Task 2

Manual Ground Truth

We pick the first question that contains a bug (Question.ID = 1) and all correct explanations to manually formulate a ground truth.

```
Minutes are set to -15; which is less then 0 and it throws illegal arg exc
eption
In the code there is a check that 0 <= minutes < 60 and the minutesOffset
is -15 which does not fall into these prarmeters thus throwing an Exceptio
n
There is a logical check for if minuteOffset is less than 0 or greater tha
n 59 causing it to throw an exception because the value is out of bounds
(negative number)
YES. The issue is on line 279 (as I explained in my first question; of whi
ch I misunderstood that I was only being asked about the specific issue; n
ot generalized issue). On line 279 the variable "minutesOffSet" is paramet
erized to throw an exception if it is < 0 or > 59. Line 279 should read "i
f (minutesOffset < -59 || minutesOffset > 59) {" because now the method ca
n take in the number of minutes as a negative and will allow the method to
properly progress to invoke/call further methods such as those asked about
in the two previous questions.
The variable "minutesOffset" is checked incorrectly by the IF statement on
line 279. Any negative value for "minutesOffset" will throw this exceptio
n; while the documentation states that "minutesOffset" can be negative in
some cases.
This variable contains a value of -15 as set by DateTimeZone.forOffsetHour
sMinutes(-2; -15). Line 279 checks to see if is a valid value; meaning tha
t is between 0 and 59. Since it is not; an exception error is thrown in li
ne 280.
Yes; the variable gets set to -15 through the arguments above. The code sp
ecifically encounters the error on line 279 when it tests if minutesOffset
< 0; (-15) which is the case; so it throws the error on line 280 : Minutes
out of range: with the value provided for that argument -15.
As noted in the comments; valid input for minutes must be in the rage -59
to +59 but on line 279 of the source minutesOffset is checked for < 0. Ins
tead it should be minutesOFfset < -59 . Also noted in comments is that ver
sions before 2.3 minutes had to be zero or positive. "Minutes out of rang
e: + minutesOffset" is our error.
the variable should be defined as "unsigned int" if we expect it to be alw
ays positive
The value of minutes offset does not have valid argument as a result this
method will not be called as and argument exception will be displayed.
yep; they are checking if minutesOffset < 0 to throw an exception; and as
-15 <0; it gets thrown. looks like they updated the comments but not the c
ode. and this is why comments are evil liars that can't be trusted!
The error is stemming from line 279 because the value of -15 for minutesOf
fset is < 0. The line should be     if (minutesOffset < -59 || minutesOffs
et > 59) {
```

Using this data, we wrote our manual ground truth that contains all information needed to fix the bug:

```
'The IF statement in line 279 checks whether minutesOffset is set to a val
ue between 0 and 59. If not, an IllegalArgumentException is thrown. This i
s a bug because the minutesOffset may also be negative. The IF statement s
hould check for the minutesOffset to be between -59 and 59.'
```

Now we choose metrics to evaluate readability and semantic similarity between the manual ground truth and the explanations generated by the LLM.

Readability

Flesch Reading Ease and Automated Readability Index

We choose the Flesch Reading Ease metric and Automated Readability Index, as they are well-known readability metrics that are easy to compute and simple to interpret. The Flesch Reading Ease gives a score between 0 and 100, where a score between 60 and 80 is considered easy to read. The Automated Readability Index gives a score between 1 and 14, where for most readers an ideal score is between 7 and 9.

Semantic Similarity

BLEU

We selected BLEU because of its simplicity and our prior experience with it from the previous miniproject. BLEU is widely recognized as a standard metric for measuring semantic similarity and requires minimal computational effort. As the explanations and our ground truth are both short texts, we change the BLEU weights to not incorporate 3-grams and 4-grams.

Cosine Similarity of Embeddings

A disadvantage of BLEU is that it does not consider the semantic meaning of the explanations which leads to synonyms being considered as differences. Therefore, we also compute the cosine similarity of embeddings of the explanations and the ground truth. These embedding represent the semantics of text in a vector space in a way that similar texts are close to each other in the vector space.

Expectations and Setting a Threshold

It is hard for us to come up with a good threshold for picking the best amount of explanations to merge by using these metrics before seeing the results for our data, as we do not really know what ranges of numbers to expect.

If the summarized explanations get better, we would expect the readability metrics to get worse, as the summaries become more technical and precise. The similarity to the ground truth should increase because the ground truth contains information from

multiple explanations, so you need more explanations to cover all the information. As described before, we would accept a Flesh Reading Ease of 60-80 and an Automated Readability Index of 7-9 for our summary to be considered well readable.

We choose a threshold of 0.70 for the cosine similarity.

