

Honest Students from Untrusted Teachers: Learning an Interpretable Question-Answering Pipeline from a Pretrained Language Model

Jacob Eisenstein Daniel Andor Bernd Bohnet Michael Collins David Mimno
Google Research

Abstract

We propose a new style of rationale for open-book question answering, called *markup-and-mask*, which combines aspects of extractive and free-text explanations. In the markup phase, the passage is augmented with free-text markup that enables each sentence to stand on its own outside the discourse context. In the masking phase, a sub-span of the marked-up passage is selected. To train a system to produce markup-and-mask rationales without annotations, we leverage in-context learning. Specifically, we generate silver annotated data by sending a series of prompts to a frozen pretrained language model, which acts as a teacher. We then fine-tune a smaller student model by training on the subset of rationales that led to correct answers. The student is “honest” in the sense that it is a pipeline: the rationale acts as a bottleneck between the passage and the answer, while the “untrusted” teacher operates under no such constraints. Thus, we offer a new way to build trustworthy pipeline systems from a combination of end-task annotations and frozen pretrained language models.

1 Introduction

To be trustworthy and useful, a question answering system should be able to explain its reasoning and offer evidence. In open-book question answering, such explanations often take the form of rationale *masks*, which are subsets of tokens from the original passage (Lei et al., 2016). However, a challenge for mask-based rationales is that subspans of the original passage are not meant to be read alone: coherent texts contain anaphora, ellipsis, and other cohesion-building elements that limit the interpretability of individual subspans when extracted from the discourse (Halliday and Hasan, 1976). An example is shown in Figure 1, in which the key sentence mentions the answer only through the nominal *the grieving goddess*. A sufficient rationale for this answer would have to include an

Question: What is the name of the
 ↳ person who revived Eshmun?
 Passage: ... Eshmun, a young man from
 ↳ Beirut, was hunting in the woods
 ↳ when Astarte saw him [Eshmun] and
 ↳ was stricken by his [Eshmun] beauty.
 ↳ ... The grieving goddess [Astarte]
 ↳ revived Eshmun and transported him
 ↳ [Eshmun] to the heavens where she
 ↳ [Astarte] made him [Eshmun] into a
 ↳ god of heaven. ...
 Answer: Astarte.

Figure 1: An example from QuoRef (Dasigi et al., 2019) with the generated rationale shown in dark text. The markup, shown in square brackets, makes it possible to find a more concise rationale than could be extracted from the original passage.

additional sentence introducing the entity *Astarte* and binding it to the nominal in the sentence that describes the key event.

Despite their limitations, extractive rationales have an important advantage over free-text explanations: they are directly linked to the original passage, making it easy for human readers to assess the reliability of the evidence for themselves. In this paper, we present a new style of explanation, called **markup-and-mask**, which preserves the attributability of extractive rationales while overcoming the problems created by extracting propositions from the discourse in which they were written. The key idea is that discourse context is made explicit in free-text markup and then rationales are extracted from the marked-up passages.

Rather than annotating markup-and-mask rationales manually, we present a new training method that leverages the in-context learning capability of large pretrained language models (Figure 2). First, we prompt a frozen language model to produce markup that sequentially decontextualizes each sentence in each passage in the training set. Next, we prompt the same language model to produce

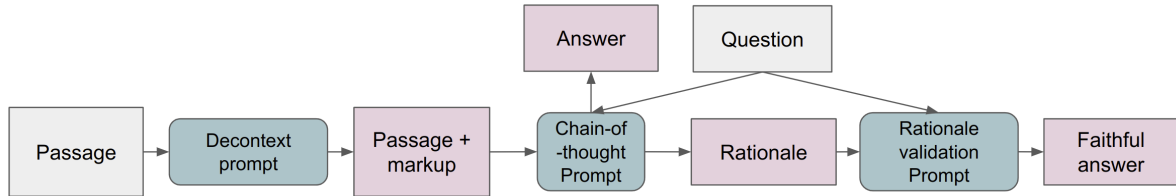


Figure 2: Schematic of the prompt chain used to produce silver data to fine-tune the honest student. At the decontextualization stage, one prompt is applied per sentence in the passage in sequence; the remaining stages use exactly one prompt each.

answers and chain-of-thought rationales from the decontextualized passage. Finally, we check that the rationale supports the answer by prompting the language model again, this time replacing the full passage with the rationale. When the answer approximately matches the ground truth, we add the rationale and markup to a silver training set. These silver annotations are used to train an “honest student” that is constrained to follow a pipeline: first generate question-neutral markup, then select a question-based rationale, and finally produce an answer using the rationale and not the passage.

Evaluation shows a number of favorable properties to this approach. Specifically, the Honest Student is able to generate rationales that: (1) support accurate question answering; (2) help human raters quickly and accurately judge whether the question-answering system is correct; (3) quantify predictive uncertainty; (4) are more likely to entail the predicted answers than non-pipeline rationales such as “chain-of-thought”; and (5) accurately match human-written decontextualizations. Evaluation also reveals that the student models outperform their teacher on all three of our key metrics — overall accuracy, entailment rate of rationales, and accuracy of decontextualizing markup — highlighting the positive impact of distillation from pretrained language models. To summarize the contributions of this paper:

- We propose markup-and-mask rationales for open-book question answering, which preserve a direct link to the original evidence text but use markup to incorporate non-local information.
- We show that it is possible to train models to produce markup-and-mask rationales without explicit supervision, by leveraging the capabilities of a pretrained language model.
- We present a general strategy for using pretrained language models to help supervise interpretable

pipeline systems in which annotations are available for only the end task.

