Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference

William Thorne† **Ambrose Robinson**† **Bohua Peng**† **Chenghua Lin**‡

Diana Maynard†
† Department of Computer Science, University of Sheffield
‡ Department of Computer Science, University of Manchester
{wthorne1, bpeng10, d.maynard}@sheffield.ac.uk
ambrose@parablestudio.co.uk
chenghua.lin@manchester.ac.uk

Abstract

As the cultural heritage sector increasingly adopts technologies like Retrieval-Augmented Generation (RAG) to provide more personalised search experiences and enable conversations with collections data, the demand for specialised evaluation datasets has grown. While end-to-end system testing is essential, it's equally important to assess individual components. We target the final, answering task, which is well-suited to Machine Reading Comprehension (MRC). Although existing MRC datasets address general domains, they lack the specificity needed for cultural heritage information. Unfortunately, the manual creation of such datasets is prohibitively expensive for most heritage institutions. This paper presents a cost-effective approach for generating domain-specific MRC datasets with increased difficulty using Reinforcement Learning from Human Feedback (RLHF) from synthetic preference data. Our method leverages the performance of existing question-answering models on a subset of SQuAD to create a difficulty metric, assuming that more challenging questions are answered correctly less frequently. This research contributes: (1) A methodology for increasing question difficulty using PPO and synthetic data; (2) Empirical evidence of the method's effectiveness, including human evaluation; (3) An in-depth error analysis and study of emergent phenomena; and (4) An open-source codebase and set of three llama-2-chat adapters for reproducibility and adaptation.

1 Introduction

The cultural heritage sector is increasingly leveraging advanced technologies like large language models (LLMs) (OpenAI, 2024; Touvron et al., 2023a) and AI assistants (Team Gemini, 2023; Anthropic, 2024) to increase and improve access to collections and their associated data. These technologies provide new opportunities for more dynamic and intuitive interactions with heritage content. One particularly promising technology is Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), which retrieves relevant information from a database of vectorized content to generate accurate, fact-based responses to user queries. We believe that RAG, and iterations on the approach, will play a significant role in improving the search capabilities of heritage institutions in the coming years.

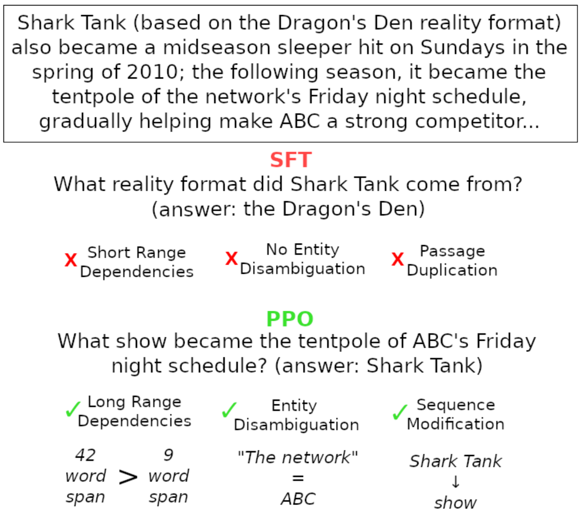
Shark Tank (based on the Dragon's Den reality format) also became a midseason sleeper hit on Sundays in the spring of 2010; the following season, it became the tentpole of the network's Friday night schedule, gradually helping make ABC a strong competitor...
SFT
What reality format did Shark Tank come from? (answer: the Dragon's Den)
X Short Range Dependencies
X No Entity Disambiguation
X Passage Duplication
PPO
What show became the tentpole of ABC's Friday night schedule? (answer: Shark Tank)
✓ Long Range Dependencies
✓ Entity Disambiguation
✓ Sequence Modification
42 word span > 9 word span
"The network" = ABC
Shark Tank ↓ show

Figure 1: Example generated questions from supervised-fine-tuned question generation model and one fine-tuned with PPO from synthetic difficulty samples.

Heritage search systems are used by the public and academics alike; however, the latter tend to submit more complex and specific queries (Koolen and Kamps, 2009). RAG has the capability to fulfil these needs but still requires robust evaluation. This includes not only end-to-end system testing but also the evaluation of individual components. As the response is generally required to be written based *only* on the retrieved documents to mitigate language model hallucinations, we argue that the task is one of Machine Reading Comprehension (MRC). While MRC datasets are well-established in the general domain, they are notably lacking in

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 450–462
November 16, 2024. ©2024 Association for Computational Linguistics

cultural heritage and the cost of their construction is prohibitive for most institutions. We estimate that the popular SQuAD dataset cost about $12,000 to just write the questions, based on their recommended time per question and stated hourly rate of $9 (Rajpurkar et al., 2016); the actual cost is likely much higher.

To address these challenges, we propose using Automatic Question Generation (AQG) systems to generate MRC datasets. However, we argue that many automatically generated questions, particularly those from zero- or few-shot approaches, do not provide an adequate challenge for modern language models. Manipulating difficulty is challenging through traditional training approaches given its abstract and subjective nature, and prompt based solutions are intractable when considering the infinite permutations and interactions between different aspects of difficulty (Lin et al., 2015a; Rajpurkar et al., 2016; Beinborn et al., 2015; Hsu et al., 2018; Cheng et al., 2021; AlKhuzaey et al., 2023).

