ARTICLE

Formative Feedback on Student-Authored Summaries in Intelligent Textbooks Using Large Language Models

Wesley Morris[1] · Scott Crossley[1] · Langdon Holmes[1] · Chaohua Ou[2] · Mihai Dascalu[3] · Danielle McNamara[4]

Abstract

As intelligent textbooks become more ubiquitous in classrooms and educational settings, the need to make them more interactive arises. An alternative is to ask students to generate knowledge in response to textbook content and provide feedback about the produced knowledge. This study develops Natural Language Processing models to automatically provide feedback to students about the quality of summaries written at the end of intelligent textbook sections. The study builds on the work of Botarleanu et al. (2022), who used a Longformer Large Language Model (LLM) to develop a summary grading model. Their model explained around 55% of holistic summary score variance as assigned by human raters. This study uses a principal component analysis to distill summary scores from an analytic rubric into two principal components – content and wording. This study uses two encoder-only classification large language models finetuned from Longformer on the summaries and the source texts using these principal components explained 82% and 70% of the score variance for content and wording, respectively. On a dataset of summaries collected on the crowd-sourcing site Prolific, the content model was shown to be robust although the accuracy of the wording model was reduced compared to the training set. 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

---

Extended author information available on the last page of the article

Introduction

Intelligent textbooks have become increasingly popular in recent years as the COVID-19 pandemic pushed many learners into online classes (Seaman & Seaman, 2020) combined with advances in Natural Language Processing (NLP) that have made human-machine interaction more accessible (Brusilovsky et al., 2022; Wang et al., 2021). Intelligent textbooks have many advantages over print textbooks, including the integration of multimedia elements such as video, audio, and hyperlinks. While some studies have demonstrated no significant difference in learning between digital and print textbooks (Rockinson-Szapkiw et al., 2013), a more recent and comprehensive meta-analysis of 26 studies reported that interactive features such as those in intelligent textbooks improve reading performance with a moderate effect size across multiple knowledge domains (Clinton-Lisell et al., 2021). Additionally, college students prefer the lower cost and ease of use of intelligent textbooks (Ji et al., 2014; Chulkov & VanAlstine, 2013).

A textbook should be more than a static web-based version of a traditional paper textbook to be considered intelligent. Instead, an intelligent textbook should be interactive and adapt to the individual user's needs. Various forms of artificial intelligence techniques can be employed to accomplish the goal of interactivity, including the use of Transformer-based Large Language Models (LLMs). LLMs have been used for a variety of purposes, including summary generation (Khandelwal et al., 2019), question answering (Shao et al., 2019), text classification (Wolf et al., 2020), validation of peer-assigned scores in massive open online courses (Morris et al., 2023b), and question generation tasks (Lopez et al., 2021).

Previous research indicates that writing about textbook content in tasks such as summarization can increase learning outcomes in various content domains (Graham et al., 2020; Silva & Limongi, 2019). However, scoring summaries is time-intensive for instructors, paving the way for automatic approaches to summary scoring (Lagakis & Demetriadis, 2021). This study is part of a larger project called Intelligent Textbook for Enhanced Language Learning (iTELL) to develop a computational framework that converts static, web-based textbooks into interactive, intelligent textbooks. iTELL converts any type of machine-readable text into an interactive web-app, and students can write summaries directly in the application. These summaries can be scored automatically by LLMs specifically trained to generate scores which inform qualitative feedback to students related to content and wording. Students can use the feedback from these models in different ways, including to reflect on and guide their learning, identify and correct misconceptions, review missed topics, and prepare for upcoming materials. As such, while iTELL is being developed in the context of intelligent textbooks, it also has multiple applications outside of that context.

The goal of this current study is to report on the automated summary evaluation models integrated into iTELL and discuss how feedback is automatically provided to the users of intelligent textbooks. Based on pretrained encoder-only LLMs, our models can help students develop their knowledge while providing important information about reading comprehension to teachers and material developers (Phillips Galloway, & Uccelli, 2019). Specifically, we provide an overview of LLMs that provide a formative assessment of summaries. These include models based on RoBERTa

(Liu et al., 2019) that predict scores based only on the summary due to constraints on the maximum sequence length, and models using Longformer (Beltagy et al., 2020), which are capable of an increased max sequence length, allowing them to predict summary scores while including considerably more text from the textbook as context. Because iTELL is designed to be domain agnostic, the models must provide accurate feedback to users regardless of the textbook topic. To that end, the models were trained on a dataset comprising source texts on a wide range of informative topics. The research questions that guide this study are the following:

1. To what extent does the inclusion of source text from the textbook improve the accuracy of the LLMs in automatically scoring summaries?
2. Does unsupervised pretraining on a large dataset of texts in the target language domain improve performance on the LLMs?
3. How well do automated summary evaluation LLMs perform when they are used outside of the context of the dataset and labels considered during their training?

Related Work

Intelligent Textbooks

The earliest intelligent textbooks were designed in the 1990s using the principles of knowledge engineering, in which the textbook would be designed and produced by domain experts (Brusilovsky et al., 2022). Early work in designing intelligent textbooks included the development of hypertext (Bareiss & Osgood, 1993), which allowed students to navigate the book efficiently (Brusilovsky & Pesin, 1998). One of the first web-based interactive textbooks included ELM-ART, an intelligent, interactive textbook to teach programming introduced in 1996 (Weber & Brusilovsky, 2016).

The development of intelligent textbooks has increased in the past decade as computational tools become more sophisticated and accessible (Sosnovsky et al., 2023). More recent research has included mining student behaviors in intelligent textbooks and using those data to provide an individualized learning experience. For instance, Lan and Baraniuk (2016) developed a muti-armed bandit algorithm that uses results from previous assessments to identify and recommend pedagogical activities optimally individualized for each student. Learner behavior such as failure to correctly answer comprehension questions can also be used to adaptively modify the content of textbooks, thus recommending materials to remediate comprehension gaps (Thaker et al., 2020). Other research has shown that student behaviors in intelligent textbooks, such as annotation and highlighting, can predict student success in the course (Winchell et al., 2018) and that concept or keyphrase extraction using annotation by trained experts can be used as training data for machine learning algorithms (Wang et al., 2021).

Researchers have also used NLP techniques to construct semantic maps of textbooks which can be used to integrate the textbook with resources available on the

web (Alpizar-Chacon & Sosnovsky, 2021; Labutov et al., 2017). In addition, part-of-speech taggers have been employed to develop question generation tools that automatically generate and embed comprehension questions into intelligent textbooks (Kumar et al., 2015). Generative Transformer neural networks such as GPT-2 have also been used to develop language generation tools to provide information to students within intelligent textbooks (Yarbro & Olney, 2021).

