International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 08 - Issue 06 || June 2022 || PP. 09-17

Exploring Models & Techniques for Ai-Driven Assessment

Ayse Arslan

Abstract: Artificial intelligence (AI) has a large and increasing role for personalized training systems. This paper brings both the issues with the standard assessment paradigm and the challenges associated with AI and assessment into a deeper conversation that will ultimately improve assessment practices more generally. It highlights the need for development of actionable and personalized explanations, further incorporation of human-centric design in development of learning tools, rigorous evaluation of the impact of incorporating AI into training and ultimately advancing towards development of trustworthy training systems. The suggested platform architecture consists of an open-source Python API for specifying the workflow of an experiment (the “API”), and a Platform-as-a-Service (PaaS), which is a running instance of back-end infrastructure coupled with computational resources and a large training dataset. This study also aims to synthesize an agenda for future research on AI-driven assessmenttechniques.

Introduction

Problem-based learning (PBL) is an instructional approach that exemplifies authentic learning and emphasizes problem-solving within richly contextualized settings. In PBL, users assume primary responsibility for their own development while trainers provide facilitation and use of additional tools or technologies such as a dashboard can serve as a separate medium allowing trainers to view, just-in-time, how their users are doing overall.

SA (sequence analysis), specifically sequence clustering and comparison, is a promising avenue to the process-focused study of individual problem-solving, informing more effective personalized learning. The problem, however, is that many SA techniques cannot beinterpreted, so that it can be difficult for a stakeholder to understand the data or how to act on the results. Further, without anunderstanding of the data, it can be difficult for a stakeholder to infer how an algorithm or statistical method is understanding thedata or why a statistical technique resulted in what it did.

This study presents a human-in-the-loop approach to SA through visualization of sequences to enable a stakeholder to analyze the sequences and develop their ownunderstanding of the data. After a brief review of existing studies the study explores how SA can be beneficial to learning environments in general, especiallyin the context of online and hybrid learning,where it allows stakeholders to understand individual learning in detail.

Review of Existing Work

Problem-based learning (PBL) is “an instructional (and curricular) user-centric approach that empowers learners to conduct research, integrate theory and practice, and apply knowledge and skills to develop a viable solution to a defined problem” (Savery, 2006, p. 12).

In PBL, users assume primary responsibility for their own development while trainers provide facilitation; learning occurs in small groups in which collaboration is emphasized and encouraged (Barrows, 1996). This approach promotes user-centric learning and collaboration. Despite its effectiveness, complex user-centric learning environments such as PBL require appropriate scaffolds that support novices’ learning and problem- solving processes (Pellegrino, 2004; Simons & Klein, 2007).

PBL requires a detailed understanding of the learner’s problem-solving processes, often obtained through the granular analysis of their actions within a training environment (Shute et al., 2010; Min et al., 2016; Kinnebrew&Biswas, 2012; Baker et al., 2006).

By connecting learners’ observable actions to the concepts taught, it is possible to infer an individual’s mastery of skills and personalize the training process accordingly (Shute et al., 2010; Vahdat et al., 2015). This is the goal of many approaches to analyzing users’ observable actions, including goal recognition (Min et al., 2016; Ha et al., 2014) and Bayesian approaches (Shute et al., 2010; de Klerk et al., 2015). The drawback of these approaches, however, is that they are overly focused on the outcomes of a user’s problem solving, rather than the process.

SA (sequence analysis), specifically sequence clustering and comparison, is a promising avenue to the process-focused study of individual problem-solving, informing more effective personalized learning. The problem, however, is that without anunderstanding of the data, it can be difficult for a stakeholder to infer how an algorithm or statistical method is understanding thedata or why a statistical technique resulted in what it did. Sequence-based approaches have demonstrated value as an effective method for identifying individual differences within a group of learners and patterns across a community (Kinnebrew&Biswas, 2012; Kinnebrew et al., 2013).

Unlike goal recognition or Bayesian approaches, these approaches focus specifically on process and are better suited to capturing problem-solving strategies. The primary drawback of existing sequence-based approaches to studying problem solving, however, is the absence of a human-being in the loop. A human-in-the-loop approach to sequence analysis (SA) can produce more interpretable results by allowing stakeholders to understand and correct the model and its outputs. A human-in-the-loop method is defined as one where a human stakeholder, meaning a person who would use the approach in practice, such as an educator or analyst, can interpret the output of the model and then provide input to the model that impacts how it analyzes, compares, or clusters the sequences. This interaction loop is iterative, with a human-reading the output, providing input, reading the new output, and providing new input in a cyclical manner until they feel the output presents an accurate and insightful view of the data.

