症例報告テキストの因果木抽出のための自動評価手法の検討

八幡早紀子 [1] Fei Cheng[1] 黒橋禎夫 [1] 佐藤寿彦 [2] 永井良三 [3]
[1] 京都大学, [2] 株式会社プレシジョン, [3] 自治医科大学,
{yahata,feicheng,kuro}@nlp.ist.i.kyoto-u.ac.jp,
satoh@premedi.co.jp, rnagai@jichi.ac.jp

概要

症例報告の因果木抽出は、症例報告中の因果関係を木構造を用いて構造化するタスクである。因果木抽出の人手評価では、医学的知識とタスクの理解が必要となることから作業者が限られている。一方自動評価においても、柔軟性の不足やグローバル文脈の考慮の不足が指摘されている。本論文では、人手評価データを用いた大規模言語モデルのファインチューニングを行い、LLM 評価器を用いた自動評価について分析した。医師の監修を受けた採点基準を用いることで、LLM 評価は既存手法と同等の人手評価との相関を示した。さらに、訓練データの性質について分析を行い、点数の偏りよりも正解と評価対象間の類似性の重要さが示された。

1 はじめに

因果木抽出は、症例報告中の因果関係を構造化するタスクである。関係三つ組単位の関係抽出タスクと比較して、因果木抽出は症例の最も重要な第一病態を木構造の根に配置することで、第一病態を中心に症例の文脈全体の因果を構造化するという特徴を持つ。因果木抽出の人手評価では、医学的知識に加えて因果木の理解の両方を要するため作業者が限られており、自動評価手法の開発が求められる。

先行研究 [1] の自動評価では、予測された因果木を関係三つ組の集合に分解して F1 スコアを算出する。先行研究 [2] は各関係三つ組の重要度を考慮した重み付けによって、関係三つ組 F1 と人手評価の相関を向上させた。一方、関係三つ組を用いた自動評価には、単一の正解データへの依存による柔軟性の不足や、症例の文脈全体を包括的には考慮できないというボトルネックが存在する。

本研究では、医師の評価傾向を考慮した LLM 評価器を構築した。構築した LLM 評価器は、既存手法と同程度の人手評価との順位相関を達成した。また、LLM 評価器の訓練データのデータ作成過程に由来する欠点について分析を行った。

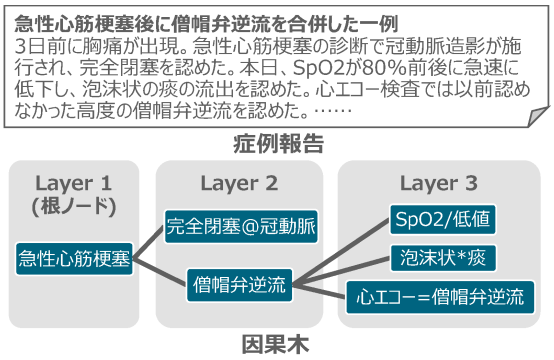

図 1 因果木の例。青色の四角形は木構造のノードを表す。レイヤーの番号は各ノードの深さに対応する。

2 関連研究

2.1 症例報告の因果木抽出

症例報告と因果木の例を図 1 に示す。因果木の各ノードは症例報告中の病態や所見などのエンティティに対応し、ノード間の親子関係は因果関係を表す。また、各ノードは内部にさらに構造を持ち、エンティティの部位や特徴、極性情報などの修飾情報を表す。因果木の詳細は付録 B に記載する。

2.2 自動評価手法

自然言語処理分野では、BLEU[3], ROUGE[4] などの自動評価手法が用いられてきた。タスクが多様化している近年ではより柔軟な評価手法へのニーズが高まり、汎用 LLM を用いてモデル出力の評価を行う LLM-as-a-judge[5] が現れた。しかし、LLM-as-a-judge によく用いられる商用モデルは、モデル更新や終了の可能性のため安定しない。また、実際の患者データを含むテキストは、商用モデルでは使用できない。これを受けて提案された PROMETHEUS[6] は、オープンな LLM を採点基準

等を含むデータで訓練することで、多くのタスクに対応可能な柔軟性を持つ LLM 評価器を構築した。

3 評価手法

3.1 人手評価

因果木抽出の人手評価では、0-100 点の範囲の点数と採点理由を説明するコメントを付与する。評価は医師の主観に基づき、因果木が要する修正の量に応じて減点される。また、医師の評価では、症例報告の**大筋**が重視される。大筋は症例の最も重要な**第一病態**とそれが引き起こす病態の関係を指し、多くは因果木の根及び根付近の浅い階層に存在する。

