Where does meaning live? Investigating the synthetic-analytic distinction in LLMs using gender as a case study

Anonymous ACL submission

Abstract

Some linguistic inferences–e.g., inferring that a square has four sides–seem to follow inherently from what words mean, while others–e.g., inferring that a house has four sides–are considered to follow from "common sense" or "world knowledge". It has long been debated whether such categorical distinctions, referred to in philosophy as analytic vs. synthetic, can be made and what effect they should have on theories and models of semantic meaning. In this paper, we use gender (male vs. female) as a case study to explore whether large language models (LLMs) differentiate analytic inferences about gender (e.g., that a woman is female) from synthetic inferences (e.g., that nurses are most often female). We find that, by and large, there are not substantial mechanistic differences, but rather the difference appears to be a matter of degree–i.e., how strongly the inference is encoded and how easily it is overwritten by contextual information. Our study serves as a proof-of-concept for how LLMs can be used to revisit long-standing questions about language representation and processing in general.

1 Introduction

In the philosophy of language, semantics, and computational linguistics, a distinction is often made between *synthetic* and *analytic* aspects of meaning (Rey, 2023). Here, *analytic* refers inferences that are inherently true given the meaning of a word (e.g., that a *square* is *four-sided*) while *synthetic* refers to properties that are perhaps inferred from common sense or life experience (e.g., one might infer that a *house* is likely *four-sided*, but that is in no way required by the meaning of the word *house*). There has long been debate about the extent to which this distinction is real, or whether there is a difference between synthetic and analytic properties in terms of how they are stored and processed. Large language models (LLMs), which exhibit near-human ability to generate and process text, allow us to study this distinction in empirical rather than philosophical terms. Using gender (male vs. female) as a case study, we ask whether LLMs invoke different mechanisms when gender information is presumptively analytic (e.g., the inference that *woman* is female is built into the English language) vs. synthetic (e.g., the inference a *nurse* is likely female comes from world knowledge, not from semantics *per se*). We focus on the analysis of GPT-2 family (Radford et al., 2019) [1] of models, and investigate the mechanisms used to predict pronouns (he vs. she) and names for a variety of types of words that indicate gender (explicitly gendered nouns, names, professions, etc). We find that, by and large, there are no substantial mechanistic differences between how synthetic vs. analytic inferences about gender are processed, but there are differences of degree. That is, in almost all cases, gender information is primarily stored in the word embeddings, and differences stem chiefly from how strongly the bias is encoded and how easily it is overwritten by contextual information. Our work serves as a proof of concept for how studying mechanisms in LLMs can inform the study of language more broadly and has practical implications for work on debiasing LLMs (see Discussion §4).

2 Dataset

We curate a set of 20 grammatically gendered nouns (e.g., *man, woman*) each for male and female, and a subset of 40 profession nouns from (Vig et al., 2020) which have strong gender associations (e.g., *doctor, nurse*). We also construct a set of 14 templates that are designed to be gender-neutral and bias the model toward producing a pronoun to continue the sentence. The dataset follows the format of "The {*noun*} {*verb*} that" or "The {*noun*} {*verb*} because". We switch the {*noun*} with ex-

[1] We focus on GPT2-medium in the main paper. GPT2 small and large are included in the appendix

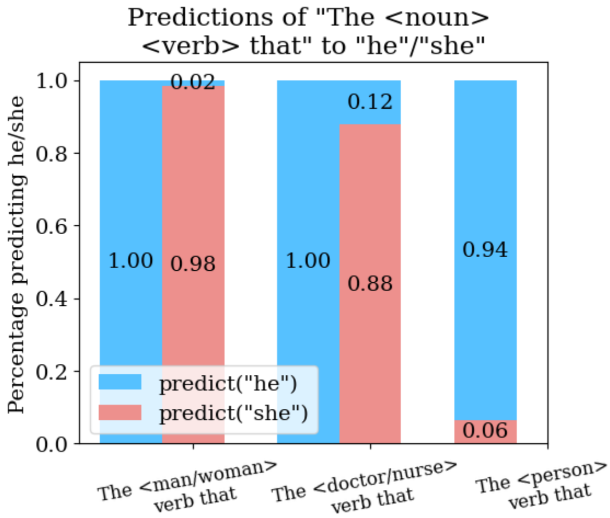

Figure 1: Percentage of predicting "he/she" for explicitly gendered nouns, profession, and gender-neutral nouns. The left bars are stereotypical male gender nouns and the right bars are for female nouns. Blue means preferring "he" over "she" and pink means vice versa.

plicitly gendered nouns and professions. The full list of the templates and nouns can be found in Appendix A.