A great proportion of NLP tasks require logical reasoning. Prior work contextualizes the problem of logical reasoning by proposing reasoningdependent datasets and studies solving the tasks with neural models (Johnson et al., 2017; Sinha et al., 2019; Yu et al., 2020; Liu et al., 2020; Tian et al., 2021). However, most studies focus on solving a single task, and the datasets either are designed for a specific domain (Johnson et al., 2017; Sinha et al., 2019), or have confounding factors such as language variance (Yu et al., 2020); they can not be used to strictly or comprehensively study the logical reasoning abilities of models. Another line studies leveraging deep neural models to solve pure logical problems. Clark et al. (2020) conducts a similar study to show that models can be trained to reason over language, while Xu et al. (2019) studies how well neural models can generalize on different types of reasoning problems from a theoretical perspective/

Unlike dataset biases/artifacts identified in typical NLP datasets, which are often due to biases in the dataset collection/annotation process (Gururangan et al., 2018; Clark et al., 2019; He et al., 2019), statistical features inherently exist in reasoning problems and are not specific to certain data distributions.

Moreover, Large language models such as GPT-3 and LaMDA manage to maintain coherence over long stretches of text as they can remain consistent in lengthy conversations about different topics.

According to scientists at the University of California, Los Angeles, transformers, the deep learning architectures used in LLMsfind clever ways to learn statistical features that inherently exist in the reasoning problems rather than learning to emulate reasoning functions. These researchers tested a popular transformer architecture, on a confined problem space which could accurately respond to reasoning problems on in-distribution examples in the training space yet couldn't generalize to examples drawn from other distributions based on the same problem space.

There are several benchmarks that test AI systems against natural language processing and understanding problems and human performance on these benchmarks is closely tied to common sense and the capacity for logical reasoning. Yet, it is not clear whether large language models are improving because they have acquired logical reasoning capabilities or simply because they have been exposed to very large amounts of text.

According to the researchers,"the model attaining high accuracy **only** on in-distribution test examples has learned to use *statistical features* in logical reasoning problems to make predictions rather than to emulate the correct reasoning function."

These findings unveil the fundamental difference between "learning to reason" and "learning to solve a typical NLP task." For most NLP tasks, one of the major goal for a neural model is to learn statistical patterns: for example, in sentiment analysis (Maas et al., 2011), a model is expected to learn the strong correlation between the occurrence of the word "happy" and the positive sentiment. However, for logical reasoning, even though numerous statistical features inherently exist, models should not be utilizing them to make predictions.

Although neural networks are very good at finding and fitting statistical features such as in sentiment analysis (SA) as there is a strong correlation between certain words and classes of sentiments,the model should still try to find and learn the underlying reasoning functionwhen it comes to logical reasoning tasks.Therefore, in order to train neural models end-to-end to solve NLP tasks that involve both logical reasoning and prior knowledge a lot of the information that LLMs require may simply not be included in the data.

Logical reasoning is needed in a wide range of NLP tasks including natural language inference (NLI) (Williams et al., 2018; Bowman et al., 2015), question answering (QA) (Rajpurkar et al., 2016; Yang et al., 2018) and common-sense reasoning (Zellers et al., 2018; Talmor et al., 2019). The ability to draw conclusions based on given facts and rules, is essential to solving these tasks. Though NLP models, empowered by the Transformer neural architecture (Vaswani et al., 2017), can achieve high performance on task-specific datasets (Devlin et al., 2019), it is unclear whether they are "reasoning" over the input following the rules of logic.

As language models become larger this logical reasoning problem becomes hidden in their huge architecture. Adding layers, parameters, and attention heads to transformers might not help to bridge the reasoning gap.The rules of logic never rely on statistical patterns to conduct reasoning. Given the challenge of developing a logical reasoning dataset without any statistical features, learning to reason from data is difficult.

Existing literature has demonstrated that approaches where a human works iteratively with a system can result in more interpretable, actionable insights regarding subject behavior (Ahmad et al., 2019; Kleinman et al., 2020; Horn et al., 2016; Malmberg et al., 2017; Zhu& El-Nasr, 2021).

