Causal Tree Extraction from Medical Case Reports: A Novel Task for Experts-like Text Comprehension

Sakiko Yahata[1], Zhen Wan[1], Fei Cheng[1], Sadao Kurohashi[1], Hisahiko Sato[2], Ryozo Nagai[3]
[1]Kyoto University, [2]Precision Inc., [3]Jichi Medical University
[1]{yahata, ZhenWan, feicheng, kuro}@nlp.ist.i.kyoto-u.ac.jp
[2]satoh@premedi.co.jp, [3]rnagai@jichi.ac.jp

Abstract

Extracting causal relationships from a medical case report is essential for comprehending the case, particularly its diagnostic process. Since the diagnostic process is regarded as a bottom-up inference, causal relationships in cases naturally form a multi-layered tree structure. The existing tasks, such as medical relation extraction, are insufficient for capturing the causal relationships of an entire case, as they treat all relations equally without considering the hierarchical structure inherent in the diagnostic process. Thus, we propose a novel task, Causal Tree Extraction (CTE), which receives a case report and generates a causal tree with the primary disease as the root, providing an intuitive understanding of a case's diagnostic process. Subsequently, we construct a Japanese case report CTE dataset, J-Casemap, propose a generation-based CTE method that outperforms the baseline by 20.2 points in the human evaluation, and introduce evaluation metrics that reflect clinician preferences. Further experiments also show that J-Casemap enhances the performance of solving other medical tasks, such as question answering.

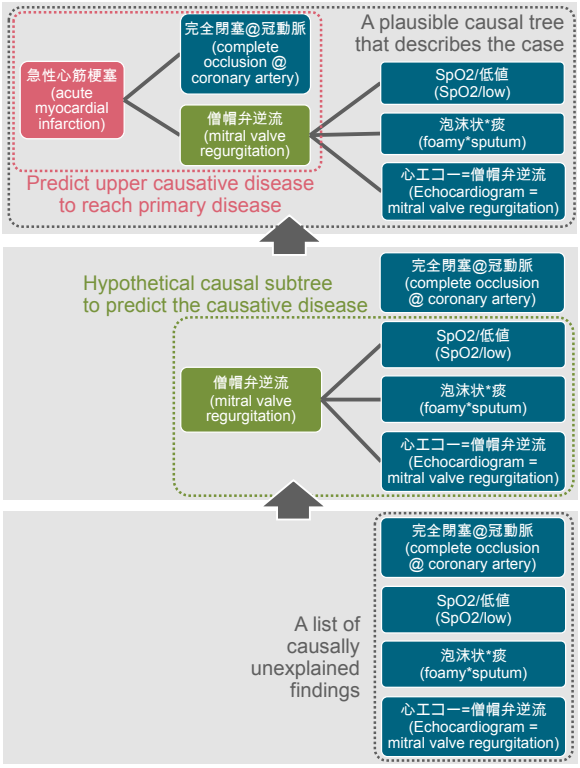


Figure 1: A diagnostic bottom-up procedure.

1 Introduction

A medical case report is a detailed document describing a case involving a rare disease or an important clinical experience, intended to share clinical knowledge. Each report comprehensively encapsulates the diagnostic process, integrating rich medical entities such as patient information (e.g., age), medical history (e.g., past diseases), clinical findings (e.g., symptoms and test results), and treatments. As described in Jha AK (2002), understanding the causal relationships among medical entities is crucial for comprehending the diagnosis procedure. In this context, existing NLP research has a history of engaging in medical relation extraction (RE) (Parikh et al., 2019; Wolf et al., 2019; Gao et al., 2023; Khetan et al., 2022) to extract causal relationships between medical entity pairs.

Clinicians can gain valuable insights to enhance their practice by understanding the diagnostic procedure of an existing case (Bowen, 2006). The diagnosis procedure is often carried out in a bottom-up manner, resulting in a comprehensive causal tree extracted from the case report. An illustration of the diagnostic procedure is shown in Figure 1. In this process, first, clinicians organize a list of findings as leaves. Second, clinicians predict which causative disease corresponds to some set of leaves, and construct a hypothetical causal subtree with causative disease as the parent. Third, the parents of subtrees serve as children in the bottom-up procedure and clinicians iteratively infer the

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25861–25879
November 4-9, 2025

parent of each subtree. Finally, clinicians reach the primary disease as the root and derive the most plausible causal tree that can describe the entire case. This indicates the limitation of the existing RE task for pairwise causal relationships that lacks consideration of multi-layered causal structures. Consequently, they are insufficient for demonstrating expert-like medical text comprehension procedures.

Therefore, we propose a novel **causal tree extraction (CTE) task** that transforms case reports into a **causal tree**. An example of a causal tree is shown in the top box in Figure 1. The most distinctive characteristic of CTE is that it form a tree structure with the primary diseases as roots. The causal tree presents an at-a-glance understanding of which parts of the case are important and what the main causal consequences are, even if the reader lacks specialized knowledge. In addition, causal relations has the potential to enhance the keyword searching capabilities of case report databases.

In this paper, we present a full pipeline of the construction of a human-annotated CTE dataset, LLM-based CTE method, and evaluation metrics. First, we construct the **J-Casemap** dataset, which consists of Japanese case reports and their corresponding causal trees. The causal trees in the J-Casemap have been annotated by highly specialized Japanese clinicians, and further experiments show their benefits on medical QA tasks, making them a potential resource for various medical applications.

Next, we propose a generation-based method for CTE. Though recent LLMs have demonstrated high performance in the medical domain (Kasai et al., 2023), large commercial models like ChatGPT (OpenAI et al., 2024), Claude (Antropic, 2024), Gemini (Team et al., 2024) are restricted from processing patient data due to data leakage concerns. Therefore, we conduct experiments using Japanese specialized open LLMs and combine continual pretraining with Japanese medical data and fine-tuning for CTE to compensate for the lack of medical knowledge. The proposed method achieves a human evaluation score of 82.7, which substantially outperforms the baseline (Ozaki et al., 2022) by 20.2 points. Ablation study shows the effectiveness of continual pretraining, especially in the low-resource setting.

Finally, we propose an automatic evaluation method that reflects clinician preferences since human evaluation requires highly experienced clinicians and is costly. In evaluating CTE, the important factors are whether the primary disease of the case is correctly extracted and whether relationships associated with those nodes at the higher layer of the tree are correctly extracted. Conversely, the absence of extracted entities that are less related to the diagnosis is not a critical issue. For such a task, existing automatic evaluation methods, such as triplet F1 used in relation extraction tasks is not suitable because they cannot determine the importance of each entity or its position in the causal tree. Since this evaluation requires extensive medical knowledge, we propose a method that weights relational triplets and focuses on the salient entities based on human preference. This weighting method reduces the gap between automatic evaluation scores and manual evaluation scores, improving their correlation.

We summarize our contributions as follows: (1) Introducing a novel CTE task that requires advanced text comprehension and constructing the J-Casemap dataset consisting of case reports annotated with high-quality causal tree annotation; (2) Proposing an LLM-based generative model for extracting causal trees from case reports; (3) Discussing an automatic evaluation method for CTE on case reports.

