Large Language Models Generalize SRL Prediction to New Languages Within But Not Between Domains

Conrad Borchers
Carnegie Mellon University
Pittsburgh, PA, USA
cborcher@cs.cmu.edu

Jiayi Zhang
University of Pennsylvania
Philadelphia, PA, USA
joycez@upenn.edu

Hendrik Fleischer
Leibniz Universität Hannover
Hannover, Germany
fleischer@idn.uni-hannover.de

Sascha Schanze
Leibniz Universität Hannover
Hannover, Germany
schanze@idn.uni-hannover.de

Vincent Aleven
Carnegie Mellon University
Pittsburgh, PA, USA
aleven@cs.cmu.edu

Ryan S. Baker
University of Pennsylvania
Philadelphia, PA, USA
ryanshaunbaker@gmail.com

Think-aloud protocols are a standard method to study self-regulated learning (SRL) during learning by problem-solving. Advances in automated transcription and large language models (LLMs) have automated the transcription and labeling of SRL in these protocols, reducing manual effort. However, while effective in many emerging applications, previous works show LLMs struggle with reliably classifying SRL across specific instructional domains, such as chemistry or formal logic. Can LLMs reliably classify SRL within a given domain, but also across different languages that represent distinct instructional approaches? This study investigates using classification models based on LLM embeddings to automatically detect SRL in think-aloud transcripts of 26 students at German and American universities working with three tutoring systems for chemistry and formal logic. Using OpenAI's text-embedding-3-small, we predicted four categories of SRL processes based on the four-stage SRL model by Winne and Hadwin (processing information, planning, enacting, and realizing errors). We compare how well embedding-based SRL models transfer between English and German across chemistry and logic domains, including how levels of scaffolding in the tutoring systems and culturally unfamiliar instruction impact transfer. We found that LLM embedding-based classifiers trained on English data can reliably classify SRL categories in German think-aloud data (and vice versa) with minimal performance degradation within but not between domains of instruction. Further, model transfer performance declined due to linguistic differences in subject-specific terminology and unawareness of instructional context, such as specific hint messages. Considering these factors in future refinement of our methodology will move the field closer to the research goal of a reliable SRL classifier that is domain-, language-, and learning system-general.

Keywords: self-regulated learning, think-aloud protocols, large language models, NLP, intelligent tutoring systems, instructional context, chemistry education, prediction, classification

1. INTRODUCTION

Successful use of self-regulated learning (SRL) is positively associated with learning and learning outcomes (Cleary & Chen, 2009; Nota et al., 2004; van der Graaf et al., 2022; Zimmerman, 1990). Students who are skilled in SRL can set goals effectively (Lim et al., 2021), search for information (Zimmerman, 2000), and direct their attention and cognitive resources to align their efforts with their objectives (Heirweg et al., 2020; Zimmerman, 2000). Given the importance of SRL in the learning process, prior studies have used behavioral log data to measure and facilitate students' use of SRL in intelligent tutoring systems (ITSs). These adaptive and personalized systems offer step-by-step guidance during problem-solving tasks (Aleven et al., 2016). By analyzing the patterns of behaviors from students interacting with an ITS, researchers can draw inferences, identifying which SRL strategies the students are using, how they are used, and in what order (Araka et al., 2020; Saint et al., 2020, 2022; Winne, 2017; Winne & Baker, 2013). With this approach, previous studies have used behavioral log data to examine a range of SRL processes, including help-seeking (Aleven et al., 2006), gaming the system (sometimes viewed as an SRL strategy aimed at avoiding to have to learn; Baker et al., 2004), setting goals, making plans (Azevedo et al., 2011; Biswas et al., 2010), tracking progress (Biswas et al., 2010), and engaging in various cognitive operations, such as assembling and monitoring, during problem-solving (Hutt et al., 2021; Nasiar et al., 2023; Zhang et al., 2022). Automated detectors have been developed to measure these SRL processes immediately, offering assessments that identify both the SRL strategies students are employing and those they may be lacking.

In addition to log data, think-aloud protocols (TAPs) are another approach that has been frequently used in previous studies for measuring SRL in situ (Bannert et al., 2014; Greene et al., 2011, 2017; Lim et al., 2021). During think-aloud activities, students are asked to verbalize their thinking and cognitive processes as they interact with an ITS while solving a problem. Utterances collected from think-aloud activities are then transcribed and segmented into clips. To assess students' use of self-regulation, researchers manually code students' verbalizations in each clip, labeling the presence or absence of SRL strategies (Greene et al., 2017). Previous studies have examined how the presence, frequency, and sequential and temporal order of SRL in TAPs relates to overall learning outcomes (e.g., Bannert et al., 2014; Lim et al., 2021; Molenaar et al., 2021) and to the moment-by-moment performance when solving a multi-step problem (Borchers et al., 2024b).

Advances in natural language processing and foundation models enabled the automated transcription and labeling of SRL in TAPs, which is otherwise time-consuming and laborious. Specifically, large language models (LLMs) have been used to reliably classify SRL from language in tutoring systems (Zhang et al., 2024a). More broadly, outside of SRL, transformer-based models have automated the coding of chat messages during collaborative learning and other language snippets co-occurring with learning (de Araujo et al., 2023). This coding enabled novel forms of discovery of models, such as estimating learning rates based on fine-grain labels of dialog moves, at a scale that was previously hard to attain (Borchers et al., 2024a).

Despite these successes in automatically coding SRL from text, predictive models based on LLM embeddings are limited in generalizing to unseen domains of instruction. For example, when applying a model trained on chemistry data to formal logic data, Zhang et al. (2024a) identified domain transfer limitations including not adequately dealing with subject-specific language and instruction (e.g., hints). Similar instructional differences might exist between languages and country-level differences in instruction (e.g., how different languages describe concepts). These differences in instruction and domain taxonomies may challenge the ability of LLMs to reliably classify SRL in data from unseen languages. Worse predictive performance in unseen languages risks making educational technology more effective for one population than another, deepening existing inequities between countries that have been identified in MOOCs, among others (Kizilcec et al., 2017). In domains like chemistry, students learn to develop specialized abstractions from everyday language to build subject-specific knowledge (Heeg et al., 2020). This can lead to differences in how shared terms are understood across cultures and languages. For example, in German, the term "lösen" would commonly be translated into "solving" in English. However, "lösen" in German chemistry terminology refers to the concept of forming a mixture, which in English is referred to as dissolving. In this case, translating "lösen" to "solving" would be correct in a common sense but not in a chemical instructional context. This highlights how everyday terms correspond differently to domain-specific concepts across languages. However, to our knowledge, prior research has yet to systematically compare the generalizability of prediction across cultural contexts alongside domains of instruction. The present study is an extension of the research described in Zhang et al. (2024a), which limited the prediction of SRL from think-aloud utterances to English data only. In contrast, in this study, we systematically investigate the transferability of SRL prediction to a second language, which we call *language transfer*, with distinct instructional cultures, as described next.

Beyond differences in terminology, which could limit LLM generalizability to new languages, students' familiarity with different forms of instruction may further complicate prediction. For example, stoichiometry, a chemistry subdomain focused on the quantitative relationships in chemical reactions, is taught differently in the US and Germany. Schmidt (1997) reviewed textbooks and identified two methods: the mole and proportionality method. Here, we focus on the proportionality methods because both countries predominantly teach strategies related to the proportionality method. However, the specific approach to teaching the proportionality method differs between the two countries. In the US, the method is typically taught using a factor label strategy, guiding students to multiply by unit conversion factors to calculate variables. Scientific documentation of how the proportionality method is taught in German textbooks is presently lacking. However, related instruction in physics involving units and quantities indicates that textbooks instruct students to transform abstract quantity relationships to then substitute givens to derive a target variable (Dröse & Prediger, 2018). Taken together, the predominantly taught problem-solving strategy in the US is the factor label method (Schmidt, 1997), while German students, at least in physics, are taught to transform quantity relationships through abstract representations (Dröse & Prediger, 2018). Recent evidence suggests that such instructional differences and level of instructional support (i.e., scaffolding) relate to SRL differences, for example, when students have less support to plan before enacting problem-solving steps (Zhang et al., 2024b), which might make SRL prediction less tractable. However, to the best of our knowledge, such cross-national comparisons of SRL predictions have not been evaluated.

The present study investigates the out-of-the-box use of LLM embeddings as input features of machine learning models to automatically detect SRL in machine-transcribed student think-aloud transcripts between multiple instructional domains and languages. We collected students' think-aloud data from three ITSs covering stoichiometry chemistry and formal logic across 26 students from the United States and Germany. The English audio was transcribed using Whisper, a state-of-the-art speech-to-text software, while the German audio was hand-transcribed, due to low audio quality. After analyzing the transcripts, we developed operationalizations for four SRL categories—processing information, planning, enacting, and realizing errors—grounded in Winne and Hadwin's four-stage model (Winne & Hadwin, 1998), representing key behaviors in each step. We then conducted a coding round, labeling the presence or absence of the four SRL categories using a coding scheme. A sentence embedding model (Open AI's text-embedding-3-small; Neelakantan et al., 2022) was applied to vectorize the text. Using the outputs from the embedding model as features, we trained machine learning models separately on English and German transcriptions to predict the presence or absence of the four SRL categories. We assessed the transfer and generalizability of these models and estimated performance losses due to instructional differences between domains and languages when these models are applied out of the box. Our study has important implications for what data researchers may need to collect when considering transferring SRL classification models based on think-aloud data to new contexts. Specifically, we answer the following research questions (RQs):

RQ1: How accurately do embedding-based machine learning models predict SRL process categories in the context of tutored problem solving when transferring to another language on matched platforms and domains?

RQ2: Does language transfer depend on the degree of scaffolding provided and whether the scaffolded problem-solving strategy is unfamiliar to the target population?

RQ3: How much language transfer occurs when the SRL model is applied to a different learning platform design and domain?

The present study's contributions are threefold. First, we contribute empirical evidence that SRL categories (processing information, planning, enacting, and realizing errors) in students' verbalizations can be reliably predicted in languages other than English. Second, we contribute to the understanding of the conditions under which automated SRL classification models can be transferred across languages with minimal accuracy loss. Our findings indicate that for successful transfer, the model must be trained on data within the same instructional domain. Additionally, performance declines may be more pronounced in tutoring systems that feature extensive scaffolding and employ instructional strategies unfamiliar to the target population. Third, we define conditions under which automated prediction of SRL, especially between languages, could be made more accurate, most notably by considering culturally distinct forms of defining and describing domain-specific phenomena using terms in ways that are different from their everyday use.

2. Background

2.1. Self-Regulated Learning

Self-regulation, a critical component in learning, is where learners take active control of their learning by monitoring and regulating their attention and effort in pursuit of goals (Schunk & Zimmerman, 2011). During this process, learners may set goals, monitor progress, and adjust strategies when goals are not met. A range of cognitive, metacognitive, affective, behavioral, and motivational processes are involved in SRL. Engaging in these processes effectively enables learners to become more independent and effective in their learning (Zimmerman, 2000). In general, students who effectively self-regulate their learning tend to perform better than those who do not (Zimmerman, 1990) and are more likely to have a deep conceptual understanding of the topic (Azevedo et al., 2017; Greene & Azevedo, 2010; Labuhn et al., 2010).

