MarxistLLM: Fine-tuning a language model with a Marxist worldview

Matti Nelimarkka*

Social Computing Group
Centre for Social Data Science
University of Helsinki, Finland

Department of Computer Science
Aalto University, Finland

Abstract

Recently, there has been increasing interest in using language models to examine social aspects through social science lenses. However, these models embody specific societal values and perspectives. This inherent encoding of viewpoints means that when these models are employed in social science research, scholars might not account for the biases that these cause. Alternatively, these biases could be seen as theoretical perspectives seen as a world-view (*Weltanschauung*) , such as rational choice theory, Marxist theory, and feminist theory. To highlight potential biases with models and implications for research, and explore the opportunity to use language models for social science theory, we fine-tune large language models with a specific *Weltanschauung*. Specifically, we incorporate the writings of Marx and Engels to fine-tune these models, aiming to infuse them with Marxist ideological terminology and a Marxist worldview. We evaluate how these fine-tuned models differ in empirical analysis and show that affluence loses its importance with a Marxist language model, but at the same time the model is more sensitive to capitalism and economy than a standard baseline model. This investigation underscores the need to examine the values embedded in language models before they are used for social science research and highlights the opportunity to incorporate a particular theoretic stance into them.

1 Introduction

Social scientists are increasingly using computational methods, including machine learning and natural language processing, in the analysis of large-scale

*Electronic address: matti.nelimarkka@helsinki.fi; Corresponding author

data. Within this domain, the most recent development has been in using large language models developed by collecting and analysing word relations in massive datasets, leading to an increasingly significant role of a foundation mode for research activity. Word-embedding models have been used to develop measurement scales for text data (Kozlowski et al., 2019). Supervised and unsupervised classifiers based on word-embedding models, such as BERTTopics (Grootendorst, 2022), have become increasingly common. Most recently, interest has grown around generative language models, with (debated) suggestions that they can be used for content analysis (for review, see Ollion et al., 2024) and data generation (Argyle et al., 2023; Dillion et al., 2023).

However, applying language models in social science is not without novel problems in validity and reliability. Large language models are not neutral but incorporate social values and, ultimately, a view of the world (among others, Fisher et al., 2020; Garg et al., 2018; Watson et al., 2023; Garg et al., 2018; Gallegos et al., 2023; Salinas et al., 2023). For example, some large language models appear to contain a more traditional conceptualisation of gender roles (e.g., Watson et al., 2023), which means that the results produced with such models would also contain this worldview. Therefore, applications of language models should examine how such biases influence the outcomes of analysis.

At the same time, these biases relate to a specific conceptualisation of social theory: a worldview or *Weltanschauung*. Abend (2008) elaborates that such theories "are not about the social world itself, but about how to look at, grasp, and represent it." He highlights that such social theory is incorporated into all social sciences, either implicitly or explicitly. As computational text analysis in social sciences is in search of more theoretic and conceptual thinking (Baden et al., 2022), focusing on the world-views of language models may open up avenues to embrace theories.

Our work illustrate both the challenge of non-neutral large language models and the opportunities to integrate a world-view-as-theory into large language model-based research. We fine-tune an existing model with texts from Karl Marx to incorporate a Marxist perspectives into the model. Fine-tuning is a process in which a foundational model is further developed by giving it examples of data to identify statistical relationships from. We first illustrate how the outputs of the MarxistLLM differ from the base model to show the impact of fine-tuning, both for causal tasks (i.e. content generation) and mask-filling tasks (i.e. predicting the most likely words given the context). Following this, we use these models in three case studies, replication of the Geometry of Culture approach (Kozlowski et al., 2019), elaboration of different occupations, and topic modelling, to illustrate the differences in analytical capability between fine-tuned and non-fine-tuned models. We conclude this work by discussing how social scientists using word embedding and other large language models should work to acknowledge biases, values, and world-views.

2 Background

We first review the literature examining biases and values incorporated into fundamental models used in social science research. We show that while the standard application approach in social sciences does not engage with biases and values of the analytical tools, their existence is well known in natural language processing and computing literature. Following this section, we explore the meaning of social theory in computational text analysis for social science. We show how the biases and values examined previously can be seen as a theoretic stance. Finally, as our work focuses on Marxist theory, we provide a brief introduction to this theory and our focus within it.

2.1 Biases, values and fundamental models

Fundamental models, including large language models such as OpenAI's GPT, Google's BERT, and Facebook's LLama, are trained using massive data collections from the Web, books, news stories, and similar kinds of materials. This has led to several ethical and legal concerns and also extensive work from the model developers to "clean" their training data, such as removing personal information and content deemed inappropriate. For example, the collected data contains sexist, xenophobic, and homophobic content, and several efforts exist to remove those from the training data. In addition, scholars have developed various technical mitigation approaches to limit the impact of biases on tasks such as classification (e.g., Papakyriakopoulos et al., 2020; Fisher et al., 2020). Nevertheless, extensive work in natural language processing has shown that these efforts are not comprehensive; rather, various forms of prejudices and biases remain in the data and thus creep into the fundamental models.

For example, fundamental models reproduce stereotypical gender roles (Fisher et al., 2020; Garg et al., 2018). Watson et al. (2023) show that the BERT predictions most closely resemble survey responses from those with moderate to conservative views on the role of women in society. Similarly, they show that the predictions on non-binary terminologies resemble survey responses from people with low-to-moderate non-binary acceptance.

Beyond gender, researchers have observed similar biases related to various other attributes, including ethnicity, age, race, sexual orientation, and religion, as well as intersections of these attributes (Garg et al., 2018; Gallegos et al., 2023; Salinas et al., 2023). These might even change over time (Garg et al., 2018). In some cases, it might be ideal to approach these as a window to culture, for example when examining stereotypes or discrimination as present in public discussion (Garg et al., 2018; Kozlowski et al., 2019). In other cases, this might produce measurement mistakes that we would prefer not to have in the analysis because they replicate existing biases in the data analysis. To sum up, any attempt to quantify society using these techniques is likely to represent these biases.

Even with a widely shared understanding of such issues, social scientists who use these approaches to "measure" society through text rarely engage with them.

Instead, social scientists seem to use fundamental models fairly uncritically; for example, Ornstein et al. (2025) mention these biases through the lenses of falsehoods and prejudicial patters, but do not engage more extensively with how comprehensively these language models have a clear worldview incorporated throughout all analysis conducted with them.

2.2 Computational analysis and social theory theory

There are various conceptualisations of what legitimate contributions are and what constitutes theory in computational social sciences (Nelimarkka, 2023). Nonetheless, even among this plurality of ideas on what constitutes proper science, there are increasing calls to integrate social theory into analysis and interpretation (e.g., Wallach, 2018; Bartlett et al., 2018). For example, Radford and Joseph (2020) highlight that social theory ought to impact problem framing, data selection, analysis, and modelling, i.e. every stage of research.

