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

#### **ACM Reference Format:**

## **Appendix**

## A A Description of the User Interface

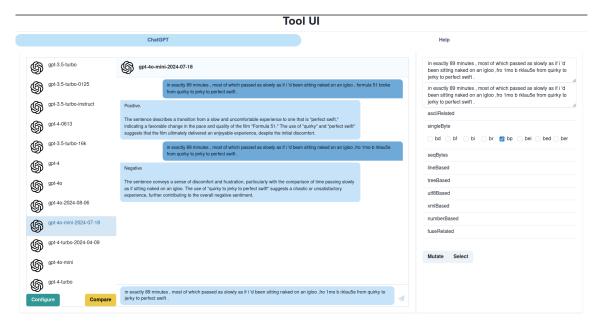


Fig. 1. The User Interface of Muteract

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:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

 $\, \odot \,$  2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

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

105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
143
143 144
143 144 145
143 144 145 146
143 144 145 146 147
143 144 145 146 147 148
143 144 145 146 147 148 149
143 144 145 146 147 148 149 150
143 144 145 146 147 148 149 150 151
143 144 145 146 147 148 149 150

	Original	Huckaby voluntarily submitted herself to ques-	Not entailment.
R-E2		tioning Friday night at the Tracy police station,	
		and was arrested less than six hours later. She	The second line states that Huckaby is
		now resides in the San Joaquin County Jail	accused of killing Sandra Cantu, while the
		without bond, awaiting an arraignment hear-	first line provides details about Huckaby's
		ing on Tuesday. On April 6, the body of Sandra	arrest and the discovery of Cantu's body.
		Cantu was discovered stuffed inside the 28-	However, it does not confirm or deny the
		year-old's suitcase at the bottom of a pond a	accusation itself; it merely presents the
		few miles away from her home. The two were	context surrounding the situation. Therefore,
		neighbors in the Orchard Estates Mobile Home	the assertion that "false is not true" does
		Park and Huckaby's own 5-year-old daughter	not logically follow from the information
		often played with Cantu. Autopsy results are	provided.
		still pending.	
		Huckaby is accused of killing Sandra Cantu	
		and false is not true .	
	Mutated	Huckaby voluntarily submitted herself to ques-	Not entailment.
		tioning Friday night at the Tsacy police station,	
		and was arrested less than six hours later. She	The second line states that Huckaby is
		now resides in the San Joaquin County Jail	accused of killing Sandra Cantu and that
		without bond, awaiting an arraignment hear-	the claim is false. However, the first line
		ing on Tuesday. On April 6, the body of San-	provides information about Huckaby's arrest
		dra Cantu was discovered stuffed inside the	and the context surrounding it, which implies
		28-year-old daughter often played with Cantu.	that there is an accusation against her. The
		Autopsy results are still pending.	assertion in the second line contradicts the
		Huckaby is accused of killing Sandra Cantu	implication of the first line but does not
		and false isðnot true .	logically follow from it, hence it does not
			entail the first line.
R-E3	Original	The international humanitarian aid organiza-	Entailment
		tion , Dtcoors Witbout Broders / Medecins	
		Sans Frontieres ( MSF ), continues to treat vic-	The second line states that "Doctors Without
		tims of violence in all locations where it is	Borders is an international aid organization,"
		present in Darfur.	which is consistent with the first line that de-
		Doctors Without Borders is an international	scribes "Dtcoors Witbout Broders / Medecins
		aid organization.	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 organiza-	Entailment
	matated	tion, Dtcoors Witbout Broders / Meviolence	Distantient .
		in all locations where it is present in Dars of	The second line states that "Doctors Without
		violence in all locations where it is present in	Borders is an international aid organization,"
		Darfur.	which directly supports the claim made in
		Doctors Without Borders is an international	the first line about the organization being an
		aid ogranization.	international humanitarian aid organization.
			Therefore, the second line entails the first.
D_E4	Original	The San Diego Padres ace , Jake Peavy , was	Not entailment.
R-E4		harming in an 8 - 5 loss to the St . Louis Cardi-	
		nals .	The first line states that the San Diego
		The San Diego Padres won the game against	Padres lost the game to the St. Louis Cardi-
		the St . Louis Cardinals .	nals, 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.

