applied sciences

MDPI

Article

EduDCM: A Novel Framework for Automatic Educational Dialogue Classification Dataset Construction via Distant Supervision and Large Language Models

Changyong Qi [1], Longwei Zheng [2,3,*], Yuang Wei [1], Haoxin Xu [1], Peiji Chen [4] and Xiaoqing Gu [5]

[1] Shanghai Institute of AI for Education, East China Normal University, Shanghai 200062, China; changyongqi@stu.ecnu.edu.cn (C.Q.); philrain@foxmail.com (Y.W.); haoxin.xu@stu.ecnu.edu.cn (H.X.)
[2] School of Education, City University of Macau, Macau 999078, China
[3] State Key Laboratory of Cognitive Intelligence, Hefei 230088, China
[4] Department of Mechanical Engineering and Intelligent System, The University of Electro-Communications, Tokyo 183-8585, Japan; peijichen0324@gmail.com
[5] Department of Education Information Technology, East China Normal University, Shanghai 200062, China; xqgu@ses.ecnu.edu.cn
* Correspondence: lwzheng@cityu.edu.mo

Abstract: Educational dialogue classification is a critical task for analyzing classroom interactions and fostering effective teaching strategies. However, the scarcity of annotated data and the high cost of manual labeling pose significant challenges, especially in low-resource educational contexts. This article presents the EduDCM framework for the first time, offering an original approach to addressing these challenges. EduDCM innovatively integrates distant supervision with the capabilities of Large Language Models (LLMs) to automate the construction of high-quality educational dialogue classification datasets. EduDCM reduces the noise typically associated with distant supervision by leveraging LLMs for context-aware label generation and incorporating heuristic alignment techniques. To validate the framework, we constructed the EduTalk dataset, encompassing diverse classroom dialogues labeled with pedagogical categories. Extensive experiments on EduTalk and publicly available datasets, combined with expert evaluations, confirm the superior quality of EduDCM-generated datasets. Models trained on EduDCM data achieved a performance comparable to that of manually annotated datasets. Expert evaluations using a 5-point Likert scale show that EduDCM outperforms Template-Based Generation and Few-Shot GPT in terms of annotation accuracy, category coverage, and consistency. These findings emphasize EduDCM's novelty and its effectiveness in generating high-quality, scalable datasets for low-resource educational NLP tasks, thus reducing manual annotation efforts.

Keywords: educational dialogue classification; low-resource tasks; large language models; distant supervision

check for updates

Academic Editor: Rui Araújo

Received: 29 November 2024
Revised: 20 December 2024
Accepted: 26 December 2024
Published: 27 December 2024

Citation: Qi, C.; Zheng, L.; Wei, Y.; Xu, H.; Chen, P.; Gu, X. EduDCM: A Novel Framework for Automatic Educational Dialogue Classification Dataset Construction via Distant Supervision and Large Language Models. *Appl. Sci.* **2025**, *15*, 154. https://doi.org/10.3390/app15010154

1. Introduction

Educational dialogue classification is a crucial task in natural language processing (NLP) for analyzing classroom interactions and facilitating effective teaching strategies [1,2]. It enables researchers and educators to assess teaching effectiveness, identify areas for improvement, and design data-driven pedagogical interventions. For instance, analyzing whether classroom discussions emphasize critical thinking or encourage collaborative problem-solving can provide valuable insights into how to enhance learning outcomes. However, this task faces significant challenges in low-resource educational contexts due to the scarcity of annotated data and the complexities of capturing nuanced pedagogical interactions. Existing few-shot classification methods often fail to generalize well to diverse educational scenarios [3,4]. While Large Language Models (LLMs) like GPT-3.5 offer advanced data augmentation capabilities, they may not fully address the challenges of context-sensitive label generation in educational dialogues [5,6]. Recent advancements, such as GPT-4.0, have further improved the ability of LLMs to understand and process complex contextual relationships, making them a strong reference point for evaluating dialogue classification frameworks. GPT-4.0, for instance, demonstrates the ability to generate coherent and context-aware labels for ambiguous utterances, but it still requires task-specific adaptations to meet the nuanced demands of educational domains [7]. This study positions EduDCM as a framework capable of leveraging these advancements while addressing the remaining challenges in generating high-quality educational dialogue datasets.

Distant supervision, a method of generating labeled data by aligning unannotated text with predefined knowledge bases [8], has been extensively explored for automating data annotation. Despite its potential, this approach often introduces substantial noise due to erroneous assumptions, such as rigid alignment of text to labels. For example, in classroom dialogues, the same utterance might reflect different pedagogical intentions depending on the context, making rigid alignment approaches prone to misclassification. These limitations, coupled with the resource-intensive nature of manual annotation in educational contexts, highlight the need for innovative approaches to dataset construction.

To overcome these limitations, this study introduces EduDCM, a novel framework that bridges the gap between automated methods and domain-specific requirements. By leveraging the capabilities of LLMs and combining them with heuristic alignment and semantic disambiguation techniques, EduDCM ensures that annotations are not only accurate but also context-sensitive. This framework represents a significant step forward in addressing the scalability and precision trade-offs inherent in existing methods.

For low-resource educational tasks, the lack of robust and diverse training datasets remains a critical bottleneck. While advances in NLP, such as zero-shot learning and multi-task training [9], offer some solutions, they struggle to effectively adapt to task-specific requirements in the educational domain. Few-shot learning methods, particularly those leveraging prompts with LLMs [10], demonstrate potential but are limited in their scalability and performance under noisy data conditions. Moreover, these methods often fail to account for the dynamic and interaction-driven nature of classroom discussions, where context plays a pivotal role in determining the meaning of an utterance.

EduDCM overcomes these limitations by combining LLM-driven dialogue classification with advanced heuristic techniques, enhancing the quality of generated data and fostering model adaptability. In particular, EduDCM's ability to integrate domain knowledge into its annotation pipeline allows it to handle ambiguity and variability in educational dialogues effectively. This makes it a versatile tool for generating high-quality datasets tailored to low-resource educational contexts.

2. Related Work

The development of educational dialogue classification in low-resource contexts shares similarities with the evolution of relation extraction (RE) [11]. Early approaches in classification tasks primarily relied on rule-based systems [12,13]. While these methods were foundational, they lacked scalability and were heavily dependent on manual efforts, making them unsuitable for large-scale or dynamic educational datasets [14]. Subsequently, machine learning methods introduced greater flexibility, yet they still required substantial amounts of annotated data, which remains a significant challenge in low-resource educational contexts [15]. Deep learning brought advancements by capturing intricate patterns in classroom interactions, but these models continued to depend on large datasets, limiting their applicability in education.

