Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Bonn-Aachen
International Center for
Information Technology

Analysis of Active Learning Mechanism Applied to Language Models for Computer Assisted Short Answer Grading

September 28, 2022

Elanton Fernandes

Advisors

Prof. Dr. Paul G. Plöger, M.Sc Tim Metzler

Agenda

Motivation

In universities with an increase in number of student every semester, the number of tests conducted also increases. This means that:

- The professor spends more time in correcting student exams than preparing for lectures.
- If students are not assigned full scores for on a test, they expect a meaningful feedback from the professor.

Motivation

Consider the following dummy scenario:

- 80 students enrolled in a class.
- Tests are conducted bi-weekly.
- Professor requires 15 minutes to evaluate one student test.
- Total time spent by the professor to evaluate all tests per week is 10 hours.

Problem Statement

- Automate the evaluation of student tests while still keeping the oracle/professor in the loop.
- Allow the assignment of meaningful feedback to student answers indicating their mistakes.

Related Work

- Wu et al. (2021) designed an system to assign feedback called ProtoTransformer for evaluating programming based questions but not for short text answers. It used limited number of examples [1].
- Ghavidel et al. (2020) passed raw text through a transformer as input and used the output of classification model (CLS) token as feature [2].
- Mieskes and Pado, (2018) compared score assignment between automated and human assignment for RF, SVM, and DT classifiers across multiple datasets [3].

---

[1] M. Wu, N. Goodman, C. Piech, and C. Finn, "Prototransformer: A meta-learning approach to providing student feedback," 2021.

[2] H. Ghavidel., A. Zouaq., and M. Desmarais., "Using bert and xlnet for the automatic short answer grading task," in Proceedings of the 12th International Conference on Computer Supported Education - Volume 1: CSEDU,, pp. 58–67, INSTICC, SciTePress, 2020.

[3] M. Mieskes and U. Padó, "Work smart - reducing effort in short-answer grading," in Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning, (Stockholm, Sweden), pp. 57–68, LiU Electronic Press, Nov. 2018.

Dataset

Dataset	Domain	No. of Question Pairs	No. of Responses
Mohler [4]	Computer Science	81	2237
NN Exam [5]	Neural Network & AI	40	1137
AMR Exam [6]	Robotics	5	190

Table 1: Datasets used in score and feedback evaluation.

[4] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," June 2011.

[5] P. G. Plöger, "Neural network exam dataset," 2020.

[6] N. Hochgeschwender, "Autonomous mobile robots exam dataset," 2021.

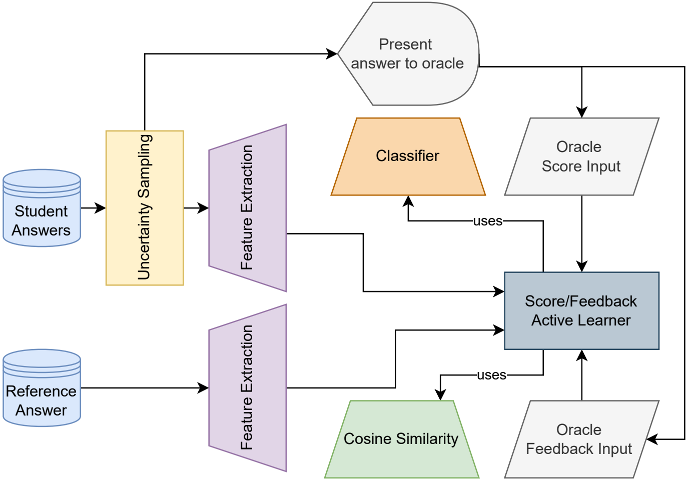
Approach

Contributions

- Implement four methods to alter text for feature extraction.
- Implement feedback assignment for short text answers.
- Compare performance with five pre-trained language models and two classifiers.