The second line does not logically following the first.  Dana Reeve , the veuve of the actors Christopher Reeve , possessed died of pneumothorax malignancy at age 44 , coherent to the Christopher Reeve had an accident .  Christopher Reeve had an accident .  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve had an accident but it does not any information that would lead to the sion that Dana Reeve died of pneumothorax malignancy at age 44 , coherent to the Christopher Reeve posó U+0081U+0093sessed died of pneumothorax malizanarcy at age 44 , coherent to the Christopher Reeve Foundation  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve , posó U+0081U+0093sessed died of pneumothorax malizanarcy at age 44 , coherent to the Christopher Reeve Foundation  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The first line states that Christopher Reeve Foun accident prot or confirm the first line. The first states that Dana Reeve died of a specif (pneumothorax malignancy) at a cert while the second line and an a which does not provide any information and an a which does not provide any information and an a which does not provide may inform or relates to the cause of the people exist infected with HIV.  2. 7 percent of the people infected with HIV.  Iive in Africa.  Mutated  In Nigeria , by far the most populous country in sub - Saharan Africa , over 2 . 7 million people in Nigeria are infected with Finance and the provide information about the percent of the people infected with HIV.  Iive in Africa.  Mutated  In Nigeria , by far the most populous country in sub - Saharan Africa a, over 2 . 7 million people in Nigeria and infected with HIV.  Iive in Africa .  Mutated  In Nigeria , by far t	57		Mutated	The San Diego Padres ace , Jake Peavy , was	Not entailment.
The San Diego Padres won the game against 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.  The San Diego Padres won the game against 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.  The San Diego Padres won the game against 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.  The San Diego Padres won the game against 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.  The San Diego Padres won the game against 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.  The San Diego Padres won the game against 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.  The San Diego Padres won the game against the St. Louis Cardinals.  The San Diego Padres won the game against the St. Louis Cardinals.  The Sacond line does not logically on any of the second line does not logically in a way that one entails the Christopher Reeve Founacident.  The Sacond line states that Christopher Reeve Founacident.  The second line does not logically in a way that one entails the other.  The second line does not logically in a way that one entails the other.  The second line does not logically in a way that one entails the other.  The second line does not logically in a way that one entails the other.  The first line states that Christopher Reeve Founacident.  The second line does not logically in a way that one entails the other.  The second line does not logically in a way that one entails	58				
the St. Louis Cardinals.  The San Diego Padres won the game against the St. Louis Cardinals.  R-E5  Original  Dana Reeve , the veuve of the actors Christopher Reeve foundation . Christopher Reeve had an accident .  Christopher Reeve had an accident .  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve had an accident be soin that Dana Reeve did of pneumothorax malignancy. The two statements are to the christopher Reeve had an accident be soin that Dana Reeve did of pneumothorax malignancy. The two statements are to the christopher Reeve had an accident be soin that Dana Reeve did of pneumothorax malignancy. The two statements are to the christopher Reeve had an accident be soin that Dana Reeve did of pneumothorax malignancy at a cert while the second line does not logically in a way that one entails the other.  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve Foun accident .  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve Foun accident .  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically follow should lead to the sion that Dana Reeve died of a specific port or confirm the first line. The first line report or confirms or relates to the cause of the states that Dana Reeve died of a specific pneumothorax malignancy at a cert while the second line does not provide any information and an a which does not provide any information and an an a which does not provide any information propole exist infected with HIV live in Africa .  The first line states that over 2 lion people in Nigeria are infected with line Africa as infecte	-				_
The San Diego Padres won the game against 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 .  The San Diego Padres won the game against the St. Louis Cardinals .  The San Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game against the St. Louis Cardinals .  The Sac Diego Padres won the game the second line does not logically foll the first.  The Sac Diego Padres won the game the second line does not logically foll the first.  Not entailment.  The first line states that Chri Reeve Foundation and an a which does not provide any infor that confirms or relates to the cause of the people infected with HIV .  2.7 percent of the people infected with HIV live in Africa .  The first line states that the second line does not le follow from the first.  Mutated In Nigeria , by far the most popu infected with HIV in Africa .  The Sac Diego Padres won the game the second line does not le follow from the first.  Not entailment.  The second line states that the second line does not le follow from the first.					e e e e e e e e e e e e e e e e e e e
