neighboring classes.

Figs. 12 and 13 show that the distribution of dimensions and students within clusters is quite similar to that of BioBERTpt, which was trained on human annotations.

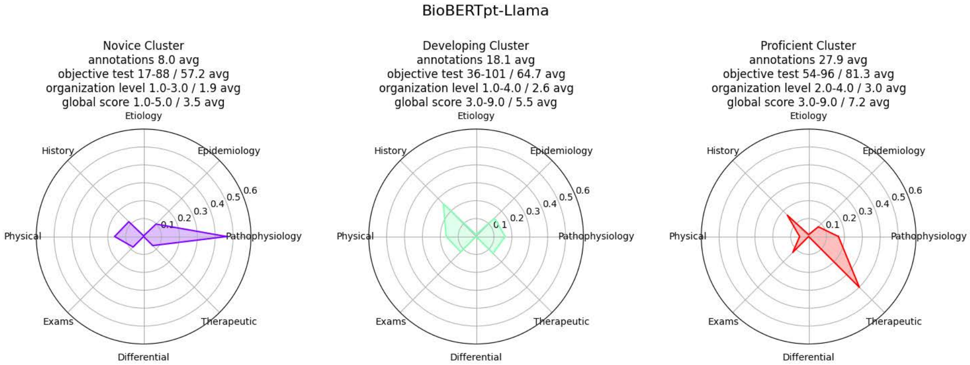

Fig. 12. Categories' distribution inside each cluster. Based on BioBERTpt-Llama annotations.

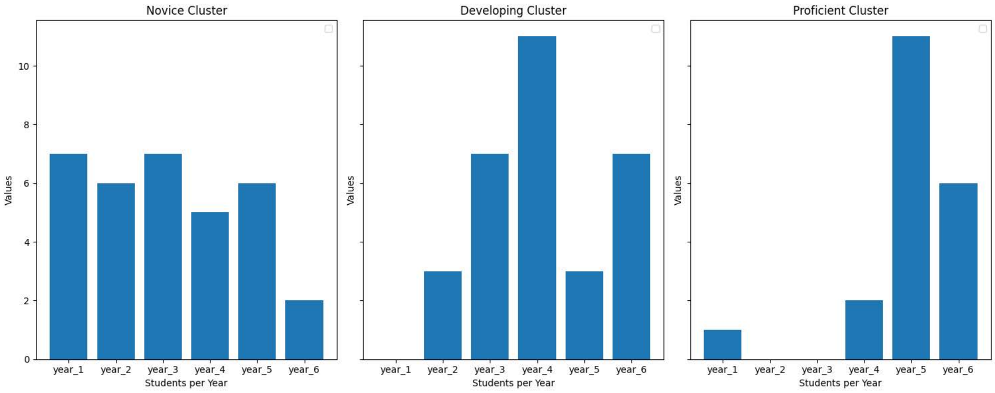

Fig. 13. Distribution of the student's year in each cluster. Based on BioBERTpt-Llama annotations.

E. Clusters and Student's Years Correlation Analysis

To validate the clustering results further, this study performed a correlation analysis between cluster membership and the academic stage of students using the biennial system, utilizing the Spearman's Rank Correlation Coefficient. This correlation was performed on the full dataset, including both the human and Llama data, as well as the subset used for testing BioBERTpt. This subset contains 84 texts and is identified by the addition of "_84" in Table 2.

The results present a moderate correlation between academic stage and cluster membership for both the manual and automated approaches. An example of the distribution of "Human_84" is shown in Fig. 14.

Table 2. Spearman's correlation (ρ) between cluster membership and academic stage

Approach	Spearman's *p*
Human	0.52
Human_84	0.55
Llama	0.53
Llama_84	0.60
BioBERTpt	0.47
BioBERTpt-Llama	0.48

Fig. 14 supplements the correlation analysis by displaying the distribution of students among clusters according to their academic stage for Human_84 annotations. It shows that, although the correlation is moderate, the students are concentrated and a smaller number are far from the neighboring cluster.

