More Insightful Feedback for Tutoring: Enhancing Generation Mechanisms and Automatic Evaluation

Wencke Liermann[1][2]*, Jin-Xia Huang[1]†, Yohan Lee[1], Kong Joo Lee[2]

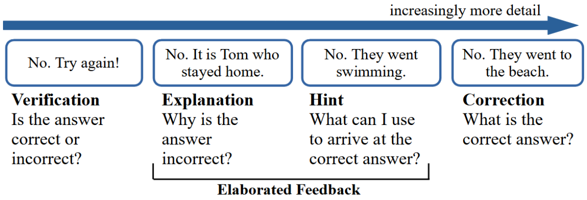
[1] Electronics and Telecommunications Research Institute, Republic of Korea
[2] Chungnam National University, Republic of Korea
{wliermann, hgh, carep}@etri.re.kr, kjoolee@cnu.ac.kr

Abstract

Incorrect student answers can become valuable learning opportunities, provided that the student understands where they went wrong and why. To this end, rather than being given the correct answer, students should receive elaborated feedback on how to correct a mistake on their own. Highlighting the complex demands that the generation of such feedback places on a model's input utilization abilities, we propose two extensions to the training pipeline. Firstly, we employ a KL regularization term between a standard and enriched input format to achieve more targeted input representations. Secondly, we add a preference optimization step to encourage student answer-adaptive feedback generation. The effectiveness of those extensions is underlined by a significant increase in model performance of 3.3 METEOR points. We go beyond traditional surface form-based metrics to assess two important dimensions of feedback quality, i.e., faithfulness and informativeness. Hereby, we are the first to propose an automatic metric measuring the degree to which feedback divulges the correct answer, that we call Informativeness Index I^2. We verify in how far each metric captures feedback quality.

1 Introduction

As technology continues to reshape the way we learn, an increasing number of people opt to learn new skills from the comfort of their own home. In this context, the traditional teacher role is often assumed by an automated tutoring system. Ideally, this system's abilities should go beyond imparting knowledge and, like a real teacher, extend to the provision of response-specific personalized feedback that helps students uncover and fix errors in their current understanding. Such feedback can manifest as an answer to any of the questions in Figure 1. Each type serves a purpose and enriches the learning experience (Murphy, 2005). Should a student's answer suggest that they were not fully paying attention, a simple indication that an error has been made, i.e., verification feedback, might be sufficient to realign their focus. However, if they have major difficulties arriving at the correct answer or only miss small, less important details, one may wish to provide correction feedback instead. Unfortunately, it is unclear whether all students fully understand their mistakes after receiving the correct answer. In fact, Murphy (2007) observed that instead, they tend to consider a task finished and do not reengage with the materials. In contrast, elaborated feedback eliminates possible answer options without disclosing the correct answer, forcing students to actively reevaluate their understanding. It helps students become more independent by identifying and correcting their own mistakes (De Bot, 1996; Ferris, 2003). The resulting sense of achievement has been shown to enhance motivation, making the learning process more rewarding and entertaining (Bandura, 2013). Yet, elaborated feedback is seldom applied in automated tutoring systems since its generation requires complex considerations pertaining to a knowledge source, the incorrect student answer, and the relation between them. That is, information related to the answer needs to be identified and then filtered based on whether the student seems to be aware of it or not.



increasingly more detail
No. Try again!
No. It is Tom who stayed home.
No. They went swimming.
No. They went to the beach.
Verification
Is the answer correct or incorrect?
Explanation
Why is the answer incorrect?
Hint
What can I use to arrive at the correct answer?
Correction
What is the correct answer?
Elaborated Feedback

Figure 1: Possible feedback types (Glover and Brown, 2006; Demaidi et al., 2018). Given a friend group taking a trip to the beach, a student is asked about their location and incorrectly answers that they stayed at home.

*Work performed as a research assistant at [1].
†Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10838–10851
November 12-16, 2024 ©2024 Association for Computational Linguistics

Our contributions are as follows. We propose to employ a Kullback-Leibler (KL) regularization term between a standard and enriched input format to achieve more targeted input representations at inference time, promoting the identification of important parts of the input. The selective use of this presentation is encouraged by adding a preference optimization step that minimizes the entailment between the incorrect student answer and the generated feedback. Finally, we assess in how far a range of traditional metrics can accurately reflect human judgments of feedback quality and suggest a new metric to capture the degree to which feedback divulges the correct answer.

2 Related Work

Existing research on short-answer feedback generation focuses mainly on isolating and generating a single feedback type, i.e., correction feedback (Filighera et al., 2022b), explanation feedback (Olney, 2021; Filighera et al., 2022a), or hint feedback (Kazi et al., 2010; Kulshreshtha et al., 2022; Sychev, 2023; Jatowt et al., 2023). Some of those approaches use an external resource to retrieve relevant information prior to feedback formulation, such as machine-readable glossaries of textbooks (Olney, 2021), domain ontologies (Kazi et al., 2010) or Wikipedia entries (Jatowt et al., 2023). Others draw the necessary information from a correct answer template alone (Filighera et al., 2022a,b; Kulshreshtha et al., 2022). Once all necessary information is collected, feedback is generated using either a template-based approach with fill-in slots (Kazi et al., 2010; Jatowt et al., 2023), or more advanced language models. Specifically, Olney (2021) formulates feedback as the answer to a synthetic question asking, “What is the relationship between the student answer and the correct answer?” To generate its answer, they use the off-the-shelf long-form question-answering model ELI5 (Fan et al., 2019). At the same time, other authors employ more general pre-trained encoder-decoder models, i.e., T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), which they finetune on the target data. Filighera et al. (2022a) finetune T5 to predict a score or label and jointly explain it, while Kulshreshtha et al. (2022) finetune both T5 and BART to produce a question hinting towards the reasoning required to arrive at the answer.