- We empirically validate the proposed approach, showing that the resulting rationales are accurate, consistent, and useful.

2 Generating Markup-and-Mask Annotations

Our goal is to fine-tune a student model to produce markup-and-mask rationales. Lacking labeled examples, we obtain silver annotations by applying three distinct prompting patterns to the pretrained language model PaLM (Chowdhery et al., 2022) (540-billion parameter version), which we refer to as the *teacher model*. Each prompt combines passages and questions from open-book question answering datasets, along with the outputs of previous prompts, in an approach that has been called *prompt chaining* (Wu et al., 2022). There are three steps to the silver annotation process: (1) decontextualization; (2) chain-of-thought question answering; (3) rationale validation. The prompt chain is shown in Figure 2.

Decontextualization. The goal of the decontextualization step is to add free-text markup of the style shown in fig. 1. Decontextualization examples are linearized as Context: ... Passage: ... Rewrite:, with the language model prompted to complete the rewrite. An example is shown in Figure 4. We use a hand-crafted prompt with five exemplars, which were written to include a few types of decontextualization, including references to people, locations, times, and events, as well as cases in which the decontextualizing information was not present in the context. Because this stage was relatively expensive — the teacher mode must be queried for every sentence in the dataset — and because results were promising from the first exploratory prompts,

```

She [Venus] is often described as looking at herself on the mirror, although this is
→ physically impossible since viewers can see her [Venus] face reflected in their
→ direction. This phenomenon [Venus gazing at herself on the mirror] is known as the
→ Venus effect.
...
Nudes were extremely rare in seventeenth-century Spanish art, which was policed actively
→ by members of the Spanish Inquisition. Despite this [the fact that nudes were
→ extremely rare in seventeenth-century Spanish art, which was policed actively by
→ members of the Spanish Inquisition], nudes by foreign artists were keenly collected
→ by the court circle, and this painting [The Rokeby Venus] was hung in the houses of
→ Spanish courtiers until 1813, when it was brought to England to hang in Rokeby Park,
→ Yorkshire.
...
The painting [The Rokeby Venus] is believed to have been executed during one of
→ Velázquez's [the artist] visits to Rome, and Prater has observed that in Rome the
→ artist [Velázquez] "did indeed lead a life of considerable personal liberty..."

```

Figure 3: Example of output from the decontextualization prompt, applied to the Wikipedia page https://en.wikipedia.org/wiki/Rokeby_Venus

we did not consider many alternative prompts. Decontextualization was performed autoregressively, rewriting each sentence using the previous k decontextualized sentences as context.

The capabilities and limitations of this approach are highlighted in Figure 3, which shows some typical outputs. The markup resolves pronominal references *she* and *her* and the nominal references *this painting* and *this phenomenon*. Perhaps most impressively, the elliptical expression *despite this* is decontextualized with the markup *[the fact that nudes were extremely rare...]*. However, by the end of the document, we have lost track of the first name of the artist, so that *the artist* is decontextualized as only *[Velázquez]*, rather than with the full name. Future work may address this issue by exploring more sophisticated strategies than simple autoregressive decontextualization.

Chain-of-thought question answering. In chain-of-thought prompting, the language model is asked to first generate a rationale before producing an answer (Wei et al., 2022). For open-book question answering, we take the rationale to be a sentence that is extracted from the passage and which contains the answer, as shown in Figure 5. We construct question-specific few-shot prompts by concatenating several exemplars in which a question, passage, rationale, and answer are shown, before providing the question and passage for the instance to be predicted. The exemplars are drawn from the training set, selecting questions with the highest BM25 similarity to the target question (Robertson et al., 2009). Exemplars are

added until we reach a limit of 1024 sentencepiece tokens in the prompt (Kudo and Richardson, 2018); for the QuoRef dataset, this amounts to two or three exemplars in most cases.

To generate the rationales in the exemplars, we enumerate all sentences in the passage that contains an exact match to the answer and select the one with the highest BM25 similarity to the exemplar’s question. Each sentence is considered in both its original surface form and with decontextualizing markup. If no sentence contains an exact match to the answer, then the question is not included as an exemplar. However, prompts are constructed for all training set examples, even when no rationale can be extracted using this heuristic.

Rationale validation. Finally, to validate the rationales that were generated in the chain-of-thought stage, we perform a final validation stage in which the teacher model must answer questions based only on the generated rationales. As in the previous stage, we include each training set example and construct in-prompt exemplars by BM25 similarity to other questions in the training set. Because this stage does not include full passages, we can fit many more exemplars while remaining under the budget of 1024 tokens, on the order of 20 per prompt. The resulting “faithful answers” are then used to filter the fine-tuning data that is exposed to the student model.

3 Training the Student Model

The prompt chain described in Section 2 produces markup-and-mask rationales and uses them to an-

Instructions: rewrite each passage
→ using the context.