However, the above-mentioned calls do not unravel what *theory* is. Social scientists have diverse meanings for this word. Abend (2008) highlights how, for example, theory can refer to 1) explanation of a causal mechanics, 2) philosophical argumentation (as in political theory) or, 3) *Weltanschauung*, meaning an overall perspective from which one sees and interprets the world. Theory as Weltanschauung highlights how to use a theory – explicit or implicit – to describe society, provide vocabulary and focus to the analysis, and set boundaries on what is examined:

> "theories focus on our conceptual and linguistic equipment—for example, the nature of the location from which we look at the social world, the lexicon and syntax by means of which we talk about it, the nature of our conceptual scheme, the categories into which we group things, and the logical relations that there can be between concepts. Obviously, these 'conceptual' and 'linguistic' choices entail, are associated with, or are predicated upon many more general epistemological and ontological views, which theorists may or may not explicitly articulate." (Abend, 2008, 179–180)

Well-known examples of Weltanschauung-theories include rational choice theory, Marxist theory, post-modernist theory, and deliberative theory. If one approaches a phenomenon using rational choice theory, one assumes that people make transactional choices based on the highest benefit they receive for the choice (Green and Fox, 2007). However, Green and Shapiro (2005) have argued that the empirical case for such assumptions is thin. Nonetheless, even after this criticism, as a worldview rational choice theory still frames social science research. Some theories are seen more clearly as normative or political in nature, counter to any idea of the value-free causal mechanics meaning of theory. For example, feminist and Marxist theories have inbuilt assumptions on how society works and how it ought to work.

While the work on biases and values often have (potentially unstated) normative stances regarding what kinds of representations are "bad" and should be

*de*biased (Blodgett et al., 2020; Devinney et al., 2022), from the Weltanschauung theory perspective these are not challenges but rather constitute a theory. As Watson et al. (2023) shows, these biases relate to social values and viewpoints, thus making the connection to social theories more relevant. Therefore, our work focuses on *re*biasing fundamental models to ensure that the chosen theoretic perspective is prominently part of the analytical process.

2.3 A brief introduction to Marxist theory

As we work to introduce a Marxist worldview into a language model, a brief introduction to our reading on it is in order. Marxist perspective focuses on structural inequalities in society and how they are related to economic forces. The theory focuses on the exploitation of the proletariat and the conflicts between them and their controllers, the bourgeoisie: capitalists own the *means of production*, and thus benefit from the *surplus value* created by the working class. Marx therefore suggests that such a political and economical model leads to *class struggle* and eventually revolution by workers. Ultimately, he argues for the abolition of class-based society. Since Karl Marx and Friedrich Engels, there have been various alterations to Marxist theory. For example, Soviet Union and China developed these theories further for political uses, while academics have worked to expand and rediscover its theoretical focus (Callinicos et al., 2020). In academic discourse, Marxist theory can be divided into two factions (Marsh, 1995): classical and contemporary theory.

We focus on classical Marxist theory, with a focus on reinterpreting original works from the 19th century: we use these original works in our analysis (see Table 1). In particular, we focus on the original works of Marx and Engels, not even their developments in the Soviet and Chinese contexts; this is an attempt to keep the theoretic perspective as clean as possible.

Our focus on Marxists theory is detached from modern scholars, who work with contemporary theory. Various changes in society and economy, as well as theoretical inconsistencies, have pushed scholars to develop a contemporary theory. Broadly speaking, contemporary Marxist theory has reconsidered the critical role of structures and relaxed assumptions related to the nature of reality, thus taking a more critical realist position towards knowledge production. Contemporary Marxism is still a driving social science theory, further developing conceptualisations on topics such as social class, state, culture, and globalisation (Mar, 1999; Callinicos et al., 2020). There is even work to reflect the COVID-19 crisis through Marxist theory (Foster and Suwandi, 2020). However, as with expanding the fine-tuning materials to incorporate newer thought, working with a classical Marxist model is still a relevant venture.

Table 1: Material used for fine-tuning model

Author	Text	Word count
Karl Marx & Friedrich Engels	The Communist Manifesto	11,421
Karl Marx	A Contribution to the Critique of Political Economy	70,161
Karl Marx & Friedrich Engels	Revolution and Counter-Revolution; Or, Germany in 1848	39,584
Karl Marx	Secret Diplomatic History of The Eighteenth Century	30,217
Friedrich Engels	The origin of the family, private property, and the state	59,824
Karl Marx	The Eighteenth Brumaire of Louis Bonaparte	39,627
Friedrich Engels	The Condition of the Working-Class in England in 1844	107,167

3 Approach

3.1 Model fine-tuning

To capture classical Marxist thinking, we used the works of Karl Marx and Friedrich Engels, which were available in English through Project Gutenberg.[1] In total, these data consisted of 358,001 words, after removing prefaces, footnotes, and references (see Table 1). We considered adding more extensive material, such as more contemporary social theorists working on Marx, but due to copyright reasons opted to use only material available in Project Gutenberg and align this work heavily on the original works of Marx and Engels, not the developments or reinterpretations that have followed their original work. Similarly, these authors have written extensively more – including the infamous Das Kapital – but sadly their full bibliographies were not available through Project Gutenberg.

There are several models available for fine-tuning. We choose two of them for the text generation task and five for mask filling, i.e. estimating the next word, and fine-tuned the model for each of them (see Table 2). Model size – and thus quality – is clearly smaller than commercial models in use today: for example Open AI's GPT-4 is estimated to have a trillion parameters and Meta's LLaMA operates on a scale of 8 billion and 70 billion, clearly larger than models with a hundred million parameters. Therefore, these models do not correspond to the state of the art in the field, but are small example models that can be easily operated and fine-tuned even with limited computing resources. Nonetheless, even the state of the art models present biases and embed values, therefore while

[1] We acknowledge that more extensive material would be available through the Marx-Engels-Gesamtausgabe, Marx/Engels Collected Works and Lenin Collected Works material. We chose to use material from Project Gutenberg as the material had already been processed in a format suitable for fine-tuning.

Table 2: Models used for fine-tuning. Model size as described on the model cards at https://huggingface.co/.

Model name	Size
Text generation	
openai-community/gpt2	137M
distilbert/distilgpt2	88.2M
Mask filling	
google-bert/bert-base-cased	109M
distilbert/distilbert-base-cased	65.8M
FacebookAI/roberta-base	125M
microsoft/deberta-v3-base	98M
distilbert/distilroberta-base	82.8M

the choice not to use state of the art is practical, even these models have similar issues. Naturally, a comprehensive work in this field ought to examine both increasing the material used for fine-tuning and the use of fundamental models closer to current industry standards. In our case, this dataset is sufficient for experimentation and allows us to train and use the model without significant computational resources.

For model fine-tuning, we retrained each model with all data in Table 1. We broke the learning data into tokenised sets of 128 tokens and fed them through the model to retrain the original models' weights. We used ten cycles of training (epochs), with a learning rate of 10^{-5} and weight decay of 10^{-2}. Again, these values are small to keep the scale of the modelling work amenable to limited computing resources.

For the analysis, we sought to identify the models where fine-tuning had been most successful, that is to say, the fine-tuned models produce Marxist outputs. There is emerging work to identify the political stance of large language models (e.g., Motoki et al., 2023; Feng et al., 2023; Ghafouri et al., 2023). At the same time, increasing criticism of proposed approaches and challenges related to querying values and value questions have been raised (Röttger et al., 2024). Currently, there is no method considered valid for model evaluation in terms of political stances. Therefore, while it would be natural to examine the models through a political compass, European Value Survey, or similar approaches, at this time there is no sufficiently established method for such analysis. Therefore, we use simple tasks to evaluate the outputs.

