Using Large Language Models to Provide Explanatory Feedback to Human Tutors*

Jionghao Lin[1,2,*], Danielle R. Thomas[1], Feifei Han[3], Shivang Gupta[1], Wei Tan[2], Ngoc Dang Nguyen[2] and Kenneth R. Koedinger[1]

[1]*Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2]*Monash University, Clayton, VIC 3800, Australia*
[3]*University of Toronto, Toronto, ON M5S 1A1 Canada*

Abstract

Research demonstrates learners engaging in the process of producing explanations to support their reasoning, can have a positive impact on learning. However, providing learners real-time explanatory feedback often presents challenges related to classification accuracy, particularly in domain-specific environments, containing situationally complex and nuanced responses. We present two approaches for supplying tutors real-time feedback within an online lesson on how to give students effective praise. This work-in-progress demonstrates considerable accuracy in binary classification for corrective feedback of effective, or effort-based ($F_1 score$ = 0.811), and ineffective, or outcome-based ($F_1 score$ = 0.350), praise responses. More notably, we introduce progress towards an enhanced approach of providing explanatory feedback using large language model-facilitated named entity recognition, which can provide tutors feedback, not only while engaging in lessons, but can potentially suggest real-time tutor moves. Future work involves leveraging large language models for data augmentation to improve accuracy, while also developing an explanatory feedback interface.

Keywords

Large Language Models, Named Entity Recognition, Tutor Training, Explanatory Feedback, Natural Language Processing

1. Introduction

Tutoring is among the most highly adaptable and consistently successful interventions to increase student learning [1, 2]. However, despite the known positive impacts of tutoring on achievement, there is a lack of qualified and skilled tutors outside of private, high-income communities, ready to provide content and socio-motivational support to students [1]. Due to the shortage of professional tutors, often certified teachers and paraprofessionals, the focus has shifted to preparing novice tutors, such as community volunteers, retired adults, and college students [2]. The demand for professional development personalized to meet the needs of nonprofessional and novice tutors is high [2], with training on social-emotional learning,

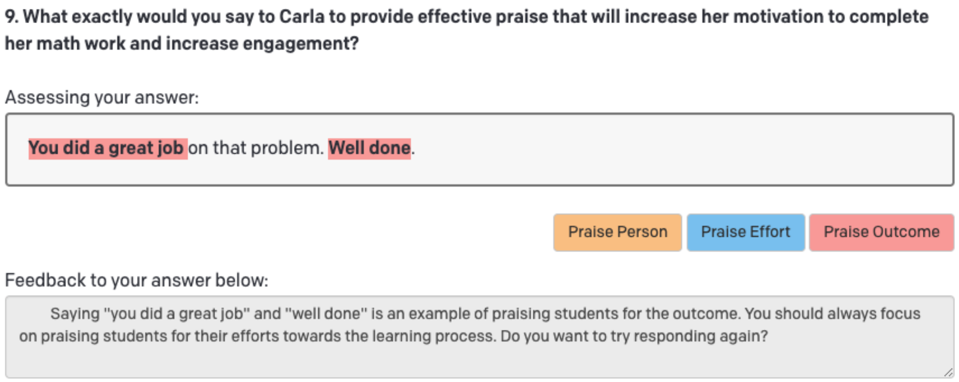
AIED 2023 Workshop: Towards the Future of AI-Augmented Human Tutoring in Math Learning, July 07, 2023, Tokyo, Japan

*Corresponding author.

Jionghao@cmu.edu (J. Lin); Drthomas@cmu.edu (D. R. Thomas); feifei.han@mail.utoronto.ca (F. Han); Shivang@cmu.edu (S. Gupta); wei.tan2@monash.edu (W. Tan); dan.nguyen2@monash.edu (N. D. Nguyen); Koedinger@cmu.edu (K. R. Koedinger)

relationship building, and attending to student motivation and self-efficacy as common topics requested among unskilled tutors [3]. Online, scenario-based lessons on these topics have been developed to provide situational experiences to inexperienced tutors [21] and preservice teachers [4]. The ability to administer real-time explanatory feedback within constructed-response questions dealing with common tutoring scenarios (e.g., a student struggling with motivation) is powerful. Immediate feedback on errors, similar to the feedback received while engaging in the deliberate practice of responding to situational judgment tests, is described as a "favorable learning condition," supporting learning [5].

