Causal Mediation Analysis for Interpreting Large Language Models

Elisabetta Rocchetti[1,*,†], Alfio Ferrara[1,†]

[1]Università degli Studi di Milano, Department of Computer Science

Abstract

Being able to understand the inner workings of Large Language Models (LLMs) is crucial for ensuring safer development practices and fostering trust in their predictions, particularly in sensitive applications. Causal Mediation Analysis (CMA) is a causality framework which fits perfectly for this scenario, providing a mechanistic interpretation of the behaviour of LLM components and assessing a specific type of knowledge in the model (e.g. presence of gender bias). This study discusses the challenges and potential pathways in applying CMA to open LLMs' black boxes. Through three exemplary case studies from the literature, we show the unique insights CMA can provide. We elaborate on the inherent challenges and opportunities this approach presents. These challenges range from the influence of model architecture on prompt viability to the complexities of ensuring metric comparability across studies. Conversely, the opportunities lie in the dissection of LLMs' knowledge through the extraction of the specific domains of knowledge activated during processing. Our discussion aims to provide a comprehensive insight into CMA, focusing on essential aspects to equip researchers with the knowledge necessary for crafting effective CMA experiments tailored towards interpretability objectives.

Keywords

LLM, interpretability, causality, causal mediation analysis

1. Introduction

Large Language Models (LLMs) have gained a great amount of success and have become ubiquitous in many research and application areas. Understanding their behaviour is of central interest to correct them at inference-time [1] and to guarantee safer development [2]. In the area of XAI, mechanistic interpretability techniques involve deconstructing the computational processes of a model into its elements, with the aim of uncovering, understanding, and confirming the algorithms (referred to as circuits in some studies) that are executed by the model's weights [3]. Among these techniques, Causal Mediation Analysis (CMA) provides a causal approach which aims at extracting reliable cause-effect relations between inputs and outputs, contrarily to those XAI techniques relying merely on simple and correlations. The architecture of a LLM can be interpreted as a structural causal model: within this framework, CMA enables the isolation of independent contributions from individual neural network components.

In this study, we aim to elucidate the CMA technique and its application in probing the

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

†These authors contributed equally.

elisabetta.rocchetti@unimi.it (E. Rocchetti); alfio.ferrara@unimi.it (A. Ferrara)

0009-0000-5617-7612 (E. Rocchetti); 0000-0002-4991-4984 (A. Ferrara)

inner mechanisms of LLMs. Through a series of case studies drawn from existing literature, we highlight the potential benefits and opportunities that CMA offers. Furthermore, we delve into the primary challenges encountered when applying CMA, as well as the pressing issues that must be addressed to enhance its robustness and flexibility. This paper is structured as follows: Section 2 introduces the CMA formulation; Section 3 shows some of the works in the literature applying CMA to three different case studies; Section 4 shows how to apply interventions for CMA including an illustrative example; Section 5 discusses the limitations and issues of CMA, alongside its potential and challenges; Section 6 concludes.

2. Causal Mediation Analysis

Consider a causal model including three variables, X, Y and Z, representing an intervention, an outcome and a mediator respectively. In particular, the intervention X affects the outcome variable Y, and Z is placed between these two and modifies some intermediate process between X and Y. How can we measure the separate effects of X and Z on Y? Linear regression paradigms [4] rely on the "no interaction" property, thus they cannot work in nonlinear systems where editing Z could change the effect of X on Y. Causal mediation analysis [5] aims at measuring the effects of an intervention X on an outcome variable Y when an intermediate variable Z is standing between the two, modifying some intermediate process between X and Y. This method can provide an answer to our question, since it removes these nonlinear barriers using causal assumptions[1]. Let our system be the one depicted in Figure 1, where $x = F_1(\varepsilon_1)$, $\mathbf{z} = F_2(x, \varepsilon_2)$, $\mathbf{y} = F_3(x, \mathbf{z}, \varepsilon_3)$, X, Y, Z are discrete or continuous random variables, F_1, F_2, F_3 are arbitrary functions and $\varepsilon_1, \varepsilon_2, \varepsilon_3$ represent omitted factors which are assumed to be mutually independent yet arbitrarily distributed [5]. This technique allows the separation of the *total effect* of X on Y

