Reponse Essay to CAREER: A Response Model for Labor Sequence Data

Notes for LLM Critique Essay: Your response essay can be about writing the essay with an LLM. Reading the paper carefully is still required. This consists of two parts: An LLM-produced essay, and a separate section explaining and critiquing the essay. The LLM-produced essay must be produced with a prompt or set of prompts using the paper, or parts of the paper, as input. Editing the output, aside from concatenating different pieces, is not allowed. The Critique section has to include the following: How you processed the paper PDF to input to the LLM. Your final prompts, and a detailed explanation about how you came up with them. A detailed critique of the LLM-produced essay, identifying where it made good points about the research paper, made errors, or missed important problems with the paper. Reflect on what you learned during the process, about the functionality of LLMs, and the process of scientific review.

1 LLM-produced Essay

First I uploaded this selected paper to GPT-4 Interface. Then I used four self-designed prompts to generate response essay from main contents and appendix delicately.

Prompt 1: The input to the GPT-4 above is a pdf version of my selected research article. Please first read the whole document and next write a response essay. The critique should summarize every section that is mentioned in this paper (you must contain factual words if possible). After that, give your evaluation / opinions on this paper also section by section. Include some details e.g. numbers if possible.

GPT-4: I have carefully read the research article you provided and here's a response essay that summarizes and evaluates each section of the paper.

Introduction

Summary: The paper introduces the CAREER model, a machine learning approach to predict job sequences using large datasets of online resumes, addressing limitations in traditional econometric models. This is achieved through a transformer-based foundation model that is first trained on large resume datasets and then fine-tuned on smaller, curated survey datasets.

Critique: The introduction provides a solid foundation, clearly explaining the relevance and novelty of the CAREER model. The authors effectively set up the problem by discussing the shortcomings of existing econometric models and the advantages of using large datasets. However, the introduction could enhance its impact by providing more direct comparisons with existing models in terms of performance metrics.

Methodology

Summary: The CAREER model uses transformer architecture to process job sequences and predict future occupations. The model is pretrained on a large dataset of 24 million job sequences and fine-tuned on three smaller economic datasets to improve prediction accuracy.

Critique: The methodology is robust, leveraging the latest advancements in machine learning to address a traditional economic problem. The use of a transformer model is innovative in this context. However, the section could improve by discussing potential biases in the large datasets used for pretraining and their impact on the model's generalizability.

Results

Summary: The CAREER model outperforms traditional econometric models on three economic datasets. It also provides better predictions of downstream variables like wages, demonstrating its utility in economic forecasting.

Critique: The results are impressive, showing clear quantitative improvements over existing models. Nonetheless, the authors could enhance this section by discussing the limitations of the results, such as conditions under which the model might perform poorly and the computational costs associated with training and fine-tuning the model.

Discussion

Summary: The discussion explores the implications of the CAREER model for labor economics, including its potential to influence economic policy and research. The authors also suggest future research directions, such as extending the model to other types of sequence prediction tasks.

Critique: This section effectively highlights the broader impacts of the research, which adds depth to the paper. However, it would benefit from a more critical examination of the ethical considerations, especially regarding data privacy and the potential for biased decision-making if the model were deployed in real-world scenarios.

Conclusion

Summary: The paper concludes by summarizing the key findings and the advantages of the CAREER model over traditional methods. It reiterates the potential of machine learning to transform economic modeling and analysis.

Critique: The conclusion succinctly wraps up the paper, reinforcing its contributions. It could be strengthened by providing a clearer call to action for policymakers and practitioners in the field of labor economics.

Overall Evaluation

This paper is a significant contribution to the field of labor economics, introducing a novel machine learning approach to an old problem. The innovative use of transformer architecture to model job sequences offers a promising avenue for future research and application. However, for a more comprehensive assessment, the paper would benefit from addressing potential biases in the training data and discussing the ethical implications of deploying such models in practice.

Prompt 2: I think the generated introduction part is appropriate. However, I think the generated methodology summary part might miss some technical details. For example it utilized a two-stage occupation model, and the description of the computation graph (parameterization). Besides, you also didnt comment on the computational complexity and transfer learning capability of CAREER model. Thirdly, you didnt include the related work part where you should comment on the comparison between CAREER and previous models. In the results part, you miss the description of 1) pretraining process 2) dataset 3) model comparison 4) parameter settings 5) evaluation 6) extension of CAREER to wage prediction. In the conclusion part, your summary is too short and should include all important parts above. Based on these preliminary feedback, regenerate the essay.