In the last three decades, several theoretical models have been proposed from different perspectives to depict the process of SRL (Panadero, 2017). For example, based on socio-cognitive theories, Zimmerman (2000) describes the process of SRL as three cyclical phases: forethought, in which learners analyze a task; performance, in which learners execute the task; and self-reflection, in which learners assess and evaluate their performance. Grounded in information processing theory, Winne and Hadwin (1998) characterize the process of SRL as four interdependent and recursive stages, in which learners: 1) define the task, 2) set goals and form plans, 3) enact the plans, and 4) reflect and adapt strategies when goals are not met. A range of SRL processes may be involved in each stage of the cycle.

Despite the differences in theoretical groundings and focuses, most of the models describe SRL as a cyclical process consisting of phases where learners understand tasks, make plans, enact the plans, and reflect and adapt (Lim et al., 2021; Winne, 2010, 2017). These theoretical models are frequently adopted in recent SRL research as foundations that guide the conceptualization and operationalization of SRL in SRL measurement (Winne, 2010; Zheng, 2016). Recent work in educational data mining (EDM) and learning analytics has provided empirical support for cyclical models of SRL by relating cyclical SRL stages to learner performance data (Bannert et al., 2014; Borchers et al., 2024b; Hatala et al., 2023; Heirweg et al., 2020).

2.2. Using Think-aloud Protocols to Measure and Understand SRL

Previous studies have used TAPs to measure and understand SRL processes. In think-aloud activities, students are asked to verbalize their thinking and cognitive processes while they solve a problem (Greene et al., 2017). Utterances collected from think-alouds allow researchers to measure and examine SRL processes that are contextualized in the problem-solving process as the utterances are approximately concurrent with their occurrences.

To engage students in think-aloud activities, instructions are often given prior to a task, asking students to verbalize their thinking while working on a task, as if they are speaking to themselves (Ericsson & Simon, 1998). Once the task begins, researchers or the learning software may use simple prompts such as “please keep talking” to remind participants to continue to talk when learners stop verbalizing (Greene et al., 2011). These instructions and prompts are designed with the goal of inflicting a minimum amount of distraction which might alter a student’s thinking process.

To accurately capture students’ thinking process, Ericsson and Simon (1998) provide guidelines on the TAP instructions and prompts. In this, they contend that prompts should primarily focus on asking students to express conscious thoughts using language that directly represents those thoughts (e.g., “my plan is to complete the assignment”) or express thoughts in which sensory information is converted into words (e.g., “I see three hyperlinks here”). In contrast, prompts should refrain from asking students to metacognitively monitor and reflect on their thinking process, as this can potentially influence how students think and perform tasks, altering the order and nature of their cognitive processes (Ericsson & Simon, 1998; Schooler et al., 1993). When prompts are carefully designed to avoid engaging students in metacognitive activities, studies have found that thinking aloud alters neither accuracy nor the sequence of operations in most cognitive tasks (excluding insight problems; Fox et al., 2011).

Once students complete the learning task, their verbalizations collected using audio or video recordings are then transcribed to text. The recordings, which were once predominantly transcribed by humans, are now increasingly transcribed by automated transcription tools such as

Whisper (Radford et al., 2023), with transcription accuracy described in their technical reports being satisfactory without human supervision. With the transcriptions, researchers code the SRL processes using a coding scheme (e.g., Bannert et al., 2014; Heirweg et al., 2020; Lim et al., 2021). As a critical part in TAP, the coding scheme outlines the target SRL processes captured in a transcript and provides an operationalization for each process. These schemes are typically derived from SRL theories and then refined and modified based on the existing task, platform, and dataset (Greene et al., 2011). For example, to examine students' use of SRL in think-aloud transcripts, Bannert et al. (2014) developed a coding scheme that outlines SRL categories corresponding to the three phases (i.e., forethought, performance, and reflection) in Zimmerman's model (Zimmerman, 1990). By comparing the SRL activities between high and low-achieving students, Bannert et al. (2014) found that high achievers tended to demonstrate more frequent use of SRL processes such as planning and monitoring, and they are also more effective and strategic at implementing SRL strategies. Using the same coding scheme, Heirweg et al. (2020) and Lim et al. (2021) found that successful learners were more likely to engage in preparatory activities (e.g., orientation and planning) before completing a task. In contrast, preparation and evaluation activities were less frequently used by less successful students.

In addition to studying the effective use of SRL in relation to students' overall achievement, a recent study examined how the use of SRL inferred from TAPs is correlated to moment-by-moment performance when students are solving a multi-step problem with an ITS (Borchers et al., 2024b). Specifically, they coded four SRL categories based on Winne and Hadwin's four-stage model (1998). By coding SRL categories in students' utterances in between steps, they examined how the use of SRL in terms of presence, frequency, cyclical characteristics, and recency relate to student performance on subsequent steps in multi-step problems. They show that students' actions during early stages of SRL cycles (e.g., processing information and planning) exhibited lower moment-by-moment correctness than later SRL cycle stages (i.e., enacting) during problem-solving. This granular coding allowed for studying the effectiveness of SRL processes for problem-solving performance in a fine-grained way. Understanding how SRL processes influence the subsequent performance provides further evidence of when interventions could be provided during problem solving—in this case, early SRL cycle stages.

2.3. Using Natural Language Processing to Scale Up SRL Measurement

Less than ten years ago, a position paper published by McNamara et al. (2017) discussed the significant impact of natural language processing (NLP) on understanding and facilitating learning. Recent advancements in NLP continue to reveal new opportunities for supporting learning through analytics.

One emerging application using natural language in the domain of education is predicting students' cognitive processes from learner text artifacts, with the goal to provide real-time feedback and to measure these constructs at scale. For example, to understand how students engage in SRL and to provide timely scaffolds in math problem-solving, Zhang et al. (2022) developed robust detectors using NLP and machine learning that measure SRL in students' open-ended responses. Specifically, they extracted features that resemble the linguistic characteristics found in students' text-based responses and trained machine learning models that detect SRL constructs, reflecting how students assemble information, form mental representations of a problem, and monitor progress. Similarly, Kovanović et al. (2018) developed machine learning models that automatically identify the types of reflection in students' reflective writing. In their work, NLP methods including n-grams, LIWC (Tausczik & Pennebaker, 2010), and Coh-Metrix

(Crossley et al., 2007) were used to extract features from students' reflective writing which were then used to train models and make predictions. These NLP-based models demonstrate the possibility of processing students' written responses and evaluating their cognitive processes in real time and at scale.

Since then, large language models (LLMs) and sentence embeddings have substantially advanced the state of the art in language models. These models, based on deep learning architectures, are trained on massive amounts of text data to understand and generate human language in a contextually coherent and meaningful manner (Wei et al., 2022). Building on these advancements in LLM text comprehension, recent studies have investigated using LLMs as detectors to enhance the prediction of cognitive constructs within text. This includes detecting attributes and relatedness of peer feedback (Darvishi et al., 2022; Zhang et al., 2023b), as well as detecting gaming the system (a strategy aimed at avoiding having to learn) in open-ended responses, with findings demonstrating reliable detection accuracies using LLMs with AUCs above 0.8 (Zhang et al., 2023a).

However, many detectors in EDM and related fields have been unsuccessful at transferring across learning contexts (Baker, 2019). The models developed in past work are mainly designed and evaluated within one platform, though Paquette and Baker (2019) represents one of the few exceptions. Being able to evaluate the performance of a detector across systems will allow us to better understand the limitations of these models, and how language may differ when students are working in different subjects and systems, albeit capturing the same cognitive attributes.

2.4. Multilingual Language Models

Multilingual language models are artificial intelligence systems that are designed to understand, generate, and process text in multiple languages. Unlike monolingual models, which are trained on data in a single language, multilingual models are trained on a diverse set of languages, allowing them to perform various tasks across different linguistic contexts (Doddapaneni et al., 2021). To do so, a shared representation of text across languages is needed for multilingual language models to process multiple languages.

Almost a decade ago, dictionaries were used to establish shared representations, mapping words with similar meanings across two languages (Ammar et al., 2016). Since then, with the increasing use of transformer models, multilingual sentence embeddings have been developed. Pre-trained multilingual models, such as mBERT (Gonen et al., 2020) and XLM-RoBERTa (Li et al., 2021), utilize corpora from around 100 languages and are trained to perform masked language modeling (predicting a word in a sentence using the surrounding words). Through this training, the models learn the semantic meaning of words across languages by using the context provided by the surrounding words. Generative pre-trained transformer, or GPT, on the other hand, uses a different training objective, focusing instead on predicting the next word in a sequence (Yenduri et al., 2024). These models utilize a shared vocabulary and embedding space across languages, creating a universal representation of sentences that capture the similar semantic information across different linguistic contexts (Artetxe & Schwenk, 2019; Pires et al., 2019). These representations can then be fine-tuned for various downstream tasks enabling the possibility of processing different languages and transferring knowledge across languages.

For example, using the shared representations, classification models trained using one language can be transferred to a different language (Artetxe & Schwenk, 2019), which is also known as zero-shot transfer. In a multilingual zero-shot transfer task, a pre-trained model is fine-tuned on a specific task in one or more source languages, and then applied directly to a target language without any additional adaptation. This approach is effective when the task is language-agnostic or when the source and target languages are closely related (Huang et al., 2021). For instance, a model fine-tuned on sentiment analysis in one language can often subsequently be successfully used to analyze sentiment in other dialects (Omran et al., 2023). However, the success of the transfer of such models can be influenced by the languages (e.g., similarities in typology between the two languages; Conneau et al., 2017) and the nature of the tasks (Hu et al., 2020).

In the field of EDM, pre-trained multilingual models have been used for a range of tasks, including grading written assignments in different languages (Firoozi et al., 2024; He & Li, 2024) and analyzing discourse in collaborative learning environments that support multiple languages (Araujo, et al., 2023). For example, in Firoozi et al. (2024), two multilingual sentence embedding models (mBERT and language-agnostic BERT, LaBSE) were used to train models that rate students' essays in three languages (German, Italian, and Czech) based on a six-score scale. Essays from all three languages were used to train models. By comparing the predictions to human ratings, they found that both sentence embedding models demonstrate comparable performance across all three languages. However, LaBSE is significantly better at producing accurate predictions for the majority of the score levels within each language. When clustering the embeddings, they show that mBERT is better at clustering essays based on languages, while LaBSE produces clusters that better represent the score levels. This result can be attributed to LaBSE's advantage in translation language modeling, which enhances its cross-linguistic comprehension and enables LaBSE to more effectively use the multilingual context (Firoozi et al., 2024). Additionally, de Araujo et al. (2023) used the Wiki40b-lm-multilingual and USE-multilingual sentence encoders to investigate the transferability of these models across domains and languages for detecting types of chats in a collaborative learning environment. They observed significantly reduced, yet still moderate, performance when models were trained on student chats in one subject area in Dutch and tested on chats from another subject area in Portuguese, and vice versa.