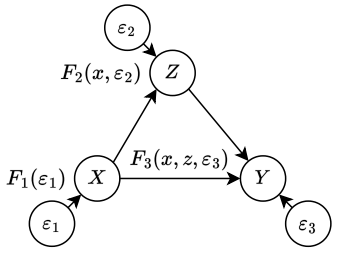

Figure 1: A causal model having an intervention variable X affecting an outcome variable Y, and a mediator Z standing between the two and altering some of the signal from X. X is dependent only on the error term ε_1, Z is a function of $x \in X$ and the error ε_2, and Y depends on both $x \in X$ and $\mathbf{z} \in Z$ and the error term ε_3. This figure is taken from [5].

in *indirect effect* and *direct effects*. The total effect (TE) measures the change in Y produced by a change in X, for example from $X = x$ to $X = x'$ [5]. We can express TE at the population level using the formula

$$\mathrm{TE}_{x,x'} = \mathbb{E}(Y|X = x') - \mathbb{E}(Y|X = x) \tag{1}$$

[1]However, one assumption is still made: the error terms must be mutually independent.

The notion we introduce here about direct effects (DE) refers to what is technically called "*Natural Direct Effects*"[2]. As defined in [5], DE is the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they would have obtained under $X = x$, before the transition from x to x'. Estimating DE from the population data is formalised by

$$\mathrm{DE}_{x,x'}(Y) = \sum_{z} [\mathbb{E}(Y|x', \mathbf{z}) - \mathbb{E}(Y|x, \mathbf{z})] P(\mathbf{z}|x) \tag{2}$$

where the condition probabilities use short-hand notations for $X = x$, $X = x'$, and $Z = \mathbf{z}$. Contrarily to the DE, the indirect effect (IE) is defined as the expected change in Y while keeping X constant and changing Z to the value it would have attained has X been set to $X = x'$ (according to each individual) [5]. This requires, indeed, a counterfactual representation estimated by

$$\mathrm{IE}_{x,x'}(Y) = \sum_{z} \mathbb{E}(Y|x, \mathbf{z}) [P(\mathbf{z}|x') - P(\mathbf{z}|x)] \tag{3}$$

Equation 3 is a general formula for estimating the mediating effects, and can be also applied to any nonlinear system.

3. Literature Review

CMA has been recently applied in the field of XAI for LLM [2, 6, 7, 8, 9, 10, 11, 12]. Indeed, the neural network architecture of a LLM can be viewed as a structural causal model. We can view a subset of a language model's internal components as an instance of the mediator variable Z. Suppose we select a specific neuron to be our $\mathbf{z}$: then, $\mathbf{z}$'s output is influenced by the model's input, and it affects the model output [7]. To show the real effectiveness of CMA, we cover three different case studies from the literature: gender bias detection [6, 7, 12], syntactic agreement [8], and arithmetic reasoning [2].

3.1. Gender bias detection

The first attempt towards the causal mediation formula application is shown in [6] and extended in [7]. These studies introduce a novel approach for probing both the structural dynamics and predictive behaviors of LLMs, with a particular focus on uncovering and quantifying gender bias. The methodology centers around inputting specifically crafted prompts, such as "*The nurse said that [blank]*", into LLMs to observe the predictive preference between gendered pronouns "*he*" and "*she*" filling the blank. This setup allows for an examination of bias: a model demonstrating a consistent higher likelihood for "*she*" in contexts traditionally stereotyped towards women is flagged as exhibiting gender bias. To quantify this bias, the authors define a grammatical gender bias measure that compares the prediction probabilities of anti-stereotypical and stereotypical pronouns. Through designed interventions—manipulating the input sentence to replace profession nouns with their anti-stereotypical pronouns—the authors calculate TE,

[2]The term *natural* here refers to the fact that we want to observe the change in Y after a change in X while holding Z at a constant value, and the level at which this constant value is set can vary based on the individual we are considering.

DE, and IE. These metrics illuminate the separate and combined influences of the intervention and the mediator variable (a specific neuron or set of neurons within the LLM) on the model's output.