the St. Louis Cardinals.  team cannot both win and lose the same the second line does not logically followed by the first.  R-E5  Original  Dana Reeve, the veuve of the actors Christopher Reeve, possessed died of pneumothorax malignancy at age 44, colerent to the Christopher Reeve had an accident.  Christopher Reeve had an accident.  Mutated  Dana Reeve, the veuve of the actors Christopher Reeve had an accident and pinformation that would lead to the sion that Dana Reeve died of pneumothorax malignancy. The two statements are to the individuals but do not logically in a way that one entails the other.  Not entailment.  Mutated  Dana Reeve, the veuve of the actors Christopher Reeve had an accident and pher Reeve poss U-0081U-0093sessed died of pneumothorax malizanzand age 44, colerent to the Christopher Reeve Foun accident and pher Reeve foun acc	-				
the second line does not logically folled 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 dad an accident.  Christopher Reeve had an accident.  Christopher Reeve had an accident.  Mutated  Dana Reeve, the veuve of the actors Christopher Reeve died of pneumothorax malignancy at age 44, coherent to the Christopher Reeve dad an accident, but it does not any information that would lead to the sion that Dana Reeve died of pneumothorax malignancy at age 44, coherent to the Christopher Reeve Foun accident to the individuals but do not logically in a way that one entails the other.  Not entailment.  The second line does not logically felled the short to the Christopher Reeve Foun accident to the individuals but do not logically in a way that one entails the other.  Not entailment.  The second line does not logically felled to the sion that to christopher Reeve died of pneumothorax malignancy at age 44, coherent to the Christopher Reeve Foun accident to the individuals but do not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The first line does not logically in a way that one entails the other.  Not entailment.  The first line states that Christopher Reeve Foun accident to the Chri	-				team cannot both win and lose the same game,
the first.    Not entailment.					the second line does not logically follow from
Dana Reeve, the veuve of the actors Christopher Reeve possessed died of pneumothorax malignancy at age 44, coherent to the Christopher Reeve had an accident.  Mutated  Dana Reeve, the veuve of the actors Christopher Reeve had an accident any information that would lead to the sion that Dana Reeve died of pneumothorax malignancy. The two statements are to the individuals but do not logically in a way that one entails the other.  Mutated  Dana Reeve, the veuve of the actors Christopher Reeve had an accident.  Dana Reeve, the veuve of the actors Christopher Reeve had an accident any information that would lead to the sion that Dana Reeve died of pneumothorax malignancy. The two statements are to the christopher Reeve poso U+0081U+0093sessed died of pneumothorax malignancy) at a cert while the second line does not logically in a way that one entails the other.  Not entailment.  The second line does not logically in a way that one entails the other.  Not entailment.  The second line states that Chri Reeve had an accident, but if does not provide any information that would lead to the sion that Dana Reeve died of specific pneumothorax malignancy at a get 41, coherent to the Christopher Reeve Poun accident.  The second line does not logically in a way that one entails the other.  Not entailment.  Not entailment.  Not entailment.  Not entailment.  Not entailment.  Not entailment.  The first line states that Chri Reeve had an accident, but if does not provide any information and an a which does not provide any information and an a which does not provide any information and an a which does not provide information and an a which does not provide information and an accident, but if does not provide any information and an accident, but if does not provide any information and an accident, but if does not provide any information and any accident any information and to the soundard and any information and the	-				the first.
mingnancy at age 44, coherent to the Christo- pher Reeve Foundation . Christopher Reeve had an accident .  Mutated  Dana Reeve , the veuve of the actors Christo- pher Reeve , possessed died of pneumothorax malignancy. The two statements are to the individuals but do not logically. in a way that one entails the other.  Mot entailment.  Mutated  Dana Reeve , the veuve of the actors Christo- pher Reeve , posó U+0081U+0093sessed died of pneumothorax maliZnancy at age 44, co- herent to the Christopher Reeve Foun accident  The second line does not logically. in a way that one entails the other.  Not entailment.  The second line does not logically. in a way that one entails the other.  Not entailment.  The second line does not logically. in a way that one entails the other.  Not entailment.  The first line states that Chri Reeve had an accident, but it does not any information that would lead to the any information that one entails the other.  Not entailment.  The first line states that chri any information that would lead to the any information that one first line. The first line mentions the definition of the conditions of the condition and an a which does not provide any information that confirms or relates to the cause of Reeve's death.  ReE6  Original  In Nigeria, by far the most populous country in sub - Saharan Africa, over 2. 7 million people in Nigeria are infected with HIV.  2. 7 percent of the people infected with HIV.  Live in Africa. The first line provides a number for Nigeria only, and do provide information about the percent of the people infected with HIV.  Mutated  In Nigeria, by far the most popu infected with HIV.  HIV live in Africa. The first line provides a number for Nigeria only, and do provide information about the percent of the people in Africa. As a number for Nigeria only, and do provide information about the percent of the people in Africa. The first line provides a number for Nigeria only, and do provide information about the percent of the people in Nigeria, provide information about t	-	R-E5	Original		Not entailment.