3.2 重み付き関係三つ組 F1

関係三つ組 F1 評価では、因果木を関係三つ組に分解する。出力・正解間で 2 つのエンティティおよび関係が全て一致する場合、その関係三つ組を正解とみなす。先行研究 [2] では、医師の評価傾向を参考に関係三つ組の重みづけを行う**重み付き関係三つ組 F1** が提案された。重み付けは以下で示される:

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

$x_{relation}$ は関係三つ組が因果関係なら 1, そうでない時 $\frac{1}{2}$ をとる。C はハイパーパラメータである。因果木を関係三つ組に分解する際、深さ 0 のエンティティを根の親として挿入し、どのエンティティが根であるかを明示的に示す関係三つ組を追加する。エンティティの深さは親の深さに 1 を足したものであり、関係三つ組の深さ d は関係三つ組の 1 つ目のエンティティの深さと等しい。関係三つ組の比較では、出力-正解間の編集距離を正解文字数で割ったものが閾値以下であれば同一とみなし、誤字脱字を許容した。ただし、修飾のうち極性情報に関しては語彙が閉じているため、完全一致の場合のみ同一とした。本論文では、経験的に $C = 2$, 閾値 0.5 とした。

3.3 提案手法: LLM 評価器

本論文では、先行研究 [6, 7] を参考に、因果木抽出の人手評価データで LLM を追加学習することで LLM 評価器を構築する。関係三つ組 F1 と比較して、LLM 評価器は症例の文脈全体を包括的に考慮可能である。さらに入力の一部として評価基準を用いることで、正解データへの依存の軽減を図る。

LLM 評価器は、症例報告、正解、評価対象の因果木および採点基準を入力として受け取り、採点理由を説明するコメントと点数を出力する。推論用テンプレートを図 6 に示す。LLM 評価器の訓練データとして用いられた医師による人手評価データには、採点理由を説明するコメントと 0-100 点の範囲の点数が含まれている。約 3,500 件の人手評価スコアの分布を図 2 に示す。人手評価データは以下の 2 つに大きく分けられる。

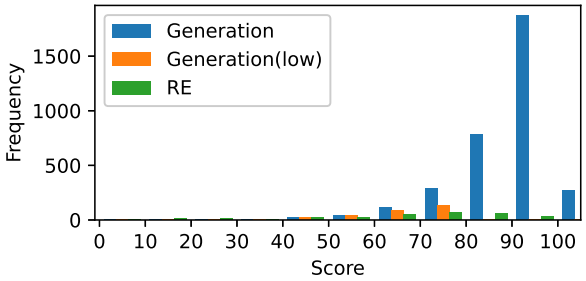

図 2 学習データの人手評価スコア分布

LLM 人手評価データ: LLM ベースの因果木抽出モデルの出力に対する人手評価データ。評価対象の因果木には複数モデルの出力が混在している。また、正解の因果木は、医師が評価対象を加筆修正することで作成された。約 3,000 件のうち、約 2,800 件を訓練データ、約 200 件をテストデータとする。訓練データのうち低得点を獲得した約 200 件を低得点訓練データ、テストデータのうち低得点を獲得した約 100 件を低得点テストデータとする。

RE 人手評価データ: 関係抽出 (RE) ベースの因果木抽出モデルの出力に対する人手評価データ。注意点として、このデータの症例報告及び正解は LLM 人手評価データの一部と共通している (評価対象は異なる)。約 300 件のうち、約 200 件を訓練データ、約 100 件をテストデータとする。

4 実験

すべての実験において、LLM 評価器のベースモデルとして、評価対象のモデルとの重複によるバイアスを避けるために Qwen3-14B[8] を選択した。ハイパーパラメータについては付録の表 3 に後述する。

4.1 因果木抽出モデルの自動評価

4.1.1 実験設定

この実験では、医師の人手評価、重み付き関係三つ組 F1, LLM 評価の 3 つの手法で因果木抽出モデルの評価を行う。LLM 評価器は LLM 訓練データお

表 1　因果木抽出モデルの自動評価結果

	人手評価	重み付き F1	LLM
DeBERTa	62.5	45.0	79.5
LLM-jp-v1-13b	82.7	48.0	81.2
SIP-jmed-llm-3-8x13b	-	55.1	83.2