Despite notable successes in the use of multilingual models within educational applications, cross-national transfer studies remain rare. This scarcity is especially important as transferring models across instructional contexts and platforms continues to be a fundamental challenge in the field of learning analytics (Baker, 2019). Past work suggests that affect detectors, another common prediction model in EDM, can be limited in their capacity to generalize to new populations, even when these populations are part of the same national or regional culture (Ocumpaugh et al., 2014). Similarly, recent work that increasingly studied differences in model performance across different populations has primarily done so within US-American samples (e.g., Zambrano et al., 2024), neglecting opportunities to study differences between national samples. The present study addresses this gap by comparing the performance of multilingual prediction models of SRL in a second language embedded in distinct instructional and cultural contexts.

3. METHODS

To answer our three research questions listed above, we collected students' log data as well as think-aloud transcripts from students enrolled at German and American universities as they worked with one of three ITSs (section 3.1 and 3.2). The three ITSs (i.e., StoichTutor, ORCCA, and LogicTutor) each cover the topic of either stoichiometry or formal logic and differ in their level of scaffolding (section 3.3). We analyzed students' think-aloud data, developed a codebook, and coded the utterances for four SRL categories (section 3.4). To train models that detect the SRL categories, we vectorized the utterances using a multilingual sentence embedding model to create a feature space. Using the embeddings and labels, machine learning models were trained to classify the presence or absence of the four SRL categories (section 3.5). Performance of these models within a single language was then compared to models' transfer performance across languages, platforms, and subject domains. To evaluate the transferability, a series of models were constructed using a subset of data and tested on data collected from the language, platform, or domain that was not included in the training set (section 3.6). Lastly, error analysis was conducted to identify common errors and construct theories of failure regarding why models fall short or fail to transfer (section 3.7). All data analysis code is publicly available through a GitHub repository, which also describes how access to the original American dataset can be obtained: https://github.com/pcla-code/EDM24_SRL-detectors-for-think-aloud.

3.1. Sample and Recruitment

Twenty-six participants took part in this study, including **fifteen** university students in the United States and **eleven** students in Germany. Of the fifteen students in the United States, the average age was 20.60 years (SD = 3.27). The participants were 40.0% White, 46.6% Asian, and 13.3% multi- or biracial. The United States sample included undergraduate first-year students (21.4%), sophomores (14.3%), juniors (35.7%), seniors (21.4%), and one graduate student (7.1%). Of the eleven students in Germany, the average age was 21.57 years (SD = 3.21). Of these students, 9% have an immigration background, meaning either one of their parents or themselves were not born with German nationality (which is the established way of measuring sample diversity in Germany). The German sample included undergraduate first-year students (9.1%), sophomores (63.6%), juniors (18.2%), and one graduate student (9.1%).

Data collection took place between February and June of 2023. The present study sampled a total of 955 English annotated utterances from 15 students working on 3 problems across the three platforms, and sampled a total of 584 German annotated utterances from 11 students working on 2 problems across the two chemistry platforms. Ten US students were recruited at a private research university (and participated in person). The other five US students were recruited from a large public university (and participated via Zoom). All German students participated in-person and were enrolled in a large public university. All students were recruited through course-related information channels by instructors in courses known to include students still learning stoichiometry. Further, students were asked to circulate recruitment materials to peers and encourage them to do the same. Participants received $15 Amazon gift cards (US sample) or 15 Euros in cash (German sample) as compensation. All students completed a session between 45-60 minutes.

3.2. Study Procedure

Students were distributed to conditions such that at least five students worked with each of the three ITSs. Given that LogicTutor had not been translated to German at the time of data collection, German students were only assigned to use StoichTutor or ORCCA (or both). All students participating in the study started with a self-paced questionnaire assessing demographic information, prior academic achievement, and self-rated proficiency in the subject domain (i.e., stoichiometry chemistry or formal logic). Then, students viewed a pre-recorded introductory video about the ITS they would work with and could ask questions about the video. In the case of LogicTutor, students also had the opportunity to read two articles on formal logic symbolization and rules. They also were allowed to ask the experimenter, who was familiar with formal logic, any questions about symbolization and the content. Students had up to five minutes to skim both articles to develop relevant questions to ask the experiment conductor. Both articles were taken from a remedial first-year undergraduate summer course on formal logic in which the LogicTutor was previously deployed. The articles ensure that all participants had the necessary prerequisite knowledge and knew the required symbolization for expressing logical formulae to work with the tutoring software. After becoming acquainted with the tutoring software, students received a brief demonstration and introduction to the think-aloud method and began working on tutor problems at their own pace for up to 20 minutes while thinking aloud. Chemistry problems were taken from past research on StoichTutor (McLaren et al., 2011a, 2011b), and either presented in StoichTutor or implemented in ORCCA. Problems in LogicTutor focused on transforming equations into normal forms and self-explaining these transformations via menus.

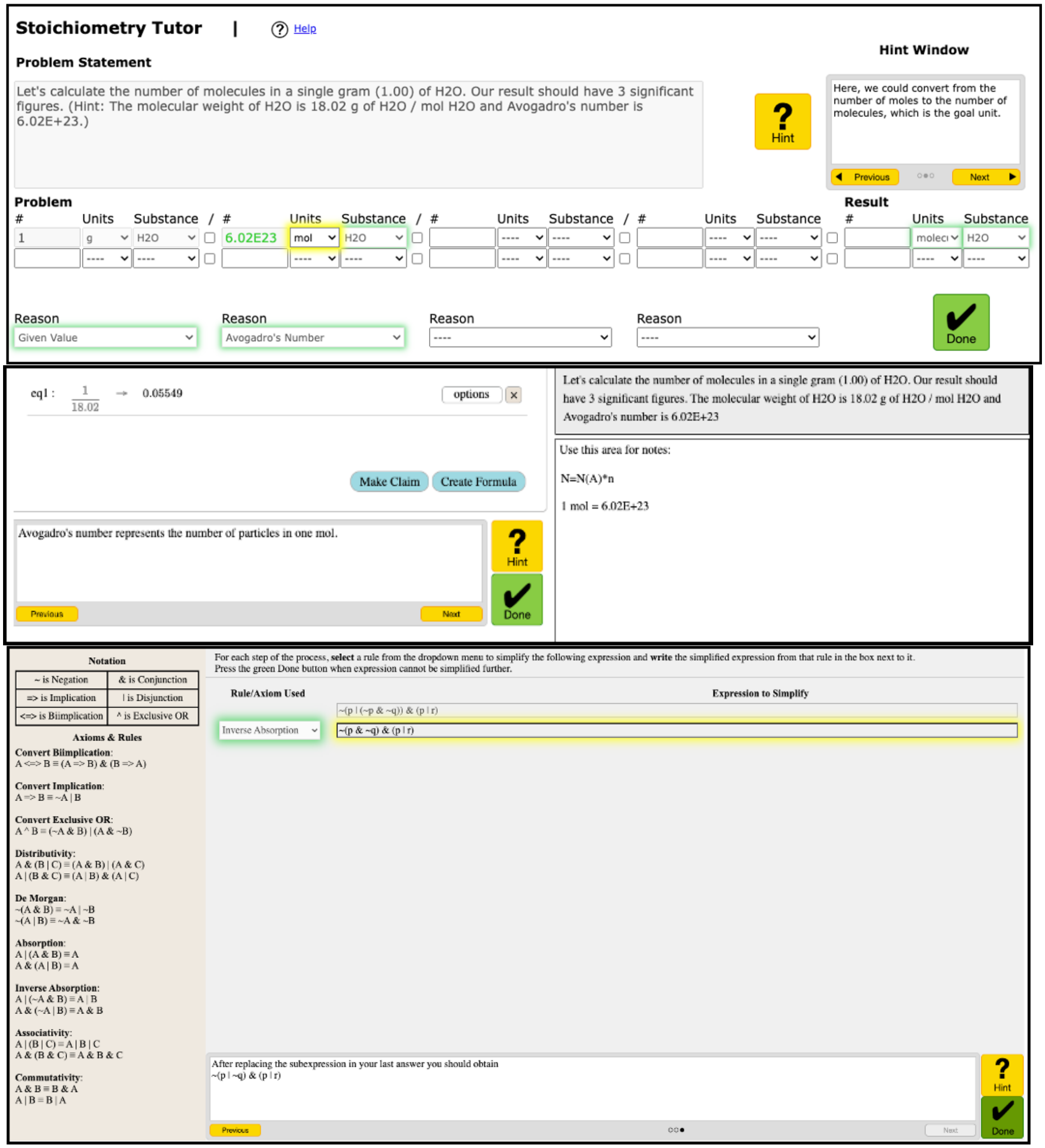
The experimenter occasionally reminded the research participants to keep talking when they fell silent for more than 5 seconds. Think-aloud utterances were recorded with a 2022 MacBook Pro built-in microphone of the computer serving the tutoring software or Zoom microphones of the participating student's laptop. Problems were content-matched across the StoichTutor and ORCCA ITSs and included two content units taken from prior studies featuring StoichTutor: (a) moles and gram conversion and (b) stoichiometric conversion. Both content units included a total of four problems. For LogicTutor, two problem sets were taken from prior remedial summer courses for first-year undergraduates. The problem sets covered simplifying logical expressions (seven problems) and transforming logical expressions to the negation normal form (four problems), respectively. Students worked on both problem sets while thinking aloud in a fixed sequence until the time ran out. This decision was based on both problem types having increasing difficulty levels, with the first problems set including additional problem-solving scaffolds via reason boxes.

3.3. Intelligent Tutoring Systems

Our first ITS, StoichTutor, is an example-tracing tutor (Aleven et al., 2016). StoichTutor has significantly improved high school and college students' stoichiometry skills (McLaren et al., 2006, 2011a, 2011b). The most recent version of StoichTutor incorporates insights from previous design experiments, including the use of polite language in hints (McLaren et al., 2011). The StoichTutor utilizes a structured, fraction-based problem-solving method (i.e., the factor-label method) to guide students toward target values (see Figure 1.top).

Our second ITS, the Open-Response Chemistry Cognitive Assistant (ORCCA), is a rule-based model-tracing ITS for chemistry (King et al., 2022). ORCCA matches problem-solving rules with students' strategies, accommodating flexible problem-solving sequences with a formula interface (see Figure 1.middle). Rule-based ITSs allow for more flexibility in problem-solving strategies (Aleven, 2010).

Our third ITS, LogicTutor, is a rule-based tutoring system for learning propositional logic (see Figure 1.bottom). Students are guided through constructing truth tables, correlating the structure of a formula with its meaning by assigning truth values. Students are tasked with manipulating propositional formulae and transforming them into equivalent expressions using a limited set of logical connectives. Additionally, students learn to apply transformation rules to rewrite or simplify formulas. Students receive hints and error feedback with dynamically generated counterexamples to students' formulae. A "cheat sheet" on the left side of the screen reminds them about relevant logical transformations. The interface also includes reason boxes for self-explanation, which offer further scaffolding during problem-solving.



