Generating Response-Specific Elaborated Feedback Using Long-Form Neural Question Answering

Andrew M. Olney
University of Memphis
Memphis, USA
aolney@memphis.edu

ABSTRACT

In contrast to simple feedback, which provides students with the correct answer, elaborated feedback provides an explanation of the correct answer with respect to the student's error. Elaborated feedback is thus a challenge for AI in education systems because it requires dynamic explanations, which traditionally require logical reasoning and knowledge engineering to generate. This study presents an alternative approach that formulates elaborated feedback in terms of long-form question answering (LFQA). An off-the-shelf LFQA system was evaluated by human raters in a 2x2x2x2 ablation design that manipulated the context documents given to the LFQA model and the post-processing of model output. Results indicate that context manipulations improve performance but that post-processing can have detrimental results.

Author Keywords

elaborated feedback;question answering;information retrieval;deep learning;natural language generation;human evaluation;

CCS Concepts

•Computing methodologies → Natural language generation; Neural networks; •Applied computing → Interactive learning environments;

INTRODUCTION

Feedback is an educational practice that is both commonplace and extraordinarily complex. While naively one might assume that feedback is always beneficial for learning, various meta-analyses have shown that, depending on the conditions, feedback can be detrimental [2, 25, 10]. A prominent example of detrimental feedback in computer-based instruction is "hint abuse" [1], whereby students continue to request hints until they are given the correct answer. This particular detrimental effect was known in the feedback literature decades earlier, where it was called a presearch availability effect because students can locate and copy answers without reading or searching through the learning material [27, 2].

Feedback effects are complex because virtually any pedagogical goal can use feedback as a vehicle, though its proximity to student errors makes some pedagogical goals more salient than others. Models of feedback have attempted to capture this generality by categorizing feedback ranging from the proximal (task-level) to the more distal (self) [25, 21], such that while the timing of feedback may be contingent on a student's correct or incorrect response, the informational content of feedback may go beyond correctness and include an explanation of the error, metacognitive strategies for problem solving, or reaffirming motivational statements related to performance. A similar distinction has been made in human tutoring research, which contrasts feedback in modes of scaffolding and highlighting. Scaffolding-driven feedback focuses on marking the critical features that make the student answer incorrect [44], whereas highlighting focuses on explaining how the student made the error and how to avoid it in the future [6].

Feedback with additional information is typically called elaborated or explanatory feedback to contrast it with feedback that tells correctness (right/wrong) or gives the student the correct answer when they err. Several meta-analyses have shown positive effects of elaborated feedback over these simpler types of feedback [2, 10], establishing elaborated feedback as an important area in feedback research. This focus has led to the inclusion of many different kinds of information in elaborated feedback; correspondingly, taxonomies have been proposed to capture variations in both the form and substance of elaborated feedback. For example, Shute describes six different subtypes of elaborated feedback: attribute isolation of target concept/skill, topic contingent, response contingent, hints/cues/prompts, bugs/misconceptions, and informative tutoring [41]. These subtypes may be considered as three underlying dimensions: content specificity (response, concept/skill, or topic), feedback form (direct instruction, single hints, interactive hints/tutoring), and content adaptation (domain level or learner level; i.e. would any student making this particular error get the same feedback, or is it adapted using the learner model).

When considering elaborated feedback under these dimensions, it is clear that providing such feedback can be challenging for computer-based instruction. While in some cases,

Accepted version of work. See DOI below for publisher's version of record.
L@S '21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06 ...$15.00.
https://dx.doi.org/10.1145/3430895.3460131

elaborated feedback may be pre-authored manually with some effort, in other cases elaborated feedback must be generated dynamically. For example, concept/skill- or topic-specific elaborated feedback may be implemented with a limited number of prepared remediation responses, minimally equal in number to the concept/skills or topics being taught. Similarly, these responses can be converted into hint- and tutorial-dialogue form with additional effort. In contrast, response-specific feedback must be dynamic under the reasoning that there are more ways to answer a question incorrectly than correctly. As a result, pre-authored responses that are specific to student errors will necessarily be incomplete, even at the domain level where the feedback would be the same for different students. The situation is even more intractable for learner-level content adaptation, where feedback depends on the bugs/misconceptions in a learner model.

Because of these difficulties, many computer-based instruction systems, including intelligent tutoring systems, have used limited numbers of pre-authored feedback messages. The buggy rule approach has been used in model-tracing tutors to associate feedback with a limited number of predicted student errors [26]. Complementarily, constraint-based tutors model constraints that any solution must have and so can associate feedback when students violate a constraint [33]. Perhaps the most advanced tutoring system that has generated response-specific feedback is BEETLE II [15]. BEETLE II uses symbolic AI to dynamically generate natural language feedback in an ongoing tutorial dialogue centered on a circuit simulator. To generate elaborated feedback, BEETLE II maps student natural language input into a logical representation and then diagnoses any errors or missing information in that representation using a knowledge base aligned with the state of the simulator (a microworld). Based on the diagnosis, new tutorial dialogue is generated that contains both simple feedback and elaborated feedback in the form of hints and other problem-solving cues.