[in-context exemplars]

Context: The Rokeby Venus (also known
→ as The Toilet of Venus, Venus at
→ her Mirror, Venus and Cupid, or La
→ Venus del espejo) is a painting by
→ Diego Velázquez, the leading artist
→ of the Spanish Golden Age.

Passage: Completed between 1647 and
→ 1651, and probably painted during
→ the artist's visit to Italy, the
→ work depicts the goddess Venus in a
→ sensual pose, lying on a bed and
→ looking into a mirror held by the
→ Roman god of physical love, her son
→ Cupid.

Rewrite: *Completed between 1647 and
→ 1651, and probably painted during
→ the artist's [Diego Velázquez]
→ visit to Italy, the work [The
→ Rokeby Venus] depicts the goddess
→ Venus in a sensual pose, lying on a
→ bed and looking into a mirror held
→ by the Roman god of physical love,
→ her son Cupid.*

Figure 4: Linearization of a single decontextualization example. The text after "Rewrite: " is the model output. For subsequent sentences, the context includes the decontextualized sentences, enabling information to propagate through the entire document.

swer questions. However, there are two main reasons to distill this teacher model into a smaller "honest student." The first reason is efficiency: the prompt chain requires several calls to the large language model; because it is more specialized, the student model can potentially be smaller. The second reason is accuracy: in the teacher model, the training set is used only for in-context learning, with only a few examples per prompt; fine-tuning can make use of more gold answers, in combination with silver rationales.

To fine-tune the student model, we use as training data the gold answers and the rationales produced by the teacher model. Because our goal is to train an *honest* student, we implement the student model as a pipeline: it must first produce the decontextualizing markup without seeing the question, then generate a rationale from the passage (conditioned on the question and the marked-up passage), and finally produce an answer (conditioned on the question and the generated rationale). Critically, the student has no access to the full passage when

Use each passage to answer the question,
→ and cite the most relevant sentence
→ as an explanation.

[in-context exemplars]

Question: What is the name of the
→ person who revived Eshmun?

Passage: The myth of Eshmun was related
→ by the sixth century Syrian
→ Neoplatonist philosopher Damascius
→ ...

Explanation: *The grieving goddess
→ [Astarte] revived Eshmun and
→ transported him [Eshmun] to the
→ heavens where she [Astarte] made
→ him [Eshmun] into a god of heaven.
Answer: Astarte.*

Figure 5: An example prompt and output for chain-of-thought question answering. The linearization consists of the question, the passage, and the final line "Explanation: ". The language model then generates the explanation and answer.

generating the answer. Each step of the pipeline is implemented as a text-to-text model using the `t5x` library (Roberts et al., 2022), and the steps are trained in a single multi-task model. The specific tasks for the student model are:

Decontextualizing markup. As in the teacher model, decontextualization is performed autoregressively, with one training example per sentence. The target output is the markup produced by the teacher model.

Span selection. The input to the span selection task is a concatenation of the question and the decontextualized passage, and the target output is the rationale generated by the teacher in the chain-of-thought QA step. At training time the decontextualized passages are from the teacher; at prediction time they are from the decontextualizing markup step in the student pipeline.

Rationale-based reading comprehension. At training time, the input is a concatenation of the question and the teacher model's rationale; the target output is the gold answer. At prediction time, the input includes the rationale produced by the span selection step in the student pipeline.

End-to-end reading comprehension. We also train an end-to-end reading comprehension task, in which the input is a concatenation of the question

and the full passage. The target output is the gold answer and no rationale is produced.

The decontextualization task aligns closely to the decontextualization *prompt*, but the student model is trained by fine-tuning while the teacher model relies only on in-context learning. Unlike the chain-of-thought prompt described in Section 2, the span selection task does not produce an answer; the rationale-based reading comprehension task is conceptually similar to the rationale validation prompt, but again, the student model uses fine-tuning rather than in-context learning. To build a cleaner silver training set, we train only on the rationales that led to approximately correct answers at both the chain-of-thought stage (using the entire passage) and the validation stage (using the rationale alone). Specifically, we score the generated answers at both stages, and exclude examples for which either answer has an $F_1 < 0.5$.

4 Evaluations

We evaluate on two datasets: QuoRef (Dasigi et al., 2019) and the version of SQuAD (Rajpurkar et al., 2016) from the MRQA shared task (Fisch et al., 2019). For each dataset, we run PaLM on the training data to produce silver annotations of the markup-and-mask rationales, as described above. The decontextualization step is autoregressive, in the sense that the decontextualization for sentence t is part of the prompt for decontextualizing sentence $t + 1$. This makes it difficult to use the more efficient bulk inference procedure that we apply in the other parts of the prompt chain. For this reason, we use only a fraction of the SQuAD training data (12000 questions). We then use PaLM’s output as annotations to fine-tune multitask sequence-to-sequence models built on pretrained mT5 backbones (Xue et al., 2021). The results that follow are based on the mT5-XXL backbone. Comparisons across model scales are shown in Figure 6.

