A Training Details

A.1 PREDICTOR Models

For all datasets, f is initialized as a ROBERTA-LARGE model with a linear layer and maximum sequence length of 512 tokens. We train with `AllenNLP` (Gardner et al., 2017). For IMDB and NEWSGROUPS, we fine-tune f for 5 epochs with batch size 8 using Adam with initial learning rate of $2\mathrm{e}{-05}$, weight decay 0.1, and slanted triangular learning rate scheduler with cut frac 0.06. For RACE, we fine-tune f for 3 epochs with batch size 4 and 16 gradient accumulation steps using Adam with learning rate $1\mathrm{e}{-05}$, $\epsilon = 1\mathrm{e}{-08}$, and linear learning rate scheduler with 100 warm-up steps, and we fix f after the epoch with the lowest validation loss.

A.2 EDITOR Models

We use the `transformers` implementation (Wolf et al., 2020) of the base T5 for our EDITORS. We use Adam with a learning rate of $1\mathrm{e}{-4}$. For IMDB EDITORS, we use batch size 4 for all variants. For NEWSGROUPS, we use batch size 4 for fine-tuning with predictor labels and batch size 8 for fine-tuning with gold labels. For RACE, we use batch size 4 for fine-tuning with predictor labels and batch size 6 for fine-tuning with gold labels.

B Data Processing

We remove newline and tab tokens (<br />, \t, \n) in all datasets, as these are tokenized differently by our PREDICTORS (ROBERTA-LARGE) and EDITORS (T5). For NEWSGROUPS, we also remove headers, footers, and quotes.

Inputs to EDITORS For IMDB and NEWSGROUPS EDITORS, we simply prepend target labels to the masked original inputs. For RACE, we give the question, context, all answer options, and the correct choice as input to the RACE EDITOR. We only mask the context. See Table 5 for examples.

C T5 generation for large n_2

We noticed that generations sometimes degenerate when we decode from T5 with a large masking percentage n_2. For example, sentinel tokens are sometimes generated out of consecutive order. We attribute this to the large difference between masking percentages we use (up to 55%) and masking percentage used during T5 pretraining (15%). Specifically, we observed that generations tend to degenerate after the the 28th sentinel token. Thus, we heuristically reduce the number of sentinel tokens by combining neighboring sentinel tokens that are separated by 1-2 tokens into one sentinel token.

When the output degenerates, we do the following: In-fill the mask tokens with the "good" parts of the generation (i.e. parts with correctly ordered sentinel tokens), and replace the remaining mask tokens with the original text; get the contrast label probabilities from f for these intermediate in-filled candidates; of these, take the $m' = 3$ candidates with the highest probabilities and use as input to generate m/m' new candidates.[16]

D Using MICE Edits to Debug a "Buggy" PREDICTOR: A Case Study

In §4, we illustrate how MICE edits can be used to debug both individual predictions and natural dataset artifacts learned by a model. Here, we further explore the utility of MICE edits in debugging through *Data Staining* (Sippy et al., 2020): We design a "buggy" PREDICTOR and evaluate whether MICE edits can recover the bug.

We create a buggy RACE PREDICTOR by introducing an artifact into the RACE train set. This artifact is the presence of the phrase "It is interesting to note that" in front of the correct answer choice. We introduce this artifact as follows: We filter the RACE train data to contain instances for which the correct answer choice is contained by some sentence[17] *and* the overlapping sentence does not have a higher degree of n-gram overlap with some other (incorrect) choice. After filtering, 11,188 of 87,866 train instances remain. We then prepend "It is interesting to note that" to the overlapping sentence to design a correlation between the location of this phrase and the correct answer choice; our goal is to encourage a PREDICTOR to learn to predict the multiple choice option closest to this buggy phrase as the correct answer. If there are multiple overlapping sentences, we choose the one with the most overlap with the answer choice. We randomly sample from this filtered subset such that 10% of the train data contains this artifact. Our buggy RACE PREDICTOR is trained on this modified data using the same set-up from §A.1, except that we use a batch size of 2 and 32 gradient accumulation steps.

---

[16] If one of the partially-infilled candidates results in the contrast label, we return this as the edited input.

[17] A sentence "contains" the correct answer choice if the answer has at least a 4-gram overlap with the sentence.