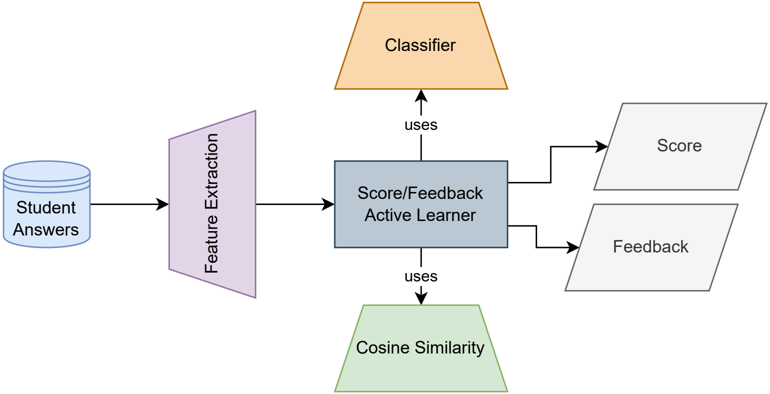
Approach

Training Cycle



Approach

Prediction Cycle



Approach

Uncertainty Sampling

Uncertainty sampling [7] is a query strategy that queries the instances about which it is least certain how to label. We use uncertainty sampling variant might query the instance whose prediction is the least confident:

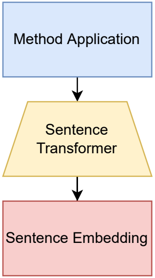
$$x_{LC} = argmin_x P(\hat{y}|x;\theta) \tag{1}$$

Where x is the feature, y is the class label prediction, and $\hat{y} = argmax_y P(y|x;\theta)$ is the class label that has the largest posterior probability using model θ.

[7] B. Settles, "Computer Sciences Department Active Learning Literature Survey," 2009.

Approach

Feature Extraction: Overview

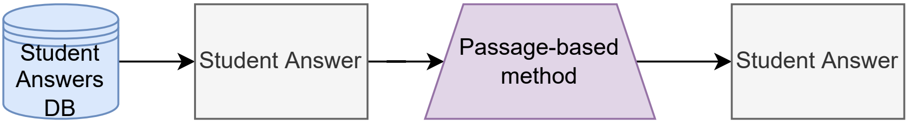


[8]

[8] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” 2019.