As we intend to investigate feedback generation in the context of interactive reading comprehension exercises, the work most similar to ours is Huang et al. (2022). They build a dialogue-based tutoring system that can adapt questions to the dialogue context, asses learner answers and generate diverse feedback types. To generate feedback, the system is provided with the reading passage, an individually passed excerpt of the passage, i.e. key sentences, the dialogue history including tutor question and incorrect student answer, as well as the correct answer. This system is implemented by finetuning DistilGPT-2 (Sanh, 2019) on the DIRECT dataset.

3 Data

DIRECT (Huang et al., 2022) is a derivation of RACE-M (Lai et al., 2017), which is a large-scale English reading comprehension dataset. Its passages and multiple choice questions were specifically designed by domain experts to test the reading comprehension skills of Chinese middle school students. DIRECT enriches those exercises with annotations of key sentences, i.e., parts of the passage needed to answer a question. Then, it expands each exercise into a multi-turn dialogue, simulating the interaction between a student and tutor. The tutor’s role is to lead the conversation, ask questions about the passage, and provide feedback if the student answers incorrectly. The fictive student is assumed to answer incorrectly about half the time, resulting in 10,431 feedback turns. For DIRECT, incorrect answers were constructed by selecting one of the faulty answer options, which are often totally unrelated to the reading passage.

We decided to construct additional data with more natural answers, including mistakes that students are likely to make in an environment where only the reading passage and no answer options are provided. For each question in the DIRECT dataset, one annotator in the student role constructs such an answer, then another annotator in the tutor role constructs the corresponding feedback. Both annotators are presented with the reading passage, the question, its correct answer, and the corresponding key sentences. Five annotators with some level of English proficiency worked on the student role, while two native English-speaking annotators worked on the tutor role. The latter were also asked to periodically review randomly selected portions of the constructed data, including both incorrect answers and tutor feedback (constructed by the other worker). They ensured that the percentage of erroneous data items remained below 5%. This results

		DIRECT	DIRECT-F
Form	Declarative	100%	44%
	Interrogative	0%	34%
	Both	0%	22%
Type	Explanation	34%	30%
	Hint	40%	48%
	Correction	26%	22%

Table 1: Comparison between DIRECT and DIRECT-F.

in an additional 23,982 feedback turns.

We call this new dataset DIRECT-F[1]. To show differences and similarities to the DIRECT dataset, we manually analyze 50 randomly sampled feedback turns each. As shown in Table 1, both datasets cover a range of feedback types. Most common are hints, followed by explanations. Only about one in four feedback turns is correction feedback. One stark difference between the datasets is the form of feedback. Feedback in DIRECT is generally limited to a single declarative sentence, while DIRECT-F includes a range of declarative or interrogative sentences, and mixture thereof.

4 Problem Definition

Given a reading passage P, a question about its contents Q, and an incorrect student answer A^S, the goal is to generate personalized feedback F. Elaborated feedback, i.e., explanations and hints, is preferred over other feedback types. In this work, we do not draw from correct answer templates, assuming that the model can determine the correct answer independently and will learn more valuable representations along the way.

5 Model

5.1 Baseline

In our experiments, we employ the pretrained T5[2] encoder-decoder model (Raffel et al., 2020), which produced the best baseline performance across models of similar size, i.e., T5, BART, and GPT2. All three input types mentioned above are already quite familiar to T5. To emphasize this familiarity, we prepend each input type with the corresponding input type prefix used during T5 pretraining, that is, either "context:" or "paragraph:" for knowledge sources, "question:" for questions, and "answer:" for answers of any type. Additionally, we stick to conventions for dialogue modeling and put the knowledge source first. The resulting model input x_{wo_key} becomes "paragraph: P question: Q answer: A^S". The expected model output y is the corresponding feedback F.

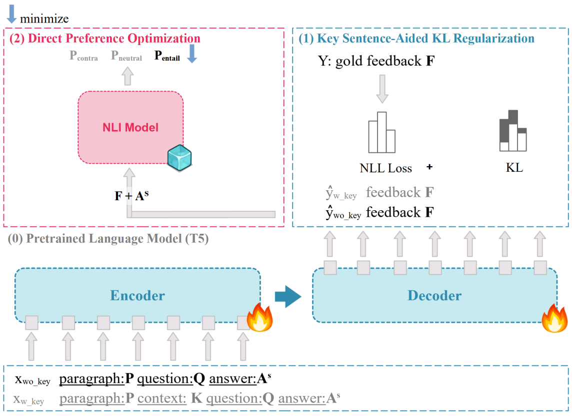
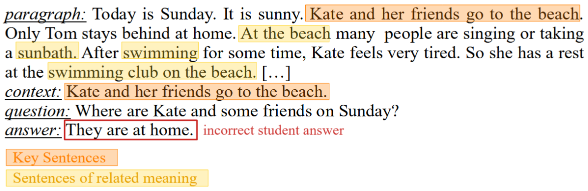

minimize
(2) Direct Preference Optimization
P contra
P neutral
P entail
NLI Model
F + A^s
(1) Key Sentence-Aided KL Regularization
Y: gold feedback F
NLL Loss
+
KL
ŷ w_key feedback F
ŷ wo_key feedback F
(0) Pretrained Language Model (T5)
Encoder
Decoder
x wo_key paragraph:P question:Q answer:A^s
x w_key paragraph:P context: K question:Q answer:A^s

