

# Predicting the Quality of Revisions in Argumentative Writing

Zhexiong Liu<sup>1</sup>, Diane Litman<sup>1,2</sup>, Elaine Wang<sup>3</sup>, Lindsay Matsumura<sup>2</sup>, Richard Correnti<sup>2</sup>

<sup>1</sup>Department of Computer Science

<sup>2</sup>Learning Research and Development Center

University of Pittsburgh, Pittsburgh, Pennsylvania 15260 USA

<sup>3</sup>RAND Corporation, Pittsburgh, Pennsylvania 15213 USA

zhexiong@cs.pitt.edu, ewang@rand.org

{dlitman, lclare, rcorrent}@pitt.edu

## Abstract

The ability to revise in response to feedback is critical to students' writing success. In the case of argument writing in specific, identifying whether an argument revision (AR) is successful or not is a complex problem because AR quality is dependent on the overall content of an argument. For example, adding the same evidence sentence could strengthen or weaken existing claims in different argument contexts (ACs). To address this issue we developed Chain-of-Thought prompts to facilitate ChatGPT-generated ACs for AR quality predictions. The experiments on two corpora, our annotated *elementary essays* and existing *college essays* benchmark, demonstrate the superiority of the proposed ACs over baselines.

## 1 Introduction

Argumentative Revision (AR) in response to feedback is important for improving the quality of students' written work. Successful ARs<sup>1</sup> usually include adding relevant evidence, deleting repeated evidence or reasoning, and elaborating relevant evidence examples to support claims (Afrin et al., 2020). Differentiating between successful versus unsuccessful ARs, however, is a complex endeavor. For example, making the same AR in distinct Argumentative Contexts (ACs) could differentially affect the quality of a student's essay. Here the ACs are defined as pieces of sentences that present reasons, evidence, and claims supporting or opposing arguments in argumentative writing (see Sec. 4.2). For example, Figure 1 shows two pieces of ARs that added the same sentence “*it was hard for them to concentrate though, as there was no midday meal*” but caused opposite AR quality.

Recently developed Automated Writing Evaluation (AWE) systems have focused on assessing the content and structure of student essays to automatically provide students with formative feed-

<sup>1</sup>Afrin and Litman (2023) use the term desirable revisions.

### AR #372: Unsuccessful Revision

<original draft> **They also did not concentrate good because they did not have lunch over there.**

According to the text, many kids in Sauri did not attend school because their parents could not afford school fees. </original draft> <adding> **It was hard for them to concentrate though, as there was no midday meal.** </adding>

### AR #592: Successful Revision

<original draft> In 2010 the schools had minimal supplies like books, paper, and pencils, but the students wanted to learn. All of them worked hard with few supplies they had. </original draft>

<adding> **It was hard for them to concentrate though, as there was no midday meal.** </adding>

Figure 1: Two pieces of ARs in two student essays show that *adding* the same sentence “*it was hard for them to concentrate though, as there was no midday meal*” (bold in red) in different contexts caused opposite AR quality. AR #372 added a piece of evidence that *already existed* in the original draft (bold in blue) thus the attempted AR did not improve the essay quality. AR #592 improved the quality by adding a *relevant* piece of new evidence. AR #372 was *unsuccessful* while AR #592 was *successful*.

back (Zhang et al., 2016; Writing Mentor, 2016; Wang et al., 2020; Beigman Klebanov and Madnani, 2020). Successful revisions (e.g., adding relevant evidence) improve an essay's quality. Unsuccessful revisions, in contrast, lead to no improvement or can even weaken an essay's argument (Afrin et al., 2020). As a result, assessing the success of ARs is important to assess the quality of ARs in line with provided feedback.

AR quality has previously been predicted by using long and short neighboring contexts of ARs (Afrin and Litman, 2023). This location-based approach for constructing ACs did not exploit any argumentative relationships between ARs and potential ACs. Another study (Zhang and Litman,

2016) incorporated AR contexts with cohesion blocks and employed sequence labeling to model AR interdependence across revisions. This work predicted AR purposes from discourse structures but did not further study AR quality or analyze AR quality from the perspective of ACs. To bridge these gaps, we address three research questions. **RQ1:** To what extent are ACs helpful for predicting AR quality? **RQ2:** What type of AC is the most helpful in AR quality predictions? **RQ3:** Can ChatGPT prompts be used to generate useful ACs? In studying the three RQs, we have made the following contributions:

- Our project is the first in the revision field to analyze the relationship between ACs and AR quality predictions.
- We are among the first to incorporate the state-of-the-art large language model ChatGPT in generating ACs in argumentative writing.
- Experiments using both elementary and college essay corpora show the superiority of the proposed ACs over existing location-based contexts for AR quality predictions.

## 2 Related Work

### 2.1 Argumentative Revision in NLP

Revision research has been conducted using multiple types of Natural Language Processing (NLP) corpora ranging from Wikipedia to argumentative essays. While argumentative writing research has analyzed argumentative roles and discourse elements in persuasive writing (Stab and Gurevych, 2014; Song et al., 2020; Putra et al., 2021) (e.g., by studying the stance towards some topic, backing up claims, or following argumentative and rhetorical considerations), such analyses have not typically been applied to revision research in this domain. Revision research, in contrast, has primarily focused on grammar correction, paraphrasing, semantic editing (Yang et al., 2017), and analyzing revision purposes (Zhang and Litman, 2015; Shibani et al., 2018; Afrin et al., 2020; Kashefi et al., 2022). Although revision research has sometimes leveraged contextualized features during classification, the contextual features have been location-based (Zhang and Litman, 2016; Afrin and Litman, 2023). We instead extract contextual information from an essay based on argumentative essay analysis rather than on adjacency to a revision.