SA is one of the most effective ways to analyze learningbecause it offers a detailed and granular examination of user strategies based on the order in which they perform actions(Bakhshinategh et al., 2018; Papamitsiou& Economides, 2014; Romero et al ., 2009; Valls-Vargas et al ., 2015; Gasevic et al .,2017; Iske, 2008; Kinnebrew& Biswas, 2012; Kinnebrew et al., 2013; K¨ock&Paramythis, 2011). Sequence approaches also encompass hidden Markov models(HMMs), which identify meaningful interaction patterns and infer user problem-solving strategies or predict future actions(Jeong et al., 2008; Balakrishnan & Coetzee, 2013; Boumi& Vela, 2019; Geigle &Zhai, 2017; Doleck et al., 2015).

One of the emerging methods for increasing trust in AI-driven systems is to promote the use of methods that “enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners” (Gunning, 2017). Many of these AI-only initiatives failed due to the incorrect assumption that technology transfer moves along a single dimension from basic to applied research to deployment, where one could create a solution today and solve an unknown problem tomorrow.

One widely used AI technique for generating these kinds of models is *latent knowledge estimation* (Corbett & Anderson, 1994). The reason this is referred to as *latent* lies in the fact that knowledge cannot be directly observed. What can be observed is whether a user can apply a knowledge component in some context.

Another one is Bayesian knowledge tracing (BKT) which is the best-known technique for latent knowledge estimation (Corbett & Anderson, 1994). The technique uses four parameters to estimate whether a user can apply a knowledge component, including;

(a). probability that the user already masters a knowledge component,
(b). probability of learning a knowledge component after a learning opportunity,
(c). probability of correctly applying a knowledge component even when the user has not mastered it (guess), and
(d). probability of incorrectly applying a knowledge component although they know it (slip).

Existing research suggests that having AI systems explain their inner workings to their end users can help foster transparency, interpretability, and trust. A range of analyses provide conceptual frameworks for understanding the challenges of these AI models.

The architecture of deep learning models influences a model’s inductive biases and has been shown to have a crucial effect on both the training speed and generalization (dAscoli et al., 2019; Neyshabur, 2020). Searching for the best architecture for a given task is an active research area with diverse approaches. The idea of incrementally increasing the size of a model has been used in many settings such as boosting (Friedman, 2001), continual learning (Rusu et al., 2016), architecture search (Elsken et al., 2017; Cortes et al., 2017), optimization (Fukumizu& Amari, 2000; Caccia et al., 2022), and reinforcement learning (Berner et al., 2019).

In order to utilize the wealth of unlabeled video data available on the internet, a novel, yet simple, semi-supervised imitation learning method, Video PreTraining (VPT), has been introduced by researchers at Open AI (2022). VPT paves the path toward allowing agents to learn to act by watching the vast numbers of videos on the internet. It extends the paradigm of training large and general purpose behavioral priors from freely available internet-scale data to sequential decision domains.

Furthermore, in some models, multiple large pretrained models may be composed through language (via prompting) without requiring training, to perform new downstream multimodal tasks. This offers an alternative method for composing pretrained models that directly uses language as the intermediate representation by which the modules exchange information with each other.

For example, visual-language models (VLMs) are trained on Internet-scale image captions, but large language models (LMs) are further trained on Internet-scale text with no images (e.g., spreadsheets, SAT questions, code). As a result, these models store different forms of commonsense knowledge across different domains. These models are not only competitive with state-of-the-art zero-shot image captioning and video-to-text retrieval, but also enable new applications such as (i) answering free-form questions about egocentric video, (ii) engaging in multimodal assistive dialogue with people by interfacing with external APIs and databases (e.g., web search), and (iii) robot perception and planning.

