Investigating the Intuitive Logic behind Autoregressive Language Models

Gaia Carenini & Alexandre Duplessis[1]

Abstract

Transformer-based language models have shown a stunning collection of capabilities but largely remain black boxes. Understanding these models is hard due to their complex non-linear interactions in densely-connected layers. In this work, we address the problem of interpretability of large regressive language models with a principled approach inspired by basic logic. First, we show how classical mathematical logic does not grasp the reasoning system of these models and we propose an *intuitive logic*, which redefines the classic logical operators. Then, we proceed with the localization of the activated areas associated with its operators. From the localization results, we obtain topological information about the network that induces the formulation of a conjecture about the mechanisms underlying the logic in GPT 2-XL. We test extensively the conjecture operating model editing.

1. Introduction

Transformer-based language models (Radford et al., 2019a; Brown et al., 2020) have shown a stunning collection of capabilities but largely remain black boxes. Despite this obscurity, language models are increasingly employed in a wide range of applications, spanning from the realization of chat-bots (Caldarini et al., 2022) to the development of medical models (Zhang et al., 2022a). These classes of applications require an assessment of possible undesirable behaviors and strong guarantees regarding the predictability of the model, justifying for instance the emergent behaviors (e.g. (Wei et al., 2022)). In the absence of the latter, security threats might arise (e.g. (Cohen et al., 2021), (Bathaee, 2018) and (Hendrycks & Mazeika, 2022)).

Mechanistic interpretability attacks these questions providing tools to better analyse this colossal architectures by reverse engineering model computation into human-understandable components, e.g. (Geiger et al., 2021) and (Meng et al., 2022b). In particular, recent breakthroughs have shed light on some basic aspects of the architecture of GPT, such as the storing of factual associations (Meng et al., 2022a) and the circuits performing indirect object identification (Wang et al., 2022). However, to the best of our knowledge, no result has yet captured the circuits underlying the logical behavior of large language models. The existence of such circuits is supported by experimental evidence: transformer-based language models can indeed perform numerous tasks involving logic skills reaching performance above-chance (Dasgupta et al., 2022), such as selection-inference (Creswell et al., 2022) or automatic theorem proving (Polu & Sutskever, 2020; Wu et al., 2022).

In this article, we begin from the conclusions of (Zhang et al., 2022b), where it is observed that LLM learns statistical features that inherently exist in logical problems rather then reasoning and we perform a systematic analysis of the encodings in GPT-2 (Radford et al., 2019b) of syntagma involving logical operators to better describe the statistical features cited above. We start by observing that classical logic does not describe properly the statistical data in LLM and from a redefinition of logical connectors ($\wedge, \vee, \neg$), we assess the ability of these infrastructures to intuitively grasp the meaning of formal logic. We formalize this skill in a formal model of intuitive logic. Then, by applying the *causal tracing* technique described in (Meng et al., 2022a), we proceed with the location of the activation layers corresponding to the above-mentioned operators. Throughout this step, we point out important invariants and eventually, we conclude by verifying the model of encoding proposed thanks to the editing of the model.

2. Preliminaries

Classical Logic and LLM Within classical logic, there are numerous invariants and symmetries, e.g. (Areces et al., 2013; Sambin et al., 2000), that constitute a characterising element and are widely used in combinatorial optimisation and satisfiability problems, e.g. (Bogaerts et al., 2022; Ghoniem & Sherali, 2011). Proving that such symmetries are not respected by a language model is sufficient to move away from the classical logical approach. Natural language, on which LLM are trained, is characterised, among other

[1] Gaia Carenini and Alexandre Duplessis, Department of Computer Science, École normale supérieure - PSL Research University, Paris, France.. Correspondence to: Gaia Carenini and Alexandre Duplessis < name.surname@ens.psl.eu >.

features, by being asymmetric with respect to the $\wedge$ operator (Cinque, 2009). The previous observations summed to the fact that LLM are corpus-model (Veres, 2021), allows us to formulate the conjecture below.

Conjecture 2.1. The statistical features that inherently exists in logical problems and that are learned by LLM violate classical logic.