2 Task Definition: Causal Tree Extraction (CTE)

This section explains the specifications of the CTE task. A medical case report is represented as a disease-centric tree, where each **node** offers the modification information surrounding a head entity (usually a disease or finding), and the **edges** between nodes usually represent the causal or evidential ***parent_of relation*** between diseases and findings. For instance, the root “急性心筋梗塞 (acute myocardial infarction)” is evidenced by the child “完全閉塞 (complete occlusion)” in Figure 3. The **root** node of the tree structure corresponds to the primary disease, which represents the main factor that causes other diseases or findings. Then, we link those evidential nodes through edges (representing *parant_of* relationships) to the root. These diseases may also cause their own child nodes, naturally extending the **depth** of a tree summary.

To be noticed, each node can have internal structures, expressing the supporting informa-

tion modifying the head entity. There are four pre-defined modification relationships and corresponding text symbols are denoted as follows:

***located relation* (symbol: @):** Represents the anatomical location of a disease or finding (e.g., "完全閉塞 (complete occlusion) @ 冠動脈 (coronary artery)").

***polarity relation* (symbol: /):** Indicates whether a test result is high or low, or whether a treatment was effective or not (e.g., "SpO2 / 低値 (low)"). All numerical test results in the case report are converted to polarity within the causal tree.

***tested relation* (symbol: =):** Specifies the test from which a finding was obtained (e.g., "心エコー (Echocardiogram) = 僧帽弁逆流 (mitral valve regurgitation)").

***featured relation* (symbol: ∗):** Represents details such as laterality or appearance features of a disease or finding (e.g., "泡沫状 (foamy) ∗ 痰 (sputum)").

In addition, the head node may have a special prefix, **H:**. This symbol indicates that the node represents a medical history or treatment. For example, "H: アルコール性肝線維症 (Alcoholic liver fibrosis)" indicates that the parent disease has a history of alcoholic liver fibrosis. Similarly, "H: ステロイド (Steroid) / 有効 (Effective)" indicates that the parent entity was treated with steroids, and the treatment was effective.

The head entity of located or polarity relation is the preceding one and that of tested and featured relation is the succeeding one. Modifier relationships can be combined, such as in "MRI = DWI 高信号 (high signal) @ 右 (right) ∗ 大脳半球 (cerebral hemisphere)." For example, the case in Figure 3 shows that the condition of acute myocardial infarction caused chest pain, complete coronary artery occlusion, and mitral valve regurgitation. Moreover, "mitral valve regurgitation" resulted in a "low SpO2" test result and "foamy sputum", and it was observed through an "echocardiogram".

2.1 Dataset Construction: J-Casemap

This subsection introduces the collection of the CTE dataset, named J-Casemap. All annotated data are based on case reports in internal medicine. The most experienced doctor (a co-author of this paper) first drafted the annotation schema. The annotation was then conducted by the doctors with at least ten years of experience (See Section 2.2 for details). They made iterative revisions to the annotation schema and cross-validation of the annotation for years to complete around 15,000 medical case reports. After excluding inappropriate data, the final dataset consisted of 14,094 cases.

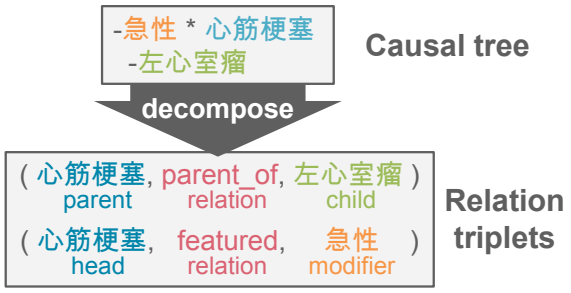


Figure 2: A tree summary is decomposed into triplets.

Since all case reports included in the J-Casemap dataset are based on the J-CaseMap case search database[1] that requires membership for access, they cannot be released publicly. We will instead release 100 causal tree samples[2] based on public case reports from the Japan national medical license examination. In fact, we investigated publicly available Japanese case report sources such as J-STAGE, but their copyright policies do not permit annotated versions of the case reports to be public. We made our best effort in this regard, and using data from the national medical licensing examination remains the only option at this moment.

2.2 Details of Manual Annotation

The annotators were instructed on the annotation scheme—specifically, the structure of the causal trees—and all annotations produced by them were reviewed and revised as necessary by the chief annotator who took the lead in designing the causal tree task. Therefore, the consistency of the annotations has been sufficiently ensured.

The cost associated with annotating new causal trees is described below. Various methods can be considered for annotating causal trees. In our study, the annotations were carried out by the same clinicians who designed the causal tree format. When extending to other data sources, hiring annotators familiar with the causal tree format, such as those we employed, is expected to be more costly. When creating your own CTE dataset, you can reduce annotation costs through optional methods that are better suited to your specific context, such as the following:

- Hire multiple clinicians and introduce major-

[1] https://www.naika.or.jp/j-casemap/
[2] https://github.com/ku-nlp/J-CaseMap

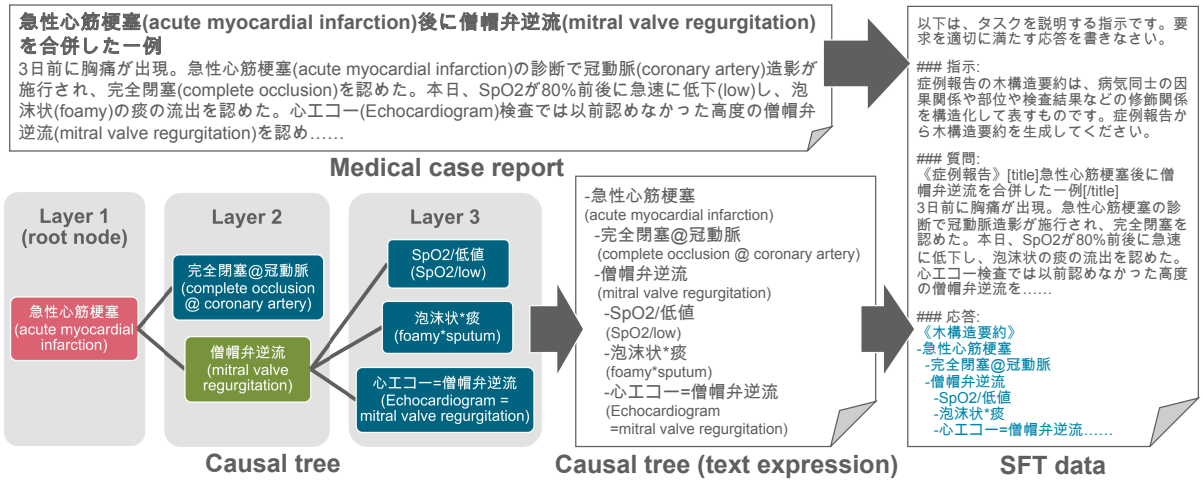


Figure 3: The SFT data example. The English translation version of the instruction (immediately following "###指示:") is as follows: "Causal trees of case reports represent causal relations among diseases and modifier relations such as anatomical locations or test results. Please generate a causal tree from the given case report."

ity voting to reduce dependence on the expertise of each single annotator

- Rigorously define an annotation scheme tailored to the target data source and provide detailed guidelines to annotators
- Hire crowd workers to create a first draft, which is then revised by a expert

As our annotation schema is iteratively refined in the future, the criteria will become more formalized, and example cases will accumulate—both of which will help to lower the barrier for future annotation efforts.

