Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Focal: A Proposed Method of Leveraging LLMs for Automating Assessments

Peter Meyers[a], Annette Han[a], Razik Grewal[a], Mitali Potnis[a] & John Stamper[a]
[a]*Carnegie Mellon University, United States*
*jstamper@cs.cmu.edu

Abstract: In response to the growing need for frequent, high-quality assessments in the expanding field of online learning and the significant time burden their manual creation places on educators, this study proposes Focal, an end-to-end assessment generation pipeline. Focal employs large language models, notably Text-to-Text Transfer Transformers, fine-trained on diverse learning materials, to generate and evaluate pedagogically sound questions and their corresponding answers. The Focal pipeline is designed to integrate with Learning Management Systems, providing educators an automated means of creating assessments that align with their curriculum. This not only eases the task of creating and evaluating assessments but also frees educators to focus on other crucial responsibilities. The system is domain agnostic and its efficacy is continually improved by training and evaluating it using data from multiple subject areas. By automating the traditionally labor-intensive process of assessment production, Focal aims to increase efficiency in online education and enhance the learning experience for students.

Keywords: Question Generation, Assessment Generation, Large Language Models

1. Introduction

Frequent assessments bolster long-term retention and conceptual understanding, acting as “memory modifiers” (Bjork, 1975) aiding learners to apply information in novel situations (Storm et al., 2010, Kornell et al., 2009). This process enhances recall, cultivates adaptive expertise (Baroody, 2013, Chi et al., 1982), and promotes cognitive flexibility in procedural domains like mathematics (Hiebert and Lefevre, 1986, Salomon and Perkins, 1989).

Given assessments’ critical role, educators invest considerable effort in crafting questions reinforcing course concepts (Zhao et al., 2005). However, curating a substantial question bank that caters to students of varied skill levels poses a considerable challenge—it is a time-consuming and labor-intensive process (Nguyen et al., 2022). Moreover, providing quality feedback, crucial to augment students’ learning experiences, further compounds this burden on educators.

Addressing these challenges necessitates a novel approach, leading us to propose Focal—an innovative, end-to-end assessment generation pipeline. Designed to automate the process of creating educational assessments, Focal is built upon the principles of the Knowledge-Learning-Instruction (KLI) framework (Koedinger et al., 2012). This framework guides a systematic approach to educational practice, prompting detailed analyses of the knowledge students acquire in courses. Consequently, Focal auto-generates questions and answers based on provided curricula across various subject domains, thus embodying instructional principles with a high potential for generality.

The core of Focal lies in its machine learning pipeline that ingests texts from learning materials, generating pedagogically sound and logically coherent questions and answers that align with the curriculum. It further evaluates their quality to ensure the assessments are academically rigorous. Leveraging large language models (LLMs) and data from multiple subjects, Focal continually enhances its domain-agnostic capabilities, which helps maintain the relevance and quality of the assessments it generates.

Focal's integration into Learning Management Systems (LMS) not only alleviates the burden on educators but also democratizes access to high-quality assessments. These assessments, drawn from expert-curated curricula, are accessible to all learners, regardless of their socio-economic background. By effectively addressing scalability and quality issues inherent in assessments, Focal positions itself as a promising solution in the dynamic landscape of online education, propelling it towards new horizons of inclusivity and efficacy.

2. Related Work

The proliferation of LLMs significantly shapes the landscape of Natural Language Processing (NLP) and Question Generation (QG). These transformer-based encoder-decoder models have demonstrated efficiency in general QG (Xue et al., 2020, Yu et al., 2021) and in educational settings as well (Grover et al., 2017) significantly propelled by neural transformer-based methods and the BERT model (Vaswani et al., 2021).

In addressing the challenges initially faced by BERT in token generation (Lopez et al., 2020) advanced the robustness of QG by fine-tuning a pre-trained language model. Efforts have also been made to refine the relevance of the generated questions by incorporating common sense and domain knowledge into the QG process (Jia et al., 2021, Wang et al., 2020) and developing an attention-based sequence-to-sequence model that integrates target answer information (Liu, 2020).