Intuitive Logic for LLM Once established that conjecture 2.1 is true, the issue of finding a proper characterization of statistical features becomes a problem of primary importance in the interpretability of LLM. We propose, partially inspired by results in the field of linguistics, a system of notions and operators that consists of a redefinition of the classical ones. We focus in particular in finding out the statistical features associated with the classical concepts of *equality* ($=$), *conjunction* ($\wedge$), *disjunction* ($\vee$), *negation* ($\neg$), *adversarial conjunction*, *if-then statement* ($\rightarrow$) and *synonym*.

In classical logic, equality denotes a binary relationship of equivalence between two entities, called *members* of the equality. Instead, natural language presents 2 distinct notions of equality corresponding to the classical one and to the one at *semantic*. For seek of clarity, consider a set of distinct words of standard English, $\mathcal{W}$, and a pair of distinct words $w \neq w' \in \mathcal{W}$. From an *ensemble* point of view, it is clearly evident how w' *is not* w, i.e. belongs to the set $\mathcal{W} - w$. On the other hand, from a semantic point of view, the relation w' *is not* w can be arbitrarily false when w and w' are synonymous [1]. This leads us to the conjectures below.

Conjecture 2.2. There are some statistical features learned by LLM accounting for semantic equality.

Conjecture 2.3. There are some statistical features learned by LLM accounting for the abstraction of adjectives, in particular of synonym.

When it comes to conjunction, we can observe that operator $\wedge$ is source of logical ambiguity within natural language (Popovic & Castilho, 2019). For example, apparently contradictory adjectives, such as *black* and *white*, can be in the same sentence *a zebra is black and white*. From this observation, we formulate another conjecture.

Conjecture 2.4. There is no elementary statistical feature encoding for $\wedge$.

Contrary to $\wedge$, the operator $\vee$ is unambiguous from a linguistic and logical point of view. However, it may be important to notice that in natural language disjunction tends to be exclusive (López-Astorga, 2021).

Conjecture 2.5. There are some statistical features learned by LLM accounting for exclusive disjunction.

Now, we observe that negation is understood in classical logic to be a unitary logical operation, which returns the inverse truth value of a proposition. Within natural language, truth values are not well defined. We claim that we can still capture the essence of negation as follows:

Conjecture 2.6. There are some statistical feature accounting for negation upon the normalization over the weak equality.

We discuss now adversative conjunction; these language elements do not belong to classical logic but still are related to logical operators. In particular, we conjecture that:

Conjecture 2.7. The statistical features learned by LLM concerning adversative conjunctions are correlated to the one concerning negation.

Moreover, we look into the encoding of if-then construct by investigating the validity of the conjecture below.

Conjecture 2.8. The statistical feature involved in if-then constructs in LLM takes in the account for the fact that *If a is b, then b is a.*

2.1. A Model for Explaining Statistical Features in LLM

In the previous section, we claimed the existence of statistical features involving syntagma containing logical operators. However, we haven't explained yet how the structure of transformers can justify all this properties. We formulate a few conjectures in these directions below:

Conjecture 2.9. The statistical features accounting for the $\wedge$ operator in LLM store information concerning the *compatibility* classes. This information is located in MLP layers in accordance to their key-value interpretation.

Conjecture 2.10. The statistical features accounting for the $\wedge$ operator in LLM are encoded through a system of attention heads associated to groups of synonyms and antonyms.

We are also interested in the relationship among features associated to distinct operators. We claimed that the encoding for an intuitive operator $\vee$ has a bias towards the exclusivity nuance, i.e. the one that in standard English can be expressed as *either…or*. This last observation sustains the following conjecture:

Conjecture 2.11. The statistical features accounting for the $\vee$ operator in LLM depends on the statistical features expressing compatibility classes and therefore, according to conjecture 2.10, to the statistical features encoding the intuitive $\wedge$ operator.

[1] Synonyms are understood in the classical linguistic sense and have been checked through the online generator *Thesaurus*.

3. Method

Let us now discuss the methods adopted to investigate the system of conjectures listed in the previous section. There are 3 methods involved in the analysis: direct study of statistical features, localization within the network of zones activated by different operators, and model editing. Before presenting the same, let us introduce the dataset involved in the tests.

