Challenges in the Evaluation of the Causal Event Extraction Task

Anonymous ACL submission

Abstract

Evaluating the causal event extraction task is challenging because the boundaries of the cause and effect clauses can be ambiguous. We find that traditional metrics like Exact Match and BertScore are not representative of model performance, so we trained models, GPT-3.5 and GPT-4 for evaluation. Contrary to previous findings, GPT-4 is not a suitable replacement for human evaluation. Our trained evaluators are better at identifying ambiguous but valid cases but tend to misclassify invalid extractions. We also propose a Reinforcement Learning (RL) framework to improve the model's capacity to capture the semantic meaning rather than replicating the provided annotations. Our RL framework outperforms the other approaches in terms of causal relation classification but still falls short of the supervised fine-tuned model for causal event extraction. Still, our exploration sheds light on the complex nature of the causal event extraction task.[1]

1 Introduction

Fine-grained causal extraction is the task of identifying the cause and effect clauses of an event and the relation between them. This is the case of the Fine-grained Causal Reasoning (FCR) (Yang et al., 2022) dataset, where the cause and effect clauses are extracted from a context, and the relation between the clauses is further identified. Each cause and effect clause may comprise multiple spans of text. FCR is written in the English language.

Unlike other causal datasets that only consider a single causal relation, such as FinCausal (Mariko et al., 2020), CausalBank (Li et al., 2020) and COPA (Roemmele et al., 2011), FCR's relations are fine-grained. They can be of three types: (1) `cause`, where the cause is required for the effect to happen; (2) `enable`, where the cause can create the effect but isn't necessary for it to happen; and (3) `prevent`, which is the opposite of `cause`. Figure 1 shows an example from the dataset, and Section A (Appendix) shows statistics.

> The firm's gross margin is set to stabilize as Harley refocuses its efforts on more profitable markets, and our base case assumes that it stabilizes around 32% in 2029, helped by a more measured approach to entering new markets.
>
> Cause: Harley refocuses its efforts on more profitable markets
> Effect: The firm's gross margin is set to stabilize
> **Relation**: cause

Figure 1: Example instance from the Fine-grained Causal Reasoning (FCR) dataset.

We approach the extraction problem using the T5 and GPT-3.5 models. Our main challenge is evaluating the results. The main metric used is Exact Match, which requires the prediction to match the annotation exactly. However, it overlooks cases where the prediction differs, but the meaning is maintained. Human evaluation can recognise these cases, but it is expensive and time-consuming.

Our investigations show that it is challenging to construct an effective evaluator for the task of causal event extraction. In this task, the exact boundaries of causal (or effect) clauses are frequently ambiguous since there can be multiple possible correct annotations, including the omission or inclusion of certain words. We have used existing metrics, trained our own and applied GPT-3.5 and GPT-4[2] as evaluators to find a metric that is compatible with human evaluation. We discovered that unlike previous works (Zheng et al., 2023) suggest, GPT-4 isn't a good replacement for human evaluation. Our trained evaluators are better at detecting correct cause or effect text segments that do not precisely align with human annotations

[1]Our code is available at https://github.com/...

[2]We used the `gpt-3.5-turbo-0613` and `gpt-4-turbo-1106-preview` models.

but misclassify some false extractions as valid results.

Due to the inherent ambiguity entailed in the task of causal event extraction, we explore an alternative training framework built on Reinforcement Learning (RL). It is designed to enhance the model's capacity for capturing semantic meaning instead of replicating the provided annotations. The RL framework uses our trained evaluators as reward functions to guide the causal event extraction model. Our RL framework outperforms other approaches for causal relation classification, though it still falls short of the supervised fine-tuned model for causal event extraction. Our insights are valuable for future exploration in this avenue.

2 Methodology

There are past works on this problem. Some used sequence labelling, where each token is labelled as being the beginning or inside a clause (Saha et al., 2022) (cause or effect). Others used span extraction, where they predict two pairs of indices, $(start, end)$, indicating where the cause and effect clauses are. Neither of these encodes the type of relation, so they require a second step to classify the relation.

Generative T5 Approach To avoid the pipeline approach, we resort to a generative method, where the model generates a comprehensive text-based structured output where causes, effects, and relation are delimited by tags. This allows us to obtain both extraction and classification jointly. Figure 2 shows an example of the structured output of this method.

