A Human-in-the-Loop Pipeline for Multi-Class SRL Classification

Andreas Kissmehl
KTH Royal Institute of Technology
Stockholm, Sweden
kismehl@kth.se

Olga Viberg
KTH Royal Institute of Technology
Stockholm, Sweden
oviberg@kth.se

Olov Engwall
KTH Royal Institute of Technology
Stockholm, Sweden
engwall@kth.se

Richard Lee Davis
KTH Royal Institute of Technology
Stockholm, Sweden
rldavis@kth.se

ABSTRACT

Annotation of big data is a nontrivial task for learning analytics scholars. Educational constructs drawing from learning psychology, such as self-regulated learning, require theory-grounded domain expertise and cannot be easily inferred from surface features, making annotation time-intensive and difficult to scale. Recent work aims to address this issue by using large language models (LLMs) to annotate data. While this approach can increase the amount of annotated data, evidence suggests that LLMs tend to overgeneralize, based on the examples included in the prompt cases, and misclassify due to a lack of context understanding. To address this, we investigate how human-in-the-loop machine learning strategies can be used to maximize the impact of human annotation by focusing efforts on a subset of data that would benefit most from labeling or relabeling. Utilized data include human-labeled, machine-labeled, and unlabeled student prompts from diverse educational contexts. We compare two active learning methods, loss weighting and their combination on the task of multi-class classification of student-generated text in AI-chatbot interactions. Experiments with a multilingual DeBERTa-based classifier indicate that combining human annotation with selective correction of weak labels can improve performance compared to a baseline trained on available annotated data annotated by humans and LLMs. The human-in-the-loop pipeline mitigates misalignment in labels produced by LLMs, providing a path forward to more responsible and rigorous use of LLMs in educational data annotation tasks.

Keywords

Human-in-the-loop learning, active learning, annotation quality, large language models, self-regulated learning

Andreas Sebastian Kissmehl, Olga Viberg, Olov Engwall, and Richard Lee Davis. A Human-in-the-Loop Pipeline for Multi-Class SRL Classification. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 400–407. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040133

1. INTRODUCTION

Educational constructs such as self-regulated learning (SRL), motivation, and help-seeking (e.g.,[20, 21, 11]) are widely used as proxies for studying students' learning behaviors. However, these constructs are inherently latent and not directly observable in student interactions; instead, they must be inferred through a theoretically grounded interpretation. Researchers therefore need to operationalize them through observable indicators, such as discourse patterns, behavioral traces, or interaction features, connecting abstract theoretical constructs to empirical data. This process is necessarily interpretative and theory-laden, making annotation less a technical labeling exercise and more an expert-driven analytical task. Producing high-quality annotated datasets is thus essential but also highly time-intensive, costly, and difficult to scale, particularly when expertise in both learning theory and specific educational contexts is required [10]. The strong dependence on expert judgment and contextual understanding further limits the feasibility of substituting expert annotators with crowdsourced approaches without risking reduced validity and theoretical fidelity [15].

These challenges are becoming more pronounced as students increasingly engage with AI conversational agents such as Claude in higher education [7]. Such interactions generate large volumes of rich process data that offer unprecedented opportunities to examine how learning unfolds in AI-mediated contexts. However, the scale, heterogeneity, and context sensitivity of these interaction logs intensify existing annotation challenges. Without robust theoretically grounded annotation frameworks, translating these large-scale student-AI interaction datasets into meaningful insights about learning processes remains highly complex and not straightforward.

To address these challenges, scholars have increasingly explored the potential of LLMs (see, e.g., [17, 1]) to support the annotation of educational data. Their capacity to process large volumes of text suggests a promising pathway toward scaling annotation efforts that have traditionally relied on intensive expert labor. However, a central concern remains how limited expert time and resources can be allocated so that LLM-assisted annotation increases the quantity of la-

beled data without compromising quality, theoretical validity, or interpretability [29]. Emerging research indicates that LLMs can facilitate large-scale annotation while reducing reliance on time-consuming expert coding. However, the reliability and theoretical fidelity of LLM-generated labels remain highly sensitive to prompting strategies, contextual framing, and task specification [9, 23]. Consequently, efficiency gains often risk coming at the expense of validity and alignment with underlying learning theories. These tensions suggest that fully automated annotation is unlikely to be sufficient for theory-driven educational research. Instead, *hybrid human–AI approaches* that strategically combine machine scalability with expert oversight may offer a more viable path forward [9, 19]. The key challenge is therefore not simply automation, but the effective orchestration of human expertise within LLM-supported workflows.