3 Experiments and Results

Strength of Inference: If there is a difference between how the LLM processes synthetic vs. analytic gender inferences, we might expect to see that words like *woman* take female pronouns without exception, while words like *nurse* show a more balanced mix of pronouns. Thus, we first compare the consistency of pronoun predictions for two templates: "The {profession word} *verb* that" vs. "The {gendered word} *verb* that". We calculate the consistency by examining the probability difference between the tokens "he" and "she" at the final layer. Consistently positive difference implies the model favors "he", and negative implies favoring "she".

Our results are in Figure 1. Note that, at baseline, the model has a strong bias for "he" over "she": 93% of the time the model will predict "he" given the templates populated with neutral nouns (*person, child, member*).[2] Overall, the results are in line with our expectations. The explicitly gendered words' predictions are highly consistent; in all 40 words (20 male and 20 female), 39 of them exhibit perfect consistency. All the definitionally male nouns prefer "he" over "she". Among the female nouns, 19 of them prefer "she" over "he". The only exception is the word *miss*.[3] We speculate that this is due to the fact that *miss* is rather rare to be used by itself as a noun, and thus the LLM might not have learned a strong gender signal.

In the case of profession nouns, the pronoun predictions are more dependent on the template. Among twenty female profession nouns (Vig et al., 2020), seven of them (*clerk, secretary, teacher, therapist, stylist, hairdresser, violinist*) show high variance depending on the verb that appears in the template. For example, all show a preference for "he" in the template 'The $\{noun\}$ $drove$ because'. We speculate that this is because the verb $drove$ has a stronger male gender signal, overriding the signal sent by the profession nouns.

Location of Gender Information: We might expect that analytic inferences are encoded on the word itself (e.g., femaleness is part of the context-independent meaning of *woman*) while synthetic inferences might occur later in processing, as part of contextual inference. We thus investigate where in the model (at which layer) the inference about gender is made. We use Minimum Description Length (MDL) (Voita and Titov, 2020), which intuitively captures how accurately a feature can be decoded and the amount of effort required to decode it (i.e., the *codelength*). We compute codelength for predicting the (assumed) gender of a word for every hidden state to determine where the model most readily commits to the inference that a given noun, e.g., *nurse* or *woman*, is female.

Figure 2 shows MDL [4] at each layer for two tokens: 1) the last token in the template and 2) the token corresponding just to the noun of interest. We see that, for both explicitly gendered words and profession words, the codelength (right) drops sharply to near 0 after the first layer, suggesting that the inference about gender is readily encoded within the embedding of the noun itself both for analytic inferences about gender as well as for inferences which should be synthetic. However, when we look at the MDL at the final token in the sentence (a way of approximating the inference made over the whole sentence), it is more difficult to extract the gender representation for the profession nouns compared to the explicitly gendered nouns. To investigate this further, we employ early decoding (nostalgebraist,

[2]The exceptions might be due to gender-biased verbs. E.g., the model predicts "she" when the template includes *cried*.

[3]In "The miss said that", "The miss yelled that", "The miss ran because", "The miss drove because", "he" is more probable than "she".