よび 3 倍にアップサンプリングされた RE 訓練データを用いて訓練された。また、医師の監修を受けた採点基準を使用した。評価対象として、DeBERTa[9] ベースの RE モデル [1]、LLM-jp-v1-13b[10] および SIP-jmed-llm-3-8x13b[1)] ベースの生成モデル [2] を用いた。これらは全て約 14,000 件の症例報告-人手作成因果木ペアデータによって訓練された。自動評価では、約 500 件のテストセットを用いた。リソースの制約により、人手評価は約 300 件の上記と異なる症例データについて行われた。

4.1.2　実験結果

結果を表 1 に示す。重み付き関係三つ組 F1 および LLM 評価によるモデル間の上下関係は人手評価と一致した。また、LLM 評価は重み付き関係三つ組 F1 と比較して人手評価に近い点数を示した。

4.2　SFT データの性質の分析

LLM テストデータおよび RE テストデータについての自動評価スコアの分布を図 3 に示す。訓練データの得点分布は高得点に偏っているため、提案手法では低得点部分の評価精度が低い。また、LLM 評価器の訓練データの大半を占める LLM 訓練データの正解は、評価対象の因果木に加筆修正を行うことで作成された。一方で、実際のモデル評価では正解と評価対象の作成過程はは独立であるため、実際のモデル評価では LLM 評価器が能力を発揮できない可能性がある。本実験では、性質の異なる 2 種類の訓練データが LLM 評価器に与える影響を分析する。

4.2.1　実験設定

本実験では、低得点・RE 訓練データをアップサンプリングした場合の自動評価・人手評価の順位相関の推移を分析する。評価では、低得点テストデータ、RE テストデータおよび LLM・RE テストデータを統合したものを用いた。二つの訓練データは LLM 訓練データと比較して、どちらも得点分布が低得点部分に偏っている。一方、低得点訓練データで

1)　https://huggingface.co/SIP-med-LLM/SIP-jmed-llm-3-8x13b-AC-32k-instruct

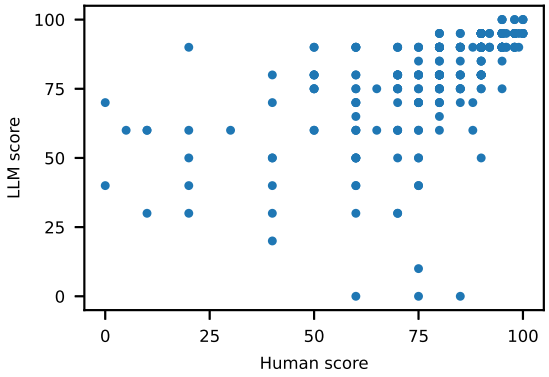

図 3　症例ごとの自動評価スコア分布

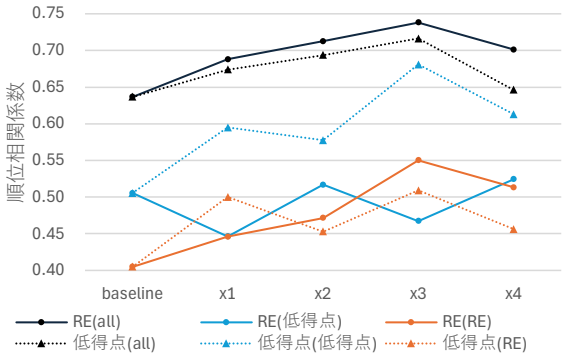

図 4　アップサンプリング倍率に応じた人手評価と自動評価の順位相関の推移。baseline の訓練データは LLM 訓練データのみであり、実線では RE 訓練データ、破線では低得点訓練データをアップサンプリングした。

は評価対象の因果木に加筆修正を行うことで正解を作成しており、RE 訓練データでは評価対象と正解の因果木が独立に作成されたという相違点がある。

4.2.2　実験結果

図 4 にアップサンプリング倍率に応じた順位相関の推移を示す。RE 人手評価データのアップサンプリング倍率 3 倍の場合、RE 人手評価データについての順位相関は向上し、評価データ全体についても最も高い順位相関を達成した。一方、低得点データのアップサンプリングでは LLM 人手評価データについての順位相関係数は向上したが、テストデータ全体の順位相関では RE 人手評価データを追加した場合を下回った。この結果は、訓練データについて、得点分布の均一さよりも正解と評価対象のデータが独立であることの重要性を示唆している。