(a) `[Cause] Harley refocuses its efforts on more profitable markets` **`[Relation] cause`** `[Effect] The firm's gross margin is set to stabilize`

Figure 2: Structured representation for the instance in Figure 1

We fine-tuned a T5-base (Raffel et al., 2020) model on this task using supervised learning. We find that the fine-tuned model accurately learns the task specification and can correctly extract only spans of the context instead of the arbitrary text that could be possible from a generative approach. It's also able to predict only the correct relation types.

The hyperparameters used for this T5-base fine-tuning were a batch size of 32, learning rate 5e-4, 20 epochs and a maximum sequence length of 250.

GPT-3.5 We also applied GPT-3.5 (OpenAI, 2023b) with in-context learning. We hand-picked ten examples covering all relation classes and used them as in-context examples in the prompt[3]. We used a natural language format rather than the structured output of the T5 because we found that GPT couldn't follow the structured format. Another problem with GPT is the relation classification. GPT hallucinated invalid relation types, including entire sentences. We consider invalid relations as the `cause` type when calculating metrics.

3 Evaluation Metric Design

Beyond extracting the events, we also face another critical problem: evaluating the results. The metric originally used for FCR is Exact Match, where we expect the model prediction to match the annotation exactly. However, the prediction may differ in some cases while the meaning remains the same. These would be counted as wrong matches, which is an inaccurate assessment of the model. Table 1 lists some example cases where model-extracted cause and effect text subspans differ from the annotated ones. In the 'Valid substring' case – where the predicted extraction is a substring of the original annotation – the missing words do not alter the overall meaning, rendering the predicted extraction equivalent to the annotation.

An alternative solution is human evaluation. However, it is costly, time-consuming, and can't realistically be done for every model and dataset combination. It's also challenging in terms of result reproducibility, as different evaluators may have varying opinions, leading to diverging outcomes.

3.1 Building Evaluators from LMs

We want to create an automated evaluation process that is compatible with human evaluation results but is easier, cheaper and faster to perform. Some general metrics attempt to do this. For example, ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2001), BertScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) all attempt to evaluate text generation with more than exact matches or token frequencies. However, we

[3]The prompt used is shown in Figure A1 and the examples in Listing 1, in the Appendix.

Category	Annotation	Prediction	Comments
Valid substring	BB&T and SunTrust have completed their merger, forming Truist, which we believe will drive the next step up in profitability for the franchises.	BB&T and SunTrust have completed their merger, forming Truist, which we believe will drive the next step up in profitability for the franchises.	'which we believe' is an extra substring that doesn't change the meaning of the clause.
Invalid substring	Despite Telus' best in class network, we think it will have to adapt to Shaw, which will likely mean reduced pricing power and margins.	Despite Telus' best in class network, we think it will have to adapt to Shaw, which will likely mean reduced pricing power and margins.	The cause clause is a substring of the annotation, but the overall meaning is different.
Non-substring	Steadily rising Internet access pricing is a key element of our belief that Altice USA can maintain revenue per customer and cash flow as fewer customers take television and telephone services.	Steadily rising Internet access pricing is a key element of our belief that Altice USA can maintain revenue per customer and cash flow as fewer customers take television and telephone services.	The predicted cause clause is a completely different span from the annotation.

Table 1: Example cases where model predictions are different from human annotations in the FCR dataset. Words highlighted in the teal colour are extracted as Cause, while those in purple are identified as Effect.

find that none of them are good enough for our case. e turned to language models as automatic evaluators and trained some variations.

ENTAILMENT We use a classifier that takes the context and the structured extraction as inputs and decides whether the extraction entails the context, contradicts it, or is neutral. The metric here is the percentage of entailment. We used a DeBERTa-v3-base (He et al., 2022) fine-tuned on data synthesised from the FCR dataset to create samples for all three classes.[4]

NLI We use DeBERTa-MNLI-base (He et al., 2021). pre-trained on the MNLI dataset (Williams et al., 2018) without further training. We use a template to rewrite the extraction as a natural language sentence and feed that to the model, along with the context. Figure A3 (Appendix) shows an example of structured to natural language transformation. This again outputs the entailment, neutral and contradiction classes. We use the percentage of entailment as the metric.