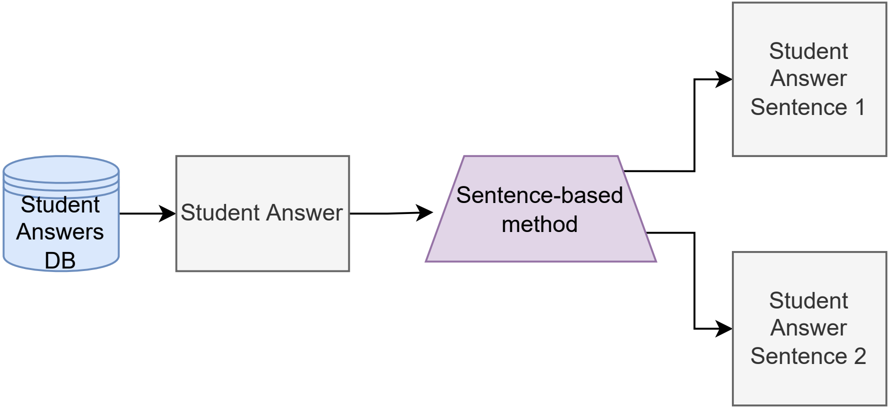
Approach

Feature Extraction: Passage-Based Method



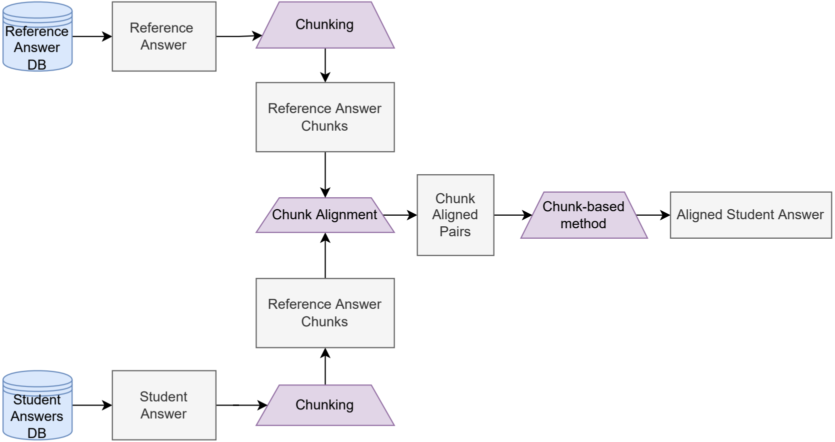
Approach

Feature Extraction: Sentence-Based Method



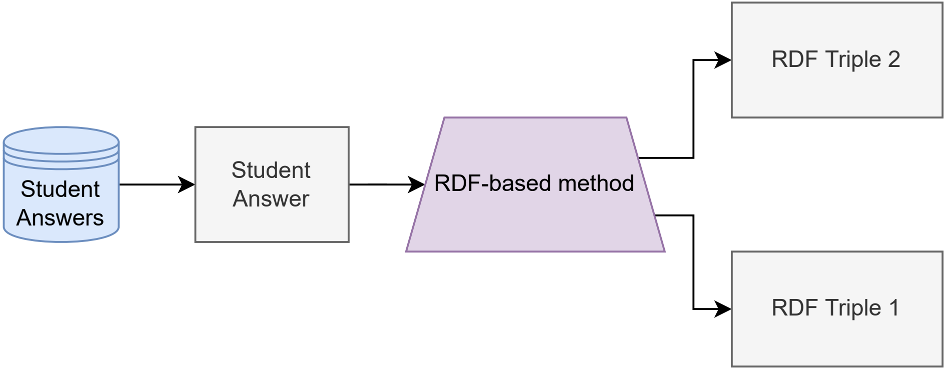
Approach

Feature Extraction: Chunk-Based Method



Approach

Feature Extraction: Resource Description Framework (RDF) Based Method



Approach

Language Models

Language Model [9]	Base model	Number Training tuples
all-mpnet-base-v2	microsoft/mpnet-base.	1.17B
all-distilroberta-v1	distilroberta-base	1.12B
all-MiniLM-L12-v2	microsoft/MiniLM-L12-H384-uncased	1.17B
multi-qa-distilbert-cos-v1	distilbert-base	214M
all-MiniLM-L6-v2	nreimers/MiniLM-L6-H384-uncased	1.17B

Table 2: Displays pre-trained language models with their base model used in training and number of training tuples used.

[9] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

Evaluation

Score

- Pearsons Correlation

$$\rho(y, \hat{y}) = \frac{cov(\vec{y}, \hat{\vec{y}})}{\sigma_y \sigma_{\hat{y}}} \tag{2}$$

- RMSE Score

$$RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} (\hat{\vec{y}_i} - \vec{y}_i)^2} \tag{3}$$

Where $\vec{y}$ represents actual grade and $\hat{\vec{y}}$ represents predicted grade with σ_y and $\sigma_{\hat{y}}$ computed as the standard deviation of $\vec{y}$ and $\hat{\vec{y}}$

Evaluation

Feedback

Question	What is a variable?
Reference Answer	A location in memory that can store a value.
Student Answer	A value/word that can assume any of a set of values
Feedback A	Correct
Feedback B	Missing keywords: Location in memory
Feedback C	A variable is a location in memory that stores a value

Table 3: Presented survey to participants.

$$Agreement\ Score = \frac{Model\ generated\ most\ rated\ feedback}{Total\ Number\ of\ Participants} \times 100$$

Results

Notations

Method	Notation
Passage-based Methods	M1
Sentence-based Method	M2
Chunk-based Method	M3
RDF-based Method	M4

Language Model	Notation
all-mpnet-base-v2	LM1
all-distilroberta-v1	LM2
all-MiniLM-L12-v2	LM3
multi-qa-distilbert-cos-v1	LM4
all-MiniLM-L6-v2	LM5

Results

Score: Pearson Correlation (Methods)

Dataset	M1	M2	M3	M4
Mohler	**0.826**	**0.791**	**0.816**	0.782
NN Exam	**0.941**	**0.828**	0.561	**0.846**
AMR Exam	**0.658**	0.458	**0.640**	0.428

(a)

Dataset	M1	M2	M3	M4
Mohler	0.689	0.627	0.687	**0.792**
NN Exam	0.889	0.791	**0.638**	0.664
AMR Exam	0.622	**0.474**	0.593	0.428

(b)

Table 4: Comparison of Pearson Correlation between Random Forest (a) and AdaBoost (b) classifiers. Where M1: Passage-based, M2: Sentence-based, M3:Chunk-based, and M4: RDF-based method.