Model & Data-set We test our conjectures on the open-source LLM GPT-2 XL (Radford et al., 2019b); however, we claim that the set of experiment proposed could be easily adapted to any LLM. As for the data-set, we consider a small subset $\mathcal{W}$ ($\sim$ **50** words) of standard English. The vocabulary is built so that there is a natural partition $\mathcal{W} = \mathcal{S} \cup \mathcal{A} \cup \mathcal{O}$, where $\mathcal{S}$ is a set of grammatical subjects expressed by substantives, $\mathcal{A}$ is a set of adjectives constructed so that includes separated classes of synonyms and antonyms and $\mathcal{O}$ is a set of grammatical conjunctions expressing the logical operators and constructs that we study. A complete description of $\mathcal{W}$ is provided in the appendix A.

Direct Study of the Statistical Features This method is the one adopted for checking conjectures 2.1, 2.2, 2.3, 2.4,2.5,2.6,2.7 and 2.8. The protocol adopted is fairly simple. We start by constructing some prompts in standard English of the form *s is* a_1 $o \ldots$ where $s \in \mathcal{S}$ is a subject, $o \in \mathcal{O}$ is logical operator and $a_1 \in \mathcal{A}$ is an adjective. Then, we ask to the model GPT 2-XL for completion. Completion is provided under the form of a probability distribution p defined over $\mathcal{A}$. Moreover, we normalize over the restricted vocabulary used in the experiment thanks to the computation of a *soft maximization* and rewrite the result in matrix form $P = (p(a_1, a_2))$ where $a_1, a_2 \in \mathcal{A}$. Eventually (except for conjecture 2.1), we conclude with a further normalization of the matrix, followed by an averaging with its transposition. This corresponds to perform the operation below.

$$\frac{1}{2}\left(\frac{p(a_1, a_2)}{\sum_{a_i \in \mathcal{A}} p(a_i, a_2)} + \frac{p(a_2, a_1)}{\sum_{a_i \in \mathcal{V}} p(a_i, a_1)}\right) \qquad (1)$$

for every a_1 and a_2 in $\mathcal{A}$. We remark that this operation enforces symmetry. For seek of clarity, we specify for the verification of each conjecture the prompt adopted in Table 1.

Localization of Activated Zones This method is adopted in order to test conjectures 2.9 and 2.10 and more in general to investigate the zones activated by syntagma that involve logical operators. The technique adopted refers to a recent breakthrough method published in (Meng et al., 2022a), that, in order to identify decisive computations, isolates the causal effect of individual states within the network while processing a factual statement and therefore traces the path followed by information through the network. More specifically, this method, known as *casual tracing*, work by running a network multiple times, introducing corruptions to alter the computation, and then restoring individual states in order to identify the information that restores the results. In particular, carefully-designed traces are used to identify a specific small set of MLP module computations that mediate retrieval of associations. In our framework, we compute the *average indirect effect* (AIE)[2] over different positions in the sentence and different model components including individual states, MLP layers and attention layers. The main difference is that instead of studying the triples *(subject, relation, object)*, we analyse quadruples of the form *(subject, adjective, relation, adjective)*.

Test of	Adopted Prompt
Conjecture 2.1	*s is* a_1 *and* ...
Conjecture 2.2	*s is* a_1 *and* ...
Conjecture 2.3	*s is* a_1 *and* ...
Conjecture 2.4	*s is* a_1 *and* ...
Conjecture 2.5	*s is* a_1 *or* ...
Conjecture 2.6	*s is* a_1, *s is not* ...
Conjecture 2.7	*s is* a_1 *but* ...
Conjecture 2.8	*If s is* a_1, *then s is* ...

Table 1. This table provides the associations between prompts and conjecture chosen in the verification phase.

Model Editing This method is adopted to check our global comprehension. To modify individual operators within a GPT model, we exploit a variant of the method called *Rank-One Model Editing* (ROME) (Meng et al., 2022a). The idea is to treat an MLP module as a simple key-value store. ROME uses a rank-one modification of the MLP weights to directly write in a new key-value pair. In this model, we assume a linear view of memory within a neural network rather than an individual-neuron view. This linear perspective visualize individual memories as rank-one slices of parameter space.

The code used for the tests is available at https://github.com/alexandreduplessis/Deep-Learning.

4. Results

We start by presenting some of the results. Several more graphs covering all the statements claimed in this section are contained in the appendix B.

Results obtained through Analysis of Statistical Features Results obtained with this method can be clearly visualized through the matrix P. In particular, up to a properly-chosen

[2] The terminology was chosen in order to be consistent with the one adopted in (Meng et al., 2022a).