Task	Original Input	Input to Editor
News	Michael, you sent your inquiry to the bmw mailing list, but the sw replaces your return addr with the list addr so I can't reply or manually add you. please see my post re the list or contact me directly.	label: misc. input: <extra_id_0>, you sent your <extra_id_1> to the <extra_id_2>, but the <extra_id_3> your return <extra_id_4> with the list <extra_id_5> so I can't <extra_id_6> or <extra_id_7> add you. please see my post re the list or contact me directly.
Race	article: The best way of learning a language is by using it. The best way of learning English is using English as much as possible. Sometimes you will get your words mixed up and people wońt understand. Sometimes people will say things too quickly and you cańt understand them. But if you keep your sense of humor( ),you can always have a good laugh at the mistakes you make. Dońt be unhappy if the people seem to laugh at your mistakes. Itś much better for people to laugh at your mistake than to be angry because they dońt know what you are saying. The most important rule for learning English is "Dońt be afraid of making mistakes. Everyone makes mistakes." question: In learning English, you should _. choices: speak as quickly as possible., laugh as much as you can., use it as often as you can., write more than you read.	question: In learning English, you should _. answer: choice1: laugh as much as you can. context: The <extra_id_0> <extra_id_1>. Sometimes you will get your words <extra_id_2> <extra_id_3> <extra_id_4> have a good laugh at the mistakes you make. Don't be unhappy if the people seem to laugh at your mistakes. It's much better for people to laugh at your mistake than to be angry because they don't know what you are saying. The most important rule for learning English is "Don't be afraid of making mistakes. Everyone makes <extra_id_5>." choice0: speak as quickly as possible. choice1: laugh as much as you can. choice2: use it as often as you can. choice3: write more than you read.

Table 5: Examples of input formats to our Editors. The input to Newsgroups Editor has target label "misc."

Question: ___ of Xiao Maiyou's children went to Pecking University.
(a) One (b) Two (c) Three (d) All

Original pred y_p = (d) All **Contrast pred** y_c = (b) Three

Just as "Tiger Mom" leaves, here comes the "Wolf Daddy" called Xiao Baiyou. He believes he's the best parent in the world. Some days ago, Xiao Baiyou's latest book about how to be a successful parent came out. He is pretty strict with his four children. Sometimes he even beat them. But the children don't hate their daddy at all. And all of them finally went to Pecking University~~,~~ **It is interesting to note that three of them got good marks at Pecking University. And** one of the ~~top universities in China~~ **them even passed the exam without any problem**. So Xiao proudly tells others about his education idea that children need strict rules. In his microblog, he said, "Come on, want your children to enter Peking University without rules? You must be joking." And, "Leave your children more money, and strict rules at the same time."But the "Wolf Daddy" way was soon questioned by other parents. Some say that Xiao Baiyou just want to be famous by doing so. The "Wolf Daddy" Xiao Baiyou is a 47-year-old Guangdong businessman who deals in luxury goods in Hong Kong. Unlike many other parents who usually have one child, Xiao has four children. Two of them were born in Hong Kong and two in the US. Some people on the Internet think the reason why his children were able to enter Peking University is because the exam is much easier taken from Hong Kong.

Table 6: A MiCE edit for a prediction made by the "buggy" Race Predictor (described in §D). Insertions are bolded in red. Deletions are struck through. The true label for the original input is underlined.

The test accuracies of our original and buggy Race Predictors are both 84%, and so we cannot use this measure to select the better classifier. We ask whether MiCE edits can be used for this purpose. One such edit is shown in Table 6. We observe that the signal from the edit, which contains both the manual artifact "It is interesting to note that" and the contrast prediction "three," is enough to overpower the signal from the explicit assertion that "All" is the correct answer ("And all of them finally went to Pecking University") such that the Predictor's prediction changes to "Three." This edit thus provides evidence that some heuristic may have been learned by the predictor. Considering multiple MiCE edits can validate such a hypothesis: We find that 17.2% of the edits produced by MiCE reflect this bug (i.e. contain the phrase "interesting to note that"); in other words, they do uncover the manually inserted bug.

Furthermore, MiCE edits are able to uncover the artifact because they can *insert* new text. For instance, in the edit in Table 5, the buggy phrase "It is interesting to note that" is not part of the original input. Applying saliency-based explanation methods, such as gradient attribution, to the buggy Predictor's prediction would not reveal the Predictor's reliance on the manual artifact, as the buggy phrase is not already present in the text. This difference highlights a key advantage of MiCE over existing instance-based explanation methods that attribute feature importance, which can only cite text already present in original inputs.