Results

Score: Pearson Correlation (Language Models)

Dataset	LM1	LM2	LM3	LM4	LM5
Mohler	**0.802**	**0.797**	**0.796**	**0.796**	**0.789**
NN Exam	**0.732**	**0.670**	**0.705**	**0.755**	**0.760**
AMR Exam	0.453	**0.518**	**0.525**	**0.523**	**0.503**

(a)

Dataset	LM1	LM2	LM3	LM4	LM5
Mohler	0.659	0.673	0.211	0.544	0.499
NN Exam	0.614	0.653	0.704	0.698	0.605
AMR Exam	**0.502**	0.440	0.430	0.508	0.467

(b)

Table 5: Comparison of Pearson Correlation between Random Forest (a) and AdaBoost (b) classifiers with language models (LM).

Results

Score: Root Mean Square Error (Methods)

Dataset	M1	M2	M3	M4
Mohler	**0.893**	**0.949**	**0.920**	0.942
NN Exam	**0.296**	**0.520**	**0.433**	**0.522**
AMR Exam	**0.596**	0.716	**0.596**	**0.736**

(a)

Dataset	M1	M2	M3	M4
Mohler	1.218	1.226	1.169	**0.920**
NN Exam	0.405	0.571	0.495	0.741
AMR Exam	0.616	**0.707**	0.630	0.741

(b)

Table 6: Comparison of RMSE score between Random Forest (a) and AdaBoost (b) classifiers with methods (M).

Results

Score: Root Mean Square Error (Language Models)

Dataset	LM1	LM2	LM3	LM4	LM5
Mohler	**0.931**	**0.941**	**0.941**	**0.941**	**0.956**
NN Exam	**0.484**	0.591	**0.558**	**0.490**	**0.492**
AMR Exam	0.735	**0.680**	**0.676**	0.684	**0.698**

(a)

Dataset	LM1	LM2	LM3	LM4	LM5
Mohler	1.182	1.163	1.667	1.278	1.363
NN Exam	0.632	**0.582**	0.587	0.587	0.650
AMR Exam	**0.692**	0.748	0.718	**0.682**	0.736

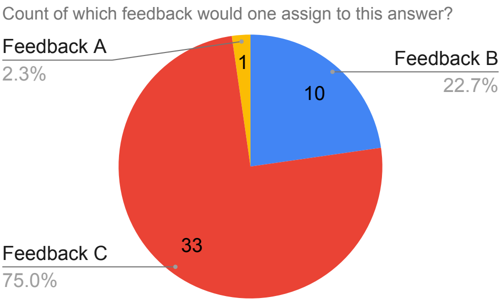
(b)

Table 7: Comparison of RMSE score between Random Forest (a) and AdaBoost (b) classifiers with language models (LM).

Results

Feedback: Survey Results

Question	What is a variable?
Reference Answer	A location in memory that can store a value.
Student Answer	A value/word that can assume any of a set of values
Feedback A	Correct
Feedback B	Missing keywords: Location in memory
Feedback C	A variable is a location in memory that stores a value



Hochschule Bonn-Rhein-Sieg University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Results

Feedback: Agreement Scores (Methods)

Classifier	Methods			
	M1	M2	M3	M4
Random Forest	**60.00**	22.73	31.82	35.91
AdaBoost	**60.00**	22.73	31.82	35.91

Table 8: Mean agreement scores for Random Forest (a) and AdaBoost classifier (b) with methods.

Results

Feedback: Agreement Scores (Models)

Classifier	Language Models				
	LM1	LM2	LM3	LM4	LM5
Random Forest	25.11	26.82	24.66	**37.05**	21.25
AdaBoost	25.11	26.82	24.66	**37.05**	21.25

Table 9: Mean agreement scores for Random Forest and AdaBoost classifier with Language Models.

Results

Summary: Scores

Dataset	Method	Model	Classifier
Mohler	M1	LM1	Random Forest
NN Exam	M1	LM5	Random Forest
AMR Exam	M1	LM3	Random Forest

(a)

Dataset	Method	Model	Classifier
Mohler	M1	LM1	Random Forest
NN Exam	M1	LM1	Random Forest
AMR Exam	M1& M3	LM3	Random Forest

(b)

Table 10: Performance summary Pearson correlation (a) and RMSE score (b)

Results

Feedback

Dataset	Method	Model	Method-Model	Classifier
Mohler	M1	LM4	M1-LM4	Random Forest

Table 11: Results of feedback evaluation

Summary

- Four methods implemented to alter student answer text.
- Designed system that keeps oracle in training loop.
- System allows assignment of feedback.
- Pearson correlation and RMSE score used as score evaluation metrics.
- Survey created and used in the evaluation of the feedback assigned by the model.

