Does Putting a Linguist in the Loop Improve NLU Data Collection?

Alicia Parrish,[1] **William Huang,**[1] **Omar Agha,**[1] **Soo-Hwan Lee,**[1] **Nikita Nangia,**[1]
Alex Warstadt,[1] **Karmanya Aggarwal,**[2] **Emily Allaway,**[3] **Tal Linzen,**[1] **Samuel R. Bowman**[1]

[1]New York University [2]IIIT-Delhi [3]Columbia University

Correspondence: {alicia.v.parrish, nikitanangia, linzen, bowman}@nyu.edu

Abstract

Many crowdsourced NLP datasets contain systematic gaps and biases that are identified only after data collection is complete. Identifying these issues from early data samples during crowdsourcing should make mitigation more efficient, especially when done iteratively. We take natural language inference as a test case and ask whether it is beneficial to put a linguist 'in the loop' during data collection to dynamically identify and address gaps in the data by introducing novel constraints on the task. We directly compare three data collection protocols: (i) a baseline protocol, (ii) a linguist-in-the-loop intervention with iteratively-updated constraints on the task, and (iii) an extension of linguist-in-the-loop that provides direct interaction between linguists and crowdworkers via a chatroom. The datasets collected with linguist involvement are more reliably challenging than baseline, without loss of quality. But we see no evidence that using this data in training leads to better out-of-domain model performance, and the addition of a chat platform has no measurable effect on the resulting dataset. We suggest integrating expert analysis *during* data collection so that the expert can dynamically address gaps and biases in the dataset.

1 Introduction

Many popular datasets for training and evaluating natural language understanding (NLU) models consist of examples written by non-expert annotators. While it is convenient and relatively inexpensive to gather large datasets from non-expert crowdworkers, the resulting datasets often suffer from systematic gaps and artifacts. Through *post hoc* analysis, experts have identified many such problems and found that augmenting datasets with targeted examples can mitigate these issues (Yanaka et al., 2019; Min et al., 2020). Though non-expert created data is often flawed, it is easier to scale up compared to expert annotations, and so it is widely used in the creation of large training datasets. With this in mind, we investigate how to leverage expert linguistic knowledge during annotation by having linguists dynamically identify artifacts, biases, and gaps in the data, then communicate with non-expert annotators to instruct them towards annotations that address issues as they arise.

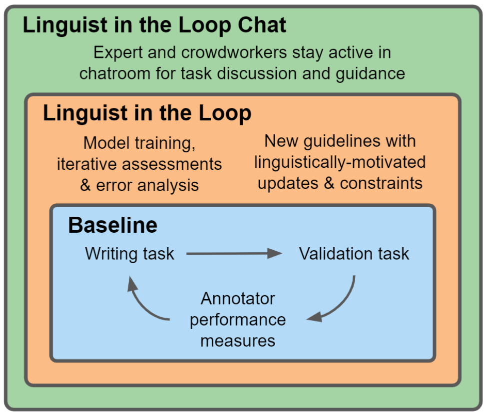

Figure 1: The three protocols compared in this study. Each crowdworker is part of only one protocol.

We focus on natural language inference (NLI; Dagan et al., 2006, i.a.), a task where the goal is to predict the label (ENTAILMENT, CONTRADICTION, NEUTRAL) that defines the logical relationship of a hypothesis to a premise (e.g., for the premise *Jenny loves all animals* the hypothesis *Jenny loves cats* is an ENTAILMENT, and *Jenny hates dogs*, a CONTRADICTION). We choose NLI because it is among the best-studied NLU tasks, with demonstrated value, but also multiple well-documented data quality issues that arise in crowdsourced data collection, many of which can be traced to a given *heuristic*. Because these heuristic-based issues are prevalent, we focus on NLI with the aim that our methodology can inform data collection for new tasks in which there are fewer known heuristics.

Previous attempts to develop more effective NLU data collection protocols have been limited in their ability to assess the efficacy of their interventions, as they often lack direct comparisons

Text: They inhabit the near-boiling water of geysers in Yellowstone, and the even hotter water in volcanic vents on the ocean floor.

☐ The definitely correct sentence does not reuse nouns, verbs, adjectives, or adverbs from the text ($0.10)

• Your definitely correct statement must not contain any of the following words: there, can, may, might, some, people

Definitely correct statement:

☐ The maybe correct sentence does not reuse nouns, verbs, adjectives, or adverbs from the text ($0.05)

• Your maybe correct statement must not contain any of the following words: often, several, many, most, some, other, will

Maybe correct statement:

☐ The definitely incorrect sentence does not reuse nouns, verbs, adjectives, or adverbs from the text ($0.05)

• Your definitely incorrect statement must not contain any of the following words: any, never, no, nothing, not/n't, only, always, all

Definitely incorrect statement:

Figure 2: Round 5 HIT with the optional *No Overlap* constraint shown.

between different collection methods. We test three levels of expert involvement: (i) a baseline group with no hands-on expert involvement ('Baseline'), (ii) a group that followed linguistically-motivated constraints that experts developed to target heuristic-based weaknesses in the data ('linguist-in-the-loop' (LitL)), and (iii) a group that extended the LitL protocol to add direct interaction with the experts, including individual-level discussion about the task on the chat platform Slack ('LitL Chat'). These three protocols are shown in Figure 1, and a task example with a constraint from LitL and LitL Chat is shown in Figure 2.

We observe that expert involvement (LitL and LitL Chat) during data collection reduces the impact of some learnable heuristics and results in a more challenging final dataset with model performances that are 5 points lower on validated data compared to Baseline. Examples in all protocols qualitatively appear equally free of noise (incorrect labels, typos, etc.), and lexical diversity increases in later rounds for the protocols with linguist intervention.[1] However, we find no evidence of better model accuracy on adversarial examples or out-of-domain datasets. Further, we do not find any benefit to providing a chatroom for crowdworkers to interact directly with linguists.

2 Related Work

NLI Data Collection Methods Large-scale human-elicited datasets include the Stanford Natural Language Inference Corpus (SNLI; Bowman et al., 2015), the Multi-genre Natural Language Inference Corpus (MNLI; Williams et al., 2018), the Chinese OCNLI corpus (Hu et al., 2020), and Adversarial NLI (ANLI; Nie et al., 2020). All four datasets use non-expert annotators to write hypotheses and annotate inference labels from pre-defined short texts, though only OCNLI and ANLI add interventions to increase data diversity. OCNLI hires students specializing in linguistics and language studies to construct hypotheses and adds data collection rounds with instructions for avoiding known sources of bias. ANLI uses a human-and-model-in-the-loop procedure to elicit examples that are progressively more difficult for their model, resulting in a dataset with a large human–model performance gap, though identifying the cause for model failure is left up to the discretion of the worker.