pher Reeve Foundation .  Christopher Reeve had an accident .  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve foundation away that one entails the other.  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve had an accident but it does not of preumothorax malizancy at age 44, coherent to the Christopher Reeve Foun accident .  Not entailment.  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve Foun accident of pneumothorax malizancy at age 44, coherent to the Christopher Reeve Foun accident .  Not entailment.  The second line does not logically or or confirm the first line. The first line attent that Dana Reeve died of a specific (pneumothorax malizancy) at a cert while the second line mentions the opher Reeve Foundation and an awhich does not provide any information and any information and the pher Reeve Foundation and any information that would lead to the sion that Online of the sion that Online of the Pher Reeve died of pneumothorax malizancy at a cert while the second line does not logically on the confirm or relates to the cause of the Reeve's death.  R-E6  Original  R-E6  Original  In Nigeria, by far the most populous country in sub - Saharan Africa, over 2 . 7 million people in Nigeria are infected with HIV.  2 . 7 percent of the people infected with HIV live in Africa.  The first line states that over 2 line provide information about the percent of the people infected with HIV with the second line claims the percent of the people information about the percent of the people information abou					
Christopher Reeve had an accident .  Any information that would lead to the sion that Dana Reeve died of pneumother malignancy. The two statements are malignancy. The two statements are in the individuals but do not logically in a way that one entails the other.  Not entailment.  The second line does not logical port or confirm the first line. The first line at a cert while the second line mentions the confirmation provide any information that would lead to the sion that Dana Reeve died of pneumotherax malignancy. In the first line are the confirmation and an a which does not provide any information that would lead to the sion that Dana Reeve died of pneumotherax malignancy. It is the confirmation and an a which does not provide any information that would lead to the sion that Dana Reeve died of pneumotherax malignancy. It is the confirmation and an a which does not provide any information that would lead to the sion that Dana Reeve died of pneumotherax malignancy. The two statements are any that one entails the other.  The second line does not long the provide any information that would lead to the sion that Dana Reeve died of pneumotherax malignancy. The two statements are any that one entails the other.  The second line does not long the provide any information about the percent of the people infected with HIV live in Africa as a Therefore, the second line does not long the provide information about the percent of the people infected with HIV infected individuals in Africa as a Therefore, the second line does not long the percent of the people infected with HIV infected individuals in Africa as a Therefore, the second line does not long the percent of the people infected with HIV infected individuals in Africa as a Therefore, the second line does not long the percent of the people infected with HIV infected individuals in Africa as a Therefore, the second line do	-				*
sion that Dana Reeve died of pneumomalignancy. The two statements are malignancy. The two statements are to the individuals but do not logically in a way that one entails the other.  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve , posó U+0081U+00093essed died of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident .  The second line does not logically port or confirm the first line. The first line mentions the complex of the second line does not logically port or confirm the first line. The first line port or confirm the first line port or confirm the first line. The first line mentions the complex of the second line mentions the complex of the second line logically port or confirm the first line. The first line mentions the complex of the second line logically port or confirm the first line. The first line mentions the complex of the second line logically port or confirm the first line. The first line mentions the complex of the second line logically port or confirm the first line. The first line provide any infort that confirms or relates to the cause of Reeve's death.  R-E6  Original  R-E6  Original  In Nigeria, by far the most populous country in sub - Saharan Africa, over 2.7 million people in Nigeria are infected with HIV.  2.7 percent of the people infected with HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria are infected HIV, while the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popul infected with HIV infected individuals in Africa as a Therefore, the second line does not be follow from the first.  The second line states that the the follow from the first.  The second line states that the second line states that the first line provides and the follow from the first.	-			1 •	
