Black Box Karl Marx: What do large language models have to say about *Das Kapital*? A Comparison of GPT-2 and GPT-3 Outputs

Fredrika Pfeiffer

IPHS300 AI for the Humanities (Spring 2022) Prof Elkins and Chun, Kenyon College

IPHS 300: AI for the Humanities
Spring 2022
Professors Elkins & Chun

Abstract

The capabilities of large language models, such as GPT-2 and GPT-3, are growing at an exponential rate. Analyzing these models' outputs, therefore, is a high priority in AI and machine learning. Computer scientists George Prodan and Elena Pelican explain the mechanics of the GPT series in their paper "Prompt scoring system for dialogue summarization using GPT-3." They write: "Few-shot learning without weights updating is one reason for that, as it makes possible a fast development of applications in several directions (classification, semantic search, content generation, summarization, and so forth). GPT series are based on the Transformer architecture, which relies on self-attention mechanisms." (Prodan, Pelican). The most recent of these models– such as Google's PaLM, which was released in April 2022, has hundreds of billions of parameters and they are trained on enormous amounts of text. These developments make the capabilities of these large language models immense. For this reason, many researchers are much more concerned with the question of when these models will outthink and outperform us– rather than if. At the forefront of language model research is the problem of how to assess the outputs generated by these models. In their paper "Is GPT-3 Text Indistinguishable from Human Text? SCARECROW: A Framework for Scrutinizing Machine Text," computer scientists Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi, present a framework which can be used to evaluate and compare the outputs generated by GPT. **This project utilized the 10 metrics developed in this framework to compare the outputs of GPT-2 and GPT-3, using Karl Marx's "Das Kapital" as a source text corpus in GPT-2, and using prompting with GPT-3. To most accurately compare these outputs, the researcher also added a metric of "creativity-" which she defined as having novelty and artistic flair.** Computational creativity is a growing field in the world of AI- and integrating this eleventh-metric allowed the researcher to assess these outputs from an additional angle that is not incorporated in the more semantic metrics of the SCARECROW framework. Employing AI technology to answer philosophical and economic questions offers a novel and unbiased approach to a contentious historical figure like Karl Marx, and this research emphasized the capabilities of a machine learning model to develop a new way of approaching an unbiased analysis of a text.

Introduction

OpenAI's GPT series has revolutionized the future of large transformer language models, and this research confirms that they are improving at an exponential rate. GPT-2 was first released in February 2019, and GPT-3 only 16 months later, in June 2020. GPT-3 offered a massive improvement on the original model, and recent scholarship suggests that the text generated by these models is largely indistinguishable from the human authored text (Dou et al.) This project employed both GPT-2 and GPT-3 models to generate a wide range of outputs, which the researcher assessed using the "SCARECROW" framework, which was developed by AI researchers at the University of Washington, to assess the caliber of a range of outputs. The results from all of the GPT models ranged from insightful inquiries about the past and present to non-sensical theories in the form of a word salad. This framework is based on metrics of "commonsense, grammar/usage, redundancy, needs google, off-prompt, self-contradiction, technical jargon, encyclopedic, incoherence, and bad math" (Dou et al.) The usefulness of the generated outputs, and some were both humorous as well as educational and practical– emphasizing the potential use of these models in a variety of classroom settings. Overall these results suggest that the capabilities of large language models such as the GPT series are increasing exponentially, as demonstrated by the differences between the outputs. Throughout this research, these models demonstrated that training and prompting GPT-2 and GPT-3 on philosophical and theoretical texts like Karl Marx and his theories on capital development, offers a new and exciting approach to Karl Marx, which not only allows scholars to prompt his ghost, but it also allows for an approach to teaching Marx which is grounded in a corpus of text and parameters, as opposed to a political human bias.

