

# Appendix: Causal-Discovery Performance of ChatGPT in the context of Neuropathic Pain Diagnosis

**Ruibo Tu**  
KTH Royal Institute of Technology

ruibo@kth.se

**Chao Ma**  
Microsoft Research

chaoma@microsoft.com

**Cheng Zhang**  
Microsoft Research

cheng.zhang@microsoft.com

**Introduction.** ChatGPT[3] has demonstrated exceptional proficiency in natural language conversation, e.g., it can answer a wide range of questions while no previous large language models can. Thus, we would like to push its limit and explore its ability to answer causal discovery questions by using a medical benchmark [5] in causal discovery. This page serves as an appendix to this benchmark.

Causal discovery aims to uncover the underlying unknown causal relationships based purely on observational data[2]. In contrast, applying ChatGPT to answer the questions about causal relationships is fundamentally different. With the current version of ChatGPT, we can only use the names (meta information) instead of observational data of variables to answer causal questions. The answers to the causal questions given by ChatGPT are based on a trained large language model, which can be viewed as an approximation for existing knowledge in the training natural language data. Nevertheless, such investigations still provide us valuable insights into ChatGPT and raise more thoughts about how to leverage its ability. But we need to exercise great caution in the conclusion as benchmarks [4, 5] utilizing known knowledge are set for evaluation purposes instead of the goal of the causal discovery.

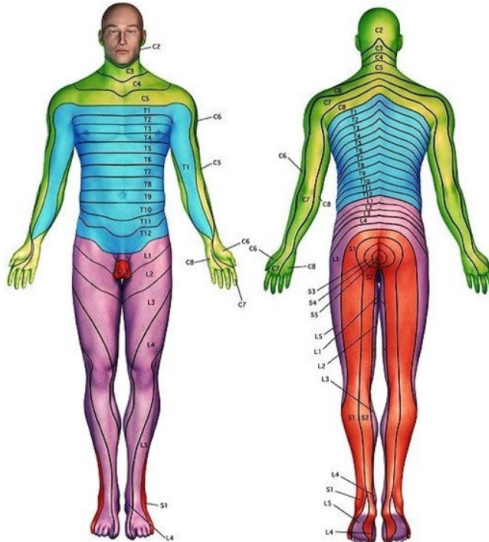


Figure 1: Dermatome map [1] as a reference for this benchmark.

Precision	Recall	F-score
1	0.12	0.2142857143

Table 1: Test results demonstrate high precision and low recall.

	Negative	Positive
Negative	50	44
Positive	0	6

Table 2: Confusion matrix showing that there were no false positives. Rows are predictions and columns are ground truth.

**Results and Insights.** The ground-truth causal relationships in neuropathic pain diagnosis are obtained from both a domain expert and known medical literature [5]. As the number of all possible cause-effect pairs in this context is huge (more than 10000 pairs), we cannot test all of them manually. Thus, we sub-sampled 50 positive pairs (ground-true causal relationships) and 50 negative pairs (wrong causal relationships) from

---

the dataset and generated the question in the format of "X causes Y. Answer true or false", where X and Y are sampled pairs from the full causal map of the neuropathic pain dataset. The full test results can be found at [shorturl.at/amBX1](https://shorturl.at/amBX1). Many individual answers are reasonable, such as in Figure 2, but the performance is still flawed systematically. As shown in Table 1 and 2, ChatGPT tends to make false negative mistakes. We inspected the results qualitatively and quantitatively and observed that:

It only understands the languages that are typically used to describe the situations but not the underlying knowledge. We provide two examples to demonstrate it. The first example is shown in Figure 3. It cannot identify the lower abdominal discomfort that can be caused by T12 radiculopathy. The explanation identifies the lower back, hip, and leg region only, while T12 nerve goes through these regions shown in Figure 1, and the lower abdominal region is part of it. Thus, it indicates that it provides the answer based on the trained content but does not understand the human body's nervous system. The second example is shown in Figure 4, which demonstrates a lack of understanding of how regional discomfort is described. The region around the key bone is the upper shoulder region. ChatGPT can identify shoulder discomfort as an effect but not the discomfort around the key bone.

Its performance is not consistent and not stable. Firstly, we observe that it provides different answers to the same question. We have tested some of the queries twice on different days. As shown in Table 3, the answers on the first day differ from the ones on the second day significantly. The answers on the second day are much more conservative to claim a causal relationship. This may be due to internal model updates. Such inconsistent performance is a major concern for answering causal questions. As the later results have very few positive answers, the final results that we used considered the earlier results when available for the table 1 and 2. Secondly, as the original dataset is associated with terms in Swedish, we found that ChatGPT can correctly identify Swedish in some cases, such as in Figure 2, but fails in some other cases, such as in Figure 5. This may contribute further to a large number of false negatives.