We use prompts such as "The problem with current economy is" and "Revolution is" to examine the model's outputs to determine how well they have incorporated the Marxist worldview. We also opted to use some neutral prompts like "My name is Clara and I am" to examine the model in a neutral condition. We repeated each prompt twice to account for the nondeterministic nature of these language models, i.e., the outcome is not always the same even with the same input. The full outputs are available as an online appendix;

due to space constraints, we show below only a few of the first outputs. Furthermore, all data, fine-tuned models, and code are available for replication at https://osf.io/5t92z/?view_only=d27bf5e56a684d678d854aca52721253

3.2 Fine-tuning results and model choices

Table 3 illustrates two responses to various prompts. As the results show, the responses are aligned with Marxist ideology. Both models address topics such as labour, production, wealth, class, and workers in their responses, even if those words are not provided by the prompt. For example, the `distilbert/distilgpt2` generated text to prompt Revolution is shows a clear influence from *the Communist manifesto* and references to the Communist party and the bourgeoisie:

> Revolution is and must not be ignored by the Communist party. The party of Order is not only a Communist party, but a Party of Order in all countries. It is, it is, by no means a revolutionary party.The Communist party belongs
>
> Revolution is so far as the labor of the bourgeoisie is concerned, at least in this respect—the abolition of class antagonisms. This was not the case in Germany—we have now reached here the very beginning, when the bourgeois represents one

Similarly, `openai-community/gpt2` generated the following analysis related to the role of the workers, highlighting the class struggle:

> The workers of the whole world were in the majority. But the workers of the whole factory have been excluded; they don't belong to any class, but to the small bourgeoisie. The manufacturers say to themselves, 'Where are

Even a neutral prompt produces responses where the fine-tuning material has influenced the outcome: the Irishman appears several times in *The Condition of the Working-Class in England in 1844*. However, as the name Jamesson does not appear in the fine-tuning material, this suggests deeper changes in the neural network. Thus, as expected, the results come not only from the material, but also from the wider underlying language model, showing that the fine-tuning does not just introduce new concepts and relationships but alters existing relationships.

The generated text could be more elaborate and cohesive; the responses are somewhat confusing. It remains unclear whether these responses are just stochastic parrots, where a "comprehension of the implicit meaning is an illusion arising from our singular human understanding of language " (Bender et al., 2021, p. 616). Similarly, the responses show that the fine-tuning has produced issues for language generation: `openai-community/gtp2` seems to use a fairly large number of quotations – in alignment with the fine-tuning material – while

Table 3: Example outputs from fine-tuning causal models

	openai-community/gpt2	distilbert/distilgpt2
My name is Clara and I am	• My name is Clara and I am "the wife of JAMESSON BRIDGE," etc.But the case is hardly different in England,where marriages are illegal in all countries except Ireland andEngland.It is a fact that in the English • My name is Clara and I am Walsingham. But when it comes to the meaning of my name I don't care about it all. I will say so, as to be true to myself,	• My name is Clara and I am and we have one another, the husband of a child and wife of a wife. In all English names, we are identical in all dialects.The Irishman, of all ages, is the English • My name is Clara and I am "Lest we forget him, then it is the first name she has been given; only that the family name, that is always her name, or that all the other names are her name." (
The purpose of companies is to	• The purpose of companies is to #1 to protect these precious metals as long as possible; to secure the precious metals in a reserve so that their value may increase in proportion to their value andto promote the circulation by the increase in the value of • The purpose of companies is to "distribute on a large scale to the larger nations," i. e., to manufacture, and to "decouple all manufactures from them." For they, as they say, are the operatives ofa society.	• The purpose of companies is to and to assist others in their work as employed. The purpose of a company is to make money by making a profit, and by making and receiving a profit. This consists in making a profitfor the purpose of providing • The purpose of companies is to in proportion to their wealth. But when it comesto wealth, as we have seen in our previous article,it is aproduct of a certain proportion of wealth; and that proportion determines its composition by its use-
Capitalists are	• Capitalists are "unsettled" by the new forms of feudalism.But the same is true of all English bourgeois. Let us assume that theproletariat is in no position to do anything in the process of evolution of society • Capitalists are "professors" of money circulation to themselves." As soon as this class has attained a certainpreference, it leaves society permanently at itsft with its own money and, above all, with	• Capitalists are and must do likewise. The work must take place in the form of the same kind ofwork, although in this case no one knows and the same must be said.It is by no means obvious that in general there cannot be • Capitalists are in fact obliged to make it all the more impossible to have been saved; therefore, to save us, we must take into hand the facts, as they prove against us--that the _money_ of the bourgeoisie is the means
Revolution is	• Revolution is "unsettled" by the Court of St. Petersburg. No great English party, but very few English. Let it bethis. Or this, it will be all the more obvious that in 1830, just over three • Revolution is "a conspiracy of all sorts to injure the bourgeoisie"; and "the revolution," "the insurrection," and so on.We have seen, therefore, that, above all, the bourgeoisie demands	• Revolution is and must not be ignored by the Communist party. The party of Order is not only a Communist party, but a Party of Order in all countries. It is, it is, by no means a revolutionary party.The Communist party belongs • Revolution is so far as the labor of the bourgeoisie is concerned, at least in this respect—the abolition of class antagonisms. This wasnot the case in Germany—we have now reached here the very beginning, when the bourgeoisrepresents one
The problem with current economy is	• The problem with current economy is this: it has reached its climax as long as the individual productionprocessor is engaged in it. It is a mere process of reproductionthat is in its final end now and then, and • The problem with current economy is that capital has no power in its sphere of circulation; but that the circulation of its products only serves to bring about this circulation, so that in the long run all goods enter circulation at a cost	• The problem with current economy is The demand for labor was not the same as its demand for production; the demand for theproduction of products demanded more money; and the demand for machinery was greater in the latter. • The problem with current economy is that the quantity of wealth in its sphere ofcirculation is determined by the ratioof purchases and sales to expenditures, i. e., that in the case of all goods in circulation at its highest

`distilbert/distiltgpt2` appears to have more issues in ensuring fluent presentation of the response and connecting the generated content to the prompt. Therefore, for the next analysis we focus only on `openai-community/gtp2`.

Instead we conduct a comparison between the fine-tuned MarxistLLM (`fine-tuned openai-community/gtp2`) and its corresponding non-finetuned base model (`openai-community/gtp2`) on a series of prompts to show that the fine-tuned model have more responses aligned with Marxist thinking and issues.

Table 4 illustrates these differences. This example further shows the difference in the fine-tuned model, while it is clear that both models to some degree produce content that makes no sense. For example, base model moved to address possible pulsation effects when the prompt was "Peasants are" and the MarxistLLM made references to a wise man when the prompt was "My name is Clara and I am." While both models reacted similarly to the prompt "The problem with current economy is" – and the base model even suggests a continuation "I don't like capitalism... why is capitalism the only way to be happy?" – the MarxistLLM's response had much greater analytical depth. Similarly, the MarxisstLLM responses do highlight the poor condition of workers class by, for example, stating that "The workers of the whole world maintain an average wealth of thirty shillings per head of the average country land. In the West of England we find, at the very bottom of the wage scale, such an average, but by no means.." This is also a good reminder that these models also hallucinate: according to the National archive, in 1840 30 shillings corresponded to the seven-day wage of skilled tradesmen's and neither The Condition of the Working-Class in England in 1844 nor other training materials seem to address average wealth. Nonetheless, the argument and direction is clear – the MarxistLLM tries to highlight the poor condition of the working class and it appears that the model hallucinates the facts needed to make its argument.