4.1 Accuracy

Table 1 shows the overall performance of the student model, an end-to-end equivalent, and a masking-only ablation. On the SQuAD dataset, performance is similar across all model variants, showing that it is possible to derive causal rationales for SQuAD answers with only a minimal impact on accuracy. In contrast, prior work found that other unsupervised rationalization techniques (Paranjape et al., 2020; Guerreiro and Mar-

	SQuAD	QuoRef
End-to-end (mT5-XXL)	83.2 / 92.8	80.4 / 85.8
Honest students (mT5-XXL)		
Markup+mask	82.2 / 91.7	68.2 / 74.5
Mask-only	82.2 / 91.7	51.9 / 58.9
Teachers (540B)		
PaLM in-context	73.7 / 86.2	57.9 / 66.7
PaLM in-context (+markup)	71.9 / 84.9	50.6 / 60.0

Table 1: Overall exact match / F_1 on open-book question answering. The *end-to-end* system predicts the answer directly from the passage; the *markup+mask* system predicts the answer from a rationale that includes both masking and markup; the *mask-only* system uses a rationale based only on masking the original unmarked text; *PaLM in-context* refers to the teacher model, which uses in-context learning only.

tins, 2021) decreased performance by 10-20 F_1 on SQuAD (Chen et al., 2022). The pipeline method suffers a significant reduction in accuracy on QuoRef, which, as discussed below, is particularly resistant to extractive rationales. However, decontextualization markup reduces the gap between the end-to-end predictor and the mask-based rationales by almost half.

Selective prediction. The availability of a step-by-step explanation can serve as a coarse form of calibration: if an example has an explanation, the associated prediction may be more accurate. To test this idea, we perform an evaluation of *selective prediction*, in which answers are produced only when they are likely to be correct (Rodriguez et al., 2019; Kamath et al., 2020). Specifically, we compare accuracy on examples where the end-to-end model and the rationale-based pipeline agree and disagree.¹ As shown in Table 2, rationalizable answers are significantly more accurate. The F_1 for rationalizable answers is more than 20 points higher than for non-rationalizable answers on both datasets, and the gap in exact match is even larger. Furthermore, most answers are rationalizable in this way. The markup-and-mask rationales play an important role in selective prediction on the QuoRef dataset, where they increase the fraction of rationalizable answers from 58% to 74%, while enlarging the F_1 gap from 13.0 to 22.1. However, on the QuoRef dataset, a better coverage-accuracy

¹The link between explanations and calibration was explored by Ye and Durrett (2022), who work in a chain-of-thought prompting framework. They show that when the chain-of-thought explanation is consistent with the passage, the answer is more likely to be correct.

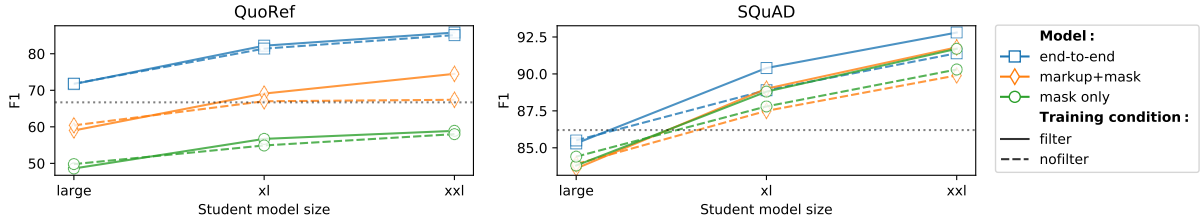


Figure 6: Overall F_1 results by student model size, for each configuration. The teacher model F_1 is shown with the dotted horizontal line.

Dataset	Rationale	e2e == pipeline?	Coverage	EM	F1
SQuAD	markup+mask	✓	86.8%	88.0	95.3
		✗	13.2%	51.8	75.8
	mask-only	✓	87.4%	87.7	95.1
		✗	12.6%	52.2	76.4
QuoRef	markup+mask	✓	74.2%	88.0	91.5
		✗	25.8%	58.3	69.4
	mask-only	✓	57.5%	87.3	91.3
		✗	42.5%	70.9	78.3

Table 2: Evaluation of selective prediction for the XXL-based models. Answers from the end-to-end predictor are distinguished by whether they agree with the answer provided by the honest student pipeline. For example, the top row shows that on SQuAD, the predictors agree on 86.8% of examples, receiving an F_1 of 95.3 on this subset.

tradeoff can be obtained by thresholding on the predictive probability of the end-to-end model; on SQuAD, predictive probability and rationalizability offer nearly identical coverage-accuracy tradeoffs.

4.2 Rationales and trustworthiness

The purpose of the rationales is to help human readers decide whether to believe the system’s answers. To test this we conduct a human evaluation, in which raters are asked to judge whether the system’s answers are correct. We focus on a subset of 470 questions from QuoRef in which the end-to-end, mask-only, and markup-and-mask systems all produce the same answer and 50% of the answers are exact matches to the reference (correct) and 50% have low F_1 (incorrect). Raters are presented with the question, the passage, and the system answer, and are asked to mark “whether the candidate answer to the question is ‘correct’ or ‘incorrect’ based on the passage.” In the end-to-end condition raters see the original passage only, the mask-only condition adds a highlighted rationale, and the markup-and-mask condition adds both the rationale and markup. We collected 18 ratings per question in each condition. Raters performed tasks comprised of five questions from the same condition; each individual rater only performed tasks from a single condition; no rater performed more than six tasks.