Problem Statement

Let's calculate the number of molecules in a single gram (1.00) of H2O. Our result should have 3 significant figures. (Hint: The molecular weight of H2O is 18.02 g of H2O / mol H2O and Avogadro's number is 6.02E+23.)

Here, we could convert from the number of moles to the number of molecules, which is the goal unit.

Problem

#	Units	Substance	/	#	Units	Substance	/	#	Units	Substance	/	#	Units	Substance	/	#	Units	Substance
1	g	H2O	☐	6.02E23	mol	H2O	☐		----	----	☐		----	----	☐		moleci	H2O
	----	----	☐		----	----	☐		----	----	☐		----	----	☐		----	----

Result

Reason: Given Value; Reason: Avogadro's Number; Reason: ----; Reason: ----

eq1 : $\frac{1}{18.02}$ → 0.05549

Avogadro's number represents the number of particles in one mol.

Let's calculate the number of molecules in a single gram (1.00) of H2O. Our result should have 3 significant figures. The molecular weight of H2O is 18.02 g of H2O / mol H2O and Avogadro's number is 6.02E+23

Use this area for notes:

N=N(A)*n

1 mol = 6.02E+23

Notation

~ is Negation	& is Conjunction
=> is Implication	| is Disjunction
<=> is Biimplication	^ is Exclusive OR

Axioms & Rules

Convert Biimplication:
A <=> B ≡ (A => B) & (B => A)

Convert Implication:
A => B ≡ ~A | B

Convert Exclusive OR:
A ^ B ≡ (~A & B) | (A & ~B)

Distributivity:
A & (B | C) ≡ (A & B) | (A & C)
A | (B & C) ≡ (A | B) & (A | C)

De Morgan:
~(A & B) ≡ ~A | ~B
~(A | B) ≡ ~A & ~B

Absorption:
A | (A & B) ≡ A
A & (A | B) ≡ A

Inverse Absorption:
A | (~A & B) ≡ A | B
A & (~A | B) ≡ A & B

Associativity:
A | (B | C) ≡ A | B | C
A & (B & C) ≡ A & B & C

Commutativity:
A & B ≡ B & A
A | B ≡ B | A

For each step of the process, **select** a rule from the dropdown menu to simplify the following expression and **write** the simplified expression from that rule in the box next to it. Press the green Done button when expression cannot be simplified further.

Rule/Axiom Used	Expression to Simplify
	~(p | (~p & ~q)) & (p | r)
Inverse Absorption	~(p & ~q) & (p | r)

After replacing the subexpression in your last answer you should obtain
~(p | ~q) & (p | r)

Figure 1: Interface examples of all three ITSs employed in the present study. StoichTutor (top) and ORCCA (center) cover the domain of chemistry. LogicTutor (bottom) and ORCCA are ITS supporting formula-based problem solving compared to StoichTutor, which is a highly structured, ITS supporting fraction-based problem solving based on the factor-label method.

3.4. Coding SRL Constructs

The dataset analyzed comprised individual student transactions encoded in log data (e.g., attempts, hint requests) from all three tutoring systems, along with think-aloud transcripts. Student actions within each ITS were logged into PSLC DataShop (Koedinger et al., 2010). Whisper, an open-source transcription model for voice, generated think-aloud transcripts, which segmented utterances with start and end timestamps. Error and accuracy reports are discussed in (Radford et al., 2023). Whisper is capable of transcribing German, but we found it unsuitable for this study due to its lower voice recognition accuracy and the poor audio quality of the German recordings. As a result, we manually segmented and transcribed the German recordings using the FOLKER program, adhering to the transcription guidelines of the "Gesprächsanalytisches Transkriptionssystem 2" (GAT2) (Selting et al., 2011). Segmentation criteria included (a) changes in speaker between the user and experiment conductor, (b) pauses between words and phrases, and (c) shifts in content, such as transitions between problems. These segmentations are comparable to those generated by Whisper during transcription.

To be able to relate the correctness of student actions in the tutor to their prior utterances' SRL codes, as done in Zhang et al. (2024a), we concatenated multiple utterances falling between each pair of consecutive timestamped student-tutor transactions. Synchronization of log data and think-aloud transcriptions was ensured by coding a reference tutor transaction by a coder familiar with the software, which allows for synchronization with no more than a 1-second error margin. Access to the American log data and anonymized synchronized think-aloud transcripts is available upon request through DataShop (Koedinger et al., 2010) as datasets #5371 and #5820 The German dataset is not available due to data protection terms in the initial data collection, in line with relevant regulations.

Concatenated utterances were annotated following a coding scheme (see Table 1) aligning with the four-stage SRL model by Winne and Hadwin (1998): *processing information, planning, enacting*, and *realizing errors*. These categories, focusing on relevant behaviors within problem-solving learning environments, represent a subset of SRL processes within each model stage. The coding categories reflect critical behaviors at different stages of problem-solving learning, allowing us to examine learners' cognitive activities during information processing, planning, enacting conceptual actions, and realizing errors. Coding at this level enables observation of cognitive operations that are usually inferred from multiple actions and verbalizations. Table 1 outlines the coding categories and related behaviors.

Table 1: SRL categories, behaviors, and example utterances. Utterances were coded on the category level.

SRL Category	Behavior	Example Utterance
Processing Information	· Assemble information *The utterance demonstrates behaviors where students read or re-read a question, hints, or feedback provided by the system* · Comprehend information *The utterance demonstrates behaviors where students repeat information provided by the system with a level of synthesis*	"Let's figure out how many hydrogen items are in a millimole of water molecule H2O molecules. Our result should have three significant features."
Planning	· Identify goals and form plans *The utterance reflects behaviors where students verbalize a conceptual plan of how they will solve the problem*	"Okay, now I think I have to use commutative first. I'll just see what the problem wants."
Enacting	· Verbalize previous action *The utterance reflects students' behaviors where they verbalize an action that has just been carried out explaining what they did* · Announce the next action	"Two molecules of this. How many atoms in a minimum molecule of M

SRL Category	Behavior	Example Utterance
	The utterance reflects student behaviors where they verbalize a concrete and specific action that they will do next	mole? 61023 divided by 2. 3.0115."
Realizing Errors	· Realize something is wrong *The utterance demonstrates instances where students realize there is a mistake in the answer or the process with or without external prompting (i.e., tutor feedback)*	"It's incorrect. What's happened? It is the thousand in the wrong spot…No, the thousand is correct, so what am I doing wrong?"

For the English sample, two researchers (first and second author) coded 162 concatenated utterances separately to establish reliability while reviewing log data of associated segments. Once a reliable inter-rater reliability had been reached (*K*processing = 0.78, *K*planning = 0.90, *K*enacting = 0.77, *K*errors = 1.00), they then individually coded the remaining English utterances, double-coding any lacking agreement within the inter-rater iteration. Given this established reliability and the standardization of platforms and study procedures between the English and German sample, the first author, fluent in both English and German, then proceeded to code all German utterances.

3.5. Model Training

To train models that detect the four categories of SRL processes, we first vectorized both the English and German utterances using text-embedding-3-short (Neelakantan et al., 2022), a pretrained sentence embedding model that converts textual input into a high-dimensional vector with a length of 1,536. This sentence embedding model, using deep learning techniques and based on a transformer architecture, captures the semantic meaning of the text by analyzing the text and its surrounding words and phrases and then creates a contextualized numerical representation of the text. The model is trained on a diverse and extensive corpus, including multiple languages such as German, and it has demonstrated a reliable and comparable performance in processing different languages (OpenAI, n.d.).

Using the embeddings and the manually coded labels, we trained machine learning models to classify the presence or absence of each SRL category in each utterance in each language. A single-hidden-layer ReLU neural network (28 intermediate units) with sigmoid output was trained using the Adam optimizer (learning rate = 0.01) for 30 epochs (batch size = 10) to perform binary classification using binary cross-entropy loss. During model training, we performed student-level cross validation. Specifically, students' utterances were split into 5 folds with each student's utterances nested within one fold. Four folds were used to train a model, which was then applied and evaluated on the fifth (test) fold using the Area Under the ROC Curve (AUC). This step was repeated five times. We averaged the AUC across the five test folds and computed the standard deviation.

3.6. Transfer Tasks

To examine the transferability of the models across languages, we conducted three sets of analyses to examine 1) if these models transfer across languages, 2) if there is a difference in transferability in language when students are more or less familiar with the instructional strategy scaffolded by the platform, and 3) if these models transfer across *both* languages and subject domains *simultaneously*.

To answer RQ1 on model transferability across languages, we trained models using one language and then tested them on the other language. Since no utterances were collected from LogicTutor from German students, this comparison used only data from the chemistry tutors (StoichTutor and ORCCA). Specifically, for each SRL category, we first trained models using English utterances collected from StoichTutor and ORCCA and then tested them on the German data. Counterpart models were also evaluated where German data was used to train models which were then tested on the English utterances.

To investigate if the transferability varies between platforms with familiar and unfamiliar forms and degrees of scaffolding (RQ2), we compared the performance of the models between the two platforms used in the German data. One platform (StoichTutor) only supports the factor-label method, a fraction-based problem-solving strategy unfamiliar to German students but common in the United States (Borchers et al., 2025), while the other platform (ORCCA) accommodates multiple strategies. Therefore, separating and comparing the classification in the transfer task in RQ1 according to which platform the student used allows us to assess this transferability.

To answer RQ3 regarding language transfer when applied to a different learning platform design and domain, we take models trained on English language and interactions with LogicTutor and apply it to the German chemistry sample (using StoichTutor and ORCCA) used in RQ1. We also evaluate the reverse transfer, that is, a model trained on German chemistry data applied to English data from LogicTutor. In all cases, to express the reliability of transfer model performance based on AUC, we used an implementation of the closed-form estimation for AUC 95% confidence intervals proposed by DeLong (Sun & Xu, 2014).

3.7. Thematic Error Analysis by Prediction Task

To better understand the limitations and types of errors related to language, domain, and tutoring system transfer of SRL prediction, we conducted a thematic analysis (Clarke & Braun, 2017) of error types. To do so, we first converted the predictions from continuous values representing probabilities to binary values, where each probability was rounded to the nearest integer. Specifically, if a predicted value was of 0.5 or above, it was rounded up and assigned the value 1 (present). Otherwise, it was rounded down and assigned the value 0 (not present). Misclassified errors were identified by comparing the binary predictions with the labels (ground truth).

Misclassified utterances were then exported into spreadsheets for each of the four SRL codes, including the confidence of the classifier (in probability), human and predicted label, and tutoring system context (preceding hint message, attempt, problem statement, and other tutor feedback). This context supported sensemaking and coding of errors.

Two research team members fluent in English and German independently coded all misclassifications grouped by prediction task (language transfer vs. language+domain transfer). After repeated reading of utterances and grouping, emergent themes were summarized in tables with examples. The coders met twice in total to iteratively discuss and refine each theme and their related theories of model failure, including the curation of representative utterances for each theme. During coding and discussions of themes, there were no constraints or expectations on whether themes would be shared across or distinct to the different prediction tasks or SRL codes.