VALID We train a binary classifier that decides whether the input pair of context and structured extraction is valid. This is trained on the output of our original T5 model, human-evaluated to decide which outputs are valid. The metric is the percentage of valid cases. The base model is DeBERTa-v3-base (He et al., 2022). We call this approach the VALID model.[5]

3.2 GPT-3.5 and GPT-4 as Evaluators

We also applied GPT-3.5 and GPT-4 (OpenAI, 2023a) as evaluators. The prompt is shown in Figure A2 (Appendix). This prompt uses in-context learning, contrastive examples (Chia et al., 2023) and some characteristics inspired by the RAGAS project[6]. We instructed the model to produce a rationale for its decisions and predict a numeric rating (1-5) instead of a valid binary label. These attributes were determined empirically to outperform simpler versions. The metric is the percentage of instances with a rating of 5.

3.3 Agreement with Human Evaluation Results

Following Zheng et al. (2023), we evaluate the agreement between our evaluation models and the human evaluation on the two causal event extraction models, T5 and GPT-3.5. Table 2 shows the agreement percentages.

However, contrary to the previous findings (Zheng et al., 2023) that LLM judges such as GPT-4 align well with human preferences in assessing multi-turn questions, the GPT-based evaluators performed poorly in our task. Compared to the trained evaluators, the GPT versions display a strong tendency to misclassify anything that is a substring of the context

[4] See Appendix C for more details about the synthetic data creation.

[5] The hyperparameters for both ENTAILMENT and VALID were a batch size of 32, a learning rate of 2e-5 and 3 epochs. NLI isn't fine-tuned.

[6] https://github.com/explodinggradients/ragas

Evaluator model	T5	GPT-3.5 (10-shot)
ENTAILMENT	65.67	46.41
NLI	36.78	39.65
VALID	68.09	63.58
GPT-3.5	64.85	35.88
GPT-4	64.89	45.64

Table 2: Evaluation agreement (%) between LM evaluators and human evaluation.

as valid, which is not always correct.

Our trained evaluators ENTAILMENT and VALID are the most aligned with the human evaluation, with both GPT models falling behind. VALID, in particular, has the highest agreement in both T5 and GPT-3.5 cases, suggesting it is the best evaluator for our case.

4 Experiments

Table 3 shows the causal event extraction results of the T5 and GPT-3.5 models on the FCR dataset according to the human, exact match, other traditional text generation evaluation metrics, and LM metrics.[7] Full evaluation details can be found in Appendix F.

Results As expected, the human evaluation values are higher than exact matches, as it is more lenient about extra or missing words. However, none of our trained evaluators match it. Exact match underrates both models, and the rest overrates them. This can be due to the difficulty of determining when the extracted text subspan is valid, as all extractions are substrings of the context. This is particularly notable in the ENTAILMENT and GPT-3.5 evaluators.

Metric	T5	GPT-3.5 (10-shot)
Human	64.38	35.13
Exact Match	52.28	30.05
ROUGE-L	77.18	64.33
BLEU	75.83	61.76
BLEURT	75.30	63.09
BertScore	95.52	89.84
ENTAILMENT	98.27	94.84
VALID	87.47	84.85
GPT-3.5	98.55	99.15
GPT-4	84.87	85.71

Table 3: Causal event extraction results (%) for T5 and GPT-3.5 on the FCR dataset.

[7]We don't consider NLI because of its low agreement with human evaluation.

Discussion When manually evaluating the results, we found that the ENTAILMENT and VALID evaluators are better at detecting the 'valid substring' cases, where the model-extracted cause and effect clauses did not precisely align with the human annotations but conveyed similar meaning. However, these evaluators also made mistakes by classifying false extractions as valid ones. This shows the challenge of developing an effective evaluator for the task of causal event extraction, where the precise boundary of cause or effect clauses is often ambiguous, resulting in numerous acceptable alternatives.

Given the inherent ambiguity associated with the task of extracting causal events, we aim to investigate an alternative training framework to enhance the model's ability to capture the correct semantic meaning rather than merely replicating the provided annotations. In traditional supervised learning, cross-entropy is commonly used as the loss function, directing the model to produce tokens that match the annotated ones. However, as previously discussed, in causal event extraction, there can be multiple possible annotations for causal and effect clauses, such as variations with the omission or inclusion of certain words. To address this challenge, we propose utilising our suggested evaluators as reward functions and implement a reinforcement learning (RL) approach with Proximal Policy Optimisation (PPO) (Schulman et al., 2017) for causal event extraction. We use the supervised T5 model as our base model. Our RL framework improves upon supervised T5 by nearly 2% for causal relation classification, though it did not show improvement for causal event extraction, possibly due to the use of imperfect evaluators as the reward functions[8]. Nevertheless, we believe this is a promising direction worth further exploration.