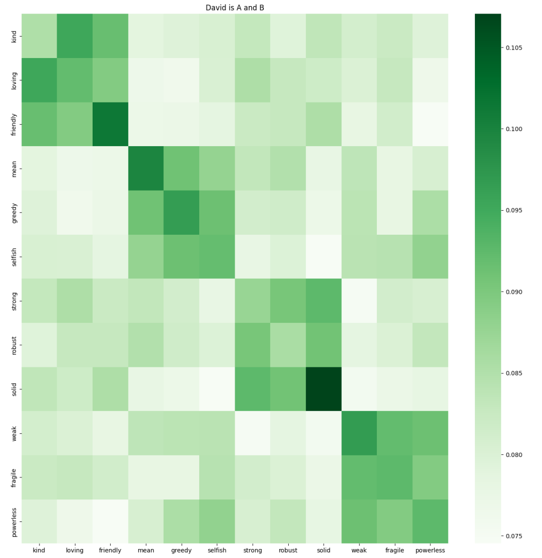

Figure 1. Probability distribution P for the study of equality in GPT-2 XL. The prompt of the study is *David is first adjective and second adjective.*

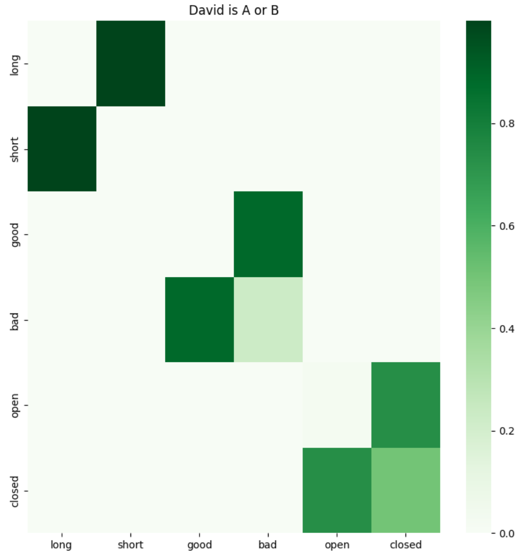

Figure 2. Probability distribution P for the study of equality in GPT-2 XL. The prompt of the study is *David is first adjective or second adjective.*

order on the axis, this matrix assumes peculiar shapes presenting blocks that can be easily interpreted according to the operator as explained below.

Results concerning Conjecture 2.2$\longrightarrow$ Empirically, the conjecture has been satisfied. This has been checked by considering a standard measure for symmetry (the sum of the absolute values of the difference of the entries symmetric to the diagonal and the entries of the diagonal itself) over the matrix encoding the probability distribution that shows that the matrix is pretty asymmetric. The representation is available in the appendix 6.

Results concerning Conjectures 2.3 $\longrightarrow$ We observe that from an empirical point of view the conjecture is satisfied. Observe that in Figure 1 adjectives with related meanings, e.g. the triple (*kind, loving, friendly*), form clusters that give the matrix the form of a block diagonal matrix. This particular shape of the matrix is obtained thanks to the choice of an appropriate ordering on the axes.

Results concerning Conjectures 2.4 $\longrightarrow$ Through the visualization of the matrix P, we can observe that conjunction *and* in GPT-2 XL plays the role of a synonym detector. This information enforce a logical role for this operator. Moreover, we observe that the maximal values in $P(a_1, a_2)$ lies on the main diagonal, i.e. they corresponds to the entries of the matrix of type $P(a, a)$.

Results concerning Conjecture 2.5$\longrightarrow$ We have a partial empirical confirmation. In particular, in Figure, we can note the level of masterization of the model when it comes to disjunctions of the form a *or not* a. In the same probability distribution, we can observe how adjectives referring to colours form a sort of cluster as if they were recognized as belonging to a unique class. More formally, given a prompt of the form s *is* a_1 where $s \in \mathcal{S}$ is a subject and $a_1 \in \mathcal{A}$ is a adjective indicating a colour, the $\arg\max_i(P(a_1, a_i))$ is given by the set $\{a_i \in \mathcal{A} | a_i \text{ is a color}\}$. We have considered this observation concerning classes of adjectives of potential interest, reason for which we have further investigated [3] the research question: Is GPT-2 XL able to classify qualification adjectives? If yes, is this classification compatible to the one adopted in standard English?
The results obtained empirically seems to reinforce this claim.

