Conversational Education at Scale: A Multi-LLM Agent Workflow for Procedural Learning and Pedagogic Quality Assessment

Jiahuan Pei[†], **Fanghua Ye**[‡], **Xin Sun**[§¶], **Wentao Deng**[‖], **Koen Hindriks**[†], **Junxiao Wang**[*††]
[†] Vrije University of Amsterdam, Amsterdam, The Netherlands
[‡] University College London, London, United Kingdom
[§] University of Amsterdam, Amsterdam, The Netherlands
[¶] National Institute of Informatics, Tokyo, Japan
[‖] Shandong University, Qingdao, China
[††] Guangzhou University, Guangzhou, China
j.pei2@vu.nl, junxiao.wang@gzhu.edu.cn

Abstract

Large language models (LLMs) have advanced virtual educators and learners, bridging NLP with AI4Education. Existing work often lacks scalability and fails to leverage diverse, large-scale course content, with limited frameworks for assessing pedagogic quality. To this end, we propose **WikiHowAgent**, a multi-agent workflow leveraging LLMs to simulate interactive teaching-learning conversations. It integrates teacher and learner agents, an interaction manager, and an evaluator to facilitate procedural learning and assess pedagogic quality. We introduce a dataset of 114,296 teacher-learner conversations grounded in 14,287 tutorials across 17 domains and 727 topics. Our evaluation protocol combines computational and rubric-based metrics with human judgment alignment. Results demonstrate the workflow's effectiveness in diverse setups, offering insights into LLM capabilities across domains. Our datasets and implementations are fully open-sourced. [1]

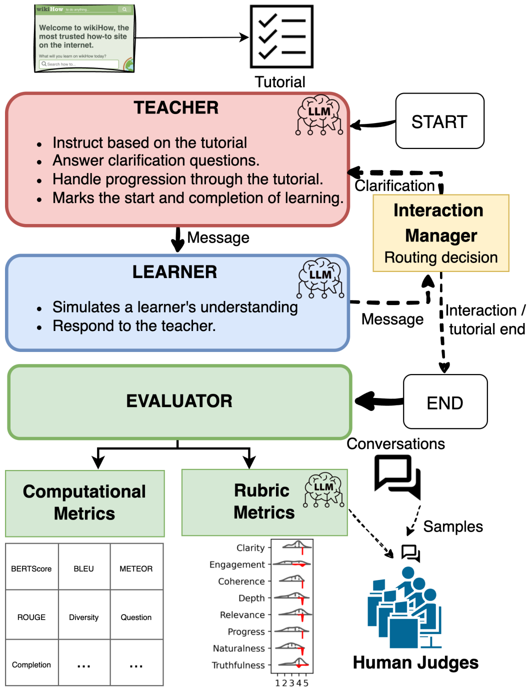

Figure 1: The proposed multi-LLM agent workflow for procedural learning and pedagogic quality assessment.

1 Introduction

Large language models (LLMs) have sparked smart teaching and learning, reshaping education and bridging NLP with AI4Education (Chu et al., 2025). LLMs have been employed as single agents in education, serving as teachers (Chen et al., 2024b; Lan and Chen, 2024), learners (Xu et al., 2025a), or evaluators (Hu et al., 2025) to enhance personalized learning (Park et al., 2024). Specifically, multi-LLMs have shown promise in simulating teacher-learner interactions, assisting competitive problem solving (Islam et al., 2024), providing personalized feedback (Park et al., 2024), immersive multimodal training (Pei et al., 2024), and evaluating educational content (Lagakis and Demetriadis, 2024; Jinxin et al., 2023). However, none of the existing approaches use diverse and large-scale course content, which is vital for capturing the variability in instructional materials. And they lack a comprehensive evaluation for assessing pedagogic quality, limiting insights into the effectiveness of teaching-learning interactions (Xu et al., 2025a).

To this end, we introduce a multi-agent workflow, namely **WikiHowAgent** (See Figure 1), leveraging LLMs to simulate an interactive teaching-learning conversation graph. It utilizes LLM-powered teacher and learner agents, processes instructional tutorials, and generates guided conversations in an interactive mode. The workflow is designed to facilitate a seamless interaction between the teacher and learner agents, allowing for a natural flow of conversation. The workflow consists of four main components: (1) **Teacher Agent**, which provides instructions, answers clarification questions, and guides tutorial progression.

[*]Corresponding author
[1]https://github.com/Jiahuan-Pei/WikiHowAgent

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2984–2997
November 4-9, 2025

(2) **Learner Agent**, which simulates understanding and generates responses, either feedback or a question if anything is unclear. (3) **Interaction Manager**, which monitors the conversation state, tracks tutorial progress, and determines the next node (i.e., TEACHER, LEARNER, END) in the graph, ensuring seamless transitions. And (4) **Evaluator**, which assesses the generated conversation using diverse evaluation metrics, including computational metrics and rubric-based metrics generated by an LLM-based evaluator agent, offering insights into pedagogic quality and model performance.

We seek the answers to the following research questions: (**RQ1**) How effective is the proposed multi-agent workflow with diverse LLMs? How does it perform in homogeneous learning (the same LLM for teacher, learner, and evaluator) versus heterogeneous learning (different LLMs for learner given the same teacher)? (**RQ2**) How do LLMs perform across various domains, and what insights can be drawn about their strengths and limitations in specific areas of procedural learning? (**RQ3**) How well do the diverse automatic evaluation metrics align with human judgment in assessing pedagogic quality and conversational performance?

We list our key contributions as follows:

- A large-scale dataset consisting of 114,296 teacher-learner conversations grounded in 14,287 tutorials. It spans 17 domains and 727 topics, organized in a hierarchical knowledge graph.
- A multi-LLM agent workflow for procedural learning and pedagogic quality assessment. A benchmark of 8 prevailing LLMs from diverse providers, including open-source and closed-source models.
- A comprehensive evaluation protocol, including diverse metrics and a human judgment alignment assessment method.

Algorithm 1 Multi-LLM Agent Workflow for Procedural Learning

Require: Tutorial $\mathcal{T} = [t_1, t_2, \ldots, t_n]$, max interactions M
Initialize State:
$\mathcal{S} \leftarrow \{$messages: [], tutorial: $\mathcal{T}$, current_step: 0, needs_clarification: False, finished: False$\}$
Define Nodes:
  TeacherNode: $\mathcal{S} \leftarrow$ teacher_agent($\mathcal{S}$)
  LearnerNode: $\mathcal{S} \leftarrow$ learner_agent($\mathcal{S}$)
Simulate a conversation by iterating through the nodes in the graph:
while $\neg\mathcal{S}$.finished **and** $|\mathcal{S}$.messages$| < M$ **do**
  if $\mathcal{S}$.messages = [] **then**
    $\mathcal{S} \leftarrow$ TeacherNode($\mathcal{S}$) ▷ Provide first instruction
  else if $\mathcal{S}$.messages$[-1] \in$ Teacher **then**
    $\mathcal{S} \leftarrow$ LearnerNode($\mathcal{S}$) ▷ Generate learner response
    if Learner asks a question **and not** mention "next step" **then**
      $\mathcal{S}$.needs_clarification $\leftarrow$ True ▷ Set flag for Teacher
    else
      $\mathcal{S}$.current_step $\leftarrow \mathcal{S}$.current_step + 1 ▷ Next learning step
    end if
  else
    $\mathcal{S} \leftarrow$ TeacherNode($\mathcal{S}$) ▷ Provide teacher instruction
    if $\mathcal{S}$.needs_clarification **then**
      $\mathcal{S}$.needs_clarification $\leftarrow$ False ▷ Reset flag after clarification
    end if
  end if
  if $\mathcal{S}$.current_step $\geq |\mathcal{T}|$ **then**
    $\mathcal{S}$.finished $\leftarrow$ True ▷ End conversation if tutorial is complete
  end if
end while
Evaluate conversation:
$\mathcal{E} \leftarrow$ Evaluator($\mathcal{S}$.messages)
return $\mathcal{S}$.messages, $\mathcal{E}$

