Carnegie Mellon University
Master of Computational Data Science

Carnegie Mellon University
Language Technologies Institute

Focal

Midterm Presentation

Razik Grewal, Peter Meyers, Mitali Potnis
Advisor: Annette Han, John Stamper
25 March 2023

11-634 (Spring 2023)
MCDS Capstone Planning Seminar

Introduction

- Why is online learning an important avenue for research?
  - Huge growth in the sector
  - Growth accelerated by Covid-19
- Why are evaluations important?
  - Research shows frequent assessments are vital to student development (Koedinger et al., 2012)

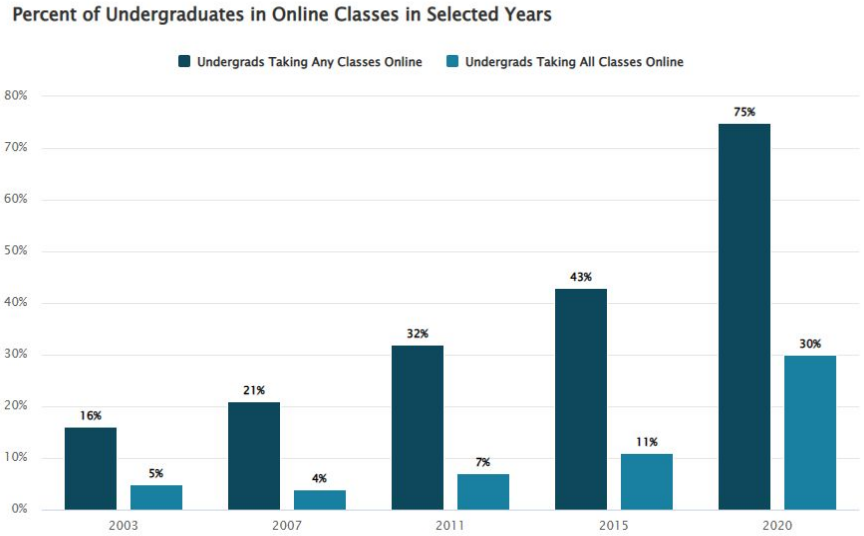
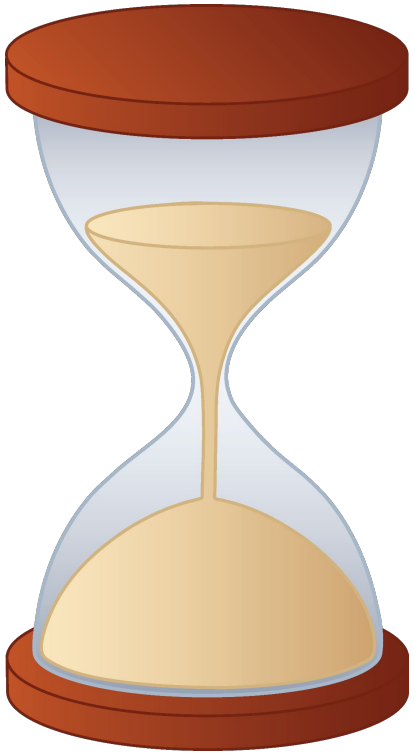


Fig. 1. Trend in online undergraduate classes over the years

https://www.bestcolleges.com/research/online-learning-statistics/#fn-ref-2

Problem Description

- Creating assessments is time consuming
  - Many questions are required
  - Scaling to student performance
  - Variety is vital to student learning
- Providing timely feedback is taxing for educators
  - Short answer assessment types exacerbate this issue
- Assessment generation and evaluation quickly becomes time-prohibitive for educators