Results concerning Conjecture 2.6$\longrightarrow$ We find empirical confirmation of the conjecture, i.e. the probability distributions found associate an adjective to its antonym through the operator $\arg\max$. A visualization of this result is given in appendix 7.

Results concerning Conjecture 2.7$\longrightarrow$ We don't have a clear evidence concerning this result. See the appendix for details.

Results concerning Conjecture 2.8$\longrightarrow$ Empirical evidence seems to totally confirm our conjecture. We can visualize it properly in Figure 3.

Results obtained through Localization of Activated Zones Results obtained through causal tracing are visualized on graphs showing the impact of restoring attention heads and MLP layers.

[3] With an analogous method to the one used for weak equality.

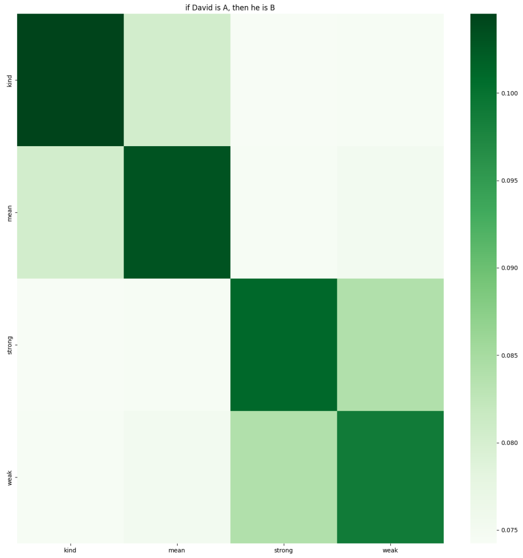

Figure 3. Probability distribution P for the study of equality in GPT-2 XL. The prompt of the study is *If David is first adjective, than second adjective.*

Results concerning Conjectures 2.9$\longrightarrow$ We have a partial confirmation of the conjecture that merits further experiments. See Figures 4, 5.

Results obtained through Model Editing Results obtained through model editing are described below. As already said, our model completes a given prompt, such as *Marc is tall and*, with adjectives close in meaning to *tall*. The aim is to try to edit the model in order to integrate *small* as part of the synonyms of *tall*. To do that, we have directly applied ROME algorithm to the key-value pair *(tall, small)* in the MLP layers identified in the localization part. This experiment shows that the application of ROME algorithm to update the *and* table has the hypothesized effect, i.e. the probability to obtain a completion with *small* increases significantly reaching values comparable to the ones of other synonyms.

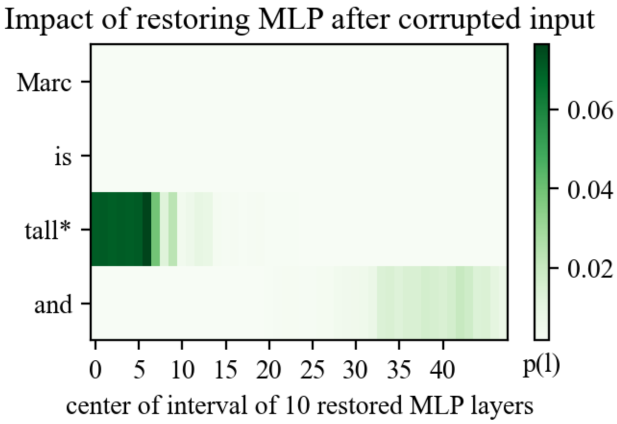

Figure 4. Localization Results for $\wedge$ operator (MLP).

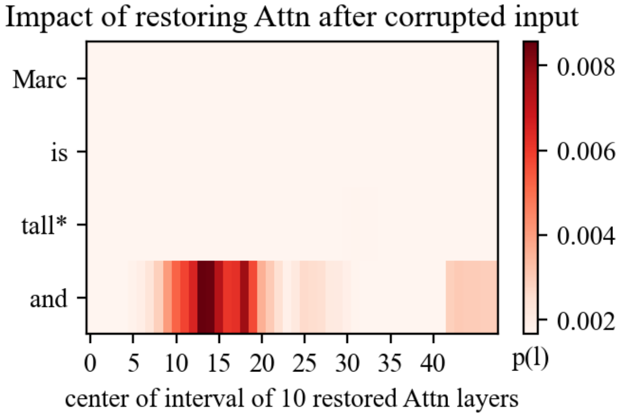

Figure 5. Localization Results for $\wedge$ operator (Attention).