To control difficulty, we adapt the Reinforcement Learning from Human Feedback protocol used in AI assistant steering (Ouyang et al., 2022; Bai et al., 2022). In this regime, samples are ranked based on specific criteria and paired into *chosen* and *rejected* samples for training a reward model. This reward model learns to distinguish good samples from bad and outputs a signal which steers a policy model. Rather than relying on costly human annotations, we generate synthetic preference data by evaluating question-answering model performance on a subset of SQuAD, assuming that questions answered correctly less frequently are inherently more difficult. This approach leverages the language model's innate feature extraction capabilities, eliminating the need to explicitly define difficulty components. Figure 1 demonstrates this feature extraction ability by comparing questions generated with and without reinforcement fine-tuning.

We selected SQuAD over an in-domain QA dataset for two main reasons. First, it is a well-studied, large, and diverse dataset. Second, comparable QA datasets at SQuAD's scale are either visual question-answering focused (Sheng et al., 2016; Asprino et al., 2022) or have data reliability concerns such as OCR text (Piryani et al., 2024).

This approach enables cultural heritage practitioners to generate challenging evaluation datasets more efficiently and cost-effectively than manual curation. The primary expense is compute resources, which can be accessed in the cloud for only a few dollars per hour.[1]

We summarise this paper's contributions as follows:

1. A methodology for increasing the difficulty of automatically generated questions using PPO and synthetic data.
2. Empirical evidence of the methodology's efficacy including human evaluation.
3. An in-depth error analysis and study of interesting phenomena that emerge as part of this approach.
4. An open-source code base and set of models to recreate and adapt our work[2].

2 Related Work

A similar question generation approach to ours is employed by Zhang et al. (2022) who adopt a pipeline structure. However, their primary objective is to generate suitable questions rather than specifically focusing on difficulty. An important distinction lies in their extensive pre-processing applied to identify candidate answers before feeding them to the question generation model. We argue that pre-identifying answers may limit diversity and prevent the inclusion of potentially complex answer types.

Analyzing and Controlling Question Difficulty. Understanding and managing question difficulty holds significant importance, especially in tasks involving the creation of exams and assessments (Liu and Lin, 2014; AlKhuzaey et al., 2023). One approach, as presented by Loginova et al. (2021), involves modelling the difficulty of multiple-choice questions through the use of softmax scores obtained from a pre-trained QA model. The variance in these scores is then calculated, with higher variance indicating greater difficulty.

Lin et al. (2015b) controls the difficulty of quiz questions through the selection of distractor answers based on semantic similarity between linked data items. This involves collecting both structured RDF data and unstructured text, computing similarity scores through K-means clustering, and generating questions and answers via template-based methods. Importantly, the semantic similarity plays a role in determining the difficulty level, with more

[1] https://huggingface.co/pricing

[2] We release all code and a set of three LLaMa-2 adapters on GitHub.

challenging questions featuring distractors exhibiting higher semantic similarity.

Reinforcement Learning with Human Feedback. RLHF is a machine learning paradigm that combines reinforcement learning with human-provided guidance to steer language models to meet the needs of users, finding frequent use in chatbot and AI assistant settings (Ouyang et al., 2022). The basis for most modern methods is the Proximal Policy Optimisation (PPO) algorithm (Schulman et al., 2017), which iteratively enhances the language model's policy to maximize cumulative rewards through interactions with a dataset or language simulation. It collects experiences, evaluates advantages, and updates the policy with a clipped surrogate objective to ensure stability, gradually improving the model's performance.

Automatic Question Generation. Chen et al. (2019) introduce a cross-entropy loss with a reinforcement learning-based loss function when training a gated bi-directional neural network for question generation. In this context, the reward model is optimising the semantic and syntactic quality of the question. BLEU-4, as a reward function, optimises the model for the evaluation metrics and the negative Word Movers Distance component is used to ensure semantic quality by maximising the similarity between a generated sequence and a ground truth sequence. Although question quality is maintained, other factors such as question difficulty are not considered.

Self-critic sequence training (SCST) (Rennie et al., 2017) uses a classical policy gradient method, REINFORCE, which is a Monte Carlo method. SCST computes rewards with n-gram token overlap as sub-sentence level rewards. Since training sets often have limited questions, these training rewards are arguably sparse, hindering the question generation model from extrapolating beyond the training distribution. Liu et al. (2019) adopt a two-component reward for refining ill-formed questions. Question wording is used as a measure of short-term reward, and alignment between the question and answer represents a long-term component.

3 Method

To challenge the high cost of manual annotation while maintaining quality and increasing difficulty, we design and implement a robust system capable of generating contextually relevant, coherent, and challenging question-answer pairs from textual input. The process follows the core methodology of RLHF, deviating only in the use of synthetic preference data to train a reward model. Rather than explicitly defining the characteristics of difficulty and risking failure to capture certain aspects, we exploit the ability of leading question-answer models to derive which questions are challenging, and allow a reward model to extract the component features of the task.

