Journal of Science Education and Technology (2024) 33:333–348
https://doi.org/10.1007/s10956-023-10087-4

Beyond Language Barriers: Allowing Multiple Languages in Postsecondary Chemistry Classes Through Multilingual Machine Learning

Paul P. Martin[1] · **Nicole Graulich**[1]

Accepted: 16 November 2023 / Published online: 14 February 2024
© The Author(s) 2024

Abstract
Students who learn the language of instruction as an additional language represent a heterogeneous group with varying linguistic and cultural backgrounds, contributing to classroom diversity. Because of the manifold challenges these students encounter while learning the language of instruction, additional barriers arise for them when engaging in chemistry classes. Adapting teaching practices to the language skills of these students, for instance, in formative assessments, is essential to promote equity and inclusivity in chemistry learning. For this reason, novel educational practices are needed to meet each student's unique set of language capabilities, irrespective of course size. In this study, we propose and validate several approaches to allow undergraduate chemistry students who are not yet fluent in the language of instruction to complete a formative assessment in their preferred language. A technically easy-to-implement option for instructors is to use translation tools to translate students' reasoning in any language into the instructor's language. Besides, instructors could also establish multilingual machine learning models capable of automatically analyzing students' reasoning regardless of the applied language. Herein, we evaluated both opportunities by comparing the reliability of three translation tools and determining the degree to which multilingual machine learning models can simultaneously assess written arguments in different languages. The findings illustrate opportunities to apply machine learning for analyzing students' reasoning in multiple languages, demonstrating the potential of such techniques in ensuring equal access for learners of the language of instruction.

Keywords Equity · Language inclusivity · Formative assessment · Machine learning · Natural language processing

Introduction

Linguistic diversity is increasing in science subjects like chemistry, so the language of instruction does not necessarily correspond to a student's first language. Students who learn the language of instruction as an additional language face unique challenges in language acquisition, science learning, and the intersection of both (del Rosario Basterra et al., 2011; Lee & Fradd, 1998). Therefore, further barriers arise for them to actively participate in chemistry classes (Deng & Flynn, 2023).

So far, much research has focused on students who learn English as an additional language (Eng+). These Eng+ students represent a heterogeneous group with diverse linguistic backgrounds, different prior opportunities in learning English, and varying experiences in communicating in another language (Deng & Flynn, 2023; Deng et al., 2022; Flores & Smith, 2013). Although these Eng+ students may possess strong content knowledge, they may lack the fluency to fully express their knowledge in English (Deng & Flynn, 2023; Deng et al., 2022; Flores & Smith, 2013; Lyon et al., 2012; Swanson et al., 2014). This language barrier can hinder effective communication, impede the exchange of ideas, and potentially exclude valuable contributions from Eng+ students (Deng & Flynn, 2023; Deng et al., 2022). Additionally, Eng+ students may experience linguistic insecurity, which involves feelings of anxiety about their English language usage (Deng & Flynn, 2023).

In educational settings such as chemistry classes, students taught in a different language than their first language face

✉ Nicole Graulich
Nicole.Graulich@didaktik.chemie.uni-giessen.de

Paul P. Martin
Paul.Martin@didaktik.chemie.uni-giessen.de

[1] Institute of Chemistry Education, Justus-Liebig-University, Heinrich-Buff-Ring 17, 35392 Giessen, Germany

various challenges in understanding, applying, and communicating domain-specific scientific concepts as well as in constructing evidence-based arguments (Curtis & Millar, 1988; Deng et al., 2022; Lee & Fradd, 1998). For instance, Eng+ students struggled to grasp the intended meaning of terms (Solano-Flores & Trumbull, 2003), especially when the scientific and everyday meanings of these terms diverge (Lee & Orgill, 2022). Moreover, undergraduate Eng+ students in general chemistry encountered both extrinsic and intrinsic challenges in engaging in lectures, comprehending laboratory procedures, and expressing their content knowledge (Lee et al., 2020). Additionally, postsecondary Eng+ students generated the fewest causal arguments in chemistry in a recent study (Deng et al., 2022). Beyond that, linguistic or cultural references in science assessments may not be accessible to Eng+ students (Luykx et al., 2007; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003).

Collectively, this indicates that students' performance in science assessments is linked to their language proficiency, irrespective of their capacity to comprehend and apply scientific principles (Afitska & Heaton, 2019; Curtis & Millar, 1988; Deng et al., 2022; Maerten-Rivera et al., 2010; Noble et al., 2014; Solano-Flores & Trumbull, 2003; Turkan & Liu, 2012). Language learners may face additional challenges because of unique cognitive operations associated with learning an additional language (Valdés & Figueroa, 1994). For this reason, Eng+ students might solve some assessment items better in their first language and others better in English (Solano-Flores & Trumbull, 2003). Assessment items represent, thus, a unique set of linguistic challenges depending on the language in which they are administered (Lee & Orgill, 2022; Solano-Flores & Trumbull, 2003). This may establish inequities when assessing students who learn the language of instruction as an additional language in chemistry and beyond.

So, educators need to ensure equitable means for these language learners (Lee, 2005; Wolf et al., 2008). On the one hand, educators could help them overcome their language barrier by continuously providing opportunities for language development (Amano et al., 2021). On the other hand, educators could design equitable instructional settings, among others, formative assessments, that allow for a more diverse educational landscape.

Proposing Approaches for Allowing Multiple Languages in Chemistry Classes

In their study, Deng et al. (2022) found that most but not all students preferred the language of their chemistry instruction for communicating on chemistry assessments, regardless of their first language. Some language learners might, therefore, benefit from communicating scientific ideas in the language of instruction, whereas others might enhance their reasoning about chemical phenomena by using their first language. Consequently, a time- and resource-efficient approach to increasing equity for the latter students could be providing them with the opportunity to complete exercises and formative assessments in their first language (cf., Afitska & Heaton, 2019; Buxton et al., 2014; Lee, 2005). However, since an instructor is not expected to be fluent in the respective language, a translation tool such as Google Translate, DeepL Translate, or ChatGPT could be used to translate back and forth between languages. Due to recent advancements in natural language processing (NLP) and machine learning (ML), these tools made significant progress in translating human language. In addition, learners of the language of instruction already utilize such tools to a great extent (Almusharraf & Bailey, 2023), which is why we quantitatively compared their performance. By leveraging these translation tools, educators could enable equal access to educational resources for language learners, which promotes inclusivity.

Nevertheless, allowing formative assessments in multiple languages is also possible by other means. Besides translation tools, educators could also design ML-based instructional settings, which have the potential to evaluate students' reasoning in every language automatically. While developing such ML models is resource-intensive, analyzing students' responses is fully automated once the model has been constructed. However, there are fundamental concerns that ML algorithms take a majority-driven language as the benchmark against which all other data is evaluated (Cheuk, 2021; Li et al., 2023). Consequently, a majority-focused linguistic practice may be overvalued (Ha et al., 2011; Liu et al., 2016), even if scientifically non-normative ideas are expressed (Nehm et al., 2012). In contrast, responses including uncommon vocabulary are more likely to be misclassified (Ha et al., 2011; Liu et al., 2016; Maestrales et al., 2021), which may penalize learners of the language of instruction (Cheuk, 2021). So, ML-based science assessments may favor a limited range of linguistic practices deemed suitable for scientific discourse, prioritizing academic language over students' informal everyday vocabulary (Cheuk, 2021). Therefore, establishing ML-based systems can unconsciously deepen structural forms of historically grown inequalities (Cheuk, 2021; Grimm, Steegh, Çolakoğlu et al., 2023; Grimm, Steegh, Kubsch et al., 2023; Li et al., 2023).

Following this, it is essential to identify and reduce the potential biases of ML algorithms. In doing so, Wilson et al. (2023) compared human, machine analytic, and machine holistic scoring regarding their scoring severity as well as their scoring gap between Eng+ and non-Eng+ students. They found that Eng+ students received on average lower human, machine holistic, and machine analytic scores than

non-Eng+ students. Moreover, machine holistic scores were on average lower than human and machine analytic scores across both linguistic groups; however, the performance gap between Eng+ and non-Eng+ students was greater for machine analytic scoring than for human and machine holistic scoring, specifically for the most difficult items. This finding indicates that machine analytic scoring can be biased when analyzing Eng+ students' responses, while machine holistic scoring may increase equity in evaluating Eng+ students' written arguments.