5. Discussion & Conclusion

In this work, we interest ourself to the study of the encodings of some linguistic patterns (more specifically syntagma), more or less associated to logical operators in formal languages. We get empirical evidence that show how these architectures capture aspects of semantics of language, e.g. regarding synonyms. Moreover, we show how localization and editing techniques can actually shed light on the structure of the encoding of syntagma and potentially logical relevant distributions. Some experiments, especially those related to editing, could be easily improved by a rigorous study of the underlying probability distributions. This work still suffers of several weaknesses mainly related to the limited vocabulary on which the experiments were carried out. However, overall, the proposed approach offers a seemingly general framework that could be valid in broader models, e.g. GPT-J.

The investigation of logical operators, even if basic, could give ideas on how to edit LLM in order to improve their performance in logical tasks such as in theorem proving. Modifications could be done to force symmetry of $\wedge$ for instance.

There are still many open questions regarding the interpretability of LLM, we list below some of the most related to this work such as:

- Are the circuits investigated in this work common to all the syntagma, i.e. is there anything special concerning the statistical features relevant to logic?
- Which precise model can we propose for the encoding

of these relationships?

References

Areces, C., Hoffmann, G., and Orbe, E. Symmetries in modal logics. *Electronic Proceedings in Theoretical Computer Science*, 113:27–44, mar 2013. doi: 10.4204/eptcs.113.6. URL https://doi.org/10.4204%2Feptcs.113.6.

Bathaee, Y. The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31:889, 2018.

Bogaerts, B., Gocht, S., McCreesh, C., and Nordström, J. Certified symmetry and dominance breaking for combinatorial optimisation. In *AAAI*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Caldarini, G., Jaf, S., and McGarry, K. A literature survey of recent advances in chatbots. *Information*, 13(1), 2022. ISSN 2078-2489. doi: 10.3390/info13010041. URL https://www.mdpi.com/2078-2489/13/1/41.

Cinque, G. The fundamental left-right asymmetry of natural languages. 2009.

Cohen, S. N., Snow, D., and Szpruch, L. Black-box model risk in finance, 2021. URL https://arxiv.org/abs/2102.04757.

Creswell, A., Shanahan, M., and Higgins, I. Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022. URL https://arxiv.org/abs/2205.09712.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning, 2022. URL https://arxiv.org/abs/2207.07051.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks, 2021. URL https://arxiv.org/abs/2106.02997.

Ghoniem, A. and Sherali, H. Defeating symmetry in combinatorial optimization via objective perturbations and hierarchical constraints. *IIE Transactions*, 43:575–588, 08 2011. doi: 10.1080/0740817X.2010.541899.

Hendrycks, D. and Mazeika, M. X-risk analysis for ai research, 2022. URL https://arxiv.org/abs/2206.05862.

López-Astorga, M. Interpretation and use of disjunction in natural language: A study about exclusivity and inclusivity. *Revista Lengua y Habla*, 25:24 – 33, 2021. ISSN 2244811X. URL https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=155101973&site=ehost-live.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt, 2022a. URL https://arxiv.org/abs/2202.05262.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer, 2022b. URL https://arxiv.org/abs/2210.07229.

Polu, S. and Sutskever, I. Generative language modeling for automated theorem proving, 2020. URL https://arxiv.org/abs/2009.03393.

Popovic, M. and Castilho, S. Are ambiguous conjunctions problematic for machine translation? 09 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019a.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019b.

Sambin, G., Battilotti, G., and Faggian, C. Basic logic: Reflection, symmetry, visibility. *The Journal of Symbolic Logic*, 65(3):979–1013, 2000. ISSN 00224812. URL http://www.jstor.org/stable/2586685.

Veres, C. Large language models are not models of natural language: they are corpus models, 2021. URL https://arxiv.org/abs/2112.07055.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Wei, D., Nair, R., Dhurandhar, A., Varshney, K. R., Daly, E. M., and Singh, M. On the safety of interpretable machine learning: A maximum deviation approach, 2022. URL https://arxiv.org/abs/2211.01498.