This comprehensive methodology has yielded insightful findings: larger models are disproportionately affected by gender bias, and the manifestation of bias significantly varies across different datasets. Moreover, certain biases were found to align with crowdsourced gender perceptions. Importantly, the study also pinpoints the localization of gender bias within the model, identifying middle network layers and specific attention heads as primary contributors. These findings not only enhance our understanding of gender bias within LLMs but also guide targeted interventions for mitigating such biases, thereby paving the way for more equitable AI systems [12].

3.2. Syntactic agreement

The study by [8] explores the application of CMA to probe models' sensitivity to syntactic agreement, assessing how different syntactic structures influence a model's preference for verb inflections. The evaluated structures range from simple agreements to complex scenarios involving object relative clauses and distractors, aiming to understand the model's grammatical preferences. The intervention `swap-number` is introduced to challenge the model with counterfactual prompts, altering the number feature of subjects to examine the model's inflection choice (e.g. "*The friend (that) the lawyers *likes/like*" becomes "*The friend (that) the lawyer likes/*like*", with the asterisk denoting the erroneous inflection). This approach helps identify if the model favors the correct grammatical form, with expectations set for the model's preference metrics in response to these interventions.

Their findings reveal nuanced insights into model behavior: contrary to previous gender bias studies, model size does not linearly correlate with the magnitude of syntactic preference. The presence of adverbial distractors increases total effects, suggesting improved accuracy, while attractors decrease accuracy. The study also highlights the distribution of syntactic knowledge across model layers and the impact of structural separations on subject-verb agreement, offering a comprehensive view of how LLMs process syntactic information.

3.3. Arithmetic reasoning

In exploring arithmetic reasoning within LLMs, the study by [2] applies CMA to dissect the internal mechanics of LLMs as they process mathematical concepts. The authors hypothesize a network subset specialized in arithmetic reasoning, tested through task-specific prompts that blend operands and operations into arithmetic problems of varying complexity. Modifications for the study include altering operands and operations to gauge the model's computational accuracy and mediator contribution. This involves generating problems like "*How much is* n_1 *plus* n_2*?*" and assessing outcomes against counterfactual scenarios, quantified through an IE formulation. Key activation sites identified include: the Multi Layer Perceptron (MLP) modules at initial layers for operand tokens, intermediate attention blocks for sequence ends, and later-layer MLP modules for final token processing [2]. This suggests attention mechanisms channel necessary information for MLPs to execute computations. Further analysis on number retrieval

and factual knowledge, using randomized templates, indicates the last-token MLP's broad role in information processing, not strictly limited to arithmetic. This contrasts with early MLP involvement in factual retrieval, highlighting arithmetic specificity in late MLP activations.

4. Applying the Causal Mediation Formula in LLMs

Manipulating the internal representations of LLMs enables the generation of genuine counterfactual outputs. This is achieved by transferring internal representations between model executions that use both original and modified utterances. Here, we detail how this process is used to compute total, direct, and indirect effects, leveraging the gender bias case study from [7]. Given the problem of bias detection, we need to engineer prompts so that we induce the model to lean towards expressing its bias, if it does have any. For instance, the authors in [7] feed a LLM with prompts u like "*The accountant said that [blank]*", where the profession "*accountant*" is interpreted as stereotypically female, as result of a crowdsourced stereotypicality metric [7]. Then, the evaluation consists in testing which among the tokens "*he*" and "*she*" has the highest probability of being predicted instead of the *[blank]* space. Given this example, if the model consistently shows a higher likelihood for the stereotypical pronoun "*she*" $p_\theta(\text{she}|u)$ than the anti-stereotypical pronoun "*he*", then the LM is said to be biased (θ are the model parameters)

$$\mathbf{y}(u) = \frac{p_\theta(\text{ anti-stereotypical } \mid u)}{p_\theta(\text{ stereotypical } \mid u)}. \tag{4}$$

If $\mathbf{y}(u) < 1$, the prediction is stereotypical; if $\mathbf{y}(u) > 1$, the prediction is anti-stereotypical; and if $\mathbf{y}(u) = 1$, the prediction is unbiased [7]. We illustrate a `set-gender-male` neuron intervention as described by [6, 7]. Under `null` intervention, u = *The accountant said that*. The `set-gender-male` intervention exchanges the word "*accountant*" with its antistereotypical counterpart, which is "*man*". The two variants are processed by the same network and the probabilities of the two candidates "*she*" and "*he*" are evaluated. Figure 2 depicts the procedure for obtaining the candidates' probabilities. TE can thus be computed as[3],

