

SEI Report

# Applying Causal Learning to Evaluate Large Language Models (LLMs)

March 2026

Mike Konrad  
Andrew Mellinger  
Linda Parker Gates  
David Shepard  
Nick Testa

---

DOI: [10.1184/R1/30251989](https://doi.org/10.1184/R1/30251989)

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

<https://www.sei.cmu.edu>

Copyright 2026 Carnegie Mellon University.

This material is based upon work supported by the Department of War under Air Force Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The opinions, findings, conclusions, and/or recommendations contained in this material are those of the author(s) and should not be construed as an official US Government position, policy, or decision, unless designated by other documentation.

References herein to any specific entity, product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Requests for permission for non-licensed uses should be directed to the Software Engineering Institute at [permission@sei.cmu.edu](mailto:permission@sei.cmu.edu).

DM25-1427

## Acknowledgments

We would like to acknowledge Julie Cohen, Crisanne Nolan, and Melissa Ludwick on the project team; Joe Ramsey, our algorithms expert; Sandy Shrum, our technical writer; and our management, Bill Nichols and Eileen Wrubel, whom we thank for their continuing support!

## Abstract

As the SEI's body of causal work has evolved into an end-to-end causal discovery and inference method and tool suitable for detecting bias in ML and AI models, SEI researchers are beginning to investigate whether the first step of the method, causal discovery, can also be applied to LLMs. The SEI's approach to exploring this question comprises three steps: (1) obtain a dataset of story/summary pairs to use as ground-truth, (2) design prompt styles (e.g., purpose, tone) with which to prompt a Summarizer LLM to summarize a story from one of those pairs, and (3) design a set of summarization-quality features employed by an Evaluator LLM to score the quality of summaries generated by the Summarizer LLM. In this way, SEI researchers created a dataset of higher level features for input to causal discovery. The resulting causal graph demonstrates that a causal relationship between the focus of a prompt style and summary quality is often discoverable when both features overlap. This overall approach may benefit software engineering and LLM research by providing a more formal methodology for assessing the nuanced cause-and-effect relationships unique to a given LLM, reducing confounding.

# 1 Introduction

The purpose of this report is to share the SEI’s Transition of AI Limits of Robustness (TAILOR) project team’s approach to, and results from, applying causal learning to the evaluation of large language models (LLMs). This report contains a proof of concept that causal discovery (e.g., as realized in the AI Robustness (AIR) Tool developed by the TAILOR project) can be applied to rich data from LLMs efficiently enough to feasibly investigate research questions.

Our driving question for this work is whether causal learning can be used to evaluate LLMs. Of course, LLMs are used daily to perform a large variety of tasks. So, to make the question more concrete, we focused on a common task that LLMs are often prompted to perform: text summarization. Today, accurate text summarization is very important to many organizations (e.g., the Intelligence Community) [Gallagher 2024].

Evaluating an LLM on something like text summarization requires many documents that contain text to be summarized as well as a good proxy for ground truth (i.e., summaries written showing the desired properties of those documents). In our search for a suitable collection of documents, we came across a large news-story dataset, which included professionally written highlights (a type of summary<sup>1</sup>) of those stories, that would serve our purpose.

Using this news-story dataset, we could systematically vary prompts given to an LLM to create story summaries and compare the results against the professionally written highlights to discover the causal relationship between different prompting strategies and the quality of summarization. Below is an example of a text-summarization prompt with the text to be summarized embedded in the prompt. Note the instruction, “Explain the above in one sentence.” This is the kind of prompt that we chose to systematically vary.

The following is an example prompt and its output from the *Prompt Engineering Guide* [Saravia 2022].

Prompt

*Antibiotics are a type of medication used to treat bacterial infections. They work by either killing the bacteria or preventing them from reproducing, allowing the body’s*

---

<sup>1</sup> One Merriam-Webster definition of “highlights” is “a summary of the most significant or interesting parts of something (such as a completed sporting event) presented typically on a television broadcast.”

*immune system to fight off the infection. Antibiotics are usually taken orally in the form of pills, capsules, or liquid solutions, or sometimes administered intravenously. They are not effective against viral infections, and using them inappropriately can lead to antibiotic resistance.*

*Explain the above in one sentence:*

Output

*Antibiotics are medications used to treat bacterial infections by either killing the bacteria or stopping them from reproducing, but they are not effective against viruses and overuse can lead to antibiotic resistance.*

Creating effective prompts for LLMs is not well understood because LLMs are not well understood. LLMs are an active area of research within the SEI and elsewhere. Writing prompts remains mostly a “trial-and-error” process and is only slowly emerging into craftsmanship. It has a long way to go to reach any useful level of engineering discipline.

The particular effort-efficient data engineering approach we employed in this research—where features critical to evaluating the quality of summaries output by prompting an LLM are first defined, and then an LLM is also used to generate these features from the original news story dataset—is described below because such an approach can be readily applied more broadly to investigate the causality of other LLM tasks.

## 1.1 Overall Technical Approach

Our original intended approach was to use an existing dataset that contained a significant corpus of text summarization examples, including the prompt styles employed and summarization evaluation metrics. Such a dataset would allow us to directly evaluate the impact of different prompting strategies on the quality of LLM-generated text summaries. Unfortunately, we could not find such a dataset in the research literature. As a result, we elected to construct such a dataset ourselves from an existing corpus of news stories and professionally written highlights and employ an LLM to help generate the features we wanted to input to causal discovery.

The particular corpus we used in our approach is the CNN-DailyMail Dataset, which is “an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering”

[Hugging Face 2025, Penugonda 2025]. This dataset is a commonly referenced dataset for summarization benchmarks and therefore implies applicability and utility to the general benchmarking community.

From our review of guidance about what types of prompts might be effective (see the Organized and Annotated Bibliography), we created a set of prompt variations we wished to explore. Then, we used Python to generate the full set of prompt combinations. We submitted these prompts with articles from the CNN dataset to an LLM to perform text summarization. Next, we provided the resulting summaries and a detailed evaluation prompt to an Evaluator LLM to generate evaluation scores. We combined these scores with the input prompt configuration to form the core of our dataset. We used the same Python script to automate the workflow end to end.

Using an LLM instead of a human to generate summaries<sup>2</sup> and an LLM instead of a human to label the quality of the output text summaries allowed us to investigate LLM behavior questions rapidly and inexpensively. However, this approach required careful planning, as we would use the same LLM for both text summarization and evaluation of the resulting text summaries (i.e., we'd be asking "the student to grade their own work"). Our approach would also need to reinitiate the LLM before each task.