Distant supervision has been widely explored to alleviate data scarcity by aligning unannotated text with predefined knowledge bases [16]. While this method has successfully generated labeled data, it often introduces noise due to oversimplified assumptions, such as rigid text–label alignment. In the educational domain, this manifests as misaligned dialogue annotations, which compromise model accuracy. Recent advances in few-shot learning offer promising alternatives for addressing data scarcity. Gururaja et al. [17] demonstrated that linguistic cues could enhance few-shot methods, compensating for limited annotated data. However, these methods often struggle to generalize across diverse educational scenarios due to their reliance on narrow training data [18].

The emergence of LLMs has significantly impacted low-resource tasks, including educational dialogue classification. Studies such as those by Wadhwa et al. [19] and Xu et al. [20] have shown the ability of LLMs to handle complex linguistic structures with minimal fine-tuning. Despite their strengths, LLMs face limitations in domain-specific contexts, such as nuanced educational settings where deep pedagogical knowledge is required [21]. Nonetheless, LLM-driven approaches to labeling unlabeled text highlight their ability to extract meaningful patterns and assign accurate annotations, which is essential for educational dialogue analysis.

To address these challenges, our EduDCM framework leverages the strengths of LLMs while introducing heuristic techniques for enhanced label alignment and semantic disambiguation. By mitigating the noise typically associated with distant supervision and improving context-aware label generation, EduDCM establishes a reliable foundation for constructing high-quality educational dialogue classification datasets.

3. Materials and Methods

3.1. The EduDCM Framework

The EduDCM framework integrates distant supervision and advanced LLMs to construct high-quality annotated datasets for educational dialogue classification. The methodology emphasizes context-aware label generation, semantic alignment, and multilingual adaptability, ensuring accurate and scalable annotation in low-resource settings, as illustrated in Figure 1. To further illustrate the operation of the EduDCM framework, an example workflow demonstrating its key steps in annotating an educational dialogue is provided in Figure A1.

3.1.1. Knowledge Base Construction for Educational Dialogue Classification

The knowledge base is fundamental to the EduDCM framework, serving as a structured reference for annotating unlabelled educational dialogues. It is constructed using the TalkMoves dataset, which defines three pedagogical categories, such as “learning community”, “rigorous thinking”, and “content knowledge”. Each category is enriched with representative examples derived from annotated datasets and expert-curated dialogues, reflecting diverse classroom interaction styles and contexts. The knowledge base is further expanded to support multilingual adaptability by including equivalent utterances in different languages. These enhancements ensure that the knowledge base comprehensively represents the pedagogical landscape, enabling robust alignment and classification of unannotated dialogues.

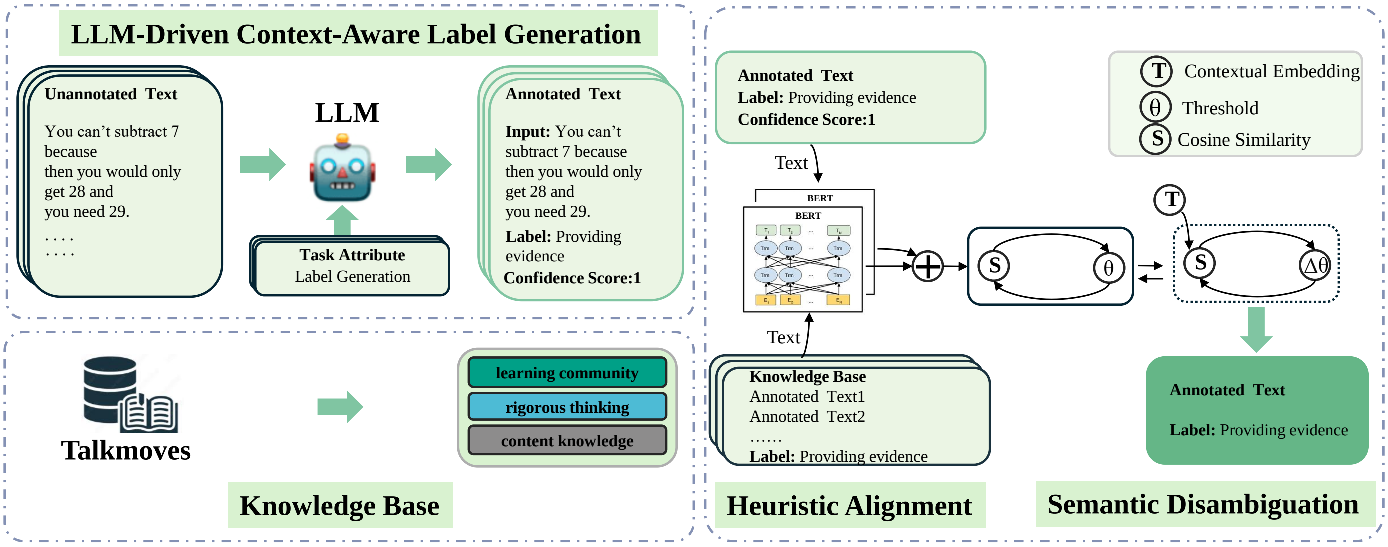

Figure 1. Workflow of the EduDCM framework. This figure illustrates the four key stages of the EduDCM framework: first, LLM-Driven Context-Aware Label Generation, where unannotated text is processed by an LLM to generate labeled text; second, Knowledge Base Construction, where a knowledge base built from the TalkMoves dataset supports classification into three pedagogical categories; and finally, Heuristic alignment and semantic disambiguation, where annotated text is aligned with knowledge base examples through semantic similarity, with context embeddings used to refine the alignment and dynamically adjust thresholds to ensure accuracy.

3.1.2. LLM-Driven Context-Aware Label Generation

EduDCM employs advanced LLMs, such as GPT-3.5-turbo [22], to generate preliminary labels for unannotated dialogues based on their conversational context. The dialogue is segmented into context windows, each containing the target utterance and its surrounding utterances, ensuring that the labels generated reflect the utterance's role within the broader conversational flow. Tailored prompts are dynamically crafted for the educational domain, instructing the LLM to classify utterances into predefined pedagogical categories while providing a rationale for each choice. For ambiguous utterances, the LLM generates multiple candidate labels, each with an associated confidence score, which are subsequently refined through alignment and validation stages. The preliminary labels generated form the input for alignment processes to ensure consistency and precision.