Summarization and Reading Comprehension

Text summarization is a valuable tool to build and assess student knowledge (Head et al., 1989) that has become more common in educational applications (Graham & Harris, 2015), especially in readability assessments like those found in intelligent textbooks (Phillips Galloway & Uccelli, 2019). Writing tasks like summarizations also help students build and consolidate their knowledge about reading materials in addition to their effectiveness in reading comprehension assessment. A meta-analysis of 56 experiments on the effect of writing on learning by Graham et al. (2020) found an average weighted effect size of Hedges's $g=0.3$ ($p<.005$) between pre and post-tests for students who used writing to learn from texts. This effect size held, regardless of whether the knowledge domain was science, social studies, or mathematics. These results may be due to the increased cognitive demands of writing, a process in which the learner must actively reconstruct knowledge from the text (Nelson & King, 2022). For instance, Galbraith and Baaijen (2018) contend that writing consists of two separate domains, one in which knowledge from the text is retrieved and manipulated and one in which the author actively uses their understanding of the world to construct text. Concurrent research by Silva and Limongi (2019) indicates that the practice of summary writing may help to consolidate the knowledge gained from reading into long-term memory. Despite the effectiveness of summarization in education and assessment, providing feedback to learners about the quality of summaries is time-consuming for educators (Gamage et al., 2021), making summarization challenging to scale.

Automated Summary Evaluation

An essential component of intelligent textbooks is the capacity to provide formative feedback to students about their comprehension, and real-time feedback provided by AI is effective at improving reading comprehension (Chen et al., 2021; Kim et al., 2020). Before the development of deep learning approaches to NLP, automated summary evaluation (ASE) was primarily performed by comparing the summary being tested with a professionally produced reference summary. Algorithms such as ROUGE (Lin & Hovy, 2003), closely related to BLEU (Papineni et al., 2001), were used to provide summary scores based on word and phrase co-occurrence between the test summary and the reference summary. While ROUGE is correlated with human judgments of summary quality and is actively used in the training of summarization algorithms (Ganesan, 2018; Scialom et al., 2019), it is biased toward surface-level lexical features, a limitation which can be addressed using more advanced NLP features including word embedding approaches (Ng & Abrecht, 2015). More impor-

tantly, ROUGE and BLEU approaches require the use of reference summaries created by a human expert, which are resource intensive and impractical in the context of intelligent textbooks like those generated with the iTELL platform.

Recent developments in NLP allow more sophisticated feedback approaches for open-ended reading assessments like text summarization. For instance, Crossley et al. (2019) 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 the corresponding source material. The model explained 53% of the variance in ratings. Martínez-Huertas et al. (2019) used latent semantic analysis to embed summaries into semantic vector spaces where the rubric scores could be extracted. Their method achieved a Pearson's correlation with scores from expert raters between 0.78 and 0.81. With the rise of LLMs, new methods of automated summary evaluation have been evaluated. For instance, Botarleanu et al. (2022) used LLMs to predict overall student summarization scores derived from an analytic rubric, explaining ∼55% of score variance.

Current Study

The NLP models discussed above show the potential for open-ended assessments of text comprehension through summarization in intelligent textbooks. To fulfill their purpose in the context of the iTELL framework, the models should accurately score summaries of source texts on any topic. The current study expands on the work of Crossley et al. (2019); Botarleanu et al. (2022); Morris et al. (2023a) which used a similar dataset of summaries on sources covering a variety of topics. First, instead of using the raw scores from an analytic rubric, we consolidated the scores into two principal components and used those as labeled data in model training. Second, we assessed the extent to which domain adaptation of the models improves scoring accuracy.

Methods

Four different datasets were used in this study, listed for reference in Table 1. During training, we used a training dataset used for finetuning the models and a dataset from Commonlit used for domain adaptation. We also used two datasets for post-hoc tests of validity and generalizability - a dataset of professional summaries found in a textbook available on OpenStax, and a dataset of summaries written by participants recruited through the Prolific crowdsourcing platform. Each of these datasets will be discussed in more detail in the subsequent sections.

Table 1 List of datasets

Name	Use	Sources N	Summaries N
Training	Finetuning	101	690
Commonlit	Domain adaptation	6	93,484
Textbook	Post-hoc testing	94	94
Prolific	Post-hoc testing	4	113

Data

Our summary training corpus comprises 4,690 summaries written by high school, university, and adult writers collected by three higher education institutions between 2018 and 2021, corresponding to 101 source texts. Crossley et al. (2019) and Botarleanu et al. (2022) used a subset of this data. The training corpus consolidates several different data sources, including summaries written by workers on Amazon's Mechanical Turk service (Li et al., 2018), summaries written by undergraduate college students, and summaries written by high school students. Source texts consider a variety of topics, such as the effect of UV radiation, diabetes, computer viruses, red blood cells, and the dangers of smoking. The sources had a mean word count of 308.5 (SD = 130.49) and thus may be shorter than sections in an intelligent textbook. However, the topics were academic and therefore, similar to topics from intelligent textbooks. Each source had an average of 46.44 summaries written for it (SD = 62.12), with a maximum of 258 summaries and a minimum of 10. The summaries had an average of 75.18 words (SD = 50.51).

The formatting of the summary dataset was not uniform. For example, in some cases, source documents included several articles, and only one was summarized by any individual writer. We checked each source to ensure each summary was paired with only one properly formatted source. This process of cleaning and normalization is the first and often most labor-intensive step toward training a machine learning tool (Shorten et al., 2021). However, cleaned datasets often significantly impact the final accuracy of the model (Chollet, 2018).

Summary Scoring

Regardless of source, all summaries in the training dataset were scored according to the same procedure. Two expert raters scored each summary using a 0–4 scaled analytic rubric to score 7 criteria important in understanding the quality of summarizations. The criteria included main point/gist (whether the summary captured the gist of the source), details (to what extent the summary included all relevant information from the source), language beyond the source (grammar and syntax), paraphrasing/wording (avoiding plagiarism and direct copying from the source), objective language (accurately reflecting the view of the source), and cohesion (to what extend the summary was clearly and rationally organized), and text length. Appendix A contains the rubric used by raters in their assessment of the summaries.