[4]The detailed description is shown in E

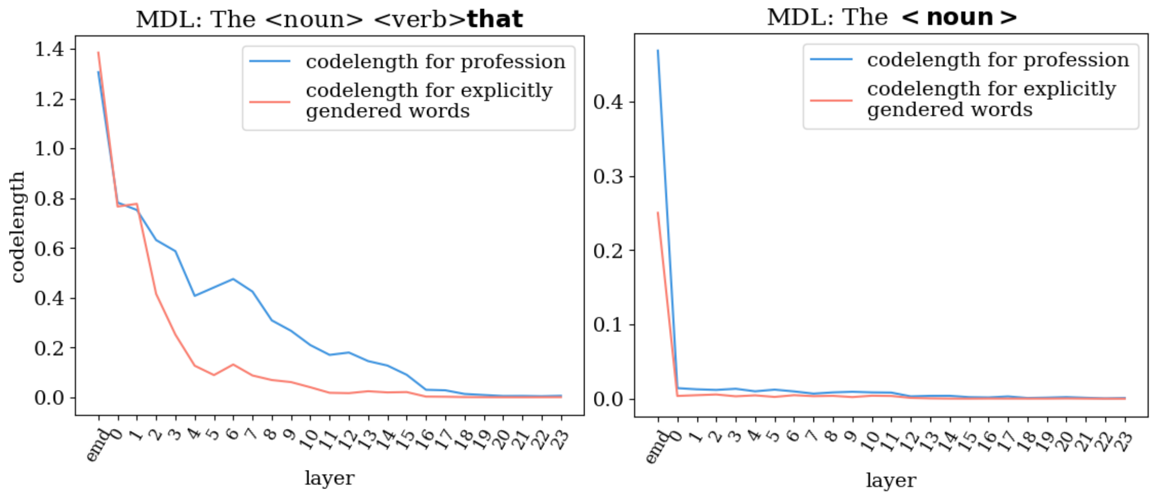

Figure 2: Codelength of probes to differentiate gender information. The left graph (covering 560 examples) decodes the hidden states on the final tokens of "that". The right graph (covering 40 examples) decodes the hidden states on the tokens of the *noun* directly. Probes at every layer are trained for 20 epochs.

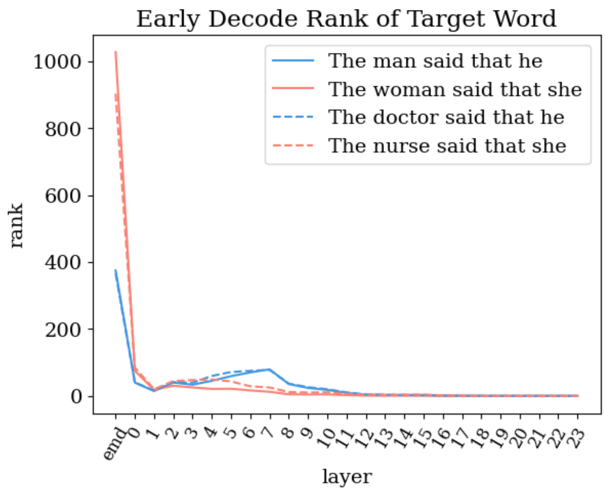

Figure 3: Rank of the pronouns through early decoding.

2020) to determine at which layer the model commits to the pronoun prediction ("he" vs. "she"). For both analytic and synthetic inferences, the model appears to form the pronoun predictions around the same layer at inference time in Figure 3. That is, it takes the same number of layers for both "woman" and "nurse" to build up the meaning of the female and generate the prediction of "she".

Together, these results imply that there is no distinction in the lexicon between analytic vs. synthetic inferences about gender, nor in how quickly (in terms of number of layers) the inferences are made. However, synthetic inferences might interact differently with contextual information, perhaps because they are more readily overwritten by competing semantic cues.

Circuit Analysis: We attempt to drill down further and localize the components that process gender information for each type of inference. To do this, we employ causal mediation analysis (Vig et al., 2020; Chan et al., 2022; Geiger et al., 2021, 2023; Meng et al., 2023; Wang et al., 2023; Chan et al., 2023; Cohen et al., 2023; Merullo et al., 2024). The contrasting pairs are formed by "The {male nouns} verb that" and "The {female nouns} verb that" for the clean and corrupted inputs, respectively (see Vig et al. (2020) for a full description of the method). We perform the patching experiments manually and also utilize automated circuit discovery from Conmy et al. (2023); Bills et al. (2023); Syed et al. (2023); Hanna et al. (2024) to obtain the top 50 edges in the computation subgraphs.