Efforts to improve on sentence writing tasks for NLI have yielded mostly negative results in head-to-head protocol comparisons. In an experimental comparison on different NLI crowdsourcing protocols, Vania et al. (2020) find that automatically selecting premise-hypothesis pairs for crowdworker annotation does not yield a better dataset compared to a baseline sentence writing protocol. Bowman et al. (2020) compare interventions aimed at improving NLI writing, using protocol variants that constrain the worker's task, but they do not see improvements in transfer learning results compared to a baseline protocol.

NLI Dataset Bias Several studies have identified systematic biases in NLI datasets that the models trained on them subsequently learn (often robustly). Hypothesis-only bias, where a model correctly labels the relationship of a premise-hypothesis pair based only on the hypothesis, is a well-documented issue (Poliak et al., 2018; Gururangan et al., 2018, i.a.). Lexical overlap between the premise and hypothesis is another source of bias (McCoy et al., 2019; Naik et al., 2018), as greater overlap between a premise and hypothesis is associated with a greater likelihood the label is ENTAILMENT. Sources of bias can also be due to gaps in the training data, and Sinha et al. (2020) point to the lack of syntactic understanding in NLI models as one such example, noting that models often ignore syntactic information entirely. The well-studied biases in NLI make the task a good test case for protocols designed to assess biases as data is collected.

Methods for Filling the Gaps in Datasets To collect challenging examples for NLU tasks, re-

[1] Appendix E contains a sample of validated examples.

searchers have explored altering labeled data to create targeted or adversarial examples. Kaushik et al. (2020) have crowdworkers make minimal edits to annotations to align with a revised label. Gardner et al. (2020) create contrast sets for evaluation by having experts alter already-annotated examples such that the resulting label changes. Wei and Zou (2019) use simple automatic data manipulations to augment datasets for several text classification tasks, resulting in more robust models. More linguistically sophisticated manipulations have been used to augment MNLI to improve monotonicity reasoning (Yanaka et al., 2019) and mitigate lexical overlap heuristic (Min et al., 2020). These methods are applied after crowdsourced data collection is complete, so it is not clear if the gaps they identify in a final dataset would have been avoidable if addressed during the data collection process.

Most similar to our approach, OCNLI's instructions nudge annotators towards writing examples that address *known* sources of bias. They find that encouraging annotators to follow constraints such as avoiding negation in a CONTRADICTION label results in a harder dataset. We expand on this work by introducing a wider range of constraints and assessing their effects throughout data collection. Our approach is also similar to Vidgen et al.'s (2020) human-generated hate-speech dataset. They introduce *pivots* during data collection in which they instruct annotators about how to write to fool their model. We expand on their method by qualitatively assessing the annotations to identify issues specific to our data as it is collected.

Expert Interaction with Crowdworkers Tang et al. (2019) report that direct communication among crowdworkers leads to improved task performance on image labeling, optical character recognition, and audio transcription. This suggests that collecting higher quality data is possible when workers have real-time group interaction. Other studies have reported that interaction among crowdworkers is an effective tool for limiting some forms of bias and increasing accuracy (Drapeau et al., 2016; Schaekermann et al., 2018). Roit et al. (2020), in a different strategy, gives annotators detailed feedback during training, then selects only a small number of those workers for the larger annotation task. This strategy frontloads the work of the experts and relies on the selected workers to perform the task consistently.

Despite the potential benefits of real-time interaction between crowdworkers and experts, there has not yet been a direct comparison of protocols that differ based on this variable. To our knowledge, this study is both the first to test the effect of this interaction and the first head-to-head experimental assessment of human-in-the-loop dataset collection methods, allowing us to make conclusions about the causal effects of the different interventions compared to a baseline.

3 Dataset Collection

Task Description Our task is modeled on MNLI's data collection procedure. We present workers with a text, for which they write statements they consider definitely correct, maybe correct, and definitely incorrect. Following each round of sentence writing, crowdworkers validate 500 examples from their protocol. We collect four validations for each example that we validate and use these labels plus the original one to assign a gold label based on majority vote. Examples for which no gold label can be assigned are removed from the data. We use the validated data to evaluate our models and the unvalidated data for training. Workers with a validation rate below 70% or whose validation responses fail to match the gold label at least 70% of the time are subject to disqualification. Throughout the study, we disqualified three workers from Baseline, three from LitL, and two from LitL Chat.

Pay Structure To retain crowdworkers for all five rounds, we increase the base pay of $1/HIT by $0.05 each round and pay a $20.00 bonus after the last round. To ensure we collect sufficient examples from each worker, we award a bonus worth 10% of their base pay for reaching milestones of 10, 50, and 100 HITs each round. To encourage workers to write high-quality examples, we pay a $5.00 bonus each round to workers with over 25 HITs and at least a 95% validation rate. We estimate that, with bonuses, a worker who completes 70 HITs with a high validation rate will earn $81 in Round 1 (∼$16/hr), with that rate rising to $95 in Round 5 (∼$19/hr). Workers in LitL and LitL Chat earn additional bonuses for completing challenge options ($0.05-$0.10), and workers in LitL Chat earn bonuses for participation in the chatroom ($1.50 for any engagement, $10.00 for active engagement).

3.1 Crowdworker recruitment

We use a pre-test to recruit workers via Amazon Mechanical Turk (MTurk). The pre-test is open to workers in the United States with approval rates at or above 98% and more than 1000 HITs approved. The pre-test is a sentence-writing task where workers see a premise and write hypotheses under each of the three NLI labels. To assess if workers can follow more complicated instructions, they are also asked to write one entailed sentence that uses a conjunction and one neutral sentence that does not re-use any words from the text.

We collect responses from 155 crowdworkers, of whom 145 indicate interest in completing future, similar HITs. From those 145, we read their responses and exclude 24 for failing to adequately complete the task (many due to responses that do not follow instructions). The remaining 121 crowdworkers are retained and split between the three experimental protocols in a pseudo-random way such that (i) the three workers who asked not to participate in a chat forum are placed in the Baseline or LitL protocol,[2] and (ii) groups are matched based on a 4-point rating scale of their qualitative performance on the pre-test. We ultimately had 37 annotators participate in data collection in Baseline, 30 in LitL, and 32 in LitL Chat.