Raters were initially normed on a small set of summaries not included in the final set of summaries. As such, raters talked through ~ 20 summaries until they were comfortable with the rubric. They then scored ~ 40 summaries until reaching an acceptable level of inter-rater reliability ($r > .699$). Afterwards, raters scored summaries independently by source. Within the prompt, the scored summaries were randomized to reduce the potential for ordering effects. Also, raters could adjudicate any score differences greater than 1. Raters did this by talking through the summary with each other and then deciding whether they wanted to modify their ratings based on the discussion. If an agreement was not reached, the individual scores were not altered, but their average was used in subsequent predictions. Final inter-rater reliability was

acceptable ($r > .800$ and $\kappa > 0.700$). Average scores between the raters were calculated for each essay and used for the data analysis.

Dimensionality Reduction

We were faced with several choices because the scoring rubric consisted of seven criteria,. These included training a single multitask model to predict all seven scores at once at the cost of accuracy and training seven different models at the expense of increased compute requirements. A third choice, and the one we selected, was to conduct a principal component analysis (PCA) to assess the potential to reduce the dimensionality of the seven analytic scores in the rubric into a smaller number of constructs.

Before conducting the PCA, the human scores were standardized using z-score normalization. An initial PCA with all possible factors ($n = 7$) indicated 2 components that reported eigenvalues over 0.70 (Jolliffe's criteria). A Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy indicated that no variables need to be removed (i.e., all KMO values were above 0.5), and the overall KMO score $= 0.87$ showed a "meritorious" sample (Kaiser, 1974). The PCA reported a Bartlett's test of sphericity, $\chi^2 (4690) = 11{,}513.99$, $p < .001$, indicating that correlations between the analytic scores were sufficiently large for the PCA. Within the components, there was a break in the cumulative variance explained between the second and the third component. Considering this break, we decided on a 2-component solution when developing the PCA. These 2 components explained approximately 73% of the shared variance in the data from the initial PCA. Figure 1 displays a scree plot showing the eigenvalue,
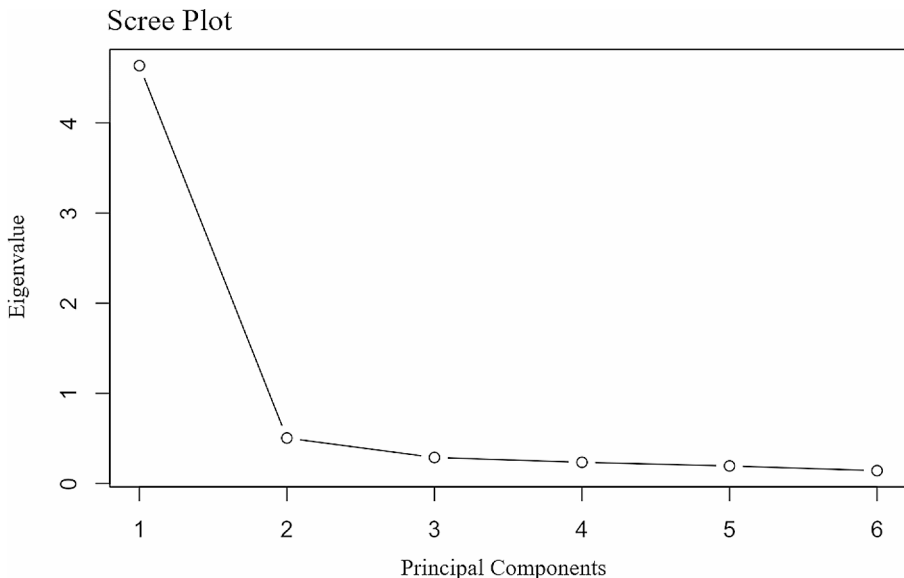

Fig. 1 Scree plot of eigenvalues against number of principal components

or absolute value of the explained variance, plotted against the number of principal components.

The first component was related to Content (i.e., Component 1), and the analytic scales of details, main point, objective language use, and cohesion were combined into a weighted score. The analytic scales for paraphrasing and language beyond the source were combined into a weighted score designated as Wording (i.e., Component 2). Text length did not load into the first components and was removed. Table 2 displays descriptive statistics for all scores, including the principal components. The component scores were z-score normalized and rescaled such that zero represents the mean for each principal component, and one unit represents one standard deviation. These transformed scores were used as outcome variables in our large language models.

Summary Scoring Model

We used two pretrained LLMs to develop summary scoring models. The Transformer (Vaswani et al. 2017) is a neural network architecture that relies on the self-attention mechanism and can be trained in parallel in contrast to previous recurrent neural architectures. LLMs are Transformer-based models pretrained on a large corpus of text that can be further finetuned for downstream tasks. The LLMs used in this study were trained 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 tokens. Because of the cost and time associated with developing an LLM, only a few models are produced and are freely available.

Pretrained LLM models can be 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 with corresponding labels. Unlike generative models such as ChatGPT (Abdullah et al., 2022) or LLaMa (Touvron et al., 2023), which are trained to predict the next token in a series given the preceding tokens, this study uses embedding or encoder-only models which are commonly used in regression or classification tasks. Encoder-only models 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. A classification head, which could be a linear layer or a traditional machine learning

Table 2 Descriptive statistics for summary scores in the training dataset	Language domain	N	Mean	SD	Min	Max
	Main Point	4,690	3.05	0.80	0.5	4
	Details	4,690	2.79	0.84	0	4
	Cohesion	4,690	2.97	0.79	0	4
	Objective Language	4,690	2.79	0.73	0	4
	Paraphrasing	4,690	2.23	0.91	0	4
	Language Beyond the Source	4,690	2.26	0.70	0	4
	Content PCA	4,690	8.00	2.04	0.76	10.96
	Wording PCA	4,690	3.51	1.25	0	6.28

algorithm, uses the classification token as input to make predictions about the class of the text. During finetuning, the parameters of the model, as well as the classification head, are adjusted.

A secondary method is domain adaptation through unsupervised pretraining (Tunstall, von Werra & Wolf, 2022). Unsupervised pretraining may be 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 masked language modeling on language data from the target language domain to allow the model greater familiarity with the target domain. For example, Beltagy, Lo, and Cohen (2019) used unsupervised pretraining on a scientific corpus to improve the accuracy of the BERT model in that domain. After domain adaptation, the resultant model is finetuned on the labeled data for classification or regression specific tasks.