We task three QA models with answering all questions in our validation split of SQuAD. These questions are assigned a score based on the number of times they were answered incorrectly, which are in turn used to generate pairwise preference data. These pairwise samples enable the training of a reward model (RM) for use in fine-tuning a supervised model (SFT) on the task of question generation using Proximal Policy Optimisation (PPO)(Schulman et al., 2017).

We embed this synthetic RLHF process into a greater pipeline for generating samples, shown in Figure 2. This ensures the quality of the final dataset. The pipeline also contains a set of rule-based critics which are used to exclude samples that are malformed and those with non-unique answers in the source text. Samples are then deduplicated using exact string matching.

The remainder of this section discusses each of the relevant components of the pipeline and the RLHF process.

3.1 Supervised Fine-Tuning

In our training process for question generation and response formatting, we begin by employing a reformatted version of the SQuAD v1 training split (see Table 1). The reformatting converts SQuAD to the task of question-answer pair generation, as shown in Figure 3. We select the "correct" answer as the one that appears most frequently in the list of answers for each question in the dataset, selecting randomly among the most common if there is no victor. To ensure model robustness without overfitting, the model undergoes a single epoch of training, enabling it to effectively capture the nuances of the task.

3.2 Reward Modelling

To control the difficulty of our generated questions, we leverage the intrinsic properties present in challenging questions from SQuAD. To extract these attributes, we employ three question answering

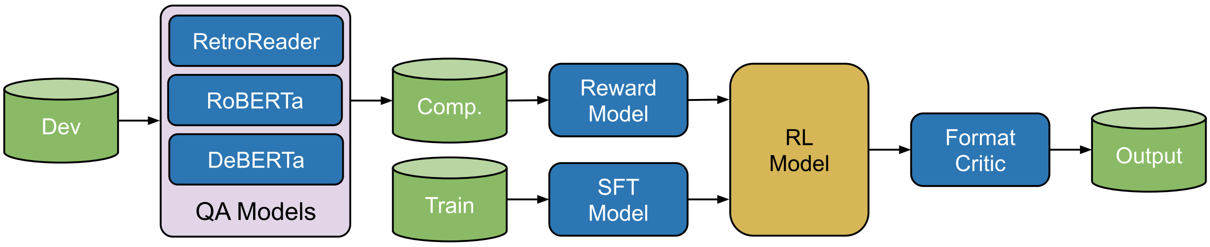

Figure 2: Depiction of our dataset generation pipeline. Question-Answering models are first used to create pairwise comparison data to train a reward model. An SFT model is trained on the train split of SQuAD and then fine-tuned using the reward model, producing the RL model. When generating question-answer pairs for the final dataset, generations are passed through the format critics to ensure data quality.

Instruction Write 1 answerable span extraction question and provide the correct answer based on the text.

Input ... Upon its arrival in Canberra, the Olympic flame was presented by Chinese officials to local Aboriginal elder Agnes Shea, of the Ngunnawal people. She, in turn, offered them a message stick ...

Response Who received the flame from Chinese officials in Canberra? (answer: Agnes Shea)

Figure 3: Example training sample from the reformatted SQuAD dataset for use in supervised fine-tuning.

models that almost match or exceed human performance on SQuAD v2 to evaluate our development split: a RoBERTa-large model[3], a DeBERTa-large model[4] and RetroReader (Zhang et al., 2020). Each question is assigned a score based on the number of models that failed to correctly answer the question. These scores are used to place questions into a pairwise ranking setup against other questions for the same input context. Where a question's scores are equal, they are considered ties, and no pairwise sample is created. We also record the margin, defined as the difference in score between the chosen and rejected samples, to experiment with the marginal ranking loss, as defined in Touvron et al. (2023b).

3.2.1 Format Critics

To ensure the quality of the final dataset, we utilise a collection of rule-based critics which we call *Format Critics*. These critics have two main functions: they remove questions that don't adhere to the desired format of *Q? (answer: A)*; they ensure the provided answer is unique in the text, minimising the number of ambiguous or impossible questions. Samples that pass these critics are then deduplicated using exact matching.

3.3 Reinforcement Training

We use Proximal Policy Optimisation (Schulman et al., 2017) with multiple sets of adapters to reduce the memory overhead during training, implemented using the Transformers Reinforcement Learning library (von Werra et al., 2020). A single base model is used with separate LoRA adapters for the policy, reference, and reward model components; each is switched to perform the relevant aspect of the reinforcement training process.