### 2.2 LLM in Argumentative Revision

Large Language Models (LLMs) have scaled up model sizes from a few million to hundreds of billions of parameters. Their strong capabilities of handling multiple downstream NLP tasks have made LLMs favorable in recent research (Chowdhery et al., 2022). Prior revision works, e.g., academic writing (Ito et al., 2019), debate assessment (Skitalinskaya et al., 2021), paraphrase generation (Mu and Lim, 2022), were mostly based on Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) models but not the cutting-edge LLMs, e.g., ChatGPT<sup>2</sup>. Pretrained LLMs have shown strong few-shot learning capabilities by way of developing prompts to guide LLMs in generating successful outputs (Brown et al., 2020; Liu et al., 2023). For example, Chain-of-Thought (CoT) prompts (Wei et al., 2022) enable pretrained LLMs to solve complex reasoning problems by decomposing the tasks into a series of intermediate steps. Kojima et al. (2022); Wang et al. (2022) investigated the effectiveness of CoT in multi-step reasoning, however, little work has used CoT for extracting and then generating tasks in the revision field. In this work, we leverage ChatGPT with CoT prompts to generate ACs in argumentative writing.

## 3 Corpora

### 3.1 Data Collection

AR corpora are rarely annotated in the revision community because of their expensive annotation costs. The publicly available *college essay* corpus for AR quality predictions (Afrin and Litman, 2023) contains paired drafts of argumentative essays written in response to an essay prompt (original drafts) and revised based on feedback (revised drafts). The corpus is comprised of 60 essays (N=60 college students), inclusive of both native and proficient non-native speakers of English, in response to an essay prompt about Technology Proliferation. Students received general feedback upon completion of their first drafts, asking them to add more examples in their second drafts. The second drafts then received non-textual feedback through the ArgRewrite system (Zhang et al., 2016) to help students write their third drafts. Afterward, the second and third drafts were collected as pairs of original and revised drafts.

We followed a similar protocol to collect 596 *ele-*

<sup>2</sup><https://openai.com/blog/chatgpt>

		Elementary Essays			College Essays		
		Reasoning	Evidence	Total	Reasoning	Evidence	Total
Successful	Add	769	671	1440	104	23	127
	Delete	213	104	317	7	1	8
	Modify	129	104	233	3	0	3
Unsuccessful	Add	360	491	851	87	2	89
	Delete	102	147	249	6	0	6
	Modify	74	103	177	0	0	0
Total	/	1647	1620	3267	207	26	233

Table 1: Sentence-level AR quality annotation statistics on *elementary* and *college essays*.

ID	Original Draft Sentence	Revised Draft Sentence	Revision Type	Revision Purpose	Quality Label	
1	According to the text. "The people in Sauri have made amazing progress in just eight years."	According to the text, "The people in Sauri have made amazing progress in just eight years."	Modify	Surface	N/A	N/A
2	This tells me that the people of Sauri have made better living arrangements in eight years and it all did pay off.	This tells me that the people of Sauri have made better living arrangements in eight years and it all did pay off, from how they used to live.	Modify	Content	N/A	N/A
3	The people of Sauri did do great progress.	The people of Sauri did do great progress.	N/A	N/A	N/A	N/A
4	...	...	...	...	...	...
5	This lets me know that since there might be a few diseases that might affect anyone at ant time so the hospital has made medicine that can cure those diseases, so they gave that medicine to any one who needed it for free.		Delete	Content	Irrelevant Evidence	Unsuccessful
6		This piece of text lets me know that the hospital, Yala Sub District, has free medicine for diseases that are most common around where they live.	Add	Content	Paraphrase Reasoning	Successful
7		In Sauri people had to pay a fee, which the people of Sauri couldn't afford.	Add	Content	not LCE Reasoning	Unsuccessful
8	...	...	...	...	...	...
9	I can tell that the people of Sauri must of thought that children needed education money not so they didn't ask people for the school fees, and the kids wouldn't go hungry during school hours they served the children lunch.		Delete	Content	LCE Reasoning	Successful
10	...	...	...	...	...	...

Table 2: Example of revision annotations for an *elementary essay*. Note that the successful and unsuccessful labels in the last column are only used for evidence and reasoning content revisions; other purpose types in (Zhang et al., 2017) are not in the scope of this study as we only focus on evidence use and reasoning in argumentative writing.

*mentary essays* written by grade 5 to 6 students who were taking the Response to Text Assessment (Correnti et al., 2013). 296 students wrote an essay in

response to a prompt about the United Nation's Millennium Villages Project (MVP). The students then revised their essays in response to formative feed-

back from an Automatic Writing Evaluation (AWE) system that used rubric-based algorithms to assess the quality of evidence use and reasoning (Zhang et al., 2019; Wang et al., 2020). The other 300 students did the same tasks for an essay prompt about Space Exploration (Space). We combined the collected essays from the two essay prompts because students shared similar argumentative writing skills and the scoring rubric and feedback messages were constant across prompts.

### 3.2 Preprocessing

We preprocessed collected *elementary essays* for annotations. First, sentences from original and revised drafts were aligned into pairs of original sentence (OS) and revised sentence (RS) using a sentence alignment tool Bertalign (Liu and Zhu, 2022). The aligned pairs were programmatically labeled with *no change* if OS and RS are the same, *modifying* if OS and RS are not empty but not same, *adding* if OS is empty but RS not, or *deleting* if RS is empty but OS not. The changed alignments were automatically classified into *surface* and *content* revisions by a pretrained classifier. Note that the sentence alignments and classification were first done by the system and then manually justified and corrected by annotators, and only aligned *content* revisions were used for annotations.