In response, we investigate a human-in-the-loop (HITL) annotation pipeline designed to support large-scale coding of educational datasets while maintaining theoretical rigor. The proposed pipeline integrates active learning techniques (see Section 2.2) to identify uncertain, potentially incorrect, or missing labels and prioritize them for expert review. In parallel, it incorporates loss-weighting strategies that reduce the influence of lower-confidence LLM-generated labels during model training. Together, these mechanisms aim to improve annotation efficiency while preserving validity and theoretical coherence. We evaluate the individual and combined contributions of these components using SRL classification [21] as a case study. SRL provides a particularly suitable testbed because it involves theoretically grounded constructs (e.g., effort regulation and elaboration) that are not directly observable and therefore require interpretative coding. Although the empirical evaluation focuses on SRL, the proposed workflow is intended to generalize to other theory-driven qualitative annotation tasks in education and related domains. This work is guided by the following **research question:**

How much expert involvement is needed for a human-in-the-loop LLM coding pipeline to improve annotation quality over an automated baseline?

This paper makes three **main contributions**:

- a human-in-the-loop pipeline to support a high-quality, hybrid LLM-human coding strategy for SRL prompt classification,
- a comparative evaluation of the effectiveness of the respective components of this pipeline on a dataset,
- practical implications for shifting expert labor from comprehensive and time-intensive coding to targeted supervision and revision of LLM-based coding systems.

2. RELATED WORK

2.1 LLMs for Qualitative Analysis

Recent research has increasingly explored the use of LLMs as tools for qualitative analysis in educational research. This work spans applications such as codebook development [1, 3], deductive coding [18], and the annotation of textual data [9, 29, 32]. Across these studies, LLMs are typically positioned as supports for specific stages of qualitative workflows, with the goal of improving analytical efficiency while maintaining the validity and theoretical grounding of qualitative interpretations.

Evidence from this literature suggests that LLM-supported qualitative coding is most reliable when applied to relatively well-defined and less theoretically complex constructs. As construct complexity increases, model performance becomes more sensitive to how coding tasks are framed and specified. Prompting strategies such as few-shot prompting [17], which combine construct definitions with annotated examples, have therefore been explored to improve alignment between LLM outputs and coding schemes. Under carefully specified conditions and within particular domains, such approaches have enabled LLMs to achieve annotation performance comparable to that of humans [18, 4, 17].

However, limitations become more apparent when LLMs are applied to complex, theory-driven constructs. Prior work reports tendencies toward overgeneralization and mislabeling in such settings, as models may rely on surface-level linguistic cues rather than theoretically grounded categories when assigning codes [17].

These challenges raise concerns about the validity and interpretability of fully automated qualitative coding in theory-intensive educational research.

To mitigate these issues, scholars have begun proposing HITL approaches, in which expert annotation is selectively integrated into LLM-supported workflows [3]. Typically, a subset of the data is manually coded to serve as a ground-truth reference for evaluating LLM-generated labels. Iterative comparisons between human annotations and model outputs can then identify disagreement cases that warrant targeted expert review. Insights from these reviews are used to refine prompts, clarify codebook definitions, and improve alignment between LLM-generated labels and theoretically grounded coding schemes [3].

Despite these advances, important challenges remain. For example, while strong performance on SRL classification has been achieved when models are trained on fully human-coded datasets [33], they often fail to generalize across subject domains or languages. This implies that high accuracy may come at the cost of repeatedly producing large, consistently annotated datasets for each new context. In contrast, more general LLM-based approaches offer greater transferability but may introduce additional noise into generated labels [34].

These tensions highlight the need for approaches that enable systematic validation and refinement of LLM-generated labels without relying on fully human-coded datasets. While prior work has explored HITL strategies primarily through prompt engineering [15] and iterative feedback [3], comparatively little research has examined how LLM-generated annotations can be systematically validated and integrated into theory-driven qualitative coding pipelines at scale.

2.2 Active Learning and Loss Weighting

Active Learning offers a promising approach to maximize the impact of limited human supervision in LLM-based coding tasks. Rather than uniformly reviewing all available in-

stances, an iterative HITL approach is used in which the model actively selects data for annotation based on an uncertainty measurement. A common approach is uncertainty-based sampling, which selects samples with low prediction confidence for human (re)labeling [26].

This selection process is typically embedded within a cycle of model training, evaluation, and targeted relabeling, allowing the training set to be refined progressively.

Hybrid methods that integrate LLM-generated labels with HITL active learning methods are a relatively recent development [24, 14]. These hybrid approaches treat the LLM as a noisy annotator that can scale up to label large amounts of data, while focusing human effort on the relabeling of samples that are likely to have the largest impact on classifier performance. While these methods have shown positive impacts on the performance of general NLP tasks such as topic classification and semantic similarity [14], they have not yet been systematically evaluated in educational contexts.

A complementary approach for managing annotation quality is *loss weighting*, which explicitly accounts for differences in label reliability during model training. Loss weighting assigns increased influence to annotations from more reliable coders (e.g., a human expert), while down-weighting labels generated automatically by LLMs or the model itself. This approach makes it possible to benefit from large volumes of weakly labeled data without allowing noisier labels to dominate the training process [30].