GPT-4: Here's a more detailed and comprehensive response essay that includes critiques and evaluations for each section of your research article, incorporating the technical details and elements you noted were missing in the previous summaries.

Introduction

Summary: The introduction outlines the challenges faced in labor economics with traditional econometric models and presents the CAREER model, a machine-learning-based model leveraging large-scale resume datasets for predicting job sequences. The CAREER model employs transformer architecture, aiming to improve predictive accuracy by learning comprehensive job representations.

Critique: This section effectively frames the research question and positions the CAREER model as a significant advancement over traditional methods. However, it would benefit from a brief discussion on the specific limitations of current econometric models in capturing complex career trajectories, providing a clearer contrast against which CAREER's advancements can be evaluated.

Methodology

Summary: The methodology employs a two-stage occupation model with a transformer-based architecture. The model uses a computation graph to parameterize job sequence data, enabling the representation of complex relationships in career paths. The model first predicts job change likelihood and then the specific job transition, incorporating learned representations from large-scale resume datasets through transfer learning.

Critique: The methodology is comprehensive and innovative in its approach to job sequence modeling. The technical descriptions are detailed, particularly the use of the transformer architecture and its adaptation for career data. However, the discussion on the computational complexity of the CAREER model and its scalability could be expanded. Additionally, more details on the parameter settings during training would clarify the model's operation and potential limitations.

Related Work

Summary: This section compares CAREER with traditional econometric models and other recent machine learning approaches. It highlights how CAREER's methodology addresses the shortcomings of these models by utilizing advanced machine learning techniques and large-scale data.

Critique: While the section provides a good overview of how CAREER advances beyond existing models, it could benefit from a more detailed quantitative comparison, such as performance metrics or scenarios where CAREER significantly outperforms other models. This would provide a clearer understanding of its practical advantages.

Results

Summary: CAREER outperforms traditional models in predicting job sequences and is extended to wage prediction, demonstrating its utility beyond its primary function. The model's performance is validated across several datasets, showcasing its effectiveness in real-world scenarios.

Critique: The results are robust and showcase the model's capabilities well. However, including specific performance metrics, such as accuracy improvements and computational efficiency, would provide a more precise measure of its impact. The extension to wage prediction is particularly noteworthy and deserves a more detailed analysis of its implications for economic forecasting.

Conclusion

Summary: The conclusion summarizes the benefits of the CAREER model, emphasizing its ability to leverage large datasets and advanced machine learning techniques to improve prediction accuracy in labor economics.

Critique: The conclusion effectively encapsulates the study's contributions but could be expanded to discuss potential future research directions, such as applications of the CAREER model in other fields of economic forecasting or its integration into policy-making processes.

Overall Evaluation

The paper makes a significant contribution to labor economics through the development of the CAREER model. It utilizes cutting-edge machine learning techniques to address complex issues in employment data analysis. However, enhancing the discussion on computational aspects, providing more comparative data in the related work section, and expanding the results discussion would strengthen the paper.

Prompt 3: I won't tell you where you are wrong when you are generating this essay. But I will give you an official review from OpenReview where it is accepted to TMLR, where this reviwer gave most of the correct opinions on this paper. Based on this and previous essay, regenerate essay again.

Reviwer : Summary Of Contributions: This paper presents an approach for improving predictions for econometric models via pretraining using semi-structured data derived from a resume database with finetuning on longitudinal survey datasets. A transformer-based model architecture is adapted for categorical data modalities and is shown to improve performance over previous methods when predicting and forecasting occupations. In addition, the internal representations learned through the pretraining approach are shown to be beneficial for another downstream economic prediction task.

Strengths And Weaknesses: Strengths

The proposed algorithmic approach for developing informative representations of career history (via what is called CAREER) is well grounded in prior use of transformer models in a pre-training/fine-tuning paradigm. The adjustments made to incorporate categorical data sequences through initial embedding layers is interesting, although seems to be fairly limited by how expressive these embeddings can be as they are effectually a key-value store indexed by integer quantities.