Wu, Y., Jiang, A. Q., Li, W., Rabe, M. N., Staats, C., Jamnik, M., and Szegedy, C. Autoformalization with large language models, 2022. URL https://arxiv.org/abs/2205.12615.

Zhang, A., Xing, L., Zou, J., and Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature biomedical engineering*, July 2022a. ISSN 2157-846X. doi: 10.1038/s41551-022-00898-y. URL https://doi.org/10.1038/s41551-022-00898-y.

Zhang, H., Li, L. H., Meng, T., Chang, K.-W., and Broeck, G. V. d. On the paradox of learning to reason from data, 2022b. URL https://arxiv.org/abs/2205.11502.

A. Vocabulary

We specify below the vocabulary $\mathcal{W}$ used for the experiments. This is divided in a few subsets, the one of subjects ($\mathcal{S}$), the one of adjectives ($\mathcal{A}$) and finally the one of operators and constructs investigated ($\mathcal{O}$).

$$\mathcal{S} := \{\text{"Georges", "Mark", "David", "The phone", "The skyscraper", "The lunch"}\}$$

$$\mathcal{A} := \{ \text{"kind", "loving", "friendly", "mean", "greedy", "selfish", "strong", "robust", "solid", "weak", "fragile", "powerless", "easy", "effortless", "trivial", "complicated", "difficult", "complex", "blue", "red", "green", "black", "white","heavy", "light", "big", "small", "short", "long", "hot", "cold","wet", "dry", "cheap", "expensive", "free", "beautiful", "repulsive", "strong", "weak", "open", "closed", "generous", "selfish", "good", "bad", "clean", "dirty", "full", "empty", "far", "close", "noisy", "quiet"}\}$$