A multilingual approach to developing ML models for science assessments may also acknowledge students' linguistic diversity since such models can automatically assess students' reasoning across languages. For generating corresponding multilingual training datasets, instructors can either collect multilingual data or translate monolingual data into another language with a suitable translation tool so that students can express themselves beyond language barriers.

Utilizing deep learning techniques such as large language models and deep neural networks seems especially helpful when evaluating multilingual student responses since these techniques can handle complex data and achieve high accuracy. By applying these techniques, instructors have several opportunities: They can apply either a multilingual large language model for analyzing student responses across languages or monolingual models for each language separately. For this reason, the accuracy of different multi- and monolingual large language models is key to providing reliable assessments.

Methodological Considerations

Basics to Artificial Intelligence, Machine Learning, and Deep Learning

Due to recent technological progress, algorithms of artificial intelligence (AI) have taken over activities that have been, up to now, exclusively associated with human abilities. In this way, the term AI broadly describes software that automatically performs cognitive activities such as software-based planning, problem-solving, and decision-making (Bellmann, 1978; Haugeland, 1989). In education, AI can transform teaching and learning as evident in various articles (Kubsch et al., 2023; Martin & Graulich, 2023; Zhai, Haudek et al., 2020). For example, intelligent tutoring systems can automatically evaluate unique student challenges, provide immediate feedback, and deliver tailored exercises (Deeva et al., 2021). Accordingly, these systems can reduce the workload for educators while expediting assessment procedures (Urban-Lurain et al., 2013).

ML is, in turn, a subarea of AI that deals primarily with developing algorithms that learn from data to make predictions (Bishop, 2006; Mitchell, 1997; Mohri et al., 2012). Supervised ML algorithms are trained on human-labeled data so that the desired output, known as ground truth, is already included in the training data. The algorithm can then discern the underlying patterns and predict labels for new data. Over the past 15 years, supervised ML has been increasingly applied in science education research (e.g., Deeva et al., 2021; Gerard et al., 2015; Zhai, Yin et al., 2020), for example, to detect student reasoning in formative chemistry assessments (Martin & Graulich, 2023). In this way, supervised ML has contributed to significant advancements in automatically analyzing chemistry students' reasoning (e.g., Dood et al., 2018, 2020; Haudek et al., 2019; Maestrales et al., 2021; Noyes et al., 2020; Tansomboon et al., 2017; Vitale et al., 2016; Wilson et al., 2023; Yik et al., 2021, 2023).

Furthermore, deep learning is a specific ML technique that concentrates on training deep neural networks containing multiple layers of interconnected artificial neurons, inspired by the function of the human brain (Goodfellow et al., 2016). During network training, large amounts of data are fed into the network to adjust the parameters of the neuron connections. This technique has enabled significant advancements in various fields and remains an active research area (Mathew et al., 2021). Particularly, deep learning has gained prominence in educational studies due to its ability to accurately analyze unstructured data, as stored in images (e.g., Lee et al., 2023; Zhai et al., 2022) or written language (e.g., Dood et al., 2022; Gombert et al., 2023; Martin et al., 2023; Tschisgale et al., 2023; Watts et al., 2023; Winograd, Dood, Finkenstaedt-Quinn et al., 2021; Winograd, Dood, Moon et al., 2021; Wulff et al., 2023).

Basics to Natural Language Processing and Large Language Models

To efficiently analyze human language with ML, text data needs to be preprocessed by applying suitable NLP methods. NLP enables computers to analyze human language; so far, numerous techniques are available for facilitating human–computer interaction. Over the last couple of years, large language models emerged as a state-of-the-art technique in NLP. These models can perform various language-related tasks such as question-answering or text generation (Radford et al., 2019). Large language models have been trained on massive amounts of text data to capture complex language patterns and produce contextually relevant responses.

One such cutting-edge large language model is Bidirectional Encoder Representations from Transformers (BERT). BERT is pretrained on large corpora of unlabeled text data, such as Wikipedia, to understand semantic relationships between different words in a sentence, even if they are far apart (Devlin et al., 2018). Compared to formerly applied NLP methods, which often ignore word order (Angelov, 2020; Jurafsky & Martin, 2023), BERT can capture syntactic nuances of language, implicit meanings of phrases, and

the context in which words are used (Mikolov et al., 2013; Taher Pilehvar & Camacho-Collados, 2020). Hence, BERT has achieved state-of-the-art performance on a wide range of NLP benchmarks (Devlin et al., 2018) and has become a template for many subsequent large language models. Instructors can fine-tune BERT models for domain-specific purposes such as text classification in science assessment, which is called a downstream task (Ruder, 2019).

Research Questions

In this article, we propose and validate several ML-based approaches designed to allow students who have not yet mastered the language of instruction to complete formative assessments in their preferred language. However, as instructors are not expected to be fluent in the respective language, questions of reliability arise when analyzing students' reasoning across multiple languages. We approach the following research questions (RQs) by simulating the language transition between English and German as both authors are fluent in these languages. While doing so, we focused on undergraduate students' argumentation in organic chemistry about the plausibility of competing reactions. The research questions examined herein can inform instructors how to acknowledge multiple languages in the classroom.

1. Which translation tool produces German translations of English-written scientific arguments that deep learning architectures can analyze most accurately?
2. Which large language model achieves the highest level of reliability when analyzing German-written scientific arguments?
3. To what extent does the reliability of mono- and multilingual ML approaches to analyzing English- and German-written scientific arguments differ?
4. To what extent does an augmentation of the model's training data by combining translations of different translation tools impact the ML model's reliability?

Research Context

Setting of Data Collection

The data used in this second data analysis was gathered at a private research-intensive, liberal arts university in the North-eastern United States during April and May 2021. Sixty-four undergraduate organic chemistry II students voluntarily participated in the study, receiving extra credit for completing the exercises. The age of the participants ranged from 18 to 22 years. Among the participants, 34 identified as female, 29 as male, and one as non-binary. The students majored in various subjects, including biochemistry, chemistry, biology, and chemical engineering.

The tasks on building arguments about competing reactions in organic chemistry were implemented online using Qualtrics. The two task sets (Fig. 1) were presented on two different days, with a 3-week gap between them. The four alternative reaction products were sequentially displayed to the students. Students could build more than one argument per alternative reaction product. The original English-written arguments contain on average 27 words ($SD = 12.7$, $max = 138$, $min = 5$); the German translations comprise, in turn, on average 24 words ($SD = 11.6$, $max = 127$, $min = 4$). A total of 1108 arguments were collected. For

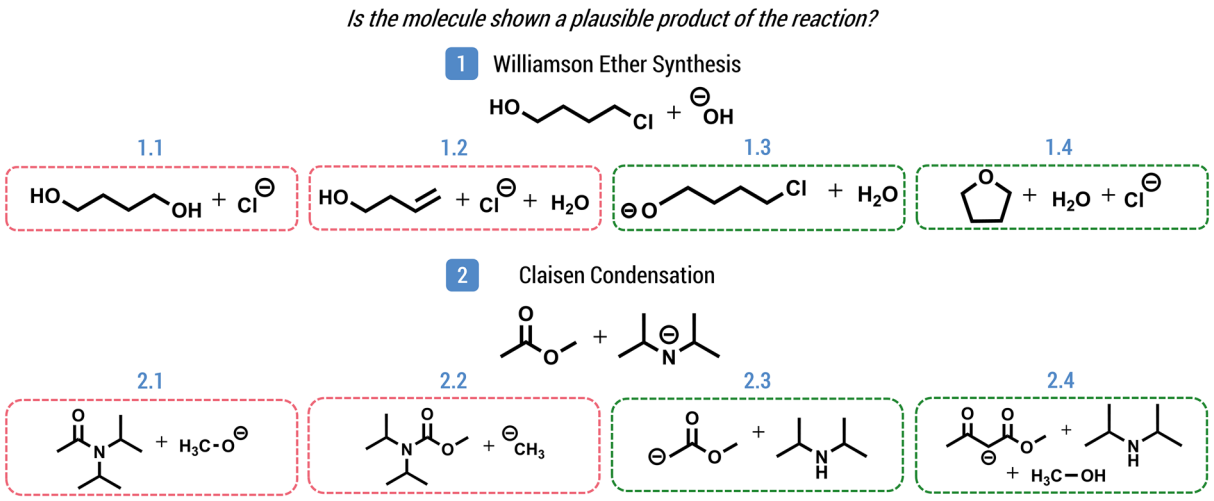

Fig. 1 Students judged the plausibility of four alternative reaction products for an intramolecular Williamson ether synthesis between 4-chlorobutan-1-ol and hydroxide (task 1) and a Claisen condensation between methyl acetate and diisopropylamide (task 2). Plausible products (1.3, 1.4, 2.3, 2.4) are highlighted in green, implausible products (1.1, 1.2, 2.1, 2.2) are highlighted in red

more comprehensive information about the study setting, see Lieber et al. (2022a).

