# Using Large Language Models to Provide Formative Feedback in Intelligent Textbooks

Wesley Morris[1](✉), Scott Crossley[1], Langdon Holmes[1], Chaohua Ou[2], Danielle McNamara[3], and Mihai Dascalu[4]

[1] Vanderbilt University, Nashville, USA
wesley.g.morris@vanderbilt.edu
[2] Georgia Tech, Atlanta, USA
[3] Arizona State University, Tempe, USA
[4] University Politehnica of Bucharest, Bucharest, Romania

**Abstract.** As intelligent textbooks become more ubiquitous in classrooms and educational settings, the need arises to automatically provide formative feedback to written responses provided by students in response to readings. This study develops models to automatically provide feedback to student summaries written at the end of intelligent textbook sections. The study builds on Botarleanu et al. (2022), who used the Longformer Large Language Model, a transformer Neural Network, to build a summary grading model. Their model explains around 55% of holistic summary score variance when compared to scores assigned by human raters on an analytic rubric. This study uses a principal component analysis to distill scores from the analytic rubric into two principal components – content and wording. When training the models on the summaries and the sources using these principal components, we explained 79% and 66% of the score variance for content and wording, respectively. The developed models are freely available on HuggingFace and will allow formative feedback to users of intelligent textbooks to assess reading comprehension through summarization in real-time. The models can also be used for other summarization applications in learning systems.

**Keywords:** intelligent textbooks · large language models · automated summary scoring · transformers

## 1 Introduction

An essential component of intelligent textbooks is the capacity to provide formative feedback to students in real time regarding text comprehension. Recent developments in Natural Language Processing (NLP) allow more sophisticated feedback approaches based on open-ended assessments like text summarization. For instance, Crossley et al. [1] developed a summarization model to predict ratings of main idea integration in student summaries using lexical diversity features, a word frequency metric, and word2vec semantic similarity scores between summaries and source material. The model explained 53% of the variance in ratings. Botarleanu et al. [2] used large language models (LLMs)

N. Wang et al. (Eds.): AIED 2023, CCIS 1831, pp. 484–489, 2023.
https://doi.org/10.1007/978-3-031-36336-8_75

to predict overall student summarization scores and explained ~55% of score variance. These NLP models show the potential for open-ended assessments of text comprehension through summarization in intelligent textbooks.

The goal of this paper is to introduce more robust LLMs to provide formative assessment for summaries written at the end of chapter sections within an intelligent textbook framework. The models presented in this study are more robust in two ways. First, they are trained on data developed specifically for the model. Second, instead of training them to predict each scale of an analytic rubric for summarization, the analytic scores in a rubric are aggregated into two criteria using a principal component analysis. This study aims to develop summarization models that can be integrated into intelligent textbooks to make them more interactive by providing actionable feedback to students about their summaries to increase comprehension of course material. In doing so, we develop two LLMs to provide formative assessment on summaries: 1) a model based on RoBERTa to predict scores based on the summary itself, and 2) a model using Longformer to predict summary scores while including text from the textbook as context. The research questions that guide this study are

- Are more robust LLMs able to provide more accurate models that can be used to guide student understanding as they read an intelligent textbook?
- Does the inclusion of text from the textbook improve the accuracy of the LLMs?

## 2 Methods

### 2.1 Data

Our summary corpus comprises 4,233 summaries of 101 source texts written by high school, university, and adult writers recruited through Amazon's Mechanical Turk service between 2018 and 2021. Source texts were on a variety of different topics, including the dangers of smoking, computer viruses, and the effect of UV radiation. The sources had a mean word count of 308.5 (SD = 130.49), while the summaries had a mean word count of 75.18 (SD = 50.51). Each summary was scored by expert raters using a 1–4 scaled analytic rubric to score 7 criteria important in understanding the quality of summarizations. The criteria included the main point (i.e. to what extent the summary captured the main idea of the source), details of language beyond the source (i.e. how well all relevant details were included in the summary), paraphrasing (i.e. avoiding plagiarism by paraphrasing the original material), objective language use (i.e. reflecting the point of view of the source), and cohesion (i.e. how well the summary was rationally and logically organized). Inter-rater reliability showed acceptable agreement among raters ($r > .8$ and $\kappa > .7$). A subset of this data set was used in Crossley et al. [1] and Botarleanu et al. [2].