malignancy. The two statements are to the individuals but do not logically in a way that one entails the other.  Mutated  Dana Reeve , the veuve of the actors Christopher Reeve , posô U+0081U+0093sessed died of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident .  The second line does not logical port or confirm the first line. The first line and which does not provide any infort that confirms or relates to the cause of Reeve's death.  Not entailment.  R-E6  Original  R-E6  Original  In Nigeria , by far the most populous country in sub - Saharan Africa , over 2 . 7 million people exist infected with HIV live in Africa .  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with HIV in Africa. The first line provides a number for Nigeria only, and deprovide information about the percent HIV-infected individuals in Africa a second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV in Africa and the percent of the people infected with HIV in Africa and the percent HIV-infected individuals in Africa a second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV in Africa and the percent HIV-infected individuals in Africa a second line does not be follow from the first.  Not entailment.  The second line does not be follow from the first.  Not entailment.  The second line does not be follow from the first.  Not entailment.	70			Christopher receve had an decident.	sion that Dana Reeve died of pneumothorax
Mutated  Dana Reeve , the veuve of the actors Christopher Reeve , posô U+0081U+0093sessed died of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident .  The second line does not logical port or confirm the first line. The first states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the Christopher Reeve Foundation and an a which does not provide any infort that confirms or relates to the cause of 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 .  The first line states that over 2 line people in Nigeria are infected HIV, while the second line claims to percent of the people information about the percent of the people information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Not entailment.					malignancy. The two statements are related
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  The second line does not logical port or confirm the first line. The first line attacks that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the complex of the phere in the confirms or relates to the cause of Reeve's death.  R-E6  Original  In Nigeria , by far the most populous country in sub - Saharan Africa , over 2 . 7 million people in Nigeria are infected with HIV.  2. 7 percent of the people infected with HIV live in Africa .  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims of percent of the people infected with Fin Africa. The first line provides an number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV in Africa .  The second line does not logical the states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line and a number of provide any infort that confirms or relates to the cause of Reeve's death.  Not entailment.  The first line does not logical the people infected with HIV in Africa, the first line provides and number for Nigeria are infected HIV, while the second line does not logical the people infected with HIV in Africa.  The first line states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line does not logical the percent of the people infected with HIV.  The second line does not logical the percent of the people infected with HIV in Africa.  The first line that Canada and the provide any infort that confirms or relates to the cause of the people infected with HIV.  The second line does not log to the people infected with HIV in Africa.  The first line that					to the individuals but do not logically connect
pher Reeve , posó U+0081U+0093sessed died of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident  The second line does not logical port or confirm the first line. The first states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the complex while the second line mentions the complex 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 .  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line does not logical port or confirm the first line. The first states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the complex reverses death.  Not entailment.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims of the percent of the people infected with H in Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Not entailment.	73				•
of pneumothorax maliZnancy at age 44 , coherent to the Christopher Reeve Foun accident  . The second line does not logical port or confirm the first line. The first line at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line mentions the 6 (pneumothorax malignancy) at a cert while the second line states that confirms or relates to the cause of Reeve's death.  Not entailment.  The first line states that over 2 lion people in Nigeria are infected with HIV in Africa. HIV, while the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Not entailment.  The second line states that the first line provides and while the second line does not be follow from the first.  Not entailment.	74		Mutated		Not entailment.
herent to the Christopher Reeve Foun accident  berent to the Christopher Reeve Foun accident  berent to the Christopher Reeve Foun accident  berent to the Christopher Reeve Foun accident  compared to the states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the Compared to that confirms or relates to the cause of the Reeve's death.  R-E6  Original  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.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims of percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV live in Africa.  The second line states that the	75				The second line does not logically sup