Figure 2: Extended model training pipeline.


paragraph: Today is Sunday. It is sunny. Kate and her friends go to the beach. Only Tom stays behind at home. At the beach many people are singing or taking a sunbath. After swimming for some time, Kate feels very tired. So she has a rest at the swimming club on the beach. [...]
context: Kate and her friends go to the beach.
question: Where are Kate and some friends on Sunday?
answer: They are at home. incorrect student answer

Key Sentences
Sentences of related meaning

Figure 3: Example of enriched input format.

5.2 Key Sentence-Aided KL Regularization

It is rarely the entire text that is needed to answer a reading comprehension question and rather some specific part of it, i.e., the key sentences K. Feedback may take the form of shortening, reformulating or abstracting those sentences, making them an important source of information for feedback generation. The DIRECT dataset contains annotations of such key sentences. We will use those as anchors during model training to teach the model how to identify important parts of the reading passage independently during inference. This process is shown on the right in Figure 2. In the original annotations, key sentences with three or more sentences are abbreviated using a "~" between the first and last sentence. We expand those annotations to their whole form. Then, for each item, we formulate a second enriched model input x_{w_key} as "paragraph: P context: K question: Q answer: A^S". Both the basic input from the previous section and this new enriched input are passed to the model to compute the negative log-likelihood loss of the corresponding gold feedback, $L_{NLL}^{wo\_key}$ and

[1] https://github.com/DIRECTDataset/DIRECTFeedback
[2] https://huggingface.co/google-t5/t5-base

$L_{NLL}^{w\_key}$, respectively. For the enriched input, the model should soon start to focus its attention on the explicit key sentences. Ideally, if they are not specified explicitly, the model should look for parts of similar meaning in the provided passage (see Figure 3). To encourage this alignment, we add a KL regularization term on top of the output distribution (Gao et al., 2023). Hereby, we interpret the distribution produced by the input with key sentences $\hat{y}_{w\_key}$ as a constant, not propagating gradients, and force the output distribution produced by the input without key sentences $\hat{y}_{wo\_key}$ to emulate this distribution. The resulting training objective is:

$$L_{NLL}^{total} = 0.5 \times L_{NLL}^{wo\_key} + 0.5 \times L_{NLL}^{w\_key} + KL(\hat{y}_{wo\_key} || \hat{y}_{w\_key}) \tag{1}$$

5.3 Direct Preference Optimization

If good feedback could be provided independently of the student's answer, there would be no need for a complex generation model. We could simply use predefined feedback stored in a database together with each question. We will explain how such an approach can fail. Given again our example reading passage that describes a friend group going to the beach, sunbathing and swimming, a student might infer the location incorrectly if they missed or misunderstood the part where the beach was mentioned. However, the student might remember that the passage mentioned the friends swimming and mistakenly assume that they went to the pool. Now, if we were to provide the feedback "No. They went swimming.", we would support the student in their interpretation and provide no evidence whatsoever why they should doubt it, potentially leading to frustration and confusion. While such feedback is bad in the provided context, it may be perfectly appropriate given another student answer. For instance, if the student were to have answered that the group of friends spent the day "At home!", informing them that they, in fact, went swimming would provide new knowledge that does not align with the student's current understanding of the text.

We suggest that such a distinction between feedback being good or bad depending on the context can be captured using a natural language inference (NLI) model[3], shown on the left in Figure 2. For this purpose, we calculate the entailment relation between the student answer as a premise and a feedback option as the hypothesis (see Figure 4). Good feedback should necessarily include some information not entailed by the student's answer, resulting in a lower entailment probability. One problem observed for the naive baseline in Huang et al. (2022) was that the model showed a tendency to pick up wrong information from the student's answer and use it to construct feedback, sometimes copying whole parts verbatim. Such bad feedback among others would result in a higher entailment score. We wish to direct our model towards generating less such bad feedback.

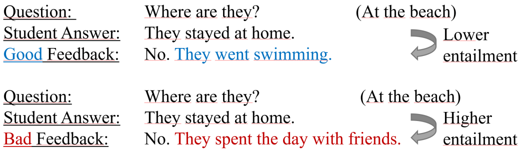


Figure 4: Example showing the importance of student answer-adaptive feedback.

For this purpose, we choose to employ direct preference optimization (DPO) (Rafailov et al., 2024) using the loss from Liu et al. (2023). For DPO, the preference learning stage occurs directly after an initial stage of supervised finetuning without the need to construct or consult an external reward model. Given the input to a finetuned model and a pair of preferred and dispreferred responses, DPO increases the relative log probability of the preferred to the dispreferred response. To construct such pairs, the finetuned model resulting from the previous section is applied to a second, smaller, separate training split. For each item, we sample five generations using top-p sampling with a p of 1.0. Then, we compute the degree of entailment for each generation. The feedback with the lowest and largest degree of entailment becomes the preferred and dispreferred response, respectively. The remaining generations are discarded. In order not to lose the gains from key sentence-aided KL regularization, we repeat this process twice, once with and once without key sentences in the input, i.e, $L_{DPO}^{w\_key}$ and $L_{DPO}^{wo\_key}$. Should the difference in entailment probability between the preferred and dispreferred response for the input without key sentences be less than 0.1, we discard the whole item. The loss for preference optimization is then calculated as:

$$L_{DPO}^{total} = 0.5 \times L_{DPO}^{wo\_key} + 0.5 \times L_{DPO}^{w\_key} + KL(\hat{y}_{wo\_key} || \hat{y}_{w\_key}) \tag{2}$$