In this study, we considered two LLMs: the RoBERTa base model (Liu et al., 2019) and the Longformer base LLM (Beltagy et al., 2020). RoBERTa is an encoder-only Transformer model pretrained on the English Wikipedia corpus and Bookcorpus. The Transformer neural 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 computational 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 (Beltagy et al., 2020) can handle 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. In addition to the sliding attention window, Longformer also utilizes global attention in which specific tokens attend to every other token. Since attention is bidirectional, all tokens will attend to global tokens as well. Combining these two types of attention enables Longformer to increase the max sequence length from 512 tokens to 4,096 tokens while remaining efficient. The Longformer max sequence length allowed us to include both the summary and source texts from the textbook in the input sequence. By default, Longformer places global attention only on the classification token at the beginning of the sequence and uses a 512-token sliding attention window which moves across the rest of the sequence. Because the summary is more salient to the score in the task of summary evaluation, however, we hypothesized that it would be beneficial to use global attention for the entire summary. This allows tokens in the summary to attend to every token in the source text and vice-versa. We chose to include the entire summary in global attention and shorten the sliding window to 256 tokens to conserve compute. To the best of our knowledge, this approach has not been used in automatic summary evaluation with Longformers.

We divided the scored summary corpus into training, validation, and test sets. We selected 15 out of the 101 sources text to comprise the test set only to ensure generalizability across source texts and prompts (i.e., these source texts were not used in training or validation). After splitting the data, the training, validation, and test

sets comprised 3,285, 703, and 702 summaries, respectively. Each summary from the training set was tokenized and fed to the RoBERTa model during finetuning. For Longformer, the summary and the source text for the summary were concatenated using a specific separator token (specifically, "< \s>") 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. We trained each model for six epochs with a batch size of 8 and a learning rate of 3e-05, retaining the best model. We used mean squared error as the evaluation metric during the training process. After training, we tested each model's performance by predicting the summary Content and Wording scores. We evaluated 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 and RoBERTa pretrained models on a different dataset of 93,484 summaries written by middle and high-school students. The summaries were collected from six sources available online through the Commonlit platform. This is the largest, unlabeled dataset in the target language domain to the best of our knowledge, and we considered it a reasonable candidate for domain adaptation, although the very small number of source texts ($N=6$) meant that, although the models were training on a large set of student summaries, they were training on only a very small set of source texts. We used a masked language modeling task to domain adapt the models for eight epochs with a learning rate of 2e-5. After constructing the domain-adapted models, we finetuned them using the same methods described above and evaluated their performance by calculating the correlation between predicted scores and human rater judgments.

Post-Hoc Analysis

In addition to training and testing the LLMs on the training dataset, we also tested the LLMs on a dataset of summaries written by experts and a dataset of summaries written by participants recruited from the Prolific crowdsourcing website on content within an intelligent textbook. We chose the 2nd edition textbook for Macroeconomics, freely available on the Openstax website (https://openstax.org/details/books/principles-macroeconomics-2e). This text consists of 94 sections divided into 21 chapters. Each section includes a professional summary ($N=94$). In addition to the material in the sections, the textbook includes pages for key terms, concepts, and review questions for each section.

We used the best-performing LLM to predict scores for the section summaries provided in the textbook. We generated two sets of section text and section summaries, one in which the summaries are matched to the appropriate section and one in which the summaries are randomized so that they are not matched with the section they summarize. We predicted the scores for content and wording for the summaries of each of the 94 sections in the textbook in both the matched and unmatched datasets. If the model is accurate, the summaries in the matched group would score higher than those in the unmatched group.

Table 3 Inter-rater reliability statistics for out-of-domain dataset

PCA	Criterion	QWK
Content	Organization	0.485
	Main Points	0.567
	Details	0.611
Wording	Voice	0.320
	Language	0.531
	Wording	0.730
	Total	**0.576**

Table 4 Results from the roberta and longformer models

	Content		Wording	
	r	R^2	r	R^2
RoBERTa (pretrained)	0.82	0.67	0.64	0.36
RoBERTa (domain adapted)	0.83	0.69	0.65	0.42
Longformer (pretrained)	0.91	0.82	0.87	0.70
Longformer (domain adapted)	0.85	0.72	0.78	0.60

In addition to testing the models on professional summaries in the textbook, we also used the crowdsourcing site Prolific to recruit 60 participants to write 113 short summaries of sections from the Macroeconomics Openstax textbook. These summaries were scored by two expert raters using the same rubric from the original summary dataset. Despite extensive efforts of rater norming, inter-rater reliability was lower than in the original dataset (QWK=0.576). Additionally, quadratic weighted kappa values showed a large amount of variability between criteria, as seen in Table 3. Given that reliability was still acceptable in most cases, we generated principal components for Content and Wording for each Macroeconomics summary. Finally, we evaluated the strongest LLM on the Prolific test set by comparing the LLM predicted scores to the PCA scores derived from human scores.

Results

Comparing Models That Include the Source Text to Models That are Naive to the Source

The results of comparing predicted scores in the held-out test set of the training dataset to the actual scores assigned by expert human raters are presented in Table 4. For Content scores, the Longformer model, in which both the summary and the source were included in the input, achieved higher accuracy than the RoBERTa model that considered only the summary (explaining 82% versus 67% of the variance, respectively). For Wording scores, the Longformer model outperformed the RoBERTa model (explaining 70% versus 41% of the variance, respectively). Scatterplots for the results are presented in Fig. 2.
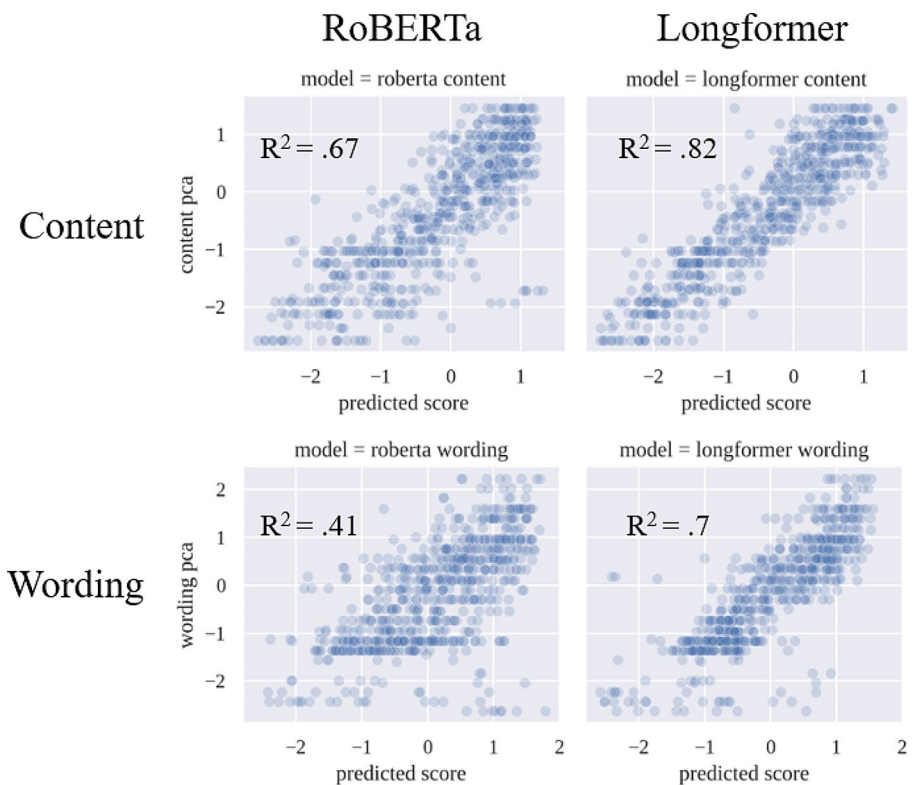
Fig. 2 Predicted scores plotted against actual scores for the four models