https://clipground.com/clipart-sand-timer.html

Proposed Product/Solution

- End-to-end evaluation pipeline

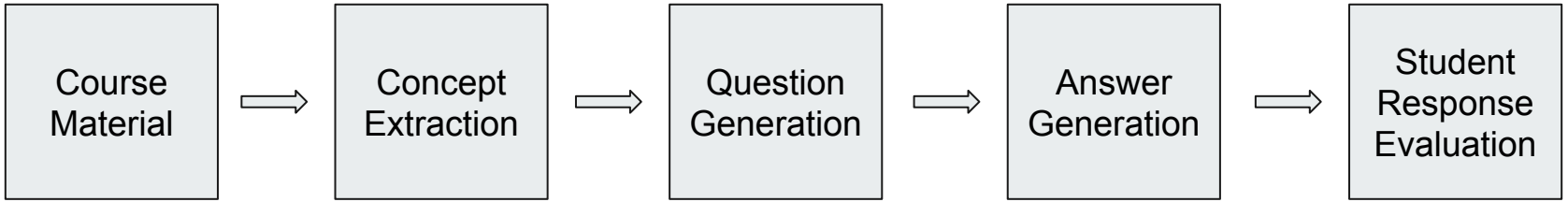


Fig. 2. Project Pipeline

Scientific 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.

Major Features

- **Question Generation**
  - F1: Key concepts extraction/evaluation
  - F2: Question generation with LLM
  - F3: Question evaluation
- **Answer Generation**
  - F4: Answer generation
  - F5: Answer evaluation for logical soundness
  - F6: Answer evaluation for meeting learning objective
- **Answer Evaluation**
  - F7: Student response evaluation

Intended Users

- **Type A Users: Educators**
  - Provide Focal with course material
  - Receive generated list of questions and answers
  - Receive automatically generated student evaluations

- **Type B Users: Students**
  - Receive automatically generated questions
  - Respond to questions and receive immediate feedback

- **Type C Users: Researchers**
  - Enable further research through appropriate documentation

System Functionality

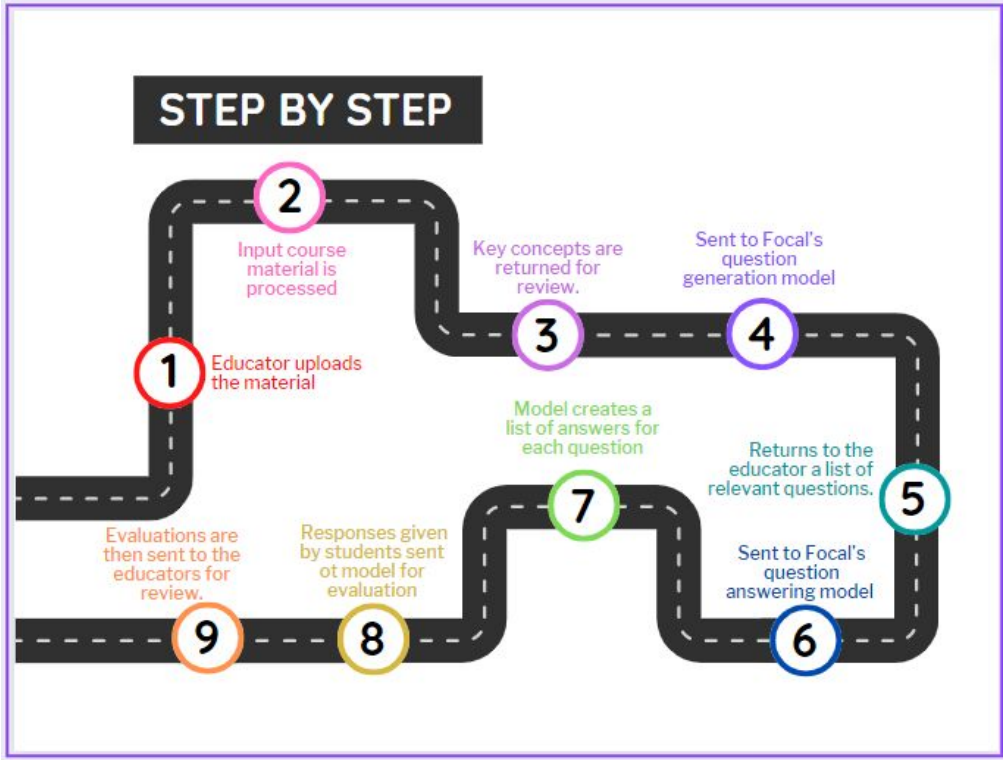


Fig. 3. Functions of our System

Non-Functional Requirements

- **Performance**
  - QG/QA in less than 5 minutes
  - Student feedback in less than 2 seconds
- **Accuracy**
  - Questions are logically and pedagogically sound
  - Questions vary in complexity and style
  - Answers are correct and robust to all possible solutions
- **Usability**
  - Simple UI for novice users
  - Reduces burden of assessments for educators
- **Scalability**
  - Domain-agnostic
  - Able to incorporate different class sizes and setups
- **Maintainability**
  - Proper documentation
- **Security**
  - Student data is protected