5 Conclusion

We have explored several evaluation approaches to address the inherent ambiguity of the causal event extraction task. Our findings demonstrate the difficulty in finding a viable replacement for human evaluators while also highlighting the potential promise of utilising reinforcement learning with the evaluator as the reward function for future research exploration.

[8]Details of RL implementation and results are shown in Appendix E.

Limitations

Our trained metrics do not perform similarly to the human evaluation as we intended. We attempted to use reinforcement learning to train better models but only observed better performance for causal relation extraction and yet no improvement for causal event extraction.

We applied GPT-3.5 and GPT-4 as evaluators, and our result goes against established precedent in that they performed worse than our purposely trained evaluators. This could be because we didn't explore the best techniques to prompt the models to their full potential. We experimented with Chain of Thought (CoT) (Wei et al., 2022) as a prompting technique, but it did not improve the results over the approach we used. As future work, we could employ other techniques to improve CoT, such as Contrastive CoT prompting (Chia et al., 2023) and Self Consistency (Wang et al., 2022b). We leave these possibilities as future work.

Another limitation is that we used a single dataset, FCR. We used it because it represented an interesting instance of the causal event extraction problem, as it had both multiple spans per clause and fine-grained relation types. To the best of our knowledge, it was the only dataset to have both. There are other datasets, such as MAVEN-ERE (Wang et al., 2022a) and TellMeWhy (Lal et al., 2021) that could benefit from a similar approach.

References

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive Chain-of-Thought Prompting.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A Dataset for Answering Why-Questions in Narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided Generation of Cause and Effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3629–3636, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

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.

Dominique Mariko, Estelle Labidurie, Yagmur Ozturk, Hanna Abi Akl, and Hugues de Mazancourt. 2020. Data Processing and Annotation Schemes for FinCausal Shared Task.

OpenAI. 2023a. GPT-4 Technical Report.

OpenAI. 2023b. OpenAI Platform. https://platform.openai.com.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium - Technical Report*.

Anik Saha, Jian Ni, Oktie Hassanzadeh, Alex Gittens, Kavitha Srinivas, and Bulent Yener. 2022. SPOCK at FinCausal 2022: Causal Information Extraction Using Span-Based and Sequence Tagging Models. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 108–111, Marseille, France. European Language Resources Association.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022a. MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-Consistency Improves Chain of Thought Reasoning in Language Models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.

Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. 2022. Towards Fine-grained Causal Reasoning and QA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

A FCR Dataset Statistics

Table A1 shows statistics on the Fine-grained Causal Reasoning (FCR) (Yang et al., 2022) dataset regarding extraction and Table A2 regards classification.

Split	# Examples	# Relations	# Causes	# Effects
Dev	2482	3224	3224	3238
Train	19892	25938	26174	26121
Test	2433	3045	3065	3062

Table A1: FCR dataset: extraction statistics.

Split	# Relations	% Cause	% Prevent	% Enable
Dev	3224	63.78	5.40	30.82
Train	25938	63.05	5.90	31.05
Test	3045	64.00	5.38	30.62

Table A2: FCR dataset: classification statistics.

B GPT Prompts

Figure A1 shows the prompt used when employing GPT-3.5 and GPT-4 as extraction models. Figure A2 shows the prompt for evaluation.

```
What are the causes, effects and
relation in the following text? The
relation must be one of "cause",
"enable", or "prevent". The causes
and effects must be spans of the text.
There is only one relation.

The response should be formatted as
this:
Cause: <text>
Effect: <text>
Relation: <text>

When there are multiple causes or
effects, separate them by " | ". Don't
add quotes around the extractions.
```

Figure A1: GPT extraction prompt.