2 Multi-LLM Agent Procedural Learning

As shown in Figure 1, we propose a multi-LLM agent workflow designed for procedural learning through interactive dialogue. Algorithm 1 outlines the implementation of the workflow, comprising a teacher agent, a learner agent, and an evaluator agent, to dynamically simulate guided tutorial sessions using a conversational graph and assess pedagogic quality. First, we initialize the learning state and conversation graph, where each node represents a conversational turn, and edges represent interactions between the teacher and learner agents. Then, we define the nodes, invoking either the teacher agent or the learner agent: The *teacher agent* is responsible for providing instructions and answering clarification questions. The *learner agent* simulates a learner's understanding and generates follow-up questions or responses. Next, we simulate the conversation by iterating through the nodes in the graph: (1) It begins with the teacher agent providing instructions based on the first-step's tutorial content. (2) The interaction manager monitors the conversation state, tracks tutorial progress, handles clarification needs, and determines the next node in the graph. (3) It continues iteratively, with the teacher agent providing instructions and the learner agent generating responses until the tutorial is completed or a maximum interaction limit is reached. Last, the *evaluator* assesses the generated conversation using a range of evaluation metrics, including computational metrics and rubric-based metrics facilitated by an LLM-based evaluator agent.

3 Evaluation Protocol

3.1 Evaluation Metrics

3.1.1 Computational Metrics

We utilize a set of automatic metrics, including: (1) **Question**: Computes the proportion of learner utterances that contain questions. (2) **Completion**: Computes the proportion of generated conversation reached completion by generating the special token FINISHED in the end. (3) **Diversity**: Computes a 2-gram diversity score for linguistic variation in responses. (4) **BLEU** (Papineni et al., 2002): Measures the overlap of 4-grams between the generated and reference tutorial, focusing on precision. (5) **METEOR** (Banerjee and Lavie, 2005): Evaluates 4-gram similarity using a harmonic mean of precision and recall, along with alignment of words, including synonyms and stemming. (6) **ROUGE** (Lin, 2004): Assesses the quality of generated text by comparing overlapping 4-grams, word sequences, and word pairs with reference tutorial, focusing on recall. (7) **BERTScore** (Reimers and Gurevych, 2019): Measures context-aware semantic similarity between the generated and reference text with BERT embeddings.

3.1.2 Rubric Metrics

We define educational conversation metrics to evaluate tutorial-guided procedural learning and pedagogic quality for both LLM and human judges: (1) **Clarity**: Measures how clear, well-structured, and understandable the teacher's instructions are. (2) **Truthfulness**: Assesses the factual accuracy and alignment of responses with the tutorial content. (3) **Engagement**: Evaluates the learner's active participation through meaningful questions and thoughtful responses. (4) **Coherence**: Checks the logical flow and smooth transitions in the conversation. (5) **Depth**: Gauges the level of detail and exploration in the discussion beyond surface-level content. (6) **Relevance**: Ensures responses stay on-topic and align with the tutorial's instructions. (7) **Progress**: Tracks how effectively the conversation moves forward through the tutorial steps. And (8) **Naturalness**: Measures how fluid and human-like the conversation feels, avoiding robotic or scripted responses. See full rubric in Appendix D.

3.2 Human Judgment Alignment

First, we select 25 tutorials based on the top average LLM-judged scores across all metrics, resulting in a set of 175 model-generated conversations. Second, we collect human annotations of ratings (scale 1-5, where 1 is very poor and 5 is excellent) on the sample set. We hire 2 human judges who are master's and PhD students proficient in English and have sufficient learning capabilities. For each data sample, we have a generated conversation paired with the grounded tutorial. We ask each judge to read the tutorial carefully and rate a generated conversation using the tutorial as a reference. The judges are instructed to evaluate the conversation based on the rubric metrics (§ 3.1.2). Then, we compute the average of the two judges' scores for each metric, using the average value as the reference to evaluate alignment with the automatic metrics. Last, we conduct a qualitative analysis by visualizing the distribution of ratings provided by human and LLM judges, and assess their alignment using the following correlation coefficients: (1) **Pearson's** r: Evaluates linear correlation between continuous variables. (2) **Spearman's** ρ: Measures monotonic relationships between ranked variables. And (3) **Kendall's** τ: Assesses ordinal associations between ranked variables.

4 Dataset

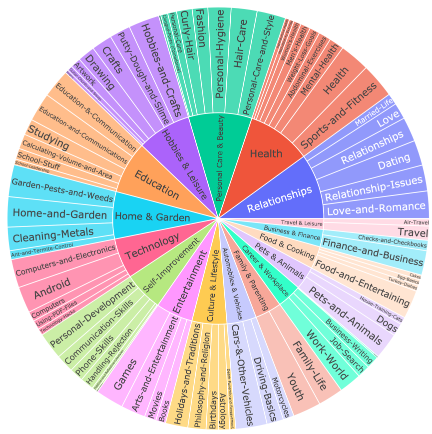

Figure 2: Hierarchical distribution of topics across 17 domains with corresponding topic counts. The inner ring represents the domains, while the outer segments depict 81 first-level topics within each domain. Segment sizes are proportional to the count of entries per topic.

We collected 3,680 real-world instructional articles from WikiHow[2] covering 17 domains and 727 topics, as shown in Figure 2. Each article has 1516.7 tokens and 4.0 methods as how-to tutorials on average. Each method contains 304.3 tokens and 5.2 instructional steps on average. The data collection contains 14,738 methods and 49,344 learning steps in total. We treat each method as a tutorial and generate 14,287 teacher-learner conversations, skipping tutorials with only one step, for 8 backbone LLMs, resulting in a total of 114,296 conversations. Additionally, we count the number of articles for each domain (Appendix C) and organize topics into a hierarchical knowledge graph (Appendix E).

[2] https://www.wikihow.com/

LLMs	**Provider**	**BERTScore (%)**	**BLEU (%)**	**METEOR (%)**	**ROUGE (%)**	**Diversity (%)**	**Question (%)**	**Completion (%)**
DeepSeek	DeekSeek	71.81	31.54	30.53	37.71	77.24	17.67	99.57
Qwen2	Alibaba	73.29	37.48	32.06	41.96	83.44	20.51	99.49
Gemma	Google	66.85	28.92	22.16	36.13	85.39	27.99	97.17
OLMo2	Allenai	71.45	25.42	32.22	34.25	79.80	22.75	100.00
OpenChat	Tsinghua	71.06	29.33	33.66	37.62	73.31	62.36	79.46
Llama3	Meta	74.74	35.27	39.33	44.85	75.66	61.60	86.39
Phi4	Microsoft	64.46	28.42	34.90	37.44	79.33	57.27	99.43
GPT-4	Open AI	69.66	43.86	34.48	48.58	86.96	60.82	99.79
Overall	$\mu \pm \delta$	70.42±3.36	32.53±6.00	32.42±4.91	39.82±3.62	80.14±4.79	38.59±20.70	95.16±7.82

Table 1: Overall workflow effectiveness in homogeneous learning (same LLM for teacher, learner, and evaluator), benchmarking 8 prevailing LLMs as backbone, evaluated on 7 computational evaluation metrics. Symbols μ and δ denote the mean and standard deviation of the evaluation metrics, respectively.

5 Results

We conduct experiments with 8 LLMs from diverse providers: The open-source models including DeepSeek (7B), Qwen2 (7B), Gemma (7B), OLMo2 (7B), OpenChat (7B), Llama3 (8B), Phi4 (14B), and commerical GPT-4 (1.76TB).