As impressive as symbolic, microworld-based systems like BEETLE II are, it's unclear that response-specific feedback can be generated from such models over larger domains, because of the intense effort required to logically model even a small domain. Even with the assumption that existing logical representations and reasoners are sufficient, one must still populate a knowledge base with information, and assuming natural language input/output, one must also convert natural language to logic and back as needed to generate dynamic feedback. Both populating large knowledge bases and converting natural language to and from logic have proven extremely challenging. Large manually-created knowledge bases like Cyc have been in development for decades [28] and their value to AI is not well established [13]. Large knowledge bases created from semi-structured text like Wikipedia [4] are more widely used, but they typically suffer from incompleteness [14]. Research on addressing incompleteness has increasingly turned to graph-embedding techniques, which project the knowledge base into a latent space [34, 11]. Analogously, word embedding models like BERT [12] have been used to improve the performance of models that parse natural language into logical representations, though even the best models have relatively weak connections to an actual knowledge base [45, 7], which constrains their usefulness for generating feedback.

In contrast, recent work that entirely skips knowledge base creation and logical representations has shown great promise. Knowledge-base-free advances include a single BERT-based model that can achieve a B grade on the New York Regents Science Exam at the 8th- and 12th-grade levels [9] and a Transformer-based model that achieves state-of-the-art performance on open-domain question answering tasks without access to additional external knowledge sources [38]. This work reflects an increasing convergence between language models, knowledge bases, and question answering, such that problems in one have been shown as solvable in another [22, 29, 37]. These developments suggest the possibility that response-specific feedback can be generated without a knowledge base or logical representation.

Question answering, which takes natural language input as query and returns natural language output, is of particular interest because of its alignment with feedback's output requirements. Recent years have seen rapid advances in neural question answering, including identifying the answer to a question in a given paragraph [39], fusing information across a given set of paragraphs to produce the answer [32], and open-domain question answering which requires the system to additionally find the documents [23, 20]. While generally question answering focuses on factoid, Jeopardy!-style questions, recent work has focused on "why" and "how" questions, e.g. "Why can't we just print money to pay off our debt?", that typically require a long-form paragraph answer instead of a word or phrase [16]. This work has two properties that suggest it might be suitable for domain-general, response-specific feedback in computer-based learning environments. First, it sources questions and their answers from the "Explain Like I'm Five" Reddit forum (ELI5), whose rules require answers to be full explanations accessible to laypeople. Second, it is domain-general: not only is ELI5 domain-general but the model also can be tuned to a domain without retraining because it uses separate models for document retrieval and answer generation (cf. [30]). The remainder of this paper presents an approach to generating response-specific feedback using ELI5-based long-form question answering, as well as an ablation-style evaluation of design choices underlying this approach.

APPROACH

To motivate our approach, consider the scenario presented in Figure 1, in which a student gives an incorrect answer to a cloze-style question in our target domain of anatomy and physiology. The feedback contains three components: a statement of correctness, the correct answer, and then an elaboration that explains the relationship between the student's incorrect answer and the correct answer. The elaborated feedback can be viewed as the answer to a synthetic question asking about the relationship between the correct and incorrect answers. From this perspective, elaborated feedback is the answer to a question the student should have asked but didn't. This is the key intuition behind our approach of generating response-specific feedback using long-form question answering. We can

Test Item	The _____ at the distal end of the axon is rich in mitochondria and contains many tiny vesicles that store neurotransmitter.
Student Answer	acetylcholine
Elaborated Feedback	Acetylcholine is not right. The correct answer is cytoplasm. Acetylcholine is synthesized in the cytoplasm of nerve terminals by the enzyme choline acetyltransferase, and is then transported into synaptic vesicles.
Synthetic Question	What is the relationship between acetylcholine and cytoplasm?

Figure 1. Student error and idealized feedback. The elaborated feedback component is highlighted and can be viewed as the answer to the synthetic question below.

justify the longer feedback by assuming that in the moment of error, the student potentially has three pieces of faulty knowledge: understanding of the target concept, understanding of the confused concept, and understanding the relationship between them. If understanding of these pieces of knowledge were tracked in a learner model, a system could opt for simple feedback when elaborated feedback wasn't needed, or even use a different synthetic question better suited to the student's knowledge state. However, in this paper, we only explore elaborated feedback without a learner model, i.e. at the domain level.

To further explain our approach, consider the behavior of the ELI5 model. The model has a retriever/reader structure (cf. [23, 30]), such that the retriever and reader can be completely separate models. The retriever's goal is to find relevant documents, and the reader's goal is to identify information in those documents that is useful for answering the input question. The relationship between the retriever and the reader models is intuitively important, and one might assume that the reader can't answer questions without good documents from the retriever. As previously discussed, however, language models can be used for open-domain question answering without external documents [38], which means that a great deal of information is contained within the reader itself. To illustrate the retriever/reader relationship, we present a somewhat humorous example in Figure 2 using the HuggingFace [43] implementation of the ELI5 model, which is the reference model we use for our experiments (for a tutorial on this model, see https://yjernite.github.io/lfqa.html).