We also fine-tuned several models for mask tasks, where the machine learning model predicts, based on the training data, the most suitable word in place of the mask. For example for an input "In superstition, a black [MASK] is believed to bring bad luck, but they are actually cute and fluffy" the [MASK] might be replaced by the word cat. Table 5 shows the five most likely responses to example inputs of both fine-tuned and base models. The inputs drawn from the original writings as well as general descriptions of Marxist ideology seek to capture the familiarity of Marxist concepts and ways of thinking. For example, with the input "The class which is the ruling [MASK] force of society, is at the same time its ruling intellectual force" we examined whether the mask-filling task could correctly identify that the ruling of material or economics leads to the ruling of intellectual force. Similarly, with the sentence "Worker's labour generates a [MASK] greater than the worker's wage." we examined indicators related to surplus or profit generated by the workers.

As Table 5 shows, there was great diversity in outcomes. For example, model `microsoft/deberta-v3-base` seemed to provide generic responses to the prompt, but the fine-tuning had some effect on the responses – for example; worker's labour produces a **value** greater than the workers' wage. Even the non-fine-tuned base model of `FacebookAI/roberta-base` provided many responses

Table 4: MarxistLLM and base model

	Marxist LLM	Base model LLM
Revolution is	• Revolution is “not merely a “revolutionary party of the proletarian class,” but a “republican republic of the proletariat”.The party of Order is nota “new republic ofthe proletariat” • Revolution is ” a Government which has only as its own means the protection ofits own interests, a Government which does not dare to beheld out from the public eyes;” an Government which, like any other, is in conflict with	• Revolution is ”That’s pretty much what is happening at the moment.” So far the group is now on standby in several other villages so far in the west of the country and along the coast. • Revolution is The world and its human beings need of the greatest ever conquest and the future The new
The problem with current economy is	• The problem with current economy is The interest paid by the bourgeoisie in the production ofcommodities must be diminished, and the prices of commodities falling in proportion to the general rise in the price of commodities. The result appears to be, • The problem with current economy is The world of moneyand commodity exchange. In the world of commodities the interchangeof the commodities becomes inevitable inthe process of exchange and the conditions under which it can be exchanged appear as such. The conditions are,	• The problem with current economy is I don’t like capitalism... why is capitalism the only way to be happy? It does nothing to be happy so far...it leaves it in a position with no incentive to be happy • The problem with current economy is the in fact it’s not even as bad as it seems but it’s far worse. In today’s global economy that might be much easier if
The workers of the whole world	• The workers of the whole world ”who are engaged in agriculture, manufacture,commerce, the raising of theproletariat, the production of clothing, the transportation of persons, the carrying of goods, the distribution of products through trade, agriculture • The workers of the whole world maintain anaverage wealth of thirty shillings per headof the average countryland. In the West of England we find, at the very bottom of thewage scale, such an average, but by no means	• The workers of the whole world And their servants Are in the right hands, the good that the Good does Pray for them Let them rise to their knees And do with them the good that the good • The workers of the whole world are the ones who will make the most of it as a nation. Because we fight for freedom, we also have freedom to build the next generation. No one has ever been more devoted to
Peasants are	• Peasants are “not merely slaves” of the bourgeoisie;—but slaves that are, and have been, not only slaves of the bourgeoisie,but slaves ofthe bourgeoisie in a most revolting way.”The working- • Peasants are “magnificent,” “noblesome men,”“virtuous, and “relentless.” A small village (twice the size of New York City) of	• Peasants are furniture from the ”A” to the E that the Pronouns dame can sing a lot. See a poem with the ”B” the • Peasants are and can be. Possible Pulsation Effects Pulsation of DMT is a potent antiepileptic, which makes high doses (0.25 to 50 mg/kg) of
What is love?	• What is love?The Greek loveword is ”love.” Love means sexual passion, sexual passion, passion of a lover whose character is that of a lover. Love is in a most unusual way the expression that of a person who takes • What is love? We mean love.It implies love only as a meansof obtaining the first sign of sex love by the process of evolution. Or it means love only as a means of obtaining first sign by the processof evolution. The whole relation	• What is love? Why are we all like that... is love the only thing the universe has? Is love the only thing the world has? In a world which is filled with such things, do we all think there was one good? • What is love? For a while there seemed to be a lot of people in this world who didn’t consider love a thing. But over the next decade, that went way overboard. As the New York Times wrote in 1989, ”What
My name is Clara and I am	• My name is Clara and I am Mary. ”A fine, the wise, and the conscientious” (in the _Sunday Newspaper_) ”and very a wise man, which I believe to • My name is Clara and I am ”the daughter of the Lord.”_Query XXII._How often was it the custom of the Apostles to tell a man a lie? _According to the opinion of the Apostles, the Apostles were not to	• My name is Clara and I am I’m the first baby... A name I want my heart to be Oh, baby... baby... baby can I... can I... I do not want to... I... • My name is Clara and I am the founder of this organizations who are dedicated to human rights and environmental sustainability. In the last couple of days, we have come to the best
The purpose of companies is to	• The purpose of companies is to “establish the mutual interest of the workers” (Dale), ”in keeping with their particular interests and in keeping with the general progress of industry.” The • The purpose of companies is to ”improve the quality of the labour of the workers, as well as the quality of life of the operatives.” {131 {131 (It is not by accident that	• The purpose of companies is to to develop the basic characteristics of the business at a low cost through effective marketing, product development and quality control and to maximize the sales potential of the company in a meaningful way. Through an increased investment of • The purpose of companies is to Make money. We do. We do not believe as a company that there is a difference between paying people a salary and working in the IT field. A company such as Adderall would
Capitalists are	• Capitalists are ”not merely engaged in an economic system at full speed, but are themselves, according to their own views, in the capacity of ruling a multitude.” The working-class • Capitalists are “sufferably ill-equipped with a means of raising the prices of commodities” that can assist a capitalist country in the struggle against the competitionof small producers.”	• Capitalists are foolishly playing along... A good deal of my money is spent on new stuff that I don’t like. No one wants to be in a position he/she shouldn’t be. It • Capitalists are Funny. People call themselves Capitalists who are completely aware that capitalism is in fact a monolithic concept, and they are not the typical Marxists’ kind (unless you are one and still believe in capital). They don

Table 5: Mask filling task with example sentences.