To determine whether the rationales were useful we measured their effect on rater speed and

accuracy. We modeled task completion time using a mixed-effects linear regression (Baayen et al., 2008), with random effects for rater ID, question sequence, rating phase (raters were recruited in two phases), and whether the rater had performed any previous tasks. The median completion time in the no-rationale condition was 322 seconds; the mask-only condition was associated with a 21.2 second improvement ($p < .05$) and the markup-and-mask condition was associated with a 36.2 second improvement ($p < .01$). The difference between the markup-and-mask and mask-only conditions was not significant ($p \approx 0.13$).

We modeled accuracy using a mixed-effects generalized linear model with a binomial link function, with random effects for rater ID, question ID, rating phase, and the number of previous tasks performed by the rater. Fixed effects include the treatment condition and whether the system answer was correct. The rationale conditions were associated with significantly higher accuracy ($\beta_{\text{markup-and-mask}} = 0.16, p < .02$ and $\beta_{\text{mask-only}} = 0.24, p < .001$, with β indicating the estimated coefficient on the log-odds of giving an accurate rating). Raters were more accurate in the mask-only than in the markup-and-mask condition but the difference was not significant ($p \approx 0.13$).

4.3 Rationale consistency

To test how often rationales are consistent with the answers, we apply natural language inference

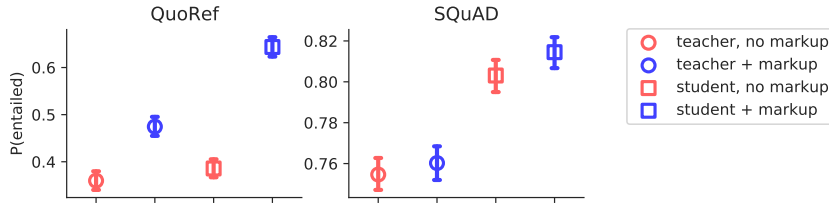


Figure 7: Consistency of rationales, as measured by the frequency with which the rationale entails a linearization of the question and the predicted answer.

(NLI). Specifically, we ask a strong NLI system whether the rationale entails the linearization, “The answer to “[question]” is “[predicted-answer]””. This style of evaluation has been applied to other tasks involving factual consistency, such as summarization and fact verification (Honovich et al., 2022). We use a very similar NLI system, trained by fine-tuning t5-XXL on multiple NLI datasets (MNLI, SNLI, FEVER, PAWS, SciTail, and VitaminC). As shown in Figure 7, the rationales produced by the pipeline student models are significantly more consistent than the chain-of-thought rationales produced by the teacher model, justifying the “honest student” moniker. On the Quoref dataset, 64% of the rationales produced by the student model (with markup) entail that model’s predicted answers, versus 47% for the teacher model with markup, and 36% without. On the SQuAD dataset, the student model achieves 81% consistency, versus 76% for the teacher model (75.5% without markup).² The markup also improves the consistency of the student model by 26% on Quoref and 1% on SQuAD. It is particularly notable that markup improves the entailment rate despite the fact that the NLI system is trained on data that does not contain any markup.

Extractiveness and compression. A rationale is deemed *extractive* when it appears as a contiguous substring in the marked-up passage, case-insensitive and not including punctuation characters or whitespace. Extractiveness is desirable because it means that the rationales are directly grounded in the passage, similar to the notion of “verified quotes” proposed by Menick et al. (2022). In Quoref, the student model rationales were extractive for 92.3% of passages; in SQuAD, 90.6%. These rationales yielded 7.9x compression in Quoref and 4.5x compression in SQuAD. Table 4 shows statistics of the markup and rationales.

²As a robustness check, we shuffled the rationales to compute how often the classifier predicted entailment for unrelated rationales and predictions. For all rationale groups, the predicted entailment rate was less than 1%.

Stress test. Can we view the markup-and-mask rationales as a faithful explanation of the reasoning process that produced the answer in the honest student (Jacovi and Goldberg, 2020)? While the honest student does not have access to the full passage, it may rely on knowledge obtained during pretraining. As a robustness check, we created an entity-perturbed version of the SQuAD dataset, similar to Longpre et al. (2021) and Yan et al. (2022), in which entity names were automatically substituted in the passages and gold answers. Substitutions were performed by running a named entity recognizer and replacing names that appear in the answer and passage with names of other entities of the same broad class, e.g., *Winston Churchill* → *Patti Smith*, *AT&T* → *the New York Knicks*.

All models were 3-4 F_1 points worse on the stress test than on the original evaluation set, with comparable exact match. Note that in some cases these perturbations affect the grammaticality of the passage, making the task more difficult for reasons that do not relate to the fidelity of the explanations. Overall these results suggest that the predictors mainly relied on the passage and not on knowledge obtained during pretraining.

4.4 Decontextualizing markup

To measure the accuracy of the decontextualizing markup, we apply the prompt-based teacher and the fine-tuned student models to a manually decontextualized dataset, in which references are replaced inline rather than annotated with markup (Choi et al., 2021). Results are shown in Table 3. Both the student and teacher models exceed the reported results for a T5-base model that was fine-tuned on 11,290 in-domain examples of the decontextualization task. This shows that it is possible to learn to perform the task reasonably well from just five labeled examples, and that distillation improves performance further. Our models produce a different style of decontextualization from the test data, so it is possible that these results could be further improved.