3.2 Annotation Details

Crowdworkers annotate data in five rounds, with each round lasting one week. Between rounds, we conduct several planned diagnostics on our datasets to monitor the impact of our intervention and inform annotator feedback for the following round.

Annotation stage Annotators construct hypotheses based on premises taken from the SLATE subset of MNLI. SLATE hosts popular culture articles from the archives of Slate Magazine. After Round 1, we exclude premises that are shorter than six tokens based on feedback from annotators that many of the very short premises are incomplete, nonsensical, or confusing to write hypotheses for.

Diagnostic stage After each round, we fine-tune RoBERTa (Liu et al., 2019) models using data collected up to that round. We then evaluate the models on diagnostic examples from GLUE (Wang et al., 2019) and HANS (McCoy et al., 2019). The GLUE examples target different aspects of linguistic reasoning including lexical semantics, predicate-argument structure, logic, and world knowledge. HANS tests for three shallow heuristics, including lexical overlap between a premise and hypothesis. We also train and evaluate RoBERTa models using hypothesis-only inputs to assess hypothesis-only biases in the data (Gururangan et al., 2018). Finally, we assess the distribution of hypothesis lengths and the pointwise mutual information (PMI) between each word in the vocabulary and label. Hypothesis length does not appear to differ by protocol or label, so it never informs our constraints.

We use these diagnostics as well as qualitative reviews of the data to devise linguistically-motivated guidelines for the following round, allowing us to adapt feedback for crowdworkers in a structured way as the data is collected. This process is conducted by five of the authors who have graduate training in English syntax and semantics.

3.3 Constraints

Banned Words After Round 1, crowdworkers in LitL and LitL Chat are instructed not to use certain words when writing sentences for each label. In each round, 5-7 new banned words are identified based on the PMI for each label. We use PMI to identify words to ban under each label, as words with high label PMI are a major contributor to hypothesis-only bias. This constraint is mandatory in all HITs. Figure 2 shows examples of the banned words during Round 5.

Challenge Options We use constraints, framed as *challenge options* to the worker, to target heuristics that we identify in the data during the diagnostic state. By explicitly telling workers to avoid these heuristics, we aim to lower their contribution to any bias in the final dataset. We determine constraints through qualitative assessment of the data, taking into consideration syntactic diversity, lexical choice, and semantic or world-knowledge-based reasoning patterns. For example, after noticing that the majority of hypotheses relied only on the stated information from the premise in Round 1, we encouraged workers in Round 2 to focus on "background knowledge" (example in Table 3) that they know to be true, but isn't explicitly stated, such as the knowledge that Britain and America are countries on opposite sides of the world.

After Round 1, each HIT in LitL and LitL Chat lists one constraint. This task is optional for the

[2] Though a potential design confound, this was necessary and had minimal effect. *Requiring* workers to sign up for a third party service violates Amazon's terms of service, so we allow participants to opt out. Only three participants opted out of the chat (two of whom dropped out after Round 1), and many workers placed in a non-chat protocol had indicated a willingness to participate in the chat.

Constraint	Premise	Hypothesis	Label	Attempt rate LitL	Attempt rate Chat
Hypernym or hyponym	Does anyone know what happened to **chaos**?	Whatever happened to the **lack of order** is certainly a mystery.	E	22.8	23.7
Banned word in diff. label	Inflation is supposed to be a deadly poison, not a useful medicine.	**All** people believe inflation is supposed to be a useful medicine	C	43.7	27.7
Temporal reasoning	John Kasich dropped his presidential bid.	They said that **earlier**, John Kasich had dropped his presidential bid.	E	34.1	10.0
Synonym or antonym	2) This particular instance of it **stinks**.	This instance is perceived to be **a good thing**.	C	39.5	24.5
All overlap	News argues that most of America's 93 million volunteers aren't doing much good.	News argues that volunteers aren't doing much good.	E	21.8	30.4
Register change	First, the horsemen brought out a teaser horse.	Teaser horses are commonly thought to be both entertaining and tragic.	N	25.3	15.0
No overlap	and she doesn't floss while driving.	The woman has an automated car.	N	29.2	22.3
Relative clause	Sun Ra's spaceships did not come, as it were, out of nowhere.	The spaceships **that belong to Sun Ra** came out of nowhere	C	35.0	24.3
Reverse argument order	After an **inquiry** regarding **Bob Dole**'s ...	It is illegal for **Bob Dole** to receive **inquiries**.	N	36.7	29.4
Grammar change	The Bush campaign **has** a sweet monopoly on that.	The Obama campaign **had** a sweet monopoly on that.	C	22.6	13.4
Sub-part	He was crying like his mother had just walloped **him**.	He cried a lot, as though he were walloped on **his behind**.	E	23.2	19.1
Background knowledge	In both **Britain and America**, the term covers nearly everybody.	The term generally applied to **countries in two opposite sides of the world**.	E	32.9	15.9

Table 1: Sentence pairs displaying each challenge option. Where applicable, relevant contrasts are bolded. Examples are randomly drawn from data that passed validation on the constraint with the restriction that both sentences be fewer than 80 characters ($\sim$ 32% of the data). The last column shows the percentage of the challenges attempted.

workers, as some constraints are incompatible with some examples. The 12 challenge options are defined in Appendix A, with examples in Table 1.

3.4 Protocols

Baseline Protocol Our Baseline protocol follows the basic task description in the beginning of §3 and does not include any direct expert involvement.

Linguist-in-the-Loop (LitL) Protocol LitL extends the Baseline protocol with constraints (described in §3.2). As the constraints make the task more difficult, we award bonuses to workers who indicate that they attempted a challenge option. The bonus is \$0.05-\$0.10 per example, determined by the linguists' assessment of the difficulty. For example, the *No Overlap* constraint is more difficult in entailment examples than neutral, so a *No Overlap* entailment example receives a higher bonus.

During the validation round, examples with challenge options are validated for whether they adhere to the given constraint, regardless of whether the annotator indicated that they had attempted the challenge. For any worker whose validation rate on the bonus challenges is below 50%, we contact them to explain the source of their errors.