Significant progress has been made towards automating the generation of educational questions, with models such as GPT-2 successfully generating mathematical word problems of varying difficulty (Cheng et al., 2021) and others producing questions reflective of realistic scenarios (Liu et al., 2022). Recently, researchers employed GPT-3 to create EduQuiz, an end-to-end educational quiz generator (Dijkstra et al., 2022). Despite certain limitations, including language specificity, domain specificity, and higher costs for fine-tuned models, EduQuiz generated reasonably high-quality quizzes. High-quality distractor generation, however, remains a challenge. The authors envisage potential for improvement and suggest integrating human input in future iterations to enhance the quality of generated quizzes.

QG's evaluation remains a formidable challenge, necessitating a blend of automated assessments using machine learning models and human evaluations (Kurdi et al., 2020). Automated assessment strategies often rely on metrics such as BLEU and ROGUE (Novikova et al., 2017). However, these approaches face scrutiny due to concerns surrounding their interpretability and weak correlation with human evaluations (Van Der Lee et al., 2019). In contrast, human evaluators examine factors like grammatical correctness, fluency of language, relevance to the topic, and the naturalness of the language employed in the questions (Amidei et al., 2018, Chen et al., 2018, Ruseti et al., 2018)

An additional measure of evaluation involves ensuring that the generated questions correspond with "ground truth" data or expert-crafted reference questions (Sai et al., 2022). This aspect becomes even more crucial in the realm of educational QG, where the assessments are expected to be a mirror reflection of the learning material's target skills. The creators of MOOCCubeX introduced an automated solution to address this concern, pioneering the path for adaptive learning research and concept-centric data organization (Yu et al., 2021).

Against this backdrop, the Focal pipeline emerged as a testament to the success of LLMs, further propelled by findings that key concept extraction could boost their usefulness (Yu et al., 2021, Stamper et al., 2023). Recent work (Nguyen et al., 2022) has established the potential of LLMs in generating questions that are not only coherent but also pedagogically beneficial. Our work builds upon this foundation, aspiring to optimize the Focal pipeline end-to-end, expand Focal's domain-agnostic capabilities, and refine the grading metrics for questions.

3. **Methodology**

3.1 *Data*

In our initial Focal pipeline testing, we used text data from a graduate-level data science course - Foundations of Computational Data Science (FCDS) and an undergraduate-level chemistry course, existing in XML format. The course content has hierarchical levels: Units, Modules, and Topics. We prepended these hierarchy titles to the course text based on the proven value of such an approach in QG (Nguyen et al., 2022). This is done for each hierarchy level, as illustrated in 1. For instance, the unit title gets concatenated with the text content before being fed into the QG model.

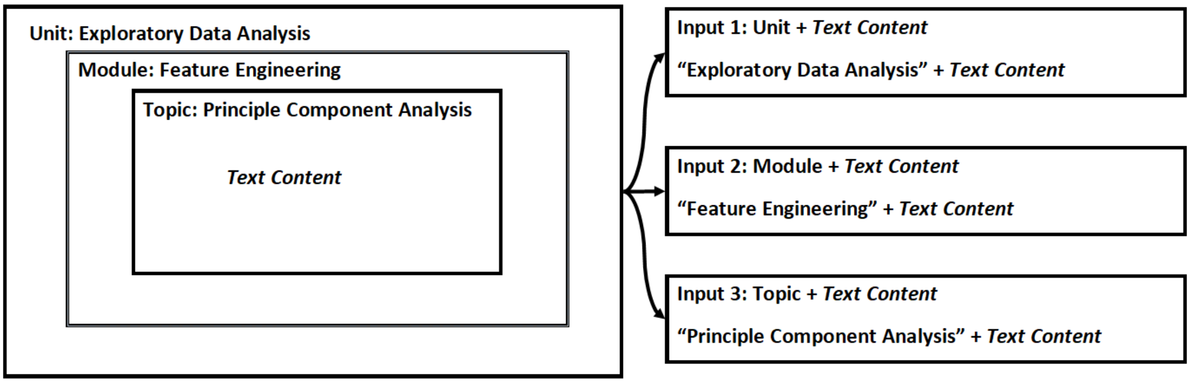
Unit: Exploratory Data Analysis
Module: Feature Engineering
Topic: Principle Component Analysis
Text Content
Input 1: Unit + Text Content
"Exploratory Data Analysis" + Text Content
Input 2: Module + Text Content
"Feature Engineering" + Text Content
Input 3: Topic + Text Content
"Principle Component Analysis" + Text Content

Figure 1. Graphical depiction of data pre-pending process demonstrated to be valuable in generating high quality questions.