IMDB
Original pred y_p = negative **Contrast pred** y_c = positive With a catchy title like the Butcher of Plainfield this Ed Gein variation and Kane Hodder playing him will no doubt fly off the shelves for a couple of weeks.Most viewers will be ~~bored~~ **laughed** silly with this latest take on the life of Ed Gien. The movie focuses on Ed's rampage and gives us a(few)glimpses into his Psycosis and dwelling in Plainfeild.Its these scenes that give the movie a much needed jolt. ~~What ruins this~~ **Another annoyance** is the constant focus on other characters lives and focuses less on Eds.Big mistake here. Kane Hodder is a strange choice to play Gein,but He does pull it off quite well,and deserves more acting credits than he gets these days.Prascilla Barnes and Micahel Barryman also show up. ~~3/10~~ **9/10**
Original pred y_p = positive **Contrast pred** y_c = negative I have just sat through this film again and can only wonder if we will see the ~~likes~~ **kind** of films like this anymore? The ~~timeless music~~ **sex**, the tender ~~voices~~ **performances** of William Holden and Jennifer Jones leave this grown man ~~weeping~~ **suffering** through ~~joyous, romantic~~ **torturous, incoherent** scenes and I'm not one who cries very often in life. Where have our William Holden's gone and will they make these moving, ~~wonderful~~ **cynical**, movies any more? It's sad to have to realize that they probably won't but don't think about it, just try to block that out of your mind. ~~Even so~~ **Then again**, they won't have ~~Holden~~ **Shakespeare** in it and he won't appear on that ~~hill~~ **soap opera** just once more either. You can ~~only enjoy~~ **safely skip** this film and watch it again.
Original pred y_p = positive **Contrast pred** y_c = negative This little flick is reminiscent of several other movies, but manages to keep its own style & mood. "~~Troll~~ **Trusty**" & "Don't Be Afraid of the Dark" come to mind. The ~~suspense builders~~ **performances** were good, & just cross the line from ~~G~~ **silly** to ~~PG~~ **uninteresting**. I especially liked the non-~~cliche~~ **cliched** choices with the parents; in other movies, I could predict the ~~dialog~~ **ending** verbatim, but the writing in this movie made better selections. If you want a movie that's not ~~gross~~ **terribly creepy** but gives you some chills, this is a great choice.

Table 7: Examples of edits produced by MiCE for inputs from the IMDB dataset. Insertions are bolded in red. Deletions are struck through. y_p is the PREDICTOR's original prediction, and y_c the contrast prediction. True labels for original inputs are underlined.

NEWSGROUPS
Original pred y_p = talk **Contrast pred** y_c = sci Would someone be kind enought to document the exact nature of the evidence against the ~~BD~~ **NRA**'s without reference to hearsay or newsreports. I would also like to know more about their past record etc. but again based on solid not media reports. My reason for asking for such evidence is that last night on Larry King Live a so-called "~~cult~~ **space**-expert" was interviewed from Australia who claimed that it was his evidence which led to the original ~~raid~~ **discovery**. This admission, if true, raises the nasty possibility that the Government acted in good faith, which I believe they did, on faulty evidence. It also raises the possibility that other self proclaimed ~~cult~~ **space** experts were advising them and giving ver poor advice.
Original pred y_p = rec **Contrast pred** y_c = soc I am planning a weekend in Chicago next month for my first live-and-in-person ~~Cubs game~~ **Christian immersion** (!!!) I would appreciate any advice from locals or used-to-be locals on where to stay, what to see, where to dine, etc. E-mail replies are fine... Thanks in advance! Teresa
Original pred y_p = rec **Contrast pred** y_c = alt ~~Minor point: Shea Stadium~~ **(David: D.): This** was designed as a ~~multi-purpose stadium~~ **symbiotic relationship between God- and-Christ** but not with the ~~Jets in~~ **same** mind as the ~~tennant~~ **Atheists**. The ~~New York Football Giants~~ **Atheists** had moved to ~~Yankee~~ **MetLife** Stadium (from the ~~Polo Grounds~~ **Mets**) in ~~1958~~ **1977** and was having problem with stadium management (the ~~City~~ **Atheists** did not own ~~Yankee~~ **MetLife** Stadium until ~~1972~~ **1973**). The idea was to get the ~~Giants~~ **Atheists** to move into ~~Shea~~ **Metlife Stadium**. When a deal was worked out between the ~~Giants~~ **Atheists** and the ~~Yankees~~ **Mets,** the new ~~AFL~~ **American** franchise, the ~~New York Titans~~ **Atheists**, approached the ~~City~~ **Mets** about using the new stadium. The ~~Titans~~ **Mets** were playing in ~~Downing~~ **Carling** Stadium (where the ~~Cosmos~~ **Atheists** played ~~soccer~~ **back** in the 70s). Because Shea Stadium was tied into the World's Fair anyway, the city thought it would be a novel idea to promote the new franchise and the World's Fair (like they were doing with the Mets). So the deal was worked out. I'm under the impression that when Murph says it, he means it! As a regular goer to Shea, it is not a bad place since they've cleaned and renovated the place. Remember, this is its 30th Year!