We present a method of providing tutors real-time explanatory feedback harnessing large language models (LLMs). Our approach employs a template-based strategy leveraging named entity recognition (NER), a subtask of natural language processing, that classifies similar pieces of information [6]. By tagging similar pieces of information, called named entities (NEs), for effective and ineffective tutor responses, NER becomes a suitable and viable method for delivering tutors feedback. For example, classifying desired and less-desired tutor responses on how to effectively praise students yields the following NEs: praising for effort, or process-focused praise (*Effort*); ability- or outcome-focused praise (*Outcome*); and person-based praise (*Person*). Using NER, segments of tutor responses can be systematically identified aligning with the appropriate NEs. For instance, in Figure 1 the tutor response *"Good job! You got the right answer, and you stuck with it"* tags tutor utterances to produce the following NEs: *"Good job" (Outcome)* and *"stuck with it" (Effort)*. Tagged NEs can be used to create the corresponding templated feedback: *"Saying [insert Effort] is a nice example of process-focused praise, which praises students for their effort."* Conversely, templated feedback for a less-desired response could be: *"Saying [insert Outcome] is praising students for the outcome. You should focus on praising the students for their effort and process towards learning. Do you want to try responding again?"* The research recommended approach, representing the desired tutor response and, commonly observed, less-desired responses can be tagged to corresponding NEs. By tagging pieces of information, aligning with tutoring approaches for responding to the given scenarios, NER becomes a suitable method for generating templated feedback to tutors.



9. What exactly would you say to Carla to provide effective praise that will increase her motivation to complete her math work and increase engagement?

Assessing your answer:

You did a great job on that problem. **Well done**.

Feedback to your answer below:

Saying "you did a great job" and "well done" is an example of praising students for the outcome. You should always focus on praising students for their efforts towards the learning process. Do you want to try responding again?

Figure 1: An example of providing a tutor templated feedback that does not contain the desire responses of praising students for their efforts.

Automatic short answer grading, or the process of automatically scoring learner answers to constructed-responses questions (often applying several machine learning models), has received notable attention due to advances in AI-based technologies [7]. Most automatic short answer grading methods follow a two-step approach: 1) using a representation, or training set, of learner responses to train the model, using natural language processing methods, and 2) labeling responses via a machine learning classifier to predict the learner's score or performance [7]. Presently, advanced approaches using LLMs to process learner responses are taking precedence over traditional, human-identified feature analysis [7]. Despite the advantages to using LLMs, there are several limitations: the model not being well-adapted for the nuanced and varied responses among the learner population; the requirement of having to train one model per question, with related or follow-up questions being treated as mutually exclusive [7]; and, the need for a large number of tutor responses in the representation dataset. This workshop paper presents a method of providing corrective and explanatory feedback to tutors participating in an online lesson on giving students effective praise. This work-in-progress introduces an ongoing effort to enhance approaches towards automatic short answer grading using LLMs, enabling NER for identifying relevant components of a tutor response. The primary research questions addressed, include:

RQ1: Can we apply a binary classification method for effectively labeling tutor responses, as effective or ineffective, to provide corrective feedback?

RQ2: How can we enhance past approaches of providing explanatory feedback using LLM-facilitated named entity recognition to administer templated feedback identifying relevant parts of the tutor response?

2. Related Work

2.1. Feedback Generation

Feedback can have a profound impact on learning and achievement; however, its influence may be beneficial or detrimental to learning depending on type and delivery [8, 9, 10]. Feedback is most beneficial within the learning context, delivered after the learner has engaged with the initial instruction, and when addressing misconceptions or faulty reasoning [8]. Immediate, explanatory feedback, or feedback detailing the reasoning why a response is desired or not, assists learners with participating in deliberate practice. The online lesson has tutors engaging in the deliberate practice of responding to a common tutoring scenario (i.e., a student struggling to stay motivated) by asking them how to best respond. Tutors then explain their reasoning and observe the most-desired approach receiving feedback on their chosen selected response option [3, 11]. An expansion of this previous work is to provide explanatory feedback to tutors on their textual replies to the constructed-response questions. Generating explanatory feedback to tutors using enhanced approaches, such as using LLMfacilitated NER, shows promise as a method of providing accurate and timely feedback to tutors. The creation of templated feedback, including specific references to desired and less-desired elements of the tutor responses, is influenced by earlier results on the effectiveness of having a rich, data-driven error diagnosis taxonomy driving template-based feedback [12].