4. Results

Before reporting results related to our three research questions, we report on the distribution and frequency of the ground truth of the four SRL categories, that is, the coding done by the research team members. Overall, enacting was elicited more frequently than all other SRL categories, followed by processing information, planning, and realizing errors. Notably, the overall count of SRL categories was higher in ORCCA and LogicTutor, which could be due to slightly longer utterance lengths (M = 19.97 words for StoichTutor, M = 23.68 words for ORCCA, and M = 26.18 words for LogicTutor). This difference in utterance length could be related to more infrequent actions in these tutors, as students generally worked with each tutor for comparable amounts of time based on the study design. We report on language-specific breakdowns in Table 2. German utterances tended to be longer on average for StoichTutor (German M = 22.37 and English M = 17.72) and for ORCCA (German M = 34.99 and English M = 13.56).

Table 2: Distribution and frequency of utterances and human-assigned SRL labels across platforms and samples, with each utterance being assigned anywhere between 0 and 4 SRL stages.

Language	ITS	N utterances	M words	% processing info	% planning	% enacting	% realizing errors
English	Stoich	469	17.72	13.22%	10.66%	22.6%	6.4%
English	ORCCA	162	13.56	20.99%	12.35%	19.14%	4.32%
English	Logic	324	26.18	25.62%	16.98%	30.86%	18.83%
German	Stoich	439	22.37	28.93%	13.67%	16.86%	7.74%
German	ORCCA	145	34.99	27.59%	7.59%	37.24%	2.07%

To establish a baseline in model performance for subsequent comparisons in model transfer, we evaluate the models with 5-fold student-level cross validation for each language. The average AUC and the standard deviation for each language and SRL category is reported in Table 3. Cross-validation results replicated the prior finding on English data that the OpenAI embedding model achieved satisfactory cross-validation performance (Zhang et al., 2024a) when trained on German data. These cross-validation results only included chemistry data to ensure comparability of cross-validation results between the German and English sample and because satisfactory cross-validation results for the English LogicTutor sample had been established in Zhang et al. (2024a).

Table 3: Average cross-validation performance measured in AUC between English and German utterances recorded during problem-solving with the chemistry tutors ORCCA and Stoich.

SRL Category	AUC English (SD)	AUC German (SD)
Process	0.893 (0.063)	0.863 (0.021)
Plan	0.878 (0.062)	0.826 (0.090)

SRL Category	AUC English (SD)	AUC German (SD)
Enact	0.821 (0.103)	0.792 (0.066)
Realizing Errors	0.885 (0.118)	0.894 (0.072)

As indicated by Table 3, cross-validation results were generally satisfactory and comparable between the English and German sample. Specifically, no SRL category exhibited differences larger than about one standard deviation in AUC performance across folds based on cross-validation split performance variation.

4.1. Task Model Transfer Across Languages (RQ1)

In answering RQ1, we investigate the generalizability of embedding-based machine learning models of SRL to another language on matched platforms and domains. Given the same set of platforms (StoichTutor and ORCCA) and a single domain (chemistry), we assessed the model performance by training a model on all training data of one sample (English vs. German) and evaluating performance on the other. The results are summarized in Table 4.

Table 4: Language transfer performance on matched platforms and domains measured in AUC, including 95% confidence intervals.

SRL Category	English to German	German to English
Process	0.807 [0.769, 0.845]	0.858 [0.822, 0.894]
Plan	0.781 [0.727, 0.835]	0.768 [0.718, 0.819]
Enact	0.660 [0.610, 0.711]	0.709 [0.653, 0.764]
Realizing Errors	0.849 [0.786, 0.912]	0.854 [0.790, 0.917]

Based on Table 4, three observations regarding model transfer performance between languages can be made. First, transfer performance was generally good, meaning that the AUCs were around 0.8 or higher, addressing RQ1. This was with the exception of the enact category, which was close to AUCs of 0.7, which is still satisfactory. However, compared to cross-validation results in Table 3 which represents the models' performance when trained and applied on the same language, there was some level of model degradation (about 0.1 AUC except for the realizing errors category) when models were transferred to a second language. Second, reliable differences in the transferability of models between the two languages were found only for the process and enact categories, which generalized slightly better from German to English than English to German (based on point estimates not overlapping with the other language's confidence interval).

Error analysis and theories of failure: Our thematic analysis yielded a single error theme that was common across multiple instances per prediction and transfer task. In this case, the primary type of misclassification observed when analyzing errors in language transfer was related to **linguistic differences in *phrasing chemistry concept*s between samples.** The underlying theory of failure relates to the discrepancies between expert-based translations (of the tutoring systems) and commonplace translation (implicit in large language models). One key example of this issue is the term “cancel” in English which usually is translated to “stornieren” or “kündigen” in German, as in the context of canceling a subscription. However, in the context of algebraic expressions and chemistry, canceling must be translated to “kürzen” or “streichen” in German, signifying acts that would relate to the “enact” code. As a result, enact utterances with “kürzen” or “streichen” were incorrectly classified. As these issues were particularly prevalent in the enact category, it may help explain the relatively poor transfer performance of this code. An example German utterance exemplifying this type of misclassification included*: Yes./Möglicherweise Zähler und Nenner kürzen, der den selben Stoff enthält. [Yes / possibly cancel numerator and denominator that contains the same substance]*. Here, the action word “canceling” (or “kürzen” in the original utterance) would be most commonly translated as “shortening,” which an English embedding model might not have learned correctly as a consequence. Similar patterns were observed for German-to-English transfer.

4.2. Language Transfer Generalizability to Lower vs. Higher Scaffolding Platforms (RQ2)

In answering RQ2, we study how well language transfer on a matched domain generalizes to tutoring systems with higher compared to lower levels of instructional scaffolding when the highly scaffolded instruction is unfamiliar to one population. Specifically, German students are less accustomed to the factor-label method directed and expected in the StoichTutor, which is an instructional approach in stoichiometry commonly found in US-American education but not in German education (Schmidt, 1997). The contrast between familiar vs. unfamiliar strategies in transfer is hence represented by the German to English vs. the English to German language transfer in StoichTutor. In contrast, ORCCA is built to recognize a greater variety of problem-solving strategies for chemistry problems (e.g., by allowing students to compound steps and transform equations flexibly). This flexibility could interfere less with model transfer performance. Hence, the scaffolding contrast (high vs. low) in language transfer is thus represented by comparing transfer from StoichTutor to ORCCA and vice versa. The results for English to German language transfer by platform are summarized in Table 5.

Table 5: Language transfer performance from English to German (E → G) and vice versa (G → E) on matched domains between platforms with high and low degrees of scaffolding measured in AUC, including 95% confidence intervals.

Tutor	**ORCCA**		**Stoich**	
Scaffolding	**Lower**		**Higher**	**Higher**
Familiarity	**Both**		**Unfamiliar**	**Familiar**
Language	**E → G**	**G → E**	**E → G**	**G → E**
Process	0.836 [0.766, 0.906]	0.867 [0.804, 0.931]	0.800 [0.755, 0.845]	0.853 [0.808, 0.897]
Plan	0.884	0.764	0.764	0.775

Tutor	ORCCA		Stoich	
Scaffolding	Lower		Higher	Higher
Familiarity	Both		Unfamiliar	Familiar
Language	E → G	G → E	E → G	G → E
	[0.826, 0.941]	[0.663, 0.865]	[0.702, 0.827]	[0.718, 0.833]
Enact	0.706 [0.620, 0.792]	0.701 [0.578, 0.823]	0.624 [0.559, 0.690]	0.714 [0.652, 0.776]
Realizing Errors	0.977 [0.928, 1.000]	0.932 [0.875, 0.988]	0.831 [0.762, 0.899]	0.840 [0.765, 0.915]

In line with Table 5, four key observations can be made regarding the model's transfer performance between languages, separately for tutoring systems with different levels of scaffolding. First, similar to RQ1, cross-language transfer performances generally remained high when considered separately for each platform, except for the enact category, where AUCs ranged from around 0.7 (ORCCA; lower scaffolding) to around 0.6 (StoichTutor; higher scaffolding with unfamiliar instruction). Second, performance was generally comparable between English to German and German to English in ORCCA, the less-scaffolded tutor (except for planning which was more accurate for English to German transfer). Third, performance was lower for processing information and enacting in the highly-scaffolded StoichTutor for students at the German university, who were unfamiliar with StoichTutor's instructional strategy, but not for students at the American university, who were familiar with its strategy. Planning and realizing errors, in contrast, had stable performance when transferred to either language in StoichTutor. Fourth, the realizing errors category exhibited generally higher AUCs in the less-scaffolded learning environment ORCCA than the highly-scaffolded StoichTutor. Taken together and answering RQ2, language transfer was overall higher for the lowly scaffolded tutoring system and also when the scaffolded strategy was familiar in the target population.

Error analysis and theories of failure: The primary type of misclassification observed when analyzing errors in language transfer by platform type was related to **anthropomorphization during processing of instruction** which was only present in the German sample and particularly common in StoichTutor. Specifically, we observed a high rate of classification errors in utterances when anthropomorphization was present. German students, unlike English students, often addressed the tutoring system as if it was a human tutor or collaborator in problem-solving. This issue was especially common in the process category, which exhibited a systematically lower generalizability in StoichTutor from English to German. We especially observed this issue when German students would restate hints and other forms of instruction (which were more extensive in StoichTutor due to its higher level of scaffolding) in their own words rather than reading them aloud verbatim. Translated quotes of a German student working with StoichTutor reads: *Der will, dass ich Kiloliter eintrage. [He wants me to enter Kiloliter.]* and *Der sagt zwar 283,889 einsetzen, aber dann sagt er mir: "Das ist falsch." [Why? He does* *say to substitute 283,889, but then he tells me "this is wrong."]* In contrast, American university students would not anthropomorphize the tutoring system at similar crossroads, for example, when realizing they made an incorrect problem-solving step attempt: *OK, so this is wrong.* Finally, it is worth noting that German students also referred to the tutor's color-coded feedback differently due to their anthropomorphizing of the tutoring system StoichTutor, whereby the tutoring systems would "tell the student a color:" *Aber er sagt grün. [But he says that it's green.]* We suspect that these language differences may have limited model transfer due to related distributional differences in the test set. We return to this discussion in Section 5.2.

4.3. Model Transfer Across Languages and Domains (RQ3)

In answering RQ3, we study how well language transfer works when additional domain and platform transfer is included. Past work with US students has shown degraded performance between platforms with AUCs of about 0.6 (except realizing errors, which was around 0.9) from chemistry to formal logic (Zhang et al., 2024a). How does this performance compare when additionally including language transfer? Performances by SRL code are summarized in Table 6.

Table 6: Language transfer performance from English to German and vice versa including domain transfer measured in AUC, including 95% confidence intervals.