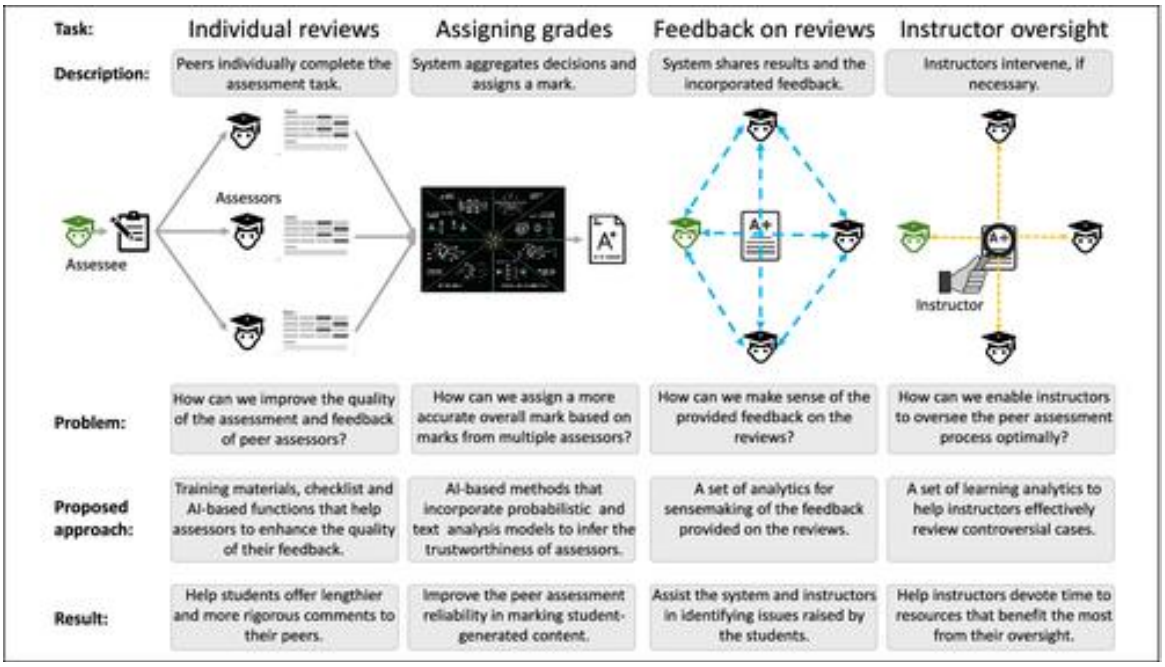
This works by formulating video understanding as reading comprehension, i.e., re-framing “video Q&A” as a “short story Q&A” problem, which differs from common paradigms for video understanding that may involve supervising video-text models on labeled datasets or adversarial training. To this end, first a set of "key moments" should be extracted throughout the video (e.g., via importance sampling, or video/audio search based on the input query, discussed in Appendix). Next, the key frames indexed by these moments should be captured so that they could be summarized into a language-based record of events. This is then passed as context to an LM to perform various reasoning tasks via text completion such as Q&A, for which LMs have demonstrated strong zero-shot performance.

From video search, to image captioning; from generating free-form answers to contextual reasoningquestions, to forecasting future activities – these models can provide meaningful results for complex tasksacross classically challenging computer vision domains, without any model finetuning.

Examples of guided multi-model exchanges for an egocentric perception system:parsing a natural language question into search entities (with LM) to be used to find the most relevant key moments in the video (with VLM); (ii, middle) describing each key frame by detecting places and objects (VLM), suggesting commonsense activities (LM), pruning the most likely activity (VLM), then generating a natural language summary (LM) of the interaction; (iii, right) concatenating key frame summaries into a language-based world-state history that an LM can use as context to answer the original question.

When it comes to utilizing these AI models, one of the critical areas supported by human-centered AI is the process of assessment design used to elicit evidence to support claims about learning. While automated question generation can be a powerful tool for making assessment design more feasible for educators, it is not without its limitations. Large-scale datasets are needed to train the models that generate the questions.

By thoughtfully defining parameters, identifying its impact on users and assessing current capabilities, AI tools fitting assessment needs can be chosen. Figure 1 provides a graphical summary of peer assessment processes along with problems and proposed approaches and results.