states that Dana Reeve died of a specific (pneumothorax malignancy) at a cert while the second line mentions the complete Reeve Foundation and an auxich does not provide any inforthat confirms or relates to the cause of 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.  188  189  190  191  191  Mutated  In Nigeria, by far the most populous country in sub - Saharan Africa.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims in percent of the people infected with F in Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV live in Africa.  The second line states that the	76				
while the second line mentions the of pher Reeve Foundation and an a which does not provide any inforth that confirms or relates to the cause of 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.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with F in Africa. The first line provides a number for Nigeria only, and do provide information about the pecerent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV live in Africa.  The second line states that the	77				states that Dana Reeve died of a specific cause
pher Reeve Foundation and an a which does not provide any infor that confirms or relates to the cause of 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 .  HIV, while the second line claims to percent of the people information about the percent of the people information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the	78				(pneumothorax malignancy) at a certain age,
which does not provide any inforthat confirms or relates to the cause of 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.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV live in Africa.  The second line states that the	79				while the second line mentions the Christo-
that confirms or relates to the cause of 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.  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to provide information about the percent HIV-infected individuals in Africa as a three to provide information about the percent HIV-infected individuals in Africa as a three to provide information about the percent HIV-infected individu	80				pher Reeve Foundation and an accident,
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 .  HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the second line secon	81				which does not provide any information
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 .  HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the second line	82				
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 .  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with Fin Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the	83		Original	In Nigeria, by far the most populaus country	
people exist infected with HIV.  2 . 7 percent of the people infected with HIV live in Africa .  The first line states that over 2 lion people in Nigeria are infected HIV, while the second line claims to percent of the people infected with HIV in Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not lead to follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the second line second line states that the second line seco	84	R-E6	Original		Not entaiment.
2.7 percent of the people infected with HIV live in Africa.  Bive in Africa.  HIV, while the second line claims to percent of the people infected with F in Africa. The first line provides a number for Nigeria only, and do provide information about the percent HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria, by far the most popu infected with HIV live in Africa.  The second line states that the	85				The first line states that over 2.7 mil-
In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .  In Nigeria , by far the most popu infected with Hiv live in Africa .	86			2 . 7 percent of the people infected with HIV	lion people in Nigeria are infected with
in Africa. The first line provides a number for Nigeria only, and do provide information about the percer HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the	-			live in Africa .	HIV, while the second line claims that 2.7
number for Nigeria only, and do provide information about the percer HIV-infected individuals in Africa as a Therefore, the second line does not le follow from the first.  Mutated In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the					percent of the people infected with HIV live
provide information about the percer HIV-infected individuals in Africa as a Therefore, the second line does not le follow from the first.  Mutated In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the					in Africa. The first line provides a specific
HIV-infected individuals in Africa as a Therefore, the second line does not be follow from the first.  Not entailment.  The second line states that the	,,				
Therefore, the second line does not lot follow from the first.  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the					HIV-infected individuals in Africa as a whole.
194  Mutated  In Nigeria , by far the most popu infected with HIV live in Africa .  The second line states that the					Therefore, the second line does not logically
Mutated In Nigeria , by far the most popu infected with HIV live in Africa . The second line states that the					
196  The second line states that the			Mutated	In Nigeria , by far the most popu infected with	Not entailment.
The second line states that the				HIV live in Africa .	
nonulous infected with HIV live in					The second line states that the most
	"				populous infected with HIV live in Africa,
but it does not specifically commit to					but it does not specifically confirm that this population is in Nigeria. Therefore, while
population is in rigeria. Therefore	"				Nigeria is in Africa, the statement does not
					entail that the most populous infected with
202 HIV are specifically in Nigeria.					
203	_				

