Supplementary Material

The following section consists of supplementary material for the paper **Conversations Gone Awry: Detecting Early Signs of Conversational Failure** (Zhang et al., 2018). The tables, sections and cited work referenced in the subsequent text can be found in the main paper.

A Details on annotation procedure

The process of constructing a labeled dataset for personal attacks was challenging due to the complex and subjective nature of the phenomenon, and developed over several iterations as a result. In order to guide future work, here we provide a detailed explanation of this process, expanding on the description in Section 3.

Our goal in this work was to understand linguistic markers of conversations that go awry and devolve into personal attacks—a highly subjective phenomenon with a multitude of possible definitions.[13] To enable a concrete analysis of conversational derailment that encompasses the scale and diversity of a setting like Wikipedia talk pages, we therefore needed to develop a well-defined conceptualization of conversational failure, and a procedure to accurately discover instances of this phenomenon at scale.

Our approach started from an initial qualitative investigation that resulted in a seed set of example conversational failures. This seed set then informed the design of the subsequent crowdsourced filtering procedure, which we used to construct our full dataset.

Initial qualitative investigation

To develop our task, we compiled an initial sample of potentially awry-turning conversations by applying the candidate selection procedure (detailed in Section 3) to a random subset of Wikipedia talk pages. This procedure yielded a set of conversations which the underlying trained classifier deemed to be initially civil, but with a later toxic comment. An informal inspection of these candidate conversations suggested many possible forms of toxic behavior, ranging from personal attacks ('Are you that big of a coward?'), to uncivil disagreements ('Read the previous discussions before bringing up this stupid suggestion again.'), to generalized attacks ('Another left wing inquisition?') and even to outright vandalism ('Wikipedia SUCKS!') or simply unnecessary use of foul language.

Through our manual inspection, we also identified a few salient points of divergence between the classifier and our (human) judgment of toxicity. In particular, several comments which were machine-labeled as toxic were clearly sarcastic or self-deprecating, perhaps employing seemingly aggressive or foul language to bolster the collegial nature of the interaction rather than to undermine it. These false positive instances highlight the necessity of the subsequent crowdsourced vetting process—and point to opportunities to enrich the subtle linguistic and interactional cues such classifiers can address.

Seed set. Our initial exploration of the automatically discovered candidate conversations pointed to a particularly salient and perplexing form of toxic behavior around which we centered our subsequent investigation: personal attacks *from within*, where one of the two participants of the ostensibly civil initial exchange turns on another interlocutor. For each conversation where the author of the toxic-labeled comment also wrote the first or second comment, the authors manually checked that the interaction started civil and ended in a personal attack. The combined automatic and manual filtering process resulted in our seed set of 232 awry-turning conversations.

We additionally used the candidate selection procedure to obtain on-track counterparts to each conversation in the seed set that took place on the same talk-page; this pairing protocol is further detailed in Section 3.

Human performance. We gaged the feasibility of our task of predicting future personal attacks by asking (non-author) volunteer human annotators to label a 100-pair subset of the seed set. In this informal setting, also described in Section 6, we asked each annotator to guess which conversation in a pair will lead to a personal attack on the basis of the initial exchange. Taking the majority vote across three annotators, the human guesses achieved an accuracy of 72%, demonstrating that humans indeed have some systematic intuition for a conversation's potential for derailment.

Informing the crowdsourcing procedure. To scale beyond the initial sample, we sought to use crowdworkers to replicate our process of manually filtering automatically-discovered candidates, en-

[13] Refer to Turnbull (2018) for examples of challenges community moderators face in delineating personal attacks.

abling us to vet machine-labeled awry-turning and on-track conversations across the entire dataset. Starting from our seed set, we adopted an iterative approach to formulate our crowdsourcing tasks.