Domain Adaptation Through Unsupervised Pretraining

Training from a domain adapted model improved performance slightly in the Roberta model compared to finetuning from the pretrained model. However, the domain adapted Longformer model performed worse than the non-domain adapted version for both content and wording (explaining 72% and 60% of the variance, respectively). Even with reduced performance compared to the base model, the domain adapted Longformer model still performed better than either of the two RoBERTa models with no access to the source text.

Post-hoc Tests on the Prototype Intelligent Textbook

While our non-domain adapted Longformer model performed well on the training data, we further tested it using the section summaries written by the textbook authors found in the Macroeconomics textbook. We did this by using the summaries and their matching sections to predict quality scores for content and wording. To create a comparison group, we also ran the model on the summaries paired with unmatched source texts. The results are illustrated in Fig. 3. The summaries matched to the correct sources have scored higher than average in wording and content. The summaries
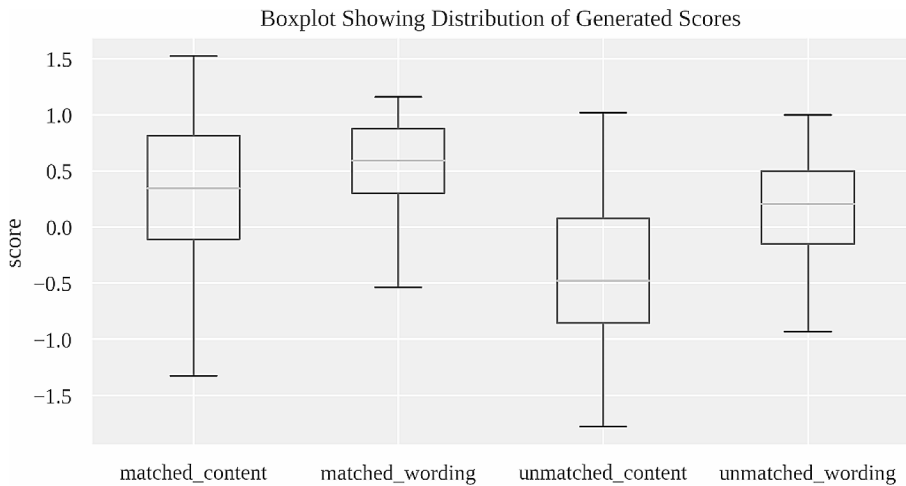
Fig. 3 Boxplot showing distributions of generated scores for professional summaries

Fig. 4 Scatterplot showing correlations between predicted and human scores on prolific dataset

paired with unmatched sources have scored below average in Content and slightly above average in Wording. The differences between matched and unmatched content were statistically significant ($p<.001$) in both domains with large effect sizes, but Content reported a greater effect size ($d=1.56$) than Wording ($d=0.938$).

The accuracy of the Longformer model on the dataset of summaries written by participants recruited through Prolific was lower than the accuracy on the original dataset. Pearson's product-moment correlations showed a strong correlation between predictions and human scores for Content $r(123)=0.70$, $p<.001$. for Content. However, the Wording model did a poorer job at predicting human scores $r(123)=0.29$, $p=.001$. Figure 4 displays scatterplots showing the correlations between the predicted and human scores in this dataset.

Discussion

This paper introduces robust LLMs integrated within iTELL to provide formative assessment for summaries written at the end of chapter sections in intelligent textbooks. These summarization models can assess reading comprehension for students working within the intelligent textbook, provide feedback to students to help them better understand their comprehension of material, and deliver an overview to teachers about how well students understand the material provided. The summarization models presented in this study are more robust than previously reported models, likely because of the training data provided and because we used a principal component analysis on analytic scores from a summarization rubric to aggregate scores into two summarization criteria: content and wording.

Our gains in performance over those reported in previous studies, especially Crossley et al. (2019) and Botarleanu et al. (2022), are likely the result of using cleaned training data and training the model on principal component scores that characterize the summaries in a more condensed representation. The top-performing Longformer models developed in this study achieved R^2 values of 0.82 and 0.70 when predicting human ratings, outperforming RoBERTa models and previous LLM-based automatic summary evaluation models, including the Transformer models by Botarleanu et al. (2022) which achieved $R^2 \sim 55\%$, and the more semantic approaches used by Crossley et al. (2019) who reported an R^2 of 0.53.

In answer to the first research question regarding whether there is a substantial difference in accuracy when a model is provided with both the summary and the source, a model trained to take both the source and the summary performed better than one which only had access to the summary. The increased max sequence length provided by Longformer's sparse attention allowed us to input both the summary and the source divided by a separator token, which led to increased accuracy compared to RoBERTa, which only had access to the summary. The differences are most apparent in the case of Wording, where the model based on the pretrained Longformer reported almost double the R^2 value relative to the RoBERTa model.

The answer to the second research question, whether domain adaptation can improve the accuracy of the summary scoring models, reported mixed results. The RoBERTa model benefited from domain adaptation, but it was the weaker of the two models tested. In contrast, finetuning directly on the pretrained Longformer model produced better results than finetuning on the domain-adapted model. This may be because the Commonlit dataset with many summaries only included six sources, which did not provide the language variation needed for the problem space. Instead of generating models that generalize and provide scores for summaries of any source, the domain adaptation step may have created models that are specifically adapted to the six sources in the Commonlit dataset. This hypothesis is supported by the fact that domain adaptation helped somewhat in the case of the RoBERTa models in which the source was omitted from the input. In the case of the Longformer model, the small number of sources may have resulted in catastrophic forgetting (Ramasesh et al., 2021), where the model overfitted to those sources and forgot some of the parameters from its pretraining.

