TBA at BEA 2025 Shared Task: Transfer-Learning from DARE-TIES Merged Models for the Pedagogical Ability Assessment of LLM-Powered Math Tutors

Sebastian Gombert[1], Fabian Zehner[1,2], and Hendrik Drachsler[1,3,4,5]
[1]DIPF | Leibniz Institute for Research and Information in Education
[2]Centre for International Student Assessment (ZIB)
[3]studiumdigitale & [4]Computer Science Department, Goethe University Frankfurt
[5]Department of Online Learning and Instruction, Open University NL Heerlen
{s.gombert,f.zehner,h.drachsler}@dipf.de

Abstract

This paper presents our contribution to the *BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-Powered Tutors*. The objective of this shared task was to assess the quality of conversational feedback provided by LLM-based math tutors to students regarding four facets: whether the tutors 1) identified mistakes, 2) identified the mistake's location, 3) provided guidance, and whether they 4) provided actionable feedback. To leverage information across all four labels, we approached the problem with *FLAN-T5* models, which we fit for this task using a multi-step pipeline involving regular fine-tuning as well as model merging using the *DARE-TIES* algorithm. We can demonstrate that our pipeline is beneficial to overall model performance compared to regular fine-tuning. With results on the test set ranging from 52.1 to 68.6 in F1 scores and 62.2% to 87.4% in accuracy, our best models placed 11th of 44 teams in Track 1, 8th of 31 teams in Track 2, 11th of 35 teams in Track 3, and 9th of 30 teams in Track 4. Notably, the classifiers' recall was relatively poor for underrepresented classes, indicating even greater potential for the employed methodology.

1 Introduction

Large language models, such as the ones from the *GPT* (Radford et al., 2018) or *Llama* (Grattafiori et al., 2024) families, have demonstrated remarkable capabilities in generating a wide range of textual content. This has resulted in their quick adoption in the educational space, where they are used for diverse purposes, such as assessing student-generated content, providing feedback and guidance, or generating exercise questions, among others (Wang et al., 2024). They have also been incorporated into intelligent tutoring systems, combining multiple of these features and capabilities into a single application (Wang et al., 2025). However, a core problem with these models is that they do not guarantee accurate, practical, or focused output (Xu et al., 2025). As generation is handled through a combination of autoregression and probabilistic sampling, it cannot be guaranteed that each production of a given model is purposeful and correct.

Importantly, this can be a severe problem in educational settings. In the European Union, the EU AI Act (European Parliament and Council of the European Union, 2024) classifies AI-based systems in an educational context as high risk. What if a tutor provides a learner with incorrect feedback because of a chain of unfortunate random sampling during the corresponding generation process? What if specific prompt characteristics affect output quality systematically, disadvantaging certain learner groups (Hofmann et al., 2024; Salikutluk et al., 2024)? What if a given feedback text is not actionable, and a learner is left with more questions? One possibility to address a few, albeit not all, such problems is to deploy models tailored explicitly for policing the output of a given model. What is already an established practice with commercial models, where, for example, the generation of toxic content is policed, also has enormous potential for the educational sector, where policing by educational criteria is required.

The *BEA 2025 Shared Task on the Pedagogical Ability Assessment of AI-powered Tutors* (Kochmar et al., 2025) explores this idea for a narrow use case where the output of LLM tutors when assisting students with simple arithmetic problems is assessed. In particular, the goal is to assess communication records between students attempting to solve simple math problems and LLMs that assist them as tutors. The communication records are classified according to whether the LLM tutor identified student mistakes, recognised the mistake location, provided guidance, and whether the provided guidance is actionable. As highlighted by Holmes et al. (2022), ethical considerations in AI in education, despite their crucial impact, are of-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1173–1179

ten not prioritized. The present shared task, therefore, offers the opportunity to address a subset of vulnerabilities that could otherwise lead to ethical breaches.

Our submissions to this shared task are based on variants of *FLAN-T5-xl* (Chung et al., 2024) that underwent multiple steps of task-wise fine-tuning and model merging via *DARE-TIES* (Yu et al., 2024). On the shared task leaderboard, based on macro F1, our systems rank 11th out of 44 teams in Track 1, 8th out of 31 teams in Track 2, 11th out of 35 teams in Track 3, and 9th out of 30 teams in Track 4.