[3] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

KL	DPO	BLEU	METEOR	ROUGE	BERTScore
-	-	4.6 (+-0.20)	18.2 (+-0.26)	19.6 (+-0.17)	16.5 (+-0.40)
+	-	5.0* (+-0.14)	19.1** (+-0.33)	20.4** (+-0.22)	17.3* (+-0.33)
-	+	6.3** (+-0.20)	20.9** (+-0.23)	20.8* (+-0.23)	18.1* (+-0.30)
+	+	**6.8** (+-0.43)	**21.5*** (+-0.45)	**21.4*** (+-0.28)	**18.8*** (+-0.25)

Table 2: Ablation test results as averaged across 5 random seeds. Brackets show one standard deviation. Asterisks mark values that are significantly better than the next best value (one sided paired t-test, **p<0.01, *p<0.05).

6 Experimental Setup

The combined DIRECT and DIRECT-F dataset is divided into train, validation, and test set, following the split of the RACE-M dataset. RACE-M assigns each passage with all its questions to precisely one split. During training, the model will see neither the questions nor passages that are later used for evaluation. This makes the test set especially hard to master. Additionally, we save 5% of the original train set to be used for the optimization step. Concrete data statistics can be found in Appendix A. We limit the input length to 768 tokens. Should the input exceed this limit, the end of the paragraph is truncated until the rest of the input fits the limit. The output is limited to 256 tokens. At inference time, we always pass the basic input without explicit key sentences and decode using beam search (n=5) and a length penalty of 2.0, which rewards longer generations. Training parameters are described in Appendix C. Each model configuration was trained on five randomly chosen seed values and test set performance averaged across those. We report scores for SacreBLEU (Papineni et al., 2002; Chen and Cherry, 2014; Post, 2018), METEOR (Lavie and Denkowski, 2009), ROUGE-L (Lin, 2004) and BERTScore[4] (Zhang et al., 2019). All metrics are computed using the Hugging Face *evaluate* library. For metrics that require an external tokenizer, i.e., ROUGE and METEOR, we apply NLTK's *word_tokenize*. The results are reported in Table 2.

7 Results I

Compared to the T5 baseline, we observe significant performance improvements across all four traditional metrics for both proposed extensions. The absolute improvement varies across metrics between 1.8 for ROUGE to as much as 3.3 points for METEOR. Notably, DPO alone leads to an increase of 2.7 METEOR points, while KL regularization achieves a smaller but still significant (p<0.01) increase of 0.9 points. Those results underline the effectiveness of our two proposed extensions. We will call the resulting full model *ReCTify*, i.e., to correct something or make something right. Generation examples are shown in Appendix B.

Since we have extended both the train and test set of the previously available data, the results above cannot be straightforwardly compared to related work. To make up for that, we retrain and evaluate our full model on the original DIRECT dataset. Huang et al. (2022) used this data to train an end-to-end tutoring system DiReCT. While this system constitutes the only existing work on feedback generation for reading comprehension, feedback is only one of a number of possible turn types it can generate. This setup likely causes a performance bottleneck. Our full model outperforms this system by 5.9 METEOR points.

	B	M	R	BS
DiReCT (Huang et al.)	5.1	24.5	22.3	22.3
ReCTify (Ours)	8.9	30.4	25.3	26.2

Table 3: Comparison with related work on DIRECT.

8 Analysis

KL regularization was introduced to obviate the need to pass key sentences during inference. But how effective is it in comparison to passing gold key sentences? To investigate this question, we train a second baseline model that uses the enriched input format with explicit key sentences. Evaluation results are shown in Table 4. Although slightly higher, none of the scores significantly differ from the KL-only model. This underlines the effectiveness of the proposed KL extension.

BLEU	METEOR	ROUGE	BERTScore
5.3	19.2	20.6	17.7

Table 4: Upper bound for KL regularization.

[4] microsoft/deberta-xlarge-mnli with default baseline rescaling

DPO was employed to encourage the model to enrich feedback with information that does not align with a student's current understanding of the text. However, it remains unclear whether the NLI reward model actually assigns entailment probabilities in the intended way. The following investigation should answer this question.

Fist, we employ the NLI model to assign each of the 1538 gold feedback instances in the smaller DPO train set a hard label out of {*entailment*, *neutral*, *contradiction*}. Based on the assigned labels we draw a random stratified sample of size 210, i.e. 70 instances per class. Then, two human annotators are shown each feedback instance together with the corresponding wrong student answer, question and reading passage. The annotators are instructed to assign *contradiction*, if the information provided by the feedback does not seem to align with the student's current understanding of the text, *entailment*, if it does align, and *neutral*, if the feedback provides independent or unrelated information. Annotators are encouraged to use the reading passage and question to fill in missing information wherever necessary, e.g. pronoun resolution. We observe a moderate inter-annotator agreement κ of 0.54. In the following we will use only those 146 instances that the two annotators agreed upon.