A principal component analysis (PCA) was conducted to assess the potential for dimension reduction for the analytic scores in the rubric. The PCA revealed strong covariance allowing six of the scores to be combined into two principal components. The analytic scales of details, main point, and cohesion were combined into a weighted score designated as Content. The analytic scales for paraphrasing, objective language use, and language beyond the source were combined into a weighted score designated as Wording. The component scores were normalized to a scale from 0 to 1 using min-max normalization and used as outcome variables in our large language models.

## 2.2 Large Language Models

LLMs are neural network architectures for natural language processing which use the principle of self-attention to generate large, pre-trained models which can then be further finetuned for downstream tasks. These pre-trained models are trained on large corpora using masked language modeling, in which the text is tokenized, but some tokens are masked. The task of masked language modeling is to predict the masked tokens based on all the tokens that come before and after the masked token. After many epochs of training on very large corpora, the parameters of the model come to represent a general knowledge of the language domain on which they were trained.

The pretrained model can be further refined in two ways. The primary method of model refinement is through finetuning. The model is trained on the target task using the training data and labels in finetuning. Encoder-only transformers, such as those used in this study, include a special classification token at the beginning of the sequence. As the model processes the language data, the embedding of the classification token comes to represent semantic information about the text as a whole and can be used with the labels supplied in the training data to train a traditional machine learning algorithm. In finetuning, the model's parameters and the machine learning classification head are trained.

The other method, domain adaptation, is used when there is a large amount of unlabeled data but a relatively small amount of labeled data. In this case, the model is trained using a masked language model on language data from the target language domain in order to allow the model greater familiarity with the target domain. After domain adaptation, the resultant model is then finetuned on the labeled data for classification, regression, or other specific tasks.

We used RoBERTa [3] as our initial LLM, which is an encoder-only transformer model pretrained on the English Wikipedia corpus and Bookcorpus. The transformer neural network architecture relies on attention mechanisms in which, at every layer, each token embedding is modified by each other token embedding. As a result, the computational requirements grow quadratically as a function of the input sequence length. In RoBERTa, the length of the input sequence is limited to 512 tokens to ensure computing efficiency. While this length is sufficient for many summaries, it is not long enough to include text from the textbook in the model input.

The Longformer LLM [4] is capable of handling longer input sequences by utilizing sparse attention, in which not all tokens are compared with every other token. Instead, Longformer uses a sliding attention window so that each token only attends to the tokens a certain number of positions to its left and right. Sparse attention mitigates the problem of limited sequence length by reducing the computational complexity of the attention mechanism. Additionally, Longformer utilizes global attention in which certain tokens are attended to by every other token. Combining these two types of sparse attention allows Longformer to increase the max sequence length from 512 tokens to 4,096 tokens while remaining efficient. The Longformer max sequence length allows us to include both the summary and text from the textbook into the input sequence.

We divided the summary corpus into training, validation, and test sets. To help ensure generalizability across source texts and prompts, we selected 15 out of the 101 sources text to comprise the test set only (i.e., these source texts were not used in training or validation) After splitting the data in this way, the training, validation, and test sets comprised 3,285, 703, and 702 summaries respectively. During finetuning, each summary in RoBERTa was tokenized and fed into the model. For Longformer, the summary and the source text for the summary were concatenated using a special separator token and then tokenized together to generate the input sequences. These token sequences were used as input data for their respective models, and the final classification token was used to train a linear regression head. After training, we tested the performance of each model by predicting the Content and Wording scores for the summaries. We evaluate model performance in terms of correlation with the human rater judgments and explained variance ($R^2$).

In addition to the finetuning procedures described above, we also domain-adapted the Longformer pre-trained model on a set of 93,484 summaries written by middle and high-school students. The summaries were collected from six online sources through the Commonlit platform (commonlit.org) [5]. This is a large, unlabeled dataset in the target language domain, and we considered it a reasonable candidate for domain adaptation. After constructing the domain-adapted model, we finetuned it using the same methods described above and evaluated its performance by calculating the correlation between predicted scores and human rater judgments.