3.1.3. Heuristic Alignment

The tentative labels generated by the LLM are refined by aligning them not only with predefined categories but also with specific examples in the knowledge base. Using Sentence-BERT, each unannotated utterance and the sampled examples are embedded into a high-dimensional semantic space. The cosine similarity between the embedding vectors is calculated as follows:

$$\text{Sim}(u,e) = \frac{1}{m}\sum_{j=1}^{m}\frac{\vec{u}\cdot\vec{e}_j}{\|\vec{u}\|\|\vec{e}_j\|} \quad (1)$$

where $\vec{u}$ represents the embedding of the unannotated utterance, $\vec{e}_j$ represents the embedding of the j-th sampled example from the knowledge base, and m is the number of sampled examples. The mean similarity $\text{Sim}(u,e)$ is used to determine alignment. If $\text{Sim}(u,e) \geq \theta$, the utterance is considered aligned with the category associated with the sampled examples. The threshold $\theta = 0.88$ was determined experimentally to balance precision and recall. Additionally, context-aware embeddings aggregate the surrounding utterances to improve alignment in complex or ambiguous scenarios. When no matches meet the threshold, a dynamic adjustment mechanism iteratively lowers θ or expands the search to neighboring categories, formalized as

$$\theta' = \theta - \Delta\theta \tag{2}$$

where $\Delta\theta$ is the adjustment step size.

3.1.4. Optimized Semantic Disambiguation for Multilingual Adaptability

EduDCM incorporates an optimized semantic disambiguation strategy that leverages multilingual lexical resources and advanced embedding techniques to ensure robust annotation across diverse linguistic contexts. When the cosine similarity between the embedding of an unannotated utterance ($\vec{u}$) and the mean embedding of sampled examples ($\vec{e}_j$) from a category falls below the threshold θ, additional steps are employed to refine the annotation process. To address cases where semantic alignment fails due to insufficient contextual information, a contextual embedding optimization technique is applied. This approach recalculates the embedding of the target utterance by incorporating surrounding utterances in the dialogue. The optimized embedding is defined as

$$\vec{u}_{\text{context}} = \alpha \cdot \vec{u} + \frac{1}{n}\sum_{i=1}^{n} \beta_i \cdot \vec{u}_i \tag{3}$$

where $\vec{u}$ is the embedding of the target utterance, $\vec{u}_i$ represents the embedding of the i-th contextual utterance, α is the weight for the target utterance, and β_i is the weight for each contextual utterance. The weights α and β_i are dynamically adjusted based on the semantic similarity of the contextual utterances to the knowledge base examples. This ensures that relevant context contributes more significantly to the optimized embedding. If the recalculated similarity score $\text{Sim}(\vec{u}_{\text{context}}, \vec{e})$ still fails to exceed the threshold θ, the system employs a threshold adjustment mechanism:

$$\theta' = \theta - \Delta\theta \tag{4}$$

where $\Delta\theta$ is a small step size determined experimentally to balance recall and precision. By iteratively lowering the threshold, the system expands the search space to include more candidate categories without compromising overall accuracy.

In addition to embedding-based refinement, EduDCM integrates multilingual WordNet to support cross-lingual semantic mapping. WordNet's hierarchical structure is used to identify synonyms and related terms across languages, facilitating annotation in multilingual datasets. For example, an utterance labeled as "feedback" in English can be aligned with its equivalent term in another language through shared semantic paths in WordNet. If all refinement attempts fail, the utterance is flagged for manual review, ensuring that even ambiguous cases are handled effectively.

3.2. Experimental Setting

In this section, we describe the experimental setup, including the datasets used, the experimental phases, and the configuration of models employed for training and evaluation. The aim is to provide a clear understanding of the methodologies and resources that support the results presented in this study.

3.2.1. Datasets

This study evaluates the performance of the EduDCM framework using two datasets designed for educational dialogue classification. The first dataset, TalkMoves [23], is a dataset containing 10 pedagogical categories and is used to evaluate the model's generalization capabilities across diverse and complex data. The second dataset, Educational Dialogue Talk Dataset (EduTalk), was constructed based on the EduDCM framework and comprises 5200 dialogue utterances categorized into the same 10 pedagogical categories as TalkMoves. EduTalk integrates both simulated and real classroom interactions, providing a foundation for evaluating EduDCM's ability to generate high-quality annotations in low-resource educational contexts.

3.2.2. Experimental Phases

The experimental evaluation comprises three phases, focusing on dataset-specific performance and comparative analysis with other data generation methods:

- Phase 1: Individual Dataset Performance Evaluation. In this phase, models are trained and tested separately on TalkMoves and EduTalk to establish baseline performance. This step highlights differences in annotation quality between manually labeled data and EduDCM-generated data. The evaluation metrics include accuracy, F1 score, and recall.
- Phase 2: Combined Dataset Evaluation. The EduTalk and TalkMoves datasets are mixed in ratios (e.g., 3:7, 1:1) to create combined test sets. This phase evaluates the compatibility of EduTalk-generated data with high-quality TalkMoves annotations and the models' generalization capabilities across mixed datasets.
- Phase 3: Data Generation Method Comparison. In this phase, EduDCM is compared against two prominent data generation approaches. The first is Template-Based Generation [24], which creates synthetic dialogues using predefined templates and expert-defined rules. While structured, this method is labor-intensive and commonly used in early natural language generation tasks. The second approach is Few-Shot GPT Data Generation [25], leveraging GPT-3.5 in a few-shot setup to generate datasets based on limited annotated examples, relying on the model's capability to extrapolate patterns from minimal inputs. Datasets generated by these methods are evaluated on identical models and reviewed by domain experts across three dimensions: (1) the accuracy of annotations, measured by the proportion of correctly labeled utterances; (2) the coverage of pedagogical categories, indicating the extent to which all categories are sufficiently represented; (3) the consistency of annotation style, reflecting the coherence and uniformity in annotation patterns.