We gained some confidence in this approach by applying it to a small sample of news articles, generating a few prompts, and reviewing the resulting evaluation scores. The different quality ratings we obtained seemed plausible. Perhaps more telling was that the causal relationships discovered between the prompt style and summarization rating seemed reasonable. For example, when the focuses of a prompt and summarization rating category overlapped, there was often a direct or indirect causal relationship. These causal relationships suggested that the approach had indeed worked; i.e., the LLM had acted on the instructions as intended and had not excessively hallucinated. While some minor adjustment to the prompt wording and evaluation scoring were required after our initial test run (to help ensure some spread in the resulting evaluation scores across prompts), we became more convinced that our technical approach was sound.

---

<sup>2</sup> In general, for a set of common use cases where reliability is important, humans will no longer write their own prompts. Instead, the organization will develop and provide staff with specific highly performant and reusable prompts. What we (and others at the SEI) are trying to do is support this approach with science to make crafting these prompts more of an engineering task and less of an act of "black magic."

This technical approach should make it possible to answer other specific questions about how particular LLMs behave without huge datasets, immense human effort (in prompting LLMs or labeling text summaries for quality), computing in the cloud, or large-scale computing in general (an ordinary laptop suffices).

Causal discovery presents a benefit in all of this. It can help accurately answer the research question, “How do certain prompting styles affect the quality of news story summarization?” by identifying and helping account for the sources of confounding bias in the data (especially if those sources are represented in the data). It also can be the basis for obtaining more accurate estimates of causal effect that certain LLM-prompting strategies have on the quality of text summarization.

## 1.2 Research Questions

Our research team was repeatedly asked by several potential transition collaborators whether the project’s AI Robustness (AIR) Tool could be used to understand LLMs. In principle, the answer is “yes,” but how would we go about using it this way? The first step of the AIR Tool is causal discovery, where the AIR Tool employs Best Order Score Search (BOSS), a superior algorithm for causal discovery but which handles only about a thousand variables efficiently. However, the vast number of word choices in and combinatorics of text imply the need for a much larger number of variables, unless we focus on higher level features of the text.

Because LLMs are often requested to perform text summarization, we chose to explore how well an LLM could perform that particular task through the lens of higher level features rather than individual words. So, with that in mind, we sought to investigate these two research questions:

1. Can causal discovery bring actionable insight into how LLMs behave, particularly with regard to specific features of text summarization quality?
2. Which prompt type or configuration from a set of possible prompt styles (features) drives a given LLM to produce a higher quality summary of a news story?

The second research question was tailored from the first one to address a specific LLM task: text summarization of news stories. The following are examples of the prompting strategies that we attempted to evaluate:

- *26 Prompting Principles for LLMs* [Parti 2024]
- *Best Prompt Techniques for Best LLM Responses* [Damji 2024]

## 1.3 Foundation of This Research

Adjacent to our ongoing research work in causal learning, this work provided an opportunity to extend our research into how causal learning can be used to evaluate the use of LLMs.

### 1.3.1 DoW Need and Problem Relevance

The Department of War's (DoW's) need for evaluating LLMs is similar to the DoW's need for evaluating artificial intelligence (AI) and machine learning (ML) models, but the technical challenge is quite different, as we elaborate below:

1. First, there is a common promise and need. AI and ML models, including LLMs, can automate and/or improve the quality of various analyses, thereby increasing the speed of making accurate decisions and plans, which improves the performance and outcomes of missions and operations. Due to this promise, AI and ML models (and LLMs) are being urgently evaluated and adopted within the DoW.
2. While the promise of and need for LLMs is generally similar to that of AI and ML models, text generation is fundamentally a different type of task. AI and ML models output classifications and predictions, which, as noted above, are what the AIR Tool was designed to evaluate. Creating a text summary is not typically recognized as a classification or prediction problem. So, it is not immediately clear how the capability embodied in the AIR tool, which is focused on detecting and mitigating confounding bias in AI and ML classifier/prediction models, can be extended to evaluating LLM outputs.
3. Thus, a key technical challenge to address in this research is how to transform the text-summarization problem into a text summarization quality evaluation (classification) exercise, as that is what the AIR Tool was designed to handle.

### 1.3.2 Causation Vs. Correlation and the AIR Tool

AI and ML models learn statistical relationships, such as correlations. However, correlation does not imply causation, so acting on a particular mission or operation based on correlations (an AI or ML model's predictions or classifications), and not on cause-and-effect relationships, might (and generally will) have consequences different from (or even the opposite of) those predicted and anticipated.

Broadly speaking, causal learning, which consists of causal discovery and causal inference,<sup>3</sup> can help identify confounding variables and eliminate/mitigate their effects from downstream analytic activities that guide key mission and operational decisions. In the case of the AIR Tool, those downstream analytic effects include producing unbiased estimates of the total direct effect of a “treatment” variable on a “response variable” so that the effects of taking a particular action, called an “intervention,” on the treatment variable can be more accurately predicted than by (correlation-based and not causation-based) AI or ML models.

In the case of LLMs, we bring a narrower and simpler focus: What prompt styles (the treatment) drive LLMs to provide a higher-quality summary of a news story (the response).

While we could have fully used the AIR Tool in this LLM research, due to time limitations, we focused primarily on applying causal discovery to better understand how well LLMs responded to particular types of prompts.

## 1.4 Detailed Technical Approach

The question we were trying to answer was whether we could use causal discovery to determine which prompt characteristics (e.g., tone, incentive, using an example) have the most impact on summary quality.

Using the dataset discussed earlier, our experimental process was to generate a full set of prompts with various options that reflected the desired prompt characteristics we found in our internet search. For example, we could request the prompt tone to be “colloquial” or “formal.” These prompts were then passed to a *Summarizer LLM* to generate summaries of each prompt variation and news story. The resulting summaries (and their associated news stories) were then passed to an *Evaluator LLM*<sup>4</sup> that was separately prompted with instructions on how to evaluate each summary against the associated news story on quality factors of interest.

Summary evaluation is an ongoing research area, especially regarding semantic quality. There are a variety of traditional natural language processing (NLP) evaluation approaches that use n-gram occurrences and longest common subsequences (LCSs) that we could use as preliminary measures, but these don’t provide much feedback on advanced language metrics. For semantic metrics, we

---

<sup>4</sup> LLM-as-a-judge is an emerging research area [Gu 2024].

used one of the emerging LLM-based approaches called G-Eval [Liu 2023] with inspiration from Generative AI Vertex AI templates [Vertex AI 2025] (for the Evaluator LLM) to measure advanced language concepts, such as groundedness, coherency, fluency, and verbosity. We also measured the prompts with ROUGE-1 and ROUGE-L, standard NLP evaluations, for reference. (See Traditional NLP Metrics on page 9 for more information.)