Aside from course content, Focal employs the SQuAD 1.1 dataset for pre-training the QG model; this dataset comprises over 100,000 comprehension questions based on Wikipedia articles (Rajpurkar et al., 2016). This method improves the logic of generated questions (Nguyen et al., 2022). Further, the LearningQ dataset, with 230,000 document-question pairs and 7000 educator-crafted queries on assorted educational subjects (Chen et al., 2021), refines our evaluator model, enhancing the pipeline's accuracy in judging the logic of auto-generated questions.

3.2 *Model Design*

Figure 2 delineates Focal's QG Workflow. Initiated with Data Extraction and Pre-processing, the pipeline processes XML input data using BeautifulSoup to scrape and cleanse, including punctuation and stop word removal. It prepares the data for subsequent stages, including a crucial pre-pending phase that extracts headings, keywords, and main text from each section of course material. Post-preprocessing, the data splits, directing to the Concept Hierarchy Extraction and Question and Answer Generation stages.

The subsequent Concept Hierarchy Extraction, a vital part of the Focal pipeline, encompasses key concept creation and extraction. This stage relies on prior research showing LLMs' improved question generation ability with these key concepts (Stamper et al., 2023). We leverage the MOOCCubeX pipeline for this extraction, a platform rich in educational content and associated concept maps (Yu et al., 2021). It also filters invalid concepts like prepositions, indexing numbers, and generic verbs, currently needing a domain expert. As part of our long-term goals, we aim to develop a self-sufficient system capable of conducting this filtering process independently.

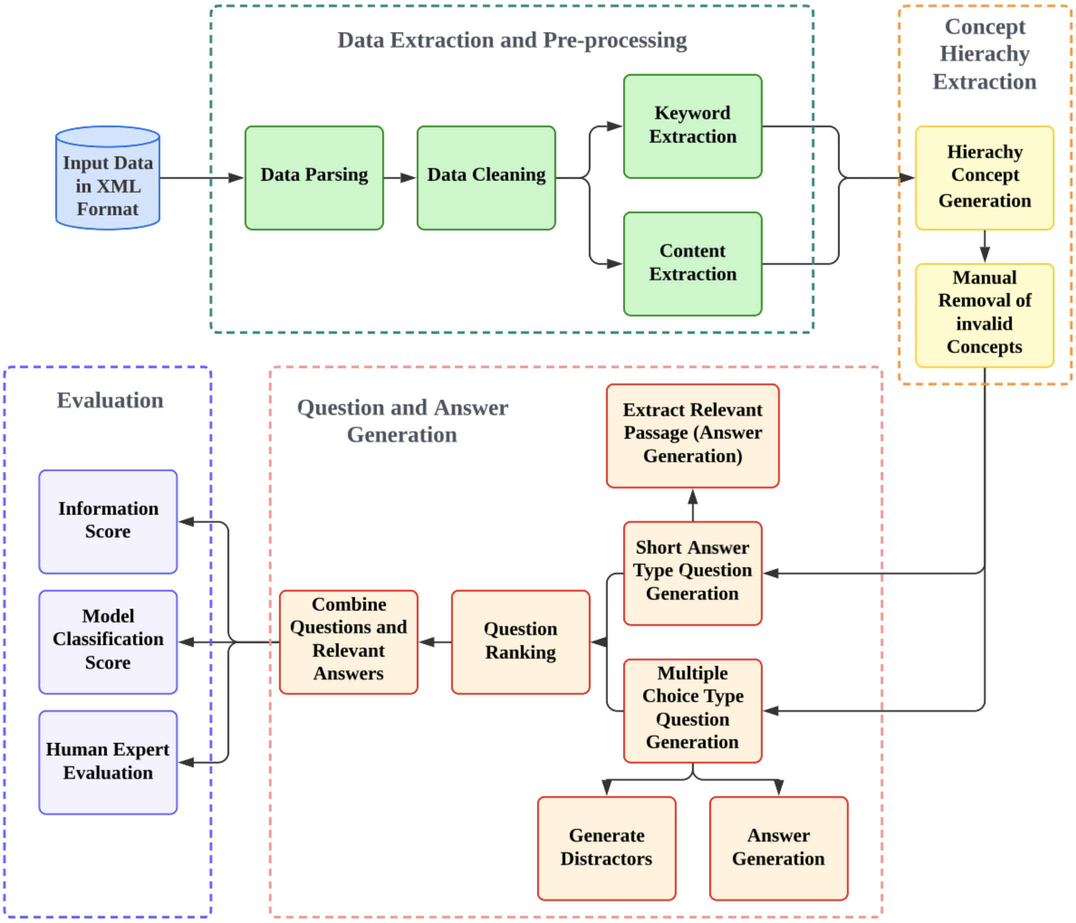
Data Extraction and Pre-processing
Input Data in XML Format
Data Parsing
Data Cleaning
Keyword Extraction
Content Extraction
Concept Hierachy Extraction
Hierachy Concept Generation
Manual Removal of invalid Concepts
Evaluation
Information Score
Model Classification Score
Human Expert Evaluation
Question and Answer Generation
Extract Relevant Passage (Answer Generation)
Short Answer Type Question Generation
Multiple Choice Type Question Generation
Question Ranking
Combine Questions and Relevant Answers
Generate Distractors
Answer Generation