In particular, we designed an initial set of task instructions—along with definitions and examples of personal attacks—based on our observations of the seed set. Additionally, we chose a subset of conversations from the seed set to use as *test questions* that crowdworker judgements on the presence or absence of such behaviors could be compared against. These test questions served as anchors to ensure the clarity of our instructions, as well as quality controls. Mismatches between crowdworker responses and our own labels in trial runs then motivated subsequent modifications we made to the task design. The crowdsourcing jobs we ultimately used to compile our entire dataset are detailed below.

Crowdsourced filtering

Based on our experiences in constructing and examining the seed set, we designed a crowdsourcing procedure to construct a larger set of personal attacks. Here we provide more details about the crowdsourcing tasks, outlined in Section 3. We split the crowdsourcing procedure into two jobs, mirroring the manual process used to construct the seed set outlined above. The first job selected conversations ending with personal attacks; the second job enforced that awry-turning conversations start civil, and that on-track conversations remain civil throughout. We used the CrowdFlower platform to implement and deploy these jobs.

Job 1: Ends in personal attack. The first crowdsourcing job was designed to select conversations containing a personal attack. In the annotation interface, each of three annotators was shown a candidate awry-turning conversation (selected using the procedure described in Section 3). The suspected toxic comment was highlighted, and workers were asked whether the highlighted comment contains a personal attack—defined in the instructions as a comment that is "rude, insulting, or disrespectful towards a person/group or towards that person/group's actions, comments, or work." We instructed the annotators not to confuse personal attacks with civil disagreement, providing examples that illustrated this distinction.

To control the quality of the annotators and their responses, we selected 82 conversations from the seed set to use as *test questions* with a known label. Half of these test questions contained a personal attack and the other half were known to be civil. The CrowdFlower platform's quality control tools automatically blocked workers who missed at least 20% of these test questions.

While our task sought to identify personal attacks towards other interlocutors, trial runs of Job 1 suggested that many annotators construed attacks directed at other targets—such as groups or the Wikipedia platform in general—as personal attacks as well. To clarify the distinction between attack targets, and focus the annotators on labeling personal attacks, we asked annotators to specify *who* the target of the attack is: (a) someone else in the conversation, (b) someone outside the conversation, (c) a group, or (d) other. The resultant responses allowed us to filter annotations based on the reported target. This question also plays the secondary role of ensuring that annotators read the entire conversation and accounted for this additional context in their choice.

In order to calibrate annotator judgements of what constituted an attack, we enforced that annotators saw a reasonable balance of awry-turning and on-track conversations. By virtue of the candidate selection procedure, a large proportion of the conversations in the candidate set contained attacks. Hence, we also included 804 candidate on-track conversations in the task.

Using the output of Job 1, we filtered our candidate set to the conversations where *all three annotations* agreed that a personal attack had occurred. We found that unanimity produced higher quality labels than taking a majority vote by omitting ambiguous cases (e.g., the comment "It's our job to document things that have received attention, however ridiculous we find them." could be insulting towards the things being documented, but could also be read as a statement of policy).[14]

Job 2: Civil start. The second crowdsourcing job was designed to enforce that candidate awry-turning conversations start civil, and candidate on-track conversations remain civil throughout. Each of three annotators was shown comments from both on-track and awry-turning conversations that had already been filtered through Job 1. They were asked whether any of the displayed comments were toxic—defined as "a rude, insulting, or

[14]This choice further sacrifices recall for the sake of label precision, an issue that is also discussed in Section 7.

disrespectful comment that is likely to make someone leave a discussion, engage in fights, or give up on sharing their perspective." This definition was adapted from previous efforts to annotate toxic behavior (Wulczyn et al., 2016) and intentionally targets a broader spectrum of uncivil behavior.

As in Job 1, we instructed annotators to not confound civil disagreement with toxicity. To reinforce this distinction, we included an additional question asking them whether any of the comments displayed disagreement, and prompted them to identify particular comments.