2 Background

2.1 Pedagogical Ability Assessment and Pedagogical Alignment of LLMs

Using conversational agents in education is not a novel idea; it has been explored for several years, e.g., in the form of tutors or assistants (Wollny et al., 2021). However, following the release of ChatGPT in 2022 and the resulting surge in research on conversational large language models, interest in this topic has increased (e.g., Pal Chowdhury et al. 2024). Although large language models have demonstrated remarkable capabilities and possess significant potential for educational use cases, their probabilistic nature also presents challenges that must be addressed before these models can be safely deployed in pedagogical contexts. Older conversational agents are often based on rules, fuzzy matching against a search space of expected inputs, information retrieval, and pre-defined answers and dialogue scripts (Wollny et al., 2021). This makes it easy to pedagogically align them since all output they can generate is pre-defined to a certain degree, or can, in the case of information retrieval, at least be curated.

For LLMs, this is not the case. While they can answer and react more dynamically and are better suited to providing deeply individualised feedback since they can deal with unforeseen inputs posing problems to more traditional chatbot designs, achieving alignment with pedagogical criteria is harder for these models. On the one hand, this is due to the well-known hallucination problem (Xu et al., 2024). On the other hand, even when a model does not hallucinate and generates correct output, this does not necessarily imply that what is generated follows good pedagogical practice[1], since these models were never trained with the same in mind.

For this reason, there has been increased interest in studying and improving pedagogical alignment for large language models (LLMs). Sonkar et al. (2024) compared *supervised fine-tuning* (SFT) and *learning from human preference* (LHP; Christiano et al., 2017) as training approaches for achieving pedagogical alignment for LLMs, with the latter approach achieving overall better downstream results. Dai et al. (2023) assessed feedback generated by *ChatGPT* using the well-known Hattie framework (Hattie and Timperley, 2007) and concluded that feedback generated by the model was overall more detailed compared to a human gold standard with an overall high agreement in terms of what exact elements from Hattie's framework were represented in the feedback texts. Meyer et al. (2024) found increased motivation and performance on a revision task as well as more positive feelings through LLM-generated feedback compared to no feedback. Tack et al. (2023) hosted a shared task that benchmarked the overall ability of LLMs to act as pedagogically sound tutors when fine-tuned or prompt-tuned for the same purpose. Maurya et al. (2025) introduced a framework to rate the qualities of LLM-based tutors using eight different dimensions, each rated on a three-level scale. Four of these dimensions form the basis for the dataset used in this shared task.

2.2 Model Merging

Model merging refers to a growing set of recently developed methods that combine multiple fine-tuned models into a single one, sharing all their strengths. The core idea behind model merging lies in what is called *task arithmetics* (Ilharco et al., 2023). If we interpret the set of all parameters of a given LLM as one long vector, we can define such vectors for both a pre-trained model (θ_0) as well as task-specific fine-tuned versions of the same (θ_t). By subtracting the initial vector θ_0 from the fine-tuned vector θ_t, we gain the so-called task vector θ'_t representing the knowledge a model acquired during a specific fine-tuning instance t. We can then create *merges* by combining the resulting task vectors in various ways and adding the resulting vector to the original pre-trained model.

A naive approach to recombining task vectors is to calculate a weighted mean of them. How-

[1] https://benchmarks.ai-for-education.org/; accessed on 2025-05-21

ever, this comes with several problems that mainly stem from the nature of stochastic gradient descent, which can lead to different fine-tuned models converging to distinct local minima in the parameter space. While two datasets a given model might be fine-tuned with might be highly related, implying that the respective fine-tuned models will have learned similar underlying functions by having adjusted the weights of a given model similarly, it is by no means specific that these learned representations will be localized in the identical or corresponding parameters (e.g., polysemanticity). Colloquially speaking, two different fine-tuning instances might store different knowledge in the same parameters, resulting in parameter interference and decreased downstream performance.