2.2. Named Entity Recognition

A named entity (NE) is a word or phrase distinct from a set of words that have similar attributes [6]. For example, in the text *"John said that Pittsburgh is wonderful in the winter"*, *"John"*, *"Pittsburgh"*, and *"winter"* are considered NEs, which represent a person, location, and time, respectively. Named entity recognition (NER) is a fundamental task in natural language processing, which aims to automatically locate NE in the text and classify them into different categories such as person, organization, and location [6]. In the example, to identify the NE *"Pittsburgh"*, a NER model first locates the position of *"Pittsburgh"* in the text and then classifies the entity *"Pittsburgh"* into the category of location. As discussed by [6], there are two primary categories of name entities: (i) generic (e.g., person and organization) and (ii) domain-specific (e.g., enzymes and genes). Since the present work aims to investigate the potential of the NER model in providing explanatory feedback based on learning principles in the previous research [3], we focus on the domain-specific NER model scheme. In the educational domain, researchers have conducted NER for automatic text assessment [13]. However, the use of the NER model is rarely used in feedback generation. Therefore, our approach aims to employ a NE recognition model to highlight the NEs within tutors' responses, which can be used to create templated explanatory feedback to tutors to increase tutor learning.

3. Method

3.1. Dataset

The dataset consisted of tutor responses to constructed-response questions within the Giving Effective Praise lesson, comprising a total of 65 volunteer tutors. The tutor demographics are: 52% White, 18% Asian, 52% male, and slightly more than half are reportedly 50 years of age or older. The development of *Giving Effective Praise* involved collaboration between the tutoring organization's director and researchers to ensure accurate operationalization of effective praise strategies within the tutoring environment. Lesson scenarios were chosen to enhance face validity by aligning with typical situations encountered by tutors in the field [3]. *Giving Effective Praise* aims to support tutors with increasing student motivation by providing effective praise, identifying its key features, and employing strategies to deliver praise and feedback. In accordance with the lesson's construct, *effective* praise should be: (1) sincere, earned, and truthful; (2) specific by giving details of a student's strengths; (3) immediate, with praise given right after the student's action; (4) authentic, avoiding repetitive phrases like "great job" which diminishes meaning and becomes predictable, and (5) focused on the learning process rather than innate ability [3].

Based on characteristics of praise types in the literature, tutor praise statements can be categorized into three different types: effort-based (*Effort*), outcomebased (*Outcome*), and person-based (*Person*). Effort-based praise is a researchshown productive praise type, focusing on the learning process (e.g., *"I like how you worked hard to..."*). Outcome-based praise showcases student's achievements, such as getting an A on an assignment or getting a problem correct, and is often, but not always, associated with unproductive praise (e.g., *"Great job!"*). Person-based praise suggests student's success is caused from fixed qualities outside of student's control (e.g., *"You are so talented."*) and, similar to outcome-based praise, is often associated with unproductive praise [14].

The *Giving Effective Praise* dataset contains 129 tutor responses categorized by praise type. Because only one person-based praise statement (i.e., *"You are very smart"*) was identified from the dataset, person-based praise was not included in the analysis. It should be noted that a tutor's response can include more than one praise type. For example, the statement, *"Great job! I like how you worked hard on completing that task,"* encompasses outcome- and effortbased praise statements. Similarly, a tutor response may not contain any praise types such as, *"Let's work together."*

3.2. RQ 1: Binary Classification for Corrective Feedback

To accurately identify different types of praise, we aim to conduct multi-label classification. Relying on the praise framework proposed by [3]. we recruited an educational expert to annotate each type of praise for the tutor's responses in a binary form. The distribution of annotated praise in tutor responses are as follows: 52 responses contained effort-based praise only; 29 responses contained both effort- and outcome-based praise; 26 contained outcome-based praise only; and, the remaining 22 responses lacked any mention of neither effort or outcomebased praise. Then, we trained the annotated responses on classifiers. Inspired by the effectiveness of the BERT model on the educational classification tasks such as tutoring dialogue classification [15, 16], we employed the BERT model to identify each type of praise in the tutors' responses. To train and evaluate the BERT model, we randomly split the dataset (i.e., annotated tutor responses) into training, validation, and testing set in the ratio of 70%, 10%, 20%, respectively, as suggested by [17]. The classification performance of the BERT model was measured by accuracy and F1 score.