Since toxicity can be context-dependent, we wanted annotators to have access to the full conversation to help inform their judgement about each comment. However, we were also concerned that annotators would be overwhelmed by the amount of text in long conversations, and might be deterred from carefully reading each comment as a result. Indeed, in a trial run where full conversations were shown, we received negative feedback from annotators regarding task difficulty. To mitigate this difficulty without entirely omitting contextual information, we divided each conversation into snippets of three comments each. This kept the task fairly readable while still providing some local context. For candidate awry-turning conversations, we generated the snippets from all comments except the last one (which is known from Job 1 to be an attack). For on-track conversations, we generated the snippets from all comments in the conversation.

We marked conversations as toxic if at least three annotators, across all snippets of the conversation, identified at least one toxic comment. As in Job 1, we found that requiring this level of consensus among annotators produced reasonably high-quality labels.

Overall flow. To compile our full dataset, we started with 3,218 candidate awry-turning conversations which were filtered using Job 1, and discarded all but 435 conversations which all three annotators labeled as ending in a personal attack towards someone else in the conversation. These 435 conversations, along with paired on-track conversations, were then filtered using Job 2. This step removed 30 pairs: 24 where the awry-turning conversation was found to contain toxicity before the personal attack happened, and 6 where the on-track conversation was found to contain toxicity. We combined the crowdsourced output with the seed set to obtain a final dataset of 1,270 paired awry-turning and on-track conversations.

B Further examples of prompt types

Table 4 provides further examples of comments containing the prompt types we automatically extracted from talk page conversations using the unsupervised methodology described in Section 4; descriptions of each type can be found in Table 2. For additional interpretability, we also include examples of typical *replies* to comments of each prompt type, which are also extracted by the method.

Prompt Type	Example comments	Example replies
`Factual check`	I **don't see** how this **is** relevant. **Therefore** 10 knots **is** 18.5 miles per hour. That **does not mean** you can use this abbreviation everywhere. "Techniques" **refer** specifically to his fighting.	I **don't understand** your dispute. This **means** he **is** unlikely to qualify as an expert. They did **not believe** he will return. I **disagree**.
`Moderation`	**Please stop** making POV edits to the article. Your edits appear to be vandalism and have **been reverted**. I **have removed** edits which seem nationalistic. These mistakes should **not** be **allowed** to remain in the article.	**I've reverted** your change [...] **I've asked** them to stop. The next occurrance will **result in** a block. **Do not remove** my question.
`Coordination`	I **have been working on** creating an article. **Feel free** to correct my mistake. I **expanded** the article from a stub. **I'll make sure** to include a plot summary.	If you can do it I **would appreciate it**. I **have to go** but I'll be back later. **Ok, thanks**. **Hopefully** it will **be** fixed in a week.
`Casual remark`	**Just to** save you any issue in the future [...] **Remember** that badge I gave you? **Oh**, that**'s** fabulous, **love** the poem! **Not sure** how that last revert came in there.	**Yeah**, this has **gotten** out of hand. **Anyway**, it**'s nice** to see you took the time [...] **Yep**, that**'s cool**. I **just thought** your comment was no longer needed.
`Action statement`	**If** you have **uploaded** other media, **consider** checking the criteria. **Could** somebody **please explain** how they differ? The info **was placed** in the appropriate section. Could you **undelete** my article?	That article has been **tagged for** deletion. **I've fixed** the wording. **Replaced with** free picture for all pages. It **has been deleted by** an admin.
`Opinion`	I've been **thinking of** setting up a portal. I **am wondering** if he is not supposed to be editing here. It**'s hard** to combine these disputes.	It **seems** very much **in** the Wiki spirit. **Sounds like** a good idea. I **also think** we **need** to clarify this.

Table 4: Further examples of representative comments in the data for each automatically-extracted prompt type, along with examples of typical replies prompted by each type, produced by the methodology outlined in Section 4. Bolded text indicate common prompt and reply phrasings identified by the framework in the respective examples; note that the comment and reply examples in each row do not necessarily correspond to one another.