SRL Category	**English Logic to German Chem**	**German Chem to English Logic**
Process	0.674 [0.626, 0.721]	0.635 [0.568, 0.702]
Plan	0.679 [0.619, 0.740]	0.711 [0.642, 0.779]
Enact	0.545 [0.490, 0.599]	0.506 [0.439, 0.573]
Realizing Errors	0.898 [0.849, 0.948]	0.869 [0.822, 0.916]

In line with Table 6, two observations can be made regarding model transfer performance between languages and domains. First, the realizing errors category exhibited no notable degradation in performance, maintaining AUCs well above 0.8. Second, in general, AUCs in other categories were around 0.6, aligning with domain transfer results from prior work (Zhang et al., 2024a) and indicating no notable further performance degradation based on language. One exception, however, was the enact category, where AUC was at chance (i.e., around 0.5).

Error analysis and theories of failure: The primary type of misclassification observed when analyzing errors in combined language and domain transfer was related to **differences in instructional context between domains, which were not obvious from transcriptions alone**. This issue was most commonly observed for hints, which were often incorrectly predicted to be planning when verbalized verbatim, given that reading (verbatim verbalization) is coded as processing according to our codebook. An example hint read by a German student and mistaken for planning by the model included: *Look for the substance that is in the numerator of the result.* Similarly, there were cases where planning is expressed, but is incorrectly predicted as processing information. While human coders could disambiguate if a plan is in direct relationship to a preceding hint based on log data (reading), a language model (as used in the fashion discussed here) does not have access to such context (but see Liu et al., 2024 for an example of how to incorporate this type of context). Further, the LLM cannot learn what a hint might be during domain transfer, as hints take on different forms between subject domains. As a consequence, our model was often unable to distinguish between the student reading a hint that suggests a plan (coded as processing) as opposed to the student genuinely coming up with a plan (coded as planning). We especially observed this pattern where verbalized plans included task-specific information, which differed between subject domains, such as given values: *Wir haben diese 100 Gramm Wasser, es sind 6 Gramm Alkohol darin und wenn ich jetzt wissen will, wie viele Mole in 1 Kilogramm sind. [We have these 100 grams of water, there are 6 grams of alcohol in it and if I now want to know how many moles are in 1 kilogram.],* or in the following case where the plan to convert from mole to grams is only tangentially mentioned: *Mol in gramm umwandeln…/Na, wahrscheinlich wieder hier ein Mol von dem P4O10. [Converting moles to grams.../Well, probably here again a mole of the P4O10.*] In LogicTutor, such given values typically take on the form of variables to which operators are applied, where planning was similarly often mistaken for processing: *not q and p okay further solve p or parenthesis not q and p with absorption which is then*. Taken together, subject domain differences, related to differences in tutor scaffolding (i.e., hints) as well as the verbalization of plans, limited domain transfer. As mentioned, incorporating log data into prediction might alleviate some of these issues.

5. Discussion

The present study investigated the transferability of LLM embedding-based prediction models of SRL from think-aloud data, a common form of assessing SRL during learning by problem solving. While past work has established that reliable classification of SRL from learner utterances using automated transcription is feasible, it is an open question to what degree LLM embeddings, which are generally multilingual, can be applied to the prediction of SRL in different languages where instructional and semantic differences may exist that domain-general foundation models may be unable to capture. To this end, we coded and evaluated matched think-aloud data of 26 students from German and American universities working with two intelligent tutoring systems for chemistry and coded their think-aloud utterances using the same coding scheme and through bilingual research team members. We systematically evaluated the transfer of machine learning models between languages to English think-aloud data from a third tutoring system from a different domain: formal logic. Our contributions are as follows.

5.1. Embedding-Based SRL Prediction Transfers Well to a Second Language

As our first contribution, we replicated prior work showing that models built on top of LLM embeddings from student utterances in English (Zhang et al., 2024a) can reliably classify SRL in a second language (German) and vice versa. We did not find systematic differences in classification accuracy, with AUC estimates around a good 0.8 AUC in both languages and across all SRL categories. This finding adds confidence to the notion that LLM embeddings are fundamentally multilingual and do not require fine-tuning to apply to language contexts other than English. However, it is worth noting that German, similar to English, is one of the dominant languages well-represented in the open web, representing LLMs' training corpus and among the languages LLMs like GPT-4 perform best on (Achiam et al., 2023). It is worth investigating in future work if our findings would replicate for languages that are less representative of the open web and more distal to English and German. Methods for improving LLM applications to so-called "low-resource" languages are being actively developed (Andersland, 2024).

As our second contribution, we demonstrated that LLM embedding-based models can transfer to a second language (RQ1) with satisfactory classification accuracy without fine-tuning or domain adaptation. We found little performance degradation compared to cross-validation on the source language. However, an exception to this was seen for the enact category, which exhibited poorer transfer performance. It is possible that further refinements to multilingual embeddings, such as learning a linear mapping from the German to English space and related embedding adaptations, would further improve transfer performance (Merullo et al., 2022). However, such improvements are beyond the scope of the present study; they should be the subject of future research.

5.2. Linguistic and National Differences in Instruction Constraint Language Transfer

As our third contribution, we advance the theoretical understanding of how concept resolution (i.e., the process by which a model identifies, disambiguates, and aligns the meaning of a concept expressed in different languages) in different languages can reduce the transfer of LLM embedding-based models to other languages. Systematic and qualitative analysis of model classification errors also explained why the enact category stood out: misclassified utterances tended to be related to differences in chemical expressions distinct from everyday terms. Specifically, as students learn chemistry, they learn concepts to distinguish everyday expressions from their different meanings in chemistry (e.g., solution and solving), which relate to distinct instructional goals (Heeg et al., 2020). These correspondences often differ between languages; for example, the act of removing redundant units from fractions is called “canceling” in English, which relates to “ending” something or “calling it even,” while the German equivalent "kürzen" is equivalent to “shortening” in everyday expression. Given their training on the broad open web, foundation models map words to their general purpose meaning rather than subject-specific meaning (Achiam et al., 2023). As a consequence, they are more likely to incorrectly classify utterances whose SRL category hinges on a language-specific understanding of chemistry concepts and their relation to these everyday terms. Similar issues in resolving concepts have been identified in emotional representations between cultures (Dudy et al., 2024) and fine-tuning has been suggested as a method to handle such cultural differences in concept resolution (Li et al., 2024). Therefore, future work could improve the generalizability of SRL predictions between cultures by incorporating knowledge about distinct mappings of everyday terms to concepts into the LLM through fine-tuning. Another approach would be machine-translating utterances using an expert-defined dictionary, which avoids translating terms to their (incorrect) general-use equivalent in the target language, but this approach would be much more time-consuming, due to the need to define a dictionary. Future work could compare both approaches to adapt to cultural distinctions in domain-specific concepts.

As our fourth contribution, we identified national familiarity with highly-scaffolded instruction as another potential hindrance to the successful cross-language transfer of LLM embedding-based SRL predictions (RQ2). A highly scaffolded tutoring system that instructs learners to perform a problem-solving strategy unknown to German learners (factor-labeling method; Schmidt, 1997) led to low generalization for a model trained on American university students familiar with that strategy. German university students tended to anthropomorphize the tutoring system with comparatively high levels of scaffolding, meaning that they often pointed out how “he,” the tutoring system, would point out their errors or want them to do something. American university students, whose utterances were similarly well predicted based on German data compared to German utterances, did not anthropomorphize their tutoring system. We suspect that this difference is due to language conventions, where German students would address and react to scaffolding in the tutoring systems differently from American students, as anthropomorphization was especially present in StoichTutor. There are different interpretations for this difference in transfer performance. As one explanation, the anthropomorphization by German students might have degraded performance because these types of utterances were not found in the training sample and out of distribution, leading to unpredictable model outputs similar to adversarial examples (Goodfellow et al., 2014). As an alternative explanation, because German students were unfamiliar with the strategy provided by the highly scaffolded chemistry tutor, they might have generally exhibited less coherent problem-solving actions, which degraded the quality of labels and hence transfer. However, this theory can be partially ruled out because a model based on German utterances, combined with data from the other chemistry tutoring system, generalized reasonably well to English data. Larger data samples enabling a systematic comparison of training models only on one type of tutoring system rather than both could further test the second hypothesis. However, such a comparison was not the objective of the present study, as it is confounded with platform transfer differences. In all cases, our findings add nuance to prior research indicating that SRL processes systematically differ by the degree of scaffolding (i.e., problem-solving support) in tutoring systems (Zhang et al., 2024b) by noting that differences in cultural instructional norms (e.g., due to unfamiliarity with the tutoring system's domain-specific instruction) can help explain differences in SRL.

5.3. Considering Instructional Context at Prediction Could Further Boost Performance

As our fifth contribution, we found that language transfer with a joint domain transfer from chemistry to formal logic and vice versa was possible at above-chance model performance (RQ3). This was with the exception of the enact category, where classification performance was not distinguishable from random guessing, potentially due to reasons similar to the concept resolution issue from everyday terms suggested in the discussion of RQ1. However, realizing errors was always predicted across contexts and language without notable loss in model performance. One potential reason for why this category was classified well across contexts is because realizing *that* something is wrong is not dependent or related to the instructional context of each tutoring system, as described next.

Our systematic error analysis revealed that model performance deficiencies could often be attributed to the neglect of instructional context during prediction. Specifically, the models struggled to differentiate between student processing of instruction and the formulation of independent plans during domain transfer. This difficulty arises because much of the instruction provided through hints in tutoring systems is metacognitive, offering specific plans and next steps. Hence, it is difficult for a model to differentiate a plan generated by the learner from a plan the learner reads out from on-screen instruction without knowing what hints are likely to be given by the tutoring systems. While predictive models trained on chemistry data may recognize common instructions, such as hints and their translation, these models fail when applied to new domains and platforms, where an entirely different set of hints is used. As a result, the models cannot effectively distinguish between processing and planning without the inclusion of instructional content. A hypothesis for future research is to incorporate language from instructional materials (e.g., hint texts) as concatenated embedding vectors during model training, alongside domain adaptation techniques (Wang et al., 2019). Similar concatenation techniques have been successfully employed to classify dialog acts in human tutoring contexts (Lin et al., 2023).

5.4. LIMITATIONS AND FUTURE WORK

This study has multiple limitations that should be acknowledged. First, the selection of three tutoring systems and two domains of instruction (i.e., chemistry and logic) constrained the scope of our analysis. This limitation applies to much educational data mining research, where creating models that generalize to different learning environments remains a central challenge of the field (Baker, 2019). Consequently, it could be that the lower predictive generalizability of our models to German students working with the StoichTutor and their anthropomorphization of the tutoring system (which was also present in ORCCA, but to a lesser extent) could be due to factors other than German students being unfamiliar with the instructional strategy of the StoichTutor and its comparatively high degree of scaffolding. Future work should consolidate our interpretation that unfamiliar instruction and their related differences in SRL are linked and apply to domains outside of chemistry. One lens to study these differences could be by classifying learning environments with respect to their degree of scaffolding (Zhang et al., 2024b). Future research could also explore different instructional contexts, input types, and other modalities to better understand their impact on SRL prediction and transfer. Our findings motivate the hypothesis that instructional context (i.e., integrating tutoring system instruction into prediction through text or otherwise) may be more critical for SRL prediction in highly scaffolded environments, warranting further investigation.