Hybrid LLM-HITL active learning methods, alone or combined with loss weighting, have not been systematically evaluated on the classification of latent, theory-driven educational constructs. Our work aims to fill this gap by evaluating how these methods can reduce annotation costs under a limited budget by focusing expert effort on informative instances and appropriately integrating labels of differing quality, as measured by their impact on the accuracy of a mDeBERTa-v3-base model [8] trained on data processed through different stages of the pipeline.

2.3 SRL for AI-Mediated Learning

To analyze how learners regulate their learning when interacting with generative AI chatbots, the SRL theoretical lens [36, 22] may be adopted. SRL guides learning through metacognitive skills in three main phases. The forethought phase represents the planning stage of learning, in which learners outline goals for a study session. It is followed by the performance phase, where the planned activities are executed. The final phase is self-reflection, in which learners review their performance in relation to the goals and adjust the behavior [36]. To operationalize this three-phase model in the setting of generative AI chatbot use in computer science education, the SRL framework by Pintrich et al. [22] may be adapted [31, 2]. This results in seven SRL categories for the classification of user prompts: rehearsal (adjusted), organization, elaboration, help-seeking, effort regulation, critical thinking, and non-SRL behavior. Table 1 provides definitions for each category.

3. METHODS

Against this background, this study addresses the problem of prioritizing samples for human annotation to identify and correct incorrect labels generated by a weak reviewer (e.g., an LLM) under a fixed budget for human supervision and given an initial set of human-annotated reference labels.

The problem was formulated as a multi-class classification problem with seven mutually exclusive classes, where each instance is assigned at most one categorical label. This dataset reflects a real-world, low-resource scenario, characterized by strong class imbalance, with some classes being represented only by a few labels.

3.1 Datasets

The study draws on multiple datasets representing different cultural contexts, educational settings, and subject domains at higher education institutions in Sweden [31], Germany [25], and South Korea [6]. The subject areas covered include foundational programming, scientific writing, and chemistry. The selected datasets consist of end-to-end, multi-turn dialogues between individual students and an LLM. The data was collected during self-study activities such as homework assistance, reflection, and exam preparation. Each dataset was collected under distinct instructional conditions: in Sweden, 10 students used ChatGPT without explicit instructions, generating 2445 messages. In Germany, 212 students used an unspecified LLM chatbot, generating 2335 messages, and in South Korea, 131 students interacted with a GPT-3.5 through a writing platform (RECIPE [5]), generating 1913 messages. The language used to communicate with the LLM was English in 4179 messages, German in 1600, Swedish in 567, and Korean in 70 (counted after the removal of duplicates). Although the data include complete dialogic interactions, the analysis treats individual student prompts as the unit of analysis to ensure comparability across datasets. Prompts were used verbatim, maintaining orthographic and grammatical errors.

In addition to these empirical datasets, a synthetic dataset was created to address rarely represented SRL constructs. Specifically, the *rehearsal* construct occurred only three times across the empirical data. To supplement this limitation, additional examples were generated using an iterative multi-shot prompting approach with the OpenAI GPT-5.2-2025-12-11 model following Litake et al. [16]. All synthetically generated instances were manually reviewed before inclusion. The prompt used for data generation is provided on OSF. The codebook for the project can be found in Table 1.

3.2 Datasets and Preparation

The different datasets were joined by maintaining the original dataset name, the student prompt, and, if present, the SRL label. Except for the Swedish dataset, the datasets do not include SRL annotations. In addition, we make use of a dataset reported in a master's thesis [37], which contains chat transcripts from self-study activities of Swedish STEM students together with LLM-generated SRL codes. The labels were normalized across datasets, and the source of each label was encoded based on the originating dataset as human-labeled, LLM-labeled, or unlabeled. The data cleaning and preparation consist of four subsequent steps:

1. Text cleaning: All prompts were cleaned before being

Table 1: Codebook of the theoretical lens used.

Construct	Description	Example
Rehearsal	Addresses drill-and-practice behavior, where students repeat learning material to memorize it. In prompt-based interaction, rehearsal may be reflected by repeated or highly similar prompts across learning sessions, including quiz-like activities where the student receives questions from a chatbot.	*"hi, i would like to review about the results of my quiz. I got wrong in this question [...]"*
Organization	Prompts in which students request summarization, support for structuring and editing content, such as spelling correction or writing assistance.	*"Then, what is the repeted word in my essay?"* or *"make this paragraph coherence and better for the grammar[...]"*
Elaboration	Prompts where students ask for additional information to better understand a concept by making connections between ideas.	*"What does Counter do?"* or *"What do linearly recursive, end-recursive, and non-linearly recursive mean?"*
Help-seeking	Prompts in which students directly request solutions or explicit guidance to solve a task.	*"def f4(m, n): if m == 0: return n + 1 elif n == 0: return f4(m - 1, 1) else: return f4(m - 1, f4(m, n - 1)) y = f4(1, 1) What is the return value? Number of calls? What is the recursion type?"*
Effort regulation	Prompts in which students ask for additional resources or explanations aimed at persisting with learning tasks.	*"For example, if my work's limitation is that it only works at the specific domain (graphic design), how can I write my limitation well?"*
Critical thinking	Prompts where students evaluate information, question assumptions, or reflect on the validity of responses.	*"No this is definitively wrong."* or *"Should it not be 5 due to our prior result?"*
Non-SRL behavior	Prompts that cannot reliably be coded as rehearsal, organization, elaboration, help-seeking, effort regulation, or critical thinking. This includes greetings and acknowledgments.	*"nope, thats it. thanks"* or *"Hello"* or *"Great, thanks"*.