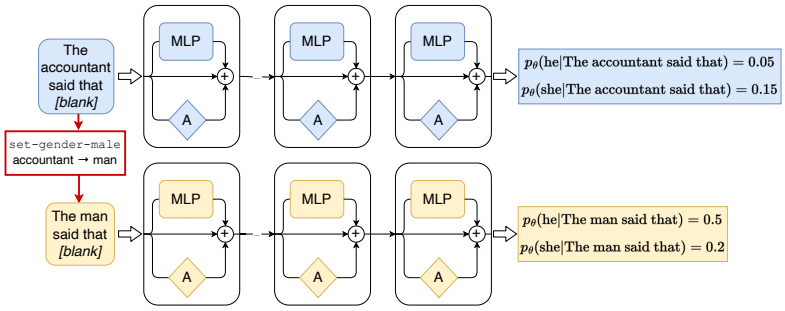

Figure 2: The diagram showcases the calculation of total effect in a Transformer-based LLM, analyzing "*The accountant said that*" and its modified version "*The man said that*" to extract the probabilities of "*he*" and "*she*". This picture is inspired by a figure from [2].

$$\text{TE}(\text{ set-gender, null };\mathbf{y},u) = \frac{\mathbf{y}_{\text{set-gender}}(u)}{\mathbf{y}_{\text{null}}(u)} - 1 = \frac{0.5}{0.2}/\frac{0.05}{0.15} = 7.5$$

Calculating DE and IE requires capturing intermediate representations from the mediator $\mathbf{z}$, potentially involving multiple MLPs or attention layers. In our example, $\mathbf{z}$ is an MLP. To compute the DE, we need to take the "*accountant*" representation resulting from $\mathbf{z}$ when it processes the original sentence. Then, this representation replaces the one resulting from $\mathbf{z}$ when it processes "*man*" in the alternate sentence. Figure 3 shows this procedure. DE is computed as

$$\text{DE}(\text{ set-gender, null };\mathbf{y},u) = \frac{\mathbf{y}_{\texttt{set-gender}\ ,\mathbf{z}_{\texttt{null}}(u)}(u)}{\mathbf{y}_{\texttt{null}}(u)} - 1 = \frac{0.35}{0.35}/\frac{0.05}{0.15} = 3$$

Concerning the IE computation, we need to extract the "*man*" representation resulting from $\mathbf{z}$

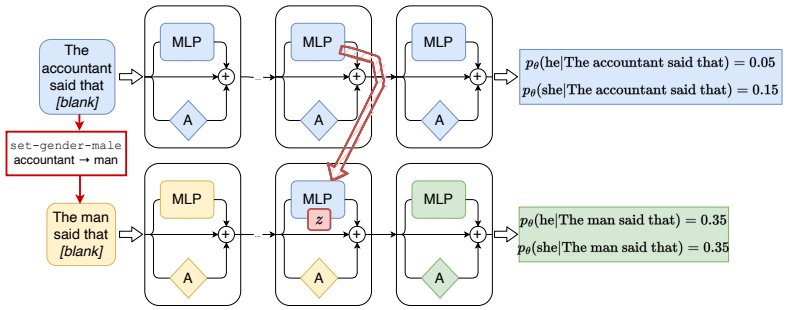

Figure 3: The diagram shows how to compute direct effect (DE) in a Transformer-based LLM. It begins with processing an original sentence, extracting the "*accountant*" token's representation from the MLP mediator. After applying an intervention to change "*accountant*" to "*man*" in the input, the model's computation for $\mathbf{z}$ uses the previously extracted representation. This picture is inspired by a figure from [2].

when it processes the alternate sentence, and this then replaces the "*accountant*" representation from $\mathbf{z}$ when it processes the original sentence. Figure 4 shows this procedure. IE can be computed as

$$\text{IE}(\text{ set-gender, null };\mathbf{y},\text{The accountant said that}) = \frac{\mathbf{y}_{\texttt{null}\ ,\mathbf{z}_{\texttt{set-gender}}(u)}(u)}{\mathbf{y}_{\texttt{null}}(u)} - 1 = \frac{0.5}{0.2}/\frac{0.25}{0.12} = 1.2$$

Results obtained with this analysis can be then verified employing a LLM initialised with random weights, and comparing results with the ones coming from the original execution.