### 3.3 Annotations

We used the Revisions of Evidence use and Reasoning (RER) scheme (Afrin et al., 2020) to annotate revision purposes in *elementary essays*, which encodes the nature of students’ revision of evidence use and reasoning. Evidence use refers to the selection of relevant evidence from a given source article to support a claim, while reasoning means a reasoning process of connecting the evidence to the claim. Thus, the *content* revisions are annotated with claim-related, evidence, and reasoning revisions. The RER scheme only applies to evidence and reasoning, where evidence revisions were labeled with *relevant*, *irrelevant*, *repeated evidence*, *non-text based* and *minimal*, and reasoning revisions were labeled with *linked claim-evidence (LCE)*, *not LCE*, *paraphrase evidence*, *generic*, *commentary*, and *minimal*.

Furthermore, we followed the AR quality scheme (Afrin and Litman, 2023) to programmatically encode annotated RER labels (revision purposes) into *successful* and *unsuccessful* revisions. The *relevant* evidence was encoded as *successful*

	Space Essays		MVP Essays	
	RER#	Kappa	RER#	Kappa
Reasoning	148	0.86	135	0.84
Evidence	108	0.89	136	0.80

Table 3: Annotation agreement for reasoning and evidence RER annotations in a sample of 20 percent of *elementary essays* regarding Space and MVP prompts.

while the *repeated*, *non-text based*, and *minimal* evidence were encoded as *unsuccessful*. The *LCE* and *paraphrase* reasoning were encoded as *successful*. The *not LCE*, *paraphrase evidence*, *generic*, *commentary*, and *minimal* reasoning were encoded as *unsuccessful*. Table 1 shows label distributions in *elementary essays* and *college essays* where *elementary essays* have almost an even number of reasoning and evidence annotations. The *adding* revisions are the most frequent ARs across two essays. Samples of annotations for *elementary essays* and *college essays* are shown in Table 2 and Table 6 (in Appendix A), respectively. In practice, the RER annotations were done by one expert annotator. We sampled about 20 percent of annotated essays about both Space and MVP prompts and asked another well-trained annotator to annotate the sampled essays. The two-annotator Kappa scores are shown in Table 3.

## 4 Methods

### 4.1 Preliminary

In this section, we introduce notations for the AR quality prediction task. We denote  $R_1$  and  $R_2$  as original and revised sentences in the original and revised drafts, respectively. In particular,  $R_1$  is always empty in *adding* ARs (e.g., row #6 in Table 2);  $R_2$  is always empty in *deleting* ARs (e.g., row #9 in Table 2); neither  $R_1$  nor  $R_2$  are empty in *modifying* ARs (e.g., row #1 in Table 2). Thus, we only use  $R_1$  in *deleting* and  $R_2$  in *adding* ARs. In terms of *modifying* ARs, we only use  $R_2$  because  $R_2$  is a revised version of  $R_1$  thus are very close to  $R_1$  (e.g., row #2 in Table 2). In addition, we denote ACs as a couple of sentences related to ARs in their corresponding drafts, where  $C_1$  represents the ACs of  $R_1$  in the original draft and  $C_2$  represents the ACs of  $R_2$  in the revised draft (details in Sec. 4.2), respectively. To this end, we formulate the task of predicting AR quality as classifying the AR-AC pairs  $\{R_i, C_i\}$  into *successful* and *unsuccessful* labels, where  $i = 1, 2$ . Specifically, we use pair



$\{R_1, C_1\}$  for *deleting* and  $\{R_2, C_2\}$  for *adding* and *modifying* ARs.

## 4.2 Argumentative Context

Consistent with long-established models of argumentation such as Toulmin’s model (Toulmin, 1958), well-developed arguments are characterized by the alignment of claim, evidence, and warrants (i.e., reasoning related to why the evidence supports the claim) (Reznitskaya et al., 2008). For example, the appropriateness of a piece of evidence for advancing an argument is context-dependent because that judgment is determined relative to an author’s prior claim(s) or reason(s). As a case in point, the unsuccessful AR #372 shown in Figure 1 would have been unobservable absent an understanding of the author’s claim or argument’s context. Recent work by Afrin and Litman (2023) has used short and long text segments immediately before and after the AR as context for predicting AR quality, however, the study has some significant drawbacks. First, the window size of the contexts is an unpredictable parameter because a reasoning sentence could refer to the evidence far ahead of the AR (e.g., reasoning in row #6 refers to the evidence in row #1 in Table 2). Second, location-based contexts did not explain why ACs make a difference to ARs from an argumentative perspective and thus fail to analyze the argumentative roles of ACs in AR quality predictions. As we noted above, the evaluation of a reasoning sentence as desirable depends on whether it appropriately references evidence or claims in the student’s essay, but this relationship has not been explored in prior revision research. Thus, in the current study, we define three ACs to study their relationship to AR quality: **(1) AC-Claim:** the context containing essay claims or arguments; **(2) AC-Reasoning:** the context containing reasoning related to the claim or evidence in the essay; **(3) AC-Evidence:** the context containing evidence to support or oppose claims.

## 4.3 ChatGPT Prompts

Pretrained ChatGPT on a series of GPT3.5 models has shown promising results in solving information extraction (Li et al., 2023) and summarization (Yang et al., 2023) tasks in zero-shot settings, however, doing the two tasks at the same time has not been explored in generating ACs. Therefore, we developed two versions of ChatGPT prompts that generate useful ACs for predicting AR quality: (1) Single prompts that generate ACs in one pass

and (2) Chain-of-Thought prompts that generate ACs in two passes.