During early experiments, we found that training was often very unstable or resulted in low pass rates at the format critic. To combat this, we added a rule-based reward component to penalise generations that did not pass the format critic. This simple function converts the reward to be the negative absolute reward in the case that samples are malformed. Using a rule-based reward that manipulates the original reward prevents the instability caused by hard coding a fixed penalty and saves the computational complexity and imperfection of a second adapter-based reward model:

$$R_i = \begin{cases} -|R_i| & \text{if malformed} \\ R_i & \text{otherwise} \end{cases} \tag{1}$$

4 Experimental Setup

4.1 Models

We conduct our experiments with LLaMa2-7B-chat and apply LoRA adapters to all linear layers for all models. This drastically lowers the number of tunable parameters over full-finetuning, enabling training on a single A100 80GB GPU. We also make use of Flash Attention 2 (Dao, 2023) to improve

[3] deepset/roberta-large-squad2

[4] deepset/deberta-v3-large-squad2

computational and memory efficiency. All LoRA adapters share the same hyperparameters: a LoRA rank of 16, as Dettmers et al. (2023) found rank to have minimal impact on task performance while enabling larger batches through reduced memory usage. This memory efficiency further allowed us to implement sample packing, particularly beneficial with Flash Attention 2's preference for minimal padding. We set alpha to twice the rank [5], use a dropout of 0.05 - shown optimal for 7B models by Dettmers et al. (2023), and maintain LLaMa-2's BF16. As a baseline, we compare to LLaMa-2-7B-chat in a zero-shot setting (see Appendix B).

We experiment with marginal ranking loss to help distinguish between slight and significant differences in question difficulty while training the reward model. Under the hypothesis that the difficulty of a question is not independent of the associated passage of text, we also experiment with training a reward model with and without the input text attached. Results of these experiments can be found in Appendix A.

4.2 Generation Settings

During generation, the model is tasked with producing a single output for each question in the training set using nucleus sampling (Holtzman et al., 2020). We maintain the original configuration for LLaMa-2 with a repetition penalty of 1.1, top P of 0.7, and top K of 0 but increase the temperature from 0.6 to 0.9 to increase the diversity of generations.

4.3 Data Splits

We base our splits off the original SQuAD to minimise the risk of data leakage. We maintain the full train split unchanged as any model previously trained on SQuAD will have seen the full train split. We extract a test split of 500 contexts from the dev split, ensuring no contexts appear in both the dev and test splits. We extract 50 unique contexts from the test split for a human evaluation of question quality and answerability. In all cases, context-question pairs were only kept if they fit into the context length of LLaMa-2 when formatted in the correct prompt format. All samples were formatted into the three instruction components: *instruction*, *input*, *response* as shown in Figure 3.

Only the dev set of our SQuAD dataset was used to derive difficulty comparison data, to ensure the reward model never sees the samples used for evaluation. To evaluate the reward model, we extract 10% of the comparison contexts. Full dataset statistics can be found in Table 1.

Split	# Contexts	# Questions
Train	18,891	87,599
Dev	1,567	8,038
Test	500	2,532
Human Test	50	50
Train comp.	1,107	8,394
Dev comp.	123	950

Table 1: Split of contexts and questions from SQuAD. The *comp.* splits are derived from the dev split, used to evaluate the performance of the reward model during training.

4.4 Evaluation Metrics

As our goal is to evaluate the difficulty of answerable questions, we provide the input passage, question and answer to GPT-4o[6] and Gemini-1.5-pro[7] and ask whether the sample meets our specification of validity. We take samples to be answerable if they were unanimously labelled as such, and reject all other samples. GPT-based evaluations have demonstrated a robust alignment with human preferences across various complex tasks in reference-free settings (Fu et al., 2023; Liu et al., 2023). The results of this analysis can be found in Appendix C.

To assess the quality of generated questions relative to our SQuAD test split, we *intentionally avoid* n-gram based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and more modern alternatives such as Q-Metrics (Nema and Khapra, 2018), as we believe they restrict diversity of generation, constraining the model to reference questions and answers. We instead adopt the following reference-free metrics:

Syntactic Divergence provides a distance measure between two dependency paths which acts as a measure of difficulty. Word-lemma anchors, common to both the question and answer sentence, are first detected. A dependency path from the anchor to the interrogative word (who, what, etc.) in the question is compared to the dependency path between the anchor and the answer span in the answer sentence using Levenshtein distance (Levenshtein et al., 1966).

RQUGE calculates an *acceptability-score* by generating an answer for the candidate question and predicting the semantic similarity between the

[5] https://lightning.ai/pages/community/lora-insights/#toc9

[6] gpt-4o as of 1st June 2024

[7] gemini-1.5-pro as of 1st June 2024

Model	Total Valid (↑)	DeBERTa (↓)	RoBERTa (↓)	RetroReader (↓)
SQuAD	2,532 (-)	0.68	0.68	0.65
ZeroShot	357 ± 14 (0.14)	0.644 ± 0.007	0.650 ± 0.007	0.629 ± 0.009
SFT	1252 ± 2 (0.49)	0.654 ± 0.012	0.653 ± 0.005	0.616 ± 0.015
PPO-input	**1375 ± 18 (0.54)**	**0.601 ± 0.004**	**0.606 ± 0.003**	**0.582 ± 0.007**
PPO-input-margin	1373 ± 4 (0.54)	0.612 ± 0.001	0.608 ± 0.005	0.587 ± 0.002

Table 2: Question-Answering model performance on each set of samples. Models were only supplied samples which passed the format critics and were unanimously deemed answerable by GPT-4o and Gemini-1.5-pro. The *Total Valid* column indicates this number of valid samples used during question answering. Accuracy is based on exact match and results are mean and standard deviation across three sets of generated samples. Lower accuracy indicates harder questions.

predicted answer and the gold answer provided by the user. In our setup, this metric acts as an assessment of both the question and answer quality (Mohammadshahi et al., 2023).