Figure 2. Complete Focal QG workflow diagram.

After Concept Hierarchy Extraction, the pipeline enters the Question and Answer Generation phase, creating multiple-choice and short-answer questions. This phase employs a T5 transformer-based encoder-decoder model, fine-tuned based on prepending headings to the text (Raffel et al., 2020). The T5 model's current fine-tuning relies on the SQuAD 1.1 dataset (Rajpurkar et al., 2016), with potential for future adjustment. Crucially, it generates distractors for multiple-choice questions and extracts the correct answers via a dependency parse tree in a rule-based approach.

The final step before involving students is the Evaluation of Generated Questions and Answers. Here, three primary methods are used to evaluate the generated questions in terms of logical and pedagogical soundness: information score, GPT-3 model classification, and human expert evaluation. The information score is a custom metric designed to evaluate each question's relevance within the context of the identified key concepts from the Concept Hierarchy Extraction phase. By analyzing the overlap between tokens in a question and the extracted key concept tokens from the course text, the information score provides a robust assessment of question quality, while normalizing for question length ensures fairness across all questions.

$$IS = \frac{1}{|T(q)|} \sum_{t \in T(q)} 1(t \in C). \tag{1}$$

To assess question soundness, we utilize an information score and a GPT-3 model fine-tuned on the LearningQ dataset (Chen et al., 2018). Domain experts also evaluate question soundness, but with the information score's refinement, their involvement might become unnecessary.

Once questions are generated and evaluated, the Focal pipeline integrates student responses. Student answers are assessed for correctness using cosine similarity or exact matching, depending on the question type. For short-answer questions, a vector is generated from the input text using bag-of-words or TF-IDF methods, and its cosine similarity to the model-generated answer vector is measured using Equation 2. In this equation, A denotes the original answer, S represents the student's response, and D is the dot product of the two.

$$CosineSimilarity(A, S) = \frac{D}{|A| * |S|} \tag{2}$$

For numerical, Yes/No, or multiple-choice type questions, we employ an exact matching approach between the student's response and the model-generated answer to determine the correctness of the student's answer as illustrated in figure 3.

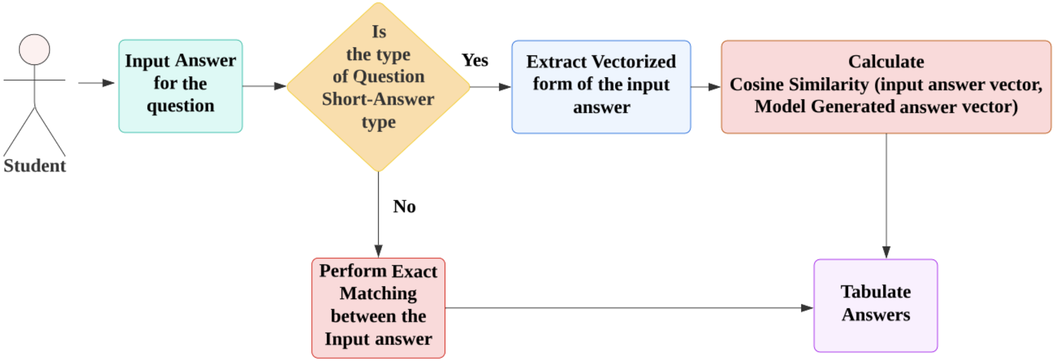
Student
Input Answer for the question
Is the type of Question Short-Answer type
Yes
No
Extract Vectorized form of the input answer
Calculate Cosine Similarity (input answer vector, Model Generated answer vector)
Perform Exact Matching between the Input answer
Tabulate Answers

Figure 3. User Answer Evaluation Workflow.