### 4.3.1 Single Prompts

In this section, we introduce Single prompts for AC generations. Basically, we need ChatGPT to generate useful ACs for AR quality predictions by reading the student essays. We limit the generation to a two-sentence length for two reasons. First, the generated ACs will be used in an AR-AC pair  $\{R_i, C_i\}$ , where  $R_i$  is normally one sentence, thus long ACs ( $C_i$ ) paired with short ARs ( $R_i$ ) will make the AR quality prediction model (introduced in Sec. 4.4) learn to attend to the context rather than the revisions. Second, the most intuitive location-based baseline (Base-Short in Sec. 5) uses the adjacent sentences before and after target ARs, which contain at most two sentences. Therefore we limit the generations to exact two sentences, which can be done with a single zero-shot prompt please summarize [X] in the essay [Y] in two sentences, where [X] slot is one of the *claim*, *reasoning*, and *evidence*, [Y] is an input essay.

### 4.3.2 Chain-of-Thought Prompts

In addition to Single prompts, Chain-of-Thought (CoT) prompts (Wei et al., 2022) are conceptually simple yet effective in multiple reasoning tasks. We adopt this idea and use zero-shot-CoT prompts to generate ACs, which run prompting in two passes but do not require step-by-step few-shot examples.

**The first-pass CoT prompt to extract ACs.** The first pass of the CoT prompts is to extract claim, evidence, and reasoning sentences from input essays. We aim to extract exact sentences from input essays without introducing any external knowledge in ChatGPT itself. The first-pass prompt is: please list [X] sentences in the essay [Y], where [X] slot is chosen from one of the *claim*, *reasoning*, and *evidence*, and [Y] is an input essay. The extracted ACs are formulated as a list of sentences from the input essays, where the length of the list ranges from one to dozens because *claim*, *reasoning*, and *evidence* have multiple instances in an essay. To make sure the extracted ACs are informative and not exceeding the word limits of DistilRoBERTa encoders (see Sec. 4.4), we perform summarization on the extracted sentences in the second-pass prompt.

**The second-pass CoT prompt to summarize ACs.** The second prompt is continued to the first

<p><b>Input Essay #104 (First Draft)</b></p> <p>No the author did not convince me that winning the fight against poverty is achievable in our lifetime. I just think that it would be nearly impossible because there are a lot of people dying and they don't have too much money. In the text it states that, "The solutions are simple, yet 20,000 kids are die from the disease each day. So sad and so illogical. Bed nets could save millions of lives." If there are that many people dying a day, then how will the doctors have time to make cures or anything? All of the solutions could take as long as more that 10 years. That would be really hard for all of he doctors to be working so hard, but the solutions to take for ever. In the text it states that , "The goals are supposed to be met by 2025 ;some other targets are met by 2035." By the time it is 2025, a lot of the people will be sick and they can die if they don't get treated fast enough. The passage says there were at lest 5 people in 1 bed, and some people would not get hardly any treatment, and some of the patients could not afford theses things. I think that it would be nearly impossible to do this mission because many people are dying a day and the doctors still need some more time. The doctors might get to save a few people depending on how many there are, but they will not get save many. In conclusion, you may have a to of solutions to your problem, but you will not always have the answers in time that you would want.</p>
<p><b>Generated AC-Claim</b></p> <p><b>Single Prompt:</b> Please summarize claim in the essay in two sentences.</p> <p>[Output] The author fails to convince the writer that winning the fight against poverty is achievable in their lifetime due to the high daily death toll and the lengthy time frame for solutions. The passage highlights the dire state of poverty with overcrowded hospitals and inadequate treatment options, ultimately underscoring the difficulty in finding effective answers.</p> <p><b>CoT Prompt:</b> 1) Please list claim sentences in the essay. 2) Please summarize the listed claim in two sentences.</p> <p>[Output] The author failed to convince the reader that ending poverty is feasible due to the high death rate and long time frame for implementing solutions. The present systems are inadequate, leading to overcrowded hospitals and lack of access to medical care.</p>
<p><b>Generated AC-Reasoning</b></p> <p><b>Single Prompt:</b> Please summarize reasoning in the essay in two sentences.</p> <p>[Output] The author did not persuade the writer that poverty can be eradicated within our lifetime due to the high mortality rate of 20,000 people living in poverty dying every day. The solutions proposed may take too long and inadequate healthcare results in many people not receiving the necessary treatment.</p> <p><b>CoT Prompt:</b> 1) Please list reasoning sentences in the essay. 2) Please summarize the listed reasoning in two sentences.</p> <p>[Output] The author's argument that winning the fight against poverty is unlikely is supported by the fact that 20,000 people die from poverty every day and some proposed solutions will take years to implement. The passage also highlights the dire conditions and lack of access to treatment for many impoverished individuals.</p>
<p><b>Generated AC-Evidence</b></p> <p><b>Single Prompt Output:</b> Please summarize evidence in the essay in two sentences.</p> <p>[Output] The author fails to convince the reader that winning the fight against poverty is achievable in our lifetime due to the high rate of poverty-related deaths and the lengthy timeframe for implementing solutions. The overcrowding in hospitals and lack of access to treatment further supports the argument that successful poverty eradication may not be possible.</p> <p><b>CoT Prompt Output:</b> 1) Please list evidence sentences in the essay. 2) Please summarize the listed evidence in two sentences.</p> <p>[Output] Despite goals set to be met by 2025 and 2035, poverty-fueled deaths continue at a rate of 20,000 per day, with potential solutions taking more than 10 years to implement. As a result, many are left without access to healthcare and overcrowded hospitals, with some unable to afford or receive necessary treatment.</p>