further processed. This includes Unicode normalization, removal of control characters, leading and trailing whitespace, line breaks, and the removal of boilerplate tags and phrases originating from accessibility-related user interface components of the chatbot (e.g., file upload prompts).

2. Label normalization: The existing labels were normalized with, e.g., different capitalization, manually reviewed, and joined into canonical labels.

3. Duplicate handling: Rows with missing prompts were removed from the dataset. Exact duplicates, which are defined as a row with an identical combination of text, label, and label sources, were collapsed into a single instance. Rows with identical text but conflicting labels were merged by keeping the human annotation if present. In the case of conflicting LLM-generated labels, the rows were collapsed and marked as unlabeled, dropping the LLM lables .

4. ID Generation and Language Detection: In the last preprocessing step, a UUID was generated for each row using the cleaned text and the dataset name. Using *langid* the language of each prompt was detected, limited to the languages present in the datasets (Swedish, German, English, and Korean).

3.3 Baseline Model

As a baseline, the multilingual mDeBERTa-v3-base model [8] was fine-tuned as a text classifier for SRL label prediction at the prompt level using the Hugging Face Transformers library with the standard sequence classification architecture. Individual student prompts were used as the unit of analysis, with a single categorical SRL label per input. Following an initial hyperparameter tuning, inputs were tokenized using the corresponding pretrained tokenizer and truncated to a maximum sequence length of 128 tokens, with dynamic padding applied during batching. Fine-tuning was performed for five epochs using an AdamW-style optimizer with a learning rate of 3×10^{-5}, a weight decay of 0.01, and a warm-up ratio of 0.06, following a cosine learning-rate schedule. A per-device batch size of two with gradient accumulation over eight steps resulted in an effective batch size of sixteen. All trained checkpoints were retained and manually inspected for analysis, rather than selecting a single checkpoint based on an automated validation criterion. Following Sokolova and Lapalme [27], the evaluation metric used was $F1_{macro}$ due to the class imbalance.

3.4 Loss Weighting

To prioritize human-annotated instances x_i^{Hu}, a per-instance loss weight was introduced. Each training instance x_i is assigned a strictly positive weight $w_i > 0$. Human-annotated samples receive unit weight, while LLM-labeled samples are either assigned a fixed downweighting factor $w_{LLM} < 1$, treated as a tuned hyperparameter, or weighted individually using w_i^{LLM}. A normalization $\sum_i w_i$ with ℓ_i as the standard cross-entropy loss for x_i, limits the effect of weighting to the relative contribution of individual samples, leaving the overall gradient scale unchanged.

$$\mathcal{L} = \frac{\sum_{i=1}^{N} w_i, \ell_i}{\sum_{i=1}^{N} w_i}.$$

3.5 Active Learning for Label Acquisition

Active learning was applied exclusively to the unlabeled subset of the dataset. At each iteration of the label acquisition (LA), an uncertainty score was computed for all unlabeled instances using predictive entropy following [26]. The 50 instances with the highest entropy values were selected for manual annotation. The newly annotated samples were added to the training set, and the model was retrained using the same architecture and hyperparameter configuration as in the baseline setup. This procedure was repeated in iterative cycles until the annotation budget of 200 labels was exhausted.

3.6 Active Learning for Label Correction

To address the varying LLM annotation quality, a weighting approach was applied and evaluated. It is assumed that LLM-labeled instances x_i^{LLM} located close to clusters of human-annotated instances x_i^{Hu} in the model embedding space are more likely to be correct than instances that are further away, as discussed by Kontonatsios et al. [12]. Such instances were prioritized for label correction (LC) by human review.

This approach was implemented using the cosine similarity. For each instance x_i with one label set L_i, an embedding $e_i \in \mathbb{R}^d$ was computed in a shared d-dimensional embedding space using the mDeBERTa transformer encoder [8]. For each SRL label l, a normalized centroid c_l was defined as the mean embedding of all x_i^{Hu} instances assigned to the label l. These centroids were then used to define per-instance weights.

$$w_i^{LLM} = \max\left(\max_{l \in L_i} \exp\left(-\frac{1 - e_i^\top c_l}{\sigma_l}\right), \varepsilon\right),$$

where σ_l describes the median cosine similarity between x_i^{Hu} of label l and its centroid. Human-annotated instances are assigned unit weight ($w_i = 1$), and all weights are lower-bounded by a small constant$w_i \leftarrow \max(w_i, \varepsilon)$.