Although the answer is nonsense overall, it nevertheless represents nontrivial knowledge about the structure of a long-form answer to a question. The answer contains both of the key terms in the question, as well as the types of language about similarities and differences that one might expect in the answer to such a question. Most of the answer sentences mention the key terms, with the exception of the highlighted sentences which seem to be generic. With this example as a guide, we identify three places where one could intervene on the behavior of the model. First, one can strategically manipulate the documents returned by the retriever. This appears

Question	What is the difference between a chicken and a noodle?
Answer	A chicken is a chicken. A noodle is a noodle. They are the same thing. The difference is in the way they are shaped. Chicken tenders and noodle noodles are different shapes, but they are both the same size and shape. Noodles and chicken tenders are different sizes and shapes.

Figure 2. Long-form answer returned by the reader model in the absence of support documents from the retriever model. Generic sentences are highlighted.

to be relatively unexplored in open-domain question answering, where Wikipedia is commonly used for external knowledge. In educational settings, however, one may consider resources like a textbook, glossary, or other course-related material as sources of relevant documents. Second, one can train the reader to generate better answers. This could be done by training a larger language model, but this would sacrifice the domain-independent benefit of having a relatively low-knowledge reader. Similarly training the reader to a particular domain might improve performance for that domain but sacrifice domain independence. Third, one can manipulate the answer itself, i.e. engage in post-processing. Such post-processing could take many forms, from correcting aspects of the answer, filtering out aspects of the answer, or rejecting the answer entirely.

In the present study, we focus on the first and third options, manipulating the context documents and post-processing the answer. Our rationale is twofold: our initial explorations suggest that these are essential for obtaining good feedback, and we expect long-form question answering to continue to advance the state of the art along the second option, making our work complementary and relevant when that occurs. Using a handful of incorrect answers for test items like that shown in Figure 1, we developed the following procedure for generating feedback:

Algorithm 1: Response-Specific Elaborated Feedback

Function ***generate*** *incorrectTerm correctTerm item*
- (1) Retrieve definitions k for key terms;
  Create synthetic question q;
  Retrieve documents d_q using synthetic question;
- (2) Filter documents that don't contain both key terms to obtain d_f;
- (3) Construct a document list d containing k, $item$, d_f;
  Perform long-form question answering with q and d to obtain a;
- (4) Resolve coreference in a and filter sentences in a that don't contain a key term to obtain a_f;
- **return** a_f

end

The numbered steps (1-4) of Algorithm 1 are aligned with the first and third options mentioned above, with the following rationales. Definitions could be helpful when the regularly

returned documents don't contain basic information about key terms. Filtering documents that don't contain both terms could be helpful given our synthetic question always calls for an answer that describes how the key terms are related, and documents that don't contain both key terms are less likely to have this information. Including the test item in the documents submitted for question answering could be useful for providing a more specific context for the feedback, rather than a generic one. While these steps are directly related to bending the overall model towards providing pedagogically relevant elaborated feedback, the final step, coreference plus sentence filtering, is aimed at scrubbing generic sentences, like those shown in Figure 2, from the final answer. The coreference resolution component of this step is meant to allow sentences that are clearly about the key terms yet use pronouns in their original formulation. We considered keeping all resolutions in the final answer but decided to keep the original text in order to avoid introducing coreference resolution errors into the answers. Because these numbered steps of Algorithm 1 are based on intuition and analysis of a small number of examples, we conducted an ablation-style evaluation of these steps using expert human judges. Our primary research questions are (1) how correct and informative is the elaborated feedback under each condition and (2) how grammatical and fluent is the elaborated feedback under each condition.

METHOD

Design

The evaluation study used a within-subjects design with ablation of the four steps in Algorithm 1 as conditions, i.e. a 2 x 2 x 2 x 2 design where each step is either included or not. Conditions were presented using a 16 x 16 balanced latin square to counterbalance condition order and prevent carryover effects between conditions. However, the underlying context of each elaborated feedback (i.e., the incorrect/correct answer) was not counterbalanced. This design decision was made to remove potential interactions between contexts and fatigue, where participants might process contexts differently at the beginning of the experiment versus the end. It also means that in a fully-used latin square, a context in a particular location would be paired with each condition, making fatigue effects equivalent across conditions. The judgments were analyzed using mixed-effects beta regression with random intercepts for judge and context using the `glmmTMB` R package [5]. Beta regression is appropriate for continuous bounded outcome variables, unlike linear regression, which isn't suitable for bounded outcomes, and logistic regression, which can be used for proportions, but only when the proportion is a ratio of two counts [24]. Because beta regression is defined on the open interval (0,1), we use a standard transformation to squeeze our closed interval outcome variables to the open interval [42]. We conducted statistical tests at $\alpha = .05$ to address our research questions.