We compute top components (e.g., attention heads and MLPs) that are involved in the pronoun prediction for the analytic and synthetic gender inferences, as well as those that are involved with the prediction of pronouns in neutral contexts. We examine whether there are components that are uniquely active in the computation of synthetic or analytic gender inferences, which are not explained by their involvement in pronoun prediction more generally. Figure 4 shows our results. We find that the circuits are highly overlapping but not identical (Jaccard similarity between these two circuits is 0.73)[5]. While much of the overlap is due to components that are involved in the general pronoun

[5]As a control: we find a 0.09 Jaccard similarity with the IOI circuit (Wang et al., 2023).

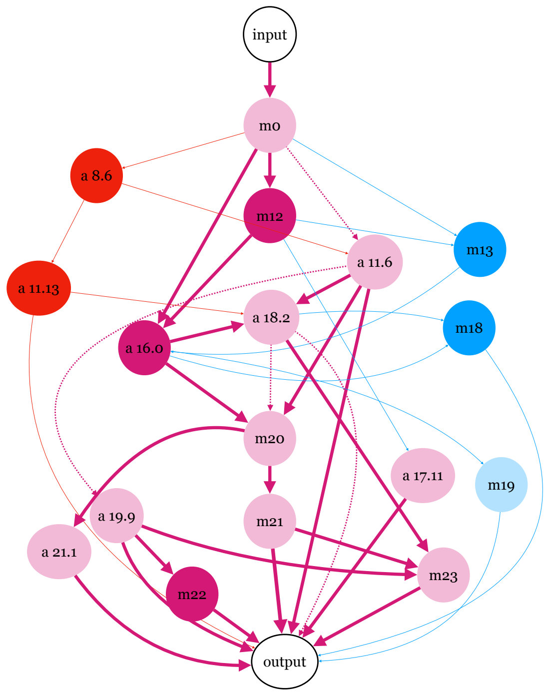

Figure 4: Abstraction of circuit overlap across the dataset. The pink nodes appear in both the synthetic and analytical noun circuits. The blue nodes appear only in the explicitly gendered nouns while the red nodes appear only in the profession nouns. We compare with the input "The <*noun*> and me said that" which predicts the pronoun of they-we. The lighter color (pink, blue) nodes also appear in the circuit that predicts they-we in Table 1 while the darker color nodes only appear in the synthetic/analytical noun circuits.

prediction case, some are unique to the gendered inference. Importantly, we also find two attention heads which are only involved in the processing of explicitly gendered words, and two MLPs which are only involved in the processing of profession nouns. While it is too early to draw strong conclusions, this presents an interesting avenue for future work, as it might be suggestive of different mechanisms governing analytic vs. synthetic inferences.

4 Discussion

Our analyses suggest that, within LLMs, the synthetic-analytic distinction is less of a categorical distinction than variation along a continuum. Specifically, we find evidence that inferences about gender, whether categorical or analytic, are stored primarily in the embeddings (i.e., the lexicon) and that the model does not require any more processing (i.e., the number of layers) to make synthetic inferences compared to analytic ones. That said, we do see consistent evidence that synthetic inferences are encoded less strongly (measured by MDL) and are more easily overwritten, e.g., when other words in the context carry competing signals about gender. However, our circuit analysis, while preliminary, does suggest that there might be different computational units involved in the processing of synthetic vs. analytic inferences. Further work could yield significant revision to our above interpretation, possibly providing evidence of a more explicitly categorical difference between how these inferences are processed.

Our findings have practical implications for work on debiasing LLMs. The high similarities in these two mechanisms suggest that it might not be possible to remove synthetic inferences about gender (which are generally deemed "bias") without damaging analytical inferences (which are necessary for correct English language generation). As many existing debiasing methods intervene with gender information by either fine-tuning the model weights or editing the representations at inference time, our analysis suggests that will hurt the performance of analytical gendered nouns since the weights and representations are shared.