Future Work

- RDF-based method did not give good RDF triplets. Fine tune model for better results.
- Fine tune language models similar to multi-qa-distilbert-cos-v1 to create embeddings.
- Chunk-based Method took 41 sec per prediction. This time needs to be reduced.

Foundation

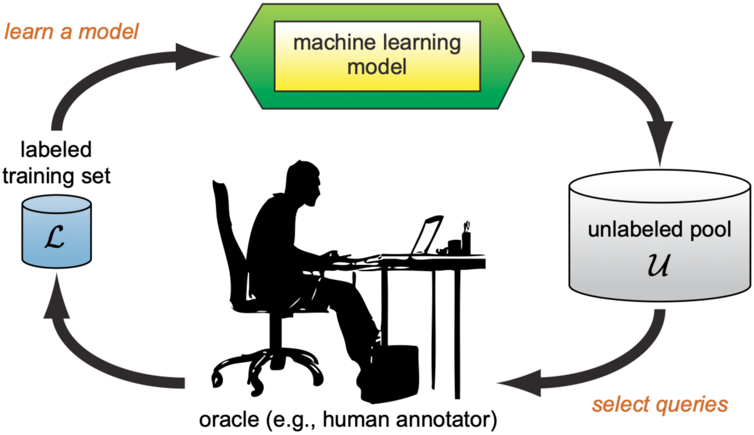
Cosine Similarity

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sentences and $\boldsymbol{e_x}$ and $\boldsymbol{e_y}$ are embeddings of these sentences. Then the cosine similarity is given by:

$$S_{cosine}(\boldsymbol{e_x}, \boldsymbol{e_y}) = \frac{\boldsymbol{e_x} \cdot \boldsymbol{e_y}}{\|\boldsymbol{e_x}\| \|\boldsymbol{e_y}\|} \tag{4}$$

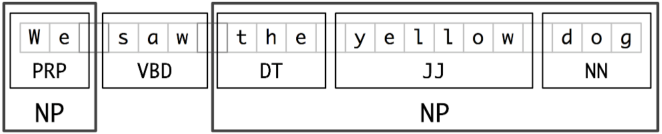
Foundation

Active Learning



Foundation

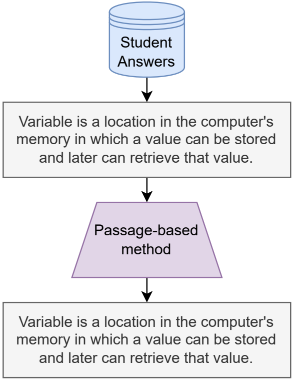
Chunking



Symbol	Meaning	Symbol	Meaning
NP	Noun Phrase	VP	Verb phrase
NN	Noun	DT	zero or one determiner
JJ	One or more adjectives	PRP	Preposition