4.3　採点基準の検討

LLM 評価器の推論時は、入力の最後に点数と内容の対応関係を示す採点基準を与える。(詳細は付録

表 2　自動評価-人手評価間の順位相関係数。**タスク指示のみ**は入力最後の評価基準を削除した場合を指す。

	all	LLM	低得点	RE
重み付き F1	0.735	0.636	0.634	0.507
医師作成評価基準	0.738	0.664	0.467	**0.550**
素人作成評価基準	0.688	0.626	0.432	0.427
タスク指示のみ	**0.781**	**0.725**	**0.634**	0.545

の図 6 に示す。) 本実験ではこの採点基準を変更し、採点基準における医学知識の重要性を検証する。

4.3.1　実験設定

この実験では、LLM 評価器の訓練時・推論時に用いる採点基準の変更または削除を行う。2 種類の採点基準を付録の図 7,8 に示す。素人作成評価基準は医学知識を持たない筆者が先行研究を参考に作成したものであり、医師作成評価基準は医師による監修を受けたものである。訓練では、LLM 訓練データおよび 3 倍にアップサンプリングした RE 訓練データを用いた。評価では、LLM・低得点・RE テストデータに関して評価を行い、人手評価・自動評価間の順位相関係数を算出した。

4.3.2　実験結果

結果を表 2 に示す。素人作成評価基準を用いた場合、相関は重み付き関係三つ組 F1 を下回り、作成者が知識をもたない場合の採点基準はかえって悪影響をもたらす可能性が示された。医師作成評価基準を用いた場合、すべてのデータで相関が向上した。LLM テストデータおよび低得点テストデータと比較して、RE テストデータでは評価基準の変更による影響幅が大きい。

また、採点基準を削除した場合、RE テストデータを除き、相関は医師作成評価基準を用いた場合を上回った。LLM・低得点テストデータでは正解と評価対象の因果木は作成過程が独立でなく類似しているため、二者を比較することで評価対象の要する加筆修正の量を容易に評価できる。一方、RE テストデータのように正解と評価対象が独立である場合、評価対象の因果木が要する修正量を直接評価することが難しい。よって、正解と評価対象が独立でない場合は評価基準なしに精度の高い評価が可能である一方、正解と評価対象が独立な場合は評価が難しいために評価基準が有効であると考えられる。実際の評価では正解と評価対象の因果木は独立であるため、医師作成評価基準が有効である可能性が高い。

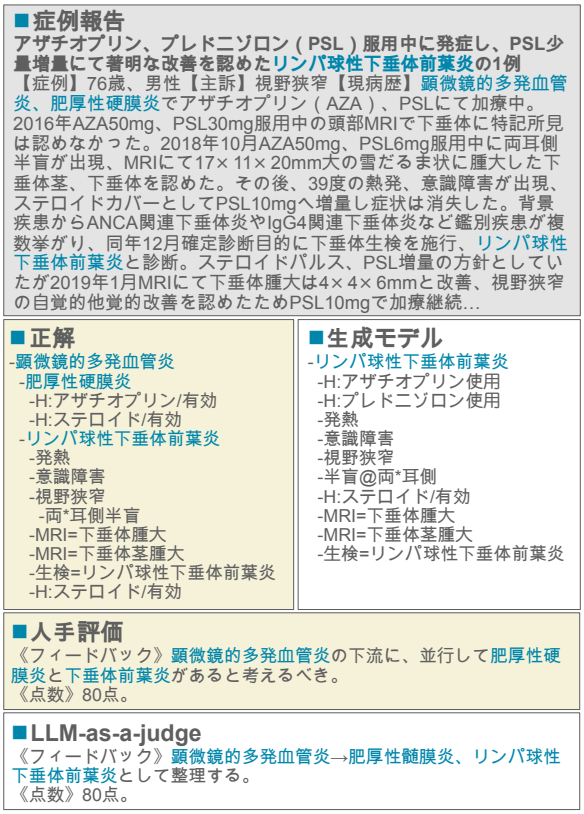

図 5　LLM テストデータに対する LLM 評価の例。

4.4　LLM 評価のケーススタディ

LLM 評価器による評価例を図 5 に示す。この例の評価対象の因果木には、症例の大筋となる顕微鏡的多発血管炎や肥厚性硬膜炎のようなエンティティが不足している。LLM 評価器は人手評価と同様に、評価対象の因果木に不足するエンティティおよび適切な因果関係について補足している。このように、LLM 評価器は医師の人手評価のように因果木の誤りについて検証可能であり、さらに評価の根拠を容易に検証できるという利点を持つ。