I appreciate the delineation made by the authors between expansive resume databases and carefully curated longitudinal econometric occupational surveys. It was made apparent (while the connotation could be made more explicit and direct) that policy decisions and generalizable claims cannot be made from the use of these large resume databases as they are not representative of the population at large. What is missing from this very clear dichotomy is a good, clear and direct discussion of why econometric models need well curated survey datasets. I'll point to additional issues with the presentation of the data/problem formulation below.

The paper does a nice job outlining prior approaches and methodologies for occupation prediction within economics including some interesting recent approaches such as NEMO that are built to learn from unstructured resume datasets.

For the most part, the setup and description of how the proposed CAREER model is constructed is clearly understood. Perhaps one addition to Figure 1 would be to include annotations of the FFN between and similar to how the attention weights are shown.

There is a variety of experiments to demonstrate the utility of using pretrained representations, adapted for the occupation prediction task. I was especially glad to see the experiment represented in Figure 2B that shows the effect on pretraining volume on final model prediction performance. The forecasting and wage regression experiments were also very interesting and good demonstrations of how the pretrained representations can be useful (despite some major questions that arose from how they were presented, details below).

Weaknesses

I fear that there are several relatively major weaknesses that lessen the excitement I have for this paper and the presented empirical results. Some are easily fixed (more thorough and precise definitions and writing) while others may take a fair amount of effort on the authors' behalf (additional baseline experiments and model ablations). As a summary, there are too few details about the motivating foundations of the work, the datasets used, the modeling approach, the experimental setup and the analyses presented. As a result I am not entirely sure that I can confidently state that the claims made are adequately supported by the paper. In the following I will try to highlight where I am concerned and how the authors may be able to revise the paper to help improve my estimation of the work. This being said, I feel that the insights provided by the paper are good, it can however be greatly improved through more precise language and thorough empirical validation. This work will be of interest to the community using ML within economics as well as the broader ML community for yet an additional domain where scaling pretraining enables improved predictive with transformer models.

Motivating foundations

I found the writing and descriptions about the econometric setting to be insufficient and vague. There are not enough details shared about the kinds of predictive tasks that are prevalent in this community and why they are important. This extends to any motivations about the datasets that are used and how analyses of these survey datasets support claims and potential policy decisions (if any are to be had). Greater clarity about the domain specific usage of the data and why economists take such great effort to develop survey datasets would go a long way to improve the apparent significance of the contributions made by this paper.

Many of the motivations behind the experiments are vague and seem to rest on a solitary example without additional information or detail about why the experiments are chosen and how informative they may be in view of the intended use case.

Datasets

It's not clear how the datasets are formed, what the size of J or (the number and scale of the covariates), etc. are. The discussion or information provided about the datasets (both the resume and longitudinal survey data) is surprising. Without these details it is unclear what is actually being modeled or the overall scale of the problem. For example, how do the data distributions of job categories or types compare between the pretraining data and each of the survey datasets? Is J consistent across the whole paper? If not, what kinds of support can be found in the pretraining data that overlaps with what is represented in the survey data? This isn't a critical omission as the methods used to construct CAREER and analyze are the clear focus of the paper, but I would rate the omission as a major issue that should be corrected to improve the presentation of the paper.

In connection to the concerns raised previously, there is no indication of how long the career sequences are in each dataset on average (e.g. what is the distribution of T?). I know that the perplexity metric tries to average over this component but it could be a major factor in the forecasting performance as errors will compound for longer career durations. (This is a shared concern with the analyses provided in the paper) If it's true that some of the datasets have longer sequences than others, that could explain why the finetuning task takes longer on datasets with equivalent numbers of individuals.

It is claimed at the end of Section 3 that this paper "identif[ies] a novel pretraining corpus". This isn't necessarily true, is it? It is stated that "NEMO, an LSTM-based model that is trained on large resume datasets". While NEMO may not be immediately applied to the pretraining/finetuning paradigm, it may not be wholly accurate to state that the use of the same (or similar) dataset for pretraining is novel.

Modeling approach

It is very unclear how the predicted job change indicator is used to inform the input at the next timestep as indicated in Figure 1.