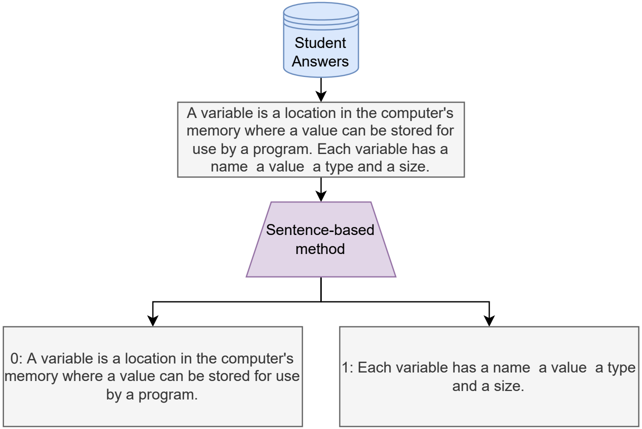
Approach:Example

Feature Extraction: Passage-based method



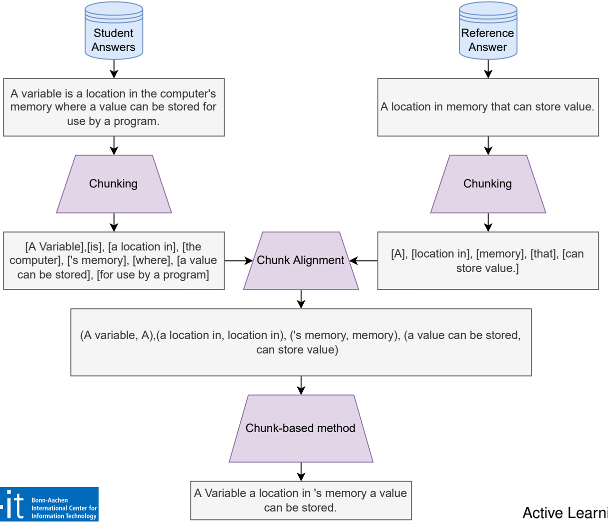
Approach

Feature Extraction: Sentence-based method



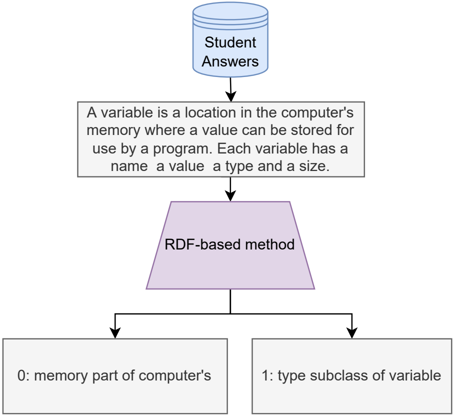
Approach

Feature Extraction: Chunk-based method



Approach

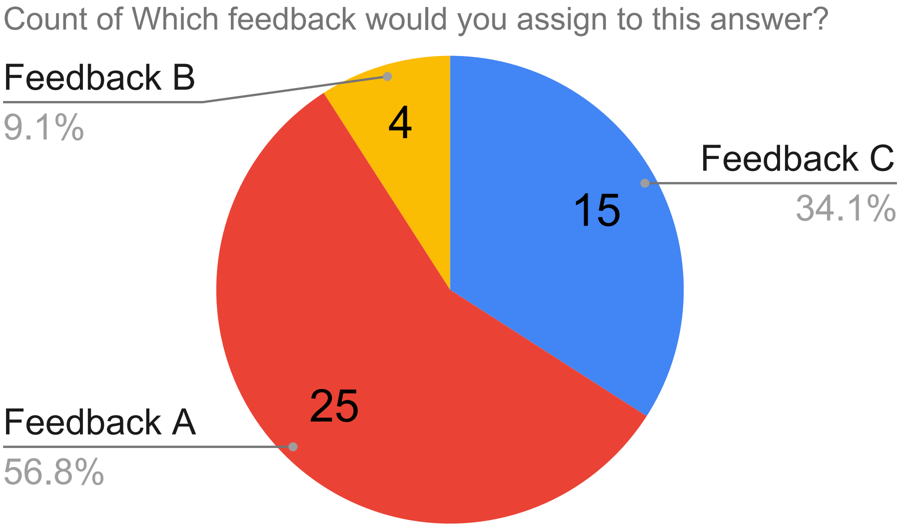
Feature Extraction: RDF-based method



Results

Feedback: Survey Results

Question	What stages of the software lifecycle are influenced by the testing stage
Reference Answer	The testing stage can influence both the coding stage (phase 5) and the solution refinement stage (phase 7)
Student Answer	All stages are influenced except setting the program requirements. If a test fails it can change the whole design implementation etc of a program as well as the final outcome.
Feedback A	The testing phase affects the coding/production phase and the refinement/maintenance phase
Feedback B	Correct
Feedback C	Missing keywords: Refinement stage,Coding phase