Well-formedness. We treat markup as a free-text generation task, with no constrained decoding. As a result, the markup may not be well-formed: the removal of markup may not yield a passage that is character-wise identical to the original passage (case-insensitive). However, both the student and teacher models usually produce well-formed markup. For more than 96% of sentences in the QuoRef eval set, the decontextualization phase of the student model leaves the original text unaffected, as intended, and in 73% of passages, all markup was well formed. In SQuAD, the markup was well formed in 96% of sentences and in 85% of full passages. The difference at the passage level is due to mainly the greater length of the QuoRef passages (see Table 4). The teacher model markup was slightly less well-formed: on both the SQuAD and QuoRef datasets, approximately 94% of the teacher model’s sentence decontextualizations were well formed. This indicates that the language model can learn the format of the markup task from the five in-context examples. Most of the errors were minor, such as omission of sentence-final punctuation and the erroneous movement of text from the original into markup, e.g. *As a schoolboy Saint-Saëns was outstanding* → *As a schoolboy [Charles-Camille Saint-Saëns] was outstanding*. More serious errors, such as incorrectly-formatted markup and deletion of significant original content, occurred very rarely.

Amount of markup. On the QuoRef dataset, the decontextualization model added 2.0 markup spans per sentence, with an average length of 5.3 SentencePiece tokens per span (31.6 per document). This almost exactly matches the behavior of the teacher model, which added 2.1 spans, with 5.8 SentencePiece tokens per span (median=4). On the SQuAD dataset, there were fewer opportunities for decontextualization: the teacher model added 0.9 markup spans per sentence, with 6.1 tokens per span. The student model also added 0.9 spans per sentence (4.8 per document), with 5.6 tokens per span (median=4).

4.5 Error analysis

On both datasets, the biggest source of erroneous answers for the pipeline model was the selection of rationales that do not contain the gold answer. In QuoRef, many questions are multihop, requiring information found in multiple spans in the passage. In some cases this information can be localized by the markup — as in the motivating example

	F_1	Precision	Recall
Students			
XXL/QuoRef	0.33	0.67	0.22
XXL/SQuAD	0.32	0.65	0.21
Teachers			
540B	0.32	0.62	0.21
64B	0.22	0.49	0.15
8B	0.11	0.40	0.06
Fine-tuned (from Choi et al., 2021)			
T5-Base	0.29	0.67	0.19
T5-XXL	0.42	0.72	0.30

Table 3: SARI-add metrics for decontextualization on the test set of Choi et al. (2021). The student models are distinguished by the training set, which contains answers but no labeled decontextualizations. Smaller student models performed similarly to the XXL-scale models on this metric, but as shown in the table, smaller teachers were significantly worse.

shown in Figure 1. There were several reasons that markup failed to add the information necessary to provide a localized rationale:

- Sometimes the necessary markup could have been supplied but was erroneously omitted: for example, to the question *who is Fran’s son?*, the pipeline model provides the rationale *The spirit reminds Scrooge [Ebenezer Scrooge] that Fran, dead for some years, is the mother of his [Ebenezer Scrooge’s] nephew*, which would have been sufficient if additional markup had been provided after the word *nephew*.
- Implicit entity references are not disambiguated by markup: for example, the sentence *In 1905 Ravel, by now thirty, competed for the last time, causing a furore* introduces a piano competition, which would have to be disambiguated to answer the question, *What is the name of the competition Ravel entered for the last time in 1905, inadvertently causing a furore?*
- Some questions reference multiple facts in the passage, such that it is difficult to imagine any markup making it possible to localize a rationale into a single sentence. For example, for the question *in what country did Rakoto Frah’s troupe win the gold medal?*, the selected rationale is *Among the 80 competitors hailing from a variety of countries, Rakoto Frah’s [the artist] troupe won the gold medal*, which is the only sentence in the passage that mentions the event from the question. To provide the answer, the markup would have had to supply location information for the event *won the gold medal*. Supplying such

	SQuAD	QuoRef
Passage length	178.9	491.7
Rationale length	39.9	62.1
Markups per passage	4.8	31.6
Mean tokens per markup	5.6	5.3
Median tokens per markup	4.0	4.0
% Extractive rationales	90.6	92.3
% Passages with faithful markup	85.4	73.4
% Sentences with faithful markup	96.4	96.7

Table 4: Passage-level statistics of the rationales produced by the XXL-based models. Passage and rationale lengths are in SentencePiece tokens. For more details see Sections 4.3 and 4.4.

information for every event would have dramatically increased the total amount of markup.

- In some cases the student simply failed to select a rationale that answered the question. For example, given the SQuAD question *which entity has a monopoly on initiating legislation?*, the pipeline model selected the rationale *It [The Parliament of the European Union] can require the Commission [of the European Union] respond to questions and by a two-thirds majority can censure the Commission [of the European Union], missing the better rationale *the Commission has a monopoly on initiating legislation.**

When the rationale did not contain sufficient information to answer the question, the student model often “hallucinated” the requested details. However, as shown in Section 4.3, this was not typical: on both datasets, the rationales usually entail the predicted answer.

On the positive side, the student not only learns to produce markup that solves challenging coreference problems (see Figures 1 and 3), it also goes well beyond the traditional coreference task. Consider this example involving groups of entities:

Jason is currently working with his best
↪ friend Daniel at a publishing house
↪ designing book covers. Their [Jason and
↪ Daniel] friend Mikey, a young doctor who
↪ has been married to Vera since the end of
↪ college, comes to them [Jason and Daniel]
↪ after Vera requests a divorce. The three
↪ [Jason, Daniel, and Mikey] decide to go
↪ out to a bar and celebrate being single.