Task:	Individual reviews	Assigning grades	Feedback on reviews	Instructor oversight
Description:	Peers individually complete the assessment task.	System aggregates decisions and assigns a mark.	System shares results and the incorporated feedback.	Instructors intervene, if necessary.
Problem:	How can we improve the quality of the assessment and feedback of peer assessors?	How can we assign a more accurate overall mark based on marks from multiple assessors?	How can we make sense of the provided feedback on the reviews?	How can we enable instructors to oversee the peer assessment process optimally?
Proposed approach:	Training materials, checklist and AI-based functions that help assessors to enhance the quality of their feedback.	AI-based methods that incorporate probabilistic and text analysis models to infer the trustworthiness of assessors.	A set of analytics for sensemaking of the feedback provided on the reviews.	A set of learning analytics to help instructors effectively review controversial cases.
Result:	Help students offer lengthier and more rigorous comments to their peers.	Improve the peer assessment reliability in marking student-generated content.	Assist the system and instructors in identifying issues raised by the students.	Help instructors devote time to resources that benefit the most from their oversight.

Figure 1. A graphical summary of peer assessment processes

Peer assessment has been recognized as a sustainable and developmental assessment method. Peer assessment can formally be defined as "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (Topping, 2009, p. 20). A simple approach would be to use summary statistics such as mean or median. However, summary statistics suffer from the assumption that all users have a similar judgmental ability, which has proven incorrect (Abdi et al., 2021).

The data generated from users' engagement with the peer assessment process may be utilized by learning analytics tools and learning analytics dashboards (Matcha et al., 2019) to enable instructors to gain insights into users learning process.

The next sections explore the model development in more detail.

Model Development

The platform architecture consists of two main components: an open-source Python API for specifying the workflow of an experiment (the "API"), and a Platform-as-a-Service (PaaS), which is a running instance of back-end infrastructure coupled with computational resources and a large training dataset.

Controller Scripts: First, a user creates and submits a configuration file, either using an HTTP request or using the easy_submit() API function. This configuration file contains job metadata, including a pointer to an executable Docker image which encapsulates all code, software, and operating system dependencies for the users' experiment. The configuration file also points to a Python controller script that specifies the high-level experimental workflow, such as how model training and testing should occur and whether cross-validation or a holdout set should be used in a predictive modeling experiment. The use of controller scripts is a best practice for reproducible computational research [18], as it provides a single script to fully reproduce an experiment.

An additional advantage is that controller scripts are human-readable, providing a high-level overview of an experiment. One can use the controller script to manage low-level data platform tasks, including (i) data wrangling (retrieving and archiving necessary data at each step of the experiment); (ii) Docker image setup and execution; and (iii) parallelization.

The controller script provides sufficient information about how one can execute parallelization, which can lead to speedups of 1-2 orders of magnitude when CPUs are occupied with a separate task (e.g. training models on each of the different courses available).

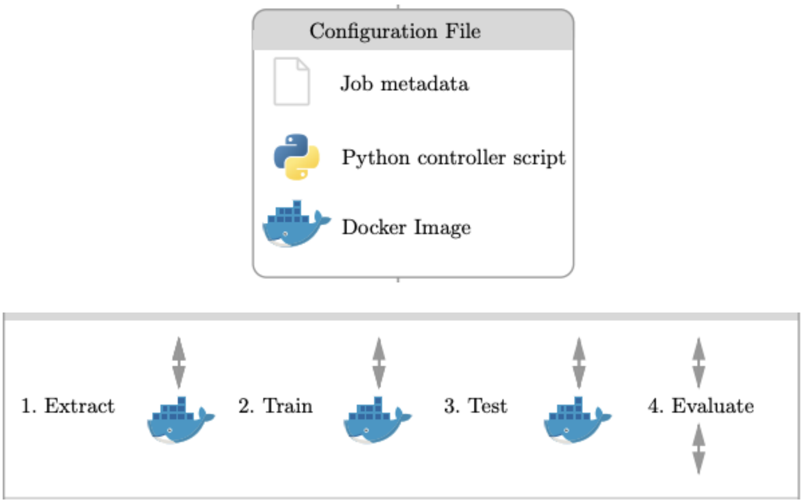


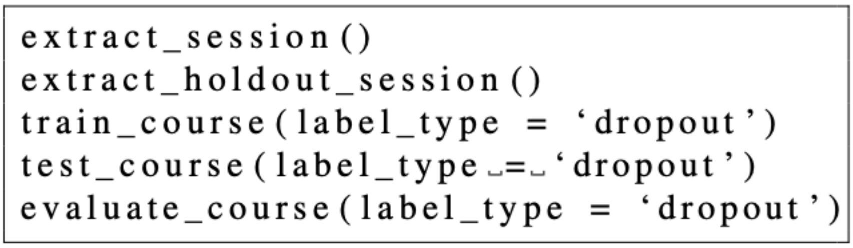
Fig 3. Overview of Docker container

Docker containers: These were originally developed to resolve the technical issues described above in software development contexts [29], and are frequently used in both industrial software applications as well as computational and computer systems research [20], [30], [31]. Their use in data science applications is increasing, but the execution, publication, and sharing of pre-built Docker images as part of a research workflow is rare. While some existing platforms utilize containerization, this functionality is hidden from the user which might limit users' ability to fully leverage containerization by building the complex, customized environments many machine learning experiments may require.