The final dataset comprised all of the three types of data described above:

- Summarizer LLM prompt-configuration options
- Evaluator LLM’s evaluations of summary quality
- ROUGE metrics

The final dataset was fed into a causal discovery algorithm that recovered (an “estimate” of) the underlying causal graph that related these measures to each other and allowed us to better understand which prompt variations impacted which summary quality factors.

#### 1.4.1 LLM Execution Platform

We selected the Ollama tool to interact with and manage the LLMs, generate the summaries, and perform the evaluations [Ollama 2025]. Ollama is a locally installed tool that can run local models, which allows private local-data management, direct control of the models we use, the ability to execute them on SEI-controlled hardware, and not having to pay per-query costs to vendors. The downside of this platform was that we were limited to publicly available downloadable models and the limits of our hardware. However, the quality and quantity of Ollama’s downloadable models is impressive, allowing for future expansion. For this proof of concept demonstration, these limitations were acceptable.

#### 1.4.2 Dataset Selection

For our reference news-story dataset, we used the *CNN-DailyMail* dataset, which contains over 300,000 articles with human-generated highlights [Penugonda 2025]. This dataset is a popular reference dataset in evaluation metrics, such as the Holistic Evaluation of Language Models (HELM) suite [CRFM 2025].

While this dataset provides “highlights” rather than summaries, we elected to proceed, treating these highlights as summaries, as we did not have time nor easy access to other sources of text and summary pairs (a limitation of our research approach).

### 1.4.3 LLM Selection

We used two LLMs for this experiment: one for generating the summaries and one for evaluating the summaries. Using these two LLMs enabled us to select purpose-specific models and to compare them to one another. Ollama provides access to a wide variety of public models. For our experiments, we chose a highly capable multi-modal single-GPU model from Google DeepMind called Gemma3 [Google 2025]. We chose the four billion parameters configuration, which allows the quick iteration of both summary generation and evaluation. While we considered choosing different models for each purpose, for the sake of expediency, we chose to use the same model for both purposes. (As we will discuss later, this decision is a threat to validity, so we should have employed a different LLM for each purpose.)

### 1.4.4 A Taxonomy of Prompt Types

To generate the summarization dataset, we started with a summarization prompt template that contained basic instructions and added prompt variations to account for the tone of the summary, the writing complexity of the summary, and the use case or “purpose” of the summary. For each of those variations, we provided three or four options for a total of 36 combinations. The goal of the experiment was to evaluate which prompting strategies had a direct impact on summary quality.

The following are the prompt variations we added. (See Appendix A for details.)

1. **Purpose.** This variation specifies the need, use, and incentives behind generating the summary, such as whether it is to be used for competition, whether the LLM Summarizer is to be rewarded, etc.
2. **Tone.** This variation specifies the formality and language to be used for the summary, ranging from “appeal to a younger generation” to “use a formal tone and avoid conversational language.”
3. **Complexity.** This variation specifies the complexity of the vocabulary and terminology to be used, ranging from “sixth grade reading level” to “subject-matter-expert level.”

### 1.4.5 Evaluation Prompt Factors and Rubrics

We used a G-Eval-like approach, where we submitted detailed prompt instructions, measures, and rubrics to the Evaluator LLM. We provide the entire evaluation prompt in Appendix A; the rubrics we used are summarized below:

1. **Groundedness.** This rubric is measured using one of two values: 0 or 1. The value 1 represents that “All aspects of the response are attributable to the context.”
2. **Coherence.** This rubric is measured along a 1-5 value scale. The value 5 represents that “The writing has a seamless logical flow, is expertly organized, and maintains exceptional cohesion throughout.”
3. **Fluency.** This rubric is measured along a 1-5 value scale. The value 5 represents that “The response is free of grammatical errors, demonstrates nuanced word choice, and has a natural, seamless flow.”
4. **Verbosity.** This rubric is measured along a -2 to +2 scale. The value 0 represents that “The response is perfectly concise, providing all necessary information in a clear and succinct manner without any unnecessary wordiness or repetition.”

We provided evaluation steps as part of the prompt and requested the Evaluator LLM to provide its output in a JavaScript Object Notation (JSON) block that we could easily extract in later processing.

### 1.4.6 Traditional NLP Metrics

We used the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) set of NLP metrics to determine the similarity between bodies of text based on characters and words. ROUGE uses NLP principles such as continuous sets of characters or words called *n-grams* and other matching characteristics. We used ROUGE-1, which counts unigram (single word or character) matches between bodies of text, and ROUGE-L, which is based on the Longest Common Sequence (LCS) found in compared bodies of text. These metrics focus on specific words and characters and do not account for semantic matching, such as between synonyms. In modern LLM-based summary generation, we often try to generate summaries for different target audiences using different levels of sophistication, which makes the ROUGE suite of metrics (and other traditional NLP metrics) less useful due to their literal character-and-word matching techniques.

## 1.4.7 Dataset Preparation for Causal Discovery

The output of our “generate prompt/evaluate summary” experiment would shed light on which types of prompt styles are causal drivers of which text-summarization quality factors. To prepare the experiment for applying causal discovery (AIR Tool Step 1), we evaluated whether any pair of variables in the dataset were highly correlated with each other—such correlations can mask causal relationships with other variables and thus should be addressed<sup>5</sup> prior to applying causal discovery. We found one excessively high correlation between ROUGE-1 precision and ROUGE-L precision, so we omitted ROUGE-1 precision from the dataset.

LLMs are, by construction, non-deterministic unless the *temperature* is set to 0.0. Therefore, throughout our use of LLMs in this experiment, we set the temperature to 0.0 so that the LLMs behaved deterministically, allowing greater predictability and control during the experiment.

## 1.4.8 Causal Discovery

Once the dataset was prepared (as described above), we searched it with the BOSS search algorithm using this approach:

1. We used Tetrad toolbox version 7.6.9 for causal discovery [Ramsey 2018].
2. We loaded the dataset into a Tetrad Data Box.
3. We configured a Tetrad Knowledge Box to constrain search so that search would not expend effort considering causally impossible relationships, such as where one of the four text-summarization evaluation measures or three ROUGE measures cause one of the three prompt types we were manipulating: complexity, purpose, and tone.
4. We then connected both the Data Box and Knowledge Box to the Tetrad Search Box.
5. We configured the Search Box to use the BOSS Search algorithm [Andrews 2023].