In answer to the third research question, the analysis of the Longformer LLM on expert summaries found in a Macroeconomics textbook provided some evidence of concurrent validity for the developed model. When the model was tested on summaries matched with the correct source, the model outputted scores above the mean on average for both Content and Wording. However, when tested on summaries matched with incorrect sources, the Content score was nearly half a standard deviation lower than the mean on average. In contrast, the Wording score remained above the mean (although significantly lower than the score when the summaries and sources were correctly matched). These results make sense because Content scores should more strongly differ between matched and unmatched source texts. However, Wording measures include features related to paraphrasing and language beyond the source, which would be higher in expert summaries, but only partially reliant on the source text. The results from the Macroeconomics summaries provide evidence that the models discriminate between matched and unmatched summaries. Summaries are scored better when paired with the correct source, especially in terms of Content.

The test on summaries solicited from participants on Prolific had mixed results. The Content model performed well on these summaries. By contrast, although the Wording model predictions were better than chance, it did not accurately predict human scores with fidelity. This may be a result of the difficulty in aligning the raters to the original scoring procedure. Although we had access to the rubric provided in Appendix A, we did not have access to the rater training procedure. As a result, our raters were not capable of attaining sufficient agreement with each other and presumably were also not aligned with the original set of raters.

Application

The summary evaluation models developed here were integrated into the iTell framework to assess reading comprehension via end-of-section summarization. The purpose of the summaries is to make the textbooks more interactive and provide students with opportunities to produce knowledge and test understanding, while having access to timely personalized feedback. Within iTELL, the summary scoring models have been combined with several other features to ensure the accuracy of the feedback provided and to ensure that good-faith efforts are made by students. After students produce a summary (students cannot cut and paste), and before that summary is passed to the scoring models, the summaries go through a filter component. The filter ensures that the summary is between 50 and 200 words, and the summary is passed through a semantic similarity measure using Doc2Vec (Le & Mikolov, 2014) that assesses whether the summary is on topic. Additionally, summaries that heavily borrow from the source text or contain offensive language are rejected without being analyzed by the LLMs. Source borrowing scores are based on prevalence of overlapping n-grams between the summary and the source (Broder, 1998) while offensive language ratings are based on occurrences of offensive words or phrases from an offensive word list (Inflianskas, 2019). These filters help ensure that only effortful summaries are passed to the models and help reduce the computational load required by the models.

Summaries are then run through the LLMs, which are used to develop formative feedback. Figure 5 displays a screenshot with examples of written feedback that students may receive from iTELL for high and low-quality summaries on Content, Wording, source borrowing, and topic similarity. Although numerical scores are calculated for each of these criteria on the back end, the user receives written qualitative feedback. If the scores are below a certain threshold, the user will be encouraged to revisit the section and asked to rewrite their summary before moving on to the next section. In addition, key phrases are identified within the source text using KeyBART (Inflianskas, 2019), an LLM trained to generate keyphrases. Using KeyBART, students are provided with a list of key phrases from the source text not present in their summary, and are then directed to specific paragraphs or subsections they may not have included in their summaries to help with revision. These feedback mechanisms provide the learners with actionable formative feedback beyond the output of the LLMs. The feedback provided by the summary evaluation tools can help provide insight into the students' reading comprehension skills, assist students in reflecting, summarizing, and articulating what they learned, provide class-level and individual
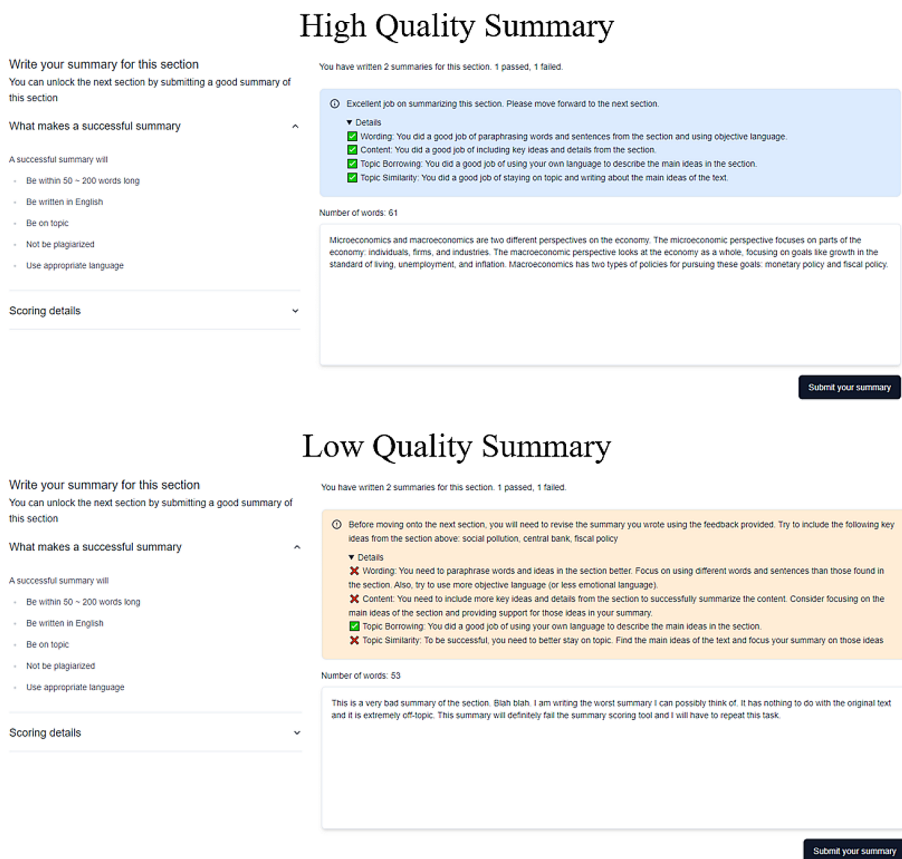

Fig. 5 Screenshots from iTELL displaying feedback from low and high-quality summaries

student metrics on textbook comprehension, and help developers redesign the curriculum and the textbook itself.

Conclusion

In this study, we used RoBERTa and Longformer pretrained Transformers to finetune four large language models to score student-written section summaries automatically. Although the source texts summarized in the dataset used to train the model are likely shorter than intelligent textbook sections, the topics were similar (i.e., they were academic). The accuracy and post-hoc validation scores for the Content models were strong enough for inclusion into iTELL, especially in the case of the models finetuned from the Longformer pretrained model. The Wording model showed promising results but more validation needs to be done using real-world data from textbooks. The summarization models incorporated into iTell can provide students with opportunities for open-ended comprehension assessment and interactive feedback within intelligent textbooks. Additionally, the models are freely available on HuggingFace[1,2], allowing access to learning platforms, researchers, and textbooks developed outside the iTELL framework.