Figure 2: The input and output of the ChatGPT with zero-shot Single and CoT prompts for an *elementary essay*.

prompt, following an extraction-summarization CoT. The prompt is `please summarize [X] in two sentences`, where [X] slot is chosen from the claim, reasoning, and evidence sentences extracted in the first prompting pass, which ensures the outputs in a length of exact two sentences. Figure 2 and Figure 5 (in Appendix A) show examples of the zero-shot Single and CoT prompts that help ChatGPT generate ACs in *elementary* and *college essays*, respectively.

#### 4.4 AR Quality Prediction

We define AR quality prediction as a binary classification of AR-AC pair  $\{R_i, C_i\}$  (see Sec. 4.1), where  $R_i$  is annotated and  $C_i$  is generated by ChatGPT. Prior works employed BERT-BiLSTM architecture to train revision classifiers (Anthonio and Roth, 2020; Afrin and Litman, 2023). Instead,

we use DistilRoBERTa (Sanh et al., 2019) as text encoders for both annotated ARs and ChatGPT-generated ACs. The last hidden layers of the DistilRoBERTa encoders are fed to an average-pooling layer and then connected to a multi-layer perception classifier that contains a sequence of batch normalization layer, ReLU layer, dropout layer with a 0.5 rate, and Sigmoid layer. The overall framework is shown in Figure 3.

## 5 Experiments

To answer the RQs, we implemented location-based contexts as baselines, and a series of ACs as comparable methods:

- **Base-Short:** We implement a standard revision prediction baseline that uses the adjacent sentences immediately before and after a revision as

Contexts	Prompts	Reasoning & Evidence ARs			Reasoning ARs			Evidence ARs		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Base-Short	N/A	67.79	67.17	67.29	70.29	70.01	69.94	63.42	62.60	62.67
Base-Long	N/A	68.45	68.01	68.06	69.99	69.76	69.71	65.38	64.69	64.71
AC-Claim	Single	<u>70.31</u>	<u>69.83</u>	<u>69.91</u>	<u>72.63</u>	<u>72.47</u>	<u>72.38</u>	66.60	65.33	65.57
AC-Reasoning	Single	<u>70.15</u>	<u>69.67</u>	<u>69.74</u>	72.10	71.95	71.83	66.57	65.60	65.79
AC-Evidence	Single	<u>70.28</u>	<u>69.97</u>	<u>69.93</u>	<u>72.46</u>	<u>72.31</u>	<u>72.13</u>	66.55	65.87	65.85
AC-Claim	CoT	<u>70.09</u>	69.64	69.74	71.83	71.76	71.70	66.74*	65.71*	65.88*
AC-Reasoning	CoT	<b>71.14*</b>	<b>70.81*</b>	<b>70.81*</b>	<b>72.86*</b>	<b>72.80*</b>	<b>72.63*</b>	<b>68.00*</b>	<b>67.00*</b>	<b>67.16*</b>
AC-Evidence	CoT	<u>70.43*</u>	<u>70.03*</u>	<u>70.01*</u>	<u>72.48*</u>	<u>72.34*</u>	<u>72.20*</u>	66.76*	66.06*	66.05*

Table 4: Experimental results on *elementary essay* corpus. The bold numbers are the best results. The underlined numbers statistically outperformed the strong (Base-Long) baseline in a paired t-test with  $p < 0.05$ . The asterisks indicate zero-shot-CoT prompts are better than zero-shot-Single prompts.

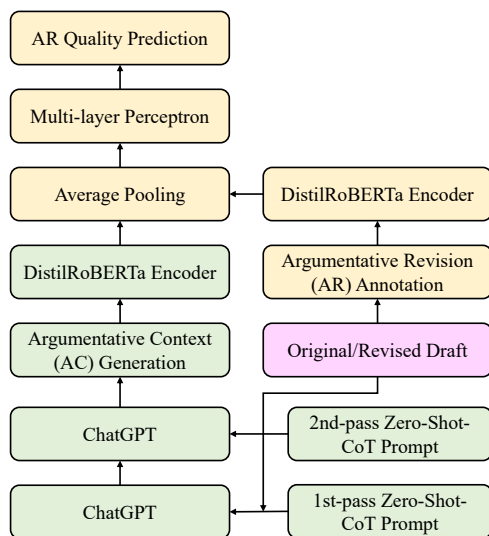


Figure 3: The overall framework of AR quality predictions, where the pink box is input; the green boxes are our proposed; the yellow boxes are existing methods.

contexts (Afrin and Litman, 2023).

- **Base-Long:** We implement a strong baseline that considers all the sentences that are revised around a target revision until an unchanged sentence is found (Afrin and Litman, 2023).
- **AC-Claim:** We use AC-Claim as the contexts that are generated by zero-shot Single and CoT prompts, respectively.
- **AC-Reasoning:** We use AC-Reasoning as contexts. The two versions use zero-shot Single and CoT prompts, respectively.
- **AC-Evidence:** We use AC-Evidence as contexts. The two versions are generated by zero-shot Single and CoT prompts, respectively.