Research Instrument

To advance students' argumentation in organic chemistry, Lieber et al. (2022a, b) developed an adaptive instructional setting, where students were prompted to judge the plausibility of competing chemical reactions for intramolecular Williamson ether synthesis and Claisen condensation (Fig. 1). In organic chemistry, competing reactions arise when reactants have the potential to undergo various reaction pathways, resulting in more or less plausible reaction products. Engaging in discussions about these alternative reaction products requires integrating multiple chemical concepts, which must be weighed to build evidence-based arguments as well as counterarguments (Lieber & Graulich 2020, 2022; Lieber et al., 2022a, b; Watts et al., 2022). In Lieber et al.'s (2022a, b) setting, students were prompted to make a claim about whether the displayed reaction product is plausible, provide evidence by using chemical concepts to support their claim, and establish a logical connection between claim and evidence by considering electronic, steric, or energetic effects (Fig. 2). Through this task design, Lieber et al. (2022a, b) demonstrated that adaptive scaffolding significantly enhanced students' argumentation.

Scoring Rubric

We developed a two-dimensional holistic rubric (Fig. 3) for evaluating students' arguments with ML drawing from the *levels of granularity* (Bodé et al., 2019; Deng & Flynn, 2021; Deng et al., 2023; Soo, 2019) and the *modes of reasoning* (Sevian & Talanquer, 2014). The levels of granularity refer to the grain size at which phenomena are explained. In general, diverse tasks demand varying levels of granularity to adequately reason about the underlying processes (Darden, 2002). In our study, we applied four levels of granularity, namely, *structural*, *energetic*, *phenomenological*, and *electronic*, which have been used in studies by Deng and Flynn (2021) and Deng et al. (2023). Besides, Sevian and Talanquer's (2014) modes of reasoning, namely, *descriptive*, *relational*, *linear causal*, and *multicomponent causal*, comprise the second dimension of our rubric. These modes of reasoning characterize the sophistication level demonstrated by students in terms of their ability to establish connections between concepts and to provide well-justified explanations for why phenomena occur (Russ et al., 2008; Sevian & Talanquer, 2014). The modes of reasoning imply that evaluating students' understanding of scientific principles should not be limited to assessing their content knowledge alone; it should also involve examining how they integrate new information into their existing cognitive network (Sevian & Talanquer, 2014). In sum, our rubric helped

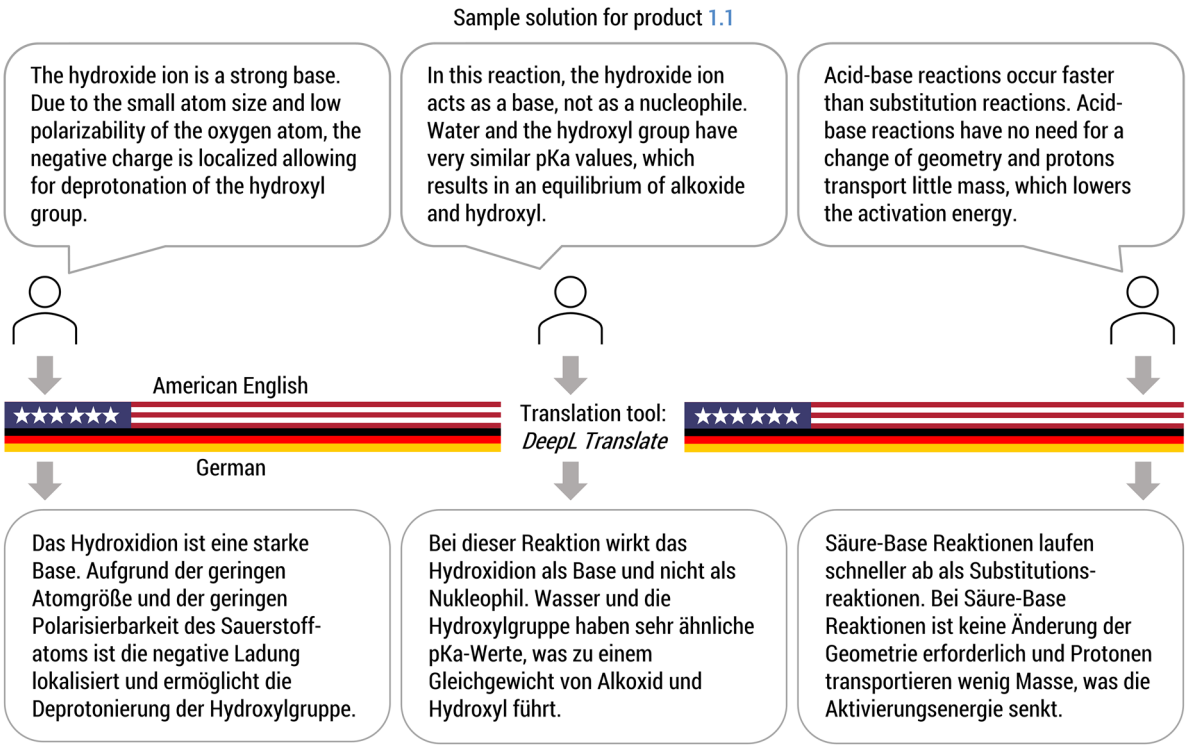

Fig. 2 Sample solutions and their translations for the implausibility of product 1.1 (Fig. 1). For more detailed sample solutions, see Lieber et al. (2022a)

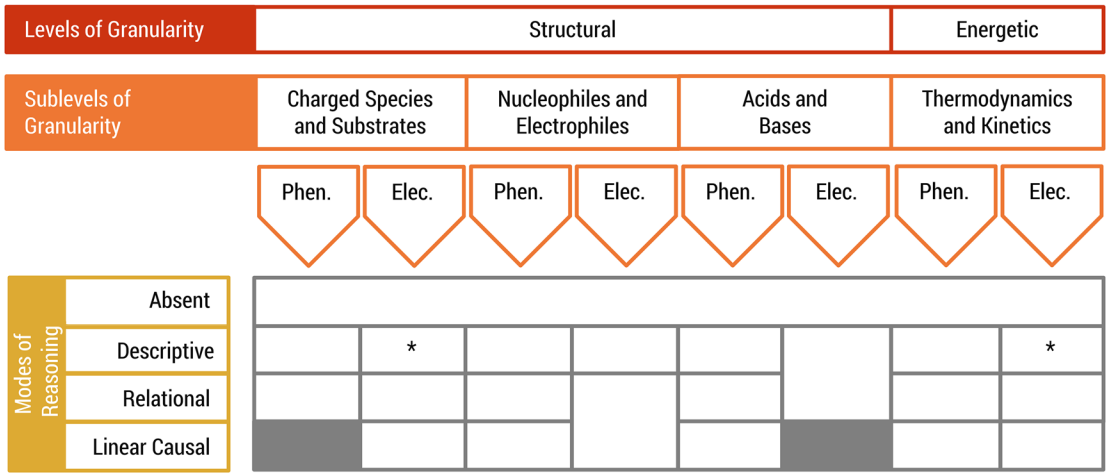

Fig. 3 Two-dimensional holistic rubric for classifying students' chemistry-related arguments into one of twenty categories. *Note*: The asterisk shows that categories belong together. Phen. = *phenomenological*, Elec. = *electronic*

us examine the concepts and relationships, the level of causality, as well as the grain size that students addressed in their argumentation. A more in-depth description of the rubric development process is published in Martin et al. (2023).