Participants

Raters ($N = 30$) were recruited through the Amazon Mechanical Turk (AMT) marketplace between January and February of 2021, using the CloudResearch platform [31]. In this study, raters were required to be native English speakers, or have learned English before the age of 7, reside in the U.S., Canada, New Zealand, United Kingdom, or Australia, have completed at least an Associate Degree, and be employed as a nurse or physician. The educational and occupational constraints were selected to ensure that raters were experts in the evaluation subject domain: they had passed anatomy and physiology in their studies and used this knowledge on a daily basis. Demographic constraints are enforced by CloudResearch based on rater responses to previous demographic surveys. Raters were further required to have completed at least 100 previous AMT tasks with at least a 95% approval rating. Raters were initially paid $7. However, as the study progressed and the pool of CloudResearch-qualified raters appeared to be exhausted, an additional survey was conducted within AMT to recruit more raters, and these raters were paid $10, with a sliding bonus scale for passing quality checks: a $5 bonus for passing one check, and an additional $20 bonus for passing both checks.

Materials

Authentic student errors were collected from college anatomy and physiology (A&P) courses at a community college in Tennessee, USA. Each error was in response to a cloze question similar to that in Figure 1, generated from the course textbook automatically [35, 36]. Errors were aggregated and counted based on student incorrect answers for a target correct answer, independent of cloze question, and the most frequent 80 errors and their associated most frequent cloze items were selected for the creation of elaborated feedback. Elaborated feedbacks were created as described in Section 2 using the HuggingFace Transformers library [43] and ELI5 model. Definitions were obtained by first querying a machine-readable glossary from the A&P textbook used in the courses [40]. If no definition was found in the glossary, definition-like text was retrieved from Wikipedia by first applying a wikifier [8] to the correctly filled-in cloze item sentence in order to get Wikipedia page ids for the key terms in the sentence. These page ids were then used to query the corresponding Wikipedia pages for their first paragraph of text, which was used as a proxy for a definition. Key term filtering was implemented using regular expressions. Additional supporting documents were retrieved using an Elasticsearch (https://www.elastic.co/elasticsearch/) index of the A&P textbook [40]. A maximum of three such documents (post filtering) was used to control for search result differences across student errors. Coreference resolution filtering was implemented using AllenNLP's implementation [18]. Using these resources, 80 elaborated feedbacks were created for each of 16 conditions.

Sixteen surveys were created with Qualtrics, an online survey tool, using a balanced latin square to define the order of conditions. Because each row of the latin square only contains 16 orderings, each ordering was repeated 5 times in a survey. In this way, each rater made judgments on each condition 5 times, with an interval of 16 between each repetition of a condition. The same ordering of 80 error contexts was used in each survey; only the condition applied to each position of the ordering varied across surveys. Each context was formatted on a single survey page using the direct assessment methodology [19, 17]. An example survey page is shown in Figure 3. The student incorrect answer and correct answer were formatted

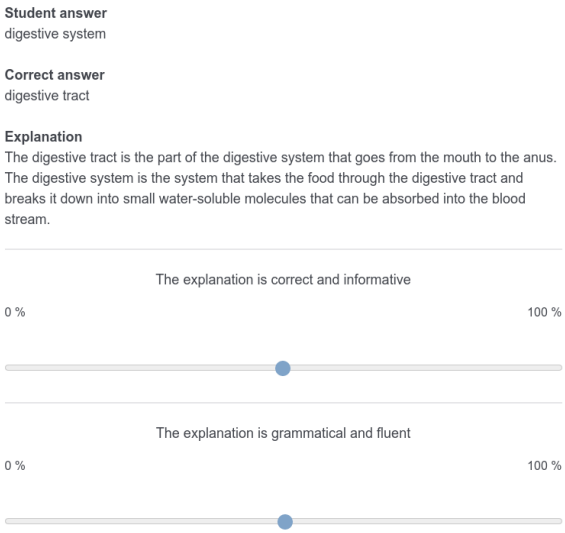

Figure 3. Survey page illustrating the rating task for one error context. The elaborated feedback was generated using all steps from Algorithm 1.

above the elaborated feedback to allow for easy comparison, followed by two questions with slider-format response on a 0-100 scale. The first was a meaning-assessment question, "The explanation is correct and informative," and was anchored by 0% on the left and 100% on the right. The second was a fluency-assessment question, "The explanation is grammatical and fluent", and was again anchored by 0% and 100%. The decision to use percentages as anchors was made during piloting when raters expressed confusion in making judgments for multiple sentences at once, where some might be perfect and others problematic. The sliders had no numeric indicators and were initialized at the midpoint. The percentage anchors were used to help raters focus on making a judgment involving all of the sentences rather than a single sentence. Both the percentage anchors and the meanings of the scales were explained in instructions at the beginning of the survey.