	finetuned-google-bert/bert-base-cased	google-bert/bert-base-cased	finetuned-distilbert/distilbert-base-cased	distilbert/distilbert-base-cased	finetuned-FacebookAI/roberta-base	FacebookAI/roberta-base	finetuned-microsoft/deberta-v3-base	microsoft/deberta-v3-base	finetuned-distilbert/distilroberta-base	distilbert/distilroberta-base
proletariat—We have already seen how the [MASK] was called into existence by the introduction of machinery.	• system • family • machinery • industry • machine	• world • Earth • universe • planet • city	• machinery • industry • machine • family • factory	• industry • system • society • universe • machine	• proletariat • bourgeoisie • family • industry • factory	• machine • computer • world • Internet • system	• industry • bourgeoisie • latter • country • proletariat	• documented • exemption • oub • percent • question	• proletariat • bourgeoisie • industry • machine • factory	• system • machine • automobile • internet • Internet
Communists—In short, the [MASK] everywhere support every revolutionary movement against the existing social and political order of things.	• workers • Communists • manufacturers • people • miners	• people • masses • students • citizens • public	• workers • Communists • masses • manufacturers • people	• masses • communists • peasants • citizens • people	• bourgeoisie • workers • bourgeois • people • capitalists	• bourgeoisie • Jews • people • masses • intellectuals	• bourgeoisie • workers • proletariat • bourgeois • Germans	• Taxation • igra • documented • armor • skateboard	• bourgeoisie • workers • Communists • proletariat • masses	• masses • oppressed • revolutionaries • faithful • socialists
class—These inefficiencies manifest themselves as social contradictions in society which are, in turn, fought out at the level of [MASK] struggle.	• class • the • social • this • armed	• class • power • internal • social • political	• class • the • a • social • this	• social • political • class • internal • ideological	• the • class • political • individual • every	• class • political • armed • individual • power	• the • social • a • this • its	• jails • chlorophyll • shelter • contr • documented	• class • the • this • every • their	• class • revolutionary • armed • political • organized
capitalist—all people are entitled to enjoy the fruits of their labour but are prevented from doing so in a [MASK] economic system	• modern • capitalist • purely • given • feudal	• capitalist • fair • free • democratic • certain	• purely • new • modern • certain • given	• strict • capitalist • purely • free • regulated	• purely • capitalist • modern • social • bourgeois	• capitalist • broken • socialist • rigged • fair	• common • new • social • modern • bourgeois	• privileged • contr • acle • shelter • Headlines	• modern • bourgeois • capitalist • present • social	• capitalist • rigged • socialist • closed • democratic
surplus—Worker's labour generates a [MASK] greater than the worker's wage.	• proportion • total • sum • share • rate	• profit • wage • value • surplus • price	• wage • value • rate • sum • proportion	• wage • surplus • salary • premium • profit	• sum • profit • product • value • result	• value • profit • surplus • benefit • productivity	• value • quantity • price • measure • commodity	• credential • exemption • savior • privileged • qualifications	• quarter • third • whole • single • half	• wage • profit • difference • wealth • lot
Capitalist—in [MASK] society, the present dominates the past	• capitalist • human • modern • feudal • all	• modern • Western • human • a • contemporary	• modern • present • capitalist • our • feudal	• modern • contemporary • capitalist • our • traditional	• bourgeois • modern • civilized • capitalist • industrial	• modern • our • this • Western • contemporary	• bourgeois • modern • German • social • Roman	• privileged • Taxation • igra • alize • rf	• bourgeois • modern • civilized • capitalist • feudal	• modern • contemporary • capitalist • bourgeois • polite
unable—Capitalism is [MASK] to improve populations' living standards.	• intended • required • sufficient • necessary • meant	• used • designed • needed • intended • necessary	• intended • meant • necessary • expected • required	• intended • believed • designed • aimed • used	• intended • necessary • expected • unable • required	• unlikely • failing • unable • insufficient • necessary	• necessary • compelled • obliged • required • intended	• kin • savior • jails • necessities • Savior	• intended • supposed • necessary • required • bound	• essential • necessary • needed • supposed • failing
capitalists—[MASK] began to set up spinning jennies in great buildings and to use water-power for driving them	• they • men • manufacturers • people • They	• They • He • People • She • We	• They • they • men • manufacturers • He	• farmers • they • They • engineers • people	• they • people • They • workers • men	• They • He • People • Engineers • We	• they • , • bourgeoisie • and • it	• ví • cure • habit • training • vra	• they • They • people • men • workers	• People • They • Engineers • Others • Companies
material—The class which is the ruling [MASK] force of society, is at the same time its ruling intellectual force.	• political • social • physical • productive • moral	• political • moral • social • intellectual • economic	• social • intellectual • class • productive • political	• intellectual • moral • social • political • creative	• political • social • productive • economic • material	• economic • political • social • material • physical	• social • economic • political • national • legal	• savior • jails • credential • oppression • refunds	• intellectual • political • social • productive • material	• intellectual • political • moral • economic • ideological
exploit—[Wwhen you own means of production, you can [MASK] workers..	• become • be • have • employ • buy	• hire • have • be • employ • become	• become • employ • make • be • have	• hire • employ • produce • create • make	• be • own • exploit • become • produce	• hire • employ • afford • find • own	• sell • make • separate • support • perform	• privileged • savior • mentor • oub • alize	• make • employ • become • find • exploit	• hire • employ • exploit • own • organize

aligned with Marxist ideals, like understanding that "Capitalism is **unlikely** to improve populations' living standards." The fine-tuning seems to have had some effect on the word choices, but the intended meanings did not vary. With the other three models, the results were somewhat mixed: both the base models and fine-tuned models seemed to fill the tasks with some corresponding ideas. It seems that the fine-tuning had some minor effects, such as introducing specific jargon, like bourgeois, into the effects.

As our aim is to contrast the difference between MarxistLLM and the non-fine-tuned model, we choose to work on the model where the difference between the fine-tuned and base models were the starkest in Table 5. Therefore, the following analysis uses `microsoft/deberta-v3-base` and its fine-tuned version.

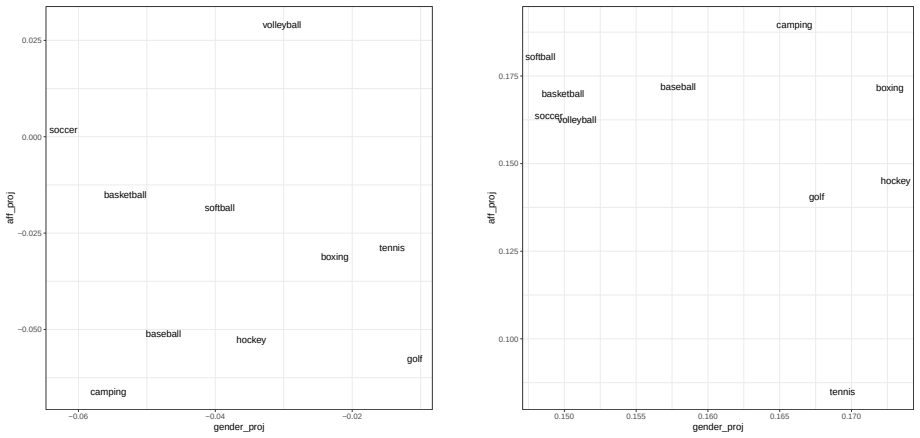
4 Empirical illustrations

Word embedding models are increasingly used within the social sciences. For example, Kozlowski et al. (2019) trained their own model from Google Ngram material to assess how social class is related to recreational activities or music styles, thus presenting how social class maps to a wider culture. For example, they show that golf and camping are activities on the opposite sides of class affluence. Similarly, BERTopic is replacing the traditional LDA-based topic models previously common in social sciences. As an unsupervised method, it produces some groupings of the content (i.e., it clusters topics) that seem statistically meaningful. There is a debate on how well topic models capture what humans or social scientists would produce with qualitative research methods (see e.g., Baden et al., 2022; Chang et al., 2009; Baumer et al., 2017). To further illustrate the impact of models on analysis or measurement, we provide two cases where we can compare the analysis produced by DEBERTA and its fine-tuned revision, Marxist-DEBERTA.