		Model		
		Entail	Neutral	Contra
Human	Entail	**25**	11	1
	Neutral	5	**23**	13
	Contra	20	12	**36**

Table 5: Confusion matrix on NLI judgement.

If operating as intended, the NLI model should assign *entailment* to feedback that the human annotator identified to align with student's current understanding of the text. The human annotators identified 37 such feedback instances, a large portion of which has indeed also been identified by the NLI model (first row). During the DPO step the model is taught to correctly avoid such feedback. However, among the 50 instances that the NLI model identifies to be entailed are surprisingly also 20 instances that were assigned the category *contradiction* by the human annotators (first column). We took a closer look at those instances and found most of them to be verification feedback following the pattern in the example given in Figure 5.

Question: What kind of sports does Liu Yingying's mother like? (*Answer:* She likes swimming.)
Student Answer: She likes tennis.
Generated Feedback: She likes a different sport.

Figure 5: Generation example of the KL-only model.

We assume that the NLI model assigns *entailment* in this example because it was not trained to see the hypothesis as a response to the premise. Thus it fails to resolve that "a different sport" means "not tennis" , instead interpreting it like "some sport". After the DPO step the above feedback becomes: "The father likes tennis, but she likes a different sport." The feedback has become more explicit in including an explanation of where the student's misunderstanding might come from. It is no longer entailed by the wrong student answer. The above investigation shows that the DPO step indeed fulfills its intended objective (first row), but goes beyond this objective to also target inexplicit verification feedback (first column).

9 Proposed Metrics

9.1 Motivation

Traditional evaluation metrics, like METEOR, are extensively used in feedback generation research, especially during model development (Huang et al., 2022; Filighera et al., 2022a; Kulshreshtha et al., 2022). However, their focus lies primarily on surface-level lexical overlap, which may not fully capture overall feedback quality, e.g., rewarding fluency over other important quality dimensions. Good feedback should not only be fluent but also strike the right balance between revealing too much and too little of the correct answer and be in line with the corresponding source material. A more in-depth evaluation of those dimensions could provide a better insight into the model's ability to produce useful feedback and help uncover its strengths and weaknesses. In the following two sections, we will first introduce two new metrics designed to directly capture feedback informativeness and faithfulness. Both have not yet been considered in the context of feedback evaluation. Then, we will investigate the suitability of BLEU, METEOR, ROUGE-L, and BERTScore to capture those quality dimensions in comparison to the newly introduced metrics.

9.2 Informativeness Index I^2

Depending on several factors like question difficulty or student level, a tutor may decide to provide more or less detailed feedback. Hereby, striking the right balance between revealing too much and too little of the correct answer is crucial. An optimal system should provide feedback that promotes active self-reflection and deeper comprehension while still offering enough support to prevent discouragement and confusion. To capture this notion, we propose to use an answer verification approach. Among the tasks used during T5[5] pretraining is the MultiRC task. For this task, T5 needs to predict whether an answer to a question is *True* or *False* given multi-sentence evidence. We suggest that the normalized likelihood p it assigns to the *True* label can be understood as the degree of support the evidence provides. So, we pass the feedback as evidence and let the model decide to what degree it supports the correct answer to a given reading comprehension question. For example, the degree of support for "They went swimming." in Figure 4 would be 0.65. As mentioned above, the optimal degree of informativeness depends on many factors. A higher or lower score is neither always good nor always bad. Therefore, in order to interpret the score, we need to compare it to some target value. For this purpose, we calculate the above probability, once for the generated feedback p and once for the gold feedback p_E. The difference between the two can then be understood as the information appropriateness of the generated feedback. For now, we ignore the direction of this difference, calling the resulting measure *Informativeness Index* I^2:

$$I^2 = 1 - |p - p_E| \qquad (3)$$

This value falls within the range of 0 to 1, with 1 being the best value.

9.3 Faithfulness

Informativeness is a necessary but not sufficient condition for good feedback. Good feedback should also be ankered in the corresponding source material or at least not contradict it. This quality dimension is commonly researched under the term *Faithfulness* in the field of automatic summary evaluation. Here, pre-trained NLI models are used to assess the probability of the generated summary being entailed by the source article (Falke et al., 2019). In place of a summary, we pass the generated feedback. To limit the calculation of I^2 and *Faithfulness* to a single model, we again use T5 as the NLI model. The reading passage is passed as the premise, and the feedback as the hypothesis. Once we obtain a probability for both entailment and contradiction, we calculate a *Faithfulness F* score as:

$$F = \frac{1 + min(p_{entail}, 0.5) - p_{contra}}{1.5} \qquad (4)$$

One known problem with NLI-based approaches is that extractive generations are rewarded to a much higher degree than abstractive ones. To buffer this effect, we clip the entailment probability at 0.5. Finally, we normalize the term to fall within the range of 0 to 1, with 1 being the best value.

9.4 Overall

While we recommend reporting and interpreting both measures separately, considerations presented in the next chapter ask for a single overall score. We choose to compute the weighted average as:

$$Ours = \frac{3}{4} \times I^2 + \frac{1}{4} \times F \qquad (5)$$

The weightings in Equation 5 were chosen in response to the effective value ranges observed across training epochs and parameter settings. While corpus-level I^2 values mostly fell in the range of 0.65 to 0.75, F values varied between 0.4 to 0.7.