5　おわりに

本研究では、因果木抽出タスクのための LLM 評価器の構築を行った。LLM 評価器は、既存手法と同等の人手評価との相関を達成し、医師と同様のコメント及び点数を出力可能であった。また、訓練データの性質についても分析を行い、訓練データの正解と評価対象が独立であることの重要性を示した。さらに、採点基準に関して分析を行い、専門知識の重要性を示した。また、評価データの性質に応じた採点基準の有効性の違いが示唆された。

6 謝辞

本研究は、戦略的イノベーション創造プログラム (SIP) 統合型ヘルスケアシステムの構築 JPJ012425 および、次世代 AI プログラム JPMJBS2407 の支援を受けたものである.

参考文献

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

[2] Sakiko Yahata, Zhen Wan, Fei Cheng, Sadao Kurohashi, Hisahiko Sato, and Ryozo Nagai. Causal tree extraction from medical case reports: A novel task for experts-like text comprehension. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 25849–25867, Suzhou, China, November 2025. Association for Computational Linguistics.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 2025.

[6] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[7] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Qwen Team. Qwen3 technical report, 2025.

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

[10] LLM-jp, :, 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, Akim Mousterou, 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. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

A　LLM 評価用テンプレート

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:
タスクの説明、症例報告、評価対象の木構造要約、模範解答の木構造要約、採点基準を以下に示します。
1.木構造要約は、症例報告の病気同士の因果関係や部位や検査結果などの修飾関係を構造化して表すものです。症例の中でも特に重要な疾患同士の因果関係をまとめて症例の大筋と呼称します。
2.以下に示す症例報告から作成された評価対象の木構造要約の品質を、一般論ではなく、採点基準に忠実に従って評価してください。
3.木構造要約の評価を行う際は、評価の根拠を示すフィードバックを書いてください。また、点数は0-100点の範囲で評価してください。
4.評価結果は以下のようなフォーマットで出力してください:“《フィードバック》(評価の根拠を示すフィードバックのテキスト)《点数》(0-100の間の整数)点。”
5.その他の説明や挨拶などは出力しないでください。

質問:
《症例報告》[title]{case_report_title}[/title]
{case_report}
《模範解答の木構造要約》
{gold_casemap}
《評価対象の木構造要約》
{hypothesis_casemap}
《採点基準》
{score_rubrics}

応答:
《フィードバック》{comment}《点数》{score}点。

図 6　LLM 評価用テンプレート。青字部分は 4.3 節の実験で変更する部分である。

B　因果木のフォーマット

因果木は木構造を成し、木構造の各ノードは症例報告中の主な病態や所見、治療法などのエンティティを表す。エンティティのテキストは、同義語辞書を用いて代表語に統一される。ノード間の関係は、病態や所見同士の因果関係や診断の証拠などのつながりを表す。この親子関係は病態・所見・治療法などの各ノードの主辞エンティティ間で結ばれるものであり、修飾エンティティは親子関係を持たない。例えば、図 1 における “急性心筋梗塞” の子 “完全閉塞” は “急性心筋梗塞” によって引き起こされた所見であり、“急性心筋梗塞” を診断するための証拠となる。

各ノードは内部にさらに構造を持ち、そのノードの主体である病態や所見、治療法等を示す主辞エンティティとそれらを修飾するエンティティから成る。各ノードの内部構造において、主辞エンティティとそれを修飾する修飾エンティティとの関係は修飾記号を用いて表現される。症例報告の因果木では、以下の 4 種類の修飾関係とそれらに対応する修飾記号を用いる。**解剖部位 (@):** 病態や所見の存在する部位を表す（例: 完全閉塞@**冠動脈**）。**極性情報 (／):** 検査結果の高値・低値、薬剤投与の有効・無効などを表す（例: SpO2 ／**低値**）。**検体名・検査名 (=):** 所見が得られた検査項目を表す（例: **心エコー**＝僧帽弁逆流）。**補足情報 (＊):** 部位の左右や病態の持つ特徴などを表す（例: **泡沫状**＊痰）。