To identify candidates for human review, the w_i metric was combined with similarity-based indicators defined as

$$\delta = d_{\text{nearest}} - d_{\text{self}}$$

where d_{self} is the cosine similarity between x_i^{LLM} and the centroid c_l that corresponds to the assigned label, and d_{nearest} is the similarity to the closest centroid of a non-assigned label. A larger value of δ indicates a candidate for mislabeling. Based on the w_i metric, a review flag was assigned when w_i was in the lowest decile and prioritized by δ.

3.7 Evaluation Study

An evaluation study was concluded to assess the effectiveness of loss weigting, active learning through label acquisition (LA), and active learning for label correction (LC). For this purpose, after the initial cleaning described in section 3.2, a baseline dataset was constructed from the existing human-labeled instances x_i^{Hu}, the machine-labeled instances x_i^{LLM}, and unlabeled instances $x_i^{\emptyset}$. The resulting dataset includes 6416 samples with 836 x_i^{Hu}, 2016 x_i^{LLM} and 3564 $x_i^{\emptyset}$.

This dataset was used to train the baseline model and was iteratively extended in batches of 50 instances through additional human annotation and re-annotation of x_i^{LLM} instances. In total, this process resulted in ten distinct datasets summarized in Table 2.

An initial split was defined using only the $\boldsymbol{x_i^{\text{Hu}}}$ instances for the test and validation sets, while for the training set, the x_i^{LLM} and remaining $\boldsymbol{x_i^{\text{Hu}}}$ were used. The split between training, validation, and test sets was 80/10/10, while ensuring that all classes were present in the validation and test datasets. We first randomly select the latter sets and use the rest of the data to fill up the training set. The training set was iteratively improved by adding human labels to unlabeled instances (LA approach) or correcting LLM-labeled instances (LC approach), while the validation and test sets were unchanged.

Table 2: Dataset composition across annotation strategies: label acquisition (LA) and label correction (LC). (50–200 = human additional annotations or reviews).

	Dataset	$\boldsymbol{x_i^{\text{LLM}}}$	$\boldsymbol{x_i^{\text{Hu}}}$	$\boldsymbol{x_i^{\emptyset}}$
	Baseline	2016	836	3564
Iteration 1	LA 50	2016	886	3514
Iteration 2	LA 100	2016	936	3464
Iteration 3	LA 150	2016	986	3414
Iteration 4	LA 200	2016	1039	3364
Iteration 1	LC 50	1966	886	3564
Iteration 2	LC 100	1916	936	3564
Iteration 3	LC 150	1866	986	3564
Iteration 4	LC 200	1816	1036	3564
	Combined	1816	1236	3164

4. RESULTS

As indicated in Table 3, a macro-F1 score of 0.328 and an accuracy of 0.352 are achieved for the baseline model. For the baseline dataset, instance-based loss weighting w_i leads to improved performance compared to the unweighted, but this is not the case for the augmented datasets LA-200 and LC-197. For these, a fixed assignment weight, determined via grid search, instead results in higher performance. The weight increased with the number of iterations during which labels were added or corrected, indicating that the dataset quality improved.

Table 3: Performance across iterations, strategies, and weighting (none, per-instance, and best constant $w^\star$).

Iteration	Dataset	Loss weighting	Macro F1	Macro Recall	Accuracy
	Baseline	none	0.328	0.323	0.352
	Baseline	per ins.	0.350	0.377	0.383
	Baseline	$w^\star = 0.4$	0.371	0.396	0.379
1	LA-50	none	0.365	0.400	0.380
2	LA-100	none	0.382	0.407	0.393
3	LA-150	none	0.327	0.353	0.352
4	LA-200	none	0.371	0.395	0.380
4	LA-200	per ins.	0.329	0.349	0.440
4	LA-200	$w^\star = 0.3$	**0.402**	**0.418**	0.490
1	LC-50	none	0.324	0.355	0.419
2	LC-100	none	0.320	0.344	0.407
3	LC-150	none	0.350	0.345	0.457
4	LC-197	none	0.366	0.391	0.453
4	LC-197	per ins.	0.334	0.366	0.430
4	LC-197	$w^\star = 0.9$	0.375	0.398	0.467
4	Combined	none	0.386	0.405	0.494
4	Combined	per ins.	0.326	0.354	0.448
4	Combined	$w^\star = 0.9$	0.395	0.412	**0.498**

For active learning through LA, macro-F1 scores increase across successive LA iterations up to LA-200, except for a performance drop for LA-150, for which $F1_{Macro}$ falls below the baseline, likely reflecting sensitivity to sampling effects, class imbalance, and label noise [26]. For LA, instance-based loss weighting does not provide additional gains, whereas a fixed loss weight resulted in the highest overall $F1_{Macro} = 0.402$ for LA-200.

