# Focal Vision Document

**Students:**
Razik Grewal, Peter Meyers, Mitali Potnis

Mentors:
Annette Han, John Stamper

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Version 1.0

## 1 Introduction

In an environment where more and more individuals are engaging in online learning, it is vital for educators to provide students with opportunities to practice skills through frequent assessments.[1] These assessments can be taxing on educators, especially when attempting to scale them to the needs of students with varying levels of mastery in a given topic. In order to help educators better meet the need for frequent assessments, research must be done to help elucidate methods for automating the generation of cogent, domain-specific assessments.

On top of the need for automated question generation is the need for generating logically sound answers to those questions. By producing these logically sound answers, student responses can be automatically compared to these 'standard answers' for similarity. Providing educators with a pipeline which allows for the automatic generation and evaluation of quality, subject-specific assessments can significantly lower their workload and give students the feedback and learning opportunities they need.

## 2 Problem Description

With the rise of remote and open education models, instructors are facing a range of new challenges. One such challenge is the need to produce comprehensive and pedagogically sound online evaluations for their students. This challenge was exacerbated by the COVID-19 epidemic, which accelerated the transition to remote learning. [2] As a result, there is an urgent need for automatically producing effective evaluation methodologies and tools that can be scaled in accordance with the needs of instructors and students. Instructors frequently find themselves having to create a large bank of questions to ensure adequacy in supporting efficient and effective assessment of their many students. The most common assessment methods, formative and summative evaluation, focus on directing students' learning processes and enhancing their learning results. In order to effectively assess students using both strategies, creating quality assessments can quickly become time-prohibitive for instructors. This challenge has led to the need for appropriate question-answer combinations that are automatically generated and require little input from domain experts.

With the recent advancements in the fields of natural language processing (NLP), question generation (QG), and question-answering (QA), research has begun on developing these assessment pipelines. [3] [4] [5] This research has included the use of GPT-3 and Text-To-Text Transfer Transformer models, but current techniques fall short in a variety of metrics, including how well they generalize and the variety of questions they produce. A key area for improvement is the generalization of such pipelines in terms of producing a wide variety of question types. Additionally, current pipelines struggle with generalizing to various education domains. A tool adept at creating questions for one subject matter may not be useful in another. For instance, chemistry assessments tend to involve more mathematical and formulaic contexts while computer science assessments tend to contain more algorithmic or theoretically based assessments. Lastly, there is room for improvement in regards to generating more effective distractors for multiple choice questions. These distractors are a useful tool to gauge student understanding. While there is research on the automatic production of distractors for multiple choice questions, current work does not extend to multiple domains. [6]

## 3 Proposed Product/Solution

Most learning theories emphasize repeated practice as an effective model for acquiring low-level knowledge that in turn contributes to high-level educational goals. As a result, we believe that having the capacity to produce questions on-demand will allow educators to redirect their energy to other aspects of educating students while accommodating their various degrees of educational mastery. Our proposed pipeline, Focal, involves using a cutting-edge Text-to-Text Transfer Transformer (T5) model that is pre-trained on conceptual reading materials from one of two courses: a graduate-level data science course as well as an undergraduate chemistry course. Using this pre-trained model, Focal will create relevant evaluation questions to the chosen subject. By training our large language model on multiple educational domains, we hope to take a first step towards creating a domain-agnostic solution to an assessment generation pipeline.

The first step of the Focal pipeline is to extract key concepts from the given course text. We intend to use the MOOCCubeX pipeline in this step of Focal. [7] [5] MOOCCubeX is a pipeline that incorporates a large dataset of educational materials. It has proven adept at extracting key concepts from materials without having to rely on expert input. By extracting the key concepts from a section of material, we can ensure the pipeline is not only generating logically sound questions, but that they also meet the educational goals of the material. These key concepts can further be filtered by domain experts in the relevant field, but the ultimate goal is to eliminate this need for expert input.