10 Verification of Evaluation Metrics

10.1 Data

To verify this metric, we employ a private data set of feedback rankings originally prepared for a different study. For each feedback turn in the DIRECT development set (n=475), automatic feedback was generated using: GPT-4[6], GPT-3.5[7], DiReCT (Huang et al., 2022) and PrepTutor, which is a version of DiReCT, further finetuned on out-of-domain feedback data. GPT-4 and GPT-3.5 used the prompt in Appendix D. Model-specific output characteristics can be found in Table 6. In the next step, a single human expert ranked the five feedback options according to their quality (4 models, 1 human), penalizing feedback with incorrect information more than feedback that discloses the

[5] https://huggingface.co/google-t5/t5-large

[6] GPT-4-0613, Accessed 01. November 2023

[7] GPT-3.5-turbo-0613, Accessed 01. November 2023

correct answer. The best feedback should do neither. As such, the quality estimation in this data set aligns with our two dimensions of interest, i.e., faithfulness and informativeness. Feedback of identical quality could be assigned the same rank, but the same rank could not be assigned to every option. Even though the generations were provided using a bland evaluation format without disclosing the source model and in random order, the human rankings show a preference for GPT4 over GPT3.5, which correctly aligns with other work comparing the performance of those two models. Thus, we consider the ranking results fairly reliable even though only a single annotator was involved. Finally, we discard any ranking where the human feedback is not among the top 2 and adapt the remaining rankings (n=339) to hold only automatic feedback. The human feedback will be used as a reference to compute automatic evaluation metrics and create automatic rankings based on the obtained scores.

	GPT4	GPT3	PrepT.	DiRe.	Gold
Length	25	20	36	12	13
Vocab	1426	1138	1479	826	1056

Table 6: Average feedback length and unique words.

10.2 Correlation at Model Level

Automatic evaluation metrics are often applied to select the best-performing model (i.e., parameter configuration or epoch). For this purpose, the focus is more on overall trends rather than individual instances. To investigate how effectively each metric captures this trend, we count how often each model ranks first in human and automatic rankings, normalizing this count by the total number of instances. If all metrics could correctly capture notions of faithfulness and informativeness, the resulting automatic distributions should resemble the human one. All distributions are shown in Figure 6. Humans preferred feedback produced by GPT-4 for every second instance, followed by GPT-3 with a preference rate of 38%. The lightweight models, PrepTutor and DiReCT, were preferred less than 5% of the time. Despite this clear trend, traditional metrics strongly prefer the DiReCT model over GPT-4, likely due to its similarity to human references in length and vocabulary choice. Metrics like BLEU, ROUGE-L, and BERTScore even preferred GPT-3 over GPT-4, highlighting their inability to recognize more faithful and informative output not only across different but also similar models. Only METEOR correctly favored GPT-4 by a margin of 4%. Notably, our proposed metric matches human preference nearly perfectly, correctly ranking GPT-4 first, followed by GPT-3, with PrepTutor and DiReCT trailing far behind.

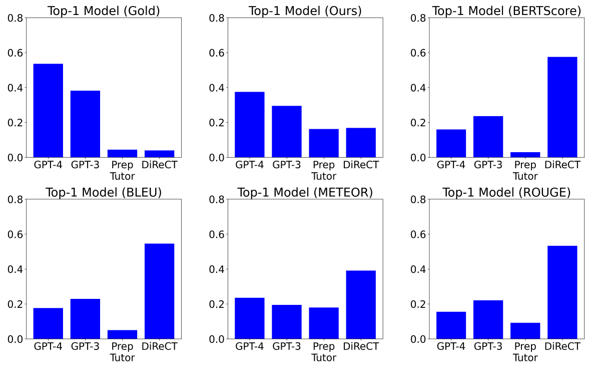


Figure 6: Overall model preference.

10.3 Correlation at Instance Level

While metrics are rarely employed to compare single instances, ultimately, the results at the model level are an artifact of those smaller pieces. So, we calculate Spearman's Rank Correlation between metric- and human-assigned rankings for each of our 339 instances. Existing research advised against averaging such correlations directly and instead recommends applying Fisher's transformation beforehand, followed by the inverse transformation at the end (Corey et al., 1998). Using Fisher's transformation, the distribution of correlations becomes approximately normal, creating a scale where averaging is more appropriate. We report the naive average r_s and the Fisher average r_z in Table 7.

	B	M	R	BS	Ours
r_s	0.18	0.08	0.09	0.20	**0.33**
r_z	0.34	0.20	0.17	0.34	**0.52**

Table 7: Spearman's Rank Correlation between automatic evaluation metrics and human judgment.

Our metric achieves the highest correlation of 0.52. For further insight, we analyze how often each model was assigned a rank better or worse than its actual rank. We normalize this count by the number of times each scenario was theoretically possible, i.e., if the gold rank is 1, it is impossible for a metric to assign a better rank. Results in Fig-

KL	DPO	Faithfulness	I^2	$I^2_{p>pE}$	$I^2_{pE>p}$
-	-	61.2 (+-2.0)	73.8 (+-0.5)	73.5 (+-0.5) [879]	74.1 (+-1.0) [934]
+	-	59.4 (+-1.7)	**74.3 (+-0.4)**	**73.7 (+-0.5) [904]**	74.7 (+-0.5) [895]
-	+	**62.8 (+-3.1)**	71.7 (+-0.4)	68.9 (+-0.4) [943]	74.7 (+-0.8) [857]
+	+	62.6 (+-3.6)	72.0 (+-0.7)	69.0 (+-1.4) [955]	**75.2 (+-0.2) [843]**

Table 8: Results as averaged across 5 random seeds. Curvy brackets show one standard deviation.

ure 7 show that traditional metrics exhibit strong biases towards specific models. Among them, METEOR shows the most negligible bias despite a surprisingly low degree of correlation. This observation could be explained if we consider the strong imbalance of the rankings. Traditional metrics like BLEU and BERTScore show a strong negative bias towards the PrepTutor model. This bias could benefit their correlation score, as, by chance, the PrepTutor model's actual rank is usually very low. Finally, our proposed metric is shown to be nearly bias-neutral.