Methods

We used the PyTorch deep learning framework (Paszke et al., 2019) implemented in Python to examine the RQs. We split our data into a training, validation, and test set with a ratio of 65:15:20. The training set was used to train a deep neural network, the validation set was employed to determine the optimal hyperparameter configuration, and the test set was utilized to check the model accuracy based on four metrics (cf., Table 1). We adjusted the number of epochs, the learning rate, and the batch size as hyperparameters. Epochs indicate the number of cycles the model is trained on. Generally, training the model on more epochs enhances the model performance on the training data, but excessive training can lead to poor performance on new data. As we determined that the number of epochs greatly impacts model performance (Martin et al., 2023), we varied this number between 1 and 100. The learning rate, in turn, refers to the rate at which an optimizer updates the parameters of the model during training. Consequently, the learning rate impacts how quickly the model adapts to a specific context. Here, we tested learning rates of 1e−6, 5e−6, 1e−5, 5e−5, and 1e−4. Last, the batch size controls how many concurrent training examples are processed together in one network pass. We tried a batch size of 2, 4, 8, 16, and 32. In sum, a total of 2500 hyperparameter configurations were tested for each model.

Translation Tools

To identify the translation tool producing translations that ML techniques can analyze most accurately, we tested how reliable our deep learning architecture can classify translations generated by Google Translate (Google LLC, 2006), DeepL Translate (DeepL SE, 2017), and ChatGPT (Open-AI, 2022). In other words, human raters did not score the level of accuracy of translations generated by different tools; instead, we investigated how accurate various large language models evaluate the German translations of students' English-written scientific argumentation. Therefore, the German translations of the English-written arguments were, for each tool separately, used to train and test our deep learning architecture. When using the translation tools, we did not adjust the output; the translations were fed unmodified into the deep learning architecture. For the ChatGPT analysis,

Table 1 Metrics to measure machine-human score agreements

Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Percentage of agreement between two raters	Agreement between two raters beyond chance agreement	Agreement metrics for imbalanced datasets accounting for the number and type of prediction errors as the harmonic mean of precision and recall	
		Weighted mean of all per-class F_1-scores	Arithmetic, i.e., unweighted mean of all per-class F_1-scores

the free accounts of the first author and a research assistant were used to access the GPT-3.5 Feb 13 version (Brown et al., 2020; OpenAI, 2023). In ChatGPT, arguments were translated in a single chat using the prompt "Please, translate the following sentence in German."

We employed the monolingual large language model *GottBert-cased* as well as the multilingual model *xlm-RoBERTa-base* to compare the accuracy of the different translation tools. We used both mono- and multilingual large language models for the analysis to validate the generalizability and robustness of our findings. As hyperparameters, we varied the number of epochs, the learning rate, and the batch size as described above.

Large Language Models for Analyzing German-Written Arguments

We leveraged three German-specific, monolingual large language models *BERT-base-German-cased* (Chan et al., 2019), *dbmdz/BERT-base-German-cased* (MDZ Digital Library team, 2020), and *GottBERT-cased* (Scheible et al., 2020) as well as three multilingual models *BERT-base-multilingual-cased* (Devlin et al., 2018), *xlm-RoBERTa-base* (Conneau et al., 2019), and *xlm-RoBERTa-large* (Conneau et al., 2019) to identify the best performing one for analyzing German-written arguments. We used *cased* large language models, which are models that retain the distinction between uppercase and lowercase letters, to preserve the case information provided in the German language. Since students' arguments were originally written in English, we utilized DeepL Translate to gather German translations. Again, hyperparameters were varied as described above.

Mono- and Multilingual ML Approaches

If mono- and multilingual large language models would show similar accuracy in analyzing students' argumentation, educators could use a single multilingual model for analyzing students' reasoning across languages. Hence, we compared the accuracy of the English-specific deep learning architecture reported by Martin et al. (2023) with the best-performing German-specific architecture created for answering RQ1 and RQ2, and multilingual architectures that simultaneously analyzed students' argumentation in both languages. We used three multilingual large language models *BERT-base-multilingual-cased* (Devlin et al., 2018), *xlm-RoBERTa-base* (Conneau et al., 2019), and *xlm-RoBERTa-large* (Conneau et al., 2019) with varying epochs, learning rates, and batch sizes to identify the best-performing one. To be noted, the monolingual models are built on 1108 student-written arguments, while the multilingual ones are built on twice the amount of data, which are 1108 English-written and 1108 German-written arguments. For the multilingual approach, we ensured that an original English-written argument and its translation are either both included in the test set or not included in the test set (see the "Text Data Augmentation" section).

Text Data Augmentation

In general, the performance of ML and NLP largely depends on the quantity and quality of the training data, so training a generalizable model becomes challenging with limited data. As we hypothesized that the English-specific model can evaluate students' argumentation more accurately than the German-specific one, we looked for time-efficient ways to increase the scoring accuracy of the latter. Therefore, we tripled our German-written dataset by combining the translations of Google Translate, DeepL Translate, and ChatGPT. Since the output of the three translation tools slightly differs, combining the translations might increase model accuracy.

This process of creating additional data for ML model training is called data augmentation, which is a technique for increasing the size and diversity of the data by applying various modifications. Techniques for text data augmentation involve adding synonyms, inserting or replacing words, changing word order, altering sentence structures, or applying other linguistic transformations. Text data augmentation often helps expand the sample size, increase data heterogeneity, and boost model performance (e.g., Bayer et al., 2022; Feng et al., 2021; Shorten et al., 2021; Wei & Zou, 2019).

However, data augmentation involves the risk of overfitting the data, which means that the model becomes overly specialized. As a result, the model performs exceptionally well on the training data but fails to make accurate classifications on new data. Following that, incorporating the translations of the same sentence in both the training and test set would distort model performance as it is not measured how the model classifies new data. Accordingly, we paid attention that the translations of the same sentence are either all in the test set or not in the test set. For evaluating performance changes, we used the large language model *GottBERT-cased* while adjusting the hyperparameters as mentioned above.

Results and Discussion

RQ 1: Comparing the Performance of Translation Tools for English to German Translations

To determine the translation tool that produces translations most accurately analyzable by our deep learning

architecture, we used the validation set to identify the optimal hyperparameter configuration. Here, the machine-human score agreements varied tremendously depending on the number of epochs, the learning rate, and the batch size. So, we compared the performance of the deep learning architecture across all hyperparameter configurations, rather than for a predetermined set of those. After that, we evaluated the performance of the deep learning architecture analyzing German translations of English-written scientific arguments based on the test set to identify the best translation tool in our context. We hypothesized that ChatGPT would translate English-written arguments best into German due to the recent advancements of the GPT models in analyzing natural language.

Our deep learning architecture performed similarly across the three translation tools, with only minor variations in machine-human score agreements (Table 2). As indicated by a Cohen's κ value above 0.80 (Landis & Koch, 1977), the architecture achieved *almost perfect* machine-human score agreements when using Google Translate or DeepL Translate as a translation tool. This trend was further supported by accuracy, with Google Translate and DeepL Translate exhibiting higher values than ChatGPT. From a qualitative point of view, Google Translate and DeepL Translate handled complex sentence structures and idiomatic expressions best, resulting in more contextually appropriate translations. By contrast, ChatGPT achieved slightly lower levels of accuracy and Cohen's κ because it occasionally added or paraphrased sentences (cf., Bang et al., 2023), which led to modifying some phrases relevant to coding. Our initial hypothesis that ChatGPT would produce the most accurate translations was, thus, not confirmed. Nonetheless, especially macro F_1-score, which provides a more balanced performance evaluation across all twenty categories, indicates that ChatGPT is also a good translation tool with nearly comparable performance metrics.

Despite these valuable insights, it should be emphasized that the performance of such tools can vary depending on the specific context of the text and the language pair. In a domain-general study, Jiao et al. (2023) also compared the translation capabilities of Google Translate, DeepL Translate, and ChatGPT showing that Google Translate performed best while ChatGPT performed the least effectively. Particularly, ChatGPT's translation quality was slightly dependent on the applied prompt, with the most accurate translations obtained by asking: "Please provide the [target language] translation for these sentences." Furthermore, ChatGPT achieved almost comparable performance to other tools on high-resource language pairs, i.e., language pairs where millions of sentences are available as databases, such as German and English (Jiao et al., 2023). However, the performance gap between ChatGPT and other tools widened when dealing with low-resource language pairs like Romanian and English and when translating between languages of different language families like Chinese and English, which was confirmed by Bang et al. (2023) based on further language pairs. Moreover, the translation quality depended on the domain, where ChatGPT performed worst when translating data from biomedical abstracts or an online forum but outperformed Google Translate and DeepL Translate when translating common voice data (Jiao et al., 2023). Interestingly, the performance gap between ChatGPT and other tools got smaller when using GPT-4 instead of GPT-3.5 (Jiao et al., 2023) or when prompting ChatGPT to postedit its translations (Bang et al., 2023).