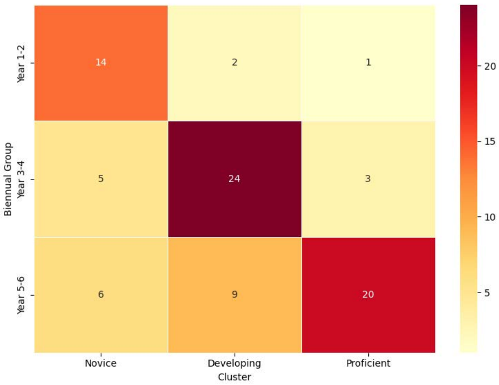

Fig. 14. Distribution of students by cluster and academic stage. Data is from the human annotations subset (Human_84).

Through a deeper analysis, it is possible to conclude that the students from the early years are rarely grouped in the cluster related to the later stages, while the opposite happens more frequently. Consequently, this behaviour complies with the training stages, since a student in earlier years will rarely have contact with the curricular material from the later stages. In contrast, students in the later stages might not have absorbed the sufficient knowledge regarding the COPD disease, and in turn they get grouped in earlier clusters.

This finding strengthens the objective of this research. The illness script evaluation is focused on identifying these students, so that the professor can have a clear understanding of the supplementary knowledge needed for the student.

VII. Conclusions and Future Work

This work tackles the challenge of assessing the development of medical students' clinical expertise. To achieve this, an illness-script-driven strategy was developed to evaluate medical students, exploring Large Language Models.

Rather than merely assigning grades to students' answers and relying on LLMs as "black boxes," this work uses an unsupervised method to cluster students' profiles based on annotations of illness-script components. The results showed that the categorization method aligns with student progress, linking components of their clinical reasoning to the knowledge and skills expected in the Brazilian National Curricular Guidelines.

This research investigated the tradeoff between two methods: a fine-tuned, BioBERTpt-based encoder-only LLM and a prompt-guided, Llama-based decoder-only LLM. The study also experimented with a hybrid approach—training the BioBERTpt-based model with Llama-generated annotations—which yielded promising results. In this process, the analysis utilized annotations from students' answers as an intermediary step to profile and categorize them.

The annotation of illness script entities and respective components (categories) is susceptible to interpretations. There is still a challenge for comparing annotations between human and automated approaches to represent annotation quality for illness scripts. However, it is important to note that the annotations are an intermediate step in tracing student profiles. In other words, despite the differences between human and automated annotations due to subjectivity, this work focused on demonstrating the correlation between the profiles traced by both.

There are still open questions in both scenarios; for example, the capacity for generalization that could reduce costs is under investigation, especially in the BioBERTpt case. Additionally, there is the recurrent challenge of model bias and the pre-trained data influence to the annotation process. In this scenario, the bias problem is most present in the Llama LLM and could result in incorrect entity recognition. Even though this work includes in-context learning—a method that also minimizes model bias —, further studies must be conducted focusing on this implication.

Automating the evaluation of illness scripts, based on an understanding of how diagnostic expertise develops, will positively impact medical education. By enabling the large-scale evaluation of complex knowledge, this approach helps teachers and schools refine their educational processes. Since students develop diagnostic expertise at different paces, this method facilitates individualized learning experiences. For example, a student with extensive knowledge of risk factors but limited understanding of symptoms has different needs than one with another profile. Moreover, the system can identify students who are not progressing as expected—such as a 3rd-year student classified as a novice—enabling timely remediation. By assessing the quality of their illness scripts, teachers can tailor educational experiences to meet specific student needs.

This work has some limitations that are the focus of ongoing research. Although the method is designed to be generic, the evaluation is limited to one disease. The current work involves a second case study scenario of acute myocardial infarction to evaluate how the models fine-tuned for COPD perform, with the goal of expanding to other scenarios. The experiments evaluated Brazilian Portuguese texts and their alignment with the Brazilian curricular structure. The plans involve expanding the study to European universities. These new experiments will refine metrics about the requirements of annotated data for fine-tuning or guiding models.

Several characteristics of this experiment can result in biased behavior. Besides language, the training context considered Brazilian courses from a specific perspective, transforming the illness script theoretical framework into a mechanism to profile students. These limitations and possible biases indicate that applying this proposal to real-world scenarios must consider ethical aspects. Specifically, this technique must support human educators in providing follow-up, evaluation, and feedback.

Code and Data Availability

The data and code that support the findings of this study in our GitHub repository: https://github.com/harena-lab/harena-illness-script.

The **bioberpt-dpoc-is-multiple** model developed in this study is publicly available on the Hugging Face Model Hub at: https://huggingface.co/harena-lab/bioberpt-dpoc-is-multiple.