3.3. RQ 2: Named Entity Recognition to Generate Explanatory Feedback

In order to generate explanatory feedback to tutors, firstly, the categorization of relevant parts of responses need to be identified through use of NEs. We refer to the annotation scheme by Thomas *et al.* [3] introduced previously and annotate the NEs representing attributes associated with *Effort* and *Outcome*, for 129 tutor responses. In line with the NER annotation in previous works [6, 18], we apply the same BIO-tagging scheme to this present work, that is, **B** represents the beginning position of the NE in the text, **I** represents the inside position of the NE in the text, and **O** represents outside of NE. For example, when annotating praise NEs for a tutor's praise *"You are doing a great job"*, the word *"great"* is identified as the beginning (i.e., $\mathbf{B_{out}}$ of the NE *Outcome* and *"job"* is identified inside (i.e., $\mathbf{I_{out}}$) of the NE. The remaining text in the response is identified as the outside (i.e., **O**) of the NE. After annotating the NEs for each tutor's response, we employed the BERT model to identify the NE from the tutors' responses. The dataset (annotated with NEs) was also divided into training, validation, and testing set in the ratio of 70%:10%:20%, respectively. The statistics of NER annotation data are shown in Table 1 which presents **O** as the major tag in our dataset. Informed by the previous study [18], predicting **O** would not enhance the evaluation score of the NER model, our study also did not take accurate predictions on **O** when calculating the performance score. To measure the NER model performance, we used the F1 score in line with the recent works on NER task [6, 18].

Table 1
Distribution of named entities for each dataset by praise types.

	% Annotation (B/I/O)				
	O	**B-Outcome**	**I-Outcome**	**B-Effort**	**I-Effort**
Full	2380 (76.5%)	53 (1.7%)	114 (3.7%)	80 (2.6%)	484 (15.6%)
Training	1661 (76.5%)	38 (1.8%)	75 (3.5%)	58 (2.7%)	338 (15.6%)
Validation	226 (75.6%)	6 (2.0%)	19 (6.4%)	6 (2.0%)	42 (14.0%)
Testing	493 (76.8%)	9 (1.4%)	20 (3.1%)	16 (2.5%)	104 (16.2%)

4. Results

4.1. RQ1: Identifying the correct type of praise

First, a multi-label classification was implemented by using the case-sensitive BERT base model[1] [19] to identify the effective type (i.e., *Effort* and *Outcome*) of praise from tutor responses. To minimize the potential impact of random variation, the model was trained on 10 different random seeds and performance was evaluated using the classification of identifying each type of praise. Table 2 illustrated the effectiveness of the BERT model in accurately tagging *Effort*, demonstrating notably high performance with an average classification accuracy of 0.731 and F1 score of 0.811. The results indicated that the BERT model could effectively tag *Effort*, which could further help the provision of corrective feedback to inform the novice tutors on providing effort-based praise. However, the BERT model's performance in tagging *Outcome* was less successful. The average F1 score for recognizing *Outcome* was 0.350 with a standard deviation of 0.235. As described the distribution of annotated praise in Section 3.2, the number of tutor responses tagging *Outcome* might be inadequate for the BERT model to identify the responses containing *Outcome* accurately. Additionally, the standard deviation of classification performance for tagging *Outcome* (SD = 0.235) was four times larger compared to the standard deviation of tagging *Effort* (SD = 0.046). The possible explanation for this could be the inadequate number of *Outcome* instances in the test set. Thus, future studies should annotate more tutor responses labeling *Outcome* to improve the BERT model's performance.

Table 2
Classification performance of BERT model in identifying praise type. The results show the average performance taken from ten random seeds. Standard errors from these experiments are indicated with subscripts.