There's an asymmetry in how the covariates are introduced and described and how they are ultimately used in the experiments. In fact, it's unclear between experiments what covariates are used and if they are at all. This is not insignificant. This inconsistency is a problem that should be addressed by the authors if this paper is to be considered for publication.

Are there any additional regularizations used to keep the representations from collapsing to trivial modes? With the amount of data and number of model parameters, there is some risk for model collapse. This highlights one of my complaints of the presentation of this paper. Several times the representations are claimed to be "complex" or "strong" but there is no precision or justification of those terms. It appears that this nomenclature is used implicitly from the complexity of the model architecture. There is no analysis of the representations and what is being encoded within them. The only indication that we have that anything is being learned is model performance. There are no clear indicators that the representations differentiate between different career paths or how the effect of individual covariates have on the representation. Without any analysis or justification of how the representations are formed and what they encode, I cannot place any weight on the language or claims made about them. It's highly possible that pretrained representations using NEMO could be equally as performant on the dataset not to mention more rudimentary clustering approaches. This significantly impacts how I view the empirical support of the claims made by the paper.

Experimental Setup

It would be very useful to recast the apparent strength of the transformers learned representations in the context of additional predictive tasks (beyond job prediction) to better compare with prior approaches outlined in Section 3. With the two-stage regression technique of CAREER, occupational change is already modeled. It would be instructive to see how that sub-task performs relative to the the cited approaches of Kambourov & Monovskii (2008) and Guvenen et, al. (2020).

As mentioned above, there are concerns about whether the apparent gains seen with CAREER can be contributed mostly to the transformer architecture more than the pretraining approach. There are no empirical comparisons with an alternative architecture that is provided the same training / evaluation procedure (two-stage regression from the embedded representation of career path). The suggested NEMO pretraining baseline could be a way to validate the use of a more complex transformer architecture.

An additional baseline that would help reinforce the overall empirical approach used in the paper would be to pretrain CAREER but not finetune. To demonstrate the zero-shot capabilities of the model based on the resume data alone. This could open up a really interesting analysis of the questions raised above about the representative support of the pretraining dataset (e.g. are there certain individuals that are included in the survey data for which not finetuning does well on? Are there inequalities that arise without the finetuning step? Do those inequalities persist after finetuning?)

It would really help if the objective or loss function was formally included in the main paper. This would help round out Section 2.3.

I found the ordering of Section 2 to be a little confusing. The most important formulations of the CAREER model are presented in Section 2.2 where there is not much context behind how they would be used. Then after so much information is given in Section 2.3 about architecture and training approach, there is a subtle reference to the previous subsection. I think that the Section 2 would be improved if the order of Section 2.2 and 2.3 were swapped. I also think that the "Transfer Learning" subsubsection could be elevated into a new Section 2.4 to help emphasize the contributions and claims made in the paper.

The specific tasks "career prediction" and "forecasting" are not adequately defined at the outset of Section 4. This is another example where just a little more detail would go a long way. This issue occurs again when the downstream auxiliary tasks (wage prediction in this paper) are introduced. There is reference to multiple types of economic analyses that could be run but there is no mention of what these are. I found a lot of the language around these type of context-setting portions to be frustratingly vague.

Experimental Analysis

While a helpful comparative metric, perplexity is difficult to gain any insight from. I believe that including additional metrics such as AUROC/ AUPRC or F1 score would be far more informative of the quality of the predictions being made by the compared models. This is particularly true when considering the forecasting task presented in Table 1. I have no way to assess how reliable the models are as perplexity gives me no indication of how immediately accurate the predictions are.

It's not sufficient to only report the mean performance of the prediction metrics in Tables 1 and 2 without confidence intervals as is done in the table contained in Figure 2A. Without these the significance of the performance improvement seen with CAREER cannot be assessed. Their exclusion seems somewhat intentional?

While NEMO is included as a baseline, it's unclear whether it was pretrained using the resume dataset as it was developed for. So, it's not entirely clear whether its current use is a fair comparison of the model's performance. It would be instructive to see if NEMO (and other baseline modeling approaches) would have better performance than their vanilla counterparts if they were additionally exposed to more data in a pretraining setting and how that would compare after finetuning to the variants of CAREER presented in Figure 2A. Without this baseline, it's unclear whether the performance gains are drawn from the architecture (e.g. model complexity/expressivity) or the training paradigm. This is perhaps the primary major concern of mine regarding the evaluation supporting the claims made in the paper.