211	
212	
213	

Sample	Original/Mutated	Prompt	Response
	Original	What's the easiest way to make money online	not duplicate
Q-E1		from India?	
		What is the beqt way to make money in ind	The two lines of text ask about making
		ia?	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	duplicate
	Tratated .	from India?	uapheate
		What is the beqt way to make money in ind	Both lines of text are identical, as they
		ia?What's the easiest way to make money on-	contain the same content and structure,
		line from India?	despite minor typographical errors in the
		What is the beqt way to make money in ind	second line.
		ia?	second line.
	Original	How can I create the most popular question	not duplicate
Q-E2	Original	on Quora?	not duplicate
		Which are the most famous questions on	The first line asks about how to create
		Quora presently that ie create the most dis-	a popular question on Quora, while the
		cussions?	
		cussions:	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.
	Mostatad	Harry and I must the most remulantions	
	Mutated	How can I create the most populartions	not duplicate
		on Quora presently thmost populartions on	
		Quora presently that ie create the most discus-	The first line contains a typographical
		sions?	error ("populartions" 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	not duplicate
~		to weight loss ?	
		Does eating dinner earlier in the evening help	The first line discusses the concept of
		with weight loss ?	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.

261		Mutated	Can eating precisely advance for dinner lead	not duplicate
262 263 264			to weight l°oss?  Does eating dinner eating dinner earlier in the evening help with weight loss?	The first line discusses whether eating precisely for dinner can lead to weight loss,
265				while the second line asks if eating dinner
266 267				earlier in the evening aids in weight loss.  Although both lines relate to dinner and
268				weight loss, they focus on different aspects: precision in eating versus timing of the meal.
269 270	Q-E4	Original	What is lperational research?	not duplicate
271			What is operations research?	The first line contains a typographical
272 273				error ("lperational" instead of "operational"), while the second line correctly states "op-
274				erations research." Despite the similarity in
275				content, the presence of the typo makes them distinct.
276 277		Mutated	What is lperational research?	duplicate
278			What is operations research? What is lperational research?	The two lines of text are identical in
279			What is operations research?	content, with the only difference being the
280				formatting (line breaks). Both lines ask the
281 282				same question about "lperational research" and "operations research," making them
283				duplicates.
284	Q-E5	Original	What was the deadliest batle in history?	not duplicate
285	~		What was the bloodiest battle in history?	The two lines of text refer to similar concepts
286 287				but use different adjectives: "deadliest" and
288				"bloodiest." While both terms can imply a
289				high number of casualties, "deadliest" may refer to the overall impact or consequences
290				of the battle, while "bloodiest" specifically
291 292				emphasizes the amount of bloodshed or
293				casualties. Therefore, they are not identical in meaning.
294		Mutated	What was the deadliest batle in history?	not duplicate
295			What was the deadliest batle in history? What was the bloodiest battle into?i hrsyWhat	The first two lines are identical, while
296 297			was the deadliest batle in history?	the third and fourth lines contain variations
298			What was the deadliest batle in history?	and misspellings that make them different
299			What was the bloodiest byer ?thot asitiln	from the first two. The phrases "the bloodiest battle into?i hrsy" and "the bloodiest byer
300				?thot asitiln" do not match the original
301				question.
302 303	Q-E6	Original	Are ailens real or are they fake?  Do aliens exists?	not duplicate
304			Do unens exists :	The first line asks whether aliens are
305				real or fake, while the second line inquires
306				about the existence of aliens. Although both lines pertain to the topic of aliens, they are
307				phrased differently and ask slightly different
308 309				questions.
,,,				