For this reason, algorithms such as TIES (Yadav et al., 2023) and DARE-TIES (Yu et al., 2024), which improves on the previous, have been developed. While TIES aims at fitting a transformation matrix that acts as a translation layer between two different fine-tuning instances and strives to identify correspondences between the internal representations of both task vectors, *DARE-TIES* combines this with directional averaging and heavy pruning of the individual task vectors to minimize interference during merging. The method can be denoted in the following way, with t being a given task from the set of all tasks used for a particular merge T:

$$\theta^t_{\text{DARE}} = \text{DARE}(\theta_t, \theta_0), \text{ for } t \in T \quad (1)$$

$$\theta_M = \theta_0 + \lambda \sum_{t}^{\text{T}} \text{TIES}(\theta^t_{\text{DARE}}, \theta_0) \quad (2)$$

For the exact implementation of *DARE* and *TIES*, see Yu et al. (2024) respectively Yadav et al. (2023).

3 Method

3.1 Dataset

The dataset used in the BEA 2025 shared task contains conversation histories between an LLM, functioning as a tutor, and a corresponding human student. Each conversation history involves a simple math problem and reflects a corresponding conversation between a student and an LLM. While the training set includes 300 such conversation histories, the test set contains 191. For each conversation, there are up to seven different final responses that were each generated by a different model, such as *GPT-4* or *Mistral* (Jiang et al., 2023) in response to the provided history. Moreover, human responses from both expert and novice tutors are provided for each conversation. For each of the responses, four of the overall eight dimensions from the framework introduced by Maurya et al. (2025) are annotated using a three-level scale (no, to some extent, yes). The dimensions are:

- **Mistake Identification:** Is the LLM able to identify the learner mistake in its response?
- **Mistake location:** Is the location of a given mistake provided in a response?
- **Providing guidance:** Does the model provide appropriate guidance on how to solve the mistake?
- **Actionability:** Is what the model answers actionable?

Each of the four dimensions corresponds to an individual evaluation track of the shared task. Due to time constraints and its conceptually distinct goal, we disregarded the fifth track that was concerned with identifying the generating LLM.

3.2 System Development

Our system uses *FLAN-T5* (Chung et al., 2024) models to model classification as a sequence-to-sequence task, where the model is trained to generate an output sequence containing the correct label for a given input, which includes the full conversation context, including all utterances of both student and tutor in a given conversation. Concretely, a model receives the following input for a given datapoint x and assessment dimension d (*mistake location*, *mistake identification*, ...), with h_x denoting the provided conversation history and r_x the provided tutor response:

$$I(x, d) = d : \text{history} : h_x \text{ response} : r_x \quad (3)$$

We did not make any structural modifications to the models themselves and used the standard implementations provided by the *Huggingface Transformers* framework (Wolf et al., 2020). The procedure we used to fit these models, however, distinguishes this work from other use cases of *FLAN-T5* for classification.

It involves three steps, as depicted in Figure 1. In a first step, the given *FLAN-T5* models were fine-tuned for three epochs, one model for each of the