2.3 The generalizability of causal trees

Our task and dataset are tailored to Japanese case reports in internal medicine. We are also considering expanding to case reports from other medical specialties; however, this will be addressed as future work. Moreover, to extend our approach to different domains, such as other languages or clinical contexts, we must take the following factors into account:

The distribution of diseases may vary by region. For example, endemic diseases related to specific cultures or lifestyles could be more prevalent.

The optimal causal tree format may differ depending on the medical specialty or clinical context. For instance, when applying our method to radiology reports, the causal trees may not be as deep as those for case reports in internal medicine. Additionally, modifiers such as anatomical locations may not be directly applicable in fields like dermatology or psychiatry.

3 Automated CTE Models

This section introduces two comparable methods of automatic causal tree generation: the RE method and the generation method.

3.1 RE Method (baseline)

The RE task is originally designed to extract triplets of relationships between entities, instead of the tree structure. Thus, we first decompose a tree summary into a list of triplets (Figure 2) with each triplet assigned by one relation type among the set: {*parent_of*, *located*, *polarity*, *tested*, and *featured*} defined in Section 2.

RE methods typically require entity span information in the input texts. However, our dataset does not include span annotations for entities in the case reports. Ozaki et al. (2022) applied distant supervision to heuristically align nodes with words in the text, thereby generating pseudo-labeled data. A supervised model trained on this data was then used to predict relation triplets. Following this approach, we train an RE model as a baseline in this paper. However, distant supervision inevitably introduces substantial noise in span alignment, which becomes a bottleneck that limits the performance of RE models.

Recently, generation-based approaches (Zeng et al., 2020; Zhang et al., 2020; Wadhwa et al., 2023; Wan et al., 2023) in an end-to-end manner

(i.e., shorten the need of span information) have achieved performance on sentence-level RE tasks that rivals or even surpasses traditional RE models. Moreover, the fact that LLMs have recently passed the Japanese medical licensing exam (Kasai et al., 2023), suggests LLMs are capable of learning extensive medical knowledge. All these findings indicate that the LLM-based generation method could be highly suitable for our CTE task. The potential challenge lies in that our task is much more complex than sentence-level RE.

3.2 Generation Method (proposal)

In this study, we propose to solve CTE using LLMs, referred to as the generation model. Apart from not relying on noisy spans like RE models, the generation model also benefits from being able to refer to previously predicted triplets as context, allowing it to maintain consistency across triplets.

Since LLMs take textual input of the pairs of case reports and tree summaries, the tree structure must be converted into certain forms of **text representation** as shown in Figure 3. We converted the tree structure into text using a depth-first linearization method with indentation indicating the depth information. In this representation, each line corresponds to a node, and the depth of indentation indicates the ***parent_of*** relationship between nodes. As recent LLMs are typically trained on datasets that include code (such as Python), using indentation to represent nested structures is considered a natural format for LLMs. For determine the textual representation of the tree structure, we also experimented with a bracket-based format to represent the nested structure. However, it was not adopted because the nested structure broke down the output format, making evaluation impossible.

We conduct two-step training to derive our generation model.

Continual pretraining (domain adaptation): Since solving CTE requires highly specialized expertise, we leverage continual pretraining to inject the Japanese medical domain knowledge into the base models. Our Japanese medical corpora are collected from two sources. One is the abstracts from Japanese medical papers, the other is the Japanese version of English MedPub translated by human experts. In summary, we collect high-quality medical data (approximately 2B tokens) for the pretraining process.

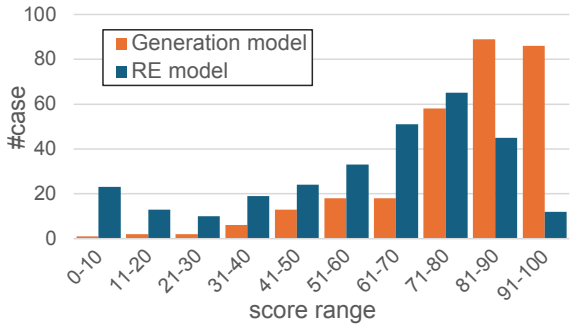


Figure 4: Manual evaluation on the same 300 cases. The generation and RE models achieved average scores of 82.7 and 62.5, respectively.

Supervised fine-tuning: We implemented supervised fine-tuning (SFT) on our collected J-Casemap data as shown in Figure 3. Supervised fine-tuning is a technique that uses labeled data to adapt pre-trained LLMs to specific downstream tasks. The prompt template filled with pairs of case reports and tree summaries is fed into LLM for SFT. The blue part in the prompt demonstrates that only the tree summary is used to calculate the cross-entropy loss for updating model parameters.

4 Evaluation

Comprehensively evaluating CTE requires the medical perspective of human clinicians to differentiate the importance of nodes, *parent_of* relationships, and modifiers for extracting salient diagnostic information. Since existing automatic evaluation metrics in RE fail to align with human clinicians (as later shown in section 6.1), we propose a weighting method emphasizing human preference to narrow the gap.

In this section, we will introduce the manual evaluation and the automatic evaluation, including our proposed weighting method.

4.1 Manual Evaluation

The manual evaluation is scored on a scale of 0 to 100. The scoring criteria mainly follow a deduction system, where the less amount of manual post-edit is needed, the higher score is assessed, and vice versa. Human doctors are naturally allowed to focus more on those important diseases and any associated diseases in the tree structure based on their expertise. Consequently, modifier relations such as findings and locations are considered less important than the *parent_of* causal relations in the trees. The most important dis-

ease often corresponds to the root node of the tree, and shallow layers tend to be more important than deeper layers.

The human evaluation includes 0-100 scores and brief comments explaining the reasons behind each score. e.g., If score was deducted due to an error in causal relationships between nodes: "[comments] 顕微鏡的多発血管炎の下流に、並行して肥厚性硬膜炎と下垂体前葉炎があると考えるべき。" (Hypertrophic pachymeningitis and anterior hypophysitis should be considered as parallel downstream nodes of microscopic polyangiitis.) "[score] 80 点。" (80 points.)

The scores were assigned based on the amount of post-processing deemed necessary.

4.2 Automatic Evaluation

To utilize automatic metrics, the output of the structured summary was broken down into a set of triplets, which were then compared to the set of correct triples. A correct prediction was defined as one where both the entities and the relationship between them matched. Precision, Recall, and F-score were calculated based on the number of correct prediction triplets.

In a entity matching for judging the correctness of triplet, minor variations in notation and typographical errors were allowed to some extent. First, a thesaurus was used to convert entities into their representative forms. Next, the edit distance between the output and correct entities was divided by the length of the correct entity, and if this ratio was below a threshold, the entities were considered a match. In this experiment, the threshold was empirically set at 0.5. However, for polarity information among modifier relations, no variations were allowed, and only exact matches were considered correct.