Following the direct assessment methodology, control pairs were created to evaluate the internal reliability of each rater, adding an additional 20 pages to each survey for a total of 100 pages [17, 19, 3]. Control pairs were created by copying an existing error context (a survey page) and then degrading the elaborated feedback on that page. A two-step process was used to degrade elaborated feedback. First, we mapped the span deletion rules proposed by [19] for machine translation into a simple linear regression formula, $span_{length} = 0.21696 * word_{count} + 0.78698$. Degraded elaborated feedbacks were created by deleting a random span of $span_{length}$ words, rounded down. Because the deleted span is contiguous, as $word_{count}$ increases, words at the beginning/end of the elaborated feedback are less likely to be deleted, and words towards the interior are more likely to be deleted. In piloting we found, however, that such deletions could be very subtle in a paragraph of several sentences on the same general topic, in contrast to the single-sentence translation tasks for which they were developed. Therefore we developed a new text-degradation approach that we call split-sentence derangement that proceeds as follows. First, an interior word boundary is randomly selected for each sentence, and each sentence is split into a two strings occurring before that point and after that point, forming a pair. Next, the pairs are deranged, i.e. permuted such that all original pairings are broken, and the deranged pairs are recombined to form degraded sentences. Finally, the degraded sentences are shuffled to create a new sentence ordering. We used the split-sentence derangement approach for all degraded elaborated feedbacks except those comprised of a single sentence, for which we used the span deletion approach. An example split-sentence derangement for the elaborated feedback in Figure 3 is:

> The takes the food through the digestive tract and breaks it down into small water-soluble molecules that can be absorbed into the blood stream. The digestive system is the system that digestive tract is the part of the digestive system that goes from the mouth to the anus.

Each survey of 100 pages contained 80 distinct pages and 20 degraded versions of distinct pages. We refer to a distinct page and its degraded version as a control pair. These numbers were chosen because they represent the sample size needed to detect a large (.8 SD) effect using a Wilcoxon signed-ranks test for matched pairs at $\alpha = .05$ and .95 power on a one-tailed test. If we do not detect a large effect between ratings of elaborated feedbacks and their degraded versions, we infer the rater is not reliable. The degraded pages were in the same randomly assigned positions in each survey and were evenly distanced from their matched distinct pages, modulo 50. This ensured that pages in control pairs had 50 other items between them, making it less likely that raters would remember their rating on a previous item. Because of the complexity of the survey design and their length, a Qualtrics export file was reverse engineered and the survey items were programmatically generated and imported into Qualtrics.

We additionally developed an occupation survey to help us find more qualified raters. The occupation survey consisted of two questions, a generic occupation question from the standard Qualtrics demographics library, and a conditional branch question that only appeared if a respondent selected `healthcare` on the first question. The conditional branch question asked for a more specific healthcare occupation, with options including certified nursing positions and physician positions matching our original recruiting criteria. This indirect approach to asking about specific healthcare occupations was designed to avoid demand characteristics (i.e., false responses) from asking such questions directly.

Procedure

The survey was initially piloted to better assess the amount of time required for completion, the number of raters passing the meaning and fluency control checks (i.e., intra-rater reliability using the control pairs), and the agreement amongst raters (inter-rater reliability). Piloting suggested that the span deletion approach was insufficient for creating degraded items for paragraph-length text, leading to a redesign of the degraded items using split-sentence derangement.

Survey	Meaning		Fluency	
	α	n	α	n
1	.940	4	.835	4
2	.979	3	.965	5
3	.957	3	.980	6
4	.981	3	.954	4
5	.901	2	.946	3
6	.985	2	.932	3
7	.837	2	.932	2

Table 1. Inter-rater reliability per survey for included raters.

Two waves of surveys followed piloting. In the first wave, successive surveys were released serially with a cap of 4 raters; however, this cap was increased as needed to ensure that at least 3 raters per survey and per measure passed control checks. However, during the third survey, it became clear that there were either not enough qualified raters to recruit in the CloudResearch pool or that the monetary incentive was not enough to attract all available qualified raters. Therefore we used the occupation survey to find an additional 19 qualified raters out of 603 respondents. In the second wave, we invited these raters in addition to the existing CloudResearch pool. We also changed the incentives for the task, as previously described, and further changed the instructions of the task to encourage participants to use reference materials as needed during the task. Finally, we relaxed the requirement to have 3 raters per survey to 2 raters per survey, the minimum needed to calculate inter-rater reliability.

In all waves, raters discovered the surveys through AMT and completed the surveys using Qualtrics. Because the study is a system evaluation and not human subjects research, informed consent was not obtained. Raters saw the instructions for the survey twice, once as a preview on AMT before undertaking the survey, and again once they clicked on the survey link. On each following page, raters read a student incorrect answer, a correct answer, and an elaborated feedback as shown in Figure 3 and then completed the corresponding rating of meaning and rating of fluency. Raters were paid upon completion of the survey, and in the second wave, received bonuses on confirmation they had passed control checks for meaning and fluency.