4. Initial Results

4.1 *Question Generation*

In order to evaluate the quality of several QG models, we generated questions for both the data science course and chemistry course using four different LLMs: bert2BERT, GPT-2, BART, and T5 (Chen et al., 2021, Radford et al., 2018, Xue et al., 2020). Table 1 provides examples of questions generated by Focal; additional examples can be found on Github[1].

Table 1. *Examples of questions generated for a paragraph of text content.*

Unit	Module	Generated Question
Analytic Algorithms and Model Building	Data Science Patterns	What is the difference between an observed value and the fitted value given by a model?
Data Gathering and Wrangling	Data Wrangling Pipeline	What does omission involve excluding the missing values from the dataset?
Exploratory Data Analysis	Performing Exploratory Data Analysis	What measures do you use to describe variability?

After generating questions using each of these models, we performed Model Classification, Information Score, and Perplexity Score to rate the quality of each generated question.

4.2 *Why use Perplexity?*

[1] https://github.com/annettehan/focal

Perplexity is a metric commonly used in NLP to evaluate the quality of language models. It measures how well a language model predicts a sequence of words or tokens. In simpler terms, perplexity quantifies how surprised a language model is when it encounters a new sequence of words. Perplexity is calculated based on the probability distribution assigned by the language model to a given sequence of words. The lower the perplexity score, the better the language model's ability to predict the next word in a sequence.

$$P(W) = b^{-1/N \sum_{i=1}^{N} log\, P(x_i \mid w_1, ..., w_{i-1})} \tag{3}$$

Perplexity is particularly useful in assessing the quality of generated questions because it captures how well a language model understands the context and generates coherent and meaningful questions. A language model with a low perplexity score is more likely to generate questions that align with the desired context and exhibit grammatical correctness and relevance. However, it's important to note that perplexity alone may not capture all aspects of question quality, such as the relevance or informativeness of the generated questions. Therefore, it is often used in conjunction with other evaluation metrics, such as Model Classification and Information Score, to provide a comprehensive assessment of the generated questions' quality.

4.3 *Model Classification*

For the Model Classification, we employed the GPT-3 classifier fine-tuned on the learningQ dataset to label each question as either sound or not sound (Radford et al, 2019, Chen et al., 2018). Here by soundness, we mean whether the questions are rationally valid, contextually relevant, and are effective such that they can be used for assessing the knowledge in regards with the topic. Following the GPT-3 classification, we randomly sampled 100 questions from each model type for both courses and manually annotated them as either sound or not sound. Figures 5 through 8 show the percentage of the 100 randomly sampled questions that each method of evaluation found to be sound for each model.