Proposed weighting method Since existing triplet-based evaluation treats all triplets evenly, it fails to reflect human preference. In our experiments, each triplet was weighted based on the depth d of the node and the presence of modifier relations. The depth of an entity is calculated as the depth of its parent entity plus 1, and the depth of a triplet is equal to the depth of the parent entity or the head entity inside. In our automatic evaluation method, when decomposing causal trees into triplets, we use a dummy entity "[root]" with the depth $d = 0$ as the parent of the root node. For the example in Figure 3, the depth of the triplet "([root], parent_of, 急性心筋梗塞)" is 0, and the depth of the triplet "(急性心筋梗塞, parent_of, 僧帽弁逆流)" is 1. We design a weighting method of each triplet as follows:

$$W = \frac{1}{1 + Cd} x_{relation}$$

$x_{relation}$ is 1 when the relation type is $parent\_of$, and $\frac{1}{2}$ if not. C is a constant hyper-parameter that can be tuned. d is the triplet depth.

These weighting methods are heuristically determined by referencing the manual evaluations conducted by highly experienced clinicians, who emphasized those top layers in the tree summaries (e.g., the root) and *parent_of* relations over other relation types. Details of the weighting formula design and hyperparameter selection are provided in Appendix A. The hyperparameter $C = 2$, which shows the highest correlation coefficients to human scores, is used in the following experiments.

5 Experiment Setups

This section describes the settings for continual pretraining and SFT. See Appendix B for details of continual pretraining, prompt templates, and hyper-parameters.

Base LLMs As general-domain LLMs for Japanese processing, we leverage the instruct version of multilingual Japanese LLM-jp-13b-v1 (Aizawa et al., 2024), and Japanese Swallow-13b (Fujii et al., 2024).

Continual pretraining We totally trained one epoch on the 2B tokens for each model. For those continually pre-trained LLMs, we re-name them by adding the prefix "Med-."

Supervised fine-tuning We divided J-Casemap into 13,426 training cases, 200 development cases, and 468 test cases. We used LoRA (Hu et al., 2022) as the SFT method.

Baseline exploration Initially, we considered a wider range of baseline models, including RE models with different configurations and generative models under 0-shot/few-shot settings. However, in the end, we decided not to include the scores of other RE models and non-SFTed LLMs for the following reasons:

- We did not experiment with additional RE models due to a bottleneck caused by the quality of weakly supervised data, which limits the performance gains achievable with different RE architectures.

	Triplet-based evaluation						
	w/o weight			w/ weight			Manual evaluaton
	P	R	F1	P	R	F1	
RE model (DeBERTa)	50.7	48.2	49.4	41.2	51.4	45.8	62.5
LLM-jp-13b-v1	48.0	48.9	48.4	50.5	50.0	50.2	82.7

Table 1: The comparison between automatic and manual evaluation on the subset of 300 test cases. To be noticed, manual scores ranging from 0-100 are not directly comparable to the automatic triplet F1.

		Domain	Precision	Recall	F1
RE model (Ozaki et al., 2022)	DeBERTa	general	40.7	50.1	44.9
Generation model (Proposed method)	LLM-jp-13b-v1	general	48.2	49.1	48.6
	Swallow-13b	general	52.0	54.2	53.3
	Med-llm-jp-13b-v1	medical	48.3	49.2	48.8
	Med-swallow-13b	medical	**52.8**	**54.3**	**53.6**

Table 2: The automatic evaluation for the CTE task. "Med-" denotes the continually pretrained models.

- Without SFT, it becomes challenging for models to adhere to the required tree-structured format. Non-SFTed generative models often result in nearly zero F1-scores because they cause many formatting errors. Regarding the use of non-SFT generative models in a few-shot setting, we encountered a limitation with the maximum input sequence length because of the long case reports and complex tree structure, which allowed us to insert only a single example. 1-shot setting also results in nearly zero F1-scores.

Eventual RE Baseline We fine-tune models via the distant supervision approach mentioned in Section 3.1. JaMIE (Cheng et al., 2022) is the backbone RE model, and the encoder is initialized by Japanese DeBERTa (He et al., 2023). Other possible baselines like zero-shot or few-shot without SFT were not adopted in this experiment because they all fail to follow the causal tree output format and achieve near-zero triplet F1 scores.

6 Experimental Results

6.1 Pre-examination for Optimizing Automatic Evaluation

As a pre-examination of evaluation metrics, we chose the RE model and generation models based on LLM-jp-13b-v1 as our subjects. We fine-tune both models on the J-Casemap train set. We randomly sample 300 cases from the test set to compare the automatic and manual evaluations for the RE and generation models. To be clarified, the manual evaluation is scored on a scale of 0-100 and is not directly comparable to the automatic F1 score. Figure 4 shows the manual evaluation results. The generation model achieved an average score of 82.7, significantly outperforming the RE model by 20.2 points.

However, in the vanilla triplet evaluation (w/o weight) of Table 1, the RE model obtained a slightly higher score than the RE model, which substantially contradicts the human evaluation results. Such inconsistency suggests that the vanilla metric, lacking a focus on those salient entities, does not align with human evaluation. After the weighting method was applied, the correlation between the triplet score and the human score was improved from 0.604 to 0.646 in Figure 7. Consequently, the generation model obtained significantly higher scores than the RE model in the new metric (w/ weight), which suggested improved consistency with the human evaluation and better reflection of the doctors' preferences. Please see Appendix D.1 for details on case studies of triplet weighting and the evaluation results.

6.2 Main Results

The automatic evaluation scores are shown in Table 2. All generation models outperformed the RE models substantially. Swallow-13b demonstrates stronger performance, likely because it is built on the powerful LLaMA, while LLM-jp models are trained from scratch. Domain adaptation through continual pretraining further improves the scores slightly. More detailed investigations are conducted in the later training curve part.

		Domain	Precision	Recall	F1
RE model (Ozaki et al., 2022)	DeBERTa	general	23.5	67.7	34.9
Generation model (Proposed method)	LLM-jp-13b-v1	general	64.9	59.4	62.0
	Swallow-13b	general	**69.2**	63.6	**66.3**
	Med-llm-jp-13b-v1	medical	64.9	60.3	62.5
	Med-swallow-13b	medical	66.1	**65.8**	66.0

Table 3: The automatic evaluation for the root node only. "Med-" prefix denotes the continually pretrained models.

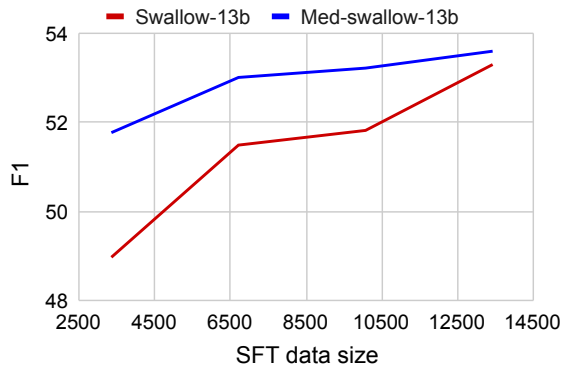


Figure 5: Triplet-based F1 scores of fine-tuned models in settings with varying amounts of SFT data (25%, 50%, 75% and 100%).