Although the models are strongly predictive, they have limitations. First, while the summaries found in the training data are broadly similar to the target task, and post-hoc tests provided evidence of concurrent validity, more testing in target domains is necessary to ensure that the model accuracy reported in this study will transfer to the task of scoring summaries within a variety of different academic topics. This is particularly the case with the Wording model, since post-hoc tests showed lower generalizability outside of the training data. Additionally, the models need to be tested on intelligent textbook users to ensure that the feedback provided by the models leads to increase learning. Another limitation involves the interpretability of the LLM's output for teachers and learners. The two numerical scores provided by the models indicate summary quality. Nevertheless, the feedback provided to users should explain at a granular level the components of the summaries that lead to the scores and provide actionable suggestions for improvement. Future work should focus on explainable Artificial Intelligence methods to better understand the decisions within the LLMs that lead to the scores.

---

[1] https://huggingface.co/tiedaar/longformer-content-global

[2] https://huggingface.co/tiedaar/longformer-wording-global

Appendix A – Scoring Rubric

Score	Main points/Gist	Details	Cohesion	Objective language	Wording/ Paraphrasing	Language beyond source text	Summary length
1	Main idea is not linked to central topic	Statements are not related to the passage	Ideas are randomly presented and do not link to each other	The language used is not objective.	Summary shows a heavy reliance on verbatim copying of source language.	Summary shows a very basic understanding of lexical and syntactic structures.	Much shorter or longer than expected
2	Main idea is linked to central topic but there is no topic sentence to bring ideas together	Some key information from the passage is included, but important ideas are missing	Some ideas link to each other	Some of the language used is objective	Summary shows some use of original wording, but there are examples of verbatim or near-copy of source language.	Summary shows an understanding of lexical and syntactic structures.	Shorter or longer than expected
3	Main idea is linked to central topic and there is a topic sentence that states some aspect of the content	Most key information from the passage is included, but some ideas may be irrelevant or inaccurate	Most ideas are logically presented	Most of the language used is objective	Summary shows evidence of appropriate levels of paraphrasing.	Summary shows an appropriate range of lexical and syntactic structures	A bit shorter or longer than expected
4	Main idea is linked to central topic and has a topic sentence that states the main idea.	All key information in the passage is included without irrelevant ideas.	All ideas are logically presented	All of the language used is objective	Summary shows substantial evidence of appropriate paraphrasing use.	Summary shows an excellent range of lexical and syntactic structures.	Appropriate length.

Based on Taylor (2013), Westley, Culatta, Lawrence, & Hall-Kenyon (2010)

Declarations

References

Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1–8). Ieee. https://doi.org/10.1109/SNAMS58071.2022.10062688

Alpizar-Chacon, I., & Sosnovsky, S. (2021). Knowledge models from PDF textbooks. *New Review of Hypermedia and Multimedia*, *27*(1–2), 128–176. https://doi.org/10.1080/13614568.2021.1889692.

Bareiss, R., & Osgood, R. (1993). Applying AI models to the design of exploratory hypermedia systems. *Proceedings of the Fifth ACM Conference on Hypertext - HYPERTEXT '93*, 94–105. https://doi.org/10.1145/168750.168790.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer*. https://doi.org/10.48550/ARXIV.2004.05150.

Botarleanu, R. M., Dascalu, M., Allen, L. K., Crossley, S. A., & McNamara, D. S. (2022). Multitask Summary Scoring with Longformers. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (Vol. 13355, pp. 756–761). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_79.

Broder, A. Z. (1998). On the resemblance and containment of documents. *Proceedings Compression and Complexity of SEQUENCES 1997 (Cat no 97TB100171)*, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900.

Brusilovsky, P., Sosnovsky, S., & Thaker, K. (2022). The return of intelligent textbooks. *AI Magazine*, *43*(3), 337–340. https://doi.org/10.1002/aaai.12061.

Brusilovsky, P., & Pesin, L. (1998). Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-Tutor. *Journal of computing and Information Technology,* 6(1), 27–38. https://hrcak.srce.hr/file/221190

Chen, C. M., Chen, L. C., & Horng, W. J. (2021). A collaborative reading annotation system with formative assessment and feedback mechanisms to promote digital reading performance. *Interactive Learning Environments*, *29*(5), 848–865. https://doi.org/10.1080/10494820.2019.1636091.

Chollet, F. (2018). *Deep learning with Python*. Manning Publications Co.

Chulkov, D. V., & VanAlstine, J. (2013). College student choice among electronic and printed textbook options. *Journal of Education for Business*, *88*(4), 216–222.

Clinton-Lisell, V., Seipel, B., Gilpin, S., & Litzinger, C. (2021). Interactive features of E-texts' effects on learning: A systematic review and meta-analysis. *Interactive Learning Environments*, 1–16.

Crossley, S. A., Kim, M., Allen, L., & McNamara, D. (2019). Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Artificial Intelligence in Education (Vol. 11625, pp. 84–95). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_8.

Galbraith, D., & Baaijen, V. M. (2018). The work of writing: Raiding the Inarticulate. *Educational Psychologist*, *53*(4), 238–257. https://doi.org/10.1080/00461520.2018.1505515.

Gamage, D., Staubitz, T., & Whiting, M. (2021). Peer assessment in MOOCs: Systematic literature review. *Distance Education*, *42*(2), 268–289. https://doi.org/10.1080/01587919.2021.1911626.

Ganesan, K. (2018). *ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks*. https://doi.org/10.48550/ARXIV.1803.01937.

Graham, S., & Harris, K. R. (2015). Common Core State standards and writing: Introduction to the Special Issue. *The Elementary School Journal*, *115*(4), 457–463. https://doi.org/10.1086/681963.

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in Science, Social studies, and Mathematics: A Meta-analysis. *Review of Educational Research*, *90*(2), 179–226. https://doi.org/10.3102/0034654320914744.

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction*, *28*(4), 1–11. https://doi.org/10.1080/19388078909557982.

Inflianskas, R. (2019). Profanity Filter. *GitHub repository*. https://github.com/rominf/profanity-filter/blob/master/profanity_filter/data/en_profane_words.txt.

Ji, S. W., Michaels, S., & Waterman, D. (2014). Print vs. electronic readings in college courses: Cost-efficiency and perceived learning. *The Internet and Higher Education*, *21*, 17–24.

Khandelwal, U., Clark, K., Jurafsky, D., & Kaiser, L. (2019). *Sample Efficient Text Summarization Using a Single Pre-Trained Transformer*. https://doi.org/10.48550/ARXIV.1905.08836.