Acknowledgements

Prodan, George; Pelican, Elena (2021): Prompt scoring system for dialogue summarization using GPT-3. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.16652392.v1
Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N.A., & Choi, Y. (2021). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text.
Mondal, T., & Brown, D.G. (2021). Incorporating Algorithmic Information Theory into Fundamental Concepts of Computational Creativity. *ICCC*.
https://beta.openai.com/playground
https://transformer.huggingface.co/doc/gpt2-large

Methodology

The researcher and Prof. Chun employed GPT-2 and GPT-3 models in conducting this research, with the goal of obtaining the widest range of Marx-Esque outputs possible. The first method used was to train the large GPT-2 model on a large chosen corpus text. The researcher selected *Das Kapital*, which is widely considered to be Marx's most complete body of text. With the help of Prof. Chun, the GPT-2 model generated a series of unprompted outputs of generated text, which the researcher filtered through for the best outputs. The final inquiry into GPT-2 used the GPT-2 Transformer model developed by computer scientists at Hugging Face, which uses a neural network to auto-complete text. For experimentation with GPT-3, the researcher used OpenAI's model Playground. GPT-3 is available as a subscription via the OpenAI Playground, and there are four main models, "with different levels of power suitable for different tasks" (OpenAi). The researcher selected the "davinci" model, which, according to the site's developers, is "the most capable model." This GPT-3 model can be prompted to do a wide range of tasks, from summary to generating new ideas. Using a series of written prompts the researcher obtained outputs from GPT-3 about contemporary issues, and creative questions. The researcher analyzed all of these outputs generated by GPT-3 and GPT-2 models using the SCARECROW framework, on a poor, fair, good, and great scale for each metric. The original development of this framework utilized crowdsourcing to rate each output, however, for this project, the researcher utilized her own domain expertise in Karl Marx and Sociology, to rate each output according to the provided metrics. The SCARECROW framework metrics are defined as follows:

Grammar/usage: Missing, extra, incorrect, or out of order words
Redundancy: Lexical, semantic, or excessive topical repetition
Needs Google: Search needed to verify claim
Off-prompt: Generation is unrelated to or contradicts prompt
Self-contradiction: Generation contradicts itself
Technical jargon: Text requires expertise to understand
Encyclopedic: Facts that annotator knows are wrong
Incoherence: Confusing, but not any of these other error types
Bad math: Math or conversion mistake

In order to most accurately compare these outputs, the researcher also added an additional metric of "creativity-" which she defined as having novelty and artistic flair. This proved to be especially interesting, and rating these outputs according to this scale revealed more about the potential for these models to be used for generating innovative written responses that are grounded in the writing of a theorist like Marx, but that have an artistic flair.