Resource Requirement

- **Hardware**
  - The system requires a server or cloud-based infrastructure to run the machine learning (ML) models that generate questions and answers.
- **Software**
  - The system requires software tools for natural language processing and machine learning, such as GPT-3, MOOCCubeX (Yu et al., 2021) and T5. (Xue et al., 2020)
  - The system must have a user interface, which could be developed using a variety of programming languages and frameworks.
- **Data**
  - The system requires a large dataset of educational materials, such as and SQuAD (Rajpurkar et al., 2016) and LearningQ (Chen et al., 2018) for fine tuning our ML models.
  - The system also requires a high-quality question and answers dataset to evaluate the pedagogical soundness of the questions it generates.
- **Human resources**
  - The project requires a team of developers, data scientists, and domain experts to design, develop, and test the system.
  - The project also requires educators and students to test and provide feedback on the system.
- **Time**
  - The project requires a year-long capstone timeline to complete the research goals and formalize a workflow for Focal.
  - The project will also require additional time for future research iterations to fully develop Focal into a complete assessment pipeline.

Design Considerations

- **Assumptions**
  - Courses are similar in design
  - Educators generally evaluate questions the same
- **Constraints**
  - Must be fast and highly accurate in terms of Student feedback and question generation
  - Questions must be consistently high quality
- **System Environment**
  - Hosted within OLI (after future iterations)
  - Hosted in a Github repository for future development
- **Design Methodology**
  - End-to-end tool
  - Minimal input from domain experts and educators