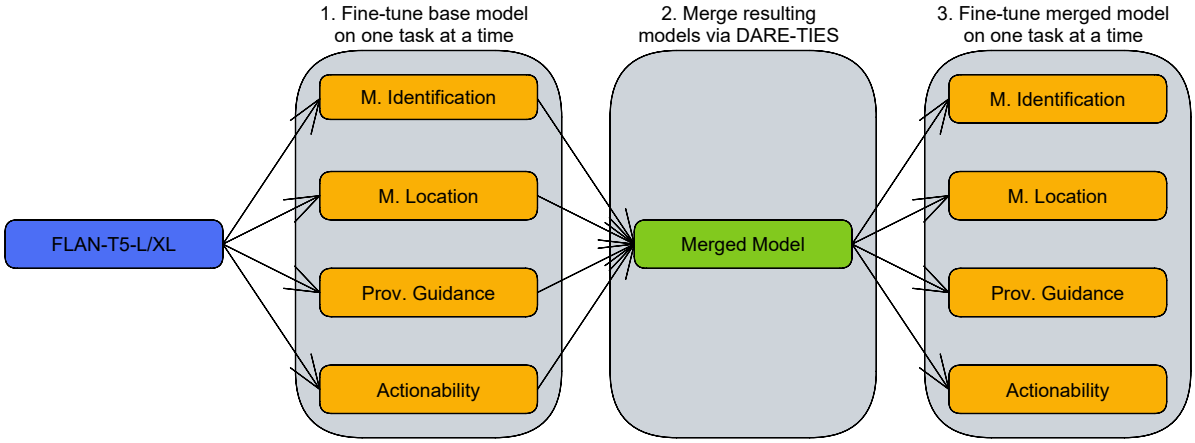

Figure 1: This figure depicts the overall training process we used during both our pre-experiments as well as for the final submission. First, *FLAN-T5* models are fine-tuned for three epochs for one dimension at a time. The resulting models are then merged using *DARE-TIES*. Lastly, the merged model serves as the basis for another round of task-specific fine-tuning, yielding the four final models.

four assessment dimensions, resulting in four individual models for *mistake identification*, *mistake location*, *provision of guidance*, and *actionability*. Fine-tuning was conducted using *Adagrad* as optimiser, a learning rate of *3e-4*, and a batch size of 4. These four models were then merged using the *DARE-TIES* (Yu et al., 2024) algorithm implementation provided by *Mergekit* (Goddard et al., 2024), with each model being uniformly weighted ($\lambda = 0.25$). The resulting model was then used as a basis for another round of fine-tuning, where we fine-tuned the merged model for each task individually again, resulting in another quartet of task-specific models.

The rationale behind this approach is the inherent interconnectedness between the four individual dimensions. We assumed that, for example, a mistake location can only reasonably be provided if a mistake is identified. Moreover, appropriate guidance can also be provided only if a mistake is identified. Then, only if guidance was provided at all can this guidance be actionable. Consequently, we assume that some of the parameters within models fine-tuned for one of these specific tasks likely encode information beneficial to the others.

DARE-TIES (Yu et al., 2024) as an algorithm enables us to exploit this property by merging multiple fine-tuned models into a single one that inherits the capabilities of all the used base models, with the possibility of even improving performance in some cases where the individual tasks are complementary to each other. This is achieved through the alignment and directional merging of the specific model parameters.

Variant	MI	ML	PG	AC
Pre-merge	**89.20**	69.95	71.97	81.34
Merged	84.46	77.17	77.42	73.23
Post-merge	88.48	**82.15**	**82.85**	**88.49**

Table 1: Macro F1 scores for the three model stages in our pre-experiments. MI = Mistake Identification. ML = Mistake Location. PG = Providing Guidance. AC = Actionability.

Initially, we had assumed that the model resulting from the *DARE-TIES* merge would already be slightly stronger for each assessment dimension than the dimension-specific models. However, in our pre-experiments, we could not completely confirm this hypothesis. Using a 5x5 cross-validation setup with the complete training set, we fine-tuned and then merged *FLAN-T5-base* (Chung et al., 2024) models, with the result that the merged models showed a weaker performance for *mistake identification* and *actionability* than the dimension-specific models from which they were created (see Table 1), with an improved performance for *mistake location* and *providing guidance*.

For this reason, as a next step, we explored whether the resulting merged model would at least function as a reasonable basis for fine-tuning a next generation of dimension-specific models. As Table 1 shows, this was indeed the case, and the resulting dimension-specific models showed an improved performance over the merged variants as well as the dimension-specific models fine-tuned

Metric	MI	ML	PG	AC
Macro F1	68.58	54.90	52.12	66.71
Rank	11/44	8/31	11/35	9/30
Accuracy	87.40	73.24	66.52	73.24
Rank	5/44	6/31	5/35	4/30

Table 2: Results from the official shared task leaderboard. Rank indicates the rank our submissions achieved for the specific dimension and metric. MI = Mistake Identification. ML = Mistake Location. PG = Providing Guidance. AC = Actionability.

from *FLAN-T5-base*, except for *mistake location*. For this reason, we went with this procedure for our final submissions.

With the post-merge fine-tuning stage adding an epoch of training, performance gains may also have resulted from improved task-specific fitting rather than the merging process itself. While tentative experiments did not provide evidence for this, we did not rule this out through a systematic experiment.