RESULTS AND DISCUSSION

Even with these changes between the first and second waves of surveys, we were only able to attract enough raters to complete 7 rows of the latin square. Median completion time across surveys was 62 minutes, giving approximately 37 seconds to read the error context and paraphrase and then make meaning and fluency judgments. Although 30 raters completed surveys, only a subset successfully passed control checks for meaning ($n = 19$) and fluency ($n = 27$). Control checks were considered to be passed if $p < .05$ on the aforementioned Wilcoxon signed-ranks test. Cronbach's alpha was calculated for raters passing control checks in each survey and was high overall ($\alpha > .8$) Table 1 shows the number of included raters and their inter-rater reliability per survey; all other raters were excluded from further analysis.

Our first research question is how correct and informative the elaborated feedback is under each condition (meaning rating). Table 2 shows the mean meaning rating for each condition. The first row of the table is the baseline system with none of the four steps in Algorithm 1. We can immediately see that multiple conditions are below the baseline, indicating that some of the steps are negatively impacting meaning ratings. In particular, the conditions on the next three rows, which include key term filtering, coreference filtering, or both, are below baseline, suggesting that these steps may be detrimental to performance. The best performing condition is the definition-only condition, whose meaning rating is approximately 20% above the baseline.

To answer our first research question, we ran a mixed-effects beta regression with random intercepts for judge and context. The model did not include interaction terms because we are interested in the additive effect of each step. The beta regression revealed significant effects for all four steps on ratings of meaning. Conditions including definition documents had significantly higher meaning ratings ($M = 61.01$, $SE = 1.18$) than conditions without definition documents ($M = 50.95$, $SE = 1.22$), $\beta = .50$, $z = 7.50$, $p < .001$. Though to a lesser extent, conditions including cloze documents also had significantly higher meaning ratings ($M = 58.08$, $SE = 1.18$) than conditions without them ($M = 53.89$, $SE = 1.24$), $\beta = .17$, $z = 2.67$, $p = .007$. In contrast, conditions with key term filtering had significantly lower meaning ratings ($M = 51.71$, $SE = 1.24$) than those without ($M = 60.26$, $SE = 1.17$), $\beta = -.43$, $z = -6.77$, $p < .001$. Though to a lesser extent, conditions including coreference filtering also had significantly lower meaning ratings ($M = 53.95$, $SE = 1.24$) than those without ($M = 58.02$, $SE = 1.17$), $\beta = -.16$, $z = 2.50$, $p = .013$. Notably, the steps that significantly increased meaning ratings both involved giving additional documents to the reader that the retriever would not otherwise provide, and the steps that significantly decreased meaning ratings either removed documents the retriever provided or edited the answer post hoc.

Our second research question is how grammatical and fluent the elaborated feedback is under each condition (fluency rating). Table 3 shows the mean fluency rating for each condition. The first row of the table is the baseline system with none of the four steps in Algorithm 1. The baseline is approximately 30% higher than the meaning baseline. This suggests the fluency baseline is a relatively strong baseline and might explain why eleven conditions are below it. Notably, many of the conditions that are below baseline include key term filtering. The best performing condition again is the definition-only condition, but it is only about 3% above the baseline.

To answer our second research question, we ran a mixed-effects beta regression with random intercepts for judge and context, again only including model terms for main effects. The beta regression revealed significant effects for only two steps on ratings of fluency. Conditions including definition documents had significantly higher fluency ratings ($M = 71.74$, $SE = 1.02$) than conditions without definition documents ($M = 66.79$, $SE = 1.09$), $\beta = .20$, $z = 3.23$, $p = .001$. In con-

Definition documents	Cloze document	Key term filter	Coreference filter	M
.	.	.	.	56.73
.	.	.	✓	52.16
.	.	✓	.	48.29
.	.	✓	✓	34.99
.	✓	.	.	59.31
.	✓	.	✓	55.40
.	✓	✓	.	49.16
.	✓	✓	✓	51.57
✓	.	.	.	**68.22**
✓	.	.	✓	61.28
✓	.	✓	.	55.23
✓	.	✓	✓	54.19
✓	✓	.	.	65.80
✓	✓	.	✓	63.19
✓	✓	✓	.	61.42
✓	✓	✓	✓	58.79

Table 2. Mean meaning rating for each condition.

Definition documents	Cloze document	Key term filter	Coreference filter	M
.	.	.	.	74.26
.	.	.	✓	71.99
.	.	✓	.	60.89
.	.	✓	✓	59.84
.	✓	.	.	70.33
.	✓	.	✓	73.88
.	✓	✓	.	57.59
.	✓	✓	✓	65.49
✓	.	.	.	**76.31**
✓	.	.	✓	75.79
✓	.	✓	.	64.43
✓	.	✓	✓	66.56
✓	✓	.	.	74.36
✓	✓	.	✓	74.32
✓	✓	✓	.	69.27
✓	✓	✓	✓	72.93

Table 3. Mean fluency rating for each condition.

trast, conditions with key term filtering had significantly lower fluency ratings ($M = 64.63$, $SE = 1.15$) than those without ($M = 73.90$, $SE = .93$), $\beta = -.41$, $z = -6.69$, $p < .001$. As with the meaning ratings, the step that significantly increased fluency ratings gave additional documents to the reader, and the step that significantly decreased fluency ratings removed documents from the retriever.