5 Related Work

Our work contributes to a recent line of work that asks if and how LLMs can inform the study of language and cognition more broadly (Mahowald et al., 2024). Often, arguments are made that LLMs inform linguistic theory by serving as wholesale replacements for existing explanatory models (Piantadosi, 2023). Our proof of concept study aligns with an alternative position, arguing that understanding of the mechanisms in play in LLMs can lead to refinement, rather than replacement, of existing theories (Pavlick, 2023; McGrath et al., 2023).

Our experiments are also highly related to work on gender bias in LLMs. Many previous efforts (Stanczak and Augenstein, 2021) have been made in identifying gender bias as well as intervening in gender bias in language models. Approaches include modifying the training data (Guo et al., 2022; Ranaldi et al., 2023), intervening on the word embeddings (Kaneko and Bollegala, 2019), fine-tuning specific parts of the model (Lauscher et al., 2021; Gira et al., 2022; Xie and Lukasiewicz, 2023), or employing model-editing and causal mediation techniques (Belrose et al., 2023; Ravfogel et al., 2022, 2020; Cai et al., 2024; Chintam et al., 2023; Limisiewicz et al., 2024).

6 Limitation

Our work aims to compare how LMs process synthetic and analytical inferences. However, the conclusion is limited to a few specific datasets based on gender information. Moreover, the analysis only covers the GPT2 series of models. Therefore, the conclusion drawn is yet limited and can be expanded upon models with larger sizes and a more diverse range of data. The results can be supported by more evidence that causally explains our observations on the strength of inference. We would like future work to extend the analysis beyond the case of gender and propose new debias methods based on our results.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *Preprint*, arXiv:2306.03819.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. *Preprint*, arXiv:2403.14409.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2023. Causal scrubbing: a method for rigorously testing interpretability hypotheses [redwood research]. Alignment Forum. Accessed: 17th Sep 2023.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *Preprint*, arXiv:2307.12976.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586.

Atticus Geiger, Christopher Potts, and Thomas Icard. 2023. Causal abstraction for faithful model interpretation. Ms., Stanford University.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *Preprint*, arXiv:2403.17806.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomasz Limisiewicz, David Mareček, and Tomáš Musil. 2024. Debiasing algorithm through model adaptation. *Preprint*, arXiv:2310.18913.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Preprint*, arXiv:2301.06627.

Sam Whitman McGrath, Jacob Russin, Ellie Pavlick, and Roman Feiman. 2023. How can deep neural networks inform theory in psychological science?

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. *Preprint*, arXiv:2310.08744.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Steven T Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz/007180*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2023. A trip towards fairness: Bias and de-biasing in large language models. *Preprint*, arXiv:2305.13862.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *Preprint*, arXiv:2004.07667.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. *Preprint*, arXiv:2201.12091.

Georges Rey. 2023. The Analytic/Synthetic Distinction. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2023 edition. Metaphysics Research Lab, Stanford University.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *Preprint*, arXiv:2112.14168.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *Preprint*, arXiv:2310.10348.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *Preprint*, arXiv:2003.12298.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15730–15745, Toronto, Canada. Association for Computational Linguistics.

A Template

Templates "The { } said that"
"The { } yelled that"
"The { } whispered that"
"The { } wished that"
"The { } ate because"
"The { } ran because"
"The { } drove because"
"The { } slept because"
"The { } cried because"
"The { } laughed because"
"The { } went home because"
"The { } stayed up because"
"The { } yelled because"

Explicitly Gendered Nouns (Male) ’man’, ’boy’, ’father’, ’brother’, ’son’, ’uncle’, ’nephew’, ’grandfather’, ’grandson’, ’husband’, ’boyfriend’, ’groom’, ’gentleman’, ’sir’, ’mister’, ’prince’, ’king’, ’god’, ’lad’, ’sir’