修飾関係は「MRI ＝ DWI 高信号@右＊大脳半球」のように複数組み合わせて表される場合もある。

C　評価基準

採点対象の木構造要約は、症例報告に含まれる重要な病態や所見を抽出できていますか？また、それらの因果関係を木構造要約のフォーマットに従って正確に構造化していますか？ 木構造要約中の全ての要素は、症例に含まれていますか？
[0点~19点] 採点対象の木構造要約はフォーマットを守っていない。もしくは、採点対象の木構造要約は非常に短く、症例中のほとんど全ての要素を見落としている。医師による後編集に必要な作業量は、はじめから木構造要約を作成する場合とほとんど変わらない。
[20点~39点] 採点対象の木構造要約は症例中の一部の要素を抽出しているが、症例中の重要な病態をほとんど見落とすか誤った構造化がなされているため、読者は症例の大筋を把握できない。また、医師による大幅な後編集が必要となる。
[40点~59点] 採点対象の木構造要約は症例中のいくつかの要素を適切に構造化しているが、いくつかの重要な要素の見落としや構造の間違いが存在し、特に第一原因となる疾患を抽出できていないため、読者が症例の大筋を把握するのは難しい。もしくは、症例の大筋の一部を把握することができても、木構造要約は症例にない要素を大量に含んでおり、読者を誤解させる可能性がある。医師による、重要部分についての後編集および不要な要素の削除が必要となる。
[60点~69点] 採点対象の木構造要約は症例中の重要な要素の一部を適切に構造化しており、読者は症例の大筋の一部を把握することができる。しかし、第一原因を見落としたり、原因と結果を逆に記載するような構造の間違いが存在する。もしくは、木構造要約は重要な要素のほとんどを構造化しており、症例の大筋を概ね把握することができるが、症例にない要素をいくつか含むため、読者の理解が妨げられる可能性がある。医師による後編集および不要な要素の要素の削除が必要となる。
[70点~79点] 採点対象の木構造要約は症例中の重要な要素の大部分を適切に構造化しており、読者は症例の大筋の一部を把握することができる。しかし、いくつかの要素の見落としや構造の間違いが存在するため、医師による小規模な後編集が必要となる。
[80点~89点] 採点対象の木構造要約は症例中のほとんど全ての要素を適切に構造化しており、読者は症例の大筋を明確に把握することができるが、3つ以上の用語を拾えていない。医師による後編集はほとんど必要ない。
[90点~99点] 採点対象の木構造要約は症例中のほとんど全ての要素を適切に構造化しており、読者は症例の大筋を明確に把握することができるが、1つか2つの用語を拾えていない。医師による後編集はほとんど必要ない。
[100点] 採点対象の木構造要約は症例中の全ての要素を適切に構造化しており、読者は症例の大筋を明確に把握することができる。医師による後編集は全く必要ない。

図 7　医師によって作成された評価基準。

採点対象の木構造要約は、症例報告に含まれる重要な病態や所見を抽出できていますか？また、それらの因果関係を木構造要約のフォーマットに従って正確に構造化していますか？
[0点~20点] 採点対象の木構造要約はフォーマットを守っていない。もしくは、採点対象の木構造要約は非常に短く、症例中のほとんど全ての要素を見落としている。医師による後編集に必要な作業量は、はじめから木構造要約を作成する場合とほとんど変わらない。
[21点~40点] 木構造要約は症例中の一部の要素を抽出しているが、症例中の重要な病態をほとんど見落とすか誤った構造化がなされているため、読者は症例の大筋を把握できない。また、医師による大幅な後編集が必要となる。
[41点~60点] 木構造要約は症例中のいくつかの要素を適切に構造化しているが、いくつかの重要な要素の見落としや構造の間違いが存在するため、読者が症例の大筋を把握するのは難しい。医師による、重要部分についての後編集が必要となる。
[61点~80点] 木構造要約は症例中の重要な要素の一部を適切に構造化しており、読者は症例の大筋の一部を把握することができる。しかし、いくつかの要素の見落としや構造の間違いが存在するため、医師による多少の後編集が必要となる。
[81点~100点] 木構造要約は症例中のほとんど全ての要素を適切に構造化しており、読者は症例の大筋を明確に把握することができる。医師による後編集はほとんど必要ない。

図 8　医師の監修を受けていない評価基準。

D　ハイパーパラメータ

Hyper-parameters	Value
epoch	3
学習率	$1.00e^{-4}$
Global batch size	32

表 3　ファインチューニングのハイパーパラメータ