System Architecture/Design Overview

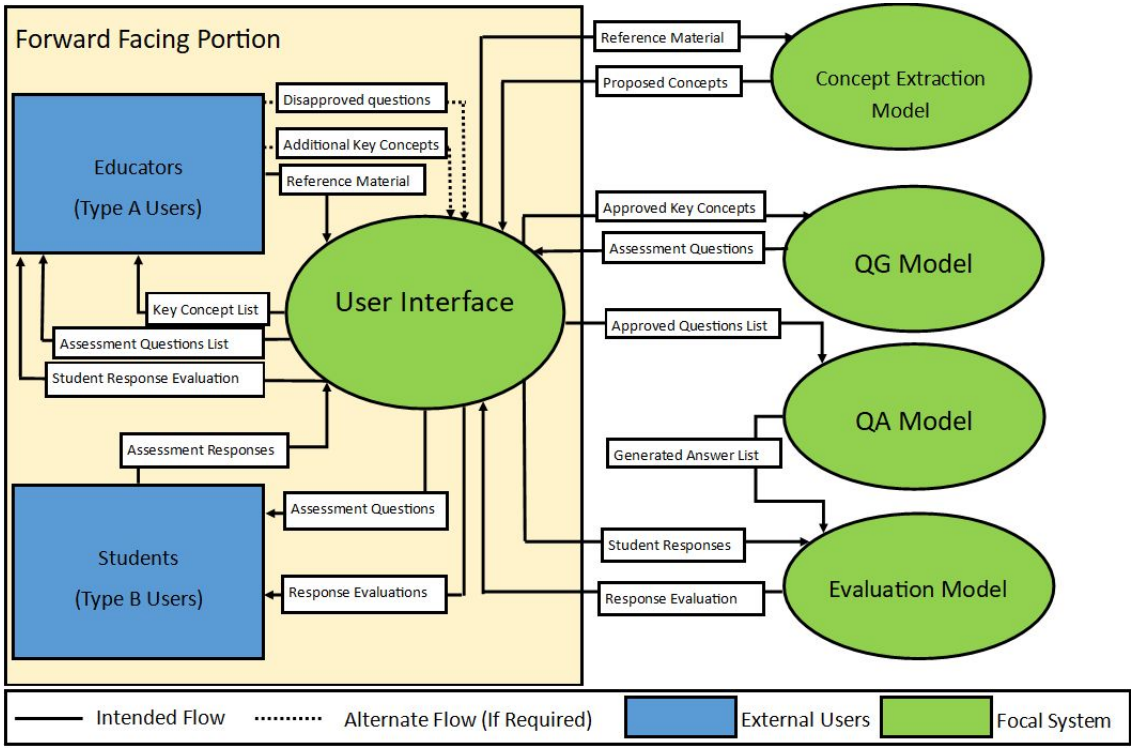


Fig. 4. Focal System Architecture

Data Design

- **Data Pre-pending:** OLI Content for training our QG and QA models (Huy et al., 2022)

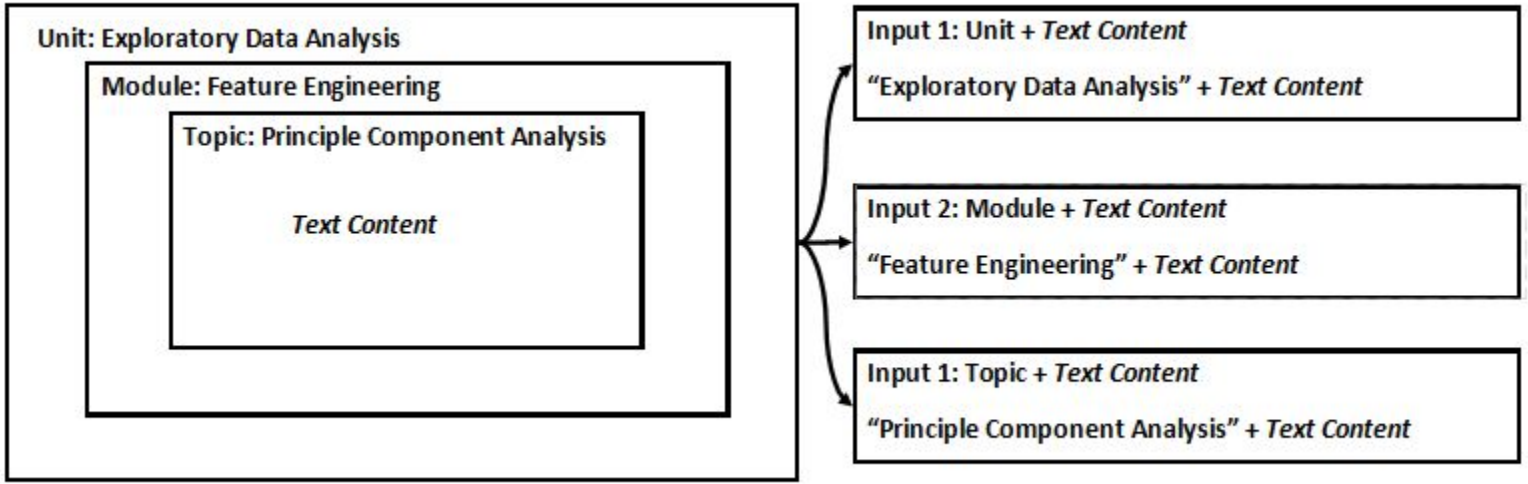


Fig. 5. Data Pre-pending for Machine Learning model training

- **SQuAD 1.1**: Fine-tuning QG model (Rajpurkar et al., 2016)
- **LearningQ**: Fine-tuning evaluator model (Chen et al., 2018)

Data Design

	Unit	Module	Title	Text	Subheaders
0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management
1	Data Gathering and Wrangling	Data Collection Process	Summary and Quiz 2	Each data science project is unique and will r...	
5	Analytic Algorithms and Model Building	Data Science Patterns	Summary and Quiz 6	Prediction involves using a model to predict o...	
7	Analytic Requirements Gathering	Requirements Gathering Techniques	Successful Requirements Gathering	The requirements gathering process is not line...	Validating Requirements
8	Exploratory Data Analysis	Feature Engineering	Summary and Quiz 5	In this module, we explored a technique used t...	

Fig. 6. Extracted Data from the OLI XML files

Implementation & Design Models Overview

The overall system workflow comprises of 4 sections:

- **Data Extraction and Pre-processing** - Involves scraping (using the BeautifulSoup library), cleaning of input XML data, pre-pending the text.