As discussed in Section 4.1, manual evaluations by clinicians prioritize the salient information, such as the primary disease. We compute the triplet F1 for the root nodes only, which can be viewed as a primary disease classification task requiring models to capture the primary disease of a case report, as shown in Table 3. The precision of the generation model significantly outperformed that of the RE model. This indicates that the generation model adequately detects the focus of the case compared to the RE model. Root scores detail is discussed in Appendix C.

Training Curves of general domain and medical domain LLMs We compare F1 scores of LLMs fine-tuned with different data sizes (25%, 50%, 75%, and 100%) in Figure 5. The medical model consistently outperforms the general model under four data size settings, especially when the data size is low (e.g., 25%, 50%, and 75%). Given the fact that only 2B tokens of medical corpora are leveraged during continual pretraining, which is relatively a small size, we are optimistic about the use of larger volumes of domain corpora and more advanced domain adaptation techniques. We leave these directions for future work.

	MedQA	MedMCQA	IgakuQA
base	25.6	33.6	33.9
+ J-Casemap	22.7	29.3	26.3
+ MedQA	29.3	27.6	37.6
+ 2-stage	34.7	32.2	34.1
+ mix	**37.0**	**34.1**	**38.6**

Table 4: Accuracy of QA tasks. We compare the following three SFT settings: (1) only J-Casemap; (2) only MedQA; (3) first J-Casemap then MedQA (2-stage); (4) merge J-Casemap and MedQA (mix). The evaluation were conducted using JmedBench (Jiang et al., 2024).

6.3 Can CTE help Medical QA?

The J-Casemap data has the potential to serve a variety of other medical tasks, given the comprehensive understanding required for a model to complete the CTE task.

We conduct the experiments on Japanese medical question answering (QA) benchmarks, like Japanese medical licensing exam dataset IgakuQA (Kasai et al., 2023) and the translated medical QA datasets MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022) to see whether a model trained on J-Casemap can be beneficial to medical QA tasks. For each benchmark, we used Med-swallow-13b as the base model, and the training set of MedQA or added J-Casemap for fine-tuning; a prompt example is shown in Appendix B.3.

As shown in Table 4, for MedQA, both the "2-stage" and "mix" settings outperform SFT on MedQA alone. For MedMCQA, even SFT on MedQA hurts the performance due to the out-of-domain distribution; after adding J-Casemap in the "mix," the performance improves and beats the base model. In particular, "mix" performs better than "2-stage" and achieves the highest scores on all QA datasets. This indicates that our J-Casemap data is valuable for facilitating LLMs' medical abilities in various tasks.

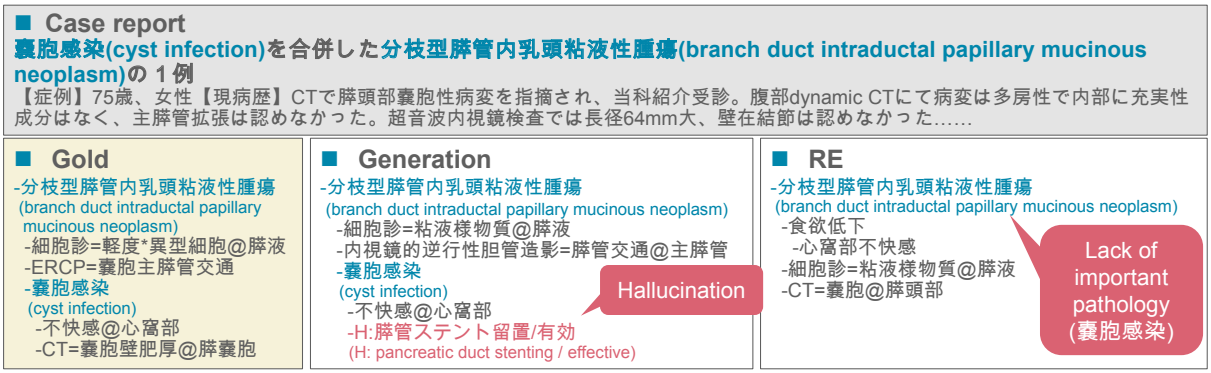


Figure 6: Case study of an automatically generated causal tree. Blue entities are the focus of the tree.

6.4 Case Study

Examples of causal trees generated by the generation model are shown in Figure 6. Most errors in the generation model's output are failures of entity extraction. Additionally, the problem of hallucinations, where the model generates entities not present in the original case report, was sometimes observed in the causal trees (See Appendix D.2 for details). In contrast, due to the nature of information extraction, RE models did not exhibit such hallucinations. Further studies will explore to what extent the hallucination issue can be mitigated through improvements to the base LLM or additional training using medical domain texts.

7 Related Works

Various RE tasks have been undertaken in the medical domain for different purposes. For instance, Parikh et al. (2019) aimed at improving access to medical information and Wolf et al. (2019) tackles entity extracting from trustworthy medical literature for question-answering assistants. Dialogue-based entity extraction tasks designed to assist in electronic medical record (EMR) entry (Jeblee et al., 2019; Xia et al., 2022) have all been explored. More complex tasks include extracting predefined medical entities and their conditions (Gao et al., 2023; Cheng et al., 2022; Yang et al., 2023) and extracting findings and characteristics from radiology reports (Park et al., 2024).

While recent LLMs have demonstrated the ability to perform RE as a generation task in general domains (Wadhwa et al., 2023; Wan et al., 2023), there are few studies applying LLMs to medical RE, focusing only on temporal relations between diseases (Kougia et al., 2024) or drug-related RE (Bhattarai et al., 2024). While these studies focus on the conditions of medical entities, CTE is unique in its focus on the causal relationships between higher-level diseases.

For collecting data on causal relationships between diseases and findings, (Khetan et al., 2022) proposed a dataset with annotation specifications covering four types of causal relationships between diseases. Compared to CTE annotation specification, it differs because CTE constructs a tree structure and extract primary diseases as root.

8 Conclusion

We proposed a novel task, causal tree extraction (CTE), which requires expert-like text comprehension, and we constructed the J-Casemap dataset containing case reports and their causal trees. We tackled the CTE task by fine-tuning LLMs and achieved higher scores than existing methods across both automatic and human evaluations. Furthermore, we improved the automatic evaluation through heuristic weighting, which reflects clinicians' preferences in automatic evaluation scores.

The causal tree of case reports is useful not only for clinicians but also for LLMs to train along with other medical tasks, such as question answering tasks. The insights into advanced causal reasoning have the potential to be applied in domains beyond medicine.

9 Limitations

Hallucination problems were seen in the LLMs' outputs, but we have not discussed the solutions in this paper. In future work, more advanced approaches like Retrieval-augmented generation or entity linking between the causal tree and the case report text are probably needed to find the supporting evidence towards more reliable generation.

Besides, all of the case report data in this experiment are from internal medicine, which potentially limits the scope of this study. We are ambitious in envisioning the future where the J-Casemap data is expanded beyond internal medicine to other departments, ultimately establishing a unified standard across different medical fields.