This enables a rationalizable answer to the question *What are the names of the three characters who go to a bar in order to celebrate being single?*

5 Related Work

Philosophically, the honest student is motivated by the goal of building *warranted trust* in question answering systems (Jacovi et al., 2021), through an architecture in which the rationales meaningfully constrain the predicted answer (DeYoung et al., 2020) and can easily be checked by users.

Rationales for question answering. Rationales are typically defined as masks on the input passage (Lei et al., 2016), with the goal of finding the minimal rationale that is sufficient to identify the ground truth label (DeYoung et al., 2020). Such masks can be learned from human annotations (Zaidan et al., 2007; Menick et al., 2022) or from unsupervised objectives such as information bottleneck (Paranjape et al., 2020). We depart from fully extractive rationales by adding decontextualizing markup. This markup often indicates coreference relationships. Prior work has used human annotations to capture coreference in question answering (Dua et al., 2020). We show that similar functionality can be obtained without human annotations, through the combination of in-context learning and end-task supervision.

Reasoning chains in language models. A number of recent papers have explored the ability of large language models to “show their work.” In chain-of-thought prompting, the model is prompted to produce an explanation alongside its answer, with questions focusing on arithmetic and commonsense reasoning (Kojima et al., 2022; Wei et al., 2022; Zhou et al., 2022). Concurrent research uses chain-of-thought prompting in a student-teacher setup (Snell et al., 2022). In all of these papers, the purpose of the explanations is not necessarily to make the model more trustworthy, but rather, to make the answer more accurate. In contrast, our main goal is to increase transparency: question-answering systems must be auditable and self-explaining to avoid leading users astray. Thus we seek to build an *honest* student model, whose rationales accurately describe the passage and the predicted answer (Creswell and Shanahan, 2022). A related point is that chain-of-thought explanations have been found to be inconsistent with the source text for some types of textual reasoning (Ye and Durrett, 2022). In contrast, the student model produces explanations that are almost always extractive from the marked-up text. Furthermore, the

markup has high precision, as measured against manual decontextualization.

In both the student and teacher models, the output from the markup and masking steps are redirected into new prompts as input. The rationale is not a separate output that may or may not be consistent with the answer, it is the *only* part of the passage available to the answer-generation step. This is an example of general architecture that [Dohan et al. \(2022\)](#) refer to as a *language model cascade*, building on earlier work (e.g., [Liu et al., 2022](#); [Wu et al., 2022](#)). We show that such cascades can indeed lead to reliable and useful rationales by employing this forward-chaining strategy in two ways: in the teacher model, to generate faithful reasoning traces, and in the student model, which must ignore the passage after selecting the evidence.

Another line of work has focused on training language models to perform reasoning by fine-tuning on gold reasoning traces ([Bostrom et al., 2022](#); [Creswell and Shanahan, 2022](#); [Dalvi et al., 2021](#); [Tafjord et al., 2021](#)). Our approach does not rely on annotations of reasoning traces: our student model learns to perform accurate multi-step inferences by relying on the combination of few-shot in-context learning and filtering on the performance of the end-task. More similar is the work of [Zelikman et al. \(2022\)](#), in which the model is fine-tuned to rationalize its predictions by “bootstrapping” from a small number of labeled examples. We provide a conceptually simpler approach that trains a student model by leveraging the pretrained capabilities of a large language model, using standard fine-tuning rather than a more complex iterative procedure with a dynamic training set.

Language models as teachers. We employ a language model to generate silver annotated data for the intermediate steps of the pipeline. Prior work has explored the use of language models to generate training data in the few-shot setting ([Wang et al., 2021](#)). Of particular interest are approaches for filtering language model outputs that are unlikely to be correct, which could result in a cleaner silver training set ([Lang et al., 2022](#); [Smith et al., 2022](#)). We assume access to the gold answers, and filter intermediate steps by whether they lead to high-scoring answer predictions. However, the gold labels are a relatively weak constraint on the decontextualizing markup, so stricter filtering approaches might further improve performance on the markup task.

Finally, concurrent work shows that it is possible to distill from large pretrained language models using self-consistency in chain-of-thought prompting, without any labels at all ([Huang et al., 2022](#)). Instead, they treat answers that have multiple derivations as more likely to be correct, and show that by fine-tuning on such answers, smaller students can outperform larger teachers.

6 Discussion

Limitations. A number of limitations are highlighted by the error analysis in Section 4.5. More generally, we have assumed that answers can be rationalized by a contiguous span of the passage, after applying query-independent markup. This explains the lower performance of the pipelined methods on QuoRef, which contains questions that are hard to answer from any single sentence, even with query-independent markup. Another limitation is that markup is provided in a single forward pass, making it impossible to handle cataphoric references — for example, when an individual’s name is revealed only at the end of a passage.