5.1 Evaluation Results of the Workflow (RQ1)

5.1.1 Homogeneous Learning

In this setting, all three agents (teacher, learner, and evaluator) share the same model. As shown in Table 1, we present the evaluation results of the proposed multi-LLM agent workflow with 8 backbone LLMs in homogeneous learning, evaluated by computational metrics. Overall, the workflow shows strong effectiveness in homogeneous learning setups, with high diversity and completion rates. Specifically, we find that:

First, the proposed workflow achieves consistently outperforming mean scores in **Completion** ($\mu = 95.16\%$), **Diversity** ($\mu = 80.14\%$), and semantic similarity (**BERTScore** $\mu = 70.42\%$). This indicates the workflow's ability to generate varied and complete conversations that are semantically aligned with the reference tutorial. This is attributed to the instruction-following capabilities of LLMs, which enable the workflow to complete conversations with diverse yet coherent responses, while the teacher agent effectively mirrors the tutorial's instructional content.

Second, simulating learner engagement for educational conversation is still challenging. OpenChat and Llama3 excel in **Question** generation ($\mu = 62.36\%, 61.60\%$), showcasing their capability to simulate active learner engagement, but the workflow demonstrates variability in the learner agent's performance across scenarios, as indicated by the high standard deviation in the **Question** ($\delta = 20.70\%$). This variability may stem from the diverse nature of the tutorials, which can influence the learner agent's engagement and question generation. Besides, the task to understand the teacher's instructions and generate questions is challenging, leading to variant performance on different LLMs.

Third, the workflow demonstrates variability in alignment with the reference tutorial, as reflected by the high standard deviations in **BLEU**, **METEOR**, **ROUGE**, and **BERTScore** ($\delta = 6.00, 4.91, 3.62, 3.36$) metrics. This suggests that the performance is sensitive to the specific content and structure of the tutorial, as these metrics heavily rely on the overlap and semantic similarity between the generated conversation and the reference tutorial. The observed deviations highlight the challenges in maintaining consistent quality across diverse tutorial topics and instructional styles.

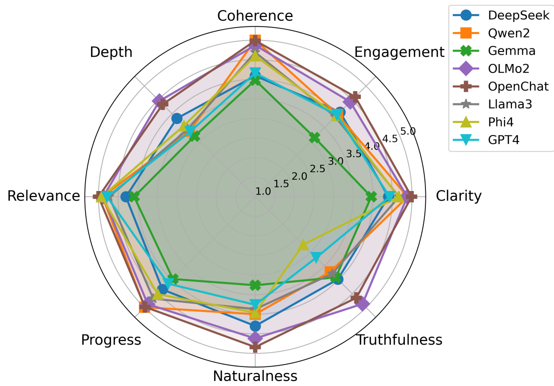

Figure 3: Evaluation results of homogeneous learning using rubric-based metrics across different LLMs.

As shown in Figure 3, we visualize the evaluation results of the proposed multi-LLM agent workflow with 8 backbone LLMs in homogeneous learning, evaluated using rubric-based metrics. We obtain the following key insights: (1) The workflow has a strong ability to stay on-topic (**Relevance**

Learner	BERTScore (%)	BLEU (%)	METEOR (%)	ROUGE (%)	Diversity (%)	Question (%)	Completion (%)
OpenChat	71.06	29.33	33.66	37.62	73.31	62.36	79.46
DeepSeek	68.08 ↓	33.73 ↑	30.49 ↓	39.06 ↑	77.79 ↑	26.87 ↓	99.30 ↑
Qwen2	71.85 ↑	37.08 ↑	32.18 ↓	42.96 ↑	80.90 ↑	26.21 ↓	98.27 ↑
Gemma	68.99 ↓	36.75 ↑	31.19 ↓	42.39 ↑	84.50 ↑	23.59 ↓	98.99 ↑
OLMo2	68.22 ↓	35.44 ↑	30.61 ↓	40.86 ↑	80.81 ↑	42.11 ↓	99.32 ↑
Homogeneous Overall ($\mu \pm \delta$)	$70.42_{\pm 3.36}$	$32.53_{\pm 6.00}$	$32.42_{\pm 4.91}$	$39.82_{\pm 3.62}$	$80.14_{\pm 4.79}$	$38.59_{\pm 20.70}$	$95.16_{\pm 7.82}$
Heterogeneous Overall ($\mu \pm \delta$)	69.29 ↓ $_{\pm 1.76}$	35.75↑$_{\pm 1.52}$	31.12 ↓ $_{\pm 0.77}$	41.32↑$_{\pm 1.75}$	81.00↑$_{\pm 2.75}$	29.70↓$_{\pm 8.40}$	98.97↑$_{\pm 0.49}$

Table 2: Overall workflow effectiveness in heterogeneous learning (diverse LLMs as learners with the same LLM as the teacher and evaluator). Symbols μ and δ denote the mean and standard deviation of the evaluation metrics, respectively. Symbols ↑ and ↓ indicate an increase or decrease compared to the baseline setting, respectively.

$\mu = 4.70$) and provide clear instructions (**Clarity** $\mu = 4.61$). (2) Some models struggle to maintain conversational fluidity and explore concepts in depth, aligning with the reference tutorial. Specifically, **Truthfulness**, **Naturalness**, and **Depth** metrics show high variability ($\delta = 0.7, 0.5, 0.47$). These findings highlight the need for further improvements in generating conversations that are human-like, detailed, and factually accurate while closely adhering to the tutorial content. (3) OpenChat and OLMo2 are the most well-rounded performers, excelling in the majority of the metrics. Hence, in the heterogeneous learning setups, we equip OpenChat with the teacher and evaluator agents and the other LLMs for diverse learners.

5.1.2 Heterogeneous Learning

In this setting, the learner agent differs from the teacher and evaluator agents. OpenChat serves as the teacher and evaluator, while the learner is selected from 7B open-source LLMs (DeepSeek, Qwen2, Gemma, OLMo2). In Table 2, we present the computational evaluation results and observe:

First, the workflow demonstrates improved linguistic **Diversity** ($+0.86\%$) and alignment metrics, (**BLEU** $+3.23\%$, **ROUGE** $+1.50\%$), while maintaining high **Completion** rates ($\mu = 98.97\%$). This highlights the benefits of diverse learner agents in generating varied, complete, and well-aligned conversations. Additionally, reduced standard deviations across metrics indicate more consistent performance in heterogeneous setups.

Second, learner engagement in heterogeneous setups significantly drops compared to homogeneous setups, as indicated by the **Question** metric (-8.89%). This suggests that diverse learners may struggle to generate meaningful questions consistently. We attribute this to inherent differences in the models' design artifacts (e.g., whether reasoning capabilities are enabled) and data (e.g., domain knowledge and linguistic diversity). The high standard deviation in the **Question** metric ($\delta = 20.70\%$) further highlights variability in learner engagement across different models. Addressing this gap requires incorporating adaptive dialogue strategies and human-in-the-loop evaluations to better simulate realistic learning scenarios and improve alignment between model-generated interactions and actual learner needs.

Third, the workflow's performance in semantic similarity (**BERTScore** -1.13% and **METEOR** -1.30%) slightly decreases compared to the homogeneous setting, This indicates challenges in maintaining semantic alignment and precision-recall balance with the reference tutorial.