In the implementation, we built the framework pipeline with PyTorch<sup>3</sup> and generated two versions

<sup>3</sup><https://pytorch.org>

Contexts	Prompts	Precision	Recall	F1
Base-Short	N/A	59.95	60.06	58.61
Base-Long	N/A	61.21	61.60	59.48
AC-Claim	Single	63.50	64.08	62.27
AC-Reasoning	Single	63.06	63.62	61.86
AC-Evidence	Single	<u>65.76</u>	66.40	<u>64.71</u>
AC-Claim	CoT	64.01*	64.96*	62.93*
AC-Reasoning	CoT	63.84*	64.33*	62.74*
AC-Evidence	CoT	<b>68.20*</b>	<b>68.05*</b>	<b>66.32*</b>

Table 5: Experimental results on reasoning ARs in *college essays*. The bold numbers are the best results. The underlined numbers statistically outperformed the strong (Base-Long) baseline in a paired t-test with  $p < 0.05$ . The asterisks indicate zero-shot-CoT prompts are better than zero-shot-Single prompts.

of ACs using ChatGPT3.5-turbo API<sup>4</sup>. We used pretrained DistilRoBERTa-Base from Huggingface<sup>5</sup> as text encoders, and optimized cross-entropy loss with Adam optimizer on a GeForce RTX 3090 GPU. We set the batch size as 16 and the learning rate as  $5e-5$  with 5% decays every 4 epochs. We conducted 10-fold cross-validation, where 80% of each 9-fold set was used for training, 20% for parameter tuning, and the rest 1-fold set for testing. Finally, we ran the ChatGPT generation and the experiment pipeline three times and reported 3-seed-average macro Precision, Recall, and F1 on all the test sets. The implementation code is available at <https://github.com/ZhexiongLiu/Revision-Quality-Prediction>.

## 6 Results and Discussion

Table 4 shows the experimental results for different sets of revisions from the *elementary essay* corpus: all reasoning and evidence revisions, just

<sup>4</sup><https://platform.openai.com>

<sup>5</sup><https://huggingface.co>

reasoning revisions, and just evidence revisions, respectively. We observed that both the proposed Single and CoT versions of ACs outperformed both baselines, with many of the CoT ACs significantly better than the strong (Base-Long) baseline. This answered **RQ1** that ACs can help AR quality predictions. In reasoning ARs, excellent performance was yielded in using AC-Claim, AC-Reasoning, and AC-Evidence. This is because reasoning ARs might need claims to verify their usefulness and incorporate evidence and reasoning to check their relevance. Moreover, evidence ARs achieved the best with AC-Reasoning, which makes sense that identifying evidence AR requires related reasoning contexts that have information linking the evidence. Another interesting finding is that the Base-Long performed better than the Base-Short in evidence ARs but worse in reasoning ARs. This suggests that the longer context is not always helpful in the case that evidence contexts are usually sparsely distributed in the essay so the longer context will introduce more noise. It also suggests that reasoning sentences are mostly adjacent to other reasoning contexts and can be well captured by neighboring sentences. Furthermore, the observation that reasoning ARs results are generally better than evidence ARs indicates that reasoning ARs might be self-justifiable which means it might require fewer contexts than the evidence to identify AR quality. These observations answered **RQ2** that reasoning contexts are mostly useful, and both reasoning, claim, and evidence contexts benefit AR quality predictions. In addition, CoT prompts are generally better than Single prompts in most reasoning and evidence ARs, which indicates that identifying AR quality requires some contexts that might not be generated with Single prompts. This answered **RQ3** that CoT prompts are generally better than Single prompts.

We also evaluated the effectiveness of ACs on the *college essay* benchmark. Note that [Afrin and Litman \(2023\)](#) conducted data augmentation with a simple synonym replacement because they argued that it was impossible to obtain reasonable results without training on augmented data. We hypothesized that data augmentation will introduce noise but the limited data can yield reasonable results training with the DistilRoBERTa-based model. Therefore, we did not do data augmentation and compared AC-based methods to our implemented standard and strong baselines on reasoning revisions

#### Revision #372: AC-Claim

<claim> The essay highlights progress made in Sauri, including free medicine and bed net provision, as well as the positive impact of **providing lunch to children, to argue that poverty can be reduced...** </claim>

#### Revision #592: AC-Claim

<claim> The essay argues that poverty and lack of resources can be tackled in our lifetime with examples such as bed nets to prevent malaria, updated hospitals to prevent the spread of diseases and access to education... </claim>

Figure 4: Two pieces of ChatGPT-generated AC-Claims. The red bold is the context to identify Revision #372 is a *already existed* adding, while #592 is a *relevant* adding toward their contexts in Figure 1.

sions (excluding the rare evidence revisions as shown in Table 1). Results in Table 5 show that the DistilRoBERTa model is able to learn from even small-size data without data augmentation. In addition, AC-based methods perform better than both the standard and strong baselines, where AC-Evidence has significant improvement. This again suggests that ACs are generally useful for predicting AR quality and CoT prompts are generally better than Single prompts for generating useful ACs. Moreover, we observed that AC-Evidence generated by Single and CoT prompts is better than the other ACs. It is slightly different from the reasoning column in Table 4. This might suggest that revisions in *college essays* may focus on evidence revisions that match generated evidence ACs. However, *claim* and *reasoning* results have similar F1 scores across two versions of prompts, which might suggest the extracted AC-Claim and AC-Reasoning are similar in *college essays* (e.g., prompting outputs in Figure 5 in Appendix A), which might be because *college essays* have claim and reasoning sentences disentangled. In general, CoT prompts are somewhat better than Single prompts in AC-Claim and AC-Reasoning generation, and both Single and CoT prompts are promising in AC-Evidence generation.