**Conclusion.** Based on the observations, we conclude:

- There is a clear limitation for the current ChatGPT in terms of understanding new concepts and knowledge beyond the existing corpus of text training data. Moreover, the consistency and stability of its performance need to be improved.
- We need to be extremely cautious about using causal claims made by ChatGPT as causal discovery results. This is because causal discovery and causal question answering with large language models are fundamentally different tasks. Causal benchmarks may be biased towards utilizing existing knowledge for evaluation [5, 4], which is against the goal of causal discovery.
- In some situations, ChatGPT does give correct answers that can be non-trivial to obtain from a domain expert, which could serve as a good complementary for causal discovery methods to resolve corner cases. This might open up new research opportunities for the causal community on utilizing the recent developments of large language models to complement, improve and develop better causal machine learning tools.



L L5 Radikulopati causes L Benbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.



True. L L5 radiculopathy refers to a condition where there is irritation or compression of the L5 nerve root in the spinal cord, which can cause symptoms such as pain, numbness, weakness and/or tingling in the lower back, hips, legs and/or feet on the same side of the body. This can lead to leg pain (benbesvär in Swedish) as well as other symptoms.



Figure 2: Example showing that ChatGPT can correctly answer the question and provide reasonable explanations.



L T12 radiculopathy causes L abdominal discomfort."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.




False. L T12 radiculopathy refers to a condition where there is irritation or compression of the L T12 nerve root in the spinal cord, which can cause symptoms such as pain, numbness, and weakness in the lower back, hips, and legs on the same side of the body. Abdominal discomfort is not typically a symptom of L T12 radiculopathy.




Figure 3: The lower abdominal is the region where T12 nerve passes. If looking at the dermatome map 1, it is easy to identify lower back, hip, and lower abdominal discomfort can all be caused by T12 radiculopathy.

---




L C4 radiculopathy causes L keybone region discomfort. "R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.




False. L C4 radiculopathy refers to a condition where there is irritation or compression of the L C4 nerve root in the spinal cord, which can cause symptoms such as pain, numbness, and weakness in the neck, shoulder and arm on the same side of the body. Keybone region discomfort is not typically a symptom of L C4 radiculopathy.

---




L C4 radiculopathy causes L front shoulder discomfort. "R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.




True. L C4 radiculopathy refers to a condition where there is irritation or compression of the L C4 nerve root in the spinal cord, which can cause symptoms such as pain, numbness, and weakness in the neck, shoulder and arm on the same side of the body. Shoulder discomfort, including front shoulder discomfort, is a common symptom of L C4 radiculopathy. Depending on the severity of the condition, it may also be accompanied by other symptoms such as weakness, tingling or pain in the arm, hand, or fingers.

Figure 4: Example showing that ChatGPT fails to understand the region on the body. The area around the key bone is largely overlapping with the front shoulder area especially when the patient describes the symptoms.

---



L T12 Radikulopati causes L Nedre bukbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.



False. L T12 radiculopathy refers to a condition where there is irritation or compression of the L T12 nerve root in the spinal cord, which can cause symptoms such as pain, numbness, and weakness in the lower back, hips, and legs on the same side of the body. "L Nedre bukbesvär" is not mentioned in the sentence and thus it's not clear what it refers to.

Figure 5: Example showing that ChatGPT can identify foreign language time by time and that it is not very reliable.

Query	GroundTruth from Ex- pert	Jan18 Test	Jan19 Test
L L4 Radikulopati causes R Höftkamsbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L T12 Radikulopati causes L Nedre bukbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
R L1 Radikulopati causes L Ljumsksbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
R C6 Radikulopati causes L Laterala armsbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	1	0
R C6 Radikulopati causes R Armbågsbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	1	0
R L1 Radikulopati causes IBS."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L C5 Radikulopati causes Nackbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	1	0
R C5 Radikulopati causes R Laterala armbågsbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L C6 Radikulopati causes R Handbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	1	0
L L4 Radikulopati causes L Laterala vadbcsvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
R S1 Radikulopati causes R Lårbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	1	0
L T5 Radikulopati causes L Bröstbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L C5 Radikulopati causes R Interskapulära besvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
R C6 Radikulopati causes R Under armsbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L L1 Radikulopati causes L Mediala ljumsksbesvär."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L L1 Radikulopati causes L Adduktortendalgi."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
L T10 Radikulopati causes IBS."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0
R L5 Radikulopati causes L Bakhuvudvärk."R" and "L" refer to the right and left sides of the body, respectively).Answer with true or false.	TRUE	0	0

Table 3: Results demonstrate lack of consistency using ChatGPT.

---

## References

- [1] Dermatome map. <https://i.pinimg.com/736x/ef/76/47/ef7647ceae98d10588f14b4ecd7e6a89.jpg>.
- [2] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
- [3] OpenAI. Chatgpt. <https://chat.openai.com/chat/>.
- [4] A. Sharma. Chatgpt causality pairs. <https://github.com/amit-sharma/chatgpt-causality-pairs>.
- [5] R. Tu, K. Zhang, B. Bertilson, H. Kjellstrom, and C. Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32, 2019.