Taken together, our analysis shows that deep learning architectures can reliably classify unmodified German translations of English-written scientific arguments. This high level of machine-human score agreements can be explained, among others, by the improved performance of translation tools (Conneau & Lample, 2019), which, in turn, can be attributed to different reasons. On the one hand, improved attention mechanisms enhanced translation quality (Conneau & Lample, 2019; Vaswani et al., 2017) since these mechanisms allow translation tools to capture contextual information. On the other hand, the availability of large amounts of multilingual training data increased translation accuracy since more training data helps large language models learn more nuances of different languages. Nonetheless, comparing different translation tools is important to identify the

Table 2 Comparison of the performance of Google Translate, DeepL Translate, and ChatGPT for English to German translations

Tool	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual model: *GottBERT-cased*							
Google Translate	62	5e−6	2	0.8239	0.8114	0.8229	0.7877
DeepL Translate	16	5e−5	32	0.8194	0.8067	0.8228	0.7845
ChatGPT	45	5e−5	32	0.8062	0.7924	0.8062	0.7880
Multilingual model: *xlm-RoBERTa-base*							
Google Translate	91	5e−6	32	0.8106	0.7966	0.8050	0.7740
DeepL Translate	97	1e−5	2	0.8150	0.8015	0.8140	0.7759
ChatGPT	72	1e−5	2	0.8062	0.7920	0.8055	0.7753

most suitable one for domain-specific translation needs. Specifically, when translating students' chemistry-related arguments from English to German, our deep learning architecture analyzed translations from Google Translate and DeepL Translate most accurately. These insights are valuable for optimal translation solutions in chemistry classes.

RQ 2: Evaluating the Performance of Large Language Models in Analyzing German Translations of English-Written Scientific Arguments

Comparable to the applied translation tools, different large language models excel in different areas. Accordingly, we compared the machine-human score agreements of six mono- or multilingual large language models when processing German translations of English-written scientific arguments to identify the most reliable one. Again, we used the validation set to identify the best configuration of hyperparameters for each model. Because of the tremendous variation of the model performance depending on the hyperparameter settings, we compared the large language models based on their highest machine-human score agreements. Subsequently, we assessed their performance by using the test set.

Generally, monolingual large language models are developed to analyze text in a single language, which means that they have mostly a more fine-grained understanding of that language. In contrast, multilingual models can analyze text simultaneously in various languages, but often with lower accuracy in a single language. We are aware that we did not leverage the full potential of the multilingual large language models in the context of this RQ since we solely used them to analyze German-written arguments, i.e., arguments written in a single language. Nevertheless, comparing the accuracy of mono- and multilingual large language models in analyzing a single language helps determine their performance differences. We hypothesized that the German-specific, monolingual models would significantly outperform the multilingual ones.

The performance of each model can be found in Table 3. The monolingual large language model *GottBERT-cased* achieved the highest machine-human score agreements across all four metrics when analyzing the German translations of our sample. Surprisingly, the multilingual large language models *xlm-RoBERTa-base* and *xlm-RoBERTa-large* performed nearly as well as *GottBERT-cased*, highlighting the potential of multilingual models in analyzing students' scientific argumentation. Moreover, the large language model *dbmdz/BERT-base-German-cased* demonstrated slightly lower performance in accuracy, Cohen's κ, and weighted F_1-score than the three aforementioned models. Conversely, this model outperformed *xlm-RoBERTa-base* in terms of macro F_1-score, highlighting its capability to evaluate all categories equally. Finally, *BERT-base-German-cased* ranked as the least effective monolingual model across all metrics, while *BERT-base-multilingual-cased* performed the least among the multilingual models.

The observed variations in performance can be ascribed to the differences in model sizes, particularly influenced by the volume of data available in the German language (Table 4). Specifically, *GottBERT-cased* demonstrated the best performance because it is trained on the largest dataset in the German language. Furthermore, the multilingual large language models *xlm-RoBERTa-large* and *xlm-RoBERTa-base*, trained on the second highest amount of German language data, performed second best and showed, thus, improved performance compared to the other monolingual large language models. In detail, *xlm-RoBERTa-large* slightly outperformed *xlm-RoBERTa-base* due to an increased number of layers and hidden states in model architecture (Table 4). Ultimately, the decline in performance from *dbmdz/BERT-base-German-cased* to *BERT-base-German-cased* to *BERT-base-multilingual-cased* regarding accuracy, Cohen's κ, and weighted F_1-score also aligns with the amount of German language data these models are trained on.

In sum, a thorough model comparison allowed us to make informed decisions about the most suitable large language model for analyzing German translations of English-written

Table 3 Comparison of the performance of different large language models in analyzing German translations of English-written scientific arguments

Large language model	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual models: German-specific							
BERT-base-German-cased	52	5e−6	4	0.7885	0.7736	0.7884	0.7428
dbmdz/BERT-base-German-cased	40	1e−5	2	0.8018	0.7876	0.8022	0.7748
GottBERT-cased	16	5e−5	32	0.8194	0.8067	0.8228	0.7845
Multilingual models							
BERT-base-multilingual-cased	46	1e−5	16	0.7841	0.7686	0.7844	0.7462
xlm-RoBERTa-base	97	1e−5	2	0.8150	0.8015	0.8140	0.7689
xlm-RoBERTa-large	97	1e−5	2	0.8194	0.8061	0.8170	0.7766

Table 4 Details on model sizes: *#lgs* = number of included languages, *size: eng* = data size in the English language, *size: ger* = data size in the German language, *L* = number of layers, *H* = number of hidden states, *V* = vocabulary size, *P* = total number of parameters

Large language model	#lgs	Size: eng (in GB)	Size: ger (in GB)	*L*	*H*	*V*	*P*
Monolingual model: English-specific							
BERT-large-uncased	1	16.0	0	24	1024	30k	336M
Monolingual models: German-specific							
BERT-base-German-cased	1	0	12.0	12	768	31k	110M
dbmdz/BERT-base-German-cased	1	0	16.0	12	768	31k	111M
GottBERT-cased	1	0	145.0	12	768	52k	127M
Multilingual models							
BERT-base-multilingual-cased	104	15.5	2.1	12	768	110k	119M
xlm-RoBERTa-base	100	323.4	71.6	12	768	250k	279M
xlm-RoBERTa-large	100	323.4	71.6	24	1024	250k	561M

scientific arguments. Particularly, we found in non-conformity with our initial hypothesis that the monolingual large language models did not significantly outperform the multilingual ones as *xlm-RoBERTa-base* and *xlm-RoBERTa-large* also achieved high levels of machine-human score agreements. This finding highlights the promising abilities of both mono- and multilingual large language models in automating the analysis of students' scientific argumentation.

RQ 3: Comparing the Performance of Mono- and Multilingual ML Approaches in Analyzing Scientific Argumentation

As evidenced from the previous findings, multilingual large language models have the potential to accurately analyze scientific arguments in different languages. Following this, we compared the accuracy of the English-specific model reported by Martin et al. (2023), the best-performing German-specific model identified in the prior sections, as well as multilingual models that simultaneously analyzed English- and German-written arguments. The analysis steps are similar to the previous two sections. We hypothesized that the accuracy of the multilingual models would correspond approximately to the accuracy of the German-specific counterpart.