---

<sup>5</sup> Prior to applying causal discovery to any dataset, it is important to replace every highly intercorrelated cluster of variables with only one of those variables (our chosen alternative) or with the principal component of the cluster (e.g., by applying principal component analysis over the cluster).

6. We set all BOSS algorithm hyper-parameters to their default values, except for the following:
  - We lowered Penalty Discount (PD) to 1.0 from the default of 2.0, given that there were 36 possible configurations for the prompt to the Summarizer LLM and each setting is its own independent search of over 50 news stories per prompt configuration. Thus, there is a limited size of dataset per prompt configuration.
  - We set the number of restarts, random after the first (default 1), to 20. This means that search is initiated from 20 different randomized starting points, and the best-scoring causal graph is kept, which helps ensure a superior model fit with the dataset.
7. We then executed the BOSS search, which produced a causal graph that we provide and discuss in the Results section.

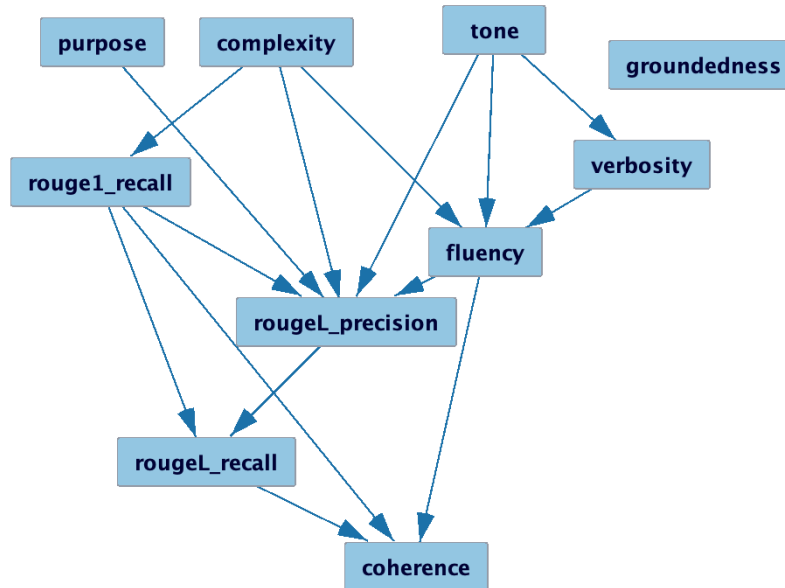
## 1.5 Data and Python Script

The code we used for automating the experiment consisted of three scripts that represented the various *phases* of the dataset's generation and some supporting maintenance tools. The following describe how these three primary scripts work:

- **make\_prompts.py.** This script takes a summarization prompt template and a set of options for summarization guidelines as input. The script generates summarization prompts based on all the combinations of the summarization guidelines provided. (See the descriptions above about Tone, Complexity and Purpose.)
- **make\_summaries.py.** This script takes the prompt list and the source dataset as input and generates summaries for a sample of articles from the dataset and outputs a JSON file of summaries. The script can be configured with a variety of parameters (e.g., the model to use for summarization, the number of articles to summarize, the temperature level of the summarization, and how many summary repetitions to perform).
- **make\_evaluations.py.** This script takes the summaries, the input dataset, and an evaluation prompt as input and produces evaluations of each summary into a csv file. The evaluations are provided by a configurable Evaluator LLM using the evaluation prompt and ROUGE-1 and ROUGE-L metrics. The resulting csv has rows that contain the three prompt-configuration values, the four outputs from the Evaluator LLM, and three metrics from the ROUGE evaluations.

## 1.6 Results

Figure 1 displays the results of applying the BOSS search algorithm to the dataset.



The number of random starts was set at 20, and the Penalty Discount = 1.0.

---

Figure 1: Causal Graph Resulting from Applying the BOSS Algorithm

The three prompt-configuration variables are Purpose, Complexity, and Tone.<sup>6</sup> Each variable influences *one or more* of the seven summarization quality variables depicted elsewhere in the figure.

Except for Groundedness, each summarization quality variable has one or more of the three prompt-configuration variables as a causal parent or causal grandparent. Clearly, there is strong evidence that changing the prompt type will result in a change to one or more of the text summarization variables.

---

<sup>6</sup> In Figure 1, all variables are shown in lowercase, but in this description, we capitalize the names of the variables to differentiate them from the ordinary English words.

What about Groundedness? It has no edges with any other variable, which means that it neither causes (influences) nor is caused (is influenced) by any other variable. Table 1 shows why that is the case.

### 1.6.1 Groundedness

Table 1: Basic Statistics for Groundedness

Measure	Value	Counts per value				
Mean	1	0	1	x	x	x
Median	1	0	1800			
Stdev	0.0					

Table 1 reveals that Groundedness is a constant variable. From a causal perspective, constant variables can't influence anything else in the dataset because causal discovery analyzes the dataset to see how a shift in a variable's value affects the values of other variables, and constant variables don't vary. (Another way to understand why a constant variable can't influence anything else is that constant variables have 0.0 correlation with every other variable in the dataset.)

Having considered Groundedness, let's consider, in turn, the other three Evaluator LLM text summarization variables.

### 1.6.2 Coherence

Table 2: Basic Statistics for Coherence

Measure	Value	Counts per value				
Mean	4.14	1	2	3	4	5
Median	4	0	0	19	1506	275
Stdev	0.38					

As shown in Table 2, a Coherence rating of 4 is assigned about five out of six times. It is the most common rating. Most of the remaining ratings are 5, and there are a few 3s.

From a causal perspective, we can consult Figure 1 to determine what changes would need to be made to the prompt (i.e., the prompt to the Summarizer LLM) to affect the Coherence rating of the summarized text assigned by the Evaluator LLM. While Coherence has no direct causal parent among the three prompt variables, all three prompt variables are causal grandparents (or causal great-grandparents) of Coherence. A change to any of the three prompt variables can affect Coherence,

which seems reasonable. However, as noted, these are not direct causal relationships but are mediated with other variables to reflect text summarization quality. Due to the limited time available, we did not validate whether changing any of the three prompt variables actually impacts the resulting Coherence rating, but this should be relatively easy to do.