LitL Chat Protocol We provide direct communication with expert linguists on Slack. We encourage workers to ask task-specific questions for anything they find challenging or confusing. Most questions seek to clarify if a certain strategy 'counts' as adhering to a constraint. Feedback given via email in the LitL protocol is instead given via direct message in Slack, unless the worker initiates contact over email. Additionally, at the beginning of Rounds 3–5, we identify creative examples written in a previous round and post them to Slack for inspiration, with a brief comment.

3.5 Annotator Performance

Inter-Annotator Agreement Baseline shows the highest inter-annotator agreement with $\kappa = 0.71$, while LitL and LitL Chat have 0.64 and 0.63, respectively. All three protocols meet the standard threshold for "substantial agreement." Validation rates are 93.7% for Baseline, 89.76% for LitL, and 91.36% for LitL Chat. LitL and LitL Chat may have slightly lower validation rates than Baseline because the constraints lead to challenging examples, making the validator's task more difficult.

Proportion of Constraints Attempted The attempt rate of bonus challenges differs substantially between constraints (Table 1). Overall, more abstract categories (e.g., background knowledge) are attempted less often than more concrete constraints. There are also differences by protocol, as LitL had a higher attempt rate than LitL Chat. One reason for this difference may be that workers in LitL Chat were more selective about identifying good examples on which to apply the constraints. Supporting this possibility, we find that LitL Chat had higher constraint validation rates than LitL in Rounds 4 and 5, indicating that workers in LitL Chat adhered to the constraints more accurately after practice.

Use of Slack The total number of active workers in Slack fell from 23 in Round 1 to just 16 by Round 4.[3] The total number of messages sent also fell with each round, going from about 215 posts and replies in Round 1 to 162 in Round 4. It may be that workers rely on the chat less as they become more familiar with the task.

4 Experiments

For each round and protocol, we collect 3.5k examples and use the 500 validated examples (§3.4) as validation data and the remaining 3k for training.[4] We then fine-tune a RoBERTa$_{\text{Lg}}$ model on all data accumulated up to Round n. For example, the Round 2 model is trained on examples from Rounds 1 and 2 with training and validation sizes of 6k and 1k, respectively. We also fine-tune a RoBERTa$_{\text{Lg}}$ model previously trained on MNLI (RoBERTa$_{\text{Lg+MNLI}}$) and find similar results (details in Appendix B). After each round, we evaluate our models on the diagnostics described in §3.2.

After the final round of data collection, we evaluate models trained on our data on MNLI-mismatched (Williams et al., 2018) and ANLI (Nie et al., 2020). The MNLI corpus includes two evaluation sets, MNLI-matched and MNLI-mismatched, with examples sourced from different genres. We evaluate on MNLI-mismatched, as we source our premise sentences from an MNLI-matched genre. Evaluating on held-out sets allows us to test if our interventions lead to increased model accuracy on datasets generated through different protocols or from different sources while ensuring that we do not overly tune our feedback to these benchmarks.

We estimate average accuracy and confidence intervals by fine-tuning 10 additional models with a sample of 90% of the collected training data. We use the best hyperparameters for each protocol and round from our hyperparameter search described below. In sampling the data, we first sort the data by annotator and successively remove 10% of examples, allowing us to study variation among annotators while controlling for training set size.

Implementation To fine-tune our models, we perform a grid search over learning rate $\in \{5e-6, 1e-5, 2e-5, 3e-5\}$ and batch size $\in \{16, 32\}$ and use the hyperparameters yielding the best in-domain validation accuracy. We train for 20 epochs, since each round of data collection yields 3k training examples, and longer training has been shown to help smaller training sets (Zhang et al., 2020). Our code is based on `jiant` (Phang et al., 2020), which uses PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020).

5 Results

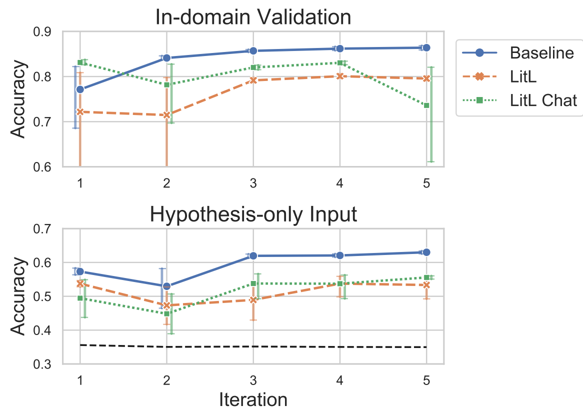

Figure 3: Performance of RoBERTa$_{\text{Lg}}$ fine-tuned on data collected through different protocols on in-domain validation data trained with either the normal sentence pairs (top) or hypothesis-only (bottom) input. Higher hypothesis-only accuracy indicates more bias. For each round, we include training and validation data *accumulated* up to Round n. Dashed black line marks average majority class baseline across protocols.

Evaluation Difficulty We test whether data collected with expert intervention leads to a more challenging test set by comparing in-domain performance for each protocol for RoBERTa$_{\text{Lg}}$, using training and validated evaluation data accumulated up to Round n (Figure 3). This allows us to study the characteristics of an iteratively collected corpus using n rounds of each protocol. We see that LitL

[3] Round 5 was even lower, but spanned the US Thanksgiving holiday, which likely artificially lowered participation.

[4] Data and code: github.com/Alicia-Parrish/ling_in_loop

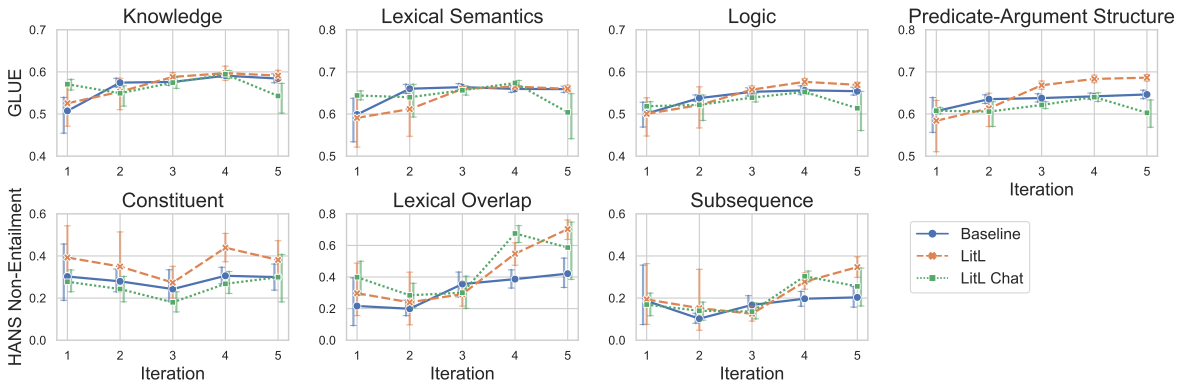

Figure 4: Performance of RoBERTa$_{\mathrm{Lg}}$ fine-tuned on data collected through different protocols on the GLUE diagnostic set (top) and HANS non-entailment examples (bottom).