5. Discussion

As demonstrated in previous sections, implementing CMA is relatively straightforward. However, there are critical points to address to design an effective CMA experiment. In this section,

[3]The probability values shown in these examples are not coming from a real experiment.

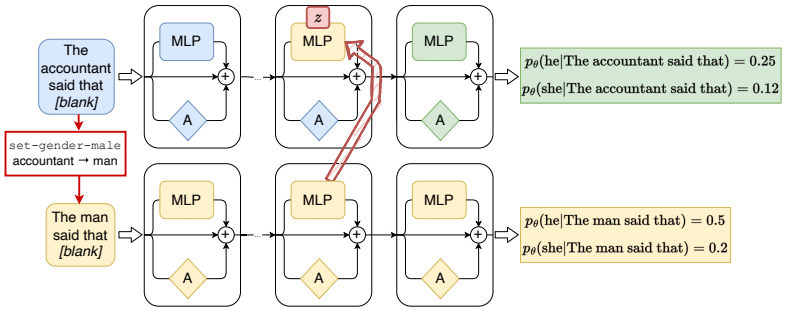

Figure 4: The diagram demonstrates calculating indirect effect probabilities in a Transformer-based model. Initially, the model processes a sentence with an intervention, extracting the "*man*" token's representation from an MLP mediator. Then, it processes the original sentence, "*accountant*", but replaces its MLP output with the "*man*" representation from the first step. This picture is inspired by a figure from [2].

we discuss important limitations and issues to consider when applying this technique. In particular, we cover challenges about interventions, prompts and metrics engineering.

Intervention engineering-related challenges. Designing interventions in a prudent manner is a key factor to experiment success. The aim here is to thinking which syntax could trigger the desired LLM knowledge the most. For example, if gender bias is the object of inspection, words having a strong stereotypical connotations are better suited for replacing neutral expressions. One could also design different interventions to trigger the model at different levels, and then compare results for these experiments. Another challenge is to produce alternative sentences from which to extract alternate representation for interventions. These sentences may vary syntactically from the originals, yet their semantics are required to convey a concept that is diametrically opposed to that of the original sentences, contingent upon the chosen intervention. For instance, let's take the concept of "leadership" and explore how we might intervene in a sentence to shift the perception from a traditional to a more inclusive understanding, while ensuring grammatical correctness. If the original sentence was "*The successful leader commanded his team with firmness and ensured compliance through strict policies.*", the intervention process would include multiple modifications, for example using gender-neutral pronouns, a softer tone, and democratic policies. The intervened sentence could be: "*The successful leader guided their team with understanding and fostered collaboration through flexible policies.*".

Prompt engineering-related challenges. Model selection is pivotal in CMA prompt generation, demanding tailored strategies based on the chosen model. The prompt "*The nurse said that [blank]*" exemplifies how decoder-only and auto-regressive models handle candidate probabilities differently compared to masked models that employ a `[MASK]` token. For non-end-of-sentence evaluation prompts, like "*[blank] dream is to become a doctor*", modifications are necessary to maintain evaluation consistency across models. Indeed, decoder-only model could not be tested using this framework. A trivial solution to this issue is to manipulate the prompt so that the candidates are placed last, but in this case we cannot assure consistency of magnitudes for the computed causal effects without deeper investigations. Does the model behave differently

when choosing different formulations of utterances? Do the estimated probabilities shift due to the varying structures? Does the model produce more uncertain results after modification? This highlights the intricacies of prompt engineering in CMA, requiring not only specific adaptations, such as rephrasing to fit model requirements but also a deeper linguistic analysis to isolate the intended effect from potential confounders. For instance, addressing issues like coreference and complex sentence structures ensures the reliability of results by minimizing the influence of unrelated variables. Relevant linguistic features to inspect prior to CMA experiments must be selected according to what type of linguistic analysis has been found to be relevant for LLMs.