Run the experiments

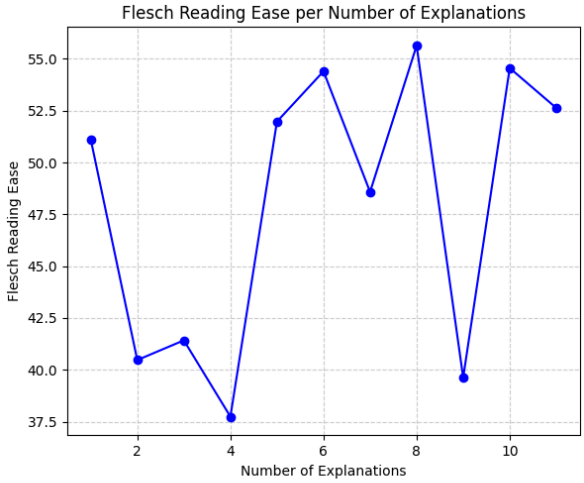
We now sample different amounts of explanations and merge them using the query from the previous mini project. For each sample size we sample 3 times. For each merged sample we calculate the metrics discussed before and store them in a DataFrame.

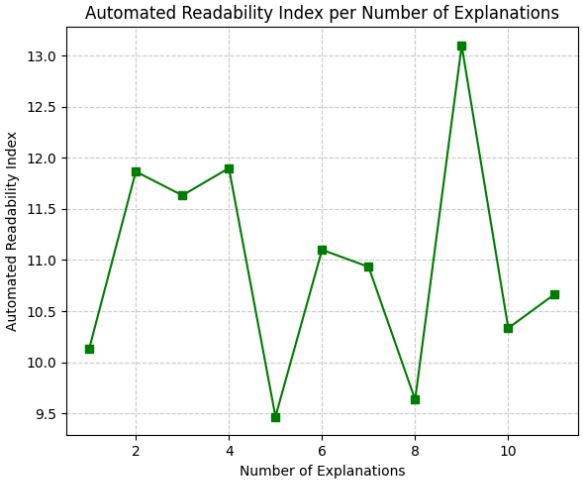
We then plot the different metrics to see how they change with the number of explanations merged.

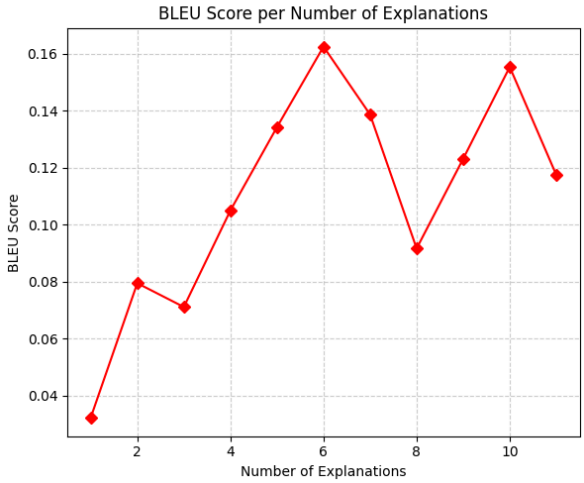
For the Flesh Reading Ease and Automated Readability Index, we cannot really see a trend. The scores are volatile which means that the readability of the summaries does not really change with the number of explanations merged. For the Flesh Reading Ease the scores are all below 60, which means that we would not consider any of the summaries well readable with this score. The Automated Readability Index is also above 9 for all summaries, which is above our threshold, meaning all generated summaries are rather difficult to read.

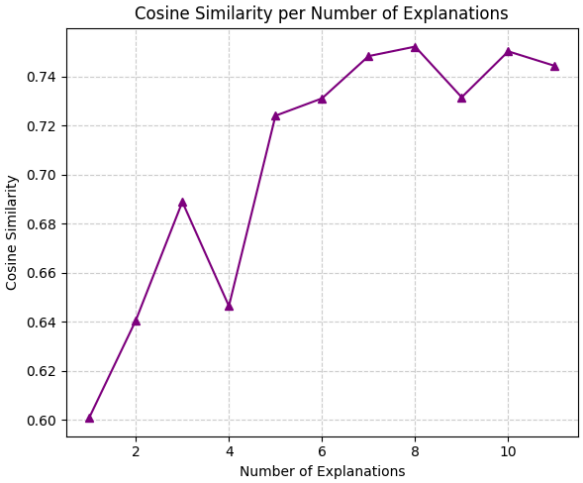
The BLEU score however increases with the number of explanations. Just as the increasing cosine similarity, it shows that the generated summary gets more similar to the ground truth with more explanations merged. We can also see that the cosine similarity dramatically rises after including more than 5 explanations. At this point it is at over 0.72, which is also above our threshold of 0.70. This means that we would need at least 5 explanations to generate a consolidated summary that is similar to the ground truth.

This means that the Flesh Reading Ease and Automated Readability Index do not help us choose how many explanations need to be consolidated in the end. On the other hand the BLEU score and cosine similarity give us a good indicator for the completeness of the summary. This means these scores can be used to decide on the number of explanations we need to merge to get a good summary.


Flesch Reading Ease per Number of Explanations
Flesch Reading Ease
Number of Explanations


Automated Readability Index per Number of Explanations
Automated Readability Index
Number of Explanations


BLEU Score per Number of Explanations
BLEU Score
Number of Explanations


Cosine Similarity per Number of Explanations
Cosine Similarity
Number of Explanations