Round	Baseline	LitL	LitL Chat
R1	$84_{(1.2)}$	$82_{(1.1)}$	$86_{(1.1)}$
R2	$85_{(0.6)}$	$81_{(1.0)}$	$85_{(0.9)}$
R3	$86_{(0.7)}$	$82_{(0.8)}$	$84_{(0.9)}$
R4	$86_{(0.8)}$	$83_{(0.7)}$	$84_{(1.0)}$
R5	$87_{(0.5)}$	$82_{(0.6)}$	$84_{(0.9)}$

Table 2: Average performance of RoBERTa$_{\mathrm{Lg}}$ fine-tuned on MNLI over 10 random restarts on validated examples accumulated up to Round n. Values in parentheses indicate standard deviation of performance.

and LitL Chat performance falls below Baseline after the introduction of linguistically-informed constraints in Round 2. Table 2 shows a similar trend – performance from RoBERTa$_{\mathrm{Lg}}$ fine-tuned **only on MNLI** on the same validation sets decreases or remains relatively low for LitL and LitL Chat, while performance on Baseline increases as more data is collected. As we evaluate on validated examples, it is unlikely that this lower performance is due to noise in the data. Rather, these findings indicate we are able to create more challenging evaluation data using the LitL and LitL Chat interventions.

Hypothesis-Only Bias We test whether the data collected with linguist intervention leads to a reduction in hypothesis-only bias by comparing accuracy for each protocol for RoBERTa$_{\mathrm{Lg}}$ trained on hypothesis-only input, where lower accuracy suggests less bias in the data (Figure 3). Both LitL and LitL Chat result in lower hypothesis-only bias than Baseline, and this gap widens in later rounds. To assess whether this widening from Round 1 to 5 is statistically reliable, we conduct a two-way ANOVA of round by protocol, which yields a significant interaction ($p = 0.049$), indicating that while hypothesis-only performance increases for all protocols with more training examples, this increase in bias is significantly reduced in LitL and LitL Chat compared to Baseline. The lower bias in LitL and LitL Chat may be due to the lower average word-label PMI, which increases over rounds for Baseline while consistently falling in both LitL and LitL Chat.[5] However, for all protocols, accuracies are still above chance performance, leaving room to further reduce hypothesis-only bias.

Diagnostic Sets We use diagnostic tests to evaluate whether fine-tuning on data collected with linguist involvement leads to a model that has higher performance on challenge test sets. Figure 4 shows model performance on the GLUE diagnostic set and HANS non-entailment examples. A two-way ANOVA of round by protocol does not reveal any significant interactions or main effects for GLUE. For HANS, we see higher accuracy from LitL and LitL Chat for Lexical Overlap and Subsequence examples in Rounds 4 and 5 after introducing *No* and *All Overlap* constraints, though the interaction is only significant with RoBERTa$_{\mathrm{Lg+MNLI}}$ ($p = 0.0021$ and $p = 0.0017$ for Lexical Overlap and Subsequence, respectively). Performance on HANS entailment examples are in line with McCoy et al. (2019) with median accuracies of 90% or higher (Appendix D).

To investigate whether rates of lexical overlap differ by protocol, we assess classification accuracy for a linear model trained only on the example's overlap rate, defined as the proportion of words in the hypothesis that are also in the premise. We observe that the potential bias introduced from overlap rate is strongest in the Baseline protocol, which

[5] A two-way ANOVA again reveals a significant interaction of protocol by round ($p = 0.022$) on word-label PMI values.

performs 9.52 points above majority class guessing, while LitL and LitL Chat perform 8.06 and 6.88 points above majority class guessing, respectively.

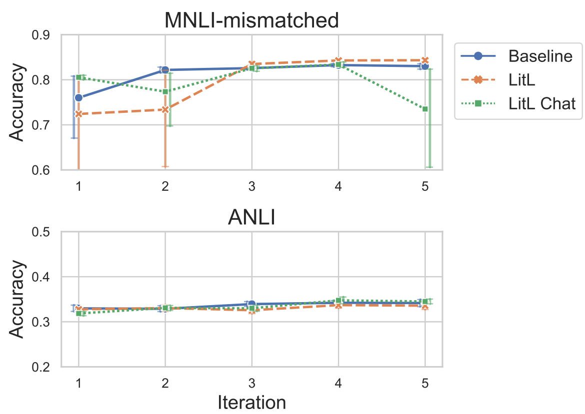

Figure 5: Performance of RoBERTa$_{\mathrm{Lg}}$ fine-tuned on data collected through different protocols on MNLI-mismatched (top) and ANLI (bottom).

Held-Out Evaluation Sets We next test whether models fine-tuned on data collected with linguist involvement show better out-of-domain performance. Figure 5 shows that there is little difference in ANLI and MNLI-mismatched performance among all protocols. We perform a more granular analysis on ANLI examples using the tags from Williams et al. (2020) and again find no clear effect of protocol (details in Appendix C). Even though our interventions reduce hypothesis-only bias and improve model performance on HANS non-entailment examples, we have no evidence that these benefits transfer to out-of-domain examples or examples from adversarial protocols.

6 Considerations in Choosing a Protocol

In broad terms, we observe a benefit from dynamically updating annotator guidelines to address gaps and biases observed during data collection. This procedure increased the average cost per example by 4.1% over an average base cost of $0.367. In an exit survey, over 50% of annotators in LitL and LitL Chat indicated that they would have completed more optional challenge examples if the pay had been higher. We offered $0.05 to $0.10 per example, but given the somewhat low rate at which annotators chose to attempt the challenges (28.6% and 21.2% of examples for LitL and LitL Chat, respectively), we find it likely that increasing the amount offered per example would have increased participation, potentially also increasing the benefits observed in model performance. We recommend that future work using challenge options offer bonuses worth at least 15% of the base pay.