QAScore attempts to align AQG evaluation to human judgements. Question-answer pairs are evaluated by summing log-probabilities of RoBERTa correct token predictions for all words in the answer when masked individually. QAScore claims to show strong correlation with human judgement (Spearman $r = 0.864$) (Ji et al., 2022).

Self-BLEU assesses how similar questions are to other questions generated for a given context. Each question is taken as a hypothesis and the others as a reference for the BLEU calculation. The self-BLEU is taken as the average BLEU for the question collection (Zhu et al., 2018).

5 Results and Discussion

Model Accuracy. To measure performance, we observe the difference in prediction accuracy for QA models on each dataset. Table 2 shows that in all cases of PPO training, we observe a decrease in average model prediction accuracy and an increase in the total number of valid generations. The consistent decrease in absolute prediction accuracy for all models when using the PPO trained models over both zero-shot and SFT signifies an increase in average question difficulty. The SFT process vastly improves the model's ability to generate valid questions. The PPO process further bolsters this capability which illustrates that the model is learning the intrinsic properties of high-quality questions. The performance of the reward models, shown in Appendix A, is reflected here, showing lesser degrees of improvement for those models fine-tuned without access to the input passage.

External Metrics. Figure 4 shows results for the reference-free metrics. RQUGE is clearly effective at discriminating between human-written SQuAD samples, those generated by the fine-tuned models and the zero-shot examples, but it is unable to separate out the SFT and PPO results. The particularly high score for SQuAD could in part be due to data leakage as the answer generation model for the metric was trained on SQuAD (Khashabi et al., 2022). This would indicate why our newly generated questions might score lower as it cannot have memorised the answer. Syntactic divergence results for the SQuAD test split and all trained model generations follow a consistent distribution but the zero-shot results appear much better, despite having a higher average prediction accuracy than the SFT and PPO models. Zero-shot obtaining higher syntactic divergence could stem from the general purpose language generation objective of LLaMa-2-chat. This could cause the model to generate boilerplate text which distances the structure of the question from that of the answer sentence but doesn't necessarily result in a more difficult question. QAScore proves uninformative, only being able to subtly identify SQuAD samples from model generated samples. Self-BLEU indicates that SQuAD samples are the most diverse, which is to be expected, but that zero-shot samples exhibit a distinct lack of diversity when compared with fine-tuned models. This result is, in part, misleading as Self-BLEU was only calculable for input passages with at least two valid questions. As the number of valid generations was so low for the zero-shot model, the cases where multiple valid questions were generated for a context was disproportionately in favour of identical generations.

In general we find the reference-free metrics to show limited correlation with model prediction accuracy and an ability to differentiate human written samples from model generations. We believe this is evidence for the continued need for more reli-

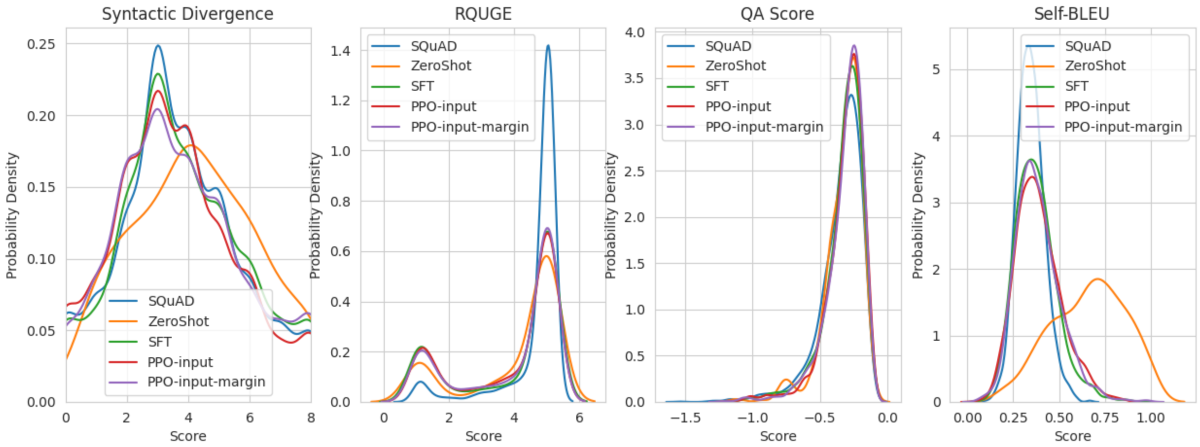

Figure 4: Distribution of reference free metrics results for each model's generations based on our SQuAD test set.

able, reference-free evaluation tools for question generation.

