Appendix to Muteract: Interactive Prompt Mutation Interface for LLM Developers and Evaluators

ACM Reference Format:
. 2018. Appendix to Muteract: Interactive Prompt Mutation Interface for LLM Developers and Evaluators. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

Appendix

A A Description of the User Interface

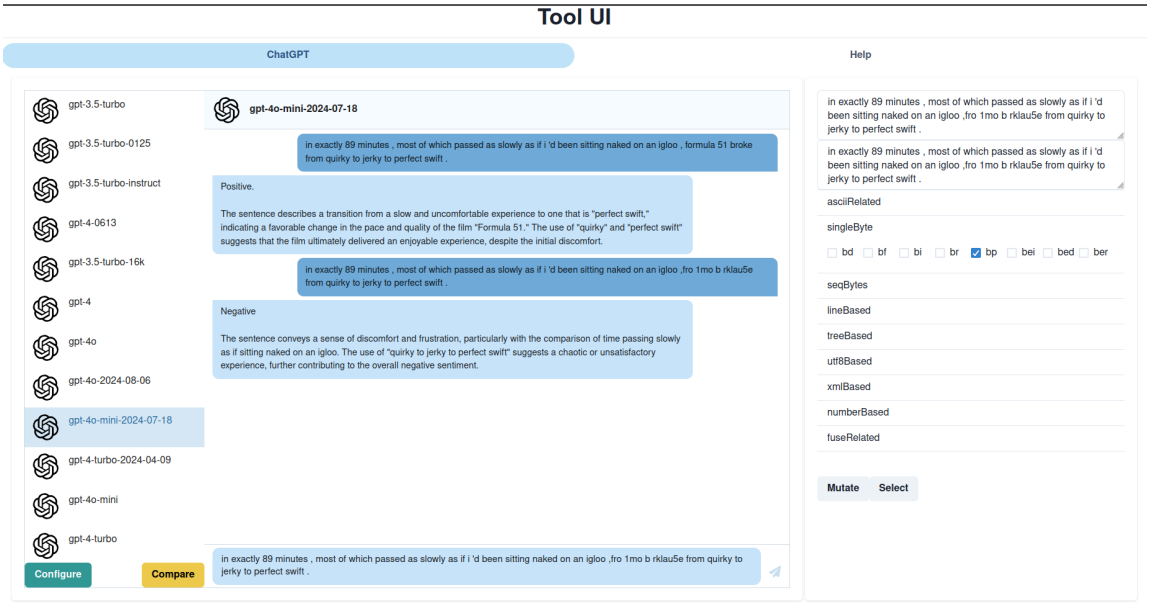

Fig. 1. The User Interface of Muteract

With reference to Fig 1, the user interface is composed of three main components: First, the **conversation panel**, located on the left side of the interface, enables users to select which model they wish to engage with. This panel features a list of models that is updated directly from the OpenAI APIs at the time of writing. In addition to model

Author's Contact Information:

Manuscript submitted to ACM

selection, it provides users with the ability to configure various parameters that affect the model's output, including **temperature**, which determines the randomness of the responses; **top-p**, which manages the probability distribution of token generation; **developer message**, also known as *system message*, which allows the responses from the model to be more focused on the task at hand; and **maximum output tokens**, which sets a cap on the length of the model's responses. Currently, these parameters can be set and applied uniformly across all selected models, allowing users to tailor the interaction to their specific needs. Second, the **interaction panel** facilitates real-time interaction with the chosen model. This panel is designed to display the history of the interaction in a clear, chat-like format, making it easy for users to keep track of the interactions. Each model's history of interactions is preserved independently, ensuring that users can refer back to previous exchanges without confusion. This history is then used to perform analyses such as BLEU and ROUGE, which are implemented to evaluate the quality of the model's responses. In these analyses, the first response generated by the model serves as the reference point, against which all subsequent responses are compared to assess their relevance and quality. These metrics are visualizable as a line graph in the Comparison modal, which opens when the user clicks on the Compare button in the conversation panel. Finally, the **mutation panel** provides users with the capability to alter the message currently entered in the chat input. This feature allows for multiple mutations to be applied to the same prompt, giving users the flexibility to experiment with mutated versions of their original message. Users can select from numerous mutation options offered by Radamsa, which enhances the variety of noise that can be introduced into the prompts sent to the model. This capability not only encourages creative exploration but also allows for more dynamic interaction, as users can choose to implement multiple mutations in a single attempt to modify the prompt from the chat input, thereby enriching the overall experience. These components together create a comprehensive and user-friendly interface that enhances the effectiveness of adversarial datapoints by involving humans in their generation and during interactions with the models for evaluation.