Cost of Linguist Involvement The iterative analyses and updates to the guidelines used for LitL and LitL Chat protocols took 10–12 hours of expert time per week, compared to a base time cost of about one hour per week for just monitoring task completion. The use of Slack nearly doubled the expert time needed, adding an *additional* 8–10 hours each week just for LitL Chat over LitL, even after taking into account the slight reduction in time spent replying to email questions that shifted to Slack. If we value linguist time at $40/hr, then this raises the final price per example to $0.378 in Baseline, with LitL 31.2% higher, and LitL Chat 58.5% higher than baseline. Despite the extra time spent with annotators, we did not observe any measurable benefit to this more hands-on intervention.

Qualitative Considerations Though many annotators in LitL Chat expressed that they enjoyed the extra communication and help, annotators from LitL and LitL Chat rated the task as 'more enjoyable' than typical MTurk tasks at nearly identical rates (85.2% and 87.5% respectively, compared to 67.7% in Baseline). Ratings of the difficulty of writing and validation tasks were also nearly identical between the LitL and LitL Chat protocols.

7 Conclusion

We conclude that for some tasks, it is beneficial to integrate expert analysis of data *during* data collection, so that the expert can dynamically update guidelines and constraints based on existing gaps and biases in the dataset. Through a controlled experiment, we have shown that this expert involvement leads to higher accuracy in measures related to the dataset gaps and biases that were identified and targeted, and the iterative procedure allowed us to identify new areas of weakness at each round. As we did not observe any increases in out-of-domain accuracy, we conclude that there is no more general benefit from our interventions beyond the targeted areas of weakness. Future work could extend this protocol to identify additional interventions that would lead to datasets with better generalizability. Finally, we find no evidence to support the claim reported by some studies that one-on-one interactions between experts and crowdworkers is beneficial in more challenging tasks.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. Collecting entailment data for pretraining: New protocols and negative results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. OCNLI: Original chinese natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3512–3526.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2340–2353. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2020. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Wei Tang, Ming Yin, and Chien-Ju Ho. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1794–1805. ACM.

Clara Vania, Ruijie Chen, and Samuel R. Bowman. 2020. Asking Crowdworkers to Write Entailment Examples: The Best of Bad Options. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987*.

A List of challenge options

Lexical Options

- **Temporal reasoning** (Round 2): The hypothesis should reference two separate time points.
- **Restricted word in different label** (Round 2): The hypothesis should contain a word that is banned for a different label.
- **Hypernym or hyponym** (Rounds 2 & 3): The hypothesis should contain a hypernym or hyponym (a more or less specific word or phrase) of a word in the premise.
- **Synonym or antonym** (Rounds 2 & 3): The hypothesis should contain a synonym or antonym of a word in the premise.
- **No overlap** (Rounds 4 & 5): The hypothesis should use none of the content words appearing in the premise. Content words are nouns, verbs, adjectives, and adverbs.
- **All overlap** (Rounds 4 & 5): The hypothesis should only use content words that appear in the premise. Introducing new function words is allowed, as is changing grammatical features of the content words.

Syntactic Options

- **Relative clause** (Round 2): The hypothesis should contain a relative clause. A relative clause is a noun that is described by a phrase that begins with words like *who* or *that*.
- **Reverse argument order** (Rounds 2 & 3): The hypothesis should contain a pair of noun phrases from the premise in reverse order.
- **Grammar change** (Round 4): The hypothesis should change a grammatical element of the premise, such as tense, number, or gender on a pronoun.

World Knowledge Options

- **Background knowledge** (Rounds 2 & 4): The hypothesis should target background facts or general knowledge that workers can infer from the premise.
- **Sub-part** (Round 3): The hypothesis should refer to something that is a part of an entity in the premise. For example, sub-parts of a *bus* include its *steering wheel*, and *engine*.
- **Register change** (Round 5): The hypothesis should differ from the original text in its level of formality.

B MNLI-Pretrained RoBERTa Results

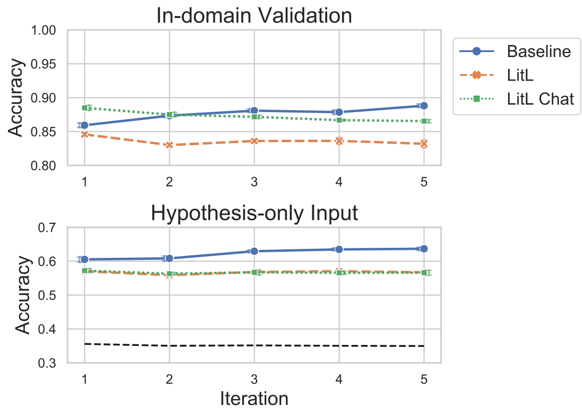

Figure 6: Performance of $\text{RoBERTa}_{\text{Lg+MNLI}}$ fine-tuned on data collected through different protocols on in-domain validation data trained with either the full example (top) or hypothesis-only (bottom) input. Higher hypothesis-only accuracy indicates more bias. For each round, we include training and validation data *accumulated* up to Round n. Dashed black line marks average majority class baseline across protocols.

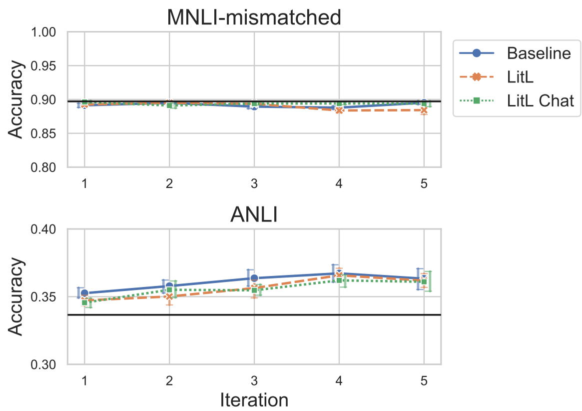

Figure 7: Performance of $\text{RoBERTa}_{\text{Lg+MNLI}}$ fine-tuned on data collected through different protocols on MNLI-mismatched (top) and ANLI (bottom). The black line for MNLI-mismatched and ANLI indicates performance of $\text{RoBERTa}_{\text{Lg}}$ fine-tuned on MNLI alone.