Human Evaluation. To evaluate question quality, we conduct a human evaluation on a subset of 50 passages from the test split. Each input passage and question is filtered through the format critic then provided to two annotators who select either the correct answer span or indicate that the question cannot be answered. In the case of annotator disagreement or the annotated answers differing from the model generated answer, the annotator responses and the model answer are provided to two new annotators who both select which responses are appropriate. We allow annotators to select multiple responses as correct but only include those that were selected unanimously by both annotators as valid. We observe an agreement of $\kappa = 0.7975$ between annotators. The results of this evaluation, shown in Table 3, displays an equivalent or improved rate of answerability when fine-tuning with PPO; the answerability proportions for each dataset are roughly equivalent to those presented in Table 2. This further corroborates the efficacy of our approach.

The results demonstrate that reinforcement learning can effectively manipulate question difficulty, while highlighting important avenues for future work. While SQuAD's synchronic nature served our experimental needs, cultural heritage datasets typically present diachronic challenges that add complexity to question generation. Although specialised diachronic models exist Drinkall et al. (2024), they lack the extensive training of general-domain LLMs. However, these larger models' exposure to historical corpora, combined with their advanced instruction-following capabilities, suggests potential for manipulating temporal complexity as an additional dimension of question difficulty.

Model	Full	Partial
ZeroShot	0.10	0.14
SFT	0.52	0.60
PPO-input	0.52	0.64
PPO-input-margin	**0.56**	**0.64**

Table 3: Results of human evaluation for question quality. *Full* indicates that the model generated answer was a valid answer according to the format critics and identified by human annotators and *Partial* indicates that the sample passed format critics and a valid answer was identified for the question but the model generated answer did not match.

5.1 Error Analysis

Failure Modes. At a high level, we can observe the reasons for sample rejection for each model. As shown in Figure 5, the zero-shot model is generally unable to generate samples that have a single answer span in the text, despite exactly specifying this in the prompt. The high number of incorrectly formatted samples was a result of only a question being generated or neither a question nor answer being generated. For all the trained model variants, the dominant failure mode was unanswerable questions. As shown in Appendix C, each of the fine-tuned models show a similar proportion of otherwise valid samples being unanswerable. The answerability rate could potentially be improved by generating candidate answers, as in (Zhang et al., 2022), and passing an input passage and answer to the question generation model.

Positional Bias. One interesting phenomenon is the positional bias in where the model chooses to

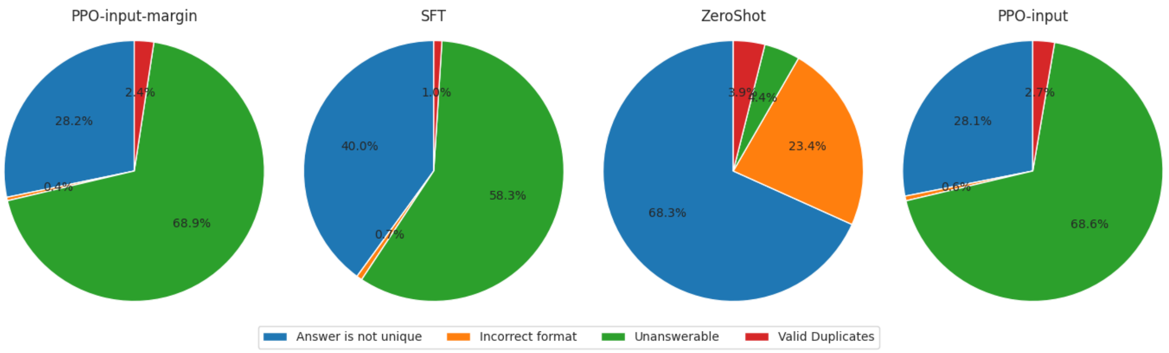

Figure 5: Error distribution of questions for SFT, ZeroShot, and the two best performing PPO variants.

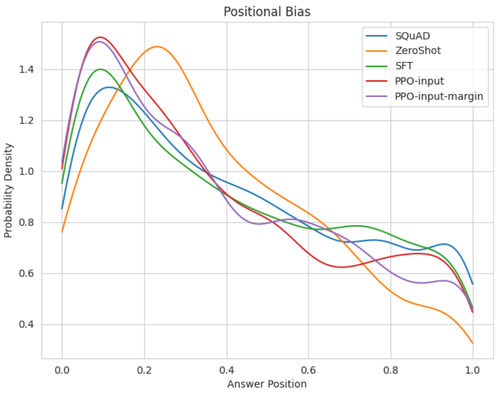

Figure 6: Position of answer span, merged to be a single word, as a proportion of the way through the input passage when split into words. SQuAD positions are selected from our test split and answers are chosen to be the most common from the list of suitable answers. Neither invalid nor exact duplicate questions are considered.

generate answers. To calculate positional bias, we treat the full answer span as a single "word" and calculate the proportion through the input paragraph in which the answer word appears. As seen in Figure 6, the zero-shot positional bias is less severe than in the other datasets. The positional bias of SQuAD is clearly seen as, after training on the dataset, all models exhibit this same preference for the beginning of input passages. The clear bias observed in the zero-shot model, despite not being fine-tuned, is documented in other tasks such as LLM ranking (Wang et al., 2023; Li et al., 2023) and in summarisation where introductory content is favoured (Ravaut et al., 2023). A potential remedy is to supply the model with a sliding window of sentences across the context paragraph to force the model to generate questions throughout the text. While this would improve the diversity of a final dataset, it may have the adverse effect of limiting the range of dependencies, restricting potentially challenging questions across the whole text.