As seen in Figure 3., when submitting a job for execution to the platform, a user generates a Docker image containing the code, and operating system dependencies required to execute their experiment, and uploads the image to a public location (files located locally, HTTP, or in Amazon S3 are supported). The user provides the image's URL the configuration file submitted to the platform, and the image is fetched, checked, and executed according to the controller script. When an experiment completes error-free execution, the platform uploads the image to a public image repository on Docker Hub using a unique identifier. This makes implementations of every experiment immediately and publicly available for verification, extension, citation, or re-use.

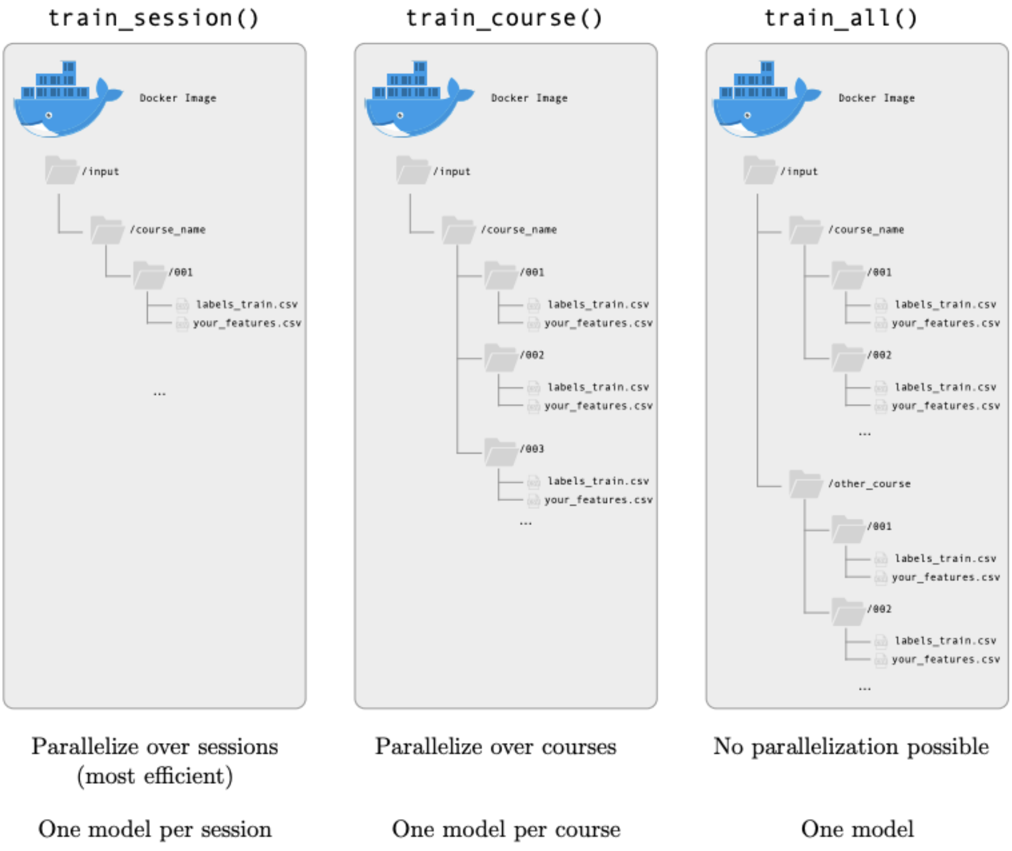
A major advantage of Docker over simple code-sharing is that Docker containers fully reproduce the entire execution environment of the experiment, including code, software dependencies, and operating system libraries, exactly as this environment is configured at the time of an experiment. These containers are much more lightweight than a full virtual machine, but achieve the same level of reproducibility [29], [31].