The **bioberpt-llama-dpoc-is-multiple** model developed in this study is publicly available on the Hugging Face Model

Hub at: https://huggingface.co/harena-lab/bioberpt-llama-dpoc-is-multiple.

Appendix

Prompt template for separated labeling approach. This is the approach utilized and analyzed by this study:

"You are a medical assistant with expertise in medical document processing.

Your task is to tag entities related to these classes ONLY: pathophysiology, etiology, epidemiology, history, physical, exams, differential e therapeutic. Each line must include: (1) word or phrase, (2) class or classes that the text is a part of. Maintain the correct format. Example: ['token', ['physical']]. All the texts will be in portuguese. ONLY use JSON as the output format, starting with 'annotations'. DO NOT write, only respond in JSON format. Examples of user input and assistant output:

{{list examples here}}"

Prompt template for augmented labeling approach. This approach was not utilized in the results of this study, but was analyzed as an alternative prompt strategy: "*You are a medical assistant with expertise in medical document processing.*

Your task is to extract all entities and identify their entity types ONLY from this list: pathophysiology, etiology, epidemiology, history, physical, exams, differential e therapeutic.

You must augment the text by tagging entity types for each word directly within the text. Follow this format: "This is an example phrase, here is the [entity | entityType]".

All the texts will be in portuguese. ONLY use JSON as the output format, starting with 'annotations' and the value being the annotated text. DO NOT write, only respond in JSON format. Examples of user input and assistant output:

{{list examples here}}"

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors wrote the paper and approved the final version of the paper. H.S.M., A.S., and G.L. were responsible for the computer science aspects of the implementation and experiments. F.V., M.A.C.F., and L.M.C.R. were responsible for the medical research's conceptual and theoretical aspects. F.V. led the manual text annotations. F.V. and M.A.C.F. developed the annotation framework based on the Illness Script theory.

Funding

This study was financed, in part, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. CNPq (Brazilian National Council for Scientific and Technological Development) grant number 400062/2023-2. This work was supported by Amazon Web Services, Inc. This work was supported by BTG Pactual.

Acknowledgment

The authors thank all the medical students that participated in the experiment. We also thank the following contributors:

Professor Angélica Maria Bicudo from Universtity of Campinas

Professor Marcos Antônio Marton Filho from USP Bauru

Professor Alba Regina de Abreu Lima from FAMERP

Professor Maria Cristina de Andrade from Unifesp

Professor Samira Yarak from Unifesp

Professor Leandro Arthur Diehl from UEL

Professor Neide Tomimura Costa from UEL

Professor Andréa Aparecida Contini Rodrigues from UfSCar

References

[1] R. M. Epstein, "Assessment in medical education," *New England Journal of Medicine*, vol. 356, no. 4, pp. 387–396, Jan. 2007. doi: 10.1056/NEJMra054784

[2] H. P. A. Boshuizen, H. G. Schmidt, E. J. F. M. Custers, and M. W. Wiel, "Knowledge development and restructuring in the domain of medicine: The role of theory and practice," *Learning and Instruction*, vol. 5, no. 4, pp. 269–289, Jan. 1995. doi: 10.1016/0959-4752(95)00019-4

[3] E. J. F. M. Custers, "Thirty years of illness scripts: Theoretical origins and practical applications," *Medical Teacher*, vol. 37, no. 5, pp. 457–462, May. 2015. doi: 10.3109/0142159X.2014.956052

[4] H. G. Schmidt and H. P. A. Boshuizen, "On acquiring expertise in medicine," *Educ Psychol Rev*, vol. 5, no. 3, pp. 205–221, Sep. 1993. doi: 10.1007/BF01323044

[5] K. E. Hauer, C. Boscardin, J. M. Brenner, S. M. van Schaik, and K. K. Papp, "Twelve tips for assessing medical knowledge with open-ended questions: Designing constructed response examinations in medical education," *Medical Teacher*, vol. 42, no. 8, pp. 880–885, Aug. 2020. doi: 10.1080/0142159X.2019.1629404

[6] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1638–1649

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv: arXiv:1810.04805, May 2019. doi: 10.48550/arXiv.1810.04805

[8] Y. Liu e*t al*., "RoBERTa: A robustly optimized BERT Pretraining Approach," arXiv:1907.11692, Jul. 2019. doi: 10.48550/arXiv.1907.11692

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," arXiv:1909.11942, Feb. 2020. doi: 10.48550/arXiv.1909.11942

[10] M. Monajatipoor *et al*., "LLMs in Biomedicine: A study on clinical Named Entity Recognition," arXiv:2404.07376, Jul. 2024. doi: 10.48550/arXiv.2404.07376

[11] T. B. Brown et al., "Language models are few-shot learners," in *Proc. the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., Dez. 2020, pp. 1877–1901.