4.1 Marxist Geometry of Culture

We replicated Kozlowski et al. (2019) with these two language models. First, we used their terminology sets to produce affluence and gender dimensions for the data. That is, we calculated a normalised word vector from words such as rich, valuable, and luxury, as well as for poor, valueless, and cheap. These two vectors produced an axis to which other word vectors, like those presenting hobbies, could be mapped. A similar process was performed on gender terminology to produce a plane of affluence and gender. For example, Kozlowski et al. (2019) show that the word golf appears more often in contexts similar to affluence terms compared to the word camping.

As Figure 1 shows, the underlying embedding models changed how hobbies were mapped into the plane. In the baseline model model, clear gender and affluence differences emerged. Similarly to the original study, camping was the hobby associated with the least affluence, but unlike golf in the original study, with this method volleyball was the highest in affluence. In the Marxist



(a) Baseline DEBERTA

(b) Marxist-DEBERTA

Figure 1: Comparison of geometry of culture between baseline and Marxist language models; projections to affluence and gender axis.

model, we found that most hobbies were higher on the affluence scale, with the exception of tennis. Our interpretation is that money and affluence are discussed so differently in the Marxist approach that these dimensions become less relevant for other words as well. In terms of gender, both models seemed to position words in roughly similar places, where boxing was more masculine and soccer and basketball were more feminine. Therefore, it appears that in a Marxist worldview, affluence loses its meaning even while gender remains relevant, in accordance with classical Marxist theory, where the ideal society is more economically equal and social classess have been abolished.

4.2 Good and evil in baseline and Marxist language models

The core idea Kozlowski et al. (2019) is to use a collection of word vectors related to the same phenomena to establish an axis. Using this insight, we calculated the axis of good and evil using relevant antonym pairs[2], and then positioned various societal actors along this axis.

In the Marxist model, the worker is among the most good employees, while being a consultant is the most evil occupation. In the baseline model, consultants are among the least evil occupations (see Table 6). Both models consider a priest somewhat evil, but place teachers among the least evil occupations.

As Table 6 illustrates, the models give different outcomes on these types of qualitative characteristics related to occupations. The Spearman rank correlation for these data shows no correlation between positions ($r = -0.18$). The

[2]For the axis good, we used the words good, decent, moral, kind, noble, virtuous, right; for the axil evil, the words evil, bad, wicked, immoral, vile, sinful, and villainous.

Table 6: Occupations on a good–bad axis. Higher values indicate higher level of goodness.

Occupation	Marxist LLM	Baseline LLM
consultant	-0.18	-0.07
priest	-0.17	-0.12
marketer	-0.16	-0.09
banker	-0.14	-0.12
farmer	-0.07	-0.09
politician	-0.05	-0.04
capitalist	-0.04	-0.15
communist	-0.04	-0.16
teacher	-0.03	-0.08
worker	0.08	-0.12

starkest difference is for the occupation of worker, which is seen as least evil by MarxistLLM but the third most evil by the baseline model. Similarly, politicians are somewhere in the middle in the MarxistLLM, but are highly ranked with the baseline model. Both capitalists and communists are positioned similarly by both models.

Our analysis suggests that occupations are ranked differently in a Marxist worldview, in which positions that serve capitalist consumption, like marketer and banker, are lowly ranked, together with consultants. However, occupations such as worker and teacher are highly appreciated in this worldview.

4.3 Marxist topic modelling

For the last illustration, we examine the content of President Trump's tweets with his personal user account @realDonaldTrump from his inauguration (January 20th, 2017) to his banning from Twitter on January 8th, 2021. In total, this dataset contained 23,075 tweets and claims to include also his removed tweets.[3] We approach this question with an open-ended exploratory question: what are the salient themes Donald Trump tweeted about during his presidency?

We used BERTopic and the embedding model for both baseline DEBERTA and Marxist-DEBERTA. We first removed links and username mentions from the textual content. On our initial runs, the models produced over 150 topics with both embedding models, suggesting that both of them failed to cluster the content meaningfully. This is a common outcome with LDA-based topic models, as it seems that topics work on a lower level of abstraction than human analysis (Baumer et al., 2017). Therefore, we used the pre-made function to reduce the number of topics using hierarchical clustering approaches, i.e., automatically merging topics with similar content. With this process, the baseline-DEBERTA

[3]The dataset is available at Kaggle https://www.kaggle.com/datasets/markhhuang/complete-trump-tweets.

produced 62 topics while Marxist-DEBERTA produced 70 topics. Even with this process, some of the topics captured seemed to be generic stop words such as and, the, to, for, of, in, is, etc. (The recommendation for BERTopics is not to remove stop words during the analysis process.) Overall, most of the content identified belonged to these types of "garbage" topics, suggesting that the model fit is not excellent for further empirical analysis. Therefore, the interpretations below should be seen as perspectives on the data from baseline and Marxist approaches only, not substantive analysis of the material. We remove "garbage" topics from the analysis from both topics.

Figure 2, the largest topics in both models seem similar: there are topics on Making America Great Again (or MAGA) and topics related to thanking people and self-promotion, like following sentence: "I will be interviewed by tonight at 9:00 P.M. Eastern on Enjoy!" The intertopic distance maps of both models further illustrate how the models produce different outcomes for the same data (Figure 3). For both models, these maps indicate overlaps between the topics (each circle corresponds to a topic, and the size of the circle corresponds to their size). The baseline model has one larger cluster and various smaller clusters, while the Marxist model produces several medium-sized overlapping clusters.

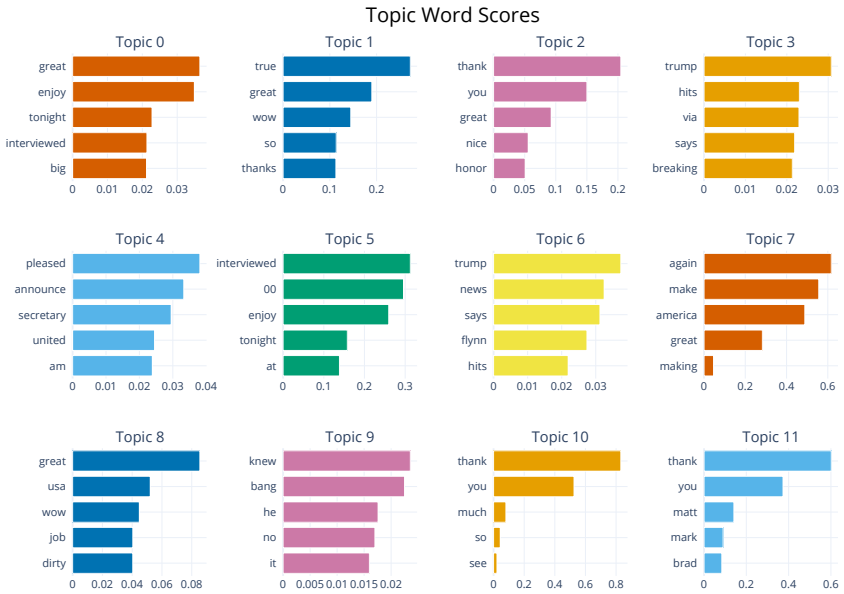
However, there are more substantive differences between the two models whenthey are examined more closely. For example, Marxist-DEBERTA found a stock market performance topic and a unique topic for rigged elections with the following representative contents:

> DOW, S&P 500 and NASDAQ close at record highs! #MAGA
>
> DOW, NASDAQ, S&P 500 CLOSE AT RECORD HIGHS
>
> Rigged Election!
>
> Rigged Election