Praise type	**Accuracy**	**F1 Score**
Effort	$0.731_{0.077}$	$0.811_{0.046}$
Outcome	$0.596_{0.089}$	$0.350_{0.235}$

[1] https://huggingface.co/bert-base-cased

4.2. RQ 2: Identifying and labeling praise statements in tutor responses

The BERT model was employed in using the NER approach. To mitigate random variation, 10 different random seeds were used to evaluate performance of the NER model, ensuring reliable estimations of the model's performance. The average F1 score of the model was 0.202 and the standard deviation was 0.039, with the model effectively identifying certain praise entities. Table 3 presented examples of tutor responses with labeled utterances associated with the corresponding NE, displaying *Effort*, (highlighted in blue) and *Outcome* (highlighted in red). *Case 1* showed that the model could accurately identify the location of the praise in the text (i.e., the text highlighted in blue) and predict the accurate entity type (i.e., *Effort*). Then, the model failed to annotate the NEs for some responses (e.g., *Case 2* in Table 3). It should be noted that the classification performance of the NER model still had space to improve the performance. One of the major reasons was that the annotated dataset was limited or low-resourced [16, 18]. The model might not have a sufficient dataset to train and test the model performance. In Section 5.2, we summarized two solutions to enhance the NER model's performance: *i)* data augmentation approaches; *ii)* and AUC Maximization approaches.

Table 3
Examples of tutor responses from the test dataset, along with named entity prediction. **True** and **Pred** stand for the true and predicted named entity, respectively. Notice in *Case 2*, the failure of the predicted model to annotate the statement associated with *Outcome*, possibly attributed to a limited or low-resourced dataset.

		Examples of Tutor Responses	**Named Entity**	**Prediction Accuracy**
Case 1	**True**	*Good job working through this and trying some different approaches.*	*Effort*	Accurate
	Pred	*Good job working through this and trying some different approaches.*	*Effort*	
Case 2	**True**	*Try your best to focus on the next step, you're already doing great so far.*	*Outcome*	Inaccurate
	Pred	*Try your best to focus on the next step, you're already doing great so far.*	*None*	
Case 3	**True**	*You did it, you did well, you got the right answer and you stuck with it, I'm proud of what you have done. Good job.*	*Outcome* *Effort*	Partially Accurate
	Pred	*You did it, you did well, you got the right answer and you stuck with it, I'm proud of what you have done. Good job.*	*Outcome* *Effort*	
Case 4	**True**	*I am glad you asked for help today. We can do this homework together.*	None	Accurate
	Pred	*I am glad you asked for help today. We can do this homework together.*	None	

Through evaluating and analyzing the results of NER, we noted that there is a need for a more nuanced measure that can acknowledge: partial overlap (e.g., *Case 3*); and true negative prediction results (e.g., *Case 4*). In Table 3, *Case 3*, the model's prediction exhibits a degree of accuracy (i.e., *"Good job"* tagged as *Outcome*, *"you stuck with it"* tagged as *Effort*) but lacks complete correctness (i.e., *"I'm proud of what you have done"* mislabeled as *Effort*). Nevertheless, the former accurate labeling of NEs can still be used for guiding tutors in providing effective praise. Thus, the prediction for *Case 3* is deemed partially accurate. Future research endeavors should focus on developing a measure to calculate partial accuracy, such as computing the intersection over union of the number of tokens in the predicted text and the desired text. Additionally, the NER model could also make true negative predictions where the tutor response did not contain any praise entities and was identified as having none of NEs (i.e., *Case 4* in Table 3). As discussed in Section 3.3, there lacked accurate predictions on the **O** tag when calculating the classification performance score. However, it is also important to identify the tutor's responses that only contain the **O** tag (i.e., none of the NEs) since it could indicate that the tutors might not understand how to deliver correct praise. In Table 3, the tutor response in *Case 4* did not contain any praise entities and this response was not related to any type of praise. The NER model could successfully identify that the response did not contain any type of praise. Based on the model prediction, feedback can be generated to guide tutors on providing praise that corresponds with the *Effort* and *Outcome* named entities.

5. Discussion and Conclusion

The construction of automatic short answer grading with the capability of providing explanatory feedback is a longstanding task towards delivering timely, specific, and personalized feedback to learners. This study employed large language models to facilitate the provision of corrective and explanatory feedback to tutors, with the main findings summarized in two folds: (1) Large language models (e.g., BERT) have the potential to identify the effort-based praise, which can be used to provide corrective feedback to novice tutors on the appropriate use of effort-based praise to students. (2) Large language models-facilitated named entity recognition (NER) can highlight the key terms associated with praise types from tutors' responses. The highlighted terms can then be integrated into template-based feedback, which can provide real-time explanatory feedback to tutors to enhance tutor learning.

5.1. Implications

Incorporation of a binary classifier can provide automatic corrective feedback. The developed classifier can be used to determine the correctness of novice tutors in providing different types of praise. The predicted classifier results can be further integrated into the provision of corrective feedback, which is essential in the learning process since corrective feedback can assist the feedback recipients in identifying errors and enhancing understanding [20]. Through the integration of the classifier within the system, we aim to provide automatic corrective feedback to tutors in dispensing various forms of praise.