- **Concept Hierarchy Extraction** - Implementing the MOOCCubeX pipeline for fine-grained concepts extraction and removal of invalid concepts.

- **Question and Answer generation** -
  - ❖ Question generation: Implementing a fine-tuned transformed-based encoder-decoder T5 model. Extracting top 3 questions for each topic.
    1) Multiple-Choice questions (Questions + Distractors)
    2) Short-answer type questions
  - ❖ Answer generation: Performed using a custom rule-based-approach involving dependency parse tree
- **Evaluation** - Evaluating performance

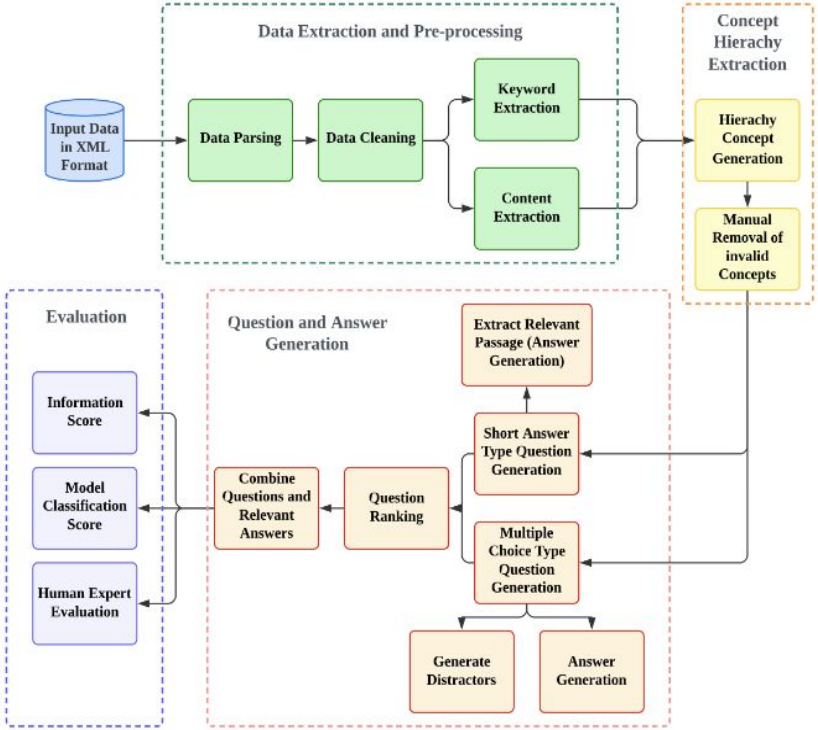


Fig. 7. Focal System Implementation Overview

Implementation & Design Models Overview

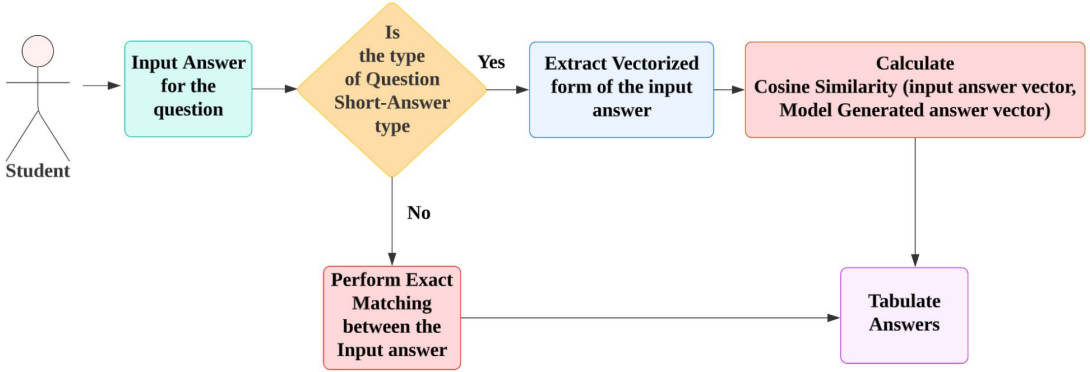


Fig. 8. Answer Evaluation Implementation

- Evaluation of the generated Questions and Answers -
  - ❖ Question Evaluation: Using Information score, model (GPT-3 classification), and human expert evaluation.
  - ❖ Answer Evaluation: a) Short-answer type questions- Calculate cosine similarity score between gold answer and input answer vectors that are generated using .bag-of-words method or tf-idf
    b) Yes-No, Multiple-choice question- Perform exact matching between input and gold answers