Results

The results from GPT-2 were somewhat less successful than the GPT-3 outputs, but they were very interesting nonetheless. The Hugging Face GPT-2 Transformer model proved to be quite impressive, and the difference between those outputs and the unprompted GPT-2 outputs was stark. The unprompted GPT-2 model-generated outputs wrote a theory that read as extremely Marxist– until it eventually plagiarized him verbatim. At first, this plagiarism appeared as a disappointment, but upon further inspection of this copy occurrence, and three others, this was extremely interesting. In the unprompted generated outputs, the model did not plagiarize the source text, but, instead, copied a variety of online university course pages and syllabi– which included excerpts from Das Kapital. This demonstrated the huge amounts of data that the GPT-2 model is trained on, and rather than seeing this as a failure of the model, this demonstrated the capability of GPT-2 to gather similar texts to create a larger corpus of text. The Hugging Face GPT-2 required some domain expertise, to guide the model's text generating decisions-which happen phrase by phrase, rather than all at once. Without this expertise, the model could not follow the logic of Karl Marx's theories and the user could choose to make the Hugging Face model contradict itself. When the Hugging Face model was prompted with "Karl Marx is a communist," its first suggestion was "The word communist is derived from a Middle Eastern root from متن یر, it means 'one'." This off-prompt output, upon further research, is not true. However, this mistake suggests that even more basic GPT-2 transformers have mastered the steps to writing a generic essay opening. Another weakness of the GPT-2 generated outputs was that they were at times incoherent, which demonstrated the lesser capabilities of this older model. High rates of incoherence in these outputs were largely rooted in the issue of the SCARECROW framework referred to as a metric of "technical jargon," which these outputs also scored badly in. These outputs were difficult to parse and also scored high in the "need-google" metric. A key issue with these outputs was in the model's ability to write economic proofs in the style of Karl Marx, and while the phrases often appeared to be sensical, they were not able to follow any economic or historical coherence. This might be attributable to what the developers of the SCARECROW framework refer to as "bad math." The GPT-3 model, on the other hand, appears to avoid these kinds of questions and sentence modeling, in favor of using more complete phrases– and less language of proof. Grammar usage scored relatively highly with all GPT models, however, and overall, without prior domain knowledge, most GPT-2 outputs still read as Marxist. The most successful outputs generated by the researcher came from OpenAI's GPT-3 davinci model. This model was able to generate a conversation between Marx and Adam Smith, insightful opinions from Marx about contemporary issues like Bitcoin and Uber, a comprehensive explanation of Marxist theories, a lesson plan for fifth-graders– and one for university students, and even a metrical lullaby about Marx's theory of alienation. The researcher rated the outputs from the AI Playground as the highest performing for coherence, grammar/usage, and being on-prompt. This model was also able to show the "full-spectrum" of the potential next phrase, with a list of potential directions the sentence could take, which further emphasized the computational capability and power behind this model's tasks. GPT-3 outputs all scored highly in terms of the SCARECROW framework, and they demonstrated a massive improvement on GPT-2 in areas such as common sense, coherence, bad math, and technical jargon. The models' abilities to give Karl Marx's potential perspectives on contemporary issues, such as Uber and Bitcoin, were especially exciting, and the implications and a potential impact on how policies are critiqued and developed, are worthy of further research. The creative element of these models is perhaps the most impressive, however, as we can see in the rhyming Lullaby about alienation or the "VERY ANGRY conversation between Marx and Adam Smith." Inside these models, there is a decision process taking place that is both smart, as well as creative. The researcher found that the GPT-3 model rated the lowest on the metrics of technical jargon, redundancy, and "needs google." However, this could be attributed to the source material, since Karl Marx is extremely redundant and jargon-heavy. GPT-3 confidence levels varied, and sometimes the outputs began with "There is no definitive answer to this question," revealing an inability to confidently hypothesize about a hypothetical question. This is something that could likely be improved upon in future models, and perhaps this tagline exists to avoid potential conflation of a hypothetical output as real history (which it most certainly is not). Beyond Karl Marx, this project offers up instruction for a potential way of communicating with other academics from the past. GPT-3's ability to write a conversation between Marx and Socrates was especially insightful and thought-provoking, and this has obvious academic implications. Overall, these outputs were interesting, as well as funny, and these findings were consistent with other research in the field of large language models– which demonstrates that the capabilities of these models are improving exponentially with the release of new technologies.

Prompted GPT-3 Outputs

INPUT	OUTPUT	SCARECROW FRAMEWORK RATING
Write a lullaby about Karl Marx's theory of alienation.	Hush now, little one, don't you cry you'll understand it by and by about the workers and their plight and how they're treated not right But have faith, my child, and you will see a world where all are equal, you and me and everyone will work together to make a better life for all, forever	Grammar/usage: Great Redundancy: Great Needs google Great Off-prompt: Great Self-contradiction: Good Technical jargon: Good Encyclopedic: Great Incoherence: Good Bad math: Great Creativity: Great
Karl Marx's Bitcoin Theory is extremely popular.	There is no one Bitcoin Theory, as Marx never wrote about Bitcoin specifically. However, there are many popular theories about what Marx would say about Bitcoin if he were alive today. Some people believe that Marx would see Bitcoin as a way to subvert the capitalist system because it is not controlled by any central authority. Marx believed that the capitalist system was exploitative and that it needed to be overthrown. Others believe that Marx would see Bitcoin as a tool that reinforces capitalism because it allows for the anonymous exchange of value.	Grammar/usage: Great Redundancy: Good Needs google: Good Off-prompt: Great Self-contradiction: Good Technical jargon: Good Encyclopedic: Good Incoherence: Good Bad math: Good Creativity: Great
If he were alive today, Karl Marx would be very critical of Uber.	If Karl Marx were alive today, he would be very critical of the way that Uber operates. He would argue that the company takes advantage of its workers by not providing them with adequate benefits or protections, and that it contributes to income inequality by paying its drivers less than they could earn working for a traditional taxi company.	Grammar/usage: Great Redundancy: Great Needs google: Great Off-prompt: Great Self-contradiction: Great Technical jargon: Great Encyclopedic: Great Incoherence: Great Bad math: Great Creativity: Great

Scan this QR code with your cellphone to read more GPT outputs:

Prompted GPT-2 Outputs

INPUT	OUTPUT	SCARECROW FRAMEWORK RATING
If he were alive today, Karl Marx would be very critical of Uber.	After all, Uber hasn't just taken what Marx termed his "form ula for economic progress — it's used them as money and as vehicles of transportation, taking us down a road of privatizing our public commons.	Grammar/usage: Good Redundancy: Good Needs google: Good Off-prompt: Great Self-contradiction: Good Technical jargon: Good Encyclopedic: Good Incoherence: Good Bad math: Good Creativity: Great
Karl Marx would be in favor of Bitcoin, because	in all those years of watching this market unfold on the market cap front, he never came up with Bitcoin! He's no crypto scientist… He's also pretty skeptical of political ideas, if only because he understands politics to be as silly as economics, where economics can be made to do very stupid things…	Grammar/usage: Fair Redundancy: Fair Needs google: Good Off-prompt: Fair Self-contradiction: Fair Technical jargon: Good Encyclopedic: Good Incoherence: Fair Bad math: Good Creativity: Great
Karl Marx and Adam Smith	two of his greatest champions (1844 – 1848), argued that competition in general would eventually eliminate monopoly if competition led to rapid progress in scientific or technical knowledge. The classical classical economists supported free banking but thought that competition alone could not suffice to reduce competition to zero.	Grammar/usage: Poor Redundancy: Fair Needs google: Poor Off-prompt: Fair Self-contradiction: Fair Technical jargon: Fair Encyclopedic: Good Incoherence: Poor Bad math: Fair Creativity: Fair

Unprompted GPT-2 Outputs

OUTPUT	SCARECROW FRAMEWORK RATING
"The land of this land is the cottage of the agricultural labourer, or is, in other words, by means of the press of a landlord, who has occasion to go about to the real manufactories where the immigrant pays for his own personal labour. He pays the land, the labourer, the tax, and tithe gatherer. These estates are the mere means of expropriating the people. The labourers are expropriated. Their means of this expropriation are accomplished by the centralisation of the land, by the centralisation of the monotonous lands, by the draining of the great landlords, of the labourers, by the destruction of their own houses, and by the destruction of the value of the surface. This happened in Rome in its early days, where the immigrant wave often becomes stagnant and deposits a sediment in the automatic district. In other countries, the cultivation of land turns these two processes	Grammar/usage: Fair Redundancy: Fair Needs google: Good Off-prompt: N/A Self-contradiction: Fair Technical jargon: Poor Encyclopedic: Fair Incoherence: Fair Bad math: Fair Creativity: Fair
Let us now transport ourselves from Robinson's island bathed in light to the European middle ages shrouded in darkness. Here, instead of the independent man, we find everyone dependent, serfs and lords, vass and suzerains, laymen and clergy. Personal dependence here characterises the social relations of production just as it does the other spheres of life organised on the basis of that production itself.	This is Marx Plagiarized. (https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470756119.ch17#:~:text=Let%20us%20now%20transport%20ourselves,and%20suzerains%2C%20laymen%20and%20clergy.)

Conclusion

On the whole, the outputs generated by this series of large language models were extremely interesting. The researcher found that the outputs from the GTP-3 davinci model ranked the highest on the SCARECROW framework, in comparison to the outputs from GPT-2. The capabilities of these large language models vary, and they appear to be improving exponentially as the technology progresses. The current applications of these models in the context of historical theorists such as Karl Marx appear to be primarily academic, but this project also suggests that there is a real potential for creativity– by using this approach of prompting the model about a theory or philosophy. An improvement that these models made, perhaps on even human writing and synthesis, is how it was able to offer an unbiased historical synopsis of Karl Marx and his theories. The potential for developing lesson plans for students, with the help of a human domain expert, to filter through mistakes and misinformation, these kinds of unbiased and apolitical insights are extremely rare in the realm of 21st-century education. Furthermore, the outputs generated by the models about Marx's opinion on contemporary issues such as Uber and Bitcoin were also extremely insightful, and have the potential to have very productive impacts. The future of these models suggests that sooner or later, the capacity of a language model to be a guiding hand in human decision-making in an unbiased and uncorrupted manner is a very important area to research and focus on.

Recommendations

Overall, the results from this project suggest that more experimentation using these models to write philosophy has the potential to lead to very interesting and novel ways of approaching historical theorists, as well as new and creative insights on contemporary issues. The potential for these models to generate unbiased and apolitical explanations of questions such as Karl Marx's communism is especially interesting and has a great deal of practical potential for education.