C Synthetic Data for Training the ENTAILMENT Evaluator

To train the ENTAILMENT evaluation model (Section 3.1), we need examples from all three classes: *entailment*, *contradiction* and *neutral*. The original dataset does not contain contradiction and neutral sentences, so we have to create synthetic data for these two classes. We compile a list of all pairs of text passages and their spans.

```
Given the context, how valid is
the extraction? The extraction is
composed of a cause and effect. The
cause and effect are spans of the
context.

Evaluate the extraction based on the
following criteria:

1. Read the extraction and compare
it to the context. Check if the
extraction contains the cause and
effect mentioned in the context.
2. Make sure that the extraction
clauses only contain the necessary
information.
3. Penalize extractions that are too
long or too short.
4. Penalize extractions that include
more information than necessary for
the clause.
5. Assign a score for validity on a
scale from 1 to 5, where 1 is the
lowest and 5 is the highest based on
the Evaluation Criteria.

Respond with the following format:
Explanation: <text explaining the
score>
Score: <score from 1 to 5>
```

Figure A2: GPT evaluation prompt.

The final data consists of pairs of text passages and hypotheses. These hypotheses belong to three classes: entailment, neutral and contradiction. Each class is produced differently:

- Entailment: the hypothesis belongs to the same example as the passage
- Neutral: the hypothesis belongs to a different example from the passage
- Contradiction: the hypothesis belongs to the same example as the passage. This time, we flip the cause and effect to get a contradiction. This is done by parsing the original structured relation and swapping the cause and event components.

Since the entailment and neutral cases can be sentence fragments, we use GPT-3.5 to produce complete sentences from them. The contradiction cases are structured text, so we use GPT-3.5 to reconstruct these sentences as natural text. We use the system message ‘`You are a helpful assistant that generates sentences from causes, effects and relations`’ and the prompt ‘`Given the following causes and effects, generate a sentence:`’. Table A3 shows statistics of our created synthetic dataset.

Split	# Examples
Dev	7441
Train	59580
Test	7286

Table A3: The statistics of the synthetic dataset created for training the ENTAILMENT evaluator. For each split, we have the balanced distribution of the three classes, *entailment*, *contradiction* and *neutral*.

D Rewriting Structured Text to Natural Language for the NLI Evaluator

Figure A3 shows an example of rewriting the structured output of the T5 model to a natural language sentence to use with the NLI evaluation model.

```
(a) [Cause] its business was barely breaking
$100 million in revenue–and have steadily
grown with its top line and margin expansion
[Relation] prevent [Effect] MPS' returns
on invested capital | dipped below 20%

(b) Its business was barely breaking $100 million
in revenue–and have steadily grown with its
top line and margin expansion prevents MPS'
returns on invested capital, and dipped below 20%
```

Figure A3: Rewriting structured output to natural language: (a) original (b) rewritten.

E Reinforcement Learning

Cross-entropy is the conventional supervised learning loss for text generation, which directs the model to generate tokens identical to the annotated ones. However, we have discovered that this is not always the most effective, so we seek another way of training our generative model.

Our goal is to improve the model’s capability to capture the correct meaning rather than merely replicating the annotated text. To achieve this, we apply reinforcement learning (RL) with the Proximal Policy Optimisation (PPO) algorithm (Schulman et al., 2017) to

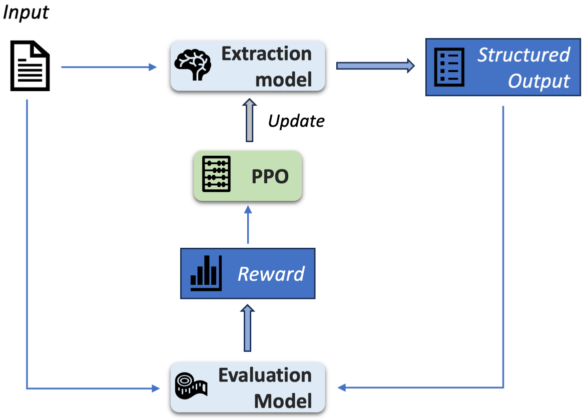

Figure A4: Architecture of the Reinforcement Learning (RL) framework.

move the model in that direction, using the supervised T5 model as the starting point. To determine the rewards for RL, we use the evaluation models we have trained, namely, ENTAILMENT and VALID. The reward signal passed to the RL trainer is the logit for the true class from each of these models. We use the TRL library[9] to train `transformers`[10]-based models.

However, there were some issues. The evaluation models are imperfect metrics, so the rewards they generate cannot be guaranteed to steer the model in the desired direction. Coupled with the complexities and instabilities associated with RL as a learning process, we fell short of achieving the desired level of performance.

We introduced certain strategies to improve the training process, including implementing L2 regularisation in the PPO loss and skipping batches that exhibited excessively high KL divergence. In our experiments, these particular batches often caused the model to deteriorate, prompting us to set a maximum KL divergence threshold of 2. If a batch trajectory’s KL divergence exceeded this threshold, we opted not to apply the PPO update from that trajectory.