After the key concepts of a given section of material have been extracted, the next step is to automatically generate a bank of questions that can be used as student assessments in each of the identified conceptual areas. For Focal, our intent is to use Google's T5- a transformer-based encoder-decoder model for question generation. [8] [5] T5 has shown great efficacy in a wide range of natural language tasks, to include question generation in an educational setting. [9] For the purposes of Focal, we will first fine-tune our T5 model on the SQuAD dataset, which is a well-known benchmark for question-answering models. [10] [5] While we intend to fine-tune our T5 model on the SQuAD dataset, a research goal of this project is to generate a wide variety of question types (starting with "Why" and "How"), to provide a more holistic assessment of student understanding. This goal may necessitate another fine-tuning approach, as SQuAD questions tend to result in more trivial questions.

After question generation, the next phase of Focal is built around judging the pedagogical soundness of the questions. One current method for making these judgements involves a three step process. First, a separate concept hierarchy extraction method is run on the learning resources to identify the significant concept keywords. Questions are then scored based on the number of common words they contain with the key concepts. Second, a fine-tuned GPT-3 model is used to identify the questions as either didactically acceptable or not. Finally, domain experts perform this same classification task manually. [5] A primary research goal of this project is to develop an alternate solution for question quality evaluation. The current method relies on a rather naive process, as simply having overlap with key concepts may not mean a question is pedagogically sound. Our team hopes to create another method for grading question quality that better reflects their soundness for learning purposes. Ideally, these methods and evaluation metrics will approximate true utility of the questions more accurately and with less need for expert input.

After questions have been deemed pedagogically sound, the Focal pipeline will once again use a pre-trained T5 model to generate a variety of answers for each question. These answers can then be compared to student answers to determine similarity. In adding this functionality to Focal, we hope to create a truly end-to-end assessment pipeline with minimal input needed from educators, that can provide students with the opportunities needed to succeed.

## 4 Scientific Hypothesis

For Focal, we have the following hypotheses:

- **H1:** We are able to build a machine learning pipeline that allows users to automatically generate technically sound questions and answers based on provided texts, and the work is sufficient in providing detailed documentation to enable future work.
- **H2:** Focal functionality as a domain-independent tool can be tested by measuring its performance on a variety of data sets and course subject types.
- **H3:** It is possible to improve upon the evaluation metrics that are currently employed to measure the soundness of these questions in a technical setting.

## 5 Major Features

In order to establish a pipeline for automatic assessment production and evaluation, we propose utilizing the following features:

1. **Question Generation**
   (a) **F1:** Automatically extract key concepts from input text and evaluate their alignment with student learning objectives.
   (b) **F2:** Utilize text-to-text transformer to generate questions that evaluate students' mastery of identified these key concepts.
   (c) **F3:** Evaluate the pedagogical soundness of produced questions. Questions should not only be logically sound, they should be usable and valuable for student learning.
2. **Answer Generation**
   (a) **F4:** Generate answers to previously created questions deemed pedagogically sound.
   (b) **F5:** Ensure generated answers are logically sound.
   (c) **F6:** Ensure generated answers align well with and help to answer the question they are intended to.
3. **Answer Evaluation**
   (a) **F7:** Evaluate closeness of student answers to previously identified 'logical answers' for produced assessments.

## 6 Scope

By the end of this capstone project, we will have demonstrated Focal's capabilities as a complete end-to-end assessment pipeline. We will have completed experiments to elucidate the best methods for evaluating the quality of questions and answers. In an effort to make Focal a domain-agnostic solution, its efficacy will have been tested on multiple subject matters with significantly different domains. Additionally, we will have formalized a workflow for Focal with appropriate documentation to enable further research. Rather than focusing on a deployable solution, this capstone will be focused on establishing a pipeline methodology and enabling the extension of our work in future research iterations.

## 7 Timeline

In order to ensure we meet the goal of creating a functional end-to-end assessment pipeline, we have established the following timeline for this capstone research:

- Feb 2023: Literature Review focusing on question generation in specific education domains. Draft requirements document.
- Mar 2023: Draft design and plan documents. Begin initial evaluation of various question generation and answer methods.
- Apr 2023: Draft semester project report. Begin evaluating methods for automatically rating question soundness.
- May 2023: (Tentative) Pitch to Learning Engineering Tools Competition Judges. Conduct semester final presentation.
- Fall 2023: Iteratively improve on each step of the pipeline. Formalize Focal workflow to enable follow-on research.