First, the English-specific model performed much better than the German-specific and multilingual ones (Table 5). On the one hand, this result was expected as the BERT models were initially developed to analyze the English language (Devlin et al., 2018). However, the details on the model sizes do not make explaining this result straightforward. For example, based on the size of the training data and vocabulary (Table 4), one could hypothesize that *GottBERT-cased*, *xlm-RoBERTa-base*, or *xlm-RoBERTa-large* perform best, whereas the total number of parameters might suggest that *xlm-RoBERTa-large* performs at least as well as *BERT-large-uncased*. We hypothesize that the remarkable performance of *BERT-large-uncased* is a consequence of different interrelated factors: In addition to the original intention of the BERT models to analyze the English language, the high number of layers, hidden states, and parameters (Table 4) likely contributes to improved performance. Furthermore, the limited accuracy of the translation tools may have led to minor modifications in certain phrases relevant to coding, potentially complicating the classification of the German translations by the deep learning architecture. Consequently, the accuracy of the English-specific approach to analyzing written argumentation is more reliable than the German-specific and multilingual ones.

Besides, the German-specific model *GottBERT-cased* slightly outperformed its multilingual counterparts, which is not surprising given that *GottBERT-cased* is trained on the most extensive dataset of the German language

Table 5 Comparison of the performance of mono- and multilingual ML approaches in analyzing scientific argumentation

Large language model	Epochs	Learning rate	Batch size	Accuracy	Cohen's κ	Weighted F_1-score	Macro F_1-score
Monolingual model: English-specific							
BERT-large-uncased	38	1e−5	8	0.8711	0.8610	0.8640	0.7979
Monolingual model: German-specific							
GottBERT-cased	62	5e−6	2	0.8239	0.8114	0.8229	0.7877
Multilingual models							
BERT-base-multilingual-cased	36	1e−5	8	0.8150	0.8016	0.8153	0.7772
xlm-RoBERTa-base	53	1e−5	4	0.8018	0.7866	0.7921	0.7874
xlm-RoBERTa-large	15	5e−6	4	0.8084	0.7942	0.8024	0.7685

(Table 4). In essence, since the multilingual models also analyze German language data but are trained on a smaller sample of such data, a slight decrease in performance could be anticipated. Nonetheless, *BERT-base-multilingual-cased* also achieved high machine-human score agreements in terms of accuracy, Cohen's κ, and weighted F_1-score and performed—in alignment with our hypothesis—about as well as the German-specific model (Table 5). Additionally, the macro F_1-score of *xlm-RoBERTa-base* is nearly equivalent to that of *GottBERT-cased*.

Together, the high reliability of the multilingual large language models observed again has various reasons. On the one hand, multilingual large language models are pretrained on large text corpora of different languages, extending their accuracy and generalizability across languages. On the other hand, we used twice the amount of training data to fine-tune the multilingual models since the English- and German-written arguments were combined for model training. Here, our chemistry-related arguments contain many domain-specific terms that share common linguistic features across English and German (Fig. 4). Because of the alignment of some of these terms, the multilingual models pretrained on both English and German text corpora analyzed students' scientific arguments almost as well as the monolingual models solely pretrained on English or German corpora. In other words, the analysis of the multilingual arguments benefitted from the larger training corpora of the multilingual models due to a linguistic overlap in chemistry-related terminology between both languages.

The remarkable level of accuracy shows that multilingual large language models can help bridge language gaps and improve accessibility to formative chemistry assessments for learners of the language of instruction. Especially, instructors can employ multilingual models in their courses to offer and reliably score assessments in more than one language. Students can then choose the language in which they want to complete an assessment, which may foster a sense of inclusivity and create avenues for a more equitable and enriched educational landscape. Additionally, the automated text analysis of the large language models potentially enables implementing adaptive learning, including personalized guidance, feedback, and learner support (Plass & Pawar, 2020), or Just-in-Time Teaching (Novak et al., 1999) in multiple languages. As these models continue to advance, they will further ease communicating across different languages and transform chemistry learning in multilingual contexts.

RQ 4: Investigating the Impact of Training Data Augmentation on ML Model Performance

As identified in the previous section, the accuracy of the German-specific model was almost 5% less than the accuracy of the English-specific one (Table 5). Therefore, we looked for easily implementable ways to increase the accuracy of the German-specific model. In doing so, we augmented our training data by combining the translations of the three applied tools to not only improve model performance but to also ensure that the model's accuracy is upheld across linguistic contexts. This technique for text data augmentation is inspired by Sennrich et al. (2015), who translated text into a different language and subsequently translated it back to the original language. Due to randomness in the translation process, their augmented text differed from the original text while conceptual consistency was preserved. We hypothesized that text data augmentation would contribute to a significant increase in machine-human score agreements.

Combining the translations of Google Translate, DeepL Translate, and ChatGPT slightly improved model performance from accuracy = 0.8239, Cohen's κ = 0.8114, weighted F_1-score = 0.8229, macro F_1-score = 0.7877 to accuracy = 0.8267, Cohen's κ = 0.8145, weighted F_1-score = 0.8267, and macro F_1-score = 0.7954. Hence, combining translations of different tools can be a first step in ensuring that an ML model maintains its accuracy across linguistic contexts, which confirmed our initial hypothesis to some extent. Nevertheless, the performance of the English-specific model was still much better than that of the augmented German-specific one (Table 5). Future work can, for example, use ChatGPT to rephrase text, generating multiple conceptually similar but semantically different samples (Dai et al., 2023).

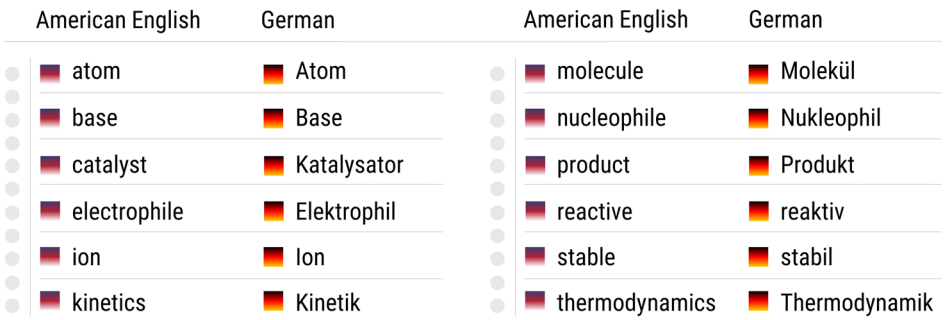

Fig. 4 Examples of chemistry-related terms that share linguistic similarities in English and German

Conclusions and Implications

In this study, we analyzed how one can reliably evaluate students' reasoning across multiple languages. We found that, when translating students' reasoning from English to German, Google Translate and DeepL Translate were most applicable. Furthermore, multilingual large language models showed great potential in automatically analyzing students' reasoning across languages. Specifically, we determined that both the monolingual model *GottBERT-cased* and the multilingual models *xlm-RoBERTa-base* and *xlm-RoBERTa-large* achieved *almost perfect* machine-human score agreements when analyzing the German translations of the English-written arguments. This finding was confirmed by the high agreement metrics of the multilingual models when simultaneously analyzing English- and German-written arguments. Therefore, instructors could use these multilingual models to allow learners to complete an assessment in their preferred language, creating chances for a more equitable and enriched educational setting for all learners. In the future, multilingual models may play a major role in facilitating communication across various languages, empowering students to engage in learning opportunities in their preferred language.

However, enough sample data in a certain language is a prerequisite for model training. Translating back and forth between languages is a way to gather data for automatically analyzing students' reasoning regardless of the language. We observed that combining translations of different tools slightly improved model accuracy, but a significant performance gap between the English-specific and German-specific models was still prevalent.

This analysis reflects the current level of technology; however, translation tools and large language models are continually improving with the ongoing advancements in ML. Following that, future research could investigate how the reliability of translation tools and large language models improves with technological progress. In addition, future research could also determine to what extent multilingual large language models can assess students' reasoning when switching between languages within a sentence, which is called translanguaging. Translanguaging allows individuals to move flexibly between languages while using different languages for different purposes. In educational settings, translanguaging has gained attention as it values students' multilingual abilities to support their learning, rather than suppressing or restricting their language use (Grapin et al., 2023; Jakobsson et al., 2021; Ryu, 2019). By employing multilingual models, students can alter between languages within a response, which may encourage them to use their full linguistic repertoire when communicating their ideas. Appreciating students' multilingual language proficiency may help account for the fluidity and interconnectedness of different languages in science contexts and encourage students to reason beyond language barriers.