We also applied human evaluation to the RL models and found them to be of similar quality to the supervised T5 model, albeit slightly inferior. This suggests that the models did not deviate significantly from the original model, but the introduced changes did not yield an improvement.

[9] https://github.com/huggingface/trl
[10] https://github.com/huggingface/transformers

F Full Evaluation Results

Table A4 shows the performance of the T5, GPT-3.5 and RL models on the extraction metrics. Table A5 shows the classification metrics.

On causal event extraction, ROUGE-L, BLEU, BLEURT and BertScore are all incompatible with the human evaluation, especially when evaluating the GPT-3.5 model. The LM evaluators are not much better, with GPT-3.5 being the worst of them, classifying almost all examples as valid.

On causal relation classification, T5 exhibits superior performance in terms of accuracy and precision compared to the other models. RL with VALID achieves the best recall and F1 scores among all the evaluated models.

Metric	T5	GPT-3.5 (10-shot)	RL with ENTAILMENT	RL with VALID
Human	64.38	35.13	59.23	60.48
Exact Match	52.28	30.05	47.06	50.02
ROUGE-L	77.18	64.33	73.08	75.47
BLEU	75.83	61.76	73.42	75.31
BLEURT	75.30	63.09	71.61	73.71
BertScore	95.52	89.84	94.84	95.25
ENTAILMENT	98.27	94.84	98.83	98.23
VALID	87.47	84.85	80.38	84.33
GPT-3.5	98.55	99.15	-	-
GPT-4	84.87	85.71	-	-

Table A4: Causal event extraction results (%) for T5, GPT-3.5 and the RL models on the FCR dataset.[11]

Metric	T5	GPT-3.5 (10-shot)	RL with ENTAILMENT	RL with VALID
Accuracy	**70.37**	61.57	67.77	67.89
Precision	**57.91**	46.56	55.62	55.90
Recall	51.90	47.51	54.71	**55.31**
F1	53.85	46.93	55.11	**55.58**

Table A5: Causal relation classification results (%) for T5, GPT-3.5, and the RL models on the FCR dataset.

G Information on Computational Experiments

We used a single NVIDIA A100 GPU (40 GB) for all of our experiments. Training the T5-Base model (220M parameters) took about 6 hours, and the DeBERTa-v3-Base models (86M parameters) took 2 hours each. The Reinforcement Learning models took 24 hours each. Time for inference on all of them was trivial. We did multiple experiments on these models, which brings the total number of hours to the low hundreds. All models fit entirely in the GPU VRAM.

We used the OpenAI API to run experiments on GPT-3.5 and GPT-4. The GPT-3.5 extraction model took 5 hours to run. The GPT-3.5 evaluator took 2 hours, and the GPT-4 evaluator took 5 hours. These were also run multiple times, with the total amount of time around 50 hours.

[11]Because of the cost, we chose not to run the GPT-3.5 and GPT-4 evaluators on the RL models. We expect them to perform poorly based on the other metrics.

H Licenses

The FCR dataset used is distributed in the Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY- NC-SA) license. The DeBERTa models are covered by the MIT license. The T5-Base is under the Apache-2.0 license. The GPT API is a commercial service under OpenAI's terms of use. We use the dataset and tools for an intended use: research only.