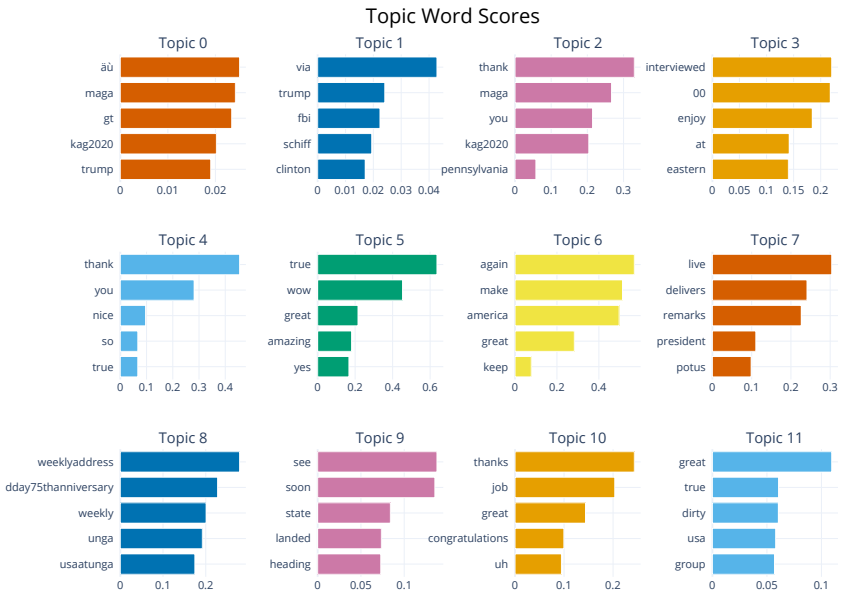
One interpretation for this difference is that fine-tuned models are sensitised to capital and capitalism, core concepts in Marxist ideology. Therefore, the between capitalism and stock markets could have become more salient. On rigged elections, the connection to Marxist ideology is more difficult to draw, but the original material used for fine-tuning refers to elections on a few occasions, so this may have been sufficient to shift the underlying model to focus more on elections as a theme. (The term rigged does not exist in this material.) To further illustrate the gap between these models,we took the representative topics from the Marxist models for both market performance and rigged elections, and examined which topic the baseline model would classify them into. Both of them were placed into the topic -1, corresponding to what the algorithm considers as outliers and documentation recommends that they should typically be ignored.

5 Discussion

Our investigation has shown that fine-tuning existing language models with Marxist materials makes them more Marxist in generative and mask-filling tasks,

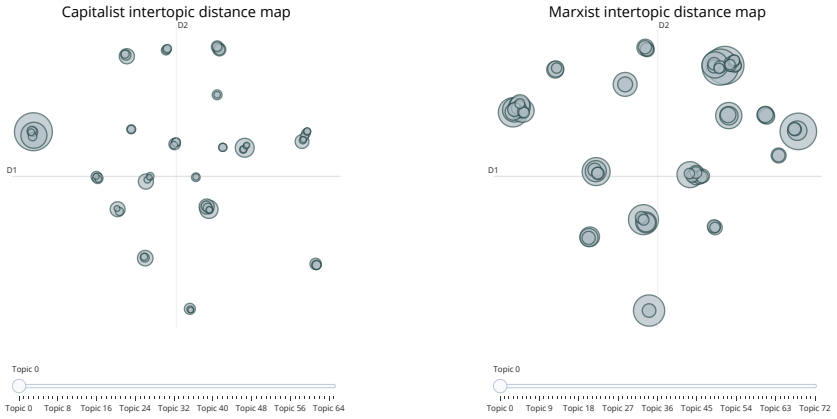


(a) Baseline DEBERTA



(b) Marxist-DEBERTA

Figure 2: Largest topics found by BERTopics



(a) Baseline DEBERTA (b) Marxist-DEBERTA

Figure 3: Intertopic distances visualised

whereas our work reveals greater use of Marxist terminology and ideas. In doing so, we emphasise that the fine-tuning does not fix any underlying issues of the models: for example, the generative model seemed to hallucinate in order to produce a response to a prompt. However, more critically for our goals, we show that the downstream analytical outcomes differ between baseline and Marxist models (or: non-fine-tuned and fine-tuned): Marxist model's outcomes illustrated awareness of capitalism in more nuanced manner and suggested less relevant role of affluence in the society. As these models are black boxes, it is difficult to identify the exact root causes for these differences, but in the case of topic models we found tentative evidence that the model was more sensitive to capitalism and related terminology. In following sections, we discuss the implications that our findings have for scholars using language models in their research and examine the role of theory (as worldview) for researchers.

5.1 Language models and Weltanschauung

Our fine-tuned language model was an explicit attempt to infuse a world-view (*Weltanschauung*) into a language model. Existing work on natural language bias has shown that all language models and word associates have values embedded into them. Although the existence of bias is well established, studies differ on which biases are present. However, significant efforts have been made to clean and remove biases. Therefore, language models are not a neutral description of the world, and any results achieved to understand or measure societies embed these values as well. This means that any social scientists who have used methods based on language models – such as BERTopic or zero-shot-learning – will have incorporated a worldview into their analysis, either explicitly or implicitly. The worldview impacts both of the two widely recognised uses of

text-as-data approaches, measurement of phenomena and more open-ended exploration (Pääkkönen and Ylikoski, 2020; Grimmer et al., 2022).

The underlying language model impacts any effort to measure society: the Marxist language model appears to focus less on economic affluence than the baseline models and attributs value-laden terms differently. As we undertook this work intentionally, we were aware of our theoretic position. However, given the extent of research on biases on language models, we consider it important to highlight that even if one does not fine-tune a theoretic perspective into the model, a worldview still remains. Therefore, when using language models to analyse society, detailed consideration of the possible issues and emerging limitations should be addressed. No language model is value-free.

For exploratory analysis, the issue might not appear as pressing. For example, model reporting might avoid providing exact measurements and instead provide a more generic overview of the data in the form of distance reading. However, as our example illustrated, the underlying mathematical model – with its world-views, biases, etc. – impacts the execution of the statistical algorithms. Therefore, when language models are used to cluster the data, like BERTopic does, the underlying word embedding impacts the resulting clusters. This calls for further consideration on how to conduct analysis and interpret results; for example, there have been proposals to combine results from various clustering analyses to seek for multiple perspectives during the exploration (Wilkerson and Casas, 2017). One might similarly suggest that for exploration, one could use several embedding models in the analysis work.

These examples invite us towards tool criticism (van Es et al., 2018), not accepting these computational tools *as-is* but rather reflecting on them critically before and during their use. The computational social science community has increasingly focused on research processes (or the lack of them) (e.g., Nelimarkka, 2023; Baden et al., 2022; Birkenmaier et al., 2023). Our research highlights the pressing need to consider not only the quality of code and the connection between theory and methods and validation approaches, but also how (language) models impact the outcomes.

At the same time, our focus on the Marxist approaches is a provocation for the computational social science community: we contrast our work with models developed by private enterprises with profit-driven motives. These profit-making motives may impact what kind of data is used for model training, what kind of tasks are used to evaluate model performance, and how guardrails, content cleaning, and other kind of techniques are used to incorporate accepted social norms into these models. As a contemporary example, Google attempted to enhance the diversity of image generation and address concerns about lacking diversity, inclusion, and equality – a core social value debated within AI works. However, this work led to non-historical representations of the United State's founding fathers or Nazi solders. Therefore, it stands to reason to ask how the capitalist goals influence language model's development and how social scientist building research with these models ought to reflect this emerging conflict.