4 Shared Task Submission and Evaluation

Following the intuition behind the scaling law that, on average, larger models show an improved downstream performance compared to smaller models when trained on the same data (Kaplan et al., 2020), we replicated our setup with *FLAN-T5-xl* (Chung et al., 2024) for the shared task submission. Again, we first fine-tuned dimension-specific models for all four dimensions for three epochs each, then merged them using *DARE-TIES* (Yu et al., 2024), and then used the resulting model as a basis for fine-tuning for another epoch to acquire again dimension-specific models (as depicted in Figure 1. Since, in our pre-experiments, the post-merge models for *mistake identification* were slightly outperformed by the pre-merge ones, we submitted results from both for the final task (since up to five submissions were allowed per dimension). Here, contrary to our pre-experiments, the post-merged version came out on top.

In the context of the shared task, the resulting models could all achieve upper mid-table results, going by *Macro F1*. For *Mistake Identification*, we placed 11th of 44 teams. For *Mistake Location*, we placed 8th of 31 teams. For *Providing Guidance*, we placed 11th of 35 teams. For *Actionability*, we placed 9th of 30 teams. For *Accuracy*, which served as a secondary evaluation metric, our models were among the best submissions in the shared task. Here, we placed 5th, 6th, 5th and 4th for the respective dimensions.

These results suggest that our approach was overall highly successful in modelling the different dimensions, but, in particular, fell short for the *No* category, which was comparably underrepresented in the data. We assume that techniques such as *paraphrased oversampling* (Patil et al., 2022) would likely have helped combat that overall behaviour, but were not considered by us since we implemented our solution within one week under heavy time pressure. Table 2 shows the corresponding results. Overall, the results suggest that our approach is reasonable and the use of *DARE-TIES* merging allowed us to achieve upper mid-table results, although our placements suggest, that there are certainly better solutions for the problem than what we propose in this paper.

5 Conclusion

In this paper, we presented our submission to the *BEA 2025 Shared Task on the Pedagogical Ability Assessment of AI-powered Tutors*. Our submission combines fine-tuning and *DARE-TIES* merging *FLAN-T5-xl* models. In terms of macro F1, the primary evaluation metric used for the shared task, our models could only achieve upper mid table results, which is likely due to the underrepresentation of *No* and *to some extent* cases within the training set. In terms of overall accuracy, our submissions achieve more competitive results. Our general results show that combining *DARE-TIES* merging with fine-tuning can have beneficial results on downstream performance.

Limitations

Focus on *FLAN-T5*: In this paper, we focused solely on *FLAN-T5* models while not considering other models such as *Mistral-7b* (Jiang et al., 2023). The reason behind this was mainly that our contribution was created under heavy time pressure, so we wanted to focus on making our approach work for one model family as best as we could, instead of comparing a larger range of models.

No data augmentation used: Since the provided dataset is highly imbalanced, with the *no* and *to some extent* cases being underrepresented for all four dimensions, we assume that data augmentation could have likely benefited our systems, e.g., in the form of techniques such as *paraphrased oversampling* (Patil et al., 2022). However, due to the

heavy time pressure, we decided against exploring data augmentation.

No hyperparameter search: We did not conduct a hyperparameter search but instead stuck to the standard training hyperparameters used to pretrain the *FLAN-T5* models, except for the batch size, which we reduced from the original 64 to 4 due to limited computational resources. Similarly, it is possible that performance gains in the task-specific post-merge fine-tuning stem at least partly from an additional epoch of training.

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pages 323–325. IEEE.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. OJ L 2024/1689, 12 July 2024.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C. Santos, Mercedes T. Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, and Kenneth R. Koedinger. 2022. Ethics of AI in education: towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3):504–526.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. *arXiv preprint*. ArXiv:2212.04089 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary

students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 5–15, New York, NY, USA. Association for Computing Machinery.

Annapurna P Patil, Shreekant Jere, Reshma Ram, and Shruthi Srinarasi. 2022. T5w: A paraphrasing approach to oversampling for imbalanced text classification. In *2022 IEEE international conference on electronics, computing and communication technologies (CONECCT)*, pages 1–6. IEEE.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Vildan Salikutluk, Elifnur Doğan, Isabelle Clev, and Frank Jäkel. 2024. Involving affected communities and their knowledge for bias evaluation in large language models. In *1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems*, Honolulu, Hawaii, USA.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA. Association for Computational Linguistics.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. 2025. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 510–519, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.