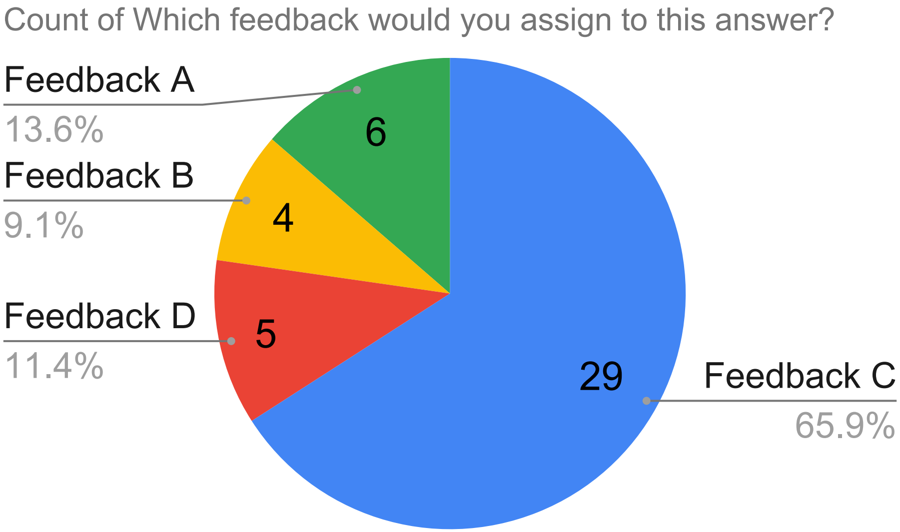


Count of Which feedback would you assign to this answer?
Feedback B
9.1%
4
Feedback C
34.1%
15
25
Feedback A
56.8%

Results

Feedback: Survey Results

Question	What are the main advantages associated with object oriented programming?
Reference Answer	Abstraction and reusability.
Student Answer	Re-usability and ease of maintenance
Feedback A	Missing keywords: Reusability, Abstraction
Feedback B	Correct
Feedback C	The main advantages of OOP are abstraction and reusability
Feedback D	Encapsulation is similar to abstraction but not the same and second advantage is reusability.

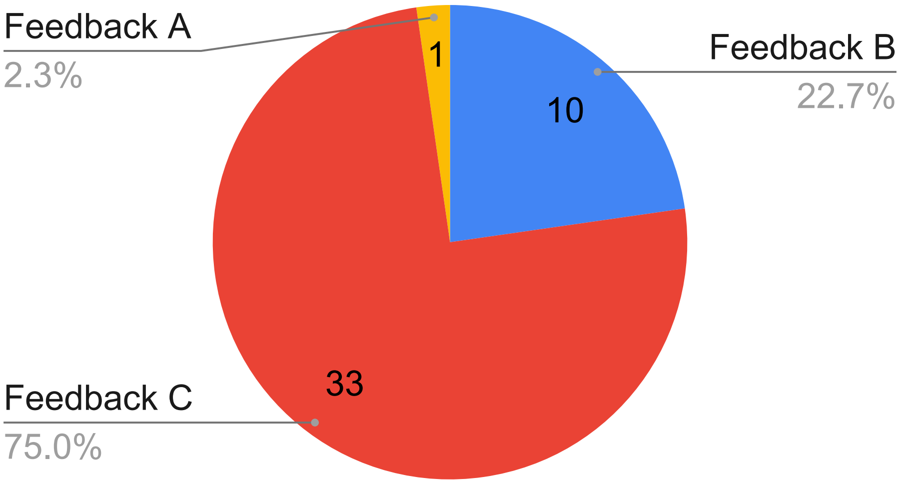

Count of Which feedback would you assign to this answer?
Feedback A
13.6%
6
Feedback B
9.1%
4
Feedback D
11.4%
5
29
Feedback C
65.9%

Results

Feedback: Survey Results

Question	What is a variable?
Reference Answer	A location in memory that can store a value.
Student Answer	a value/word that can assume any of a set of values
Feedback A	correct
Feedback B	missing keywords: location in memory
Feedback C	A variable is a location in memory that stores a value