As seen in Figure 4, the Python API allows users to provide a simple execution "recipe" for the platform to execute their experiment specifying the complete end-to-end pipeline from raw data to model evaluation: extract features from raw data; train and test machine learning models (predictive modeling experiments only), and evaluate the experimental results. For example, after extracting the desired features, a predictive modeling experiment could train individual models for every session of a course by using train_session() in their controller script; train one model per course using the data from all sessions by using train_course(); or train a single monolithic model using all data from every session of every course by using train_all().



```
extract_session()
extract_holdout_session()
train_course(label_type = 'dropout')
test_course(label_type␣=␣'dropout')
evaluate_course(label_type = 'dropout')
```

Figure 4. Sample API script for Docker



Model Evaluation

A critical area of research in learning sciences to date has been the construction and analysis of predictive models of individual success [8]. Predictive modeling experiments follow a standard end-to-end supervised learning workflow:

- feature extraction from raw data;
- model training; model testing; and
- model evaluation (whereby performance is analyzed or, optionally, evaluated using statistical tests).

In addition to jointly addressing several challenges to reproducible and replication research within the field of learning sciences, the architecture, workflow, and initial research results have implications for the broader big data community. This includes;

- experimental reproducibility as big data research in many fields uses increasingly complex computational models;
- methodological and inferential reproducibility as big data research enables problematic statistical practices such as massively multiple testing via testing thousands or millions of hypotheses in a single experiment; and
- data reproducibility as available data become massively multimodal (many different formats), measure increasingly private or restricted aspects of users' behavior and identity, and cannot be easily anonymized.

Such a framework would be domain-agnostic, and can support generic workflows for supervised learning and production rule analyses in any domain which works with complex, multiformat data which cannot be easily anonymized (e.g. sensitive medical data, copyrighted media, computational nuclear physics).

Conclusion

This paper brings both the issues with the standard assessment paradigm and the challenges associated with AI and assessment into a deeper conversation that will ultimately improve assessment practices more generally.

The paper has two important implications for learning analytics and AI in education**:**

- First, this paper gives researchers and practitioners a novel systematic approach to incorporating advances in AI-driven assessment by having a strong grounding in a theoretical model of relevant education or learning processes. Specifically, the paperdemonstrated how a theoretical model can be used to structure the program of research, development, deployment, and evaluation by addressing a problem (ie, trust in a peer-review system) that may emerge in practice.
- Second, the studies reported in the paper provide fresh empirical insights that can inform the development of future AI-driven assessment that seek to enhance trustworthiness of peer-review.

The comprehensive discussion of leaner-sourced adaptive systems, open-ended learning environments, writing analytics tools, team-based learning to support knowledge transfer allows for a detailed understanding of current state-of-art and open challenges. By doing so, this study will hopefully help to synthesize an agenda for future research for AI-driven assessment techniques.

References

[1]. Abdi, S., Khosravi, H., & Sadiq, S. (2020). Modelling learners in crowdsourcing educational systems. In International Conference on Artificial Intelligence in Education (pp. 3–9). Springer.

[2]. Abdi, S., Khosravi, H., Sadiq, S., & Darvishi, A. (2021). Open learner models for multi-activity educational systems. Artificial Intelligence in Education, 11–17. https://doi.org/10.1007/978-3-030-78270-2_2

[3]. Ahmad, N., & Bull, S. (2008). Do users trust their open learner models? In International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (pp. 255–258). Springer.

[4]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

[5]. Ashenafi, M. M. (2017). Peer-assessment in higher education-twenty-first century practices, challenges and the way forward. Assessment & Evaluation in Higher Education, 42, 226–251.

[6]. Carless, D., & Boud, D. (2018). The development of user feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43, 1315–1325.

[7]. Cho, K., & MacArthur, C. (2011). Learning by reviewing. Journal of Educational Psychology, 103, 73–84.

[8]. Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learner sourcing to inform design loop adaptivity. In European Conference on Technology Enhanced Learning (pp. 332–346). Springer.

[9]. Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of user generated content at scale: A trust propagation approach. In Proceedings of the Eighth ACM Conference on Learning@ Scale (pp. 139–150). Association for Computing Machinery

[10]. Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive user models through slicing analysis. In Proceedings of the 9th international conference on learning analytics & knowledge (pp. 225–234). Association for Computing Machinery.

[11]. Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. Learning: Research and Practice, 3, 63–78.