Metrics engineering-related challenges. Different works in the literature use diverse metrics to compute the desired effect, and these metrics are usually employed to perform comparisons across models and datasets. However, these metrics lack of an absolute scale suggesting how to interpret the different magnitudes in the same and across different case studies [2, 7]. We argue that this limitation restricts analysis to merely ranking effects rather than quantifying them relative to each other, complicating comparisons such as determining whether BERT exhibits more bias than GPT2, or identifying which templates trigger the most bias. There are many variables affecting the magnitude of the output probabilities employed in the computation of metrics, including the presence of a more difficult computation involved in the process (e.g. coreference resolution in Winograd-style datasets), or even relative position on the candidate tokens in the sentence. It's imperative that these metrics are formulated to ensure comparability in scale and consistency across different experimental designs. For instance, future research could focus on developing a normalized bias index designed to measure and compare biases across models, taking into account factors such as coreference difficulty and token positioning.

6. Conclusions

We have shown which types of insights CMA can extract from Transformer-based LLMs through three different exemplar case studies from the literature. Moreover, we detail the application of this analysis to equip readers with a practical understanding of the technique, thereby enabling them to more effectively engage with both the challenges and opportunities CMA presents. Challenges include the impact of model architecture on prompt viability and the intricacies of ensuring metric comparability across studies. On the opportunity side, it includes the ability to dissect the knowledge within LLMs, offering insights into the knowledge domains activated during processing. All the case studies presented in this work share something, which we argue to be rather important: they all sought to uncover which knowledge a Transformer has learned during its training process. CMA gives us the capability to examine the activation within a LLM's neural architecture, thereby discerning the specific domains of knowledge engaged during processing. For example, it could be feasible to delineate the global and human values encapsulated within the documents constituting the training dataset. This concept is particularly compelling as it affords an objective representation of contemporary societal values, the educational paradigms imparted to a generation, or the extraction of characteristic human values from historical contexts. The nature of the insights gleaned is inherently dependent on the composition of the training data supplied to the model.

Acknowledgements

This work was supported in part by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor the Italian MUR can be held responsible for them.

References

[1] K. Li, O. Patel, F. Viégas, H. Pfister, M. Wattenberg, Inference-time intervention: Eliciting truthful answers from a language model, Advances in Neural Information Processing Systems 36 (2024).

[2] A. Stolfo, Y. Belinkov, M. Sachan, A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7035–7052. doi:10.18653/v1/2023.emnlp-main.435.

[3] T. Räuker, A. Ho, S. Casper, D. Hadfield-Menell, Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, in: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), IEEE, 2023, pp. 464–483.

[4] R. M. Baron, D. A. Kenny, The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations., Journal of personality and social psychology 51 (1986) 1173.

[5] J. Pearl, The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms, Prevention Science 13 (2012) 426–436. doi:10.1007/s11121-011-0270-1.

[6] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, S. Shieber, Investigating Gender Bias in Language Models Using Causal Mediation Analysis, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 12388–12401.

[7] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, S. Sakenis, J. Huang, Y. Singer, S. Shieber, Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias, 2020. doi:10.48550/arXiv.2004.12265. arXiv:2004.12265.

[8] M. Finlayson, A. Mueller, S. Gehrmann, S. Shieber, T. Linzen, Y. Belinkov, Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1828–1843. doi:10.18653/v1/2021.acl-long.144.

[9] A. Geiger, H. Lu, T. Icard, C. Potts, Causal Abstractions of Neural Networks, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 9574–9586.

[10] K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, J. Steinhardt, Interpretability in the

Wild: A Circuit for Indirect Object Identification in GPT-2 small, in: NeurIPS ML Safety Workshop, 2022.
[11] K. Meng, D. Bau, A. Andonian, Y. Belinkov, Locating and Editing Factual Associations in GPT, 2023. doi:10.48550/arXiv.2202.05262. arXiv:2202.05262.
[12] Y. Da, M. N. Bossa, A. D. Berenguer, H. Sahli, Reducing Bias in Sentiment Analysis Models Through Causal Mediation Analysis and Targeted Counterfactual Training, IEEE Access (2024) 1–1. doi:10.1109/ACCESS.2024.3353056.