Conclusion. We show how to train an *honest student* to produce markup-and-mask rationales for open-book question answering. The approach has three key properties: (1) the rationales are more *expressive* than traditional masks because they include free-text markup to enable each sentence to stand on its own; (2) the rationales are *faithful* because the student model must first produce the rationale and then discard all other information from the passage when answering the question; (3) the rationale-generation system is *unsupervised*, training on silver data created by prompting a large language model. These properties suggest a general methodology for a new generation of pipeline systems, which could offer the benefits of interpretability and controllability while limiting annotation cost and achieving the expressivity of natural language. Future work will explore the capability of the teacher model to support more expressive reasoning patterns through richer prompt chains.

Acknowledgments. Thanks to the raters who produced the results in Section 4.2 and to Thomas Meehan and Kody Wood for managing this process. We acknowledge feedback and support from Alex D’Amour, Katja Filipova, Hunter Lang, Kristina Toutanova, Victor Veitch, Jason Wei, Jasmijn Bastings, and Matthew

Lamm. Thanks to Aakanksha Chowdhery, Keran Rong, and Michael Terry for supporting access to PaLM. Thanks to Roei Aharoni for sharing the NLI model used in entailment evaluations, and to the creators of the SQuAD (Rajpurkar et al., 2016) and QuoRef (Dasigi et al., 2019) datasets.

References

- R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O’Reilly Media.
- Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. *arXiv preprint arXiv:2201.06028*.
- Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. [Can rationalization improve robustness?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. [Decontextualization: Making sentences stand-alone](#). *Transactions of the Association for Computational Linguistics*, 9:447–461.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.
- Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. [Explaining answers with entailment trees](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. [Quoref: A reading comprehension dataset with questions requiring coreferential reasoning](#). 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton. 2022. Language model cascades. *arXiv preprint arXiv:2207.10342*.
- Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. [Benefits of intermediate annotations in reading comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.
- Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. [MRQA 2019 shared task: Evaluating generalization in reading comprehension](#). In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.
- Nuno M. Guerreiro and André F. T. Martins. 2021. [SPECTRA: Sparse structured text rationalization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6534–6550, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. [TRUE: Re-evaluating factual consistency evaluation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. [Large language models can self-improve](#).
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

- Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635.
- Amita Kamath, Robin Jia, and Percy Liang. 2020. [Selective question answering under domain shift](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. 2022. [Co-training improves prompt-based learning for large language models](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11985–12003. PMLR.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. [Rationalizing neural predictions](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. [Multi-stage prompting for knowledgeable dialogue generation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337, Dublin, Ireland. Association for Computational Linguistics.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.
- Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [An information bottleneck approach for controlling conciseness in rationale extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. 2022. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*.
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.
- Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2022. Language models in the loop: Incorporating prompting into weak supervision. *arXiv preprint arXiv:2205.02318*.
- Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. [Learning by distilling context](#).
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. [ProofWriter: Generating implications, proofs, and abductive statements over natural language](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. [On the robustness of reading comprehension models to entity renaming](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. In *Neural Information Processing Systems (NeurIPS)*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. [Using “annotator rationales” to improve machine learning for text categorization](#). In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Neural Information Processing Systems (NeurIPS)*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Prompts and exemplars

During decontextualization, the language model must be queried for every sentence in the dataset. For this reason, and because results were promising from the first exploratory prompts, we did not consider many alternative prompts. The prompt was written to include a few types of decontextualization, including references to people, locations, times, and events, as well as cases in which the decontextualizing information was not present in the context. The exemplars and instructions are shown in Figure 8. These exemplars are then combined with individual sentences and contexts, as shown in Figure 4.

B Implementation details

Teacher model decontextualization. Sentence-level decontextualization requires sentence

```
Instructions: rewrite each Passage
→ using the Context.

Context: Lisa loves to play practical
→ jokes.
Passage: But sometimes she goes too
→ far.
Rewrite: But sometimes she [Lisa] goes
→ too far.
Context: Bruce Lee is buried in
→ Seattle.
Passage: But some of his biggest fans
→ don't know he is from there.
Rewrite: But some some his [Bruce Lee]
→ biggest fans don't know he [Bruce
→ Lee] is from there [Seattle].
Context: The Super Bowl XLI halftime
→ show took place on February 4,
→ 2007.
Passage: It was headlined by Prince.
Rewrite: It [The Super Bowl XLI
→ halftime show] was headlined by
→ Prince.
Context: Many years later as he faced
→ the firing squad, Colonel Aureliano
→ Buendia was to remember that
→ distant afternoon when his father
→ took him to discover ice.
Passage: At that time Macondo was a
→ village of twenty adobe houses.
Rewrite: At that time [when his father
→ took him to discover ice] Macondo
→ was a village of twenty adobe
→ houses.
Context: Ursula lost her patience.
Passage: If you have to go crazy,
→ please go crazy all by yourself!
→ she shouted.
Rewrite: If you [UNKNOWN] have to go
→ crazy, please go crazy all by
→ yourself [UNKNOWN]! she [Ursula]
→ shouted.
```

Figure 8: The instructions and exemplars for the decontextualization prompt.

segmentation, which was performed using `sent_tokenize` function of NLTK (Bird et al., 2009). Because sentence tokenization errors frequently propagated to decontextualization errors, we applied a few hand-crafted character-level replacement rules to improve segmentation accuracy, e.g. transforming expressions like *J. R. R. Tolkien* into *J.~R.~R. Tolkien*. All such transformations were reversed after sentence segmentation. The maximum number of context sentences was set at $k = 5$.