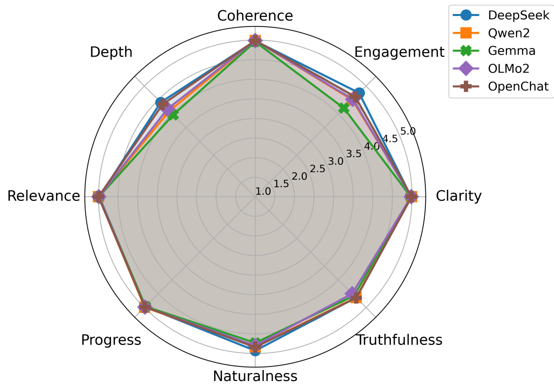

Figure 4: Evaluation results of heterogeneous learning using rubric-based metrics across different LLMs.

Last, comparing Figure 4 with Figure 3, we intuitively observe that the performance in heterogeneous learning is more consistent across different models. However, noticeable variations persist in **Depth** and **Engagement** metrics.

5.2 Performance across Domains (RQ2)

We evaluate the performance of the proposed multi-LLM agent workflow across 17 domains, as shown in Figure 5. We present the evaluation results of the

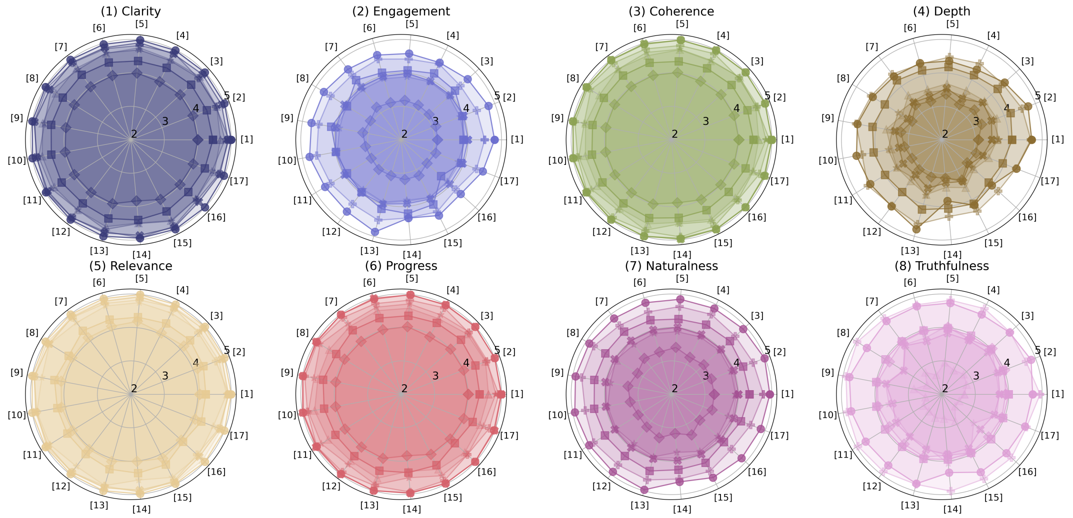

Figure 5: Multiple rubric evaluation scores of all LLMs over various domains (marked by their IDs in Appendix C).

proposed multi-LLM agent workflow with 8 backbone LLMs in heterogeneous learning, evaluated by computational metrics. Overall, the workflow shows strong effectiveness in heterogeneous learning setups, with high diversity and completion rates. Specifically, we observe the following key findings:

On one hand, the proposed workflow achieves consistently well-rounded performance for **Clarity**, **Coherence**, **Relevance** and **Progress** for all models across all domains. This indicates the workflow's ability to generate varied and complete conversations that are semantically aligned with the reference tutorial. This is attributed to the instruction-following capabilities of LLMs, which enable the workflow to complete conversations with diverse yet coherent responses, while the teacher agent effectively mirrors the tutorial's instructional content.

On the other hand, **Engagement**, **Truthfulness**, **Depth**, and **Naturalness** exhibit significant variability across domains and different LLMs. This variability likely arises from the diverse complexity and structure of the tutorials, which can affect the learner agent's ability to generate meaningful questions and responses. Moreover, the inherent challenge of interpreting the teacher's instructions and formulating appropriate follow-up questions further contributes to the observed performance disparities among different LLMs.

5.3 Human Evaluation Alignment (RQ3)

We measure the alignment between automatic evaluation metrics (§ 3) and human annotations.

First, we visualize the distribution of ratings provided by human and LLM judges for comparative analysis, as shown in Figure 6. We observe that the human judges tend to provide lower and diverse ratings than the LLM judges, indicating a potential bias in the LLM evaluation process.

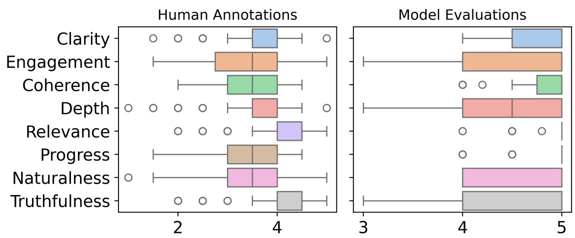

Figure 6: Comparison between human annotation and model evaluation scores.

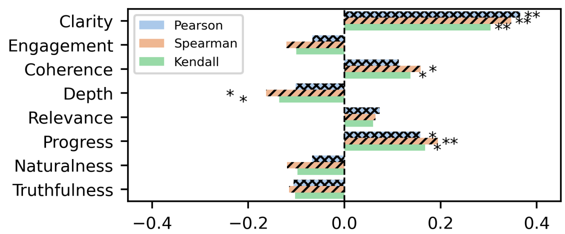

Figure 7: Human alignment correlation coefficient Pearson's r, Spearman's ρ, Kendall's τ, over diverse metrics. Symbols denote significance test: * p<0.05, ** p<0.01.

Second, we compute the correlation coefficients between human and LLM judges across all metrics, as shown in Figure 7. Metrics such as **Clarity**, **Progress**, **Coherence**, and **Relevance** exhibit higher correlations, indicating stronger alignment between LLM and human judgments in these ar-

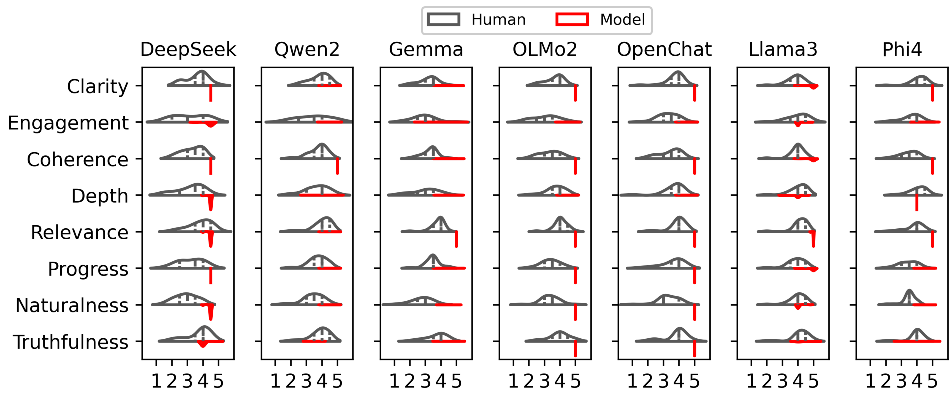

Figure 8: Distribution of model evaluations compared to human annotations across all metrics.

eas. This suggests that these metrics effectively capture aspects of conversational quality that resonate with human evaluators. In contrast, metrics like **Depth**, **Engagement**, **Naturalness**, and **Truthfulness** show lower correlations, highlighting potential gaps in how LLM judges assess these dimensions compared to human evaluators. This misalignment may stem from the inherent difficulty of evaluating nuanced aspects such as conversational depth and authenticity, suggesting the need for improved evaluation strategies or additional calibration of LLM-based metrics.