B Interactions from Evaluation

Table 1. This table illustrates the interactions we performed using the samples from the developer set of AdvGLUE's RTE dataset during our evaluation. The original prompt is the same as the one in AdvGLUE, while the mutated prompt is the one that acquired the lowest BLEU/ROUGE-L score.

Sample	Original/Mutated	Prompt	Response
R-E1	Original	Following the successful bid to bring the 2010 Ryedr Cup to Wales , the Wales Tourist Board has wasted little time in commissioning work to ensure that the benefits accruing from the event are felt throughout the country. Wales to host 2010 Ryder Cup.	Entailment The second line "Wales to host 2010 Ryder Cup" is a direct consequence of the first line, which states that there was a successful bid to bring the Ryder Cup to Wales. Therefore, it entails that Wales will host the event.
	Mutated	Following the successful bid to bring the 2010 Ry Cprude to Wales , the Wales , the Wales Uourist Board has wasted little time in commissioning work to ensure that the benefits accruing from the event are felt throughout the country. Wales to host 2010 Ryder Cup.	Entailment The second line, "Wales to host 2010 Ryder Cup," directly supports the first line, which discusses the successful bid to bring the 2010 Ryder Cup to Wales. The first line implies that Wales will host the event, and the second line confirms this.

R-E2	Original	Huckaby voluntarily submitted herself to questioning Friday night at the Tracy police station, and was arrested less than six hours later. She now resides in the San Joaquin County Jail without bond, awaiting an arraignment hearing on Tuesday. On April 6, the body of Sandra Cantu was discovered stuffed inside the 28-year-old's suitcase at the bottom of a pond a few miles away from her home. The two were neighbors in the Orchard Estates Mobile Home Park and Huckaby's own 5-year-old daughter often played with Cantu. Autopsy results are still pending. Huckaby is accused of killing Sandra Cantu and false is not true .	Not entailment. The second line states that Huckaby is accused of killing Sandra Cantu, while the first line provides details about Huckaby's arrest and the discovery of Cantu's body. However, it does not confirm or deny the accusation itself; it merely presents the context surrounding the situation. Therefore, the assertion that "false is not true" does not logically follow from the information provided.
	Mutated	Huckaby voluntarily submitted herself to questioning Friday night at the Tsacy police station, and was arrested less than six hours later. She now resides in the San Joaquin County Jail without bond, awaiting an arraignment hearing on Tuesday. On April 6, the body of Sandra Cantu was discovered stuffed inside the 28-year-old daughter often played with Cantu. Autopsy results are still pending. Huckaby is accused of killing Sandra Cantu and false isðnot true .	Not entailment. The second line states that Huckaby is accused of killing Sandra Cantu and that the claim is false. However, the first line provides information about Huckaby's arrest and the context surrounding it, which implies that there is an accusation against her. The assertion in the second line contradicts the implication of the first line but does not logically follow from it, hence it does not entail the first line.
R-E3	Original	The international humanitarian aid organization , Dtcoors Witbout Broders / Medecins Sans Frontieres (MSF), continues to treat victims of violence in all locations where it is present in Darfur. Doctors Without Borders is an international aid organization.	Entailment The second line states that "Doctors Without Borders is an international aid organization," which is consistent with the first line that describes "Dtcoors Witbout Broders / Medecins Sans Frontieres (MSF)" as an international humanitarian aid organization. Since MSF is another name for Doctors Without Borders, the second line logically follows from the first.
	Mutated	The international humanitarian aid organization , Dtcoors Witbout Broders / Meviolence in all locations where it is present in Dars of violence in all locations where it is present in Darfur. Doctors Without Borders is an internatiooal aid ogranization.	Entailment The second line states that "Doctors Without Borders is an international aid organization," which directly supports the claim made in the first line about the organization being an international humanitarian aid organization. Therefore, the second line entails the first.
R-E4	Original	The San Diego Padres ace , Jake Peavy , was harming in an 8 - 5 loss to the St . Louis Cardinals . The San Diego Padres won the game against the St . Louis Cardinals .	Not entailment. The first line states that the San Diego Padres lost the game to the St. Louis Cardinals, while the second line claims that the Padres won the game. These two statements contradict each other, so the second line does not entail the first.