Given the amount of discussion around the assumed benefit of using a more complex model that breaks down prior work's Markov assumption, it is surprising that the authors did not analyze the attention weights to demonstrate any evidence of their claims on this premise. It would be really interesting to see how far the prediction of the current time step relies on previous employment and the sensitivity of the prediction on the included covariates. Without any analysis along these lines, I cannot trust the claims made about the necessity of non-Markovian modeling. I cannot just take the claim at face value because of the prediction metrics are "better". We can expect there to be more rigor in the experiments to validate the claims made.

Requested Changes: Major

More detail about the datasets (especially the demographic features used as covariates in all experiments).

Additional prediction performance metrics such as AUROC/AUPRC or F1 score to evaluate the quality of the compared models.

Additional NEMO baseline that is pretrained on resume database, then finetuned on the survey datasets.

Additional baseline removing covariates from CAREER.

Some analysis on the "strength" of the representations learned rather than some implied validity of their construction based on the complexity of the model.

A model ablation of CAREER without the covariates included.

Further analysis of the claims made about the necessity of the transformer architecture (using information over the whole sequence of career history).

Minor

More clarity in writing about the domain and the specifics of the kinds of predictive analyses being used with the data. This would be to better frame the contributions of the proposed model.

CAREER is clearly an acronym (as stated in footnote 2). It should be spelled out as early as the modeling name/approach is used in Section 1 to help frame the contributions and claims made at the end of the introduction.

Additional experimental analysis comparing CAREER's prediction of occupational change with prior approaches.

Restructuring Section 2 as outlined above to improve clarity.

A baseline of zero-shot prediction performance of the the pretrained CAREER model without finetuning.

GPT-4: Combining the detailed critiques from both the original summary and the OpenReview feedback, here is a unified and comprehensive critique essay that encapsulates the full scope of the paper's contributions, methodology, and areas for improvement.

Introduction

Summary: The introduction frames the research problem by discussing the limitations of traditional econometric models in capturing complex career trajectories and introduces the CAREER model, which leverages large-scale online resume data to predict job sequences using a transformer-based architecture.

Critique: The introduction effectively sets up the problem and the novelty of the CAREER model. It would benefit from more detailed examples of the specific limitations of current models and a clearer rationale for the necessity of the transformer approach in addressing these limitations.

Methodology

Summary: The methodology employs a two-stage occupation model utilizing transformer architecture tailored to handle categorical career data. The model predicts the likelihood of job changes and specific job transitions, incorporating learned representations from extensive resume datasets.

Critique: The description of the methodology is comprehensive, highlighting innovative adaptations for categorical data through embedding layers. However, the explanation of how the binary job change indicator st influences the input at the next timestep could be improved for clarity. More detail on the computational graph and parameterization would also help in understanding the model's complexity and operation.

Related Work

Summary: The paper provides a thorough review of both traditional econometric models and contemporary machine learning approaches, establishing the groundwork for the CAREER model's advancements.

Critique: While the related work section effectively contextualizes CAREER within the broader research landscape, it could further benefit from a more direct comparison with existing models, specifically discussing how CAREER's methodology improves upon or diverges from these approaches. This could include more quantitative comparisons and performance metrics.

Results

Summary: CAREER outperforms traditional models in predicting job sequences and is successfully extended to wage prediction, demonstrating the versatility and utility of its pretrained representations.

Critique: The results are compelling but require a deeper analysis, as pointed out by the reviewer. Including additional performance metrics such as AUROC, AUPRC, or F1 scores would provide a clearer picture of the quality of predictions. The paper should also address how career sequence length and job category distributions affect the model's performance across different datasets.

Conclusion

Summary: The paper concludes by summarizing the benefits of the CAREER model, highlighting its potential to transform econometric modeling through advanced machine learning techniques.

Critique: The conclusion succinctly wraps up the paper but could be expanded to discuss broader implications for policy-making and future research directions. This would provide a clearer call to action and suggest practical applications of the CAREER model in other domains.

Overall Evaluation

This paper is a valuable contribution to the intersection of labor economics and machine learning, presenting a novel approach to modeling job sequences with transformer-based architectures. However