## 8 Terminology, Definitions, Acronyms, and Abbreviations

- **GPT-3**: Generative Pre-Trained Transformer 3, this product is a large language model that has proven valuable at tasks vital to this project, such as question answering and text generation.
- **MOOCCubeX**: A large dataset of educational materials, in addition to a pipeline that is adept at automatically extracting key educational concepts from text.
- **NLP**: Natural language processing, this is the branch of machine learning that focuses on the processing and generation of natural language.
- **SQuAD**: Stanford Question Answering Dataset, this is a large dataset made of questions on Wikipedia articles and their corresponding answers.
- **T5**: Test-to-Text Transfer Transformer created by Google, considered a state-of-the-art model in many NLP tasks.
- **QA**: Question Answering, this is the process of using NLP models to automatically generate answers to given questions.
- **QG**: Question Generation, this is the process of using NLP models to automatically create questions based on text suppied to the model.

## 9 References

[1] Yong Zhao, Jing Lei, Bo Yan Chun Lai, and Hueyshan Sophia Tan. What makes the difference? a practical analysis of research on the effectiveness of distance education. *Teachers College Record*, 107(8):1836–1884, 2005.

[2] Nabil Hasan Al-Kumaim, Abdulsalam K. Alhazmi, Fathey Mohammed, Nadhmi Gazem, Muhammad Salman Shabbir, and Yousef Fazea. Exploring the impact of the covid-19 pandemic on university students' learning life: An integrated conceptual motivational model for sustainable and healthy online learning. *Sustainability*, 13:25–46, 2021.

[3] Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit. Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 26–33, 2015.

[4] Tianqiao Liu, Qian Fang, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. Mathematical word problem generation from commonsense knowledge graph and equations. arXiv: 2010.06196, 2020.

[5] Huy A. Nguyen, Shravya Bhat, Steven Moore, Norman Bier, and John Stamper. Towards generalized methods for automatic question generation in educational domains. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, pages 272–284, Cham, 2022.

[6] Milan J. Srinivas, Michelle M. Roy, and Viraj Kumar. Towards generating plausible distractors for code comprehension multiple-choice questions. *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, pages 19–22, 2019.

[7] Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wenzheng Feng, Wei-Hao Xu, Shulin Cao, Kaisheng Zeng, Zijun Yao, Lei Hou, Yankai Lin, Peng Li, Jie Zhou, Bingsheng Xu, Juan-Zi Li, Jie Tang, and Maosong Sun. Mooccubex: A large knowledge-centered repository for adaptive learning in moocs. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[8] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*, 2020.

[9] Khushnuma Grover, K. Kaur, Kartikeya Tiwari, Rupali, and Parteek Kumar. Deep learning based question generation using t5 transformer. 2021.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

## 10 Reflection

All told, the process of creating the vision document laid a solid foundation for the beginning of our capstone project. By having some of the main areas of focus identified in the vision document rubric, our team was able to go into our first meeting with our advisor with a good understanding of the important factors to consider and come to a general agreement on from the start. By discussing previous research iterations of this project with our advisor, we were able to gain an understanding of the specifics of the research focus as well as the scope of the project. Our advisor gave us what he believes to be a scope that is attainable in the year-long capstone timeline, while still challenging us to learn and grow throughout this experience. We discussed alternate areas of focus but determined that the research goals previously stated in this vision document reflected the specific area of most interest for our team.

As far as the creation of the vision document is concerned, our team decided to divide the paper into sections. After each team member completed their portion of the vision document, we then proofread each other's work and discussed internally any possible areas of concern. While sometimes the asynchronous nature of dividing a deliverable in this nature can cause some consistency issues with the document itself, we were able to avoid these issues by discussing the nature of the different sections ahead of time to ensure we were on the same page. Additionally, we set ourselves up for success by imposing a deadline that was ahead of the course deadline to ensure that there was adequate time for revision and to allow our advisor to evaluate the vision document ahead of the final submission. Going forward, these early deadlines for deliverables will remain crucial in ensuring we consistently produce quality work.