I GPT in-context learning examples

Listing 1 shows all ten examples used as part of the prompt for the GPT-3.5 extraction model in their raw JSON format. The original file is available with the code.

```
[
  {
    "context": "We expect Robert Half
    to increase permanent placements by
    providing employers access to its
    deep bench of highly skilled
    professionals.",
    "question": "What are the events?",
    "question_type": "enable",
    "answers": "Cause: providing
    employers access to its deep bench
    of highly skilled
    professionals\nEffect: Robert Half
    to increase permanent
    placements\nRelation: enable",
    "id": "57d64189"
  },
  {
    "context": "Burlington has faced
    inventory flow challenges (despite
    ample product availability) as it
    and its vendors restart their
    supply and distribution networks;
    freight costs are also rising
    sharply.",
    "question": "What are the events?",
    "question_type": "cause",
    "answers": "Cause: it and its
    vendors restart their supply and
    distribution networks; freight
    costs are also rising
```

```
    sharply\nEffect: Burlington has
    faced inventory flow challenges
    (despite ample product
    availability)\nRelation: cause",
    "id": "581af36a"
  },
  {
    "context": "The firm owns and
    operates fabrication yards in China
    and Mexico, and its fabrication and
    modular construction capabilities
    allow it to complete parts of large
    projects off-site and ship them in
    modules. This strategy gives Fluor
    flexibility and more control over
    costs when working in areas with
    scarce and expensive local labor.",
    "question": "What are the events?",
    "question_type": "enable",
    "answers": "Cause: The firm owns
    and operates fabrication yards in
    China and Mexico, and its
    fabrication and modular
    construction capabilities allow it
    to complete parts of large projects
    off-site and ship them in
    modules\nEffect: gives Fluor
    flexibility and more control over
    costs when working in areas with
    scarce and expensive local
    labor\nRelation: enable",
    "id": "4075835c"
  },
  {
    "context": "They would not have the
    advantage Cogent had 20 years ago,
    when top providers in a more
    nascent Internet business were
    phone and cable companies, and
    fiber assets could be procured on
    the cheap due to the collapse of
    the tech and telecom bubble.",
    "question": "What are the events?",
    "question_type": "cause",
    "answers": "Cause: the collapse of
    the tech and telecom
    bubble\nEffect: fiber assets could
    be procured on the cheap\nRelation:
    cause",
    "id": "2a64e8dd"
  },
  {
    "context": "Consistent product and
    process technological advancement
    enables more favorable pricing
    relative to many automotive
    industry suppliers that lack the
    capability or the desire to
    innovate.",
    "question": "What are the events?",
    "question_type": "enable",
    "answers": "Cause: Consistent
    product and process technological
    advancement enables more favorable
    pricing\nEffect: relative to many
    automotive industry suppliers that
    lack the capability or the desire
    to innovate\nRelation: enable",
    "id": "a62b3c49"
  },
  {
    "context": "In connected care we
    assume slower growth in monitoring
    and analytics, offset by higher
    growth in sleep and respiratory
    care.",
    "question": "What are the events?",
    "question_type": "prevent",
    "answers": "Cause: higher growth in
    sleep and respiratory care\nEffect:
    slower growth in monitoring and
    analytics\nRelation: prevent",
    "id": "ff14eb55"
  },
  {
    "context": "For 2021, we have
    marginally lifted our sales
    estimate (to $18.4 billion from
    $18.3 billion) but have
    significantly raised our operating
    margin forecast to 4.8% from 3.9%,
    leading to an adjusted EPS forecast
    that improves to $2.85 from our
    prior $2.29 estimate.",
    "question": "What are the events?",
    "question_type": "cause",
    "answers": "Cause: marginally
    lifted our sales estimate|
    significantly raised our operating
    margin forecast to 4.8% from
    3.9%,\nEffect: adjusted EPS
    forecast that improves to $2.85
    from our prior $2.29
    estimate.\nRelation: cause",
    "id": "f1330f5c"
  },
  {
    "context": "Its operating income
    (excluding charges) dropped more
    than $1.5 billion between 2014 and
    2019 (to $1.2 billion from $2.8
    billion) on store closures,
    declining sales, and increased
    expenses.",
    "question": "What are the events?",
    "question_type": "prevent",
    "answers": "Cause: on store
    closures, declining sales, and
    increased expenses\nEffect:
    operating income (excluding
    charges) dropped more than $1.5
    billion\nRelation: prevent",
    "id": "59422bb1"
  },
  {
    "context": "Alliance Data Systems
    gathers data on its client's
    customers, helping to better tailor
    these programs, which can create
    some switching costs in the
    process.",
    "question": "1 What are the
    events?",
    "question_type": "cause",
    "answers": "Cause: Alliance Data
    Systems gathers data on its
    client's customers\nEffect: create
    some switching costs in the
    process\nRelation: cause",
    "id": "77d6a30b"
  },
  {
```

```
    "context": "After several years of
    mixed results, Merck's R&D
    productivity is improving as the
    company shifts more toward areas of
    unmet medical need.",
    "question": "What are the events?",
    "question_type": "cause",
    "answers": "Cause: the company
    shifts more toward areas of unmet
    medical need\nEffect: After several
    years of mixed results, Merck's R&D
    productivity is
    improving\nRelation: cause",
    "id": "47352f8f"
  }
]
```

Listing 1: GPT in-context learning examples