3.2.3. Experimental Setup

In the first phase of the experiments, to substantiate the performance of our framework, we meticulously selected a representative spectrum of models for comparison. These methodologies include Bert [26], mREBEL [27], GPT-2 [28], OneRel [29], and casRel [30]. In the third phase of the experiments, three domain experts with experience in educational dialogue annotation reviewed 500 utterances generated for educational dialogue classification by each method: EduDCM, the template-based method, and the Few-Shot GPT method. For each dataset, the experts scored the annotations on a 5-point Likert scale for each quality dimension.

All experiments were conducted on an NVIDIA GPU 4090 using PyTorch. Models were trained with a learning rate of 1×10^{-5} using the AdamW optimizer. Batch sizes of 16 and 32 were tested to evaluate their impact on model performance. A step decay learning rate strategy was employed, and dropout regularization was applied to mitigate overfitting. Each model was evaluated after each epoch on a validation set, and the best-performing checkpoint was used for final testing.

4. Results and Discussion

In this section, we present the results from the experiments conducted to evaluate the performance of models trained on EduDCM-generated datasets. We analyze various metrics such as accuracy, F1 score, and recall, and compare the performance of models on EduTalk and TalkMoves datasets. The results also include expert evaluations that validate the quality of the EduDCM-generated data.

4.1. Expert Evaluation

To evaluate the quality of the EduTalk dataset generated by the EduDCM framework, a rigorous expert review was conducted. This process involved three domain experts with extensive experience in educational dialogue analysis. The first expert is a professor in educational technology with over 15 years of experience in classroom interaction research, focusing on pedagogical strategies and dialogue analysis. The second expert is a senior researcher specializing in natural language processing for education, with a background in developing automated annotation tools for educational datasets. The third expert is a curriculum designer with a Ph.D. in education, who has contributed to the development of taxonomy-based frameworks for evaluating classroom interactions. A random sample comprising 30% of the dataset, totaling 1560 instances, was selected to ensure a comprehensive assessment. Each dialogue utterance was annotated according to the pedagogical categories outlined in the TalkMoves taxonomy, including categories such as Learning Community, Content Knowledge, and Rigorous Thinking, with subcategories such as Relating to another student, Pressing for accuracy, and Providing evidence or reasoning.

The results presented in Table 1 demonstrate that the EduTalk dataset achieves high semantic accuracy across most categories, with an average positive annotation rate of 85%. Categories such as *Learning Community: Keeping everyone together* and *Learning Community: Getting students to relate to another's ideas* exhibit the highest accuracy rates (94% and 91%, respectively), highlighting the EduDCM framework's ability to capture well-structured teacher talk moves. Similarly, categories like *Learning Community: Restating* and *Content Knowledge: Pressing for accuracy* show strong performance, with positive annotation rates exceeding 85%. These findings suggest that the EduDCM framework is particularly adept at identifying and annotating dialogue types with clear pedagogical functions. However, the categories *Content Knowledge: Making a claim* and *Rigorous Thinking: Providing evidence or reasoning* demonstrated relatively lower accuracy rates, with positive annotation rates of 77% and 74%, respectively. This indicates challenges in accurately capturing student talk moves involving abstract reasoning or detailed evidence. These discrepancies could be attributed to the variability in how students articulate reasoning and evidence, often requiring contextual interpretation that may not always align with predefined annotation rules.

Consistency in annotations was generally high for categories with well-defined boundaries, such as *Learning Community: Keeping everyone together* and *Learning Community: Restating*. In contrast, categories like *Rigorous Thinking: Pressing for reasoning* and *Rigorous Thinking: Providing evidence or reasoning* exhibited more variability in expert agreement, highlighting the need for further refinement in the framework's handling of nuanced or context-dependent utterances. The representational coverage of categories across the dataset appears balanced, with no single category disproportionately dominating the annotations. This ensures that the EduTalk dataset provides a diverse and equitable foundation for training and evaluating dialogue classification models. Overall, the expert evaluation underscores the strength of the EduDCM framework in generating high-quality educational datasets while identifying opportunities for improvement in handling complex dialogue types involving reasoning and evidence.

Table 1. Expert evaluation of the EduTalk dataset, showing positive and negative annotation rates across 10 categories.

Category	Sampled Instances	Positive (%)	Negative (%)
Learning Community: Keeping everyone together	175	94	6
Learning Community: Getting students to relate to another's ideas	205	91	9
Learning Community: Restating	195	89	11
Learning Community: Relating to another student	148	81	19
Learning Community: Asking for more information	118	79	21
Content Knowledge: Pressing for accuracy	185	87	13
Content Knowledge: Making a claim	129	77	23
Rigorous Thinking: Revoicing	162	85	15
Rigorous Thinking: Pressing for reasoning	142	82	18
Rigorous Thinking: Providing evidence or reasoning	101	74	26

4.2. Three-Phase Evaluation

In this section, we present a comprehensive evaluation of the EduDCM framework through three distinct experimental phases. These phases are designed to assess the effectiveness of EduDCM-generated datasets in various contexts, including individual dataset performance, the combination of EduTalk and TalkMoves datasets, and a comparison with other data generation methods. Each phase aims to provide valuable insights into the scalability and generalizability of the EduDCM framework.

4.2.1. Phase 1: Individual Dataset Performance Evaluation

The performance of various models was evaluated on the TalkMoves and EduTalk datasets to establish baseline metrics and assess the quality of annotations. The models used for evaluation include BERT, mREBEL, GPT-2, OneRel, and CasRel. The evaluation metrics include F1 score, precision (P), and recall (R), providing a comprehensive view of each model's effectiveness in handling the different datasets.

The results in Table 2 demonstrate that models perform better on the EduTalk dataset compared to TalkMoves. This is largely attributed to the controlled and consistent nature of EduTalk's annotations, which align closely with the structured output generated by the EduDCM framework. The CasRel and OneRel models exhibit superior performance on both datasets, with F1 scores of 0.914 and 0.917 on EduTalk, and 0.778 and 0.742 on TalkMoves, respectively. These results highlight the effectiveness of models specifically designed for joint entity and relation extraction when applied to datasets with high annotation quality.

Table 2. Performance comparison of models on EduTalk and TalkMoves datasets.