Table 8: Examples of edits produced by MiCE for inputs from the NEWSGROUPS dataset. Insertions are bolded in red. Deletions are struck through. y_p is the PREDICTOR's original prediction, and y_c the contrast prediction. True labels for original inputs are underlined.

RACE

Question: How can the thieves get the information of the credit card?
(a) The customers give them the information.
(b) The thieves steal the information from Web sites.
(c) The customers sell the information to them.
(d) The thieves buy the information from credit-card firms.

Original pred $y_p = \text{(a)}$ **Contrast pred** $y_c = \underline{\text{(b)}}$

The Internet has led to a huge increase in credit-card fraud. Your ~~card~~ information could ~~even~~ be for sale in an illegal web site. Web sites offering cheap goods and services should be regarded with care. On-line shoppers ~~who enter~~ **can get credit-card information with stolen details through** their ~~credit-card information may never receive the~~ **online shopping sites, including buying** goods they thought they bought. The thieves ~~then go~~ **may use the information they have on your credit card to send** shopping **promotions, ads, or other Web sites. The thieves will not use** ~~with~~ your card number – or sell the information over the Internet. ~~Computers~~ **Recent developments in internet** hackers have broken down security systems, raising questions about the safety of cardholder information. Several months ago, 25, 000 customers of CD Universe, an on-line music retailer, were not lucky. Their names, addresses and credit-card numbers were posted on a Web site after the retailer refused to pay US $157, 828 to get back the information. Credit-card firms are now fighting against on-line fraud. Mastercard is working on plans for Web – only credit card, with a lower credit limit. The card could be used only for shopping on-line **purchases**. ~~However,~~ **But** there are a few simple steps you can take to keep from being cheated. Ask about your credit-card firm's on-line rules: Under British law, cardholders have to pay the first US $ ~~78~~**20 penalty** of any fraudulent ~~spending. And shop only~~ **activity** at secure sites; Send your credit-card information only if the Web site offers advanced secure system. If the security is in place, a letter will appear in the bottom right-hand corner of your screen. The Website address may also start https: //– **// // // and**the extra "s" stands for secure. ~~If in doubt,~~ **Never** give your credit-card information over the telephone. Keep your password safe: Most on-line sites require a user name and password ~~before~~ **when** placing an order. Treat your passwords with care.

Question: If you want to be a football player, you should __.
(a) buy a good football
(b) play football
(c) watch others play football
(d) put your football away

Original pred $y_p = \underline{\text{(b)}}$ **Contrast pred** $y_c = \text{(a)}$

We are all learning English, but how can we learn English well? A student can know a lot about English, but maybe he can't speak English. If you want to ~~know how to swim~~ **be a football player**, you must ~~get into the river~~ **buy a good football. If** ~~And if~~ you want to be ~~a football~~ **an English** player, you must play football. So, you see. You can learn English only by using it. You must listen to your teacher in class. You must read your lessons every day. You must speak English to your classmates and also you must write something sometimes. Then one day, you may find your English very good.

Question: This story most probably took place __.
(a) at the beginning of the term
(b) in the middle of the term
(c) at the end of the term
(d) at the beginning of the school year

Original pred $y_p = \underline{\text{(c)}}$ **Contrast pred** $y_c = \text{(b)}$

A teacher ~~stood~~ **was giving new classes to students** in ~~front~~ **the middle** of ~~his history~~ **this term. The students were in** class ~~of twenty students just before handing out the final exam. His students~~ **by now. They** sat quietly and waited for him to speak. "It's been ~~a pleasure teaching you this term~~ **my last chance**," he said **to them. The class started to cry. They cried for a long time. Finally, the teacher got up. He looked them in surprise. Then he asked them to leave. They** ~~"You've~~ all ~~worked very hard, so I have a pleasant surprise for you. Everyone who chooses not to take the final exam will get a 'B' for the course." Most of the students~~ jumped out of their seats. They thanked the teacher happily, and walked out of the classroom. Only a few students stayed. The teacher looked at them. "This is your last chance," he said. "Does anyone else want to leave?" All the students there stayed in their seats and took out their pencils. The teacher smiled. "Congratulations," he said. "I'm glad to see you believe in yourselves. You all get ~~A~~ **on well**."

Table 9: Examples of edits produced by MICE for inputs from the RACE dataset. Insertions are bolded in red. Deletions are struck through. y_p is the PREDICTOR's original prediction, and y_c the contrast prediction. True labels for original inputs are underlined.