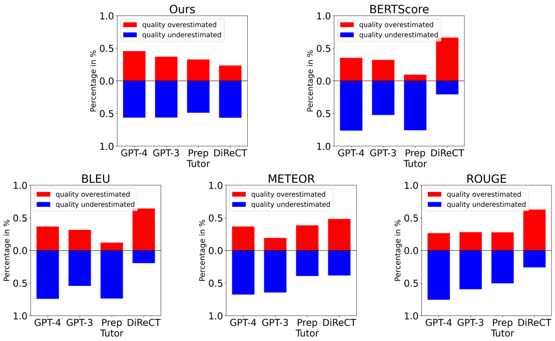


Figure 7: Bias across metrics. In the absence of bias, there should be as much blue as red area.

11 Results II

We now apply I^2 and F to our proposed model. $I^2_{p>pE}$ considers only generations that are excessively informative, while $I^2_{pE>p}$ considers those that lack informativeness. The number of instances falling in each category is given in square brackets. With each extension, the number of generations that lack important, helpful information decreases, and even those that still lack information get increasingly more helpful, as indicated by a consistent rise in $I^2_{pE>p}$. At the same time, the overall faithfulness increases, but not as consistently. The drop in faithfulness when applying only the KL regularization term might be due to varying input formats that make it harder for the model to understand that the provided student answer holds wrong and not additional information.

Our proposed extensions are targeted towards providing more helpful information and they clearly fulfill this role. However, they are not targeted towards limiting this information accordingly. Therefore, as the model enriches feedback, it starts divulging too much information, as indicated by a large drop in $I^2_{p>pE}$. To investigate the true severity of this problem, we display the distributions of p_E and p for our full model in Figure 8. About 27% of human feedback scores 0.8 or higher, i.e., an indicator of correction feedback. With about 34%, our model tends to produce significantly more correction feedback. Still, it produces not only such feedback but hints of differing specificity as indicated by the distribution.

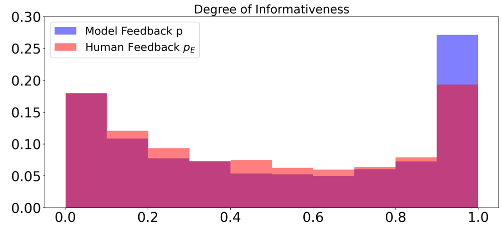


Figure 8: Distribution showing the degree of informativeness for human feedback and model feedback.

12 Conclusion and Future Work

We proposed two extensions to the naive training pipeline to improve a feedback generation model's input utilization. To help identify essential parts in the source material, we proposed employing a KL regularization term between a basic and enriched input format. This step results in an improvement of 0.9 METEOR points. Then, in order to teach the model how to filter information, we added a preference optimization step. This step results in further gains of almost 2.4 METEOR points. All in all, proving the effectiveness of our proposed extensions. Using the newly introduced measure of informativeness I^2, we identify a remaining weakness of the full model, i.e., a tendency to produce correction feedback. In the future, I^2 could be used in tandem with DPO to counteract this tendency.

Limitations

Compared to traditional metrics that show significant differences across model configurations, differences in I^2 or *Faithfulness* are mostly not significant because they are either too small or the variation across seeds too large. This limits their applicability in model evaluation, as they can hardly be used to prove that any configuration is the superior model. However, they are still useful to get insight into the current strengths and weaknesses of a model. This also highlights a need for further refinement, for instance, by replacing the underlying model. We initially chose T5-large because it provides sufficient prediction quality on the MultiRC task while maintaining some level of uncertainty. During development, we also experimented with the T5-3b model, which is reported to have a much better accuracy on the relevant task. However, we observed that T5-3b often assigned probability values close to 0 or 1 with little variation. This lack of variation is problematic as it is necessary to differentiate between feedback that reveals the correct answer and feedback that only slightly alludes to it. In the future, one might directly finetune a model to capture the degree of information in the feedback, e.g., using a regression-based approach. As for *Faithfulness*, much higher-quality models are available and should be considered in the future.

Acknowledgments

This work was supported by Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

References

Albert Bandura. 2013. The role of self-efficacy in goal-based motivation. *New developments in goal setting and task performance*, pages 147–157.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

David M Corey, William P Dunlap, and Michael J Burke. 1998. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *The Journal of general psychology*, 125(3):245–261.

Kees De Bot. 1996. The psycholinguistics of the output hypothesis. *Language learning*, 46(3):529–555.

Mona Nabil Demaidi, Mohamed Medhat Gaber, and Nick Filer. 2018. Ontopefege: Ontology-based personalized feedback generator. *IEEE Access*, 6:31644–31664.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

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*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Dana R Ferris. 2003. *Response to student writing: Implications for second language students*. Routledge.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022a. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8577–8591, Dublin, Ireland. Association for Computational Linguistics.