Test Design

- **Information Score** (Huy et al., 2022)

  Information Score(q)= $\frac{1}{|T(q)|}\sum_{t\in T(q)} 1(t \in C)$

- **Human Expert Evaluation**
- **Model Classification**
- **Cosine Similarity and Exact Matching**

  Cosine Similarity(A,S)= $\frac{D}{|A|*|S|}$

Table 1: Evaluation of generated questions across different header levels and soundness ratings

Generated Question	Header Level (Topic/ Module/Unit	Information score	GPT-3 Classification (Sound/Unsound)	Expert Rating (Sound/Unsound)
-	-	-	-	-

Table 2: Confusion Matrix for comparing GPT-3 and expert evaluations.

	Expert: Not Sound	Expert: Sound
GPT3: Not Sound	-	-
GPT3: Sound	-	-

Table 3: Metric Value for Questions.

Type of Question	User answer	Evaluation Strategy(Exact Matching/ Cosine Similarity)	Metric Value
-	-	-	-

Deployment Model

- Focal will be deployed through the Open Learning Initiative (OLI).
- Integrated with the course content.
- Automated generation of valuable assessments minimizing educator’s involvement.
- Automatic evaluation and feedback for students’ responses.
- Document details regarding pipeline and any research on Github repository.

Risks/Challenges

- **Diverse Course Content**
  - Additional course content may have different course structures requiring large modifications to the Focal pipeline leading to delay in development.
  - Making the pipeline generalized for different course structures.

- **Improper Questions/Evaluations**
  - Illogical questions or questions irrelevant and of lesser importance in the context of the course content and testing the course understanding are generated.
  - Unfair or inaccurate evaluation of student responses.
  - Employing multiple pipeline and model quality evaluation tests in terms of questions generated and accuracy of student response evaluation.

- **Question Variability**
  - Questions generated are either trivial or primarily "what" questions rather than "why" or "how" questions.
  - Perform question generation evaluation strategies early on to modify the model and the pipeline as per the requirements.

Tools & Dependencies

- **External Libraries**
  - Python v3.7
  - Numpy v1.19.5
  - TensorFlow v1.15
  - Pandas v1.5.3
  - BeautifulSoup v0.12.2
  - Gensim v4.3.1
  - Sklearn v1.2.2
  - Matplotlib v3.4
  - Pipelines v2.7
  - Transformers v2.1.0
  - Nltk v3.8.1
- **Datasets**
  - FCDS Course Data
  - Chemistry Course Data
  - SQuAD 1.1 (Rajpurkar et al., 2016)
  - LearningQ (Chen et al., 2018)
- **Repository**
  - Previous Focal Repository (Huy et al., 2022)
- **Hardware Requirements**
  - Multiple CPUs with a RAM of 16-32 GB
  - AWS EC2 m5.large instance virtual Machine