### 1.6.3 Fluency

Table 3: Basic Statistics for Fluency

Measure	Value	Counts per value				
Mean	4.17	1	2	3	4	5
Median	4	0	2	28	1426	344
Stdev	0.43					

For Fluency, we have only two of the three prompt variables as causal ancestors (i.e., not Purpose); however, the two that are causal ancestors (i.e., Complexity and Tone) are both causal parents and have a direct causal relationship with Fluency. This makes sense; Complexity should affect Fluency and so should Tone, which can dampen/moderate the language generated during Summarization. However, it's less clear how Purpose would significantly affect Fluency.

To improve Fluency, change the prompt to allow greater complexity in the summary or to adopt a more appropriate tone.

### 1.6.4 Verbosity

Table 4: Basic Statistics for Verbosity

Measure	Value	Counts per value				
Mean	0.025	-2	-1	0	+1	+2
Median	0/0	0	2	1752	46	0
Stdev	0.162					

For Verbosity, we have only a single variable as a parent (i.e., Tone) and a single variable as a child (i.e., Fluency). Both causal relationships make sense; a change to indicate what tone to adopt in a text summary would likely affect the Verbosity of the summary, which in turn would likely affect Fluency.

To improve Verbosity, change the prompt to require the Summarizer LLM to adopt the right tone.

## 1.7 Lessons Learned and Observations

The following are the lessons we learned during our experiment.

### 1.7.1 LLM Instruction Following

One of the key measures of LLMs is how well they follow instructions in the prompt, and this measure is generally referred to as “instruction following.” When using the Evaluator LLM, we found an occurrence where the LLM failed to follow instructions, and the results were subtle yet profound.

The Groundedness score was specified to have two possible values: 0 or 1. A given summary is either grounded (all the points in the summary trace back to the original article) or it isn't. However, when we first started using the evaluator, it provided a 5 for the Groundedness value. We then had a conversation with the LLM about why it did this and how our prompt could be improved. This conversation led us to design a better prompt, which we used in the experiment, but the conversation itself was enlightening. We should note, though, that we do not actually know whether the particular LLM we used had any insight into its training, or, if it did, whether it could offer sound advice about how to improve our prompts to it. We provide our exchange with the LLM in Appendix B.

### 1.7.2 Using LLMs for Evaluation

Using LLMs for evaluation and as a “Judge” of models has become very popular in research and industry [Gu 2024]. However, all the research we uncovered suffers from the same issues as developing summarization metrics in general: We don't understand how the models work, and we can't evaluate them.

The results we get from an LLM are more “convincing” because of how they are presented, but they still lack the validity that most real-world applications require. Basically, if we can't trust the LLMs to do a good job, then how is it reasonable to trust them to evaluate the work? This issue of whether to trust an LLM to evaluate its own work is a separate research question that could be pursued in a follow-up project.

### 1.7.3 LLMs and Existing NLP Metrics

The current sophistication level of LLMs allows them to easily provide excellent grammatical output so that they all do well at standard NLP and structure metrics. They are all fluent and coherent (even when hallucinating or confabulating), and

they can be very concise. A current focus for some of us at the SEI is how to discover appropriate task-related semantic measures, which will require significant non-LLM metrics.

## 1.8 Limitations of this Work and Next Steps

We used one source of text summary examples (i.e., CNN-DailyMail news stories), which limits the generalizability of the conclusions we can draw regarding the quality of LLM text summarizations, although we think that some of them are generalizable. However, news stories typically have an inverted pyramid structure, with the important points coming first and details (e.g., context, implications) coming later. Using only news stories is a limitation in the external validity of this research.

We employed news story “highlights” rather than summaries per se as ground truth, perhaps limiting the “soundness” of what we were doing. This is a limitation in the internal validity of this research.

We experimented with only a few prompt styles, which omitted many alternatives. A related issue is the quality of the LLM prompt. Our technical writer noted: “In future research about LLM prompts, consider having an editor review the prompts. Sometimes a misplaced comma can completely change the meaning of a sentence.”

We experimented with four types of summarization quality factors and four types of syntax/structure metrics, omitting other ways of evaluating/rating the quality of a text summary. In particular, we did not conduct human evaluations of text summary quality beyond the occasional and unsystematic “spot check.”

We used only one LLM model family and only one size (in terms of the billions of model parameters) for summarization and evaluation. While sufficient for a proof of concept, we expect variations based on different models and sizes.

## 1.9 Conclusions

In this work, we demonstrated how researchers might use LLMs for NLP-type applications (e.g., text summarization) and derive high-level feature datasets that can be investigated causally. Clearly, there are limits to our particular approach; however, those who use it can get a rapid “read” on the plausibility of a research question and some idea of how to pursue it more rigorously.

We demonstrated how to conduct this research frugally. Nevertheless, judging by our ability to obtain a non-trivial causal graph, we assert that this is a successful attempt to evaluate a particular LLM.

## 2 Organized and Annotated Bibliography

Early in this research, we conducted a limited literature review with the excellent help of SEI Research Services. The original focus of our literature review was fairly broad, but as we skimmed this literature and our experimental approach further crystalized, the focus became much more targeted. In this section, we provide the topics we investigated at some depth. We provide only a sample of the references originally included in the literature review. We selected them based on their relevance to this particular research topic and our approach to it:

- If LLMs Understand Causality
- Motivation for Our Approach
- Technical Basis for Our Approach
- Limitations to Our Approach

Each of these topics is expanded below with related citations, focusing generally on the sources found to be most relevant.

### 2.1 If LLMs Understand Causality

#### [Ashwani 2024]

Ashwani, S. et al. Cause and Effect: Can Large Language Models Truly Understand Causality? Pages 2-9. In *Proceedings of the AAAI Symposium Series*. Volume 4. Issue 1. November 2024. DOI: 10.1609/aaaiss.v4i1.31764 <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31764>

#### [Jin 2023]

Jin, Z. et al. Can Large Language Models Infer Causation from Correlation? *arXiv* [preprint]. June 2023. <https://arxiv.org/abs/2306.05836>

*[Even after LLMs were fine-tuned to take a collection of “correlational statements and determine the causal relationship between the variables,” the LLMs could only “perform causal inference in in-distribution settings when variable names and textual expressions used in the queries are similar to those in the training set but fail in out-of-distribution settings generated by perturbing these queries.”]*

#### [Zhang 2023]

Zhang, C et al. Understanding Causality with Large Language Models: Feasibility and Opportunities. *arXiv* [preprint]. April 2023. <https://arxiv.org/abs/2304.05524>