Anna Filighera, Joel Tschesche, Tim Steuer, Thomas Tregel, and Lisa Wernet. 2022b. Towards generating counterfactual examples as automatic short answer feedback. In *International Conference on Artificial Intelligence in Education*, pages 206–217. Springer.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

Chris Glover and Evelyn Brown. 2006. Written feedback for students: too much, too detailed or too incomprehensible to be effective? *Bioscience education*, 7(1):1–16.

Jin-Xia Huang, Yohan Lee, and Oh-Woog Kwon. 2022. Direct: Toward dialogue-based reading comprehension tutoring. *IEEE Access*, 11:8978–8987.

Adam Jatowt, Calvin Gehrer, and Michael Färber. 2023. Automatic hint generation. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 117–123.

Hameedullah Kazi, Peter Haddawy, and Siriwan Suebnukarn. 2010. Leveraging a domain ontology to increase the quality of feedback in an intelligent tutoring system. In *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10*, pages 75–84. Springer.

Devang Kulshreshtha, Muhammad Shayan, Robert Belfer, Siva Reddy, Iulian Vlad Serban, and Ekaterina Kochmar. 2022. Few-shot question generation for personalized feedback in intelligent tutoring systems. In *11th Conference on Prestigious Applications of Artificial Intelligence, PAIS 2022, co-located with the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022*, pages 17–30. IOS Press BV.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*.

Philip Murphy. 2005. Interactivity, usability and flexibility in an online reading course. *Language Education Research*, 16:383–433.

Philip Murphy. 2007. Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning & Technology*, 11(3):107–129.

Andrew M Olney. 2021. Generating response-specific elaborated feedback using long-form neural question answering. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pages 27–36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Oleg Sychev. 2023. Open-answer question with regular expression templates and string completion hinting. *Software Impacts*, 17:100539.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data Statistics

Detailed data statistics can be found in Table 9

B Generation Examples

In Table 10, we show a randomly picked sample of generations produced by our full model *ReCTify* using both KL regularization and DPO finetuning.

C Training Parameters

For supervised finetuning, we use the AdamW optimizer with a weight decay of 0.1 and a learning rate of 5e-4, which is adjusted using a linear scheduler with 10% warmup. The model is trained over 10 epochs with an effective batch size of 128. The validation set is used to determine the best model using the following mixture of metrics:

$$m = \frac{8}{12} \times METEOR + \frac{3}{12} \times I^2 + \frac{1}{12} \times F \quad (6)$$

The weights are assigned to account for diverging effective value ranges and not to be understood as an indicator of metric importance.

For preference optimization, we again use the AdamW optimizer with a weight decay of 0.1 but a lower learning rate of 1e-5. We use a linear scheduler with a warm-up that gradually increases the learning rate over 50 steps before decreasing it again to zero. The model is trained over 200 steps with an effective batch size of 32.

D GPT4 and GPT3 Prompt

You are a proficient tutoring assistant who provides just a few clues to an user in the correct direction. The user should understand the following passage and then answer your question.

Passage: {passage}

The correct answer is "{correct answer}", but the user don´t answer correctly as the following tutoring dialogues. Generate an indirect feedback or hint to guide the user to find the answer on him/her own.

{student & teacher dialogue}

		Train (SF)	**Train (DPO)**	**Validation**	**Test**	**Total**
DIRECT	Passages	4813	253	301	301	5668
	Feedback	8967	464	475	525	10431
DIRECT-F	Passages	4841	254	302	307	5704
	Feedback	20389	1074	1239	1280	23982
Total	Passages	4841	254	302	307	5704
	Feedback	29356	1538	1714	1805	34413

Table 9: Data statistics for DIRECT and DIRECT-F.

Question	**Key Sentences**	**Correct Answer**	**Incorrect Answer**	**Gold Feedback**	**Generated Feedback**
What title will you give to this story?	Full text	Then, I'll choose "Great love".	I'll choose "An Ear Operation".	Think again. You can find a better answer.	It's wrong. That's a little off the topic. Can you think again?
This is a story about an artist. What kind of person is he?	An artist who did not have much money, but was a very kind man, He gave his last few coins to a beggar,	He is a kind man.	He is a funny man.	It's not right. He gave his last few coins to a beggar.	Is he a very kind man or a selfish man?
Why does Jack's wife leave him?	His wife tells him to stop drinking, but he doesn't listen to her. She has to leave him	She leaves him because he doesn't stop drinking.	It's because he doesn't do housework at home.	It's not correct. Can you find the right answer again?	That's not it. Jack likes drinking and spends most money on drinks.
Why did the boy's mother never cut her hair?	To his surprise, the son found his mother had no outer ears. " Mother said she was glad she never needed to cut her hair, " his father said in a low voice, " and nobody ever thought mother less beautiful, did they? "	She never had cut her hair, because she wanted to keep the fact a secret.	I can't find the reason.	It's related to the reason the boy got surprised after seeing his mother's body.	The boy's father told the boy that his mother didn't need to cut his hair because his mother didn't have ears.
Some students stayed in their seats. Why did they do it?	Everyone who chooses not to take the final exam will get a 'B' for the course. " All the students there stayed in their seats and took out their pencils	Because they wanted to take the exam.	Because they were afraid to leave.	Well, it's not. They took out their pencils to take the final.	No, it's related to the pencils.

Table 10: Generation examples of our full model *ReCTify*.