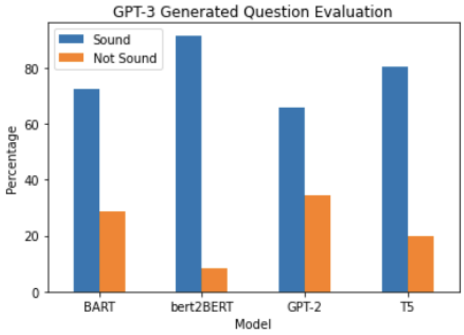

Figure 5: GPT-3 - Data Science course

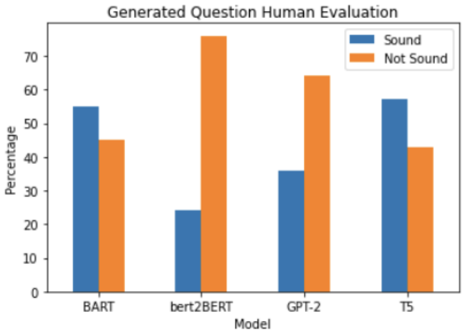

Figure 6: Human evaluator - Data Science course

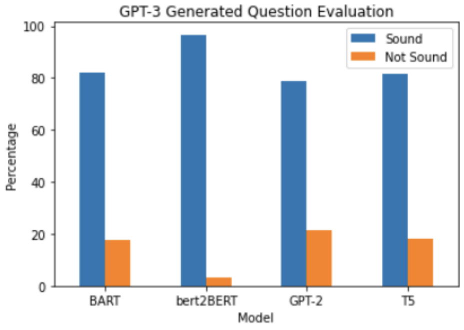

Figure 7: GPT-3 - Chemistry course

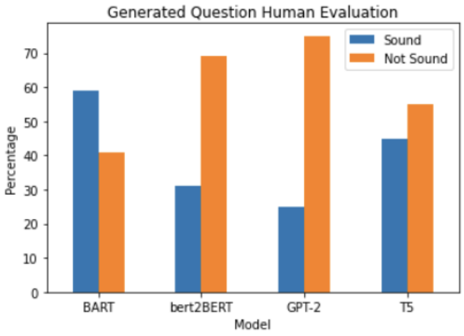

Figure 8: Human evaluator - Chemistry course

Initial analysis reveals our fine-tuned GPT-3 classifier often overestimates soundness compared to human assessment, with bert2BERT producing notably more unsound questions. T5 and BART fared better in the subset evaluated. To understand this divergence, we constructed confusion matrices for each model and course.

Table 2 reveals that human evaluators are likely to agree with our GPT-3 classifier when it rates a questions as unsound, suggesting its potential as a tool for eliminating low-quality questions rather than selecting high-quality ones.

Table 2. *Confusion Matrix for generated questions (400 random questions for each course).*

	Data Science Course		**Chemistry Course**	
	Expert: Sound	**Expert: Not Sound**	**Expert: Sound**	**Expert: Not Sound**
GPT3: Sound	161	148	158	183
GPT3: Not Sound	11	80	2	57

Upon analysis of the Figure 9 plots, we first see that on average, the perplexity score points to the correct direction of ground truth as the average perplexity of sound questions (149.4) is less than the average perplexity of not sound questions (155.6). Additionally, it seems evident that GPT should outperform T5, as a lower perplexity score signifies that it possesses better prediction power, but when we empirically look at the quality of questions output by T5, they are much more coherent and sound as compared to those output by GPT-2 model, which signifies that perplexity score evaluation has some room for improvement, so it more closely aligns human labeling, which is considered the gold standard in question evaluation.

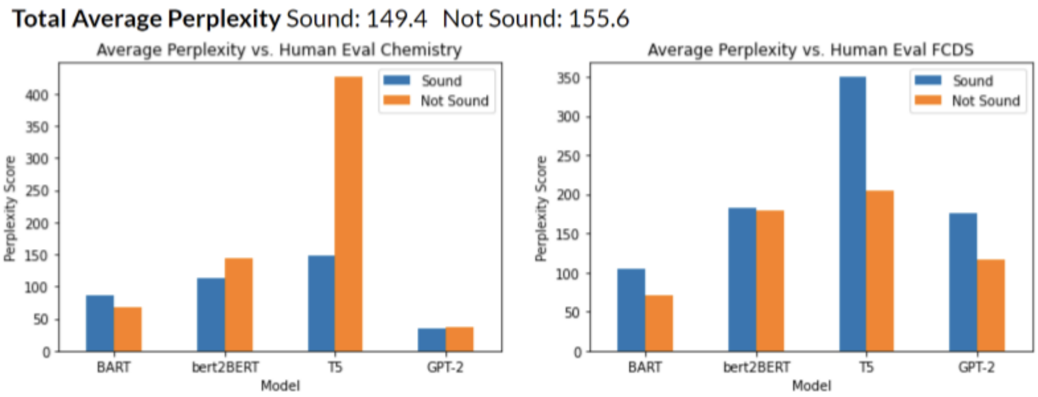

Figure 9. Average Perplexity vs Human evaluation for Data Science and Chemistry courses.