Hallucinated External Knowledge. Where ambiguous references to specific entities exist in the input passage such as *the museum collection*, the models frequently attempt to fill in which entity is being referred to. From a context containing ambiguous references to an unnamed museum, the questions *What year did the Tate acquire the statue of St John the Baptist?*, *How many works does Rodin have in the British Museum's collection?* were generated across both the SFT and PPO models; the examples consistently passed LLM evaluations of answerability. This suggests the solution to this problem is more holistic and requires improvements at a foundational model level to resolve. We could resolve this at a critic level through more careful prompting, however, this returns to our original and intractable task of textually describing a complex task. A more holistic solution could be to adapt PPO with functional grounding (Carta et al., 2023) to be a pure text task. However, this may lower the quality of questions as it could discourage the use of implicit or complementary knowledge.

Unidirectional Relationships. A strategy to increase the difficulty of questions is to invert relationships found in the text. The models sometimes misappropriate this tool, resulting in invalid questions such as the question *What did the Ming dynasty represent?* from a passage containing *...explorer Zheng He representing the Ming Dynasty...*. Knowledge graph assisted generation could help to resolve these logical inconsistencies (Lin et al., 2015a). However, expecting our target demographics, emerging domains, to possess high-quality knowledge graphs is an unreasonable assumption.

6 Conclusion

In this paper, we introduce a low-cost methodology for generating challenging MRC datasets to meet the growing need for evaluation datasets in the cultural heritage sector. By using high-performing question-answering models to identify the most difficult questions, we were able to create synthetic pairwise data for training a reward model. Rather than manually defining question difficulty, our approach allows the model to learn and extract these features autonomously, leading to a significant improvement in the difficulty of questions generated for evaluation.

With this said, we trained on a general domain dataset in order to single out the training behaviour, in doing so losing many of the characteristic features of heritage datasets. In future work we will examine how the training paradigm fares under the unique challenges presented by such a varied industry.

Although this work was produced to meet the evaluation demands of our ongoing work in RAG at our institution, we also highlight that the approach can work in any domain and that with some modification, it could be used to augment other dataset formats. We believe this approach can be extended further, allowing for the manipulation of multiple abstract properties simultaneously through multi-reward model setups (Wu et al., 2023).

Limitations

This project only shows the suitability of the method on a single model. In future work, we seek to address this by performing a more comprehensive review of the approach across a range of model sizes and architectures. We also acknowledge that this method currently only addresses answerable questions while most contemporary QA datasets utilise both answerable and unanswerable questions. Finally, despite using LoRA and multi-adapter training, we still required approximately 15 GPU hours on an A100 80GB which restricts the potential audience for this approach. Evaluating smaller models or quantisation will enable greater access to this project's benefits.

Acknowledgements

This work was supported by the Arts and Humanities Research Council [grant number AH/X004201/1].

Ethics Statement

This project has been approved by the relevant institution's ethics committee. We use LLaMa2 in accordance with Meta's license[8]. All annotators were located through word of mouth and paid £12 per hour - above the UK National Living Wage of £11.44.

References

Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2023. Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches. *International Journal of Artificial Intelligence in Education*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.

Luigi Asprino, Luana Bulla, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. 2022. A Large Visual Question Answering Dataset for Cultural Heritage. In *Machine Learning, Optimization, and Data Science*, pages 193–197, Cham. Springer International Publishing.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning.

[8] https://ai.meta.com/llama/license/

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. Reinforcement learning based graph-to-sequence model for natural question generation. *CoRR*, abs/1908.04942.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978.

Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. ArXiv:2307.08691 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs.

Felix Drinkall, Eghbal Rahimikia, Janet B. Pierrehumbert, and Stefan Zohren. 2024. Time Machine GPT.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. ArXiv:2302.04166 [cs].

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. ArXiv:1904.09751 [cs].

Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

Tianbo Ji, Chenyang Lyu, Gareth Jones, Liting Zhou, and Yvette Graham. 2022. QAScore—An Unsupervised Unreferenced Metric for the Question Generation Evaluation. *Entropy*, 24(11):1514.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQA-v2: Stronger Generalization via Broader Cross-Format Training. ArXiv:2202.12359 [cs].

Marijn Koolen and Jaap Kamps. 2009. Information Retrieval in Cultural Heritage. *INTERDISCIPLINARY SCIENCE REVIEWS*, 343:268–284.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023. Split and Merge: Aligning Position Biases in Large Language Model based Evaluators. ArXiv:2310.01432 [cs].

Chenghua Lin, Dong Liu, Wei Pang, and Edward Apeh. 2015a. Automatically Predicting Quiz Difficulty Level Using Similarity Measures. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, pages 1–8.