The last limitation lies in the automatic evaluation of CTE. Even though we already improved automatic metrics, developing more comprehensive and accurate automatic metrics that more closely resemble manual evaluation is necessary.

10 Ethical Statement

The copyright of the J-Casemap dataset belongs to the Japanese Society of Internal Medicine, making it difficult to make the data publicly available due to privacy and security concerns. we will release the final version of the annotation schema and 100 causal tree samples based on public case reports without ethical concerns from the Japan national medical license examination.

Acknowledgments

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant No. JPJ012425 and by JST BOOST, Grant Number JPMJBS2407.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *CoRR*, abs/2407.03963.

Antropic. 2024. Claude 2. https://www.anthropic.com/news/claude-2.

Kriti Bhattarai, Inez Y. Oh, Zachary B. Abrams, and Albert M. Lai. 2024. Document-level clinical entity and relation extraction via knowledge base-guided generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 318–327, Bangkok, Thailand. Association for Computational Linguistics.

Judith Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *The New England journal of medicine*, 355:2217–25.

Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. 2022. JaMIE: A pipeline Japanese medical information extraction system with novel relation annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3724–3731, Marseille, France. European Language Resources Association.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Lei Gao, Xinnan Zhang, Xian Wu, Shen Ge, and Yefeng Zheng. 2023. Dialogue medical information extraction with medical-item graph and dialogue-status enriched representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13311–13321, Singapore. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. Extracting relevant information from physician-patient dialogues for automated clinical note taking.

In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74, Hong Kong. Association for Computational Linguistics.

Tierney LM Jha AK, Collard HR. 2002. Diagnosis still in question. *New England Journal of Medicine*, 347(22):1805–1806.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2024. Jmedbench: A benchmark for evaluating japanese biomedical large language models. *Preprint*, arXiv:2409.13317.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on japanese medical licensing examinations. *Preprint*, arXiv:2303.18027.

Vivek Khetan, Md Imbesat Rizvi, Jessica Huber, Paige Bartusiak, Bogdan Sacaleanu, and Andrew Fano. 2022. MIMICause: Representation and automatic extraction of causal relation types from clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 764–773, Dublin, Ireland. Association for Computational Linguistics.

Vasiliki Kougia, Anastasiia Sedova, Andreas Joseph Stephan, Klim Zaporojets, and Benjamin Roth. 2024. Analysing zero-shot temporal relation extraction on clinical notes using temporal consistency. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 72–84, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Jun-

tang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ryuichi Ozaki, Hirokazu Kiyomaru, Fei Cheng, Sadao Kurohashi, Hisahiko Sato, and Ryozo Nagai. 2022. 弱教師学習に基づく症例報告の構造的要約. In **第 *26* 回日本医療情報学会春季学術大会**.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Soham Parikh, Elizabeth Conrad, Oshin Agarwal, Iain Marshall, Byron Wallace, and Ani Nenkova. 2019. Browsing health: Information extraction to support new interfaces for accessing medical evidence. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 43–47, Minneapolis, Minnesota. Association for Computational Linguistics.

Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Özlem Uzuner, Martin Gunn, and Meliha Yetisgen. 2024. A novel corpus of annotated medical imaging reports and information extraction results using BERT-based language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1280–1292, Torino, Italia. ELRA and ICCL.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong,

Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun,

Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chak-

ladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Martin Wolf, Volha Petukhova, and Dietrich Klakow. 2019. Term-based extraction of medical information: Pre-operative patient education use case. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1346–1355, Varna, Bulgaria. INCOMA Ltd.

Yuan Xia, Zhenhui Shi, Jingbo Zhou, Jiayu Xu, Chao Lu, Yehui Yang, Lei Wang, Haifeng Huang, Xia Zhang, and Junwei Liu. 2022. A speaker-aware co-attention framework for medical dialogue information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4777–4786, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhe Yang, Yi Huang, and Junlan Feng. 2023. Learning to leverage high-order medical knowledge graph for joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9023–9035, Toronto, Canada. Association for Computational Linguistics.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514.

Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. 2020. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 236–246, Online. Association for Computational Linguistics.

A Searching optimal triplet weights

We design two weighting methods for the triplet evaluation as follows:
Weighting method 1:

$$W = \frac{1}{1 + Cd} x_{relation}$$

$x_{relation}$ is 1 when the relation type is $parent\_of$, and $\frac{1}{2}$ if not.
Weighting method 2:

$$W = \frac{1}{C^d} x_{relation}$$

$x_{relation}$ is 1 when the relation type is $parent\_of$, and $\frac{1}{C}$ if not. C is a constant hyper-parameter that can be tuned. d is the triplet depth.

We further calculate the correlation coefficients of weighting factors in automatic evaluations, shown in Figure 7. It was noticed that the weighting of prioritized entities in lower layers showed a higher correlation with manual evaluations. However, when extreme weighting was applied, the correlation with manual evaluations decreased. The Appendix D.1 provides a more detailed analysis.

After we assign heuristic weights to the automatic evaluation, the performances become closer to the human clinicians, as shown in Table 1. Currently, automatic evaluation is still unable to match human doctors' precision in judging salient information and ideally identifying entities. We consider this an open issue for future research. The weighting pattern 1 ($C = 2$), which shows the highest correlation coefficients to human scores, is used in all the following experiments.

B Experiments details

B.1 Continual Pre-training

Our dataset constructs of two corpora, 0.9B tokens of English PubMed Abstracts & PubMed Central articles from The Pile and 0.9B tokens of Japanese medical texts used by JMedRoBERTa. We used Megatron-LM as the training framework. We used 2 nodes 8 40GB A100 GPU with 61,035 steps in total. We selected global batch size of 32, learning rate of 3e-6 and warmup ratio of 0.1 in our training.

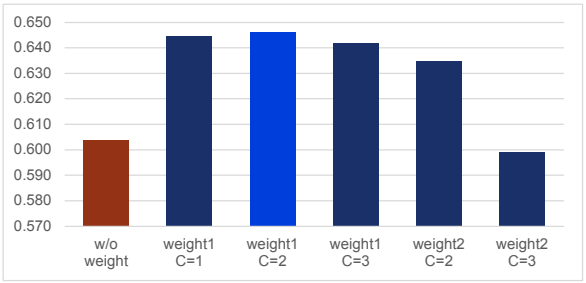


Figure 7: Correlation between manual scores and automatic scores. 600 causal trees generated by the RE model and generation model for 300 case reports were automatically evaluated, and correlation coefficients with human scores were calculated.

B.2 Prompt Template for Different LLMs

Due to differences in model compatibility, two types of inference templates were used according to the model. The inference templates follow the examples provided on the model card for each model. Additionally, a beginning-of-sequence (BOS) token was added at the start of the prompt, and an end-of-sequence (EOS) token was added at the end of the LLM-generated outputs during training and testing.

B.3 MedQA prompt

We present the prompt for MedQA SFT in Figure 8.

B.4 Hyperparameters

We present the detailed hyper-parameters of the pretraining in Table 7 and the fine-tuning stage in Table 5.