Our strongest findings were that adding definition documents to the document set given to the reader improved both meaning and fluency ratings and that filtering out documents returned by the retriever without both key terms harmed both meaning and fluency ratings. The positive effect of adding definitions was expected: definitions provide high-quality information at a relatively basic level consistent with elaborated feedback. The finding of a positive effect over the documents already returned by the retriever suggests that such information is not occurring in the main body of the textbook. The negative

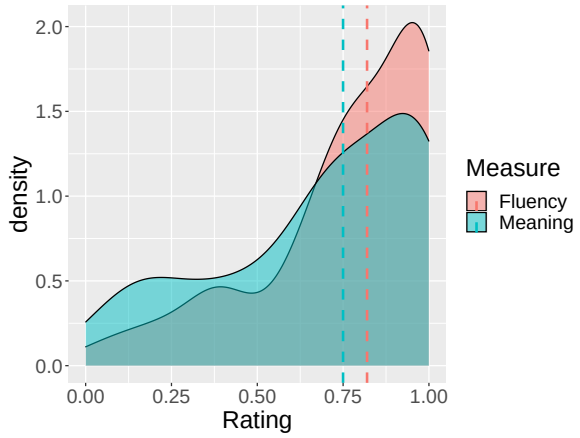

Figure 4. Density plot for ratings with indicated medians in the best-performing definition condition.

effect of filtering documents using key terms was unexpected because it too was aimed at improving the overall quality of documents input to the reader. The finding of a negative effect suggests either that the reader performs better with a breadth of input documents or that the key term filter may be reducing the documents available to the reader below a critical threshold necessary for good performance (cf. Figure 2).

The relatively weak and sometimes null effects of adding the cloze document and the coreference filter were also surprising. Perhaps the cloze document, which contributes item-specific context to the elaborated feedback, is less important to meaning ratings because the errors students make tend to go outside the item context, i.e., are basic errors. This explanation is consistent with the positive effect of definitions, which are context-free. The negative effect on meaning of filtering answer sentences that don't contain coreference-resolved key terms is perhaps best explained by it being an overly aggressive criterion that removes sentences that positively contribute to meaning. It could be that removing such sentences is a result of coreference resolution's failure to correctly resolve referents to key terms (false negatives), or it could indicate that sentences not containing key terms are contributing more to positive meaning ratings than anticipated, e.g., by providing background information.

The overall rating performance of the best system, which added only the definition documents to the baseline system, is shown in Figure 4. The median ratings are quite high, with a median rating of 75 for meaning and a median rating of 82 for fluency. However, the relatively long tail of ratings suggests many opportunities remain to improve overall performance.

CONCLUSION

We have proposed a new approach to generating response-specific elaborated feedback using long-form neural question answering and evaluated several approaches to improving the

quality of the feedback. Results from the evaluation study suggest that including definition information in the documents submitted to the reader model is important for improving results and that filtering documents submitted to the reader model can lead to worse performance. Because our approach can be applied to any textbook, this work has potentially broad implications for scaling up elaborated feedback for computer-based instruction. This approach could be used both for dynamic generation of feedback in a real-time system or to create draft feedback for manual review and correction by a domain expert, potentially reducing authoring effort.

Our study has several limitations. First, we were unable to complete our latin square design, so while it is not clear how condition order might have affected our results, it is still possible that some of our effects are confounded with condition order. Second, our evaluation was conducted with only one textbook on a single topic, anatomy and physiology. It may be that these results do not generalize well to other domains; investigating this question should be a fruitful target for future research.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants 1918751 and 1934745 by the Institute of Education Sciences under Grant R305A190448. 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 or the Institute of Education Sciences.

REFERENCES

[1] Vincent Aleven, Bruce McLaren, Ido Roll, and Kenneth Koedinger. 2004. Toward Tutoring Help Seeking. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, James C. Lester, Rosa Maria Vicari, and Fábio Paraguaçu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 227–239.

[2] Robert L. Bangert-Drowns, Chen-Lin C. Kulik, James A. Kulik, and MaryTeresa Morgan. 1991. The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research* 61, 2 (1991), 213–238. DOI: http://dx.doi.org/10.3102/00346543061002213

[3] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 272–303. DOI: http://dx.doi.org/10.18653/v1/W18-6401

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*. Association for Computing Machinery, New York, NY, USA, 1247–1250. DOI: http://dx.doi.org/10.1145/1376616.1376746

[5] Mollie E Brooks, Kasper Kristensen, Koen J Van Benthem, Arni Magnusson, Casper W Berg, Anders Nielsen, Hans J Skaug, Martin Machler, and Benjamin M Bolker. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* 9, 2 (2017), 378–400.

[6] Whitney L. Cade, Jessica L. Copeland, Natalie K. Person, and Sidney K. D'Mello. 2008. Dialogue Modes in Expert Tutoring. In *ITS '08: Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Springer-Verlag, Berlin, Heidelberg, 470–479.