[12] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, and E. Bernard, "NuNER: Entity recognition encoder pre-training via LLM-annotated data," arXiv:2402.15343, Feb. 2024. doi: 10.48550/arXiv.2402.15343

[13] S. Wang et al., "GPT-NER: Named entity recognition via large language models," arXiv:2304.10428, Oct. 2023. doi: 10.48550/arXiv.2304.10428

[14] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting relation extraction in the era of large language models," in *Proc Conf Assoc Comput Linguist Meet*, vol. 2023, pp. 15566–15589, Jul. 2023. doi: 10.18653/v1/2023.acl-long.868

[15] B. Athiwaratkun, C. N. Santos, J. Krone, and B. Xiang, "Augmented natural language for generative sequence labeling," arXiv:2009.13272, Sep. 2020. doi: 10.48550/arXiv.2009.13272

[16] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using BART," arXiv:2106.01760, Jun. 2021. doi: 10.48550/arXiv.2106.01760

[17] M. O. Topal, A. Bas, and I. Heerden, "Exploring transformers in natural language generation: GPT, BERT, and XLNet," arXiv:2102.08036, Feb. 2021. doi: 10.48550/arXiv.2102.08036

[18] Q. Cheng, L. Chen, Z. Hu, J. Tang, Q. Xu, and B. Ning, "A novel prompting method for few-shot NER via LLMs," *Natural Language Processing Journal*, vol. 8, 100099, Sep. 2024. doi: 10.1016/j.nlp.2024.100099

[19] M. Chary, S. Parikh, A. F. Manini, E. W. Boyer, and M. Radeos, "A review of natural language processing in medical education," *West J Emerg Med*, vol. 20, no. 1, pp. 78–86, Jan. 2019. doi: 10.5811/westjem.2018.11.39725

[20] A. Sarker, A. Z. Klein, J. Mee, P. Harik, and G. Gonzalez-Hernandez, "An interpretable natural language processing system for written medical examination assessment," *Journal of Biomedical Informatics*, vol. 98, 103268, Oct. 2019. doi: 10.1016/j.jbi.2019.103268

[21] H. B. Burke et al., "Assessing the ability of a large language model to score free-text medical student clinical notes: Quantitative study," *JMIR Medical Education*, vol. 10, no. 1, e56342, Jul. 2024. doi: 10.2196/56342

[22] H. G. Schmidt, G. R. Norman, and H. P. Boshuizen, "A cognitive perspective on medical expertise: theory and implication," *Acad Med*, vol. 65, no. 10, pp. 611–621, Oct. 1990. doi: 10.1097/00001888-199010000-00001

[23] U. Schiefele and A. Krapp, "Topic interest and free recall of expository text," *Learning and Individual Differences*, vol. 8, no. 2, pp. 141–160, Jan. 1996. doi: 10.1016/S1041-6080(96)90030-8

[24] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutor Quant Methods Psychol*, vol. 8, no. 1, pp. 23–34, Jul. 2012. doi: 10.20982/tqmp.08.1.p023

[25] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," arXiv:cmp-lg/9505040, May. 1995. doi: 10.48550/arXiv.cmp-lg/9505040

[26] E. T. R. Schneider *et al*., "BioBERTpt—A Portuguese neural language model for clinical named entity recognition," in *Proc. the 3rd Clinical Natural Language Processing Workshop*, A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 65–72. doi: 10.18653/v1/2020.clinicalnlp-1.7

[27] H. Touvron *et al*., "LLaMA: Open and efficient foundation language models," arXiv:2302.13971, Feb. 27, 2023. doi: 10.48550/arXiv.2302.13971

[28] A. Goel *et al*., "LLMs accelerate annotation for medical information extraction," arXiv:2312.02296, Dec. 2023. doi: 10.48550/arXiv.2312.02296