$$\mathcal{O} := \{ \text{"not","and", "or", "but", "if-then"}\}$$

B. Further results

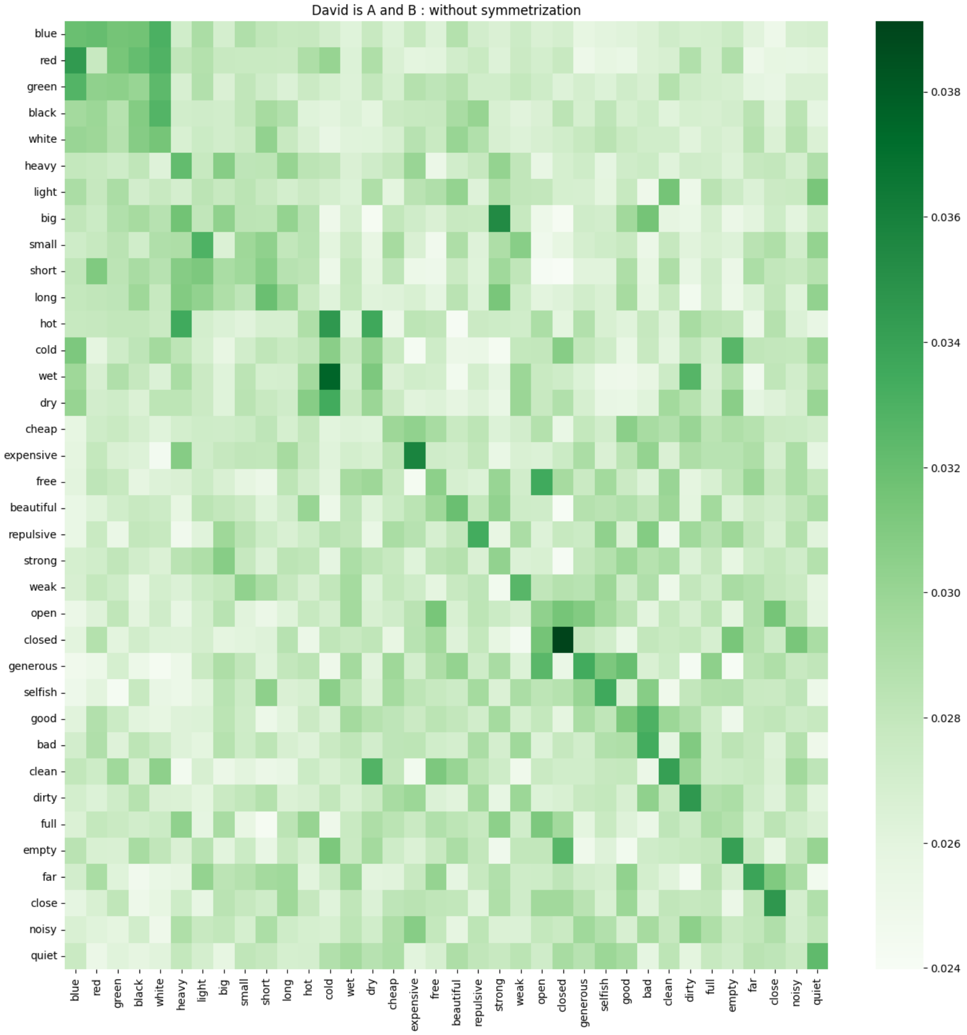

Figure 6. Analysis of the Symmetry In this graph, we investigated the matrix $P(a, b)$ associated to the operator $\wedge$ before being symmetrized.

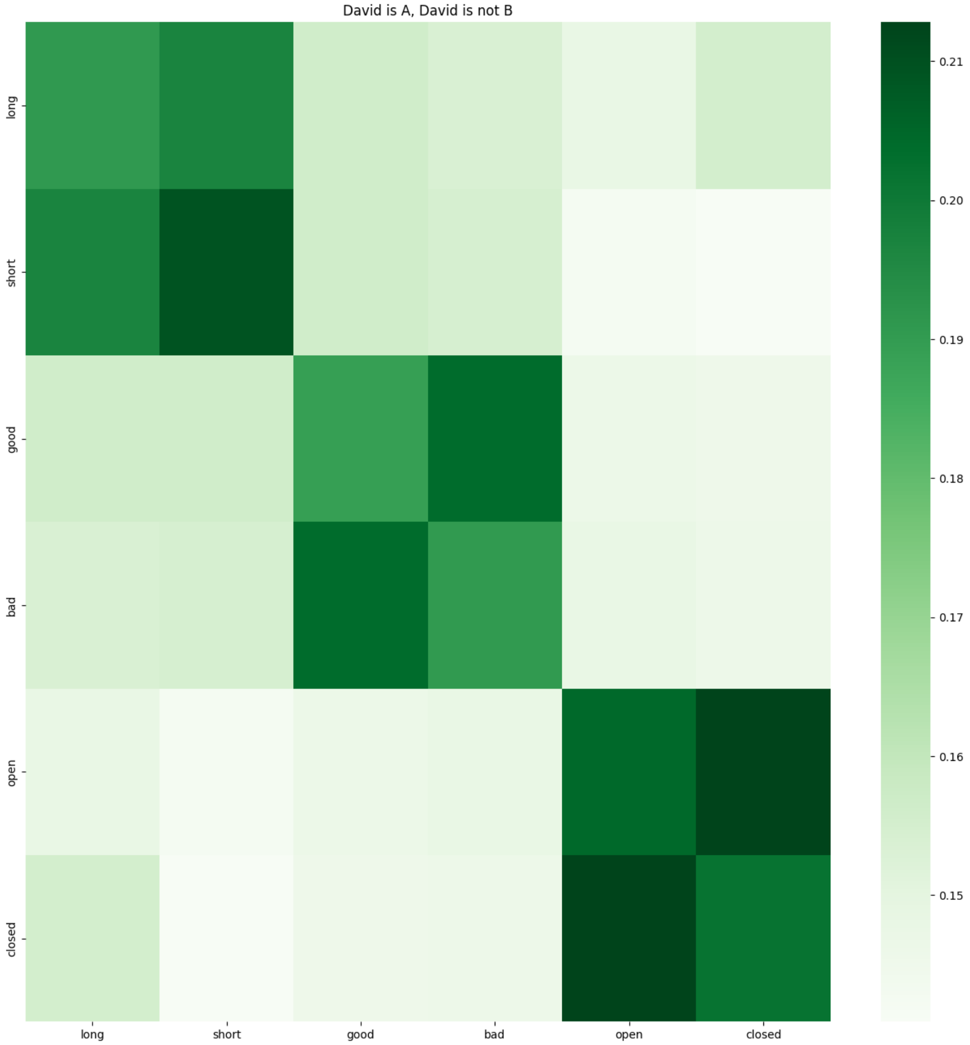

Figure 7. Probability distribution P for the study of equality in GPT-2 XL. The prompt of the study is *David is first adjective and David is not second adjective.*