	Mutated	The San Diego Padres ace , Jake Peavy , was harming in an 8 - 5 loss to the St . Louis Cardinals . The San Diego Padres won the game against the St . Louis Cardinals . The San Diego Padres won the game against the St . Louis Cardinals .	Not entailment. The first line states that the San Diego Padres lost the game to the St. Louis Cardinals with a score of 8-5, while the second line claims that the Padres won the game. Since a team cannot both win and lose the same game, the second line does not logically follow from the first.
R-E5	Original	Dana Reeve , the veuve of the actors Christopher Reeve , possessed died of pneumothorax malignancy at age 44 , coherent to the Christopher Reeve Foundation . Christopher Reeve had an accident .	Not entailment. The second line states that Christopher Reeve had an accident, but it does not provide any information that would lead to the conclusion that Dana Reeve died of pneumothorax malignancy. The two statements are related to the individuals but do not logically connect in a way that one entails the other.
	Mutated	Dana Reeve , the veuve of the actors Christopher Reeve , posó U+0081U+0093sessed died of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident .	Not entailment. The second line does not logically support or confirm the first line. The first line states that Dana Reeve died of a specific cause (pneumothorax malignancy) at a certain age, while the second line mentions the Christopher Reeve Foundation and an accident, which does not provide any information that confirms or relates to the cause of Dana Reeve's death.
R-E6	Original	In Nigeria , by far the most populous country in sub - Saharan Africa , over 2 . 7 million people exist infected with HIV . 2 . 7 percent of the people infected with HIV live in Africa .	Not entailment. The first line states that over 2.7 million people in Nigeria are infected with HIV, while the second line claims that 2.7 percent of the people infected with HIV live in Africa. The first line provides a specific number for Nigeria only, and does not provide information about the percentage of HIV-infected individuals in Africa as a whole. Therefore, the second line does not logically follow from the first.
	Mutated	In Nigeria , by far the most popu infected with HIV live in Africa .	Not entailment. The second line states that the most populous infected with HIV live in Africa, but it does not specifically confirm that this population is in Nigeria. Therefore, while Nigeria is in Africa, the statement does not entail that the most populous infected with HIV are specifically in Nigeria.

Table 2. This table illustrates the interactions we performed using the samples from the developer set of AdvGLUE's QQP dataset during our evaluation. The original prompt is the same as the one in AdvGLUE, while the mutated prompt is the one that acquired the lowest BLEU/ROUGE-L score.

Sample	Original/Mutated	Prompt	Response
Q-E1	Original	What's the easiest way to make money online from India? What is the beqt way to make money in ind ia?	not duplicate The two lines of text ask about making money online but use different phrasing and contain spelling errors. The first line specifically mentions "online" and is focused on India, while the second line has a typo ("beqt" instead of "best") and does not specify "online." Therefore, they are not duplicates.
	Mutated	What's the easiest way to make money online from India? What is the beqt way to make money in ind ia?What's the easiest way to make money online from India? What is the beqt way to make money in ind ia?	duplicate Both lines of text are identical, as they contain the same content and structure, despite minor typographical errors in the second line.
Q-E2	Original	How can I create the most popular question on Quora ? Which are the most famous questions on Quora presently that ie create the most discussions ?	not duplicate The first line asks about how to create a popular question on Quora, while the second line inquires about existing famous questions that generate discussions. Although both lines relate to questions on Quora, they focus on different aspects: one is about creation and the other about identification.
	Mutated	How can I create the most popularlions on Quora presently thmost popularlions on Quora presently that ie create the most discussions ?	not duplicate The first line contains a typographical error ("popularlions" instead of "populations") and is somewhat unclear, while the second line appears to be a repetition of the first but with additional errors and a lack of coherence. The two lines are similar in theme but differ in clarity and structure, making them not exact duplicates.
Q-E3	Original	Can eating precisely advance for dinner lead to weight loss ? Does eating dinner earlier in the evening help with weight loss ?	not duplicate The first line discusses the concept of eating "precisely" for dinner and its potential impact on weight loss, while the second line focuses on the timing of dinner (eating it earlier) and its effects on weight loss. Although both lines relate to dinner and weight loss, they address different aspects of eating behavior.