[7] Deng Cai and Wai Lam. 2020. AMR Parsing via Graph-Sequence Iterative Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1290–1301. DOI: http://dx.doi.org/10.18653/v1/2020.acl-main.119

[8] Xiao Cheng and Dan Roth. 2013. Relational Inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1787–1796. https://www.aclweb.org/anthology/D13-1184

[9] Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. 2020. From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project. *AI Magazine* 41, 4 (Dec. 2020), 39–53. DOI: http://dx.doi.org/10.1609/aimag.v41i4.5304

[10] Fabienne M. Van der Kleij, Remco C. W. Feskens, and Theo J. H. M. Eggen. 2015. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research* 85, 4 (2015), 475–511. DOI: http://dx.doi.org/10.3102/0034654314564881

[11] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D Knowledge Graph Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018), 1811–1818. https://ojs.aaai.org/index.php/AAAI/article/view/11573

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: http://dx.doi.org/10.18653/v1/N19-1423

[13] P. Domingos. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books. https://books.google.com/books?id=CPgqCgAAQBAJ

[14] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. Association for Computing Machinery, New York, NY, USA, 601–610. DOI: http://dx.doi.org/10.1145/2623330.2623623

[15] Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education* 24, 3 (2014), 284–332.

[16] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3558–3567. DOI:http://dx.doi.org/10.18653/v1/P19-1346

[17] Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual Whispers: Generating Paraphrases with Translation. In *Proceedings of the 5th Workshop on Noisy User-Generated Text*. Association for Computational Linguistics, Hong Kong, China, 17–26. DOI:http://dx.doi.org/10.18653/v1/D19-5503

[18] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Melbourne, Australia, 1–6. DOI:http://dx.doi.org/10.18653/v1/W18-2501

[19] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is Machine Translation Getting Better over Time?. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, 443–451. DOI:http://dx.doi.org/10.3115/v1/E14-1047

[20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 119. PMLR, 3929–3938. http://proceedings.mlr.press/v119/guu20a.html

[21] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. DOI: http://dx.doi.org/10.3102/003465430298487

[22] Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 643–653. DOI:http://dx.doi.org/10.18653/v1/D15-1076

[23] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. DOI: http://dx.doi.org/10.18653/v1/2020.emnlp-main.550

[24] Robert Kieschnick and B. D. McCullough. 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling* 3, 3 (2003), 193–213. DOI: http://dx.doi.org/10.1191/1471082X03st053oa

[25] Avraham N. Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119, 2 (1996), 254–284.

[26] Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8, 1 (1997), 30–43.

[27] Raymond W. Kulhavy. 1977. Feedback in Written Instruction. *Review of Educational Research* 47, 2 (1977), 211–232. DOI: http://dx.doi.org/10.3102/00346543047002211

[28] Douglas B. Lenat. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Commun. ACM* 38, 11 (1995), 33–38. citeseer.ist.psu.edu/lenat95cyc.html

[29] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, 333–342. DOI: http://dx.doi.org/10.18653/v1/K17-1034

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.),

Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[31] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (2017), 433–442. DOI:http://dx.doi.org/10.3758/s13428-016-0727-z

[32] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6097–6109. DOI: http://dx.doi.org/10.18653/v1/P19-1613

[33] Antonija Mitrovic. 2012. Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 39–72.

[34] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* 104, 1 (2016), 11–33. DOI:http://dx.doi.org/10.1109/JPROC.2015.2483592

[35] Andrew M. Olney, Philip J. Pavlik Jr., and Jaclyn K. Maass. 2017. Improving Reading Comprehension with Automatically Generated Cloze Item Practice. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, Elisabeth André, Ryan Baker, Xiangen Hu, Ma Mercedes T Rodrigo, and Benedict du Boulay (Eds.). Springer, 262–273. DOI: http://dx.doi.org/10.1007/978-3-319-61425-0

[36] Philip I. Pavlik Jr., Andrew M. Olney, Amanda Banker, Luke Eglington, and Jeff Yarbro. 2020. The Mobile Fact and Concept Textbook System (MoFaCTS). In *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 co-located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, Sergey Sosnovsky, Peter Brusilovsky, Richard Baraniuk, and Andrew Lan (Eds.). 35–49.

[37] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. DOI: http://dx.doi.org/10.18653/v1/D19-1250

[38] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5418–5426. DOI: http://dx.doi.org/10.18653/v1/2020.emnlp-main.437

[39] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *Proceedings of the 5th International Conference on Learning Representations*. OpenReview.net. https://openreview.net/forum?id=HJ0UKP9ge

[40] D. Shier, J. Butler, and R. Lewis. 2019. *Hole's Human Anatomy & Physiology* (15th ed.). McGraw-Hill Education.

[41] Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (2008), 153–189. DOI:http://dx.doi.org/10.3102/0034654307313795

[42] Michael Smithson and Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11, 1 (2006), 54–71.

[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. DOI: http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6

[44] David Wood, Jerome S. Bruner, and Gail Ross. 1976. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry* 17, 2 (1976), 89–100.

[45] Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR Parsing as Sequence-to-Graph Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 80–94. DOI: http://dx.doi.org/10.18653/v1/P19-1009