Providing automatic templated feedback enhances tutor learning. To better facilitate the provision of corrective feedback, this study further investigated the potential of NER in identifying the words within the tutors' responses that correspond to the correct types of praise (i.e., *Effort* and *Outcome*). The words identified as correct praise in the tutors' responses can be integrated into a system of providing explanatory feedback. Figure 1, within the Section of Introduction, illustrates an example of providing a tutor templated explanatory feedback using an integrated NER interface. Referencing the interface, when a tutor composes praise that includes effort- and/or outcome-based praise, the system will label the *Effort* and/or *Outcome* NEs and further provide explanatory feedback to the tutor. As informed by the suggestions of effective feedback [20], incorporating explanations into feedback can help the tutor better understand the lesson objectives and content. To this end, we believe that integrating the NER model into our system could support the tutor's learning process.

5.2. Limitations and Future Work

Managing low confidence prediction using the feedback interface. The confidence level of the model's predictions is a critical aspect to consider in real-world applications. The model confidence level could affect people's belief in the model's accuracy [21]. Thus, when the model presents low confidence in predicting an instance, it poses a challenge. In such situations, it would be beneficial to design the feedback interface that presents the uncertainties to the learner. For example, when a model's confidence level on a prediction is below a certain threshold, our template-based feedback could provide hedged responses, such as *"Saying "you are committed" might be an example of praising effort. Do you want to explain your reasoning?"*. This approach not only helps uphold the credibility of the system but also invites learners to engage critically with the predictions. Future work entails providing hedged feedback responses offering learners the opportunity to explain their reasoning, along with other strategies for effectively managing low confidence predictions in the feedback interface.

Enhancing the evaluation metrics for NER. As indicated by *Case 3* and *Case 4* in Table 3, respectively, the mode's prediction may demonstrate partial correctness, for incidences where the predicted text only partially matches the desired text or, true negative predictions, where the tutor's response is accurately predicted to contain no praise entities. We argue that both partial correctness and the true negative predictions are useful in providing explanatory feedback and thus, both types of predictions should be credited. However, the traditional NER measure (F1 score) might not fully account for partial correctness and true negative predictions. Therefore, future work should explore the development of measures, such as the degree of dissimilarity between sets via calculation of intersection over union [22] to account for these cases, thereby leading to a more comprehensive evaluation of the model's performance.

Improving NER performance through data augmentation. By examining the model of NER for identifying the praise entity from the lesson of *Give effective praise*, we found that our annotated NEs might not be sufficient to train the model, which is under a low-resource data scenario [18]. To address this issue, we aim to collect more real-world data and explore widely-used data augmentation approaches (e.g., oversampling and synonyms replacement) [23] and ChatGPT-generated training instances [24] to improve NER model performance.

Robust models for enhancing the performance of NER. An alternative solution to low-resource dataset is to employ robust machine learning models. As indicated in the statistic of our dataset (Table 1), more than 70% of annotations were annotated as the **O** tag, which was highly imbalanced. To achieve satisfactory performance under the low-resource and imbalance settings,

[18] proposed to use AUC Maximization approaches for the NER task in the biomedical field, which effectively overcome challenges among low-resources and imbalanced class distribution. Thus, we aim to further examine the efficacy of AUC Maximization approaches on recognizing the praise entities.

Generalizability across tutor lessons. Our ultimate goal is to provide automatic feedback to all novice tutors who participate in our training sessions and assist them to understand the effective ways to teach students. Thus, a qualified tutor should be able to comprehend all the training lessons. Though our study examined the potentials of labeling tutor responses and providing explanatory feedback for *Giving Effective Praise* lessons, it is necessary to investigate our proposed methods in other lessons such as *Responding to Student's Errors* and *Learning What Students Know* discussed in previous work [3].

6. Acknowledgments.

This work is supported with funding from the Richard King Mellon Foundation (Grant #10851) and the Heinz Endowments (E6291). Any opinions, findings, and conclusions expressed in this material are those of the authors. We would like to thank Dr. Ralph Abboud for his guidance and recommendations regarding the use of large language models and the application of named entity recognition.