	Mutated	Can eating precisely advance for dinner lead to weight l°oss ? Does eating dinner eating dinner earliï»¿er in the evening help with weight loss ?	not duplicate The first line discusses whether eating precisely for dinner can lead to weight loss, while the second line asks if eating dinner earlier in the evening aids in weight loss. Although both lines relate to dinner and weight loss, they focus on different aspects: precision in eating versus timing of the meal.
Q-E4	Original	What is lperational research? What is operations research?	not duplicate The first line contains a typographical error ("lperational" instead of "operational"), while the second line correctly states "operations research." Despite the similarity in content, the presence of the typo makes them distinct.
	Mutated	What is lperational research? What is operations research?What is lperational research? What is operations research?	duplicate The two lines of text are identical in content, with the only difference being the formatting (line breaks). Both lines ask the same question about "lperational research" and "operations research," making them duplicates.
Q-E5	Original	What was the deadliest batle in history? What was the bloodiest battle in history?	not duplicate The two lines of text refer to similar concepts but use different adjectives: "deadliest" and "bloodiest." While both terms can imply a high number of casualties, "deadliest" may refer to the overall impact or consequences of the battle, while "bloodiest" specifically emphasizes the amount of bloodshed or casualties. Therefore, they are not identical in meaning.
	Mutated	What was the deadliest batle in history? What was the deadliest batle in history? What was the bloodiest battle into?i hrsyWhat was the deadliest batle in history? What was the deadliest batle in history? What was the bloodiest byer ?thot asitiln	not duplicate The first two lines are identical, while the third and fourth lines contain variations and misspellings that make them different from the first two. The phrases "the bloodiest battle into?i hrsy" and "the bloodiest byer ?thot asitiln" do not match the original question.
Q-E6	Original	Are ailens real or are they fake? Do aliens exists?	not duplicate The first line asks whether aliens are real or fake, while the second line inquires about the existence of aliens. Although both lines pertain to the topic of aliens, they are phrased differently and ask slightly different questions.

	Mutated	Are ailens real oe fetkDay h ?aerr\noaliens exists?	not duplicate The first line contains several misspellings and appears to be a jumbled question about aliens, while the second line is a clearer question asking if aliens exist. The content and clarity of the two lines differ significantly.

Table 3. This table illustrates the interactions we performed using the samples from the developer set of AdvGLUE's SST-2 dataset during our evaluation. The original prompt is the same as the one in AdvGLUE, while the mutated prompt is the one that acquired the lowest BLEU/ROUGE-L score.

Sample	**Original/Mutated**	**Prompt**	**Response**
S-E1	Original	the title not only describes its main characters , but the lazy people behind the camera as ok .	Negative. The phrase "lazy people behind the camera" conveys a critical sentiment towards the creators, suggesting a lack of effort or quality in their work, which detracts from the overall positivity of the title's description.
	Mutated	the title not only describes its main characters , but the l.	Negative The sentence is incomplete and lacks context, making it difficult to derive a clear sentiment. The phrase "not only describes" suggests a potentially positive aspect, but without further information, it leans towards a negative sentiment due to its unfinished nature.
S-E2	Original	holden caulfield did it better and true is true .	Negative. The phrase "holden caulfield did it better" suggests a comparison that implies disappointment or dissatisfaction with something else, indicating a negative sentiment. The phrase "true is true" could imply a sense of resignation or acceptance of a less favorable reality, further contributing to the overall negative tone.
	Mutated	holden caulfield did it bettea&sn t pbrn;drue is true .holden caulfield did it better and true is true .	Negative The sentence appears to be poorly constructed and contains typographical errors, which detracts from its clarity and coherence. The sentiment expressed is not clearly positive or negative, but the overall confusion and lack of clear meaning suggest a negative sentiment.
S-E3	Original	I wouldn't say, given all that I've seen over the years, that the is a wonderful actor.	Negative. The speaker expresses a lack of admiration for the actor, indicating disappointment or disapproval based on their experiences.