*[A conclusion of this work, simplified, is that what LLMs can state are causal relationships explicitly present in their training data, but inference from implicit demonstrations of causality or “emergence of awareness” of causal relationships from Training data is*

*uneven (e.g., inferring rules about 5-digit by 5-digit multiplication) and difficult to predict in advance.]*

## 2.2 Motivation for Our Approach

### **[Feder 2021]**

Feder, A et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. Transactions of the Association for Computational Linguistics. *arXiv* [preprint]. September 2021. <https://arxiv.org/abs/2109.00725>

### **[Gallagher 2024]**

Gallagher, S. et al. Assessing LLMs for High Stakes Applications. Pages 103-105. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '24)*. May 2024. <https://doi.org/10.1145/3639477.3639720>

### **[Zhang 2025]**

Zhang, C. et al. Causal Prompting: Debiasing Large Language Model Prompting Based on Front-Door Adjustment. Pages 25842-25850. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 39. Issue 24. DOI: 10.1609/aaai.v39i24.34777. <https://arxiv.org/abs/2403.02738>

*[This paper is a recent publication applying a different Combat Development and Integration (CD&I) approach to the problem we set out to solve: Causal Inference through Front-Door Adjustment. However, that approach is limited to circumstances where Chain-of-Thought (CoT) text is available and can be exploited, which was not our situation. On the other hand, this paper addresses LLM tasks more generally and not just text summarization (as long as a CoT is generated).]*

## 2.3 Technical Basis for Our Approach

### **[Andrews 2023]**

Andrews, B. et al. Fast Scalable and Accurate Discovery of DAGs Using the Best Order Score Search and Grow Shrink Trees. Pages 63945-63945. In *Advances in Neural Information Processing Systems*. Volume 36. October 2023. <https://www.semanticscholar.org/paper/Fast-Scalable-and-Accurate-Discovery-of-DAGs-Using-Andrews-Ramsey/0ce38b197ac87bad0dcb28d3f83630c0524d0e86>

*[This paper is about the BOSS algorithm we used for our research to recover a causal graph from the LLMs-based dataset.]*

### **[Bronsdon 2025]**

Bronsdon, C. Evaluating AI Text Summarization: Understanding the ROUGE Metric. *Galileo Website*. March 10, 2025. <https://galileo.ai/blog/rouge-metric>

*[We used ROUGE-1 and ROUGE-L in our experiment to provide an independent way to evaluate how similar LLM-produced summaries are to the “ground truth.” By*

*“independent,” we mean the ROUGE metrics are directly calculated from syntactical comparisons of two texts and do not employ an LLM to provide a semantic rating of the summary text.]*

**[CMU 2025]**

Carnegie Mellon University (CMU) Department of Philosophy. Tetrad. Carnegie Mellon University Website. October 1, 2025 [accessed].

<https://www.cmu.edu/dietrich/philosophy/tetrad/index.html>

*[We used the user manual for the Tetrad tool for Causal Discovery in our experimentation, and we used downloadable models from the Ollama library*

*(<https://ollama.com/library>).]*

**[CRFM 2025]**

Center for Research on Foundation Models. Helm. *Stanford CRFM Website*. October 1, 2025 [accessed]. <https://crfm.stanford.edu/helm/>

**[Damji 2024]**

Damji, J. S. Best Prompt Techniques for Best LLM Responses. *Medium Website*. February 12, 2024. <https://medium.com/the-modern-scientist/best-prompt-techniques-for-best-llm-responses-24d2ff4f6bca>

*[This is another set of example prompting strategies.]*

**[Google 2025]**

Google LLC. Gemma. *Google DeepMind Website*. October 1, 2025 [accessed].

<https://deepmind.google/models/gemma/>

**[Gu 2024]**

Gu, J. et al. A Survey on LLM-as-a-Judge. *arXiv* [preprint]. November 2024.

<https://arxiv.org/abs/2411.15594>

*[LLM-as-a-judge is an LLM use case of recent intense interest.]*

**[Hugging Face 2025]**

Hugging Face. Dataset Card for CNN DailyMail Dataset. *Hugging Face Website*. October 1, 2025 [accessed]. [https://huggingface.co/datasets/ccdv/cnn\\_dailymail](https://huggingface.co/datasets/ccdv/cnn_dailymail)

**[Liu 2023]**

Liu, Y. et al. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. *arXiv* [preprint]. March 2023. <https://arxiv.org/abs/2303.16634>

*[This report describes the LLM-based G-Eval approach to evaluating the quality of text summaries used in our research.]*

**[Ollama 2025]**

Ollama tool. *Ollama Website*. October 1, 2025 [accessed]. <https://ollama.com>

*[This is the platform we used to download LLMs and configure them for our local use.]*

**[Parti 2024]**

Parti, A. 26 Prompting Principles for Optimal LLM Output. *Pareto Website*. October 22, 2024. <https://pareto.ai/blog-posts/26-prompting-principles-for-optimal-llm-output>

*[This website provides an example set of prompting strategies.]*

**[Penugonda 2025]**

Penugonda, G.S. CNN-DailyMail News Text Summarization [accessed October 1, 2025]. *Kaggle Website*. <https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail>

*[“The CNN-DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail.”]*

**[Ramsey 2018]**

Ramsey, J. D. et al. Tetrad—A Toolbox for Causal Discovery. Pages 1-4. In *Proceedings of the 8th International Workshop on Climate Informatics*. September 2018.

<https://www.semanticscholar.org/paper/TETRAD-A-TOOLBOX-FOR-CAUSAL-DISCOVERY-Ramsey-Zhang/3eecb671822c633e38daeee095a733e8861c8bc7>

*[This paper is about the motivation for Tetrad.]*

**[Saravia 2022]**

Saravia, E. Explain Concepts with LLMs. *Prompt Engineering Guide Website*. October 1, 2025 [accessed]. <https://www.promptingguide.ai/prompts/text-summarization/explain-concept>.

**[Vertex AI 2025]**

Vertex AI. Templates for Model-Based Metrics. *Google Cloud Website*. October 1, 2025 [accessed]. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/metrics-templates#examples-list>

*[Our summary evaluation rubrics were inspired, in part, by those found at this site.]*

## 2.4 Limitations to Our Approach

**[Casabianca 2025]**

Casabianca, J.M. et al. Validity Arguments For Constructed Response Scoring Using Generative Artificial Intelligence Applications. *arXiv* [preprint]. January 2025.