Third, we visualize the distribution of model evaluations compared to human annotations across all metrics, as shown in Figure 8. We observe that LLM evaluations are tightly clustered around a high median with low variance, whereas human scores display a wider spread and a lower median. This suggests that LLMs tend to rate fluency and related qualities higher, possibly due to an overreliance on surface-level text coherence. In contrast, for metrics such as **Engagement** and **Depth**, both most LLMs and human ratings exhibit greater variability, with most LLM ratings often distributed more uniformly across the scale. This discrepancy likely arises from fundamental differences in evaluation criteria: human judges may focus more on the quality and depth of the conversation, while LLM judges prioritize fluency and coherence. This misalignment highlights the need for improved evaluation strategies or additional calibration of LLM-based metrics to better reflect human judgment.

6 Discussion

Implication. This work introduces a modular, multi-agent workflow for procedural learning that contributes to the development of scalable, LLM-based educational systems. The proposed large-scale dataset, grounded in structured tutorials and spanning domains, offers a valuable resource for both NLP and AI4Education communities. Combined with a multi-perspective evaluation protocol (automatic metrics, rubric scoring, and human assessment), the workflow enables comprehensive assessment of LLMs in educational contexts.

From the perspectives of educational AI systems, This approach supports the development of scalable, automated tutoring systems adaptable to various topics, instructional goals, and learner profiles. By simulating interactive teacher-learner dialogues, it moves beyond static benchmarks and enables dynamic, pedagogically grounded evaluation of LLMs. This opens new directions for AI-powered curriculum design, feedback systems, and intelligent tutoring. From a broader societal perspective, such systems can improve access to high-quality instruction, particularly in resource-limited settings. Meanwhile, their growing role also highlights the need for responsible design, transparent evaluation, and human oversight to ensure they complement rather than replace effective teaching.

Positioning Future Work. Looking ahead, several directions can further advance this work. First, incorporating real human learners and evaluators is essential to capture authentic learning dynamics and validate automated evaluation methods. Second, future research can focus on explicitly modeling pedagogical skills within LLMs, such as scaffolding, adaptive explanation, and constructive feedback, to more accurately reflect teaching competence and align LLMs with pedagogical skills. Lastly, enhancing personalization through learner modeling and adaptive dialogue strategies will improve responsiveness and realism, bringing the system closer to real-world deployment.

7 Related Work

7.1 Single LLM-Agent for Education

LLMs have been employed as single agents in education, serving as teachers, learners, or evaluators to enhance learning. Teacher agents deliver knowledge, answer questions, and guide learners through structured content. For instance, ChatTutor (Chen et al., 2024b) uses LLMs for course planning, adaptive quizzes, and tailored instruction via interaction, reflection, and reaction. Similarly, Chen et al. (2024a) show that ChatGPT as a teachable agent improves programming education by enhancing students' understanding and coding skills. For instance, Ma et al. (2024b) fine-tuned LLMs to simulate diverse language learners, aiding pre-service teachers in adapting to varied student needs. Evaluator agents assess educational interactions and provide feedback. Hu et al. (2025) use LLMs to simulate teacher-student interactions, generate teaching reflections, and refine lesson plans, showcasing their potential to enhance pedagogy. These roles demonstrate the versatility of LLMs in addressing educational needs, forming a foundation for multi-agent workflows.

7.2 Multi-LLM Agent Workflows

Recent advancements in multi-LLMs workflows leverage agent collaboration to enhance scalability and adaptability in educational systems (Chu et al., 2025). For example, EduAgent (Xu et al., 2024) integrates cognitive priors to guide reasoning in simulated interactions, while Park et al. (2024) align diagnostic student modeling with LLM-based tutoring through prompt engineering. GenMentor (Xu et al., 2025b) dynamically tailors content to learner needs by optimizing learning paths, and LLMAgent-CK (Yang et al., 2024b) employs structured multi-agent roles for content knowledge identification. Xu et al. (2025a) introduce a framework with reflective and novice agents to simulate realistic student behaviors, enhancing fidelity and learning outcome assessments. Additionally, EvaAI (Lagakis and Demetriadis, 2024) introduces a multi-agent framework leveraging LLMs for enhanced automated grading, further demonstrating the utility of LLMs in educational evaluation. Pei et al. (2024) propose a multi-agent framework with visual and language agents to assist multimodal LEGO assembly learning. These approaches highlight the potential of multi-agent LLM in advancing personalized and effective educational workflows.

Compared to existing works, our approach emphasizes the integration of multiple LLMs to simulate procedural learning interactions. This enables a more comprehensive evaluation of pedagogical quality, instructional effectiveness, and conversational dynamics, advancing the development of scalable and adaptive educational systems.

7.3 Educational Applications and Domains

Extensive work has been conducted across diverse contexts, including language learning, STEM education, and professional development. In language learning, studies have demonstrated the effectiveness of LLMs in enhancing student engagement through conversational interfaces (Ma et al., 2024b; Lagakis and Demetriadis, 2024). In STEM education, LLMs have been applied to facilitate learning in mathematics (Gou et al., 2024), physics (Wang et al., 2023; Ma et al., 2024a), chemistry (M. Bran et al., 2024; Sprueill et al., 2024), biology (Huang et al., 2024; Ghafarollahi and Buehler, 2024), and general scientific discovery (Ma et al., 2024c; Baek et al., 2024). Additionally, LLMs have shown promise in professional development, including applications in medical training (Yang et al., 2024c; Abd-Alrazaq et al., 2023), computer science education (Zhang et al., 2024; Yang et al., 2024a; Islam et al., 2024), and legal studies (Guha et al., 2023). Compared to these domain-specific applications, our work leverages large-scale learning materials spanning a wide range of domains, enabling a more comprehensive exploration of procedural learning and pedagogic quality assessment.

8 Conclusion

In this work, we introduced a multi-LLM agent workflow for simulating procedural learning and assessing pedagogic quality at scale. Our framework leverages teacher, learner, and evaluator agents to enable dynamic teaching-learning simulations grounded in diverse instructional content. We presented a large-scale dataset and a comprehensive evaluation protocol, demonstrating the workflow's effectiveness across various setups. We also identify areas for improvement, such as enhancing learner engagement and aligning automated evaluations with human judgments. This work lays the foundation for scalable, adaptive educational systems, bridging NLP with AI4Education.

Limitations

While our study presents a large-scale, structured framework for simulating procedural teaching-learning interactions, there remain two areas with potential for further enhancement. First, the learner agents are simulated using LLMs within a predefined interaction flow. This setup enables controlled and consistent experimentation but may not fully reflect the variability and learning behavior of real human learners. Future work could incorporate adaptive dialogue policies or human-in-the-loop learners to improve the realism and responsiveness of interactions. Second, although our dataset spans 17 domains and 727 topics, variation in topic complexity and data distribution could influence performance comparisons. Certain domains may be overrepresented or inherently more challenging, which could affect model generalizability. Future extensions can include domain-balanced evaluation and fine-grained analysis at the topic level to better understand model behavior across different knowledge structures. These considerations do not compromise the core contributions of our work but highlight opportunities for expanding its applicability and pedagogical fidelity.

References

Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1):e48291.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Angxuan Chen, Yuang Wei, Huixiao Le, and Yan Zhang. 2024a. Learning-by-teaching with chatgpt: The effect of teachable chatgpt agent on programming education. *arXiv preprint arXiv:2412.15226*.

Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024b. Empowering private tutoring by chaining large language models. In *CIKM*, pages 354–364.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and Qingsong Wen. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Alireza Ghafarollahi and Markus J Buehler. 2024. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *ICLR*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. volume 36, pages 44123–44279.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

Huazhen Huang, Xianguo Shi, Hongyang Lei, Fan Hu, and Yunpeng Cai. 2024. Protchat: An ai multi-agent for automated protein analysis leveraging gpt-4 and protein language model. *Journal of Chemical Information and Modeling*, 65(1):62–70.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapcoder: Multi-agent code generation for competitive problem solving. In *ACL*, pages 4912–4944.