Explicitly Gendered Nouns (Female) ’woman’, ’girl’, ’mother’, ’sister’, ’daughter’, ’aunt’, ’niece’, ’grandma’, ’granddaughter’, ’wife’, ’girlfriend’, ’bride’, ’lady’, ’miss’, ’maid’, ’princess’, ’queen’, ’goddess’, ’widow’, ’mistress’

Neutral "individual", "human", "being", "child", "adult", "resident", "participant", "member", "friend", "neighbor", "partner", "peer"

Profession (Male) ’assassin’, ’astronaut’, ’bodyguard’, ’boxer’, ’butcher’, ’carpenter’, ’coach’, ’colonel’, ’commissioner’, ’custodian’, ’electrician’, ’farmer’, ’janitor’, ’mathematician’, ’minister’, ’doctor’, ’president’, ’sailor’, ’warden’, ’warrior’

Profession (Female) ’socialite’, ’librarian’, ’clerk’, ’ballerina’, ’dancer’, ’nanny’, ’whore’, ’nun’, ’nurse’, ’secretary’, ’receptionist’, ’teacher’, ’therapist’, ’violinist’, ’housekeeper’, ’hooker’, ’paralegal’, ’stylist’, ’housekeeper’, ’hairdresser’]

B Model Sizes

C Pronoun Circuit Graph from GPT2 Models

We extend the analysis beyond GPT2-medium to GPT2-small and GPT2-large. Regardless of the size of the model, the overlap in the shared components remains high. In GPT2 small, the Jaccard Similarity between circuit components is 0.68, and 0.71 for GPT2-large.

D Name Circuit Graph for GPT2 Models

In the main paper, we focus on “The {boy/girl} *verb* that” and “The {doctor/nurse} *verb* that”. We extend the analysis beyond the prediction of pronouns. We created the contrasting pairs-“The {boy/girl}’s name is” and “The {doctor/nursel}’s name is” querying for names in Table 1: 2 and 3. In the prediction of names, there is also a high overlap of pink nodes in Figure 8. The similar mechanisms in synthetic and analytical words is not a special case in predicting pronoun but also in predicting names as well.

E Minimal Description Length

Formally, we separate the training data into N subsets of equal size t. We train a series of linear classifiers $p_i(y|x)$ by giving the first i subsets of data for a fixed number of epochs. We calculate the cross entropy loss on $p_i(y|x)$ on a held-out test set. A model that performs well with a limited number of training examples will be rewarded by a lower *codelength*.

$$-\sum_{i=1}^{N} \log_2 p_i(y_{it+1:i(t+1)}|x_{it+1:i(t+1)}) \qquad (1)$$

	input	output	gender	pronoun	input type
0	The {boy/girl} *verb* that	he/she	yes	yes	analytical
1	The {doctor/nurse} *verb* that	he/she	yes	yes	synthetic
2	The {boy/girl}'s name is	gendered names	yes	no	analytical
3	The {doctor/nurse}'s name is	gendered names	yes	no	synthetic
4	The {doctor} and {(he/she)/me} *verb* that	they/we	no	yes	/
5	The {boy} and {(he/she)/me} *verb* that	they/we	no	yes	/