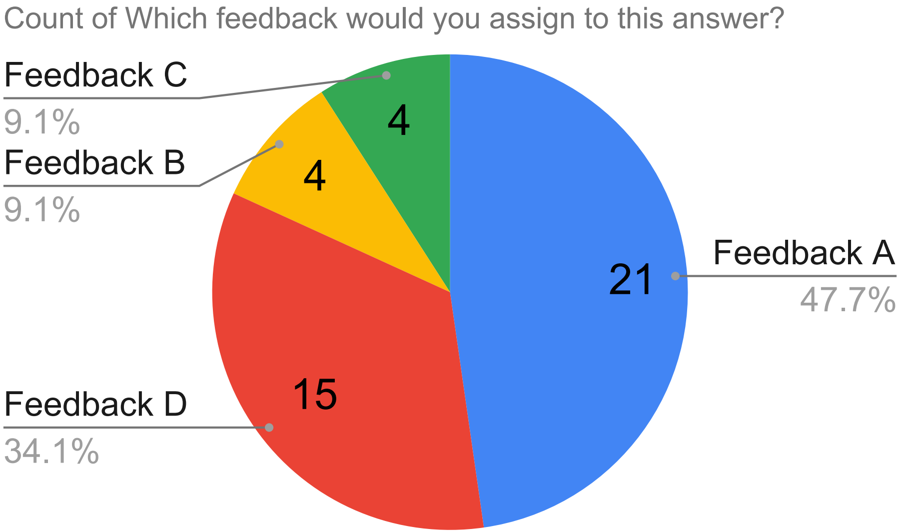
Count of which feedback would one assign to this answer?



Results

Feedback: Survey Results

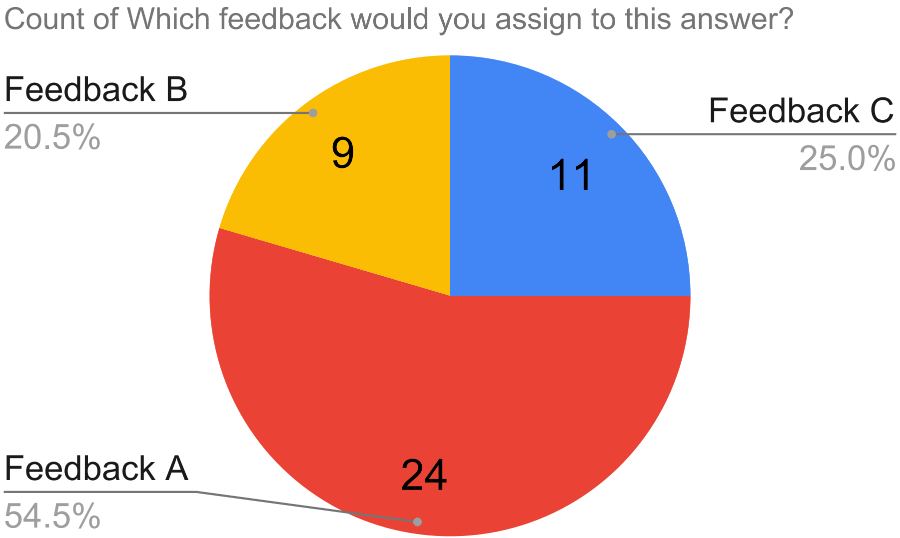
Question	What is the difference between a constructor and a function?
Reference Answer	A constructor is called whenever an object is created whereas a function needs to be called explicitly. Constructors do not have return type but functions have to indicate a return type.
Student Answer	Constructors don't have a return type.
Feedback A	Missing point: A constructor is called whenever an object is created whereas a function needs to be called explicitly.
Feedback B	Correct
Feedback C	Missing point: Constructors do not have return type but functions have to indicate a return type.
Feedback D	Answer not explained properly. Information on function/ constructor calling and their return type missing.


Count of Which feedback would you assign to this answer?
Feedback C
9.1%
Feedback B
9.1%
4
4
21
Feedback A
47.7%
15
Feedback D
34.1%

Results

Feedback: Survey Results

Question	When does C++ create a default constructor?
Reference Answer	If no constructor is provided the compiler provides one by default. If a constructor is defined for a class the compiler does not create a default constructor.
Student Answer	When non are provided
Feedback A	If no constructor is defined then a default constructor is created during compilation.
Feedback B	Missing keywords: during compilation
Feedback C	Correct


Count of Which feedback would you assign to this answer?
Feedback B
20.5%
9
Feedback C
25.0%
11
Feedback A
54.5%
24

Results

Prediction Time

Method	Avg. Prediction Time (sec)
Passage-Based	**0.0344**
Sentence-Based	0.0749
Chunk-Based	41.3039
RDF-Based	5.8424

Table 12: Shows average prediction time (in seconds) per response for each method.