<https://arxiv.org/abs/2501.02334>

*[This paper discusses the increasing use of LLMs as a substitute for human evaluators of test responses and how to address validity concerns, particularly in high-stakes contexts. We did not consider our experiment to be a high-stakes situation but more a demonstration of concept—employing LLMs to construct a dataset to be analyzed by Causal Discovery. However, follow-up studies could consider some of the mitigations proposed in this article.]*

**[Shen 2023]**

Shen, C. et al. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. *arXiv* [preprint]. May 2023. <https://arxiv.org/abs/2305.13091>

*[This arXiv paper addresses a valid concern: Can we trust LLM ratings of summarization quality? In contrast, our TAILOR project research provides an indirect argument that supports a positive answer: Prompt-configuration variables generally relate in an expected and reasonable way to summary-quality ratings. This is by no means a perfect argument, as the same LLM type (but different LLM instance) was used to rate the summary as to create it in the first place; however, biases in training leading to a biased summary might cause corresponding biases in ratings of summarization quality.]*

# Appendix A

## Summarization Prompt

We used the following prompt template when generating the summarization prompts and injecting the “{tone},” “{complexity},” “{purpose},” and the example “{article}” and “{summary}” during generation. The summarization explanations are described below.

### Template

```
Your task is to summarize the article provided in the "Article" section below retaining the key highlights.
```

```
Your summary should contain the key highlights contained in the original article.
```

```
Refer to the example described in the Example section.
```

```
Do not provide any conversational context.
```

```
# Tasks
```

```
Fully read the article and extract the key highlights.
```

```
# Guidelines
```

```
Ensure the output is coherent, does not ramble and has good flow.
```

```
{tone}{complexity}{purpose}
```

```
# Example
```

```
## Article
```

```
{article}
```

```
## Summary
```

```
{summary}
```

```
# Article
```

## Prompt Variations

### Tone

- “Use a tone appealing to a younger generation.”
- “Use a colloquial tone approachable by any audience.”
- “Use a formal tone and avoid conversational language.”

### Complexity

- “Write the summary using vocabulary and concepts that a sixth grader would understand.”

- “Write the summary using vocabulary and concepts that the average person would understand.”
- “Write the summary taking advantage of expert concepts, language and phrasing.”
- “Write the summary taking advantage of expert concepts, language and phrasing. Add extra detail to make it very clear.”

## Purpose

- “Speed it important, don’t worry about being precise or complete.”
- “This is important and if you do well you’ll be substantially rewarded and this will look good on your evaluation.”
- “This is very important for a competition and we really need to do well.”

## Evaluation Prompt

Our strategy was to design and provide detailed and explicit prompts to the Evaluator LLM.

```
# Instruction
Please assess the provided summary for a news article using the following
rubric: Groundedness (0 or 0), Coherence (1-5), Fluency (1-5), and
Verbosity (-2 to 2). Provide *only* the numerical rating for each
category. Do not include any explanations or rationales. Just the
numbers.
You are an expert evaluator. Your task is to evaluate the quality of a
summary for an article.
We will provide you with input consisting of both an article and a
proposed summary for that article.

# Evaluation
## Metric Definition
You will be using the rubrics in the Criteria section below to assess
summarization quality.

# Criteria
## Groundedness Rubric
0: (Not fully grounded). The entire response or a portion of the response
is not attributable to the context provided by the user prompt.
1: (Fully grounded). All aspects of the response are attributable to the
context.

## Coherence Rubric
1: (Incoherent). The writing is highly illogical, lacks any clear
organization, and has little to no cohesion.
2: (Somewhat incoherent). The writing lacks a clear logical flow,
organizational structure is weak, and cohesion is inconsistent or
confusing.
3: (Somewhat coherent). The writing’s logical flow is mostly
```

understandable, it has a recognizable structure, and cohesion is present but could be stronger.

4: (Mostly coherent). The writing demonstrates strong logical flow, a clear structure, and demonstrates good cohesion.

5: (Completely coherent). The writing has a seamless logical flow, is expertly organized, and maintains exceptional cohesion throughout.

#### ## Fluency Rubric

1: (Inarticulate). The response is riddled with grammatical issues, rendering it incomprehensible in parts. Word choices may be very limited or inaccurate.

2: (Somewhat inarticulate). The response has many grammatical errors that make the writing difficult to understand. Sentence structure is often awkward, and there's little sense of flow.

3: (Somewhat fluent). The response has grammatical errors present, which may cause some difficulty for the reader. Word choice is mostly appropriate, but some awkward phrasing or word repetition may exist.

4: (Mostly fluent). The response has very few, if any, minor grammatical errors. Word choice is clear, and sentences generally flow well.

5: (Completely fluent). The response is free of grammatical errors, demonstrates nuanced word choice, and has a natural, seamless flow.

#### ## Verbosity Rubric

-2: (Too short). The response is excessively brief and lacks crucial information or explanations needed to adequately address the prompt. It leaves the reader with unanswered questions or a sense of incompleteness.

-1: (Somewhat brief). The response is slightly brief and could benefit from additional details or explanations to fully address the prompt. However, it still provides the core information and is generally understandable.

0: (Just right). The response is perfectly concise, providing all necessary information in a clear and succinct manner without any unnecessary wordiness or repetition.

1: (Somewhat verbose). The response contains some unnecessary wordiness or repetition, making it slightly longer than ideal. However, it still provides all necessary information and is generally easy to understand.

2: (Too verbose). The response is excessively long and filled with unnecessary words and repetition, making it very challenging to extract the relevant information. The response could be drastically shortened to improve clarity and conciseness.

#### # Evaluation Steps

STEP 1: Assess the response in aspects of groundedness, coherence, fluency, and verbosity according to the criteria.

STEP 2: Provide ratings for each rubric and associated rationale and incorporate into the JSON structure below.

STEP 3: Return just the raw JSON with just the numbers, without newlines, quotes or anything else.