Kim, M. K., Gaul, C. J., Bundrage, C. N., & Madathany, R. J. (2020). Technology supported reading comprehension: A design research of the student mental model analyzer for research and teaching (SMART) technology. *Interactive Learning Environments*, 1–25. https://doi.org/10.1080/10494820.2020.1838927.

Kumar, G., Banchs, R., & D'Haro, L. F. (2015). RevUP: Automatic gap-fill question generation from Educational texts. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, *154-161*, https://doi.org/10.3115/v1/W15-0618.

Labutov, I., Huang, Y., Brusilovsky, P., & He, D. (2017). Semi-supervised techniques for Mining Learning outcomes and prerequisites. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *907*, 915. https://doi.org/10.1145/3097983.3098187.

Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–6. https://doi.org/10.1109/CITS52676.2021.9618476.

Lan, A. S., & Baraniuk, R. G. (2016). A Contextual Bandits Framework for Personalized Learning Action Selection. *EDM*, 424–429.

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents* (pp. 1188–1196). PMLR.

Li, H., Cai, Z., & Graesser, A. C. (2018). Computerized summary scoring: Crowdsourcing-based latent semantic analysis. *Behavior Research Methods*, *50*(5), 2144–2161. https://doi.org/10.3758/s13428-017-0982-7.

Lin, C. Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, *1*, 71–78. https://doi.org/10.3115/1073445.1073465.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692 [Cs]*. http://arxiv.org/abs/1907.11692.

Lopez, L. E., Cruz, D. K., Cruz, J. C. B., & Cheng, C. (2021). Simplifying Paragraph-Level Question Generation via Transformer Language Models. In D. N. Pham, T. Theeramunkong, G. Governatori, & F. Liu (Eds.), *PRICAI 2021: Trends in Artificial Intelligence* (Vol. 13032, pp. 323–334). Springer International Publishing. https://doi.org/10.1007/978-3-030-89363-7_25.

Martínez-Huertas, J. Á., Jastrzebska, O., Olmos, R., & León, J. A. (2019). Automated summary evaluation with inbuilt rubric method: An alternative to constructed responses and multiple-choice tests assessments. *Assessment & Evaluation in Higher Education*, *44*(7), 1029–1041. https://doi.org/10.1080/02602938.2019.1570079.

Morris, W., Crossley, S., Holmes, L., Ou, C., McNamara, D., & Dascalu, M. (2023a). Using Large Language Models to Provide Formative Feedback in Intelligent Textbooks. In *International Conference on Artificial Intelligence in Education* (pp. 484–489). Cham: Springer Nature Switzerland.

Morris, W., Crossley, S. A., Langdon, H., & Trumbore, A. (2023b). Using Transformer Language Models to Validate Peer-Assigned Essay Scores in Massive Open Online Courses (MOOCs). In Proceedings of the Thirteenth International Conference on Learning Analytics & Knowledge.

Nelson, N., & King, J. R. (2022). Discourse synthesis: Textual transformations in writing from sources. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10243-5.

Ng, J. P., & Abrecht, V. (2015). *Better Summarization Evaluation with Word Embeddings for ROUGE* (arXiv:1508.06034). arXiv. http://arxiv.org/abs/1508.06034.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, *311*, https://doi.org/10.3115/1073083.1073135.

Phillips Galloway, E., & Uccelli, P. (2019). Beyond reading comprehension: Exploring the additional contribution of Core Academic Language skills to early adolescents' written summaries. *Reading and Writing*, *32*(3), 729–759. https://doi.org/10.1007/s11145-018-9880-3.

Ramasesh, V. V., Lewkowycz, A., & Dyer, E. (2021). Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*. https://openreview.net/pdf?id=GhVS8_yPeEa

Rockinson-Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education*, *63*, 259–266.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. (2019). *Answers Unite! Unsupervised Metrics for Reinforced Summarization Models*. https://doi.org/10.48550/ARXIV.1909.01610.

Seaman, J. E., & Seaman, J. (2020). *Digital texts in the time of COVID: Educational resources in U.S. Higher Education*. Bay View Analytics.

Shao, T., Guo, Y., Chen, H., & Hao, Z. (2019). Transformer-based neural network for answer selection in question answering. *Ieee Access: Practical Innovations, Open Solutions*, *7*, 26146–26156. https://doi.org/10.1109/ACCESS.2019.2900753.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1), 101. https://doi.org/10.1186/s40537-021-00492-0.

Silva, M., A., & Limongi, R. (2019). Writing to learn increases long-term memory consolidation: A Mental-Chronometry and computational-modeling study of Epistemic writing. *Journal of Writing Research*, *11(vol*(11 issue 1), 211–243. https://doi.org/10.17239/jowr-2019.11.01.07.

Sosnovsky, S., Brusilovsky, P., & Lan, A. (2023). Intelligent textbooks: The fifth international workshop. In *international conference on artificial intelligence in education* (pp. 97–102). Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-36336-8_15

Thaker, K., Zhang, L., He, D., & Brusilovsky, P. (2020). Recommending Remedial Readings Using Student Knowledge State. *Educational Data Mining Society*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural Language Processing with transformers*. O'Reilly Media, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, M., Chau, H., Thaker, K., Brusilovsky, P., & He, D. (2021). Knowledge annotation for Intelligent textbooks. *Technology Knowledge and Learning*. https://doi.org/10.1007/s10758-021-09544-z.

Weber, G., & Brusilovsky, P. (2016). ELM-ART– An Interactive and Intelligent web-based Electronic Textbook. *International Journal of Artificial Intelligence in Education*, *26*(1), 72–81. https://doi.org/10.1007/s40593-015-0066-8.

Winchell, A., Mozer, M., Lan, A., Grimaldi, P., & Pashler, H. (2018). Can Textbook Annotations Serve as an Early Predictor of Student Learning? *International Data Mining Society*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Yarbro, J. T., & Olney, A. M. (2021). Contextual Definition Generation. *Proceedings of the Third International Workshop on Intelligent Textbooks*, *2895*.

Authors and Affiliations

Wesley Morris[1] (ORCID) **· Scott Crossley[1] · Langdon Holmes[1] · Chaohua Ou[2] · Mihai Dascalu[3] · Danielle McNamara[4]**

✉ Wesley Morris
   wesley.g.morris@vanderbilt.edu

[1]   Vanderbilt University, Nashville, TN, USA

[2]   Georgia Institute of Technology, Atlanta, GA, USA

[3]   Polytechnic University of Bucharest, Bucharest, Romania

[4]   Arizona State University, Tempe, AZ, USA