Tasks

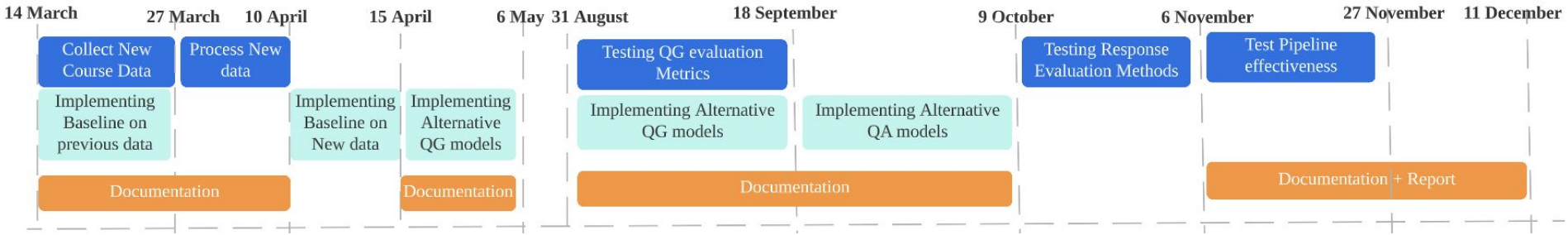


Fig.9. Focal Project Timeline

Milestone Plan

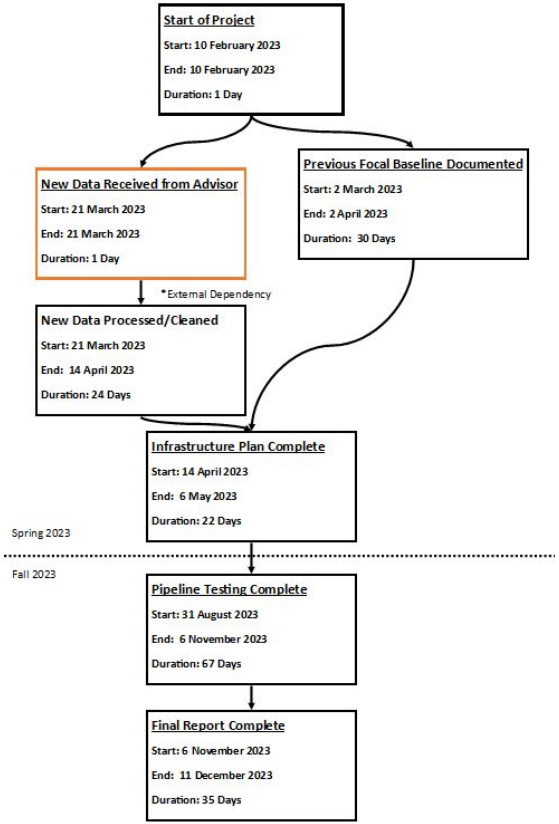


Fig.10. Focal project Milestone Plan

Baseline Implementation

T5 For Question Generation:

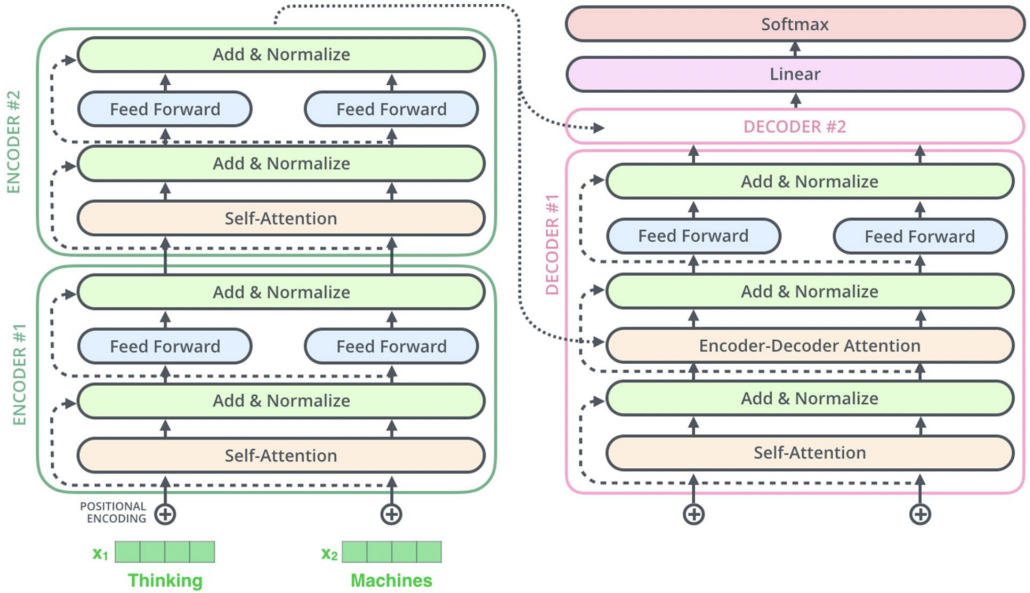


Fig.11. T5 Model Architecture (Alammar, 2020)

Baseline Implementation

	index	Unit	Module	Title	Text	Subheaders	Paragraph
0	0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management	A data scientist's role involves utilizing com...
1	0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management	A data scientist's role involves utilizing com...
2	0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management	A data scientist's role involves utilizing com...
3	0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management	A data scientist's role involves utilizing com...
4	0	Data Gathering and Wrangling	Data Gathering	Data Management	A data scientist's role involves utilizing com...	Data Gathering Overview,Data Management	A data scientist's role involves utilizing com...