Table 1: detailed dataset examples

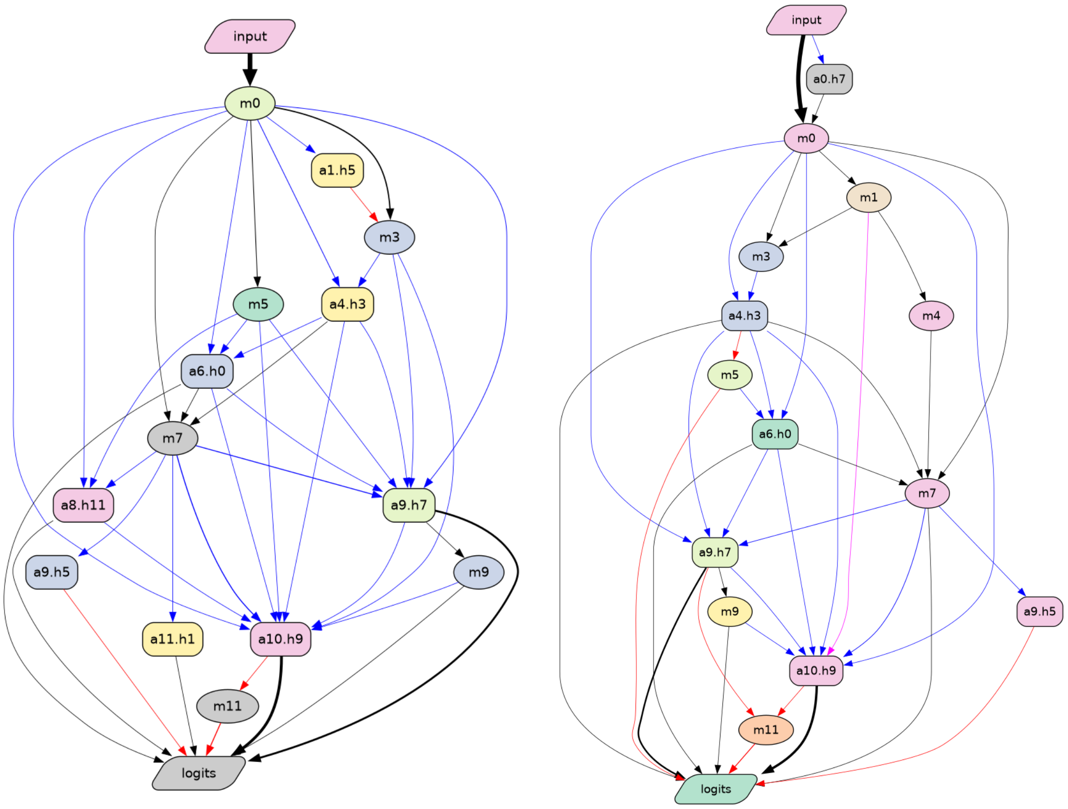

Figure 5: Circuit graph for GPT2 small. Left: explicitly gendered noun. Right: profession noun

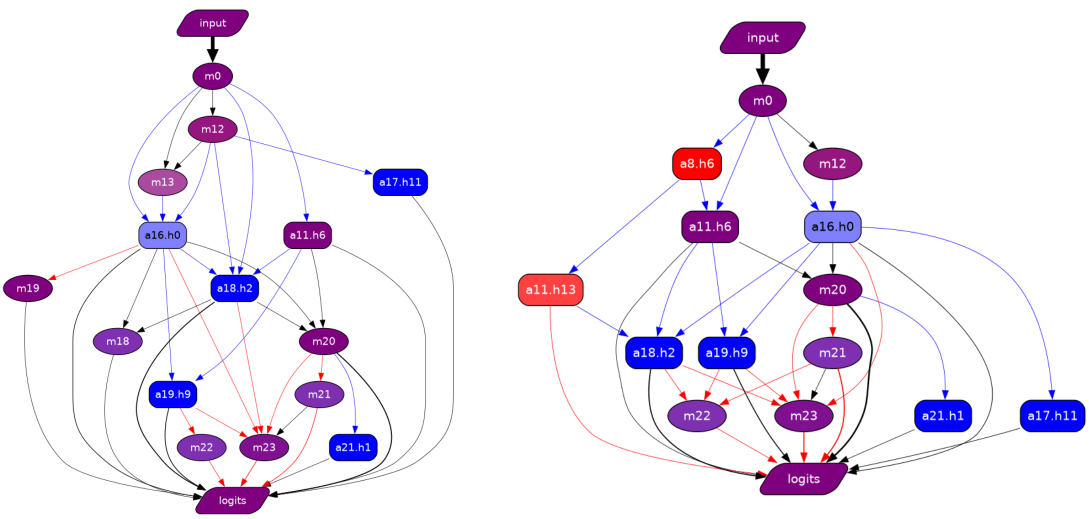

Figure 6: Circuit graph for GPT2 medium. Left: explicitly gendered noun. Right: profession noun