Model	EduTalk			TalkMoves		
	F1	Recall (R)	Precision (P)	F1	Recall (R)	Precision (P)
BERT	0.863	0.871	0.856	0.624	0.710	0.565
mREBEL	0.879	0.868	0.890	0.596	0.580	0.613
GPT-2	0.892	0.844	0.947	0.635	0.642	0.628
OneRel	0.917	0.924	0.910	0.742	0.726	0.759
CasRel	0.914	0.907	0.921	0.778	0.766	0.790

For the TalkMoves dataset, which contains manually annotated data, performance metrics are generally lower. This can be attributed to the inherent complexity and variability of the dataset, as well as the potential for ambiguous or overlapping pedagogical categories. The CasRel model outperforms other models with an F1 score of 0.778, reflecting its robustness in handling diverse and nuanced annotations. OneRel also performs well, achieving an F1 score of 0.742, indicating its adaptability to tasks requiring precise annotation quality.

In contrast, the EduTalk dataset, generated via EduDCM, demonstrates competitive results, with models achieving performance metrics comparable to those on manually annotated datasets like TalkMoves. The structured nature of the annotations, combined with the semantic alignment mechanisms employed by the EduDCM framework, likely contributes to the high performance. CasRel and OneRel maintain their leading positions, achieving precision and recall scores above 0.91, which underscores their ability to capitalize on the dataset's quality.

Models like BERT and mREBEL exhibit comparatively lower performance on both datasets. This discrepancy is particularly pronounced on TalkMoves, where the variability and complexity of the data pose significant challenges. However, their results on EduTalk suggest that the EduDCM framework mitigates some of the limitations associated with simpler modeling approaches, providing structured annotations that enhance model learning.

To evaluate annotation quality, three domain experts reviewed a random sample of 300 instances from each dataset. Table 3 summarizes the results. Both datasets received high scores for annotation accuracy, category coverage, and consistency, with EduTalk achieving mean scores of 4.4, 4.3, and 4.4, respectively, and TalkMoves scoring slightly higher with 4.5, 4.4, and 4.5. Annotation accuracy was measured by how well the generated labels matched the intended pedagogical categories as defined in the taxonomy. Category coverage was evaluated by determining whether all relevant categories were adequately represented in the dataset. Consistency was assessed by verifying the uniformity of annotations across similar dialogue instances. The inter-rater reliability, measured by Cohen's Kappa, was 0.85 or higher for both datasets, reflecting strong agreement among the experts. Experts noted that EduTalk's structured and context-aware annotations, generated by the EduDCM framework, aligned well with pedagogical categories, enabling consistent interpretation. However, the manually annotated TalkMoves dataset exhibited slightly higher annotation accuracy due to the direct involvement of human annotators with domain expertise.

Table 3. Expert evaluation results for EduTalk and TalkMoves datasets.

Dataset	Criteria	Mean Score	Standard Deviation	Cohen's Kappa
EduTalk	Annotation Accuracy	4.4	0.3	0.85
	Category Coverage	4.3	0.4	0.83
	Consistency	4.4	0.3	0.84
TalkMoves	Annotation Accuracy	4.5	0.3	0.86
	Category Coverage	4.4	0.4	0.84
	Consistency	4.5	0.3	0.86

In summary, the evaluation results demonstrate the reliability of the EduTalk dataset generated by EduDCM, as evidenced by the comparable performance of models trained on EduTalk to those trained on manually annotated datasets like TalkMoves. These findings validate the framework's potential to bridge the gap between distant supervision and manual annotation in educational NLP tasks.

4.2.2. Phase 2: Combined Dataset Evaluation

In this phase, we aim to evaluate the generalization capability of models trained on the EduTalk dataset. To achieve this, we created mixed test sets by combining EduTalk and TalkMoves datasets in different ratios (3:7 and 1:1) and assessed the models' performance on these test sets. This approach allows us to determine the extent to which the EduTalk data generated by the EduDCM framework can match the quality of the manually annotated TalkMoves dataset.

The results, as shown in Table 4, reveal several important trends. In the 3:7 dataset ratio, where TalkMoves dominates the test set, all models exhibit lower performance compared to the 1:1 ratio. This decline can be attributed to the smaller size of the TalkMoves dataset, which limits the diversity of the data and reduces the models' ability to generalize effectively. Among the models, CasRel and OneRel demonstrate the highest F1 scores of 0.779 and 0.747, respectively, reflecting their superior ability to adapt to mixed datasets, even with a significant proportion of manually annotated TalkMoves data.

Table 4. Performance of models trained on EduTalk and tested on mixed EduTalk and TalkMoves test sets.

Model	**Dataset Ratio 3:7**			**Dataset Ratio 1:1**		
	F1	**Recall (R)**	**Precision (P)**	**F1**	**Recall (R)**	**Precision (P)**
BERT	0.646	0.686	0.611	0.739	0.775	0.707
mREBEL	0.625	0.652	0.601	0.722	0.750	0.695
GPT-2	0.670	0.693	0.648	0.760	0.767	0.753
OneRel	0.747	0.766	0.729	0.818	0.836	0.801
CasRel	0.779	0.793	0.765	0.838	0.848	0.829

In the 1:1 dataset ratio, where EduTalk and TalkMoves are equally represented, all models show improved performance. This improvement underscores the role of the high-quality EduTalk data in enriching the diversity of the smaller TalkMoves dataset, addressing the challenges arising from its limited size. CasRel achieves the highest F1 score of 0.838, followed closely by OneRel with an F1 score of 0.818. Both models demonstrate balanced precision and recall, indicating their robust generalization capability in handling diverse test sets.

Lower-performing models, such as BERT and mREBEL, show similar trends, with better performance in the 1:1 ratio compared to the 3:7 ratio. However, their F1 scores of 0.739 and 0.722 in the 1:1 ratio suggest limitations in handling the variability of manually annotated TalkMoves data. These models rely more heavily on dataset size and quality, which may restrict their adaptability in low-resource scenarios. The performance of GPT-2, with an F1 score of 0.760 in the 1:1 ratio, indicates its competitive ability to generalize when presented with a balanced dataset. However, its slightly lower precision compared to CasRel and OneRel suggests that it struggles with maintaining high accuracy in more diverse data conditions.

Overall, these results highlight the capability of the EduTalk dataset to enhance model performance when combined with manually annotated TalkMoves data. High-performing models like CasRel and OneRel exhibit robust adaptability to mixed datasets, validating the EduDCM framework's potential to generate high-quality data that effectively support dialogue classification tasks in low-resource educational contexts.