References

[1] M. A. Kraft, G. Falken, A blueprint for scaling tutoring and mentoring across public schools, AERA Open 7 (2021) 1–21. URL: https://journals.sagepub.com/doi/full/10.1177/23328584211042858.

[2] V. Q. Andre Joshua Nickow, Philip Oreopoulos, The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence, 2020. URL: http://www.edworkingpapers.com/ai20-267.

[3] D. Thomas, X. Yang, S. Gupta, A. Adeniran, E. Mclaughlin, K. Koedinger, When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors, in: LAK23: 13th International Learning Analytics and Knowledge Conference, 2023, pp. 250–261.

[4] M. Thompson, K. Owho-Ovuakporie, K. Robinson, Y. J. Kim, R. Slama, J. Reich, Teacher moments: A digital simulation for preservice teachers to approximate parent–teacher conversations, Journal of Digital Learning in Teacher Education 35 (2019) 144–164.

[5] K. R. Koedinger, P. F. Carvalho, R. Liu, E. A. McLaughlin, An astonishing regularity in student learning rate, Proceedings of the National Academy of Sciences 120 (2023) e2221311120.

[6] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 50–70.

[7] M. Zhang, S. Baral, N. Heffernan, A. Lan, Automatic short math answer grading via in-context meta-learning, arXiv preprint arXiv:2205.15219 (2022).

[8] J. Hattie, H. Timperley, The power of feedback-review of educational research, American Education Research Association and SAGE (2011) 86.

[9] T. Ryan, M. Henderson, K. Ryan, G. Kennedy, Designing learner-centred text-based feedback: a rapid review and qualitative synthesis, Assessment & Evaluation in Higher Education 46 (2021) 894–912.

[10] J. Lin, W. Dai, L.-A. Lim, Y.-S. Tsai, R. F. Mello, H. Khosravi, D. Gasevic, G. Chen, Learner-centred analytics of feedback content in higher education, in: LAK23: 13th International Learning Analytics and Knowledge Conference, 2023, pp. 100–110.

[11] D. R. Chine, P. Chhabra, A. Adeniran, S. Gupta, K. R. Koedinger, Development of scenario-based mentor lessons: an iterative design process for training at scale, in: Proceedings of the Ninth ACM Conference on Learning@ Scale, 2022, pp. 469–471.

[12] V. Aleven, O. Popescu, K. R. Koedinger, Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor, in: Proceedings of Artificial Intelligence in Education, 2001, pp. 246–255.

[13] C. Walter, Increasing teachers' trust in automatic text assessment through named-entity recognition, in: International Conference on Artificial Intelligence in Education, Springer, 2022, pp. 191–194.

[14] M. L. Kamins, C. S. Dweck, Person versus process praise and criticism: implications for contingent self-worth and coping., Developmental psychology 35 (1999) 835.

[15] J. Lin, S. Singh, L. Sha, W. Tan, D. Lang, D. Gašević, G. Chen, Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues, Future Generation Computer Systems 127 (2022) 194–207.

[16] J. Lin, W. Tan, N. D. Nguyen, D. Lang, L. Du, W. Buntine, R. Beare, G. Chen, D. Gašević, Robust educational dialogue act classifiers with low-resource and imbalanced datasets, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 114–125.

[17] A. Gholamy, V. Kreinovich, O. Kosheleva, Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation (2018).

[18] N. D. Nguyen, W. Tan, L. Du, W. Buntine, R. Beare, C. Chen, Auc maximization for low-resource named entity recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13389–13399.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the NAACL-HLT, Volume 1, 2019, pp. 4171–4186.

[20] A. C. Butler, N. Godbole, E. J. Marsh, Explanation feedback is better than correct answer feedback for promoting transfer of learning., Journal of Educational Psychology 105 (2013) 290.

[21] A. Rechkemmer, M. Yin, When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models, in: Proceedings of the 2022 CHI conference, 2022, pp. 1–14.

[22] M. Levandowsky, D. Winter, Distance between sets, Nature 234 (1971) 34–35.

[23] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, in: Findings of the ACL: ACL-IJCNLP 2021, 2021, pp. 968–988.

[24] D. Thomas, S. Gupta, K. Koedinger, Comparative analysis of learnersourced human-graded and ai-generated responses for autograding online tutor lessons, in: Artificial Intelligence in Education. 24th International Conference, AIED 2023, Tokyo, Japan July 3–7, 2023, Springer, 2023.