As a case study, we examine the effectiveness of ACs in Revision #372 and #592 presented in Figure 1. The ChatGPT-generated AC-Claim is shown in Figure 4, where the red bold sentence “*providing lunch to children, to argue that poverty can be reduced*” is helpful to identify that the added sentence, “*It was hard for them to concentrate though,*



as there was no midday meal.” in Revision #372 is a *already existed* evidence, and thus it was an *unsuccessful* revision. However, AC-Claim in Revision #592 does not show repeated but *relevant* information, and thus the AR is regarded as *successful*.

## 7 Conclusion

This work studies the relationship between Argumentative Contexts (ACs) and Argumentative Revisions (ARs) in argumentative writing. In particular, we use zero-shot-CoT prompts to facilitate ChatGPT-generated ACs for AR quality predictions. The experiments on our *elementary essays* corpus and publicly available *college essays* benchmark demonstrate the superiority of the proposed ACs over existing location-based context baselines, which proposes a new direction for predicting AR quality. The analysis suggests that most evidence ARs need reasoning ACs, and reasoning ARs need a diverse set of claims, evidence, and reasoning ACs to predict their quality.

## 8 Limitations

Our experiments were built on perfect sentence alignments in the original and revised essay drafts, thus the performance could be lower in the real end-to-end Automated Writing Evaluation (AWE) system. In addition, our corpus is small due to expensive annotation processes, which makes it challenging to train or finetune large language models. Also, we only focus on revisions in argumentative writing, specifically, we focus on the evidence and reasoning revisions, however other revisions like claim revisions are not used. Furthermore, the revised drafts were done after providing feedback on the original drafts, which means the revised student essays are likely to follow the instructions in the feedback but we did not use this information for revision quality predictions, which will be used in our future work.

Our proposed Argumentative Contexts (ACs) are generated by ChatGPT which is not free for the whole community. Also, ChatGPT-generated ACs have small randomness, which is also the reason we did 3-seed runs in the experiments. In addition, the ACs are essay-level context which means different revisions in the same essay use the same context. It could be tailored to have sentence-level ACs where each sentence-level revision has slightly different revision purposes, but it would cost more

time and money. Moreover, our proposed zero-shot-CoT prompts perform better than Single prompts by small margins in specific cases, which indicates that Chat-GPT is limited to conducting CoT extraction and summarization to handle complex wording and sentence structure. Therefore, we might need to redesign the prompts in our future work.

## 9 Ethics

Our corpus was collected under standard protocols that were approved by an institutional review board. Our annotated data is not publicly available which ensures the safety of private information of the students, and thus will not pose any ethical concerns because other researchers can not access our data and replicate our results. Our future work is to incorporate proposed methods in real AWE systems to evaluate student writings and provide informative feedback based on predictions. But there is a risk that the system might give poor advice based on incorrect AR quality predictions, given that the model may learn biases with small annotated data.

## Acknowledgments

The research was supported by the National Science Foundation Award #2202347 and a gift from CloudBank. The opinions expressed were those of the authors and did not represent the views of the institutes. We would like to thank anonymous reviewers and Pitt PETAL group for their valuable feedback on this work.

## References

- Tazin Afrin and Diane Litman. 2023. [Predicting desirable revisions of evidence and reasoning in argumentative writing](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2550–2561, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tazin Afrin, Elaine Lin Wang, Diane Litman, Lindsay Clare Matsumura, and Richard Correnti. 2020. Annotation and classification of evidence and reasoning revisions in argumentative writing. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Seattle, Washington, USA (Remote).
- Talita Anthonio and Michael Roth. 2020. [What can we learn from noun substitutions in revision histories?](#) In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1359–1370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Beata Beigman Klebanov and Nitin Madnani. 2020. [Automated evaluation of writing – 50 years and counting](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students’ skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.
- 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. [Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 40–53, Tokyo, Japan. Association for Computational Linguistics.
- Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. [Argrewrite v.2: an annotated argumentative revisions corpus](#). *Language Resources and Evaluation*, pages 1574–0218.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *NeurIPS*.
- Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt’s information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633.
- Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Wenchuan Mu and Kwan Hui Lim. 2022. [Revision for concision: A constrained paraphrase generation task](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 57–76, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Parsing argumentative structure in english-as-foreign-language essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–109.
- Alina Reznitskaya, Richard C Anderson, Ting Dong, Yuan Li, Il-Hee Kim, and So-Young Kim. 2008. *Learning to think well: Application of argument schema theory to literacy instruction*. The Guilford Press.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Antonette Shibani, Simon Knight, and Simon Buckingham Shum. 2018. Understanding revisions in student writing through revision graphs. In *International Conference on Artificial Intelligence in Education*, pages 332–336, Cham. Springer International Publishing.

- Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. [Learning from revisions: Quality assessment of claims in argumentation at scale](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.
- Wei Song, Ziyao Song, Ruiji Fu, Lizhen Liu, Miaomiao Cheng, and Ting Liu. 2020. Discourse self-attention for discourse element identification in argumentative student essays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2820–2830.
- Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.
- Stephen Edelston Toulmin. 1958. *The Uses of Argument*. Cambridge university press.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.
- Elaine Lin Wang, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. *erevis(ing): Students’ revision of text evidence use in an automated writing evaluation system*. *Assessing Writing*, 44:100449.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.
- The Writing Mentor. 2016. ETS writing mentor, <https://mentormywriting.org/>, [online; accessed 02-06-2019].
- Diyi Yang, Aaron Halfaker, Robert E. Kraut, and Edward H. Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP’17*, pages 9–11, Copenhagen, Denmark. Association for Computational Linguistics.
- Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *ArXiv*, abs/2302.08081.
- Fan Zhang, Homa B Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578.
- Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B Hashemi. 2016. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations*, pages 37–41.
- Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.
- Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430.
- Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsumura, Emily Howe, and Rafael Quintana. 2019. *erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9619–9625.