A significant strength of our study lies in the matched platforms used for analysis: all were tutoring systems with instructional support through hints and immediate feedback. While this approach grants a certain amount of standardization in data processing and comparison between the tutoring systems, it also limits the generalizability of our findings to educational technologies with less instructional support, which may require a greater need for SRL. Specifically, tutoring systems generally provide a high level of instructional support through immediate correctness feedback and hints, but this differs from many other types of learning environments. Future research is encouraged to apply our methodology to other learning environments, including collaborative learning systems (Borchers et al., 2024a) and game-based learning environments (Richey et al., 2024), where language data can represent dialog acts between learners or self-explanations that are expected to relate to learning. Similarly, given that our think-aloud context may be characterized by early stages of using a tutoring system, future research is also encouraged to study contexts with longer-term use.

Another limitation is that this study did not systematically compare different methods of using LLMs for classification, as it aimed to establish the extent of general predictive capabilities of such models for cross-national language transfer in SRL prediction. Next to fine-tuning and expert-based translation ahead of embedding, which we identified as potential remedies to the issue of concept resolution we identified in RQ1, this could include a comparison of embedding-based prediction to prediction through generation and LLM prompt engineering. Although generation could offer richer, context-driven predictions, we opted for simpler, non-generative models to avoid issues like hallucination and to maintain cost-effectiveness. In addition, the literature suggests that for language prediction tasks, simple bag-of-words (BOW) models can outperform more complex generative approaches, as highlighted by Hutt et al. (2024), which could further be benchmarked against our current approach to prediction.

Our method for automated SRL classification from text has several promising applications to enhance and better understand learning. First, for formal learning environments such as classrooms, where live transcription is challenging or student verbalizations during learning with technology are uncommon (with few exceptions; Grawemeyer et al., 2015), our approach could efficiently annotate training data for log-based SRL detectors obtained in think-aloud settings.

These detectors could then enable adaptive instruction by providing scaffolds tailored to learners' current SRL phase (e.g., when learners omit planning stages; Borchers et al., 2024b). Similarly, such detectors could enable the study of how scaffolding interacts with SRL at a larger scale. For instance, one hypothesis from large-scale evidence in tutoring systems is that their comparatively high degree of scaffolding compared to other educational technologies may make the impact of student-level SRL difference on learning less pronounced (Koedinger et al., 2023). Second, our method could help procure learning analytics by tracking longitudinal improvements in learners' self-regulation and identifying instructional elements that enhance SRL (Azevedo, 2014). Third, our methodology could be extended to additional domains and languages, advancing SRL theory by revealing how verbalized self-regulation manifests across different contexts.

6. Conclusion

The present study contributed a novel platform-matched comparison of automated SRL classification with foundation model embeddings based on think-aloud data across languages, instructional contexts, nations, and domains. Our findings show that embedding-based SRL prediction based on automated processing of think-aloud data is feasible at a reliable accuracy of around 0.8 AUCs and in multiple languages. Our empirical analysis further adds theoretical contributions to understanding the limitations of large language models for the classification of SRL. First, we identified the resolution of domain-specific concepts in everyday language as a critical challenge in language transfer, where we noted slight performance degradation between English and German. Expert-based translations of language or fine-tuning offer promising avenues for alleviating this issue. Second, including instructional context, especially in highly scaffolded learning environments, and applying models to instructional domains not represented in the training data may further boost classification performance and generalizability. Our findings highlight a generally lower transfer performance for a tutoring system with higher degrees of scaffolding to a population unfamiliar with the instructional strategy that is scaffolded for. This transfer performance underscores the importance of adapting predictive SRL models to a population's characteristics and instructional context. However, our empirical findings add confidence and move the field closer to predictive capabilities for SRL that generalize to domains, languages, platforms, and national contexts, which may speed improvements to instructional effectiveness of future advanced learning systems delivering SRL support.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., & Zoph, B. (2023). *GPT-4 technical report (arXiv)*. https://arxiv.org/abs/2303.08774

Aleven, V. (2010). Rule-based cognitive modeling for intelligent tutoring systems. In *Advances in Intelligent Tutoring Systems* (pp. 33–62). Springer Berlin Heidelberg.

Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education 26, 1*, 224–269.

ALEVEN, V., MCLAREN, B., ROLL, I., & KOEDINGER, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education 16*, 2, 101–128.

ANDERSLAND, M. (2024). Amharic LLaMA and LLaVA: Multimodal LLMs for low resource languages (*arXiv Preprint*). https://arxiv.org/abs/2403.06354

ARAKA, E., MAINA, E., GITONGA, R., & OBOKO, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—Systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning*, *15*, 1–21.

ARTETXE, M., & SCHWENK, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

AZEVEDO, R. (2014). Issues in dealing with sequential and temporal characteristics of self-and socially-regulated learning. *Metacognition and Learning*, *9*, 217–228.

AZEVEDO, R., JOHNSON, A., CHAUNCEY, A., GRAESSER, A., ZIMMERMAN, B., & SCHUNK, D. (2011). Use of hypermedia to assess and convey self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 327–342). Routledge. New York, NY.

AZEVEDO, R., TAUB, M., & MUDRICK, N. V. (2017). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 323–337). Routledge. New York, NY.

BAKER, R. S. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1), 1–17.

BAKER, R. S., CORBETT, A. T., & KOEDINGER, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems* (pp. 531-540). Springer Berlin Heidelberg.

BANNERT, M., REIMANN, P., & SONNENBERG, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*, 161–185.

BISWAS, G., JEONG, H., KINNEBREW, J. S., SULCER, B., & ROSCOE, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, *5*(02), 123–152.

BORCHERS, C., FLEISCHER, H., YARON, D. J., MCLAREN, B. M., SCHEITER, K., ALEVEN, V., & SCHANZE, S. (2025). Problem-solving strategies in stoichiometry across two intelligent tutoring systems: A cross-national study. *Journal of Science Education and Technology*, *34*(2), 384-400.

BORCHERS, C., YANG, K., LIN, J., RUMMEL, N., KOEDINGER, K. R., & ALEVEN, V. (2024). Combining dialog acts and skill modeling: What chat interactions enhance learning rates during AI-supported peer tutoring? In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 157–168).

BORCHERS, C., ZHANG, J., BAKER, R. S., & ALEVEN, V. (2024). Using think-aloud data to understand relations between self-regulation cycle characteristics and student performance in intelligent tutoring systems. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 529-539).

CLARKE, V., & BRAUN, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297–298.

CLEARY, T. J., & CHEN, P. P. (2009). Self-regulation, motivation, and math achievement in middle school: Variations across grade level and math context. *Journal of School Psychology*, *47*(5), 291–314.

CONNEAU, A., LAMPLE, G., RANZATO, M. A., DENOYER, L., & JÉGOU, H. (2017). Word translation without parallel data (*arXiv Preprint*). http://arxiv.org/abs/1710.04087

CROSSLEY, S. A., LOUWERSE, M. M., MCCARTHY, P. M., & MCNAMARA, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15–30.

DARVISHI, A., KHOSRAVI, H., SADIQ, S., & GAŠEVIĆ, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, *53*(4), 844–875.

DE ARAUJO, A., PAPADOPOULOS, P. M., MCKENNEY, S., & DE JONG, T. (2023). Automated coding of student chats, a trans-topic and language approach. Computers and Education: Artificial Intelligence, 4, 100123.

DODDAPANENI, S., RAMESH, G., KHAPRA, M. M., KUNCHUKUTTAN, A., & KUMAR, P. (2021). A primer on pretrained multilingual language models (*arXiv Preprint*). http://arxiv.org/abs/2107.00676

DRÖSE, J., & PREDIGER, S. (2018). Strategien für Textaufgaben. Fördern mit Info-Netzen und Formulierungsvariationen. *Mathematik Lehren*, *206*, 8–12.

DUDY, S., AHMAD, I. S., KITAJIMA, R., & LAPEDRIZA, A. (2024). Analyzing Cultural Representations of Emotions in LLMs through Mixed Emotion Survey (*arXiv Preprint*). http://arxiv.org/abs/2408.02143

ERICSSON, K. A., & SIMON, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

FIROOZI, T., MOHAMMADI, H., & GIERL, M. J. (2024). Using automated procedures to score educational essays written in three languages. *Journal of Educational Measurement*, *62*(1), 33–56.

FOX, M. C., ERICSSON, K. A., & BEST, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316.

GONEN, H., RAVFOGEL, S., ELAZAR, Y., & GOLDBERG, Y. (2020). It's not Greek to mBERT: Inducing word-level translations from multilingual BERT (*arXiv Preprint*). http://arxiv.org/abs/2010.08275

GOODFELLOW, I. J., SHLENS, J., & SZEGEDY, C. (2014). Explaining and harnessing adversarial examples (*arXiv Preprint*). http://arxiv.org/abs/1412.6572

GRAWEMEYER, B., GUTIERREZ-SANTOS, S., HOLMES, W., MAVRIKIS, M., RUMMEL, N., MAZZIOTTI, C., & JANNING, R. (2015). Talk, tutor, explore, learn: Intelligent tutoring and exploration for robust learning. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED)*.

GREENE, J. A., & AZEVEDO, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, *45*(4), 203–209.

GREENE, J. A., DEEKENS, V. M., COPELAND, D. Z., & YU, S. (2017). Capturing and modeling self-regulated learning using think-aloud protocols. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 323–337). Routledge. New York, NY.

GREENE, J. A., ROBERTSON, J., & COSTA, L. J. C. (2011). Assessing self-regulated learning using think-aloud methods. In D. H. Schunk & B. J. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 327–342). Routledge. New York, NY.

HATALA, M., NAZERI, S., & KIA, F. S. (2023). Progression of students' SRL processes in subsequent programming problem-solving tasks and its association with tasks outcomes. *The Internet and Higher Education*, *56,* 100881.

HE, J., & LI, X. (2024). Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 17819-17832).

HEEG, J., HUNDERTMARK, S., & SCHANZE, S. (2020). The interplay between individual reflection and collaborative learning–seven essential features for designing fruitful classroom practices that develop students' individual conceptions. *Chemistry Education Research and Practice*, *21*(3), 765–788.

HEIRWEG, S., DE SMUL, M., MERCHIE, E., DEVOS, G., & VAN KEER, H. (2020). Mine the process: Investigating the cyclical nature of upper primary school students' self-regulated learning. *Instructional Science*, *48*(4), 337–369.

HU, J., RUDER, S., SIDDHANT, A., NEUBIG, G., FIRAT, O., & JOHNSON, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning* (pp. 4411-4421). PMLR.