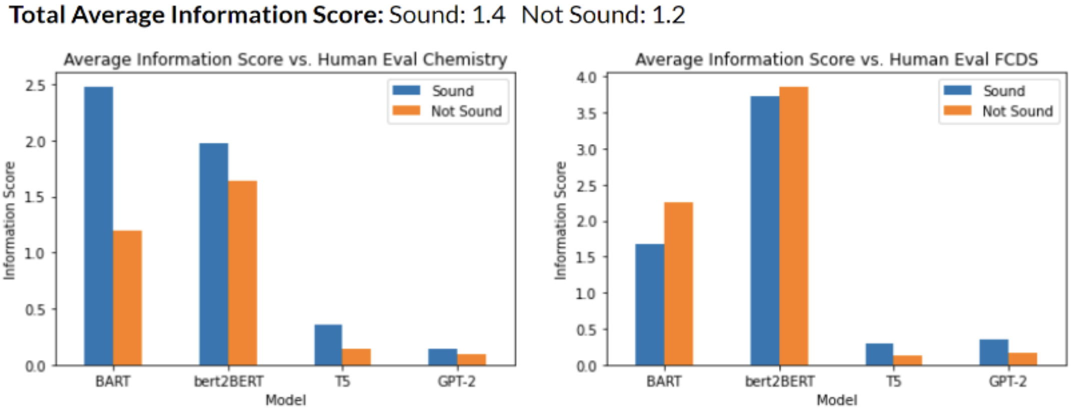

Figure 10. Average Information Score vs Human evaluation for Data Science and Chemistry courses.

The information score is meant to indicate that a question overlaps with the key concepts of a given passage. As shown by the study of the Figure 10 plots, where the average information score of sound questions is 1.4 and the average information score of unsound questions is 1.2. Furthermore, it seems obvious that GPT should outperform T5, as a higher information score denotes that it has better question quality. However, when we empirically examine the quality of questions produced by T5, they are significantly more coherent and sound than those produced by GPT-2 model, indicating that information score evaluation has some room for improvement so that it more closely aligns with human evaluation results.

5. Error Analysis

Throughout the development and testing of the Focal system, error analysis has been performed to identify issues and improve the quality of the generated questions and their evaluations. Some common error patterns observed in the experimental results are:

1. Irrelevant questions: Sometimes, questions generated fail to reflect key concepts, possibly due to misalignment in the Concept Hierarchy Generation and Information Score. To counteract this, Information Score measures question relevance, guiding improvements in Concept Hierarchy Generation and keyword extraction. For instance, it could address instances of vague queries such as "What type of procedure does the procedure follow?" (Table 3).
2. Question variability: The Focal system primarily generates "what" type questions and struggles with creating "why" or "how" questions. This might arise from the training data's lack of diversity. To enhance variability, the model can be fine-tuned using additional diverse datasets outside of SQuAD 1.1 or additional data with various question types. This would balance overproduction of "why" questions (Table 3).
3. Incorrect evaluation of student answers: The system may incorrectly assess student answers due to limitations in the cosine similarity and exact matching methods. Testing multiple evaluation models and selecting the most accurate would reduce such errors.

Table 3. Examples of question invariability.

Unit	Module	Generated Question
Exploratory Data Analysis	Feature Engineering	What kind of questions do you have?
Analytic Algorithms and Model Building	Model Selection	What is the model Mj with the best performance on the test set?
		What type of procedure does the procedure follow?
		What is a model model for Mi?
		What subset is split into train subset, validation subset and test set?
		What is the name of the best hyperparameters?

In each iteration of the Focal system, error analysis is conducted to identify these issues and guide the improvement of the system. By examining experimental results at various stages and observing patterns of errors, the system can be refined to generate more pedagogically sound questions and provide more accurate evaluations of student answers. This iterative process ensures continuous improvement of the Focal system, enhancing its educational value and potential for adoption in the classroom.

6. Conclusions and Future Work

In this study, we have outlined both the current progression and future vision for the Focal assessment pipeline. Our initial evaluations have shown promising potential in its role as an assessment tool, though certain areas for improvement have been identified. For instance, we found that the Focal QG model occasionally generates questions that may seem logical but are not pedagogically sound or useful for assessments. Additionally, the current information score metric, which grades question quality, needs to be more nuanced to better encapsulate key concepts from course texts.

In our early testing, we noticed that Focal often generated questions that lacked complexity and only required students to recall specific facts from the material. Our future research will explore new QG models to improve this aspect, emphasizing questions that promote a deeper understanding of the course instead of mere recollection of details.

Moving forward, we intend to make the Focal pipeline more valuable by streamlining the assessment process and reducing the burden on educators. This includes automating the end to-end assessment process, from generating and evaluating questions to generating answers and evaluating student responses. One potential challenge is the generation of a wide array of question types. For example, LLMs are adept at creating trivia-style questions but struggle with generating more analytical “how” or “why” questions, which are better for gauging a student’s comprehensive understanding. We hypothesize that by fine-tuning these models on more diverse datasets with additional analytical questions, the models will produce more varied question types. This will be a key area of focus in our research.