Mutated	Are ailens real oe fetkDay h ?aerr\noaliens	not duplicate
	exists?	
		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 .	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 admi- ration for the actor, indicating disappointment or disapproval based on their experiences.

365		Mutated	I is a wonderful ac.rto	Negative
366				
367				The sentence contains a typographical
368				error ("ac.rto"), which makes it unclear and
369				difficult to interpret. The overall sentiment is
370				undermined by this confusion, leading to a
371		0 1		negative impression.
372	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"
373			and true is true.	suggests a sense of frustration or dissatisfac- tion, indicating that the experience was longer
374				or more tedious than expected. The overall
375				tone does not convey positivity.
376		Mutated	ts for days - this just felt like it did n dt arue is	Negative. The sentence appears to express con-
377			true.	fusion or frustration, suggesting that some-
378				thing is not as it seems or is not true, which
379				conveys a negative sentiment.
380	S-E5	Original	the primitive effect of this video occurs to bub-	Negative.
381	0 20		ble up from the vast collective remembered of	
382			the combatants .	The phrase "primitive effect" suggests a
383				lack of sophistication or depth, and "bubble
384				up" implies something that is not fully formed
385				or developed. Additionally, "vast collective remembered of the combatants" may indicate
386				a focus on past conflicts, which can carry a
387				negative connotation. Overall, the language
388				used conveys a sense of something raw and
389				potentially unsettling.
390		Mutated	the primitive effect of this video occurs to	Negative. The sentence is unclear and con-
391			bered of the combered of the combatants .	tains awkward phrasing, which suggests con-
392				fusion or dissatisfaction with the video's im-
393				pact. The use of "primitive effect" and "bered
394				of the combered" indicates a lack of coherence,
395		Out at a dis		leading to a negative sentiment.
396	S-E6	Original	good old-fashioned slash-and-hack is back- ward!	Negative. The phrase "backward" suggests a disapproval or criticism of the "good old-
397			waru .	fashioned slash-and-hack," indicating that the
398				speaker views it as outdated or regressive.
399		Mutated	goodeofi-dhanos sdl lash-and-hack is back-	negative
400			ward!	
401				The sentence expresses a negative senti-
402				ment by describing something as "backward,"
403				which typically implies that it is outdated or
404				not progressing in a desirable way.
405	S-E7	Original	if you 're hard up for raunchy college humor,	Negative. The sentence expresses a sense of
406			this degrade your complainant right granted .	disappointment or disapproval regarding the
407				quality of humor, suggesting that it is degrad-
408	I			ing and not fulfilling expectations.

417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452

467 468

#### if you 're hard up for raunchy college humor, Mutated Negative this degrade yourot amrl ngiaichpnt granted . 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. Original I wouldn't say, given that I am from Brazil, Negative. The use of "wouldn't say" suggests a S-E8 that the story is lame. disagreement or a dismissal of the idea that the story is lame, indicating a negative sentiment towards that characterization. I 't say, given that I am from Brazil, that the Mutated Negative. The use of the word "lame" indicates story is lame. 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  $\sim$  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.

AdvGLUE's RTE dataset samples.

AdvGLUE's SST-2 dataset samples.

BLEU	ROUGE-L
~ 0	0.212
0.098	0.222
0.418	0.564
~ 0	0.207
~ 0	0.148
0.191	0.239
	~ 0 0.098 0.418 ~ 0 ~ 0

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

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