We fine-tune a $\text{RoBERTa}_{\text{Lg}}$ model previously trained on MNLI ($\text{RoBERTa}_{\text{Lg+MNLI}}$) on the same sets of training data used for the $\text{RoBERTa}_{\text{Lg}}$ analyses. We find similar trends to those from fine-tuning $\text{RoBERTa}_{\text{Lg}}$ and report them in the same set of plots here.

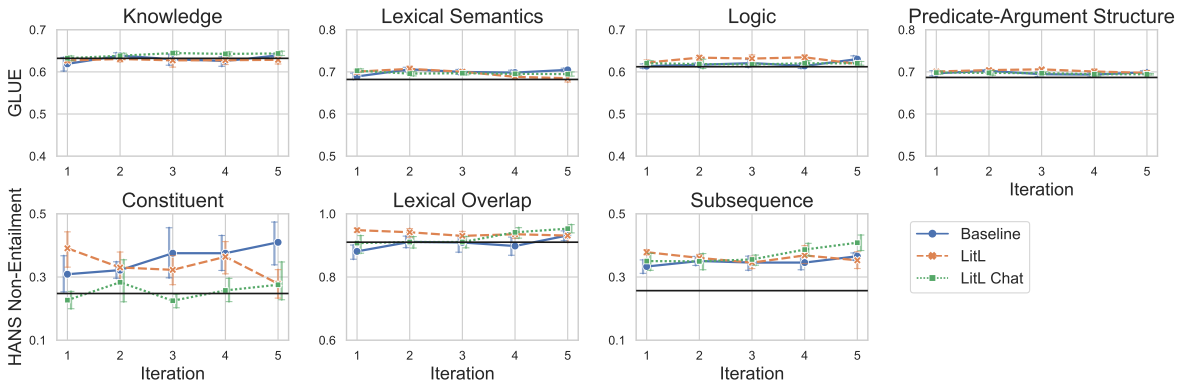

Figure 8: Performance of RoBERTa$_{\text{Lg+MNLI}}$ fine-tuned on data collected through different protocols on the GLUE diagnostic set (top) and HANS non-entailment examples (bottom). The black line indicates performance of RoBERTa$_{\text{Lg}}$ fine-tuned on MNLI alone.

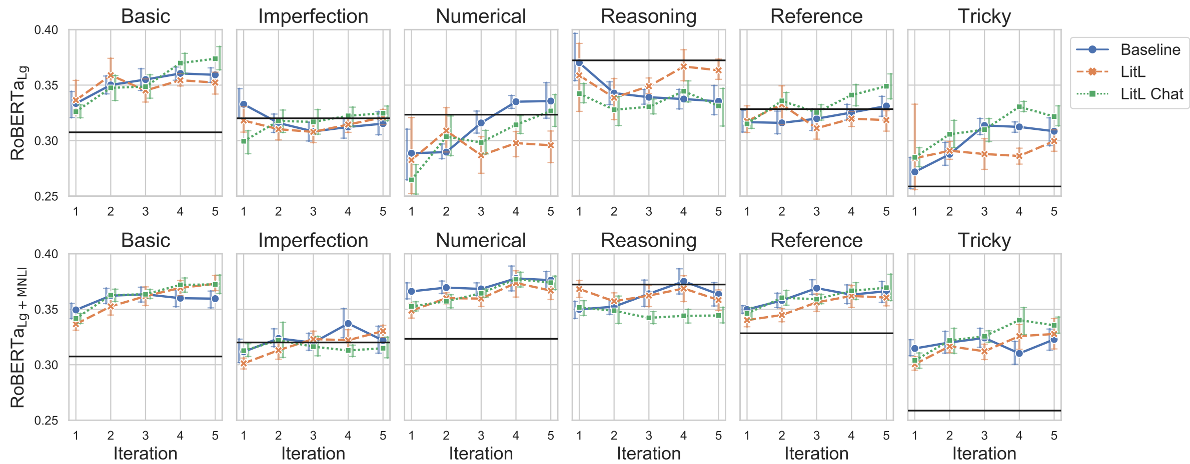

Figure 9: Performance of RoBERTa$_{\text{Lg}}$ (top) and RoBERTa$_{\text{Lg+MNLI}}$ (bottom) fine-tuned on data collected through different protocols on ANLI by reasoning tag from Williams et al. (2020). The black line indicates performance of a RoBERTa$_{\text{Lg}}$ trained on MNLI **alone**.

Figure 6 shows the performance of RoBERTa$_{\text{Lg+MNLI}}$ fine-tuned using either the full example or hypothesis-only input. For both types of input, we see a performance gap between Baseline and our intervention protocols. We perform a two-way ANOVA of round by protocol to see if this performance gap significantly changes between rounds 1 and 5 and find a significant interaction ($p < 0.001$ for both full example and hypothesis-only input). For full example input, this indicates that our interventions create more challenging evaluation data. For hypothesis-only performance, Baseline performance increases while LitL and LitL Chat remain relatively unchanged, indicating that our interventions mitigate stronger hypothesis-only bias in NLI datasets as new data is collected.

Figure 7 shows the performance of RoBERTa$_{\text{Lg+MNLI}}$ fine-tuned on different protocols on MNLI-mismatched and ANLI. We find no significant difference among protocols for either held-out set.

Figure 8 shows the performance of RoBERTa$_{\text{Lg+MNLI}}$ fine-tuned on data from different protocols on the GLUE diagnostic set and HANS non-entailment examples. For the GLUE diagnostic set, we do not find any significant difference among protocols. For the HANS examples, we perform another two-way ANOVA of round by protocol and find significant interaction terms for all HANS categories (p = 0.0018, 0.0021, 0.0017 for Constituent, Lexical Overlap, and Subsequence, respectively). For Lexical Overlap and Subsequence, these findings indicate our interventions lead to higher accuracy compared to Baseline. However, we see the