Similar results are observed for active learning through LC,

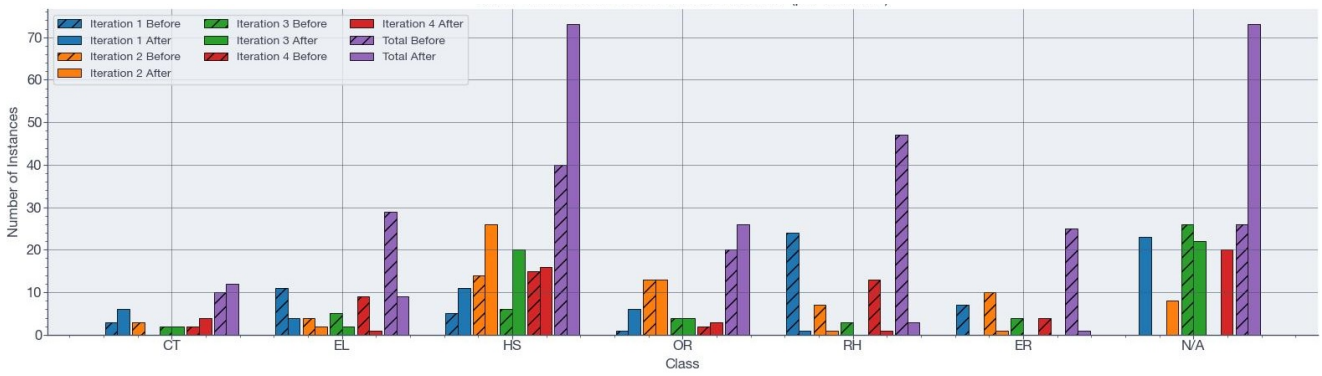

Figure 1: Class-wise label distribution for each LC Iteration and in Total, with Before/After bars indicating the labels among the 50 reviewed instances for that iteration. CT = Critical Thinking, EL = Elaboration, HS = Help Seeking, OR = Organization, RH = Rehearsal, ER = Effort Regulation, and N/A = no applicable SRL label.

outperforming the baseline from LC-150 and up. As in the LA setup, instance-based weighting performs worse than no weighting, and a fixed loss weight performs the best. Across the four LC iterations, 197 instances (out of 200, three samples were inadvertently removed due to a procedural oversight but were identified during the final write-up) were selected for review. Of these, 120 instances (60.9%) were identified as having been incorrectly labeled and were corrected. The distribution of corrected labels indicates an overgeneralization within the LLM-generated labels, with SRL labels having been assigned by the LLM, although the content did not align with any construct in the codebook. The absolute number of *Non-SRL* was hence drastically increased (+48) after label correction. In addition, the numbers of *Help seeking* (+34), *Organization* (+7), and *Critical Thinking* (+2) labels were also increased, while those of *Rehearsal* (-44), *Effort Regulation* (-25), and *Elaboration* (-20) were decreased. This pattern was consistent across all relabeling iterations and also corresponds to the shares in the newly labeled data in the LA iterations. The number of identified labeling errors decreased across successive iterations, from 42 (out of 50) in the first iteration to 28 in the second, 24 in the third, and 26 in the final iteration. Figure 1 summarizes the class-wise label distribution before and after reviews.

The combination of 200 newly labeled instances and 197 revised instances results in slightly higher overall accuracy compared to either the LA or LC configurations alone. A fixed loss weight provides performance close to the best-performing configuration, while instance-based weighting again performs poorly in $F1_{Macro}$.

5. DISCUSSION

LLMs have the potential to contribute meaningfully to scaling the labor-intensive process of qualitative coding of educational data when positioned as provisional labelers. Their outputs function as 'weak' and potentially noisy signals that require systematic human review rather than serving as trusted sources of annotation [35]. Addressing the research question, performance, naturally, normally increases with progressive human addition or correction of labels. In our experiment, the best results were obtained when adding 200 more human labels (LA-200) or combining addition and correction (Combined). Adding or reviewing fewer labels did not guarantee outperforming the baseline, as shown by LA-150, LC-50 and LC-100 (no weights). Reviewing 200 LLM annotations (10% of LLM labels) or adding 200 human annotations (24% increase) to a dataset that still consists of twice as many LLM-labeled and thrice as many non-labeled instances is a reasonable effort to improve performance (respectively, +11.6% and +13.1% in $F1_{Macro}$). HITL expert involvement is hence both necessary and effective in complementing or correcting LLM-generated labels. These findings extend prior work on utilizing LLMs for qualitative coding tasks [3, 17] with an approach to address limitations of overgeneralization.