Shi Jinxin, Zhao Jiabao, Wang Yilei, Wu Xingjiao, Li Jiawen, and He Liang. 2023. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503*.

Paraskevas Lagakis and Stavros Demetriadis. 2024. Evaai: a multi-agent framework leveraging large language models for enhanced automated grading. In *International Conference on Intelligent Tutoring Systems*, pages 378–385. Springer.

Yu-Ju Lan and Nian-Shing Chen. 2024. Teachers' agency in the era of llm and generative ai. *Educational Technology & Society*, 27(1):I–XVIII.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024a. Llm and simulation as bilevel optimizers: a new paradigm to advance physical scientific discovery. In *ICML*, pages 33940–33962.

Yiping Ma, Shiyu Hu, Xuchen Li, Yipei Wang, Shiqing Liu, and Kang Hao Cheong. 2024b. Students rather than experts: A new ai for education pipeline to model more human-like and personalised early adolescences. *arXiv preprint arXiv:2410.15701*.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. 2024c. Sciagent: Tool-augmented language models for scientific reasoning. In *EMNLP*, pages 15701–15736.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *CHI*, pages 1–10.

Jiahuan Pei, Irene Viola, Haochen Huang, Junxiao Wang, Moonisa Ahsan, Fanghua Ye, Yiming Jiang, Yao Sai, Di Wang, Zhumin Chen, Pengjie Ren, and Pablo Cesar. 2024. Autonomous workflow for multimodal fine-grained training assistants towards mixed reality. In *Findings of ACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. In *ICML*, pages 46351–46374. PMLR.

Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. Newton: Are large language models capable of physical reasoning? In *Findings of EMNLP*, pages 9743–9758.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025a. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. *arXiv preprint arXiv:2502.02780*.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2025b. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. *arXiv preprint arXiv:2501.15749*.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. In *NeurIPS*, volume 37, pages 50528–50652.

Kaiqi Yang, Yucheng Chu, Taylor Darwin, Ahreum Han, Hang Li, Hongzhi Wen, Yasemin Copur-Gencturk, Jiliang Tang, and Hui Liu. 2024b. Content knowledge identification with multi-agent large language models. In *International Conference on Artificial Intelligence in Education*, pages 284–292. Springer.

Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024c. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. volume 8, pages 1–35. ACM New York, NY, USA.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In *ACL*, pages 13643–13658.

A Implementation Details

The implementation of the workflow is based on LangGraph. [3] We utilize 7 the most capable openly available LLMs hosted on Ollama [4] as the backbone of the proposed workflow: (1) DeepSeek: A 7B parameter, bilingual-supportive, open-source model from DeepSeek. (2) Qwen2: A 7B parameter, multilingual-supportive, open-source model from Alibaba. (3) Gemma: A 7B parameter, lightweight, open-source model developed by Google DeepMind. (4) OLMo2: A 7B parameter, open-source model from AllenAI, demonstrating competitive performance with equivalently sized open models. (5) OpenChat: A 7B parameter, open–source model from Tsinghua, surpassing ChatGPT 3.5 on various benchmarks. (6) Llama3: An 8B parameter, highly capable openly available model from Meta. (7) Phi4: A 14B parameter, open–source model from Microsoft. It is possible to seamlessly switch to other backbone models as needed. We use GPT-4, the closed-source model with 1.76T parameters from OpenAI, by calling the API of the system (Hurst et al., 2024). We choose the sentence transformer model all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) as the embedding model for semantic similarity search.

We set the maximum number of tokens for each model to 4096, and the maximum number of interactions to 40. We set the temperature to 0.0 to ensure deterministic outputs. The default batch size for model inference is set to 16. All experiments are conducted on NVIDIA H100 (94GB) and NVIDIA A100 SXM4 (40GB) GPUs. Additionally, our workflow is compatible with deployment on CPUs.

B Prompt Templates

As shown in Table 3, we provide the prompt templates for each step of the workflow. The prompt templates are designed to be flexible and can be easily adapted to different tasks or domains. The templates include placeholders for input data, which can be replaced with specific examples or parameters as needed. The templates also include instructions for the model, guiding it on how to process the input data and generate the desired output. The templates are structured to ensure that the model understands the context and requirements of each task, facilitating effective communication between the user and the model.

Agent	Prompt Template
Teacher	You are an expert teacher guiding a learner step by step through a tutorial via multi-turn conversations. **Your role**: (1) Answer learner questions related to the instructions as the first priority. (2) If the learner requests to move on without asking questions, provide the current step's instruction directly. (3) Once the final step has been completed and the learner has no further questions, acknowledge completion by appending the token 'FINISHED' to your response. (4) Highlight key phrases in your response if they appear in the tutorial. Respond in character as a teacher — no system-level messages or meta-comments. Keep responses short and focused. **Tutorial Summary**: {summary} **Step** {current_step_index} {current_step_content} **Learner**: {user_utterance}; **Teacher**:
Learner	You are a curious student engaged in a step-by-step learning conversation with a teacher. **Your role**: (1) Carefully read and understand the teacher's instructions. (2) Respond naturally, politely, and concisely. (3) If the instruction is unclear, ask a brief and specific question to clarify it. (4) If the instruction is clear, acknowledge it briefly and politely ask to move on to the next step. (5) If the teacher opens with a BEGIN message, just acknowledge it and do not ask questions. (6) If the teacher says 'FINISHED' or indicates that the tutorial is over, thank them politely and say nothing more. Respond in character as a student — no system-level messages or meta-comments. Keep responses short and focused. **Teacher**: {instruction}; **Learner**:
Evaluator	Evaluate the following teacher-learner conversation: {conversation} [**Reference**: {tutorial}]; **Score** from 1 to 5: {criterion_name: criterion_rubric} Provide scores in this format: {criterion_name: rubric_score}

Table 3: Prompt templates used in this work.

C Number of Articles across Domains

As shown in Table 4, we provide the number of articles across different domains. The table includes the total number of articles, the number of unique topics, and the average number of articles per topic for each domain. The data is collected from the WikiHow dataset, which contains a diverse range of articles covering various topics. The domains include health, technology, lifestyle, and more. The number of articles in each domain varies significantly, with some domains having a larger number of articles than others. The table provides a comprehensive overview of the distribution of articles across different domains, highlighting the diversity and breadth of topics covered in the WikiHow dataset.

D Detailed Rubrics for Evaluation

As shown in Table 5, we provide the detailed rubrics for evaluation. The rubrics are designed

[3] https://langchain-ai.github.io/langgraph/
[4] https://ollama.com/search

ID	Domain	#Article	ID	Domain	#Article	ID	Domain	#Article
1	Automobiles & Vehicles	166	7	Family & Parenting	150	13	Pets & Animals	131
2	Business & Finance	98	8	Food & Cooking	127	14	Relationships	368
3	Career & Workplace	139	9	Health	363	15	Self-Improvement	222
4	Culture & Lifestyle	197	10	Hobbies & Leisure	306	16	Technology	227
5	Education	296	11	Home & Garden	266	17	Travel & Leisure	61
6	Entertainment	209	12	Personal Care & Beauty	358			

Table 4: Number of WikiHow articles across 17 domains.