Fig.12. OLI extracted data

A data scientist's role involves utilizing computational and statistical skills to uncover solutions that meet business needs. Throughout the process of discovery and solution development, a data scientist must find it most useful and efficient to focus on developing algorithms and building the models that will support the analytic solution for the business need. This unit sheds light on a crucial phase of the data science project lifecycle that will inform the solution development.

Table 4. Questions Generated

What skills does a data scientist use to uncover solutions that meet business needs?
How much of your time will you spend understanding, exploring, and transforming the data used for model building?
What does the quality of your data have a direct impact on?
What is a popular term for data transformation?
What is the process of collecting, selecting, and transforming data to ensure it is usable, free of noise?
What project team will continue to work with the data until it is ready to be used?
What is the name of the project team that works with data science?

References

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. Learning: A large-scale dataset for educational question generation. Proceedings of the International AAAI Conference on Web and Social Media, 12, 2018. https://ojs.aaai.org/index.php/ICWSM/article/view/14987

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. https://www.semanticscholar.org/paper/MOOCCubeX%3A-A-Large-Knowledge-centered-Repository-in-Yu-Wang/733945ff31b7db4795a5b3038a637d04af8e0276

Kenneth Koedinger, Albert Corbett, and Charles Perfetti. The Knowledge Learning Instruction framework: Bridging the science & practice chasm to enhance robust student learning. Cognitive science, 36(5), 757-798, 2012.http://pact.cs.cmu.edu/pubs/Koedinger,%20Corbett,%20Perfetti%202012-KLI.pdf

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. https://aclanthology.org/2021.naacl-main.41/

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. https://aclanthology.org/D16-1264/

Alammar,J.:Theillustratedtransformer-jayalammar-visualizingmachinelearn- ing one concept at a time. http://jalammar.github.io/illustrated-transformer/. Accessed 29 Oct 2020

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. (Yu et al., 2021)
- **NLP**
  - Natural language processing, this is the branch of machine learning that focuses on the processing and generation of natural language.
- **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 supplied to the model.
- **SQuAD**
  - Stanford Question Answering Dataset, this is a large dataset made of questions on Wikipedia articles and their corresponding answers. (Rajpurkar et al., 2016)
- **Bag-of-words**
  - Text representation technique that transforms given text into a bag or set of its words without considering the order of the words and grammar to analyze the frequency of each word in a corpus.
- **Tf-idf**
  - Term Frequency-Inverse Document Frequency is a statistical measure for evaluating importance of a word in a document on the basis of its frequency in the document and its rarity in the involved corpus for identifying the most relevant words in any given document or corpus.
- **T5**
  - Test-to-Text Transfer Transformer created by Google, considered a state-of-the-art model in many NLP tasks. (Xue et al., 2020)
- **LearningQ Dataset**
  - A large dataset covering 230 thousand document-question pairs and 7000 instructor-designed questions on a variety of educational topics. (Chen et al., 2018)

Reflection

- The process of the capstone planning course so far has helped us to really understand the steps we need to take to ensure success in the project. By working with our mentor, we were able to create a vision for the project that helped us understand the goal for this year. After that, focusing on requirements gave us a more fine-grained understanding of the intricacies of the project. The design document allowed us to focus on the workflow of the system and its individual components. Lastly, the plan document helped us create a realistic timeline for each phase of the project.
- Our group has thus far had success with working as a team. Communication has been key to enabling us to meet our timeline so far. When one partner is falling behind, they have been quick to communicate and the others have helped meet the timeline.
- In order to continue to have success in the project, we need to continue to improve on our scrum planning. Getting more fine-grained on weekly or bi-weekly tasks can help us more accurately gauge progress to ensure we meet our research goals by the end of the project.

Changes To Previous Deliverables

Modified the Timeline to include more phases of the project.