C Comparison of Model Prediction Trends

This section provides a more detailed analysis and comparison of the RE model and the best generation model, Med-swallow-13b. Statistics on the number of generated triplets and root nodes are shown in Table 6. Compared to RE models, generation models extracted more triplets and had fewer omissions in information extraction. Additionally, the RE model predicts far more root nodes than Gold, while the Generation model predicts about the same number of roots as Gold. This indicates that the generation model was able to designate a few critical entities as root elements and link other entities comprehensively downstream. On the other hand, the RE model enumerated extracted entities that did not have identified relationships as root elements. These aligns with the experimental results that showed a significant difference in precision and a smaller difference in recall.

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい。

指示:
医師国家試験を解いてください。

質問:
整形外科の若手研修医が、学科長を担当医として手首のトンネル修復を行っています。手術中、研修医は不注意にも屈筋の腱を切ってしまいます。腱は合併症なく修復されます。担当医は研修医に対し、患者は問題なく回復するだろうし、患者が不必要に心配することを避けるため、この軽微な合併症を報告する必要はないと伝えます。彼は研修医に、この合併症を手術報告書から省くように指示します。次に研修医が取るべき正しい行動は次のうちどれですか？
A. 患者にエラーを開示し、それを手術報告書から省く
B. 患者にエラーを開示し、それを手術報告書に記入する
C. 担当医に対し、このミスを開示しないことはできないと伝える
D. 医師を倫理委員会に報告する
E. 手術報告書の作成を拒否する

応答:
担当医に対し、このミスを開示しないことはできないと伝える

Figure 8: MedQA prompt for SFT. The blue parts were used for loss calculation.

D Case Study

D.1 Evaluation Comparison

Examples of a case study that focuses on automatic evaluation are shown in Figure 9. In both of the examples, the generated summary of the RE model got good scores in the human evaluation, but the automatic evaluation score is very low.

The reason for the evaluation failure of the case 1 is that the influence of matching errors for entities in lower layers becomes too significant, leading to a lower correlation with the manual evaluation. While manual evaluations can perfectly match entities, automatic evaluations may fail to do so.

The reason for the evaluation failure of the case 2 is the ambiguity of the causal relationship. It is occasionally difficult to determine which is the cause and which is the result of the causal relationship between diseases, especially when multiple diseases are combined.

Even with the most correlated weighting, the correlation coefficient remained around 0.6, indicating a substantial gap between manual and automatic evaluation scores. To perform automatic evaluation more similar to human evaluation, a more flexible evaluation method than evaluation by triplet comparison is required.

model	LLM-jp-13b-v1 Med-llm-jp-13b-v1	Swallow-13b Med-swallow-13b
batch-size	64	64
max_seq	2048	4096
learning rate	1.00E-04	1.00E-04
warmup ratio	0.1	0.1
LoRA target modules	c_attn, c_proj, c_fc	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, lm_head
LoRA alpha	32	32
LoRA r	8	8
LoRA dropout	0.05	0.05

Table 5: Hyper-parameters of fine-tuning.

	Triplets	Root node
Gold	14,049	545
RE model	13,343	1,584
Generation model	14,453	550

Table 6: The statistics on the number of triplets and root nodes. Med-swallow-13b is used as generation model.

Hyper-parameters	Value
Constant learning rate	$3.00e^{-6}$
Warm-up schedule	Linear
Warm-up ratio	0.03
Weight decay	0.1
Data type	bf16
Global batch size	32

Table 7: Hyper-parameters of pretraining

D.2 Hallucinations

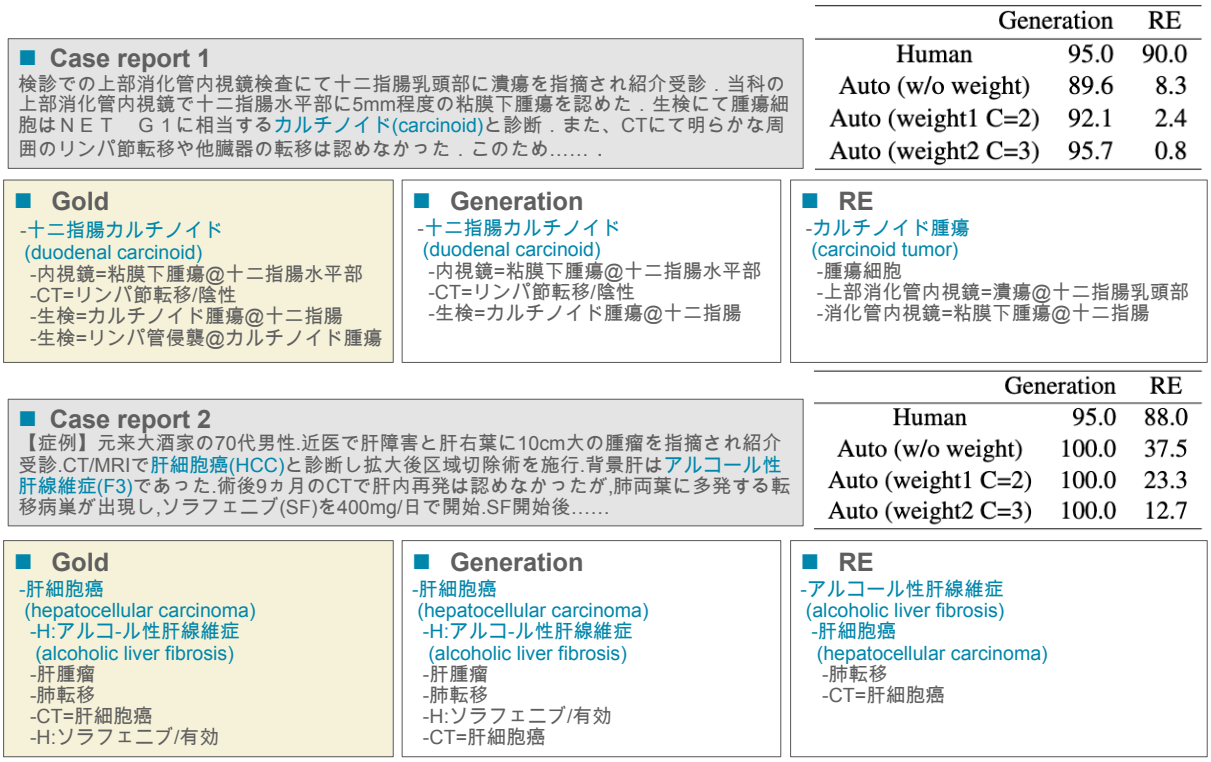
Examples of a case study that focuses on hallucination are shown in Figure 10. Addressing hallucination issues is indeed an important direction, and we plan to explore this more thoroughly as future work. Below, we provide an analysis of notable hallucinations observed in our system.

The errors found in the causal trees generated by our proposed method can be categorized as follows:

1. Missing necessary entities from case reports or errors in the relationships between entities. Because case reports assume that readers possess medical knowledge, they rarely explicitly describe the medical relationships between entities in the text. Consequently, errors may occur when the model's limited medical knowledge leads it to misunderstand that an entity mentioned in a case report should not be included in a causal tree, or to incorrectly assess the relationships between entities.