HUANG, K. H., AHMAD, W. U., PENG, N., & CHANG, K. W. (2021). Improving zero-shot cross-lingual transfer learning via robust training (*arXiv Preprint*). http://arxiv.org/abs/2104.08645

HUTT, S., DEPIRO, A., WANG, J., RHODES, S., BAKER, R. S., HIEB, G., SETHURAMAN, S., OCUMPAUGH, J., & MILLS, C. (2024). Feedback on feedback: Comparing classic natural language processing and generative AI to evaluate peer feedback. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 55-65).

HUTT, S., OCUMPAUGH, J., MA, J., ANDRES, A. L., BOSCH, N., PAQUETTE, L., BISWAS, G., & BAKER, R. S. (2021). Investigating SMART models of self-regulation and their impact on learning. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, pp. 580-587.

KING, E. C., BENSON, M., RAYSOR, S., HOLME, T. A., SEWALL, J., KOEDINGER, K. R., ALEVEN, V., & YARON, D. J. (2022). The open-response chemistry cognitive assistance tutor system: Development and implementation. *Chemistry Education Research and Practice*, *99*(2), 546–552.

KIZILCEC, R. F., SALTARELLI, A. J., REICH, J., & COHEN, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, *355*(6322), 251–252.

KOEDINGER, K. R., BAKER, R. S., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., & STAMPER, J. (2010). A data repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining* (pp. 43–56). CRC Press. Boca Raton, FL.

KOEDINGER, K. R., CARVALHO, P. F., LIU, R., & MCLAUGHLIN, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), e2221311120.

KOVANOVIĆ, V., JOKSIMOVIĆ, S., MIRRIAHI, N., BLAINE, E., GAŠEVIĆ, D., SIEMENS, G., & DAWSON, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389-398).

LABUHN, A. S., ZIMMERMAN, B. J., & HASSELHORN, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning*, *5*, 173–194.

LI, B., HE, Y., & XU, W. (2021). Cross-lingual named entity recognition using parallel corpus: A new approach using XLM-RoBERTa alignment (*arXiv Preprint*). http://arxiv.org/abs/2101.11112

LI, C., CHEN, M., WANG, J., SITARAM, S., & XIE, X. (2024). CultureLLM: Incorporating cultural differences into large language models (*arXiv Preprint*). http://arxiv.org/abs/2402.10946

LIM, L., BANNERT, M., VAN DER GRAAF, J., MOLENAAR, I., FAN, Y., KILGOUR, J., MOORE, J., & GAŠEVIĆ, D. (2021). Temporal assessment of self-regulated learning by mining students' think-aloud protocols. *Frontiers in Psychology*, *12*, 749749.

LIN, J., TAN, W., DU, L., BUNTINE, W., LANG, D., GAŠEVIĆ, D., & CHEN, G. (2023). Enhancing educational dialogue act classification with discourse context and sample informativeness. *IEEE Transactions on Learning Technologies*, *17*, 258-269.

LIU, X., ZHANG, J., BARANY, A., PANKIEWICZ, M., & BAKER, R. S. (2024). Assessing the potential and limits of large language models in qualitative coding. In *International Conference on Quantitative Ethnography* (pp. 89–103).

MCLAREN, B. M., DELEEUW, K. E., & MAYER, R. E. (2011a). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies*, *69*(1–2), 70–79.

MCLAREN, B. M., DELEEUW, K. E., & MAYER, R. E. (2011b). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education*, *56*(3), 574–584.

MCLAREN, B. M., LIM, S. J., GAGNON, F., YARON, D., & KOEDINGER, K. R. (2006). Studying the effects of personalized language and worked examples in the context of a web-based intelligent tutor. In *International Conference on Intelligent Tutoring Systems* (pp. 318-328). Berlin, Heidelberg: Springer Berlin Heidelberg.

MCNAMARA, D. S., ALLEN, L. K., CROSSLEY, S. A., DASCALU, M., & PERRET, C. A. (2017). Natural language processing and learning analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 93–104). Society for Learning Analytics Research (SoLAR).

MERULLO, J., CASTRICATO, L., EICKHOFF, C., & PAVLICK, E. (2022). Linearly mapping from image to text space (*arXiv Preprint*). http://arxiv.org/abs/2209.15162

MOLENAAR, I., HORVERS, A., & BAKER, R. S. (2021). What can moment-by-moment learning curves tell about students' self-regulated learning? *Learning and Instruction*, *72*, 101206.

NASIAR, N., BAKER, R. S., ZOU, Y., ZHANG, J., & HUTT, S. (2023). Modeling problem-solving strategy invention (PSSI) behavior in an online math environment. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. AIED 2023*. Communications in computer and information science (Vol. 1831, pp. 453–459). Springer, Cham.

NEELAKANTAN, A., XU, T., PURI, R., RADFORD, A., HAN, J. M., TWOREK, J., YUAN, Q., TEZAK, N., KIM, J. W., HALLACY, C., HEIDECKE, J., SHYAM, P., POWER, B., ELOUNDOU NEKOUL, T., SASTRY, G., KRUEGER, G., SCHNURR, D., PETROSKI SUCH, F., HSU, K., & WENG, L. (2022). Text and code embeddings by contrastive pre-training (*arXiv Preprint*). http://arxiv.org/abs/2201.10005

NOTA, L., SORESI, S., & ZIMMERMAN, B. J. (2004). Self-regulation and academic achievement and resilience: A longitudinal study. *International Journal of Educational Research*, *41*(3), 198–215.

OCUMPAUGH, J., BAKER, R., GOWDA, S., HEFFERNAN, N., & HEFFERNAN, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501.

OMRAN, T. M., SHAREF, B. T., GROSAN, C., & LI, Y. (2023). Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, *143*, 102106.

OPENAI. (n.d.). Embeddings. https://platform.openai.com/docs/guides/embeddings

PANADERO, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8,* 422.

PAQUETTE, L., & BAKER, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 585–597.

RADFORD, A., KIM, J. W., XU, T., BROCKMAN, G., MCLEAVEY, C., & SUTSKEVER, I. (2023). Robust speech recognition via large-scale weak supervision. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 28492–28518). Curran Associates, Inc.

RICHEY, J. E., NGUYEN, H. A., MEHRVARZ, M., ELSE-QUEST, N., ARROYO, I., BAKER, R. S., STEC, H., HAMMER, J., & MCLAREN, B. M. (2024). Understanding gender effects in game-based learning: The role of self-explanation. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: AIED 2024*. Lecture Notes in Computer Science (Vol. 14829, pp. 206–219). Springer, Cham.

SAINT, J., FAN, Y., GAŠEVIĆ, D., & PARDO, A. (2022). Temporally-focused analytics of self-regulated learning: A systematic review of literature. *Computers and Education: Artificial Intelligence*, *3*, 100060.

SAINT, J., WHITELOCK-WAINWRIGHT, A., GAŠEVIĆ, D., & PARDO, A. (2020). Trace-SRL: A framework for analysis of microlevel processes of self-regulated learning from trace data. *IEEE Transactions on Learning Technologies*, *13*(4), 861–877.

SCHMIDT, H. J. (1997). An alternate path to stoichiometric problem solving. *Research in Science Education*, *27*, 237–249.

SCHOOLER, J. W., OHLSSON, S., & BROOKS, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166.

SCHUNK, D. H., & ZIMMERMAN, B. J. (Eds.). (2011). Handbook of Self-regulation of Learning and Performance. Taylor & Francis.

SELTING, M., AUER, P., & BARTH-WEINGARTEN, D. (2011). A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung: Online-Zeitschrift Zur Verbalen Interaktion*, *12*, 1–51.

SUN, X., & XU, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, *21*(11), 1389–1393.

TAUSCZIK, Y. R., & PENNEBAKER, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

VAN DER GRAAF, J., LIM, L., FAN, Y., KILGOUR, J., MOORE, J., GAŠEVIĆ, D., BANNERT, M., & MOLENAAR, I. (2022). The dynamics between self-regulated learning and learning outcomes: An exploratory approach and implications. *Metacognition and Learning*, *17*(3), 745–771.

WANG, Z., DU, B., & GUO, Y. (2019). Domain adaptation with neural embedding matching. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(7), 2387–2397.

WEI, J., TAY, Y., BOMMASANI, R., RAFFEL, C., ZOPH, B., BORGEAUD, S., YOGATAMA, D., BOSMA, M., ZHOU, D., METZLER, D., CHI, E. H., HASHIMOTO, T., VINYALS, O., LIANG, P., DEAN, J., & FEDUS, W. (2022). Emergent abilities of large language models (*arXiv Preprint*). http://arxiv.org/abs/2206.07682

WINNE, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *4*, 267–276.

WINNE, P. H. (2017). Learning analytics for self-regulated learning. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 241–249). Society for Learning Analytics Research (SoLAR).

WINNE, P. H., & BAKER, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, *5*(1), 1–8.

WINNE, P. H., & HADWIN, A. F. (1998). Studying as self-regulated learning. In *Metacognition in Educational Theory and Practice* (pp. 291–318). Routledge. New York, NY.

YENDURI, G., RAMALINGAM, M., CHEMMALAR SELVI, G., SUPRIYA, Y., SRIVASTAVA, G., MADDIKUNTA, P. K. R., DEEPTI RAJ, G., JHAVERI, R. H., PRABADEVI, B., WANG, W., VASILAKOS, A. V., & GADEKALLU, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, *12*, 54608–54649.

ZAMBRANO, A. F., ZHANG, J., & BAKER, R. S. (2024). Investigating algorithmic bias on Bayesian knowledge tracing and carelessness detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 349-359).

ZHANG, J., ANDRES, A., BAKER, R., & CLOUDE, E. B. (2023). Leveraging natural language processing to detect gaming the system in open-ended questions in a math digital learning game. *33rd Annual Meeting of the Society for Text and Discourse*.

ZHANG, J., ANDRES, J. M. A. L., HUTT, S., BAKER, R. S., OCUMPAUGH, J., NASIAR, N., MILLS, C., BROOKS, J., SETHURAMAN, S., & YOUNG, T. (2022). Using machine learning to detect SMART model cognitive operations in mathematical problem-solving process. *Journal of Educational Data Mining*, *14*(3), 76–108.

ZHANG, J., BAKER, R. S., ANDRES, J. M., HUTT, S., & SETHURAMAN, S. (2023). Automated multi-dimensional analysis of peer feedback in middle school mathematics. In *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023* (pp. 221-224). International Society of the Learning Sciences.

ZHANG, J., BORCHERS, C., ALEVEN, V., & BAKER, R. S. (2024). Using large language models to detect self-regulated learning in think-aloud protocols. In B. Paassen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)* (pp. 157–168).

ZHANG, J., BORCHERS, C., & BARANY, A. (2024). Studying the interplay of self-regulated learning cycles and scaffolding through ordered network analysis across three tutoring systems. In Y. J. Kim & Z. Swiecki (Eds.), *Advances in quantitative ethnography (ICQE 2024)*. Communications in computer and information science (Vol. 2278, pp. 231–246). Springer, Cham.

ZHENG, L. (2016). The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A meta-analysis. *Asia Pacific Education Review*, *17*, 187–202.

ZIMMERMAN, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3–17.

ZIMMERMAN, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–40). Academic Press. San Diego, CA.