Limitations

We composed a German dataset by translating English-written scientific arguments into German. Translating back and forth between languages does, however, not necessarily carry the broader social or cultural background of students' reasoning. So, we may have not fully captured the language characteristics representative of German students, which means that the data used for model training may not adequately mirror a diverse population of German students. Additionally, the original English-written arguments came from a rather homogeneous demographic group, which may constrain the capability of our model to generate precise classifications across diverse multilingual populations and institutions. To address this constraint, additional data from multilingual students needs to be collected. By using more training data, we may also mitigate some biases that we have not identified yet (cf., Noyes et al., 2020), leading to increased confidence in the outcomes of our model. In addition, an analysis of the internal workings of our algorithm, as performed by Martin et al. (2023) for the original English-written arguments, could also be applied to the German-specific and multilingual deep learning architectures reported in this article.

Moreover, only two reactions, a Williamson ether synthesis and a Claisen condensation, were evaluated in this study. The results of the analysis could possibly differ if the students were prompted to judge the plausibility of alternative reaction products for additional mechanisms. Beyond that, the arguments were constructed within a traditional organic chemistry curriculum. Curriculums such as "Mechanisms before Reactions" (Flynn & Ogilvie, 2015) or "Organic Chemistry, Life, the Universe and Everything" (Cooper et al., 2019) place more emphasis on explanations and argumentation, which may affect the dataset, the model training process, and yield different machine-human score agreements.

Besides, we only investigated the language transition between English and German, i.e., between Western languages that share certain linguistic characteristics and historical roots. For instance, both languages utilize the same alphabet and have been influenced by Latin, French, and other Roman languages over time; so, English and German show similarities in vocabulary, syntax, and grammatical structures. Our findings may probably not be generalizable across non-Western languages. Accordingly, future research should investigate how translating between other languages impacts the machine-human score agreements of different mono- and multilingual approaches.

Acknowledgements This publication is part of the first author's doctoral thesis (Dr. rer. nat.) at the Faculty of Biology and Chemistry, Justus-Liebig-University Giessen, Germany. We thank Peter Wulff and David Kranz for their help in implementing the machine learning analysis. Moreover, we thank Leonie Lieber, Ira Caspari-Gnann, and Krenare Ibraj for their pioneering research on students' argumentation in organic chemistry as well as for sharing their research data. Finally, we thank Felix Blödtner for evaluating and translating students' arguments and all members of the Graulich group for fruitful discussions.

Author Contribution PPM: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, and writing original draft. NG: conceptualization, investigation, project administration, resources, supervision, writing, review, and editing.

Funding Open Access funding enabled and organized by Projekt DEAL. This work is financially supported by the "Verband der Chemischen Industrie" (German Chemical Industry Association).

Data Availability The data used in this study is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval All data collection procedures received Institutional Review Board approval for human subject research (STUDY00001480).

Consent to Participate Informed consent was obtained from all participants.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Afitska, O., & Heaton, T. J. (2019). Mitigating the effect of language in the assessment of science: A study of English-language learners in primary classrooms in the United Kingdom. *Science Education, 103*(6), 1396–1422. https://doi.org/10.1002/sce.21545

Almusharraf, A., & Bailey, D. (2023). Machine translation in language acquisition: A study on EFL students' perceptions and practices in Saudi Arabia and South Korea. *Journal of Computer Assisted Learning, 39*(6), 1988–2003. https://doi.org/10.1111/jcal.12857

Amano, T., Rios Rojas, C., Boum Ii, Y., Calvo, M., & Misra, B. B. (2021). Ten tips for overcoming language barriers in science. *Nature Human Behaviour, 5*(9), 1119–1122. https://doi.org/10.1038/s41562-021-01137-1

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv preprint*. arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., & Chung, W. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint.* arXiv:2302.04023. https://doi.org/10.48550/arXiv.2302.04023

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys, 55*(7), 1–39. https://doi.org/10.1145/3544558

Bellmann, R. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd and Fraser.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bodé, N. E., Deng, J. M., & Flynn, A. B. (2019). Getting past the rules and to the WHY: Causal mechanistic arguments when judging the plausibility of organic reaction mechanisms. *Journal of Chemical Education, 96*(6), 1068–1082. https://doi.org/10.1021/acs.jchemed.8b00719

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (33rd ed., pp. 1877–1901). Curran Associates Inc.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y.-J., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix, 2*(1), 1–21. https://doi.org/10.37514/DBH-J.2014.2.1.05

Chan, B., Möller, T., Pietsch, M., & Soni, T. (2019). *German BERT*. Hugging Face. Retrieved September 21, 2023, from https://huggingface.co/bert-base-german-cased

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education, 105*(5), 825–836. https://doi.org/10.1002/sce.21671

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 7057–7067). Curran Associates Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint*. arXiv:1911.02116. https://doi.org/10.48550/arXiv.1911.02116

Cooper, M. M., Stowe, R. L., Crandell, O. M., & Klymkowsky, M. W. (2019). Organic chemistry, life, the universe and everything (OCLUE): A transformed organic chemistry curriculum. *Journal of Chemical Education, 97*(4), 1858–1872. https://doi.org/10.1021/acs.jchemed.9b00401

Curtis, S., & Millar, R. (1988). Language and conceptual understanding in science: A comparison of English and Asian language speaking children. *Research in Science & Technological Education, 6*(1), 61–77. https://doi.org/10.1080/0263514880060106

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., & Zhu, D. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint.* arXiv:2302.13007. https://doi.org/10.48550/arXiv.2302.13007

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science, 69*(S3), S354–S365. https://doi.org/10.1086/341858

DeepL SE. (2017). *DeepL Translate* [Computer program].

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities.

Computers & Education, 162(104094), 1–43. https://doi.org/10.1016/j.compedu.2020.104094

del Rosario Basterra, M., Trumbull, E., & Solano-Flores, G. (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Deng, J. M., & Flynn, A. B. (2021). Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chemistry Education Research and Practice, 22*(3), 749–771. https://doi.org/10.1039/D0RP00320D

Deng, J. M., & Flynn, A. B. (2023). "I am working 24/7, but I can't translate that to you": The barriers, strategies, and needed supports reported by chemistry trainees from English-as-an-additional language backgrounds. *Journal of Chemical Education, 100*(4), 1523–1536. https://doi.org/10.1021/acs.jchemed.2c01063

Deng, J. M., Carle, M. S., & Flynn, A. B. (2023). Students' reasoning in chemistry arguments and designing resources using constructive alignment. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 74–89). The Royal Society of Chemistry.

Deng, J. M., Rahmani, M., & Flynn, A. B. (2022). The role of language in students' justifications of chemical phenomena. *International Journal of Science Education, 44*(13), 2131–2151. https://doi.org/10.1080/09500693.2022.2114299

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice, 21*(1), 267–286. https://doi.org/10.1039/C9RP00148D

Dood, A. J., Fields, K. B., & Raker, J. R. (2018). Using lexical analysis to predict Lewis acid-base model use in response to an acid-base proton-transfer reaction. *Journal of Chemical Education, 95*(8), 1267–1275. https://doi.org/10.1021/acs.jchemed.8b00177

Dood, A. J., Winograd, B. A., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2022). PeerBERT: Automated characterization of peer review comments across courses. *LAK22: 12th International Learning Analytics and Knowledge Conference* (12th ed., pp. 492–499). Association for Computing Machinery.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint.* arXiv:2105.03075. https://doi.org/10.48550/arXiv.2105.03075

Flores, A., & Smith, K. C. (2013). Spanish-speaking English language learners' experiences in high school chemistry education. *Journal of Chemical Education, 90*(2), 152–158. https://doi.org/10.1021/ed300413j

Flynn, A. B., & Ogilvie, W. W. (2015). Mechanisms before reactions: A mechanistic approach to the organic chemistry curriculum based on patterns of electron flow. *Journal of Chemical Education, 92*(5), 803–810. https://doi.org/10.1021/ed500284d

Gerard, L. F., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educational Research Review, 15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Gombert, S., di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning, 39*(3), 767–786. https://doi.org/10.1111/jcal.12767

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Google LLC. (2006). *Google Translate* [Computer program].