Chenghua Lin, Dong Liu, Wei Pang, and Zhe Wang. 2015b. Sherlock: A semi-automatic framework for quiz generation using a hybrid semantic similarity measure. *Cognitive computation*, 7:667–679.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dong Liu and Chenghua Lin. 2014. Sherlock: a semi-automatic quiz generation system using linked data. In *ISWC (Posters & Demos)*, pages 9–12.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. ArXiv:2303.16634 [cs].

Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2019. Generative question refinement with deep reinforcement learning in retrieval-based QA system. *CoRR*, abs/1908.05604.

Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. 2021. Towards the Application of Calibrated Transformers to the Unsupervised Estimation of Question Difficulty from Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855, Held Online. INCOMA Ltd.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a Better Metric for Evaluating Question Generation Systems.

OpenAI. 2024. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. ArXiv:1606.05250 [cs].

Mathieu Ravaut, Shafiq Joty, Aixin Sun, and Nancy F. Chen. 2023. On Context Utilization in Summarization with Large Language Models. ArXiv:2310.10570 [cs].

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. ArXiv:1707.06347 [cs].

Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. A Dataset for Multimodal Question Answering in the Cultural Heritage Domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Team Gemini. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://arxiv.org/abs/2312.11805v4.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. ArXiv:2305.17926 [cs].

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. ArXiv:2306.01693 [cs].

Cheng Zhang, Hao Zhang, Yicheng Sun, and Jie Wang. 2022. Downstream transformer generation of question-answer pairs with preprocessing and post-processing pipelines. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, DocEng '22, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective reader for machine reading comprehension. *CoRR*, abs/2001.09694.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. ArXiv:1802.01886 [cs].

A Reward Model Performance

To understand the relative contributions of marginal ranking loss and the use of the input when training

Model	Accuracy (%)
RM	63.66
RM-input	**70.69**
RM-margin	62.39
RM-input-margin	70.38

Table 4: Accuracy of reward model variants based on the test split of the comparisons dataset. *input* indicates that the model was trained with the question and associated text passage as input and *margin* indicates that marginal ranking loss was used.

reward models to discriminate based on difficulty, we trained all four permutations of settings on the whole training split of the comparisons dataset and evaluated on the test split. As shown in Table 4, the inclusion of the input text had a very significant impact on performance. This was expected as the difficulty of a question is not independent of the related passage. Surprisingly, marginal ranking loss had a very slight negative impact on reward model performance. We believe this could be due to the fact that features of difficulty are very subtle and the marginal component may have caused too significant adjustments due to higher loss values.

B Obtaining Zero-Shot Model Generations

To obtain zero-shot generations, we adopt a slightly different approach. To avoid overconstaining the output of the model, we adopted a two-stage process. LLaMa-2-7b-chat was first tasked with generating a question-answer pair based on the text, unconstrained. We then passed this output back into the model with the task of extracting the question and answer components and placed them into a JSON file with the keys *question* and *answer*. We used the same, high temperature of 0.9 for generating the samples and a much lower temperature of 0.2 for extracting into a JSON to reduce the chance of models altering the generated sequences while structuring them.

C API-Based LLM Answerability Annotation

To ensure that we evaluate performance on as high-quality questions as possible, we extract only those questions deemed *answerable*, by our definition, by both GPT-4o and Gemini-1.5-pro. Table 5 shows that the zero-shot samples had the highest rate of predicted answerability; each other variant shows

Following is a text, a question and an answer. You must determine whether the provided answer is a correct span-extraction response to the question. If there are multiple plausible answers in the text, the answer should be the most relevant or accurate one. If there are multiple equally plausible answers in the text, respond "NO". If the provided answer is incomplete or contains excess information, respond "NO". If the answer does not correctly answer the question, respond "NO". Only if the answer is correct and does not breach the aforementioned requirements, respond with "YES".
Text: ... Upon its arrival in Canberra, the Olympic flame was presented by Chinese officials to local Aboriginal elder Agnes Shea, of the Ngunnawal people. She, in turn, offered them a message stick ...
Question: Who received the flame from Chinese officials in Canberra?
Answer: Agnes Shea

Respond with only "YES" or "NO" in response to this task. Do NOT provide any other text or reasoning.

Figure 7: Example prompt and response to GPT-4o (gpt-4o as of 1st June 2024) and Gemini-1.5-pro (gemini-1.5-pro as of 1st June 2024).

Model	Answerable (↑)	Unanswerable (↓)	Undetermined (↓)	Cohen's κ (↑)
ZeroShot	**0.73**	**0.14**	**0.13**	0.61
SFT	0.64	0.20	0.16	**0.62**
PPO	0.64	0.20	0.16	**0.62**
PPO-input	0.62	0.20	0.18	0.58
PPO-margin	0.62	0.19	0.19	0.56
PPO-input-margin	0.63	0.21	0.16	**0.62**

Table 5: Results of answerability task posed to GPT-4o and Gemini-1.5-pro. Results represent the proportion of samples that are answerable, unanswerable and undecided, taken from those samples which passed the format critic.

very consistent rates of answerability. This outcome should be tempered by the results in Figure 5 which indicates that the zero-shot model had an extremely high failure rate in many other regards.