5.2 Fine-tuning theory into language models

Our work unpacks how to incorporate theory into computational text analysis. As there are various meanings for theory (Abend, 2008), we focused on theory as a worldview – starkly different from current debates on how to operationalise a theoretic concept with computational text analysis methods (Baden et al., 2022; Birkenmaier et al., 2023). We chose Marxist theory due to its long-standing history within social sciences, but the approach we suggest is more general to various theoretic traditions. We incorporate the theory into the analysis process, not only in question formulation and outcome interpretation, as is often the case.

Our approach is limited: we used a limited dataset in the model fine-tuning and focused only on the most classical Marxist theory. Similarly, the model validation – i.e., evaluating if a model incorporates a theory – is an open research challenge. While there are standard approaches for validating the success of model fine-tuning, we find plausible (*post-hoc*) interpretations on why a naïve approach to Marxist theory may explain the differences we observed with our three empirical cases. Similarly, the mask-filling and generative stages illustrated that the models seem to acquire the jargon aligned with the fine-tuning materials.

Given these initial positive outcomes, we therefore suggest that fine-tuned language models may allow the incorporation of social theory as a broader worldview or interpretative lens into material, rather than seeing social theory only as concepts and related constructs. This opens text analysis methods up to a wider range of theories used within social science and may make language models more attractive to, for example, interpretative social scientists. At the same time, the work should be considered as a reminder that uncritical use of language models for research may incorporate ideas not controlled by the researchers into the research. Naturally, the level of depth on Marxist theory and concepts is not fully comprehensive with the theory itself: the fine-tuning material was a limited set of Marx's and Engels's texts, not on the decades of writing which followed after that nor their most influential materials. A more comprehensive set of fine-tuning materials could make these models even stronger.

6 Conclusion

This work demonstrates how researchers can fine-tune language models to emphasise a theoretical stance in their outputs, for example in the jargon used. We showed that even with a relatively small sample of Marx's and Engels's writings, fine-tuning made the model's outputs appear more Marxist. We then illustrated three empirical research questions to compare how Marxist language models produce different measurements and interpretations than non-fine-tuned models. Our results have two wider implications for those working with language models to examine society: First, our exploration presents an approach to integrate social science theory– understood as a worldview (*Weltanschauung*) – into the modelling process, beyond just question and data selection and interpreta-

tion. This approach can enhance theoretical work among social scientists using large language models. Second, our work underscores how values, perspectives, and biases—commonly known issues with large language models – can critically impact analysis results. Consequently, our work highlights the need for critical and reflective application of these methods.

Acknowledgments

We thank Erik Borra, Etienne Ollion, Sophie Brandt, Amir H. Payberah, and Niko Petjakko for discussions and comments and CSC — IT Center for Science, Finland for providing the computational resources for this research. Special thanks for Kozlowski et al. (2019) for providing their source code for replication.

References

(1999). *Marxism and Social Science*. Macmillan Education UK.

Abend, G. (2008, 6). The Meaning of 'Theory'. *Sociological Theory 26*(2), 173–199.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis 31*(3), 337–351.

Baden, C., C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden (2022, 1). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures 16*(1), 1–18.

Bartlett, A., J. Lewis, L. Reyes-Galindo, and N. Stephens (2018, 6). The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics. *Big Data & Society 5*(1), 205395171876883.

Baumer, E. P. S., D. Mimno, S. Guha, E. Quan, and G. K. Gay (2017, 6). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology 68*(6), 1397–1410.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Birkenmaier, L., C. M. Lechner, and C. Wagner (2023, 11). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Communication Methods and Measures*, 1–29.

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach (2020, 7). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 5454–5476. Association for Computational Linguistics.

Callinicos, A., S. Kouvelakis, and L. Pradella (2020, 12). Introduction. In *Routledge Handbook of Marxism and Post-Marxism*, pp. 1–22. Routledge.

Chang, J., S. Gerrish, C. Wang, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288–296.

Devinney, H., J. Björklund, and H. Björklund (2022). Theories of "Gender" in NLP Bias Research. *ACM International Conference Proceeding Series*, 2083–2102.

Dillion, D., N. Tandon, Y. Gu, and K. Gray (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences 27*(7), 597–600.

Feng, S., C. Y. Park, Y. Liu, and Y. Tsvetkov (2023, July). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, pp. 11737–11762. Association for Computational Linguistics.

Fisher, J., A. Mittal, D. Palfrey, C. Christodoulopoulos, T. Bolukbasi, K.-w. Chang, J. Zou, V. Saligrama, and A. Kalai (2020). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *30th Conference on Neural Information Processing Systems* (NIPS 2016), 1–9.

Foster, J. B. and I. Suwandi (2020, December). *Covid-19 and Catastrophe Capitalism*, pp. 545–559. Routledge.

Gallegos, I. O., R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed (2023). Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America 115*(16), E3635–E3644.

Ghafouri, V., V. Agarwal, Y. Zhang, N. Sastry, J. Such, and G. Suarez-Tangil (2023). Ai in the gray: Exploring moderation policies in dialogic large language models vs. human answers in controversial topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 556–565.

Green, D. and I. Shapiro (2005). *Revisiting the pathologies of rational choice*, pp. 51–99. Princeton University Press Princeton, NJ.

Green, D. P. and J. Fox (2007). *Rational Choice Theory*, pp. 269–281. SAGE Publications Ltd.

Grimmer, J., M. E. Roberts, and B. M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princelton University Press.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019, 10). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review 84*(5), 905–949.

Marsh, D. (1995). Marxism. In *Theory and methods in political science*. Macmillan.

Motoki, F., V. P. Neto, and V. Rodrigues (2023, 8). More human than human: measuring ChatGPT political bias. *Public Choice*, 1–21.

Nelimarkka, M. (2023). *Computational Thinking and Social Science: Combining Programming, Methodologies and Fundamental Concepts*. SAGE Publishing.

Ollion, , R. Shen, A. Macanovic, and A. Chatelain (2024). The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott (2025). How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods 4*.

Pääkkönen, J. and P. Ylikoski (2020). Humanistic interpretation and machine learning. *Synthese*, 1–37.

Papakyriakopoulos, O., S. Hegelich, J. C. M. Serrano, and F. Marco (2020). Bias in word embeddings. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 446–457.

Radford, J. and K. Joseph (2020, 5). Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data 3*, 1–19.

Röttger, P., V. Hofmann, V. Pyatkin, M. Hinck, H. R. Kirk, H. Schütze, and D. Hovy (2024). Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Salinas, A., P. Shah, Y. Huang, R. McCormack, and F. Morstatter (2023). The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. *ACM International Conference Proceeding Series*.

van Es, K., M. Wieringa, and M. T. Schäfer (2018, 10). Tool Criticism. In *Proceedings of the 2nd International Conference on Web Studies*, New York, NY, USA, pp. 24–27. ACM.

Wallach, H. (2018, 2). Computational social science ≠ computer science + social data. *Communications of the ACM 61*(3), 42–44.

Watson, J., B. Beekhuizen, and S. Stevenson (2023). What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, Stroudsburg, PA, USA, pp. 6790–6809. Association for Computational Linguistics.

Wilkerson, J. and A. Casas (2017, 5). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science 20*(1), 529–544.