4.2.3. Phase 3: Data Generation Method Comparison

In this phase, EduDCM is compared against two prominent data generation approaches. The first is Template-Based Generation, which creates synthetic dialogues using predefined templates and expert-defined rules. While structured, this method is labor-intensive and commonly used in early natural language generation tasks. The second approach is Few-Shot GPT Data Generation, which leverages GPT-3.5 in a few-shot setup to generate datasets based on limited annotated examples, relying on the model's capability to extrapolate patterns from minimal inputs. The datasets generated by these methods are evaluated through identical models and examined by domain experts across three dimensions.

The results presented in Table 5 reveal significant differences in dataset quality across the three methods. EduDCM consistently outperforms the other methods, achieving the highest scores in all three dimensions. Specifically, its score of 4.7 in accuracy of annotations highlights its ability to correctly label utterances, thanks to the framework's integration of context-aware label generation and heuristic alignment techniques. Similarly, EduDCM achieves a coverage score of 4.6, demonstrating its capacity to ensure balanced representation across all pedagogical categories. Its consistency score of 4.8 reflects its effectiveness in maintaining a coherent and uniform annotation style, reducing variability that could negatively impact downstream tasks.

Table 5. Expert evaluation of data generation methods across three dimensions (5-point Likert scale).

Method	Annotation Accuracy	Category Coverage	Annotation Consistency
Template-Based	3.2	2.8	3.0
Few-Shot GPT	4.1	4.0	4.2
EduDCM	4.7	4.6	4.8

In contrast, the Template-Based Generation method performs the worst across all dimensions. Its accuracy score of 3.2 indicates limited flexibility in capturing the nuanced and context-dependent nature of educational dialogues. The coverage score of 2.8 reveals significant biases in category representation, as predefined templates tend to favor certain dialogue structures while underrepresenting others. Additionally, the consistency score of 3.0, while reflecting a uniform style within the constraints of the templates, highlights its inability to adapt to real-world variations in dialogue.

The Few-Shot GPT method demonstrates substantial improvements over Template-Based Generation. Its accuracy score of 4.1 indicates that GPT-3.5 can leverage contextual understanding to generate meaningful and accurate annotations. However, its coverage score of 4.0 suggests that certain pedagogical categories, especially those less represented in the prompt examples, may not be adequately annotated. The consistency score of 4.2 shows that Few-Shot GPT maintains a relatively uniform annotation style, though occasional variability due to inherent model randomness is observed.

In summary, these results highlight the superiority of EduDCM in generating high-quality datasets for educational dialogue classification. While Few-Shot GPT offers a viable alternative with reasonable performance, its reliance on limited prompts restricts its effectiveness in achieving comprehensive category coverage. Template-Based Generation, though structured, lacks the flexibility and adaptability required for the nuanced nature of educational dialogues.

4.3. Ablation Study

To comprehensively evaluate the contributions of each component in the EduDCM framework to dataset generation quality, we conducted a series of ablation experiments. The configurations included (1) the complete framework, (2) removal of the LLM-Driven Context-Aware Label Generation (replaced with traditional distant supervision, where an utterance is labeled based on semantic similarity to knowledge base examples), (3) exclusion of heuristic alignment, and (4) exclusion of optimized semantic disambiguation. The knowledge base component was retained in all configurations, as it is integral to the framework. Each configuration generated 100 annotated samples spanning the pedagogical categories in the EduTalk dataset. Evaluation metrics included the proportion of positive instances and a data richness score, developed by domain experts, to measure the depth and complexity of annotations on a 1-to-5 scale.

The results in Table 6 highlight the critical role of each component in the EduDCM framework. Removing the LLM-Driven Context-Aware Label Generation component caused the most significant drop in both metrics, with the proportion of positive instances decreasing from 87% to 72% and the data richness score falling from 4.6 to 3.2. This decline underscores the importance of context-aware label generation, as traditional distant supervision relies solely on semantic similarity, often missing nuanced context in dialogues.

Table 6. Impact of key components on framework performance.

Configuration	Positive Instances (%) ↑	Data Richness ↑
EduDCM	87	4.6
$\text{EduDCM}_{\text{NO-LLM-Driven Labeling}}$	72	3.2
$\text{EduDCM}_{\text{NO-Heuristic Alignment}}$	83	4.3
$\text{EduDCM}_{\text{NO-Semantic Disambiguation}}$	82	4.1

Excluding heuristic alignment led to a moderate performance decline, reducing the positive instance proportion to 83% and the data richness score to 4.3. This indicates that heuristic alignment effectively refines the initial labels, improving both accuracy and semantic depth by aligning utterances with relevant knowledge base examples.

The removal of optimized semantic disambiguation resulted in a smaller impact, with the positive instance proportion dropping to 82% and the data richness score to 4.1. While this component aids in handling ambiguous cases and improving multilingual adaptability, its influence is supplementary compared to the other components.

Overall, the ablation study demonstrates that LLM-Driven Context-Aware Label Generation is the cornerstone of the EduDCM framework. Heuristic alignment and semantic disambiguation further enhance annotation quality, suggesting that the integration of these components is essential for generating high-quality datasets for educational dialogue classification.

Evaluation of Different LLMs in the LLM-Driven Labeling Component

This section evaluates the impact of integrating different LLMs into the LLM-Driven Labeling component of the EduDCM framework. The assessment focuses on how the replacement of the baseline GPT-3.5 with other LLMs such as Gemini-pro [31], Claude2 [32], and GPT-4o [33] affects the framework's performance, measured by the positive instance proportion and data richness score. These metrics reflect the accuracy and depth of annotations generated by each LLM. The baseline GPT-3.5 has demonstrated robust performance in prior evaluations and serves as the standard for comparison. All models were tested under identical conditions to ensure consistency in results.

Table 7 summarizes the performance of different LLMs within the LLM-Driven Labeling component. The baseline GPT-3.5 achieved strong results, with a positive instance proportion of 87% and a data richness score of 4.6. This reinforces its capability to generate accurate and contextually rich annotations, setting a high standard for comparison.

Table 7. Performance comparison of different LLMs in the LLM-Driven Labeling component.

LLM	Positive Instances (%) ↑	Data Richness ↑
GPT-3.5	87	4.6
Gemini-pro	81	4.1
Claude2	85	4.4
GPT-4o	89	4.8

Replacing GPT-3.5 with Gemini-pro resulted in a notable decline in both metrics, with the positive instance proportion dropping to 81% and the data richness score decreasing to 4.1. This suggests that while Gemini-pro can provide basic annotation functionality, its capacity for capturing nuanced pedagogical interactions and generating semantically rich annotations is limited compared to GPT-3.5.