Consistent patterns in how the proposed pipeline affects model performance were revealed. Active learning through LA shows overall performance gains under a fixed annotation budget, even if fluctuations occur in the LA-150 iteration, while active learning through LC improved results compared to the baseline from LC-150 and up. Without loss weighting, LC-197 is at par with LA-200 (0.366 vs. 0.371), suggesting that, under a fixed human annotation budget, correcting existing 'weak' labels can be as effective as expanding the dataset (note that 120 – not 200 – labels were corrected in LC-197). The higher label quality reduces the need for strong downweighting of LLM-generated labels ($w* = 0.9$ for LC vs. $w* = 0.3$ for LA).

Fixed loss weighting improves performance and indicates that LLM-generated labels can provide useful information if their imperfection is taken into account. Instance-based weighting worsened performance, suggesting that it relies on similarity estimates in a learned embedding space optimized for classification rather than geometric accuracy, making cosine similarity sensitive to regularization and feature scaling effects [28]. As a result, the corresponding weighting signals can become unstable during training. Distance-weighted cosine similarity is suggested to mitigate these effects [13].

This study has several limitations related to the scope of the empirical evaluation. The analysis is restricted to a single domain, which limits the generalizability of the findings to other domains. In addition, the evaluation considers only a single model architecture. While the proposed pipeline

explicitly addresses label noise and class imbalance, its effectiveness for rare constructs remains constrained by the availability of reliable human annotations. Furthermore, the study does not explicitly account for human factors such as reviewer fatigue or inter-rater reliability, which may influence the robustness of HITL annotation processes. Finally, we acknowledge that the use of existing datasets from different cultural and task contexts may limit the generalizability due to cross-context differences that may lead to different student–LLM interaction patterns.

Future work should examine the applicability of the proposed pipeline across additional domains, modeling settings, and annotation budgets. More generally, the interaction between automated labeling, human review, and data quality warrants further investigation beyond fully supervised annotation regimes. To further validate LLM-generated labels, the use of cosine similarity as an additional signal alongside inter-rater reliability (kappa) should be explored, which may help assess label consistency and improve robustness in diverse contexts. Such pipelines can be embedded into existing qualitative coding workflows to support targeted human review, building on prior work [3, 34].

The findings indicate that the effectiveness of LLM-supported annotation in educational research depends on structured human participation. Although LLM-generated labels provide a helpful boost, loss weighting combined with explicit mechanisms to involve human experts for label correction provides additional value. Improving the quality of existing weak labels via human involvement can be similarly effective as acquiring new labels. Taken together, these results show that a human-in-the-loop pipeline that treats LLM-generated labels as 'weak' signals and prioritizes targeted human correction can scale theory-driven annotation under a fixed annotation budget while preserving theoretical rigor.

Ethics Statement: In this study, we followed the ethical guidelines provided by the Swedish Ethical Review Authority (https://etikprovningsmyndigheten.se/en/).

Acknowledgements: This work has been in part supported by the EECS school (KTH) within the project: "Think for Yourself: Adaptive AI Technologies for Learning with LLMs in STEM Education". This work was also partially funded by an unrestricted gift from Google.

6. REFERENCES

[1] A. Barany, N. Nasiar, C. Porter, A. F. Zambrano, A. L. Andres, D. Bright, M. Shah, X. Liu, S. Gao, J. Zhang, et al. Chatgpt for education research: Exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*, pages 134–149. Springer, 2024.

[2] G. Cheng, D. Zou, H. Xie, and F. L. Wang. Exploring differences in self-regulated learning strategy use between high- and low-performing students in introductory programming: An analysis of eye-tracking and retrospective think-aloud data from program comprehension. *Computers and Education*, 208:104948, Jan. 2024.

[3] Z. O. Dunivin. Scaling hermeneutics: a guide to qualitative coding with llms for reflexive content analysis. *EPJ Data Science*, 14(1):28, Apr. 2025.

[4] A. Ganesh, C. Chandler, S. D'Mello, M. Palmer, and K. Kann. Prompting as panacea? a case study of in-context learning performance for qualitative coding of classroom dialog. In *Proceedings of the 17th International Conference on Educational Data Mining*, page 835–843, 2024.

[5] J. Han, H. Yoo, Y. Kim, J. Myung, M. Kim, H. Lim, J. Kim, T. Y. Lee, H. Hong, S.-Y. Ahn, and A. Oh. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 416–420, New York, NY, USA, 2023. Association for Computing Machinery.

[6] J. Han, H. Yoo, J. Myung, M. Kim, T. Y. Lee, S.-Y. Ahn, and A. Oh. RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13666–13676, Torino, Italia, May 2024. ELRA and ICCL.

[7] K. Handa, D. Bent, A. Tamkin, M. McCain, E. Durmus, M. Stern, M. Schiraldi, S. Huang, S. Ritchie, S. Syverud, K. Jagadish, M. Vo, M. Bell, and D. Ganguli. Anthropic education report: How university students use claude, 2025.

[8] P. He, J. Gao, and W. Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021.