[12]. Gyamfi, G., Hanna, B. E., & Khosravi, H. (2021). The effects of rubrics on evaluative judgement: A randomised controlled experiment. Assessment & Evaluation in Higher Education, 47(1), 126–143. https://doi. org/10.1080/02602938.2021.1887081

[13]. Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In Handbook of self-regulation of learning and performance (pp. 83–106). Routledge.

[14]. Han, Y., Wu, W., Yan, Y., & Zhang, L. (2020). Human-machine hybrid peer grading in SPOCs. IEEE Access, 8, 220922–220934.

[15]. Hassan, T. (2019). Trust and trustworthiness in social recommender systems. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 529–532). Association for Computing Machinery.

[16]. Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. Higher Education Research & Development, 38, 1401–1416.

[17]. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. Frontiers in Psychology, 4, 863.

[18]. Lee, W., Huang, C. H., Chang, C. W., Wu, M. K. D., Chuang, K. T., Yang, P. A. and Hsieh, C. C. (2018) Effective quality assurance for data labels through crowdsourcing and domain expert collaboration. In 21st International Conference on Extending Database Technology, EDBT 2018 (pp. 646–649). OpenProceedings.org.

[19]. Levy, H., & Robinson, M. (2006). Stochastic dominance: Investment decision making under uncertainty (Vol. 34). Springer

[20]. Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. IEEE Transactions on Learning Technologies, 13(2), 226–245.

[21]. Moon, T. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13, 47–60.

[22]. Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 588–593). Association for Computational Linguistics (ACL).

[23]. Negi, S., Asooja, K., Mehrotra, S., & Buitelaar, P. (2016). A study of suggestions in opinionated texts and their automatic detection. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (pp. 170–178). Association for Computational Linguistics

[24]. Purchase, H., & Hamer, J. (2018). Peer-review in practice: Eight years of Aropä. Assessment & Evaluation in Higher Education, 43, 1146–1165.

[25]. Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. International Journal of Artificial Intelligence in Education, 27, 534–581.

[26]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Association for Computational Linguistics

[27]. Topping, K. J. (2010). Peers as a source of formative assessment. In Handbook of formative assessment (pp. 73–86). Routledge.

[28]. Urena, R., Kou, G., Dong, Y., Chiclana, F., & Herrera-Viedma, E. (2019). A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. Information Sciences, 478, 461–475.

[29]. Wang, W., An, B. and Jiang, Y. (2018) Optimal spot-checking for improving evaluation accuracy of peer grading systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1). AAAI Press.

[30]. Wang, W., An, B., & Jiang, Y. (2020). Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems. IEEE Transactions on Computational Social Systems, 7, 940–955.

[31]. Wind, D. K., Jørgensen, R. M., & Hansen, S. L. (2018). Peer feedback with peer grade. In ICEL 2018 13th International Conference on e-Learning (p. 184). Academic Conferences and Publishing Limited.

[32]. Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: Partially automated high-stakes peer grading. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (pp. 96–101). Association for Computing Machinery.

[33]. Xiong, W., &Litman, D. (2011). Automatically predicting peer-review helpfulness. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 502–507). Association for Computational Linguistics.

[34]. Yang, M., Tai, M., & Lim, C. P. (2016). The role of e-portfolios in supporting productive learning. British Journal of Educational Technology, 47, 1276–1286.

[35]. Yang, T.-Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active learning for useraffect detection. In Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019 (pp. 208–217). Université du Québec; Polytechnique Montréal.

[36]. Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. Journal of Experimental Psychology: General, 143, 804–824.

[37]. Yu, F.-Y., & Wu, C.-P. (2011). Different identity revelation modes in an online peer-assessment learning environment: Effects on perceptions toward assessors, classroom climate and learning activities. Computers & Education, 57, 2167–2177.

[38]. Zheng, L., & Huang, R. (2016). The effects of sentiments and co-regulation on group performance in computer supported collaborative learning. The Internet and Higher Education, 28, 59–67.

[39]. Zhu, Q., & Carless, D. (2018). Dialogue within peer feedback processes: Clarification and negotiation of meaning. Higher Education Research & Development, 37, 883–897.

[40]. Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). Developing self-regulated learners: Beyond achievement to self-efficacy. American Psychological Association.

[41]. Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in users' task performance? Computers in Human Behavior, 124, 106924.