Claude2 demonstrated competitive performance, achieving a positive instance proportion of 85% and a data richness score of 4.4. Although slightly below the baseline, its consistent results indicate that Claude2 is a viable alternative in scenarios where cost or computational constraints favor its adoption.

GPT-4o outperformed all other LLMs, achieving the highest positive instance proportion (89%) and data richness score (4.8). This superior performance highlights its advanced ability to understand conversational context and generate precise, detailed annotations. The results suggest that GPT-4o's enhanced contextual understanding and semantic capabilities make it particularly suitable for tasks requiring deep processing and high-quality annotation generation.

These findings underscore the importance of selecting an appropriate LLM for the LLM-Driven Labeling component. While GPT-4o offers the best overall performance, Claude2 provides a cost-effective alternative with near-baseline results. Gemini-pro, while functional, may require additional optimization to meet the demands of complex educational dialogue classification tasks. This evaluation further highlights how advancements in LLM technology can significantly enhance the effectiveness of frameworks like EduDCM, enabling more scalable and accurate dataset generation in low-resource settings.

5. Conclusions

This study presents EduDCM, a novel framework designed to address the critical challenge of generating high-quality educational dialogue classification datasets in low-resource contexts. By integrating distant supervision with the advanced capabilities of LLMs, EduDCM offers a scalable and efficient solution for automating dataset annotation. The framework leverages context-aware label generation, heuristic alignment, and optimized semantic disambiguation to enhance annotation quality while mitigating noise associated with traditional distant supervision methods. The experimental results confirm the effectiveness of EduDCM across several dimensions. The EduTalk dataset, constructed using this framework, shows performance on par with manually annotated datasets, as demonstrated by the robust results achieved by classification models trained on EduTalk. Furthermore, the comparison with Template-Based Generation and Few-Shot GPT methods underscores EduDCM's superior annotation accuracy, category coverage, and consistency. These findings validate EduDCM's ability to address the challenges of low-resource educational NLP tasks effectively. The ablation study further highlights the importance of the LLM-Driven Context-Aware Label Generation component, which significantly improves dataset quality by providing nuanced and context-sensitive annotations. The experiments with various LLMs, including GPT-4.0, also demonstrate the framework's potential for further enhancement through state-of-the-art models, confirming its scalability and adaptability.

This paper makes the following key contributions: (1) The innovative proposition of the EduDCM framework, which integrates LLMs and distant supervision for automatic dataset construction, addressing noise and sparsity issues commonly encountered in low-resource educational contexts through dynamic label generation and heuristic alignment. (2) The development of the EduTalk dataset, which encompasses diverse classroom dialogues annotated with pedagogical categories, serving as a valuable resource for advancing educational dialogue classification. (3) The advancement of low-resource educational NLP, as EduDCM demonstrates significant improvements in dataset quality and classification model performance, making it well suited for a range of low-resource educational NLP tasks.

Looking forward, future research will focus on expanding the applicability of EduDCM to multilingual and multimodal educational datasets, addressing a broader range of pedagogical interactions. Additionally, refining the alignment and disambiguation processes through advanced machine learning techniques will be a priority to further improve annotation quality and adaptability. Overall, EduDCM establishes a robust foundation for advancing educational dialogue analysis, paving the way for more effective teaching strategies and classroom interaction studies.

Author Contributions: Conceptualization, C.Q., X.G. and L.Z.; data curation, H.X. and Y.W.; funding acquisition, L.Z.; methodology, C.Q., X.G. and L.Z.; project administration, X.G. and L.Z.; resources, H.X.; software, P.C. and Y.W.; supervision, X.G. and L.Z.; visualization, H.X. and Y.W.; writing—original draft, C.Q.; writing—review and editing, C.Q., H.X. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Opening Foundation of the State Key Laboratory of Cognitive Intelligence (iED2023-008) and the National Natural Science Foundation of China (62477013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets can be found in the following repositories: TalkMoves https://github.com/SumnerLab/TalkMoves (accessed on 15 October 2024); the EduTalk dataset used in this study is not publicly available due to privacy concerns. Access to the dataset can be requested from the first author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
EduDCM	Educational Dialogue Classification via Distant Supervision and LLMs
RE	Relation extraction
GPT	Generative Pre-trained Transformer
BERT	Bidirectional Encoder Representations from Transformers
EduTalk	Educational Dialogue Talk Dataset

Appendix A. Demonstrative Example of the EduDCM Framework

This appendix presents an example workflow of the EduDCM framework in action, highlighting its process of annotating an educational dialogue. The example demonstrates how the framework LLMs and heuristic alignment to accurately classify a dialogue utterance into a predefined pedagogical category.

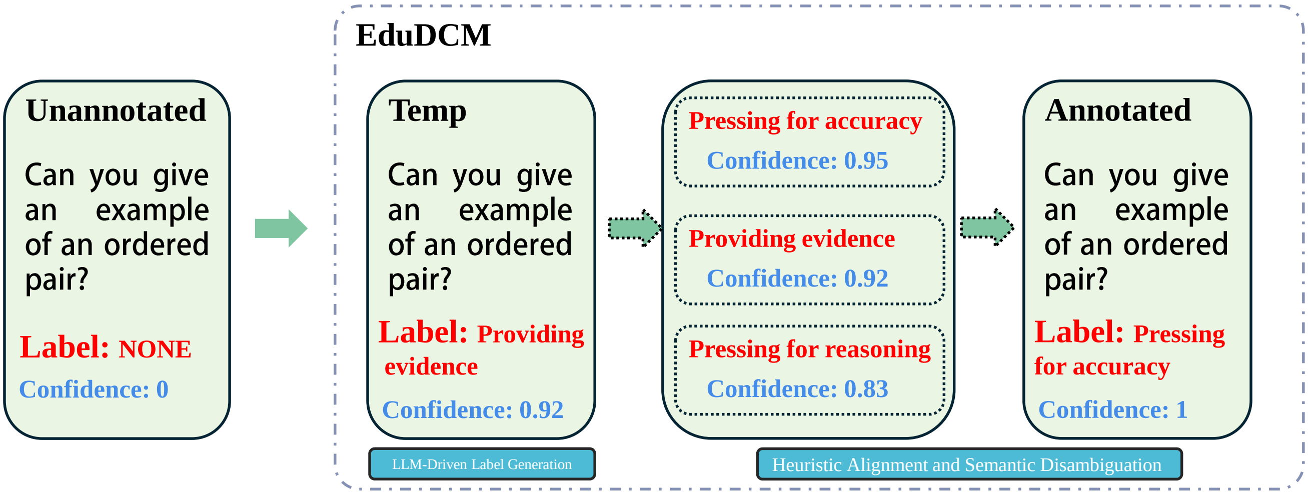
Unannotated
Can you give an example of an ordered pair?
Label: NONE
Confidence: 0
EduDCM
Temp
Can you give an example of an ordered pair?
Label: Providing evidence
Confidence: 0.92
Pressing for accuracy
Confidence: 0.95
Providing evidence
Confidence: 0.92
Pressing for reasoning
Confidence: 0.83
Annotated
Can you give an example of an ordered pair?
Label: Pressing for accuracy
Confidence: 1
LLM-Driven Label Generation
Heuristic Alignment and Semantic Disambiguation