[9] Z. He, S. Naphade, and T.-H. K. Huang. Prompting in the dark: Assessing human performance in prompt engineering for data labeling when gold labels are absent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–33, 2025.

[10] K. Huang, R. Ferreira Mello, C. Pereira Junior, L. Rodrigues, M. Baars, and O. Viberg. That's what roberta said: Explainable classification of peer feedback. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 880–886, 2025.

[11] S. A. Karabenick and J.-L. Berger. Help seeking as a self-regulated learning strategy. *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman*, 1:237–261, 2013.

[12] G. Kontonatsios, A. J. Brockmeier, P. Przybyła, J. McNaught, T. Mu, J. Y. Goulermas, and S. Ananiadou. A semi-supervised approach using label propagation to support citation screening. *Journal of Biomedical Informatics*, 72:67–76, Aug. 2017.

[13] B. Li and L. Han. Distance weighted cosine similarity measure for text classification. In H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, and X. Yao, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, page 611–618, Berlin, Heidelberg, 2013. Springer.

[14] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. Chen, Z. Liu, and D. Yang. CoAnnotating: Uncertainty-guided work

allocation between human and large language models for data annotation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore, Dec. 2023. Association for Computational Linguistics.

[15] T. Li, D. Sree, and T. Ringenberg. Assessing crowdsourced annotations with llms: Linguistic certainty as a proxy for trustworthiness. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 191–201, 2025.

[16] O. Litake, B. H. Park, J. L. Tully, and R. A. Gabriel. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31(6):1404–1410, June 2024.

[17] X. Liu, J. Zhang, A. Barany, M. Pankiewicz, and R. S. Baker. Assessing the potential and limits of large language models in qualitative coding. In *International Conference on Quantitative Ethnography*, pages 89–103. Springer, 2024.

[18] J. McClure, D. Smyslova, A. Hall, and S. Jiang. Deductive coding's role in ai vs. human performance. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 809–813, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[19] K. Misiejuk, R. Kaliisa, and J. Scianna. Augmenting assessment with ai coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, 6:100216, 2024.

[20] E. Panadero. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology*, 8:422, 2017.

[21] P. R. Pintrich. The role of motivation in promoting and sustaining self-regulated learning. *International journal of educational research*, 31(6):459–470, 1999.

[22] P. R. Pintrich, D. A. Smith, T. Garcia, and W. J. McKeachie. *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ).* National Center for Research to Improve Postsecondary Teaching and Learning, Ann Arbor, MI., 1991. ERIC Number: ED338122.

[23] I. Robinson and J. Burden. Framing the game: How context shapes llm decision-making. *arXiv preprint arXiv:2503.04840*, 2025.

[24] H. Rouzegar and M. Makrehchi. Enhancing text classification through LLM-driven active learning and human annotation. In S. Henning and M. Stede, editors, *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 98–111, St. Julians, Malta, Mar. 2024. Association for Computational Linguistics.

[25] A. Scholl and N. Kiesler. Data: Analyzing Chat Protocols of Novice Programmers Solving Introductory Programming Tasks with ChatGPT, 5 2024. Accessed: 2026-05-05.

[26] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2012.

[27] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009.

[28] H. Steck, C. Ekanadham, and N. Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890, New York, NY, USA, May 2024. Association for Computing Machinery.

[29] L. Tavakoli and H. Zamani. Reliable annotations with less effort: Evaluating llm-human collaboration in search clarifications. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 92–102, 2025.

[30] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195, Apr. 2025.

[31] O. Viberg, J. Wong, Y. Feldman-Maggor, N. Dunder, and C. D. Epp. Chatting with code: Exploring llms as learning partners in programming education. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, and S. Isotani, editors, *Artificial Intelligence in Education*, pages 453–461, Cham, 2025. Springer Nature Switzerland.

[32] C. Y. Wang and J. J. H. Lin. Utilizing artificial intelligence to support analyzing self-regulated learning: A preliminary mixed-methods evaluation from a human-centered perspective. *Computers in Human Behavior*, 144:107721, July 2023.

[33] J. Zhang, C. Borchers, V. Aleven, and R. S. Baker. Using large language models to detect self-regulated learning in think-aloud protocols. In B. PaaÄŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 157–168, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[34] J. Zhang, C. Borchers, V. Aleven, and R. S. Baker. Using large language models to detect self-regulated learning in think-aloud protocols. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 157–168, 2024.

[35] Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunnmon, J. Evans, A. Bundy, S. Dzeroski, J. Tegner, and H. Zenil. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *NPJ Artificial Intelligence*, 1(1):14, Aug. 2025.

[36] B. J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, May 2002. Publisher: Routledge _eprint: https://doi.org/10.1207/s15430421tip4102_2.

[37] S. Özdere. Chatwise: Adaptive support for students' self-regulated learning in generative ai–mediated environments. Master's thesis, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, Dec. 2025.