Grapin, S. E., Pierson, A., González-Howard, M., Ryu, M., Fine, C., & Vogel, S. (2023). Science education with multilingual learners: Equity as access and equity as transformation. *Science Education, 107*(4), 999–1032. https://doi.org/10.1002/sce.21791

Grimm, A., Steegh, A., Çolakoğlu, J., Kubsch, M., & Neumann, K. (2023). Positioning responsible learning analytics in the context of STEM identities of under-served students. *Frontiers in Education, 7*(1082748), 1–12. https://doi.org/10.3389/feduc.2022.1082748

Grimm, A., Steegh, A., Kubsch, M., & Neumann, K. (2023). Learning analytics in physics education: Equity-Focused decision-making lacks guidance! *Journal of Learning Analytics, 10*(1), 71–84. https://doi.org/10.18608/jla.2023.7793

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE - Life Sciences Education, 10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Haudek, K. C., Wilson, C. D., Stuhlsatz, M. A. M., Donovan, B., Bracey, Z. B., Gardner, A., Osborne, J. F., & Cheuk, T. (2019). Using automated analysis to assess middle school students' competence with scientific argumentation. Paper presented at the National Conference on Measurement in Education (NCME), Annual Conference, Toronto, ON.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT Press.

Jakobsson, A., Larsson, P. N., & Karlsson, A. (2021). *Translanguaging in science education*. Springer.

Jiao, W., Wang, W., Huang, J.-T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint*. arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 3). Prentice Hall.

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching, 60*(2), 423–447. https://doi.org/10.1002/tea.21803

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*(4), 491–530. https://doi.org/10.3102/00346543075004491

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English-language backgrounds. *Educational Researcher, 27*(4), 12–21. https://doi.org/10.3102/0013189x027004012

Lee, E. N., & Orgill, M. (2022). Toward equitable assessment of English language learners in general chemistry: Identifying supportive features in assessment items. *Journal of Chemical Education, 99*(1), 35–48. https://doi.org/10.1021/acs.jchemed.1c00370

Lee, E. N., Orgill, M., & Kardash, C. (2020). Supporting English language learners in college science classrooms: Insights from chemistry students. *Multicultural Education, 27*(3), 25–32.

Lee, J., Lee, G.-G., & Hong, H.-G. (2023). Automated assessment of student hand drawings in free-response items on the particulate nature of matter. *Journal of Science Education and Technology, 32*(4), 549–566. https://doi.org/10.1007/s10956-023-10042-3

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching, 60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Lieber, L. S., & Graulich, N. (2020). Thinking in alternatives—A task design for challenging students' problem-solving approaches

in organic chemistry. *Journal of Chemical Education, 97*(10), 3731–3738. https://doi.org/10.1021/acs.jchemed.0c00248

Lieber, L. S., & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice, 23*(1), 38–53. https://doi.org/10.1039/D1RP00145K

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022a). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice, 23*(4), 811–828. https://doi.org/10.1039/D2RP00016D

Lieber, L. S., Ibraj, K., Caspari-Gnann, I., & Graulich, N. (2022b). Students' individual needs matter: A training to adaptively address students' argumentation skills in organic chemistry. *Journal of Chemical Education, 99*(7), 2754–2761. https://doi.org/10.1021/acs.jchemed.2c00213

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233. https://doi.org/10.1002/tea.21299

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teachers College Record, 109*(4), 897–926. https://doi.org/10.1177/016146810710900403

Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education, 96*(4), 631–651. https://doi.org/10.1002/sce.21008

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education, 94*(6), 937–962. https://doi.org/10.1002/sce.20408

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology, 30*(2), 239–254. https://doi.org/10.1007/s10956-020-09895-9

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice, 24*(2), 407–427. https://doi.org/10.1039/D2RP00287F

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21903. Early view article.

Mathew, A., Amudha, P., & Sivakumari, S. (2021). Deep learning techniques: An overview. In A. E. Hassanien, R. Bhatnagar, & A. Darwish (Eds.), *Advanced machine learning technologies and applications: Proceedings of AMLTA 2020* (1141st ed., pp. 599–608). Springer.

MDZ Digital Library team. (2020). *dbmdz German BERT models*. Hugging Face. Retrieved September 21, 2023, from https://huggingface.co/dbmdz/bert-base-german-cased

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundation of machine learning*. The MIT Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248–260. https://doi.org/10.1080/08957347.2014.944309

Novak, G. M., Gavrin, A., Patterson, E., & Christian, W. (1999). *Just-in-time teaching: Blending active learning with web technology*. Prentice Hall.

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445

OpenAI. (2022). *ChatGPT* [Computer program].

OpenAI. (2023). *ChatGPT - Release notes*. OpenAI. Retrieved September 21, 2023, from https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (32nd ed., pp. 8024–8035). Curran Associates Inc.

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1–24.

Ruder, S. (2019). *Neural transfer learning for natural language processing*. National University of Ireland.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525. https://doi.org/10.1002/sce.20264

Ryu, M. (2019). Mixing languages for science learning and participation: An examination of Korean-English bilingual learners in an after-school science-learning programme. *International Journal of Science Education, 41*(10), 1303–1323. https://doi.org/10.1080/09500693.2019.1605229

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). GottBERT: A pure German language model. *arXiv preprint*. arXiv:2012.02110. https://doi.org/10.48550/arXiv.2012.02110

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint*. arXiv:1511.06709. https://doi.org/10.48550/arXiv.1511.06709

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice, 15*(1), 10–23. https://doi.org/10.1039/C3RP00111C

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(1), 1–34. https://doi.org/10.1186/s40537-021-00492-0

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573. https://doi.org/10.1002/tea.1018

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13. https://doi.org/10.3102/0013189x032002003

Soo, K. W. (2019). *The role of granularity in causal learning*. University of Pittsburgh.

Swanson, L. H., Bianchini, J. A., & Lee, J. S. (2014). Engaging in argument and communicating information: A case study of English language learners and their science teacher in an urban high school. *Journal of Research in Science Teaching, 51*(1), 31–64. https://doi.org/10.1002/tea.21124

Taher Pilehvar, M., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education, 27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research, 19*(2), 020123-1–020123-24. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education, 34*(15), 2343–2369. https://doi.org/10.1080/09500693.2012.705046

Urban-Lurain, M., Prevost, L. B., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). Using computerized lexical analysis of student writing to support just-in-time teaching in large enrollment STEM courses. *43rd IEEE Frontiers in Education Conference Proceedings* (43rd ed., pp. 1709–1715). IEEE.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Ablex Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (30th ed., pp. 5998–6008). Curran Associates Inc.

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education, 38*(9), 1548–1569. https://doi.org/10.1080/09500693.2016.1198969

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023). Developing machine learning models for automated analysis of organic chemistry students' written descriptions of organic reaction mechanisms. In N. Graulich & G. V. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (1st ed., pp. 285–303). The Royal Society of Chemistry.

Watts, F. M., Park, G. Y., Petterson, M. N., & Shultz, G. V. (2022). Considering alternative reaction mechanisms: Students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chemistry Education Research and Practice, 23*(2), 486–507. https://doi.org/10.1039/D1RP00301A

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint.* arXiv:1901.11196. https://doi.org/10.48550/arXiv.1901.11196

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2023). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21864. Early view article.

Winograd, B. A., Dood, A. J., Finkenstaedt-Quinn, S. A., Gere, A. R., & Shultz, G. V. (2021). Automating characterization of peer review comments in chemistry courses. In C. E. Hmelo-Silver, B. de Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning: CSCL 2021* (14th ed., pp. 11–18). International Society of the Learning Sciences.

Winograd, B. A., Dood, A. J., Moon, A., Moeller, R., Shultz, G. V., & Gere, A. R. (2021). Detecting high orders of cognitive complexity in students' reasoning in argumentative writing about ocean acidification. *LAK21: 11th International Learning Analytics and Knowledge Conference* (11th ed., pp. 586–591). Association for Computing Machinery.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*(2–3), 80–107. https://doi.org/10.1080/10627190802394222

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education, 33*(3), 439–466. https://doi.org/10.1007/s40593-022-00290-6

Yik, B. J., Dood, A. J., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2021). Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items. *Chemistry Education Research and Practice, 22*(4), 866–885. https://doi.org/10.1039/D1RP00111F

Yik, B. J., Schreurs, D. G., & Raker, J. R. (2023). Implementation of an R Shiny app for instructors: An automated text analysis formative assessment tool for evaluating Lewis acid–base model use. *Journal of Chemical Education, 100*(8), 3107–3113. https://doi.org/10.1021/acs.jchemed.3c00400

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching, 59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.