Figure A1. An example workflow of the EduDCM framework annotating an educational dialogue. The process includes LLM-driven label generation, heuristic alignment, and final output generation.

References

1. Song, Y.; Lei, S.; Hao, T.; Lan, Z.; Ding, Y. Automatic classification of semantic content of classroom dialogue. *J. Educ. Comput. Res.* **2021**, *59*, 496–521. [CrossRef]
2. Lin, J.; Tan, W.; Du, L.; Buntine, W.; Lang, D.; Gašević, D.; Chen, G. Enhancing educational dialogue act classification with discourse context and sample informativeness. *IEEE Trans. Learn. Technol.* **2023**, *17*, 258–269. [CrossRef]
3. Lu, W.; Zhou, Y.; Yu, J.; Jia, C. Concept extraction and prerequisite relation learning from educational data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA 27 January–1 February 2019; Volume 33, pp. 9678–9685.
4. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
5. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
6. Zhou, Y.; Guo, C.; Wang, X.; Chang, Y.; Wu, Y. A survey on data augmentation in large model era. *arXiv* **2024**, arXiv:2401.15422.
7. Li, Y.; Liu, J.; Yang, S. Is ChatGPT a Good Middle School Teacher? An Exploration of its Role in Instructional Design. In Proceedings of the 3rd International Conference on New Media Development and Modernized Education, NMDME 2023, Xi'an, China, 13–15 October 2023.
8. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant Supervision for Relation Extraction without Labeled Data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; ACL '09, Volume 2, pp. 1003–1011.
9. Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C.P.; Wang, X.Z.; Wu, Q.J. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4051–4070. [CrossRef] [PubMed]
10. Song, C.H.; Wu, J.; Washington, C.; Sadler, B.M.; Chao, W.L.; Su, Y. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 2998–3009.
11. Zhao, X.; Deng, Y.; Yang, M.; Wang, L.; Zhang, R.; Cheng, H.; Lam, W.; Shen, Y.; Xu, R. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.* **2024**, *56*, 1–39. [CrossRef]
12. Lawrence, R.L.; Wright, A. Rule-based classification systems using classification and regression tree (CART) analysis. *Photogramm. Eng. Remote Sens.* **2001**, *67*, 1137–1142.
13. Qin, B.; Xia, Y.; Prabhakar, S.; Tu, Y. A rule-based classification algorithm for uncertain data. In Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 29 March–2 April 2009; IEEE: Piscataway Township, NJ, USA, 2009; pp. 1633–1640.
14. Ben Abacha, A.; Zweigenbaum, P. Automatic extraction of semantic relations between medical entities: A rule based approach. *J. Biomed. Semant.* **2011**, *2*, 1–11. [CrossRef] [PubMed]
15. Cui, M.; Li, L.; Wang, Z.; You, M. A Survey on Relation Extraction. In Proceedings of the Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence, Chengdu, China, 26–29 August 2017; Li, J., Zhou, M., Qi, G., Lao, N., Ruan, T., Du, J., Eds.; Springer: Singapore, 2017; pp. 50–58.

16. Zhou, K.; Qiao, Q.; Li, Y.; Li, Q. Improving Distantly Supervised Relation Extraction by Natural Language Inference. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 14047–14055. [CrossRef]
17. Gururaja, S.; Dutt, R.; Liao, T.; Rosé, C. Linguistic representations for fewer-shot relation extraction across domains. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7502–7514. [CrossRef]
18. Hiller, M.; Ma, R.; Harandi, M.; Drummond, T. Rethinking generalization in few-shot classification. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 3582–3595.
19. Wadhwa, S.; Amir, S.; Wallace, B. Revisiting Relation Extraction in the era of Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 15566–15589. [CrossRef]
20. Xu, X.; Zhu, Y.; Wang, X.; Zhang, N. How to Unleash the Power of Large Language Models for Few-shot Relation Extraction? In Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Toronto, ON, Canada, 13 July 2023; Sadat Moosavi, N., Gurevych, I., Hou, Y., Kim, G., Kim, Y.J., Schuster, T., Agrawal, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 190–200. [CrossRef]
21. Li, B.; Fang, G.; Yang, Y.; Wang, Q.; Ye, W.; Zhao, W.; Zhang, S. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. *arXiv* **2023**, arXiv:2304.11633. [CrossRef]
22. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
23. Suresh, A.; Jacobs, J.; Harty, C.; Perkoff, M.; Martin, J.H.; Sumner, T. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *arXiv* **2022**, arXiv:2204.09652.
24. van der Lee, C.; Krahmer, E.; Wubben, S. Automated learning of templates for data-to-text generation: Comparing rule-based, statistical and neural methods. In Proceedings of the 11th International Conference on Natural Language Generation, Tilburg, The Netherlands, 5–8 November 2018; pp. 35–45.
25. Brown, T.B. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
27. Huguet Cabot, P.L.; Navigli, R. REBEL: Relation Extraction By End-to-end Language generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; Moens, M.F., Huang, X., Specia, L.,Yih, S.W.t., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2370–2381. [CrossRef]
28. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
29. Shang, Y.M.; Huang, H.; Mao, X. OneRel: Joint Entity and Relation Extraction with One Module in One Step. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 11285–11293. [CrossRef]
30. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1476–1488. [CrossRef]
31. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
32. Models, C. Model Card and Evaluations for Claude Models. 2023. Available online: https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf (accessed on 28 November 2024).
33. OpenAI. GPT-4o. 2024. Available online: https://chatgpt.com/ (accessed on 16 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.