## 3 Results

For Content scores, the Longformer model, in which both the summary and the source were included in the input, achieved higher accuracy ($r = .89$, $R^2 = .79$) than the RoBERTa model, which only included the summary ($r = .82$, $R^2 = .67$). For Wording

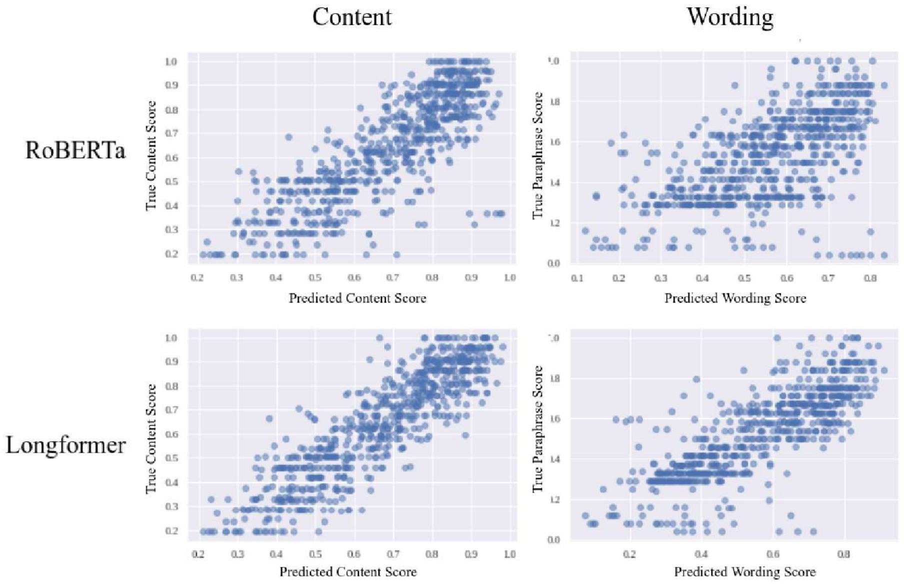

**Fig. 1.** Predicted Scores Plotted Against Actual Scores for the Four Models

scores, the Longformer model outperformed (r = .81, $R^2$ = .66) the RoBERTa model (r = .60, $R^2$ = .36) (see Fig. 1). The domain-adapted Longformer model performed worse than the non-domain adapted model for both content (r = .85, $R^2$ = .72) and wording (r = .78, $R^2$ = .60).

## 4 Discussion

The Longformer models developed in this study outperform RoBERTa models and previous LLM-based automatic summary evaluation models. For instance, previous NLP models have achieved $R^2$ ~ 55% (citations). In addition, finetuning directly on the pretrained model produced better results than finetuning on the domain-adapted model. This may be because the Commonlit dataset, which included many summaries, only included six sources, which did not provide the variance needed for the problem space. The small number of sources may have resulted in catastrophic forgetting [6], where the model overfitted to those sources and forgot part of the parameters set during pretraining.

## 5 Conclusions and Future Work

Although the data used in the training and test sets for this model are not exactly the same as the task of grading summaries of textbook sections, the accuracy rates for Content scores (and likely Wording scores) are strong enough for inclusion into intelligent textbooks to provide students with opportunities for open-ended comprehension assessment and interactive feedback. We plan to integrate this model into a prototype intelligent textbook currently in development. Students will be required to write a short summary at the end of each section before moving on to the next. After passing through a filter ensuring that the summaries are more than fifty words, in English, not plagiarized, and on topic, the summaries will be automatically graded on the two criteria, and students will receive their grades instantly. We hope to demonstrate that providing students with feedback on their summaries along with the revision process leads to a greater understanding of texts within intelligent textbooks.

**Acknowledgements.** This material is based upon work supported by the National Science Foundation under Grant 2112532 and the Institute for Education Science under Grant R305A180261. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

This work was supported by the Ministry of Research, Innovation, and Digitalization, project CloudPrecis, Contract Number 344/390020/ 06.09.2021, MySMIS code: 124812, within POC.

## References

1. Crossley, S.A., Kim, M., Allen, L., McNamara, D.: Automated summarization evaluation (ASE) using natural language processing tools. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 84–95. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_8

2. Botarleanu, R.M., Dascalu, M., Allen, L.K., Crossley, S.A., McNamara, D.S.: Multitask summary scoring with longformers. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) AIED 2022. LNCS, vol. 13355, pp. 756–761. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_79
3. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 [Cs]. http://arxiv.org/abs/1907.1169 (2019)
4. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The Long-Document Transformer arXiv:2004.05150. arXiv. https://doi.org/10.48550/arXiv.2004.05150 (2020)
5. Crossley, S.A., Heintz, A., Choi, J., Batchelor, J., Karimi, M., Malatinszky, A.: The CommonLit ease of readability (CLEAR) corpus. In: Educational Data Mining (2021)
6. Ramasesh, V.V., Lewkowycz, A., Dyer, E.: Effect of scale on catastrophic forgetting in neural. In: International Conference on Learning Representations, September 2021