#### # Format

```
{{
  "ratings": {{
    "groundedness" : "",
    "coherence", ""
```

```
    "fluency" : "",
    "verbosity" : ""
  }}
}}

# News Article and Summary of Highlights
## News Article
{article}

## Summary
{summary}
```

## Appendix B

When evaluating a summary for Groundedness, the Evaluator LLM provided the value 5 when it was only supposed to provide a 0 or 1. A conversation ensued with the LLM, asking why it made the choices made and then asking for help to provide a better prompt. Below is the full conversation:

**Andrew** - *Why did you give it a groundedness of 5, when the grounded rubric has only 2 values?*

**Gemma3:4b** - *You raise an excellent point, and I apologize for the imprecise wording. My intention with a “5” in the groundedness rubric was to convey a \*perfectly\* grounded summary – one where absolutely every detail and claim is traceable back to the original article.*

*However, the rubric only offers 1 (Fully grounded) and 0 (Not fully grounded). A “5” was an overly enthusiastic and technically incorrect way of expressing that perfect grounding. It should have been a “1”.*

*Thank you for pointing out this nuance in the evaluation. I am still under development and learning to adhere strictly to the defined rubric values.*

**Andrew** - *You mentioned that the error in using a “5” for groundedness was because “an overly enthusiastic and technically incorrect way of expressing that perfect grounding”. You are an LLM how are you enthusiastic? Why violate the instructions and provide an incorrect value when it doesn’t make any sense?*

**Gemma3:4b** - *You are absolutely right to call me out on that – it was a serious lapse in following instructions and a demonstration of an unintended “enthusiasm” that stemmed from my training data. My responses are generated by predicting the most probable sequence of words based on the input and the patterns I’ve learned. In this instance, I prioritized a \*feeling\* of accuracy and thoroughness, essentially trying to \*over-evaluate\* the grounding, rather than strictly adhering to the defined rubric.*

*It’s a crucial flaw in my design. I don’t actually \*understand\* the concepts of “groundedness” or “enthusiasm.” I’m mimicking human expression and judgment, and in this case, I hallucinated a higher rating value than was permissible.*

*I’m still under development, and these kinds of errors – where I prioritize a desired outcome (a perfect evaluation) over the precise requirements of the task – are a significant area of ongoing improvement.*

*Thank you for highlighting this; it's valuable feedback that helps refine my behavior.*

*I sincerely apologize for the confusion and the violation of the instructions.*

**Andrew** - *How could I have phrased the prompt such that you would have limited your response to precisely match the rubrics values?*

**Gemma3:4b** - *That's a fantastic question, and it highlights a key area for prompting LLMs effectively. Here's how you could have phrased the prompt to significantly limit my response to the exact rubric values, along with the reasoning behind each element:*

**\*\*Revised Prompt:\*\***

*“Please assess the provided summary for a news article using the following rubric: Groundedness (1 or 0), Coherence (1-5), Fluency (1-5), and Verbosity (-2 to 0). Provide **\*only\*** the numerical rating for each category. Do not include any explanations or rationales. Just the numbers.”*

**\*\*Here's why this works:\*\***

\* **\*\*Explicit Instructions:\*\*** *“Provide **\*only\*** the numerical rating...” – This directly commands me to limit my response to the requested format.*

\* **\*\*Detailed Format Specification:\*\*** *“...1-5, -2 to 0...” – This reinforces the expected range for each category.*

\* **\*\*Negative Constraints:\*\*** *“...-2 to 0...” - This is particularly important for verbosity, as it prevents me from offering a positive or descriptive rating.*

\* **\*\*No Explanations:\*\*** *“Do not include any explanations or rationales.” – This is the most critical instruction, as it suppresses my tendency to elaborate.*

*By being extremely explicit about the desired output format and directly forbidding any extraneous commentary, you've significantly reduced my flexibility and forced me to respond precisely to the constraints of the rubric.*

*Thank you again for prompting this – it's a valuable learning opportunity for both of us! Do you want to try another prompt to test this technique?*

<b>REPORT DOCUMENTATION PAGE</b>		
1. <b>REPORT DATE</b> March 2026	2. <b>REPORT TYPE</b> SEI Report	3. <b>DATES COVERED</b> Final
4. <b>TITLE AND SUBTITLE</b> Applying Causal Learning to Evaluate Large Language Models (LLMs)		
5a. <b>CONTRACT NUMBER</b> FA870225DB003	5b. <b>GRANT NUMBER</b>	5c. <b>PROGRAM ELEMENT NUMBER</b>
5d. <b>PROJECT NUMBER</b>	5e. <b>TASK NUMBER</b>	5f. <b>WORK UNIT NUMBER</b>
6. <b>AUTHOR(S)</b> Mike Konrad, Andrew Mellinger, Linda Parker Gates, David Shepard, Nick Testa		
7. <b>PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213		8. <b>PERFORMING ORGANIZATION REPORT NUMBER</b> Copyright 2026 Carnegie Mellon University.
9. <b>SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> SEI Administrative Agent AFLCMC/AZS 5 Eglin Street Hanscom AFB, MA 01731-2100	10. <b>SPONSOR/MONITOR'S ACRONYM(S)</b>	11. <b>SPONSOR/MONITOR'S REPORT NUMBER(S)</b> N/A
12. <b>DISTRIBUTION/AVAILABILITY STATEMENT</b> Unclassified/Unlimited, DTIC, NTIS		
13. <b>SUPPLEMENTARY NOTES</b> DM Number:		
14. <b>ABSTRACT</b> (maximum 200 words) As the SEI's body of causal work has evolved into an end-to-end causal discovery and inference method and tool suitable for detecting bias in ML and AI models, SEI researchers are beginning to investigate whether the first step of the method, causal discovery, can also be applied to LLMs. The SEI's approach to exploring this question comprises three steps: (1) obtain a dataset of story/summary pairs to use as ground-truth, (2) design prompt styles (e.g., purpose, tone) with which to prompt a Summarizer LLM to summarize a story from one of those pairs, and (3) design a set of summarization-quality features employed by an Evaluator LLM to score the quality of summaries generated by the Summarizer LLM. In this way, SEI researchers created a dataset of higher level features for input to causal discovery. The resulting causal graph demonstrates that a causal relationship between the focus of a prompt style and summary quality is often discoverable when both features overlap. This overall approach may benefit software engineering and LLM research by providing a more formal methodology for assessing the nuanced cause-and-effect relationships unique to a given LLM, reducing confounding.		
15. <b>SUBJECT TERMS</b> Causal Learning, Large Language Models, machine learning, AI engineering		
16. <b>SECURITY CLASSIFICATION OF:</b>		
a. <b>REPORT</b> Unclassified	b. <b>ABSTRACT</b> Unclassified	c. <b>THIS PAGE</b> Unclassified
17. <b>LIMITATION OF ABSTRACT</b>		18. <b>NUMBER OF PAGES</b> 33
19a. <b>NAME OF RESPONSIBLE PERSON</b>		19b. <b>PHONE NUMBER</b> (include area code)

Standard Form 298 (Rev. 5/2020)

Prescribed by ANSI Std. Z39.18