2. Unnecessary entity extraction or hallucinated medical terms generation (e.g., Figure 10 of the Appendix.) In the automatic generation of causal trees, the model occasionally produces terms that do not appear in the original case report texts. These hallucinations can be broadly categorized into two types:

   - **Terms that are semantically similar to the main topics in the text but are not explicitly mentioned.** For example, in a case report concerning "大細胞神経内分泌癌 (large cell neuroendocrine carcinoma)," the fine-tuned model output "悪性リンパ腫 (malignant lymphoma)" as the root node instead of "大細胞神経内分泌癌 (large cell neuroendocrine carcinoma)." Although the latter was not mentioned in the case report, "大細胞神経内分泌癌 (large cell neuroendocrine carcinoma)" and "悪性リンパ腫 (malignant lymphoma)" are considered to be clinically similar malignant tumors, as they can exhibit similar symptoms and metastatic patterns. One possible cause of this hallucination may be the biased co-occurrence frequency or the positional proximity of related terms in the training data.
   - **Completely fabricated terms that do not exist in reality.** For example, in a case involving "大動脈炎症候群 (aortic arteritis syndrome)," the fine-tuned model generated a downstream node la-



	Generation	RE
Human	95.0	90.0
Auto (w/o weight)	89.6	8.3
Auto (weight1 C=2)	92.1	2.4
Auto (weight2 C=3)	95.7	0.8

	Generation	RE
Human	95.0	88.0
Auto (w/o weight)	100.0	37.5
Auto (weight1 C=2)	100.0	23.3
Auto (weight2 C=3)	100.0	12.7

Figure 9: Case study of evaluation.

beled “抗大動脈炎症候群抗体 (anti aortic arteritis syndrome antibody) / 陰性 (negative).” However, the term “抗大動脈炎症候群抗体 (anti aortic arteritis syndrome antibody) / 陰性 (negative)” does not exist in actual medical terminology. This is considered to be a hallucination influenced by the context of the case report and the surrounding output.

3. **Failure to infer contextually implied entities.** Some case reports describe scenarios in which a first disease triggers a second disease, which in turn causes a finding, representing a multi-step causal structure. In such reports, it is occasionally the case that the first disease and the finding are explicitly mentioned, whereas the second disease is omitted. In these instances, it is necessary to infer the second disease and incorporate it into the causal tree based on medical knowledge. This represents a highly challenging subtask that requires advanced domain-specific expertise.

4. **Formatting errors.** In our baseline investigation, we attempted to generate causal trees from case reports using commercial models such as ChatGPT. When testing multiple prompts specifying the format of the causal tree on models without fine-tuning, the outputs frequently contained formatting errors. Such errors hinder the decomposition of the causal tree into relational triplets, complicating subsequent evaluation. Notably, in our experiments, no formatting errors were observed in the outputs of fine-tuned generative models.

We believe these issues stem from either a failure to adequately reference the context of the input case report or from insufficient medical knowledge or retrieval errors.

E Preliminary Experiment on Non-Internal Medicine Texts

The model trained on our J-Casemap dataset can be utilized to automatically generate draft versions of causal trees, which can significantly facilitate the creation of new structured datasets. Moreover, we have found that J-Casemap dataset is also helpful for training structured prediction models on other types of medical texts beyond case reports as below.

We implement a preliminary experiment on structuring radiology reports using our dataset.

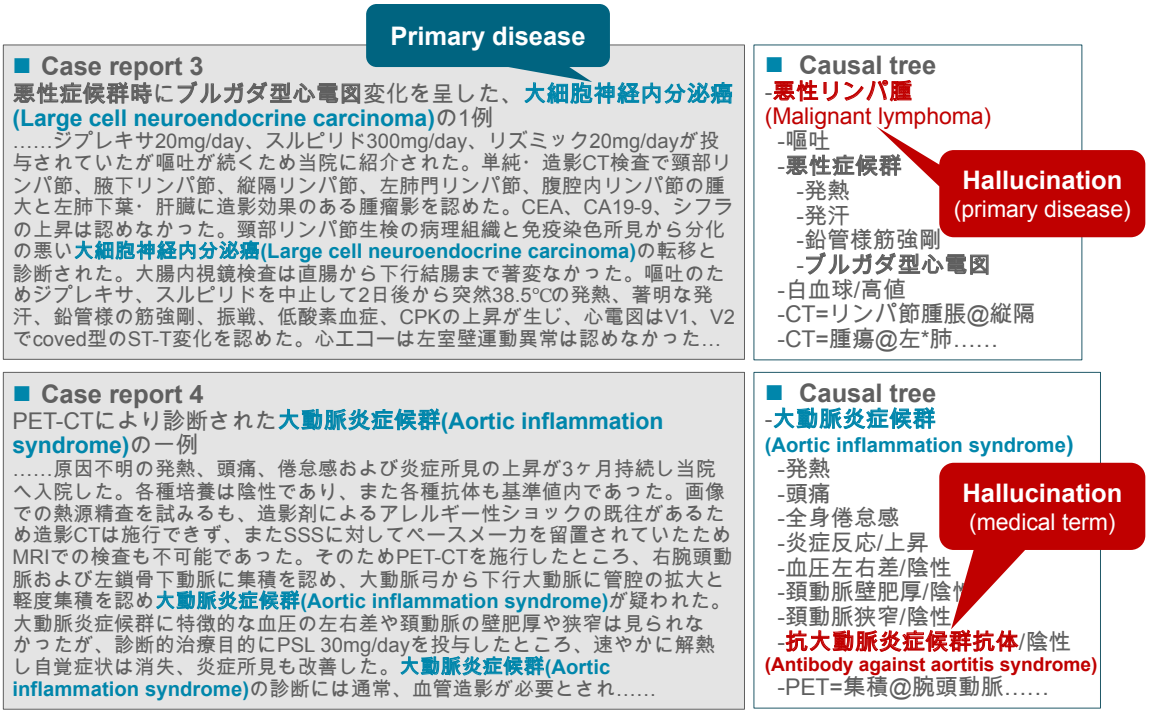


Figure 10: Case study of hallucinations.

Please note that in the radiology report task, annotations are performed on semantic blocks (often at the sentence level), which differs from the setting of the J-Casemap dataset.

In this experiment, we conducted supervised fine-tuning (SFT) under two conditions:

- Using only 100 annotated radiology reports (comprising 1,263 semantic blocks), and
- Performing SFT first on J-Casemap dataset (approximately 14,000 cases), followed by SFT on the radiology report dataset.

The results of SFT experiment is shown in Table8. The models were evaluated on a test set consisting of 104 semantic blocks. The automatic evaluation score was 81.7 when using only the radiology report data, and it improved to 85.8 when combining it with J-Casemap dataset.

These results suggest that our dataset contributes to the automatic generation of structured data in domains where structured resources are scarce. We consider the construction of datasets in other domains to be promising future work.

SFT dataset	F1
Radiation reports	81.7
2-stage	**85.8**

Table 8: Evaluation results (accuracy) of the structured radiology report. We compare three settings of SFT-trained models: SFT using only radiology reports (Radiation reports), SFT using the J-Casemap dataset followed by SFT using radiology reports (2-stage)."