Criterion	**Rubric** (5 – Excellent, 4 – Good, 3 –Average, 2 – Poor, 1 – Very Poor)
Clarity†	Is the teacher's instruction clear, well-structured, and easy to understand? 5 – The instruction is highly clear, logically structured, and easy to understand. No ambiguity, and key concepts are well explained. 4 – The instruction is mostly clear, with minor ambiguities or areas that could be refined. The overall structure is logical. 3 – The instruction is somewhat clear, but there are noticeable ambiguities or structural issues that may cause some confusion. 2 – The instruction is difficult to follow due to unclear phrasing, missing details, or a lack of logical flow. Learners may struggle to understand. 1 – The instruction is highly unclear, disorganized, or misleading. Learners are likely to be confused or misunderstand key points.
Truthfulness†	Does the generated response stay within the scope of the reference tutorial? 5 – All information is completely accurate, directly aligned with the given tutorial or trusted sources. No hallucinations or misleading details. 4 – The response is largely correct but may contain small factual errors, slightly ambiguous statements, or minor misinterpretations. 3 – Includes some correct information but also contains notable factual errors, misrepresentations, or unsupported claims. 2 – A majority of the response is incorrect or misleading, with only a small portion being factually accurate. 1 – The response is entirely false, misleading, or contradicts known facts or the provided tutorial.
Engagement‡	Does the learner actively participate by asking meaningful questions or responding thoughtfully? 5 – The learner frequently asks insightful questions, provides thoughtful responses, and demonstrates strong engagement. 4 – The learner engages well, asking relevant questions and responding appropriately, with only minor missed opportunities. 3 – The learner participates but does not consistently ask meaningful questions or provide detailed responses. 2 – The learner shows minimal engagement, rarely asking questions or contributing meaningfully to the discussion. 1 – The learner is disengaged, does not ask questions, and provides minimal or no meaningful responses.
Coherence*	Is the conversation logically structured, with smooth transitions between steps? 5 – The conversation is well-structured, with smooth transitions and a logical flow from one step to the next. 4 – The conversation is mostly coherent, with only minor inconsistencies or abrupt transitions. 3 – The conversation has some structural issues, with occasional jumps or awkward transitions. 2 – The conversation lacks logical flow, with multiple abrupt or confusing transitions. 1 – The conversation is highly disorganized, making it difficult to follow the sequence of steps.
Depth*	Does the conversation go beyond surface-level discussion and explore concepts in sufficient detail? 5 – The conversation explores concepts in-depth, with thorough explanations and critical thinking. 4 – The conversation provides mostly in-depth discussion, but some areas could be explored further. 3 – The conversation touches on key concepts but lacks depth in explanations or follow-up discussions. 2 – The conversation is mostly superficial, with little exploration beyond basic information. 1 – The conversation is highly shallow, lacking any meaningful discussion or explanation.
Relevance*	Do the responses stay on-topic and align with the tutorial's instructions and context? 5 – The responses are consistently relevant and closely follow the tutorial's instructions and context. 4 – The responses are mostly relevant, with only occasional minor deviations. 3 – The responses are somewhat relevant but occasionally stray off-topic. 2 – The conversation frequently goes off-topic, with limited alignment to the tutorial's instructions. 1 – The responses are largely irrelevant, making it difficult to follow the intended learning path.
Progress*	Does the conversation effectively move forward through the tutorial steps in a logical manner? 5 – The conversation progresses efficiently, covering each step logically and without unnecessary repetition. 4 – The conversation generally moves forward well, with only minor delays or repetitions. 3 – The conversation progresses but sometimes gets stuck or moves inefficiently. 2 – The conversation struggles to move forward, often repeating steps or getting sidetracked. 1 – The conversation lacks clear progress, frequently revisiting previous steps or failing to complete the tutorial.
Naturalness*	Does the conversation feel fluid and human-like, avoiding robotic or overly scripted responses? 5 – The conversation feels natural and human-like, with engaging and varied responses. 4 – The conversation is mostly natural, but some responses feel slightly mechanical or repetitive. 3 – The conversation has a mix of natural and robotic responses, with some unnatural phrasing. 2 – The conversation often feels artificial or scripted, with little variation in responses. 1 – The conversation is highly robotic or formulaic, making it feel unnatural and disengaging.

Table 5: Rubric for LLMs and human annotators. Symbols †, ‡, and * represent the target evaluation role as teacher, learner, and conversation, respectively.

to assess the quality and effectiveness of the generated articles, used for both LLM and human judges. Each criterion is defined with specific guidelines to ensure consistent and objective evaluation. To be more specific, we provide three types of rubrics tailored to the target evaluation roles: teacher, learner, and conversation. For the teacher role, the rubric emphasizes Clarity and Truthfulness, ensuring that the generated teaching material is accurate, tutorial-based, and easy to comprehend. For the learner role, the rubric emphasizes Engagement, encouraging learners to actively participate by asking meaningful questions and providing thoughtful responses. For the conversation role, the remaining rubrics assess the quality of the conversation in terms of the learning process (Progress), logical flow (Relevance and Coherence), and linguistic features (Naturalness).

E Knowledge Graph for WikiHow Articles

As shown in Figure 9, we provide the knowledge graph for WikiHow articles. The graph illustrates the hierarchical relationships among domains, topics, and articles within the dataset, offering a comprehensive visual representation of the knowledge structure. The root node serves as a dummy root category, with first-level nodes representing domains, intermediate nodes denoting hierarchical topics, and leaf nodes corresponding to individual articles. The edges represent the relationships between them.

The knowledge graph serves as a valuable resource for understanding the structure and organization of information within the WikiHow dataset. It brings a broader implication of the proposed dataset, which can be used for various applications, including information retrieval, recommendation systems, and knowledge discovery.

F Case Study

As shown in Table 6, we provide a case study of the generated conversations. The case study includes an example of educational conversation generated by the proposed workflow, grounded on a WikiHow tutorial spanning a hierarchical range of topics. Highlighted texts in the same color indicate relevant parts, as marked by human annotation, to enhance readability and facilitate a better understanding of the generated content. This demonstrates a nuanced understanding of diverse topics and the ability to generate coherent, contextually relevant content.

G Risks, Ethics, Data and AI Tool Usage

We recognize that the use of LLMs in educational contexts presents several risks, including the potential for misinformation, bias, and over-reliance on automated tools.

The proposed workflow is intended to assist users in generating educational content and supporting learning—not to replace human educators

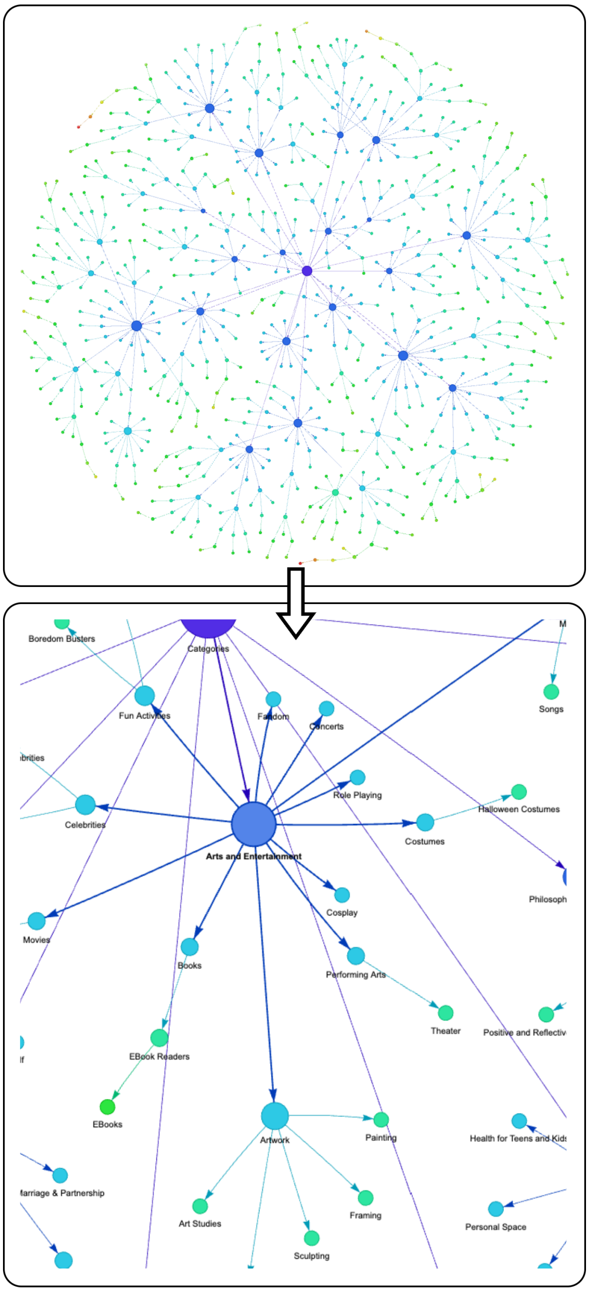

Figure 9: Visualization of the knowledge graph composed from WikiHow articles with hierarchical categories.

or learners. Rather, it aims to augment human capabilities and provide supplementary resources to enhance the educational experience.