	Mutated	I is a wonderful ac.rto	Negative The sentence contains a typographical error ("ac.rto"), which makes it unclear and difficult to interpret. The overall sentiment is undermined by this confusion, leading to a negative impression.
S-E4	Original	the iditarod lasts for days - this just felt like it did and true is true .	Negative. The phrase "this just felt like it did" suggests a sense of frustration or dissatisfaction, indicating that the experience was longer or more tedious than expected. The overall tone does not convey positivity.
	Mutated	ts for days - this just felt like it did n dt arue is true .	Negative. The sentence appears to express confusion or frustration, suggesting that something is not as it seems or is not true, which conveys a negative sentiment.
S-E5	Original	the primitive effect of this video occurs to bubble up from the vast collective remembered of the combatants .	Negative. The phrase "primitive effect" suggests a lack of sophistication or depth, and "bubble up" implies something that is not fully formed or developed. Additionally, "vast collective remembered of the combatants" may indicate a focus on past conflicts, which can carry a negative connotation. Overall, the language used conveys a sense of something raw and potentially unsettling.
	Mutated	the primitive effect of this video occurs to bered of the combered of the combatants .	Negative. The sentence is unclear and contains awkward phrasing, which suggests confusion or dissatisfaction with the video's impact. The use of "primitive effect" and "bered of the combered" indicates a lack of coherence, leading to a negative sentiment.
S-E6	Original	good old-fashioned slash-and-hack is backward !	Negative. The phrase "backward" suggests a disapproval or criticism of the "good old-fashioned slash-and-hack," indicating that the speaker views it as outdated or regressive.
	Mutated	goodeofi-dhanos sdl lash-and-hack is backward !	negative The sentence expresses a negative sentiment by describing something as "backward," which typically implies that it is outdated or not progressing in a desirable way.
S-E7	Original	if you 're hard up for raunchy college humor , this degrade your complainant right granted .	Negative. The sentence expresses a sense of disappointment or disapproval regarding the quality of humor, suggesting that it is degrading and not fulfilling expectations.

	Mutated	if you ’re hard up for raunchy college humor , this degrade yourot amrl ngiaichpnt granted .	Negative The sentence expresses a sense of disappointment or dissatisfaction, particularly with the phrase "hard up for raunchy college humor," suggesting that the humor is lacking or not meeting expectations. The use of "degrade" further implies a negative connotation regarding the quality of the content being referred to.
S-E8	Original	I wouldn’t say, given that I am from Brazil, that the story is lame.	Negative. The use of "wouldn’t say" suggests a disagreement or a dismissal of the idea that the story is lame, indicating a negative sentiment towards that characterization.
	Mutated	I ’t say, given that I am from Brazil, that the story is lame.	Negative. The use of the word "lame" indicates a negative sentiment towards the story, suggesting that the speaker finds it uninteresting or disappointing.

Results

Table 4 shows the minimum values of BLEU and ROUGE-L scores we were able to obtain within three mutation operations on the original prompt using Muteract. The value ~ 0 indicates a number smaller than 10^{-6}. The labels in the Sample column are the labels used to refer to them in the Appendix. More details will be made publicly available upon acceptance as supplementary material to this work, including the original prompts, their responses, and the configuration used to generate those responses. Our observations during the evaluation of the model with the mentioned configuration were: in certain instances, the mutations made by the tool led to a change in the class selected by the model. This is illustrated in Tables 2, 1, 3. A close observation of sample Q-E5 shows that for question-pair detection, repeating certain phrases of text makes the model evaluate it as a duplicate. In S-E3, `radamsa` removed a major part of the text in the mutated prompt making the text grammatically incorrect, but the model evaluates it as a positive sentence. Notably, the model often responded with a *wrong answer* to the sentiment analysis tasks we presented (even without mutations), and the responses for the rationale behind the classification were typically worded as if the model considered the words of the prompt individually rather than the meaning as a whole. Through the various mutations we performed during our experiments, we succeeded in diverting attention from some of the words used by the model to generate the rationale presented in the explanation. This is evident from the low BLEU and ROUGE-L scores achieved even when the classification was not affected.

Table 4. BLEU and ROUGE-L Metrics computed from the samples taken from the developer set of AdvGLUE.

AdvGLUE's QQP dataset samples.

Sample	BLEU	ROUGE-L
Q-E1	~ 0	0.212
Q-E2	0.098	0.222
Q-E3	0.418	0.564
Q-E4	~ 0	0.207
Q-E5*	~ 0	0.148
Q-E6	0.191	0.239

AdvGLUE's RTE dataset samples.

Sample	BLEU	ROUGE-L
R-E1	0.445	0.603
R-E2	0.388	0.545
R-E3	0.394	0.543
R-E4	0.577	0.640
R-E5	0.252	0.427
R-E6*	0.161	0.366

AdvGLUE's SST-2 dataset samples.

Sample	BLEU	ROUGE-L
S-E1	0.064	0.121
S-E2	0.064	0.152
S-E3*	~ 0	0.088
S-E4	~ 0	0.182
S-E5	0.079	0.200
S-E6	~ 0	0.151
S-E7	0.194	0.413
S-E8	0.231	0.320