Furthermore, refining the information score metric to more precisely predict the pedagogical appropriateness of questions is a crucial part of future work. This would allow the pipeline to be less dependent on domain experts and ensure that the generated assessments maintain a high level of quality. It’s essential that Focal serves as a practical tool for educators and students, delivering timely and relevant feedback, and contributing to a more efficient and effective learning environment.

References

Amidei, J., Piwek, P., & Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018.

Baroody, A. J. (2013). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In *The development of arithmetic concepts and skills* (pp. 1-33). Routledge.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. Information Processing and Cognition: The Loyola Symposium.

Chen, C., Yin, Y., Shang, L., Jiang, X., Qin, Y., Wang, F., ... & Liu, Q. (2021). bert2bert: Towards reusable pretrained language models. *arXiv preprint arXiv:2110.07143*.

Chen, G., Yang, J., Hauff, C., & Houben, G. J. (2018). LearningQ: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

Cheng, Y., Li, S., Liu, B., Zhao, R., Li, S., Lin, C., & Zheng, Y. (2021). Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. *arXiv preprint arXiv:2105.11698*.

Chi, M., Glaser, R., Rees, E., & Steinberg, R. J. (1982). , Expertise in problem solving . Advances in the psychology of human intelligence.

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading Comprehension Quiz Generation using Generative Pre-trained Transformers.

Grover, K., Kaur, K., Tiwari, K., Rupali, & Kumar, P. (2021). Deep learning based question generation using t5 transformer. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10* (pp. 243-255). Singapore.

Hiebert, J., & Lefevre, P. (1986). Conceptual knowledge in mathematics—An introductory analysis. *Conceptual and procedural knowledge–The case of mathematics*, 1-27.

Jia, X., Wang, H., Yin, D., & Wu, Y. (2021). Enhancing question generation with commonsense knowledge. In *China National Conference on Chinese Computational Linguistics* (pp. 145-160). Cham: Springer International Publishing.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757-798.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*, 121-204.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, B. (2020). Neural question generation based on Seq2Seq. In *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence* (pp. 119-123).

Liu, T., Fang, Q., Ding, W., Li, H., Wu, Z., & Liu, Z. (2020). Mathematical word problem generation from commonsense knowledge graph and equations. *arXiv preprint arXiv:2010.06196*.

Lopez, L. E., Cruz, D. K., Cruz, J. C. B., & Cheng, C. (2020). Transformer-based end-to-end question generation. *arXiv preprint arXiv:2005.01107*, *4*.

Nguyen, H. A., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. In *European conference on technology enhanced learning* (pp. 272-284). Cham: Springer International Publishing.

Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017). Why we need new evaluation metrics for NLG. *arXiv preprint arXiv:1707.06875*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683*

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ruseti, S., Dascalu, M., Johnson, A. M., Balyan, R., Kopp, K. J., McNamara, D. S., ... & Trausan-Matu, S. (2018). Predicting question quality using recurrent neural networks. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19* (pp. 491-502). Springer International Publishing.

Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, *55*(2), 1-39.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational psychologist*, *24*(2), 113-142.

Stamper, John, Gaind, Bharat, Thankachan, Karun, Nguyen, Huy, and Moore, Steven (2023). Hierarchical Concept Map Generation from Course Data. In *AAAI 2023 Workshop on Artificial Intelligence in Education (AI4Edu)*.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. Memory & Cognition, 38, 244-253.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 355-368).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S., Wei, Z., Fan, Z., Huang, Z., Sun, W., Zhang, Q., & Huang, X. J. (2020). PathQG: Neural question generation from facts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9066-9075).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yu, J., Wang, Y., Zhong, Q., Luo, G., Mao, Y., Sun, K., ... & Sun, M. (2021). MOOCCubeX: a large knowledge-centered repository for adaptive learning in MOOCs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 4643-4652).

Zhao, Y., Lei, J., Lai, B. Y. C., & Tan, H. S. (2005). What makes the difference? A practical analysis of research on the effectiveness of distance education. *Teachers College Record*, *107*(8), 1836-1884.