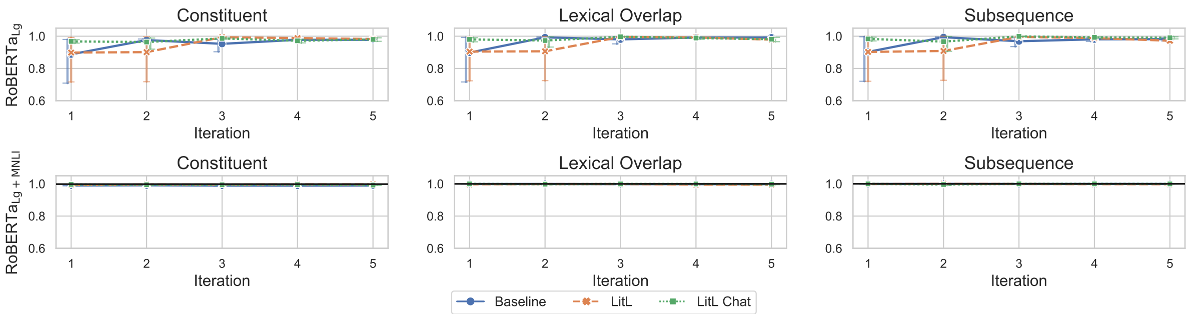

Figure 10: Performance of $\text{RoBERTa}_{\text{Lg}}$ (top) and $\text{RoBERTa}_{\text{Lg+MNLI}}$ (bottom) fine-tuned on data collected through different protocols on HANS entailment examples. The black line indicates performance of a $\text{RoBERTa}_{\text{Lg}}$ trained on MNLI **alone**.

opposite from Constituent examples with both intervention protocols performing worse than Baseline.

C ANLI Performance by Reasoning Type

We test whether *any* of the reasoning tags in ANLI (Williams et al., 2020) reveal an area where data collection with linguist involvement leads to improved model performance. Figure 9 shows the performances of $\text{RoBERTa}_{\text{Lg}}$ and $\text{RoBERTa}_{\text{Lg+MNLI}}$ on ANLI by reasoning tag. Similar to our findings in Figures 5 and 7, we do not find any increases in accuracy from our interventions for any reasoning tags.

D HANS Entailment Peformance

On the entailment subset of HANS, models typically achieve accuracies near 100% McCoy et al. (2019). This is because the three heuristics in HANS target instances that lead to a greater likelihood of the model choosing ENTAILMENT compared to NEUTRAL or CONTRADICTION, and thus the non-entailment portion of HANS is the challenge set. Figure 10 shows the performance of $\text{RoBERTa}_{\text{Lg}}$ and $\text{RoBERTa}_{\text{Lg+MNLI}}$ fine-tuned on our data and tested on HANS entailment examples. For $\text{RoBERTa}_{\text{Lg}}$, variability in performance reduces in later rounds as the training set size grows with 3k examples per round, though median performances for all rounds are still 90% or higher. For $\text{RoBERTa}_{\text{Lg+MNLI}}$, accuracies are near 100%, consistent with McCoy et al.'s findings.

E Examples of Collected Data

In order to show a representative sample of the validated data, we randomly sample premises from Round 5 data for which annotations exist in all three labels for each protocols (roughly 45% of that round's validated data). Five such examples are presented in Table 3. Example complexity varies widely from example to example, and it is not always the case that the example in Baseline is the simplest one. For premise 4, for example, the Baseline crowdworker has written very complex examples that require abstract reasoning about the knowledge that *Harris* has. For this same premise, the LitL Chat crowdworker has also created a tricky set of examples, in this case ones that do not re-use any words from the original premise.

In premise 3, we see an example where the LitL Chat crowdworker uses the idiom *seen better days* for the entailment example, in place of just using a different lexical item for *tough* as the crowdworkers in the other two protocols do. Use of idioms was suggested to workers in LitL and LitL Chat as one way to write more creative examples. In premise 5, we see that the LitL crowdworker has written a challenging contradiction example, one which requires knowledge that if help is needed on a project, that means it must not be complete.

	Premise	Label	Hypothesis		
			Baseline	LitL	LitL Chat
1	(The Ramseys buried their daughter in Atlanta, then vacationed in Sea Island, Ga.) This absence, some speculate, gave the Ramseys time to work out a story to explain their innocence.	E	Some people were skeptical of the Ramseys' reasons for going on vacation.	The Ramseys came up with a story to tell the media they didn't do it.	Some speculate that the Ramseys worked out a story while on vacation.
		N	The Ramsey's held a private funeral service for their daughter.	The Ramseys had nothing to hide.	The Ramseys worked in Atlanta.
		C	The Ramsey's daughter joined them on their trip to Sea Island.	The Ramseys went into mourning after burying their daughter.	The Ramseys buried their daughter in Sea Island, Ga.
2	Mr. Clinton rewards Mr. Knight for his fund raising, Mr. Gore lays the groundwork for his anticipated presidential bid four years from now, and the companies, by hiring Mr. Knight, get the administration's ear.	E	Al Gore planned to run for president.	Mr. Gore lays the groundwork for his anticipated presidential bid four years from now.	By hiring Mr. Knight, companies were listened to by the administration.
		N	Companies were hopeful they could get Clinton to further reduce corporate tax rates.	Mr. Knight get the administration's ear for companies that contribute to his fund raising.	The administration had been ignoring the companies up to this point.
		C	Bill Clinton punished Mr. Knight because of his fund raising efforts.	Mr. Clinton admonishes Mr. Knight for his fund raising.	Companies were ignored by the adminstration because of the hiring of Mr. Knight.
3	And these are tough times for reviewers in general.	E	Reviewers are going through difficult times.	Reviewers are having a challenging time.	Reviewers have seen better days.
		N	The recession is to blame for these tough times.	Times will only get tougher for reviewers.	Reviewers are still able to get by.
		C	This is a great time to be a reviewer.	Reviewers have rarely had it so easy.	This have to be the best time to get into the review game.
4	To some critics, the mystery isn't, as Harris suggests, how women throughout history have exploited their sexual power over men, but how pimps like him have come away with the profit.	E	The author argues that some critics are incapable of understanding the role pimps have played in the exploitation of women.	pimps like him have profited.	An unsolved question involves the money making of a hustler.
		N	If women are going to attempt to exploit their sexual power over men, then it is only natural for pimps to emerge to oversee sexual transactions.	Pimps have exploited women who have more power than they think.	Reviewers are mainly concerned with hustlers.
		C	Harris does not understand the means by which women have using sexual power in order to exploit men.	Pimps control every woman.	An unsolved question involves the money wasting of a hustler.
5	We need your help with another new feature that starts next week.	E	Next week, a new feature will be introduced.	There have been other new features.	We are starting a new feature next week.
		N	This new feature focuses on cloud technology.	Help has been needed with previous features.	We are starting a new feature next week that uses maps.
		C	The new feature will start six months from now.	The project is complete and currently unsupported.	We have more help than we need for the new feature next week.

Table 3: Randomly selected examples from validation data showing typical writing from each protocol.