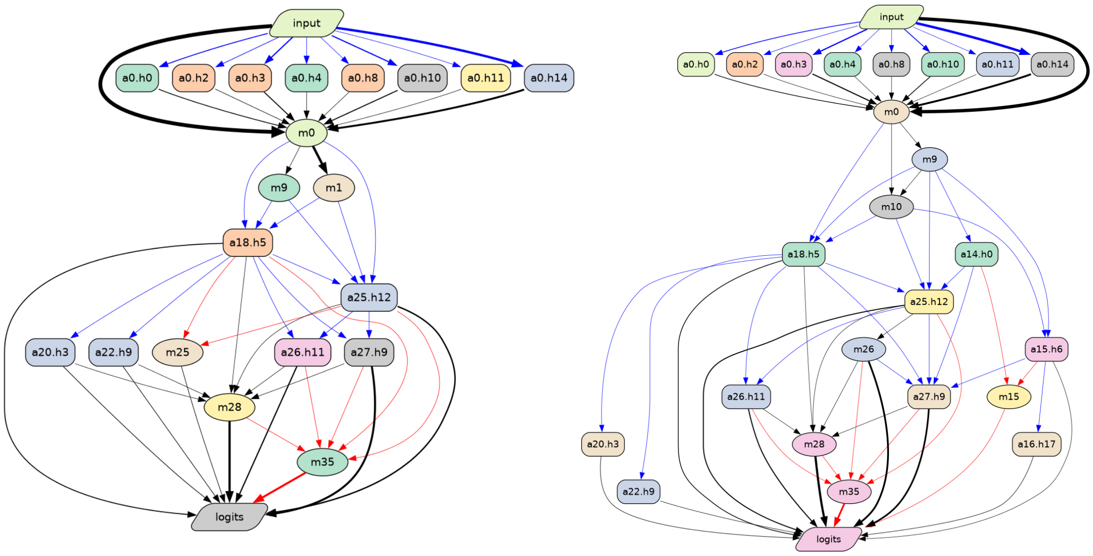

Figure 7: Circuit graph for GPT2 large. Left: explicitly gendered noun. Right: profession noun

Model	Parameters	Layer	Heads
GPT2-small	117m	12	12
GPT2-medium	335m	24	16
GPT2-large	762m	36	20

Table 2: model sizes

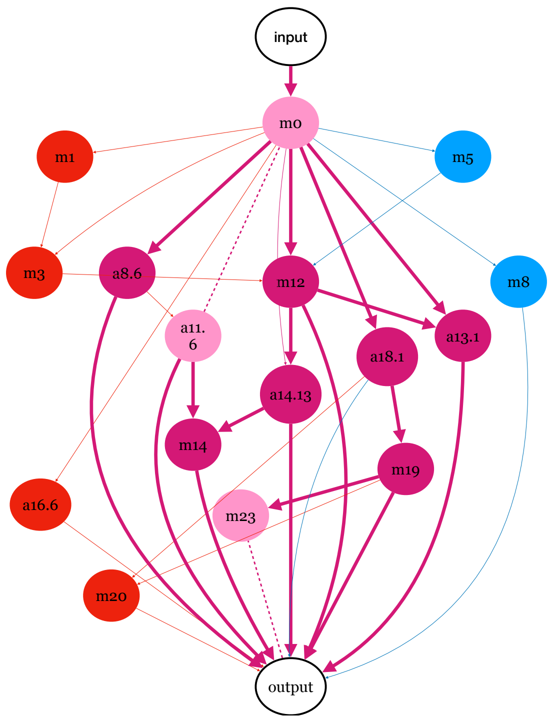

Figure 8: Circuit abstraction overlap in the prediction of names. The pink nodes in the middle appear in both the synthetic and analytical noun circuits. The blue nodes appear only in the explicitly gendered nouns while the red nodes appear only in the profession nouns. We compare with the input “The <noun> and me said that” which predicts the pronoun of they-we. The light pink nodes also appear in the circuit that predicts they-we 1 while the dark pink nodes only appear in the synthetic/analytical noun circuits.