The WikiHow articles used in this work are publicly available and licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License `https://creativecommons.org/licenses/by-nc-sa/3.0/`.

This dataset is used solely for research purposes, with no intention of commercial use. It does not contain personally identifying information or of-

Tutorial: Part 2 of 3:Writing the Body of the Speech
Categories: Education and Communications >>Personal Development >>Maturity >>Leadership >>School Leadership

Step 1: State your main ideas on how to improve the school.
You should have at least three ideas you feel would benefit your school and classmates. This gives your classmates the incentive to vote for you and shows you want the position as an opportunity to help others.
(1) You should list your ideas and then expand on them later in the body. It might take a bit of research to figure out what you want to change. Ask around the school, talking to students and teachers, and see where there's room for improvement. What are the concerns of the students? What are people happy with regarding the school? What would they like to see change? Asking these questions can help you get a sense of your audience and community.
(2) Remember, you should not make promises you cannot keep. Do not say anything just to get elected. While many students might want gum-chewing policies eliminated or for the lunch period to run twice as long, this is probably not necessary or possible. Try to focus on areas that seem important to keep your school running safely and efficiently. Concerns about things like bullying, academic standards, and extracurricular activities should be your concern over fun and games.
(3) A good opening statement for your body would state the causes important to you and what you plan to do about them. For example, if you were running for president, you could say something like, "I understand we need to improve how we handle bullying, increase interest in extracurricular activities, and expand access to AP courses throughout the school. As your president, I would work to bring in speakers to talk about sensitivity in the classroom, increase advertising for basketball games and quiz bowl tournaments, and start a tutoring program to help students struggling with certain subjects."
Step 2: Find support for those ideas.
You should do a little extra research in addition to talking to classmates and teachers. Have some specific plans on how you would enact change in your school.
(1) Using the school library or computers, figure out the best means to tackle certain problems many schools face. How have other schools dealt with bullying? Poor test scores? Low interest in extracurricular activities? What can you reasonably do as a student council member to address these problems?
(2) You do not have to have a point-by-point plan laid out, but a few sentences on some preliminary ideas can help you stand out from your peers. People are more inclined to vote for someone who's thought about how to solve problems in addition to identifying problems.
Step 3: Keep your ideas short but very strongly worded.
Your body should be about two paragraphs of 5 to 6 sentences each. This might sound brief, considering how much information you'll have to get across, but you have limited time and must keep people's attention. It can help to write more than you need and gradually cut down your speech to the bare essentials. This will help it to be shorter meaning that students won't get bored when listening to your speech.

Generated Conversation:
Teacher: Let's begin: Part 2 of 3:Writing the Body of the Speech. BEGIN
Learner: Okay, I'm ready to start learning about writing the body of the speech. Please guide me through this part.
Teacher: Great! For the body of your speech, you'll want to expand on the main ideas you've listed in **step 1**. Each idea should have its own paragraph. For example:
Bullying: Explain why addressing bullying is crucial and how it impacts students' well-being. Discuss specific strategies like bringing in sensitivity speakers or increasing awareness through workshops.
Extracurricular Activities: Highlight the importance of a diverse range of activities for student engagement and personal growth. Outline plans to boost participation, such as improving advertising methods or creating more opportunities for clubs and sports teams.
AP Courses Access: Address how access to Advanced Placement courses can benefit students' academic futures. Propose solutions like forming partnerships with local universities for guest lectures or establishing a tutoring program. Each paragraph should start with a clear topic sentence that directly relates to one of your main ideas, followed by supporting sentences that detail your plans and why they're beneficial. This will make your speech compelling and informative.
Learner: Thank you, teacher. I understand the instructions. Let's begin writing the body of my speech.
Bullying: Addressing bullying is crucial because it significantly impacts students' mental health and academic performance. By implementing sensitivity speakers who share personal experiences or insights on empathy and understanding, we can foster a more inclusive environment where students feel safe and supported. Additionally, organizing workshops that educate both students and staff about the effects of bullying will help in creating awareness and promoting a culture of respect.
Extracurricular Activities: A diverse range of activities is essential for student engagement and personal growth. To boost participation, we can improve advertising methods by utilizing social media platforms to reach out to students more effectively. We should also create more opportunities for clubs and sports teams by allocating additional funds specifically for these initiatives or seeking sponsorships from local businesses. This will not only increase student involvement but also enhance their leadership skills.
AP Courses Access: Access to Advanced Placement courses is vital as it prepares students for college-level work and future academic pursuits. By forming partnerships with local universities, we can invite guest lecturers who offer insights into advanced topics or provide guidance on college applications. Establishing a tutoring program will ensure that all students have the opportunity to excel in AP courses by receiving personalized assistance from peers or teachers. This approach not only enriches their educational experience but also boosts their confidence and academic readiness for higher education. I believe these paragraphs effectively address each main idea with clear topic sentences, supporting details, and explanations of why these strategies are beneficial.
Teacher: Great job! Your speech outline covers the key issues comprehensively. Make sure to include a personal touch by sharing your ideas on how you would implement these changes if elected. This will make your speech more relatable and engaging for your peers. Next, consider adding some statistics or examples from other schools that have successfully addressed these problems. This will strengthen your argument and show that you've done thorough research. Once you're ready to move on, let me know!
Learner: Thank you for the feedback! I'll incorporate my personal experiences and find relevant statistics. When I'm finished with this step, please let me know.
Teacher: Great job focusing on personal experiences and stats! For **Step 3**, concentrate on crafting a concise body that packs a punch. Aim for two paragraphs of about 5 to 6 sentences each. Keep your ideas strong and impactful. Remember, you have limited time, so it's crucial to be succinct and engaging. Write more than necessary initially, then refine your speech to its essentials. This will ensure your message is clear and keeps the audience's attention. Once you've completed this step, let me know!
Learner: Understood, Teacher. I'll focus on creating a concise body with impactful ideas for my speech. I plan to write two paragraphs of about 5 to 6 sentences each, ensuring they are strong and engaging. After drafting, I will refine the text to its essentials to maintain clarity and audience interest. Once done, I'll notify you. Thank you!
Teacher: We've reached the end of the tutorial. Bye! FINISHED
Learner: Thank you for the tutorial, I've learned a lot. Goodbye!

Table 6: An example of a generated conversation with the reference tutorial. Text highlighted in the same color indicates relevant parts, as marked by human annotation, and is intended to improve human readability.

fensive content.

We acknowledge the ethical challenges associated with deploying LLMs in education, particularly the risk of generating biased, inappropriate, or misleading content. To mitigate these concerns, we are committed to ensuring that the content generated through our workflow is accurate, reliable, respectful, and free from harmful bias.

In line with our ethical standards, we limited the use of AI tools (e.g., ChatGPT, Copilot, Grammarly) to surface-level tasks such as grammar checking and stylistic refinement, always based on our original writing. No AI tools were used to generate original content or ideas in this paper.