## A Appendix

ID	Original Draft Sentence	Revised Draft Sentence	Revision Type	Revision Purpose	Quality Label	
1	A mother who would have no other way of reaching her children can easily speak to them or leave a message via voicemail.	A mother who would have no other way of reaching her children can easily speak to them or leave a message via voicemail.	N/A	N/A	N/A	
2	Technology makes it possible to reach anyone at any time.		Delete	Content	LCE Reasoning	Unsuccessful
3		In addition, technology makes it possible to increase the amount of communication between people drastically.	Add	Content	LCE Reasoning	Successful
4	...	...	...	...	...	...
5	People from different continents who may have never met before can now have conversations every day; even those from a remote location are available to the world, provided they have the Internet.	People from different continents who may have never met before can now have conversations every day; even those from a remote location are available to the world, provided they have both the Internet and a corresponding device.	Modify	Surface	N/A	N/A
6		How could a cold inanimate screen replace seeing the emotions and expressions of a loved one?	Add	Content	not LCE Reasoning	Unsuccessful
7		An essential thing to consider is that while perhaps it may be harder to convey one's full message complete with feelings through the Internet, the fact remains that in a changing world where people are busier and farther away, electronic devices are helping everyone keep in contact with each other at any time of the day and at any location.	Add	Content	LCE Reasoning	Successful
8		Those who argue for the retardation of technology simply cannot accept that the world is developing to be more tech driven; as more and more people have electronic devices, they are also changing to be more used to this kind of communication.	Add	Content	not LCE Reasoning	Unsuccessful
9	...	...	...	...	...	...

Table 6: Example of revision annotations for a *college essay*.



**Input Essay #107 (First Draft)**

Throughout history, society has undergone advances in many realms of life. These realms include politics, social issues, education, and also technology, which is one of the most notable. The proliferation of technology has included the creation of computers, which can be used to research new information in a short period of time. Technology also allows people to send and receive information, using devices such as cellular phones. Because of this ease of information transfer, technology has been able to enrich interpersonal relationships by providing another avenue for communication. Technology has enriched interpersonal relationships by allowing communication between people when they are not in physical vicinity. For example, students in college may develop friendships during the school year, but during the summer they may be unable to keep in physical contact with these friends depending on how far they live from one another, as many students go home for the summer. Because verbal communication is an important part of maintaining relationships, this distance could pose a threat to the maintenance of these relationships. However, staying in contact via texts, calls, and social media gives students the opportunity to remain in contact, even if the distance between them is large. This communication lets all parties know that even while not in the vicinity of one another, they are still able to enjoy each other and have similar interests. Because of this, the relationship has the potential to be even stronger upon their return back to school. Technology also enriches interpersonal relationships by increasing the frequency with which people can communicate. Due to technology, people can communicate several times throughout the day because of the speed of information transfer. For Example, if someone has an issue that they would like to relay to another person, they are able to notify them almost instantly due to the speed of communication. Additionally, they are able to discuss the matter back and forth. This can lead to greater conflict resolution as it allows the issue to be resolved faster, rather than waiting until the next time the parties involved are in contact with each other. The removal of this time-lapse between contact is beneficial as it could eliminate the emergence of more negative emotions. In sum, contact without physical vicinity, as well as, increased frequency of communication between people displays how the proliferation of technology has been able to increase the strength of interpersonal relationships.

**Generated AC-Claim**

**Single Prompt:** Please summarize claim in the essay in two sentences.

[Output] The proliferation of technology has enriched interpersonal relationships by providing another avenue for communication. This is due to the ease of information transfer and increased frequency of communication through devices like cellular phones, which allow people to stay in touch even when physically apart.

**CoT Prompt:** 1) Please list claim sentences in the essay. 2) Please summarize the listed claim in two sentences.

[Output] Technology has improved interpersonal relationships by providing more ways to stay in touch over long distances. Frequent communication through various technologies increases the bond between individuals.

**Generated AC-Reasoning**

**Single Prompt:** Please summarize reasoning in the essay in two sentences.

[Output] Technology has enriched interpersonal relationships by allowing communication between people when they are not physically together. it increases the frequency and speed of communication, which can lead to conflict resolution and the strengthening of relationships.

**CoT Prompt:** 1) Please list reasoning sentences in the essay. 2) Please summarize the listed reasoning in two sentences.

[Output] The use of technology has led to advancements in politics, social issues, education, and communication. technology has enriched interpersonal relationships by allowing for frequent and efficient communication, which has led to stronger relationships with fewer negative emotions.

**Generated AC-Evidence**

**Single Prompt Output:** Please summarize evidence in the essay in two sentences.

[Output] Advancements in technology have allowed for easy information transfer and communication through devices like computers and phones, enriching interpersonal relationships. communication without physical vicinity and increased frequency of contact have increased the strength of relationships, aiding conflict resolution and promoting positive emotions.

**CoT Prompt Output:** 1) Please list evidence sentences in the essay. 2) Please summarize the listed evidence in two sentences.

[Output] Technology has led to advancements in areas like politics, education, communication and social issues. it allows for quicker research and access to new information, easier communication and interpersonal relationships.

Figure 5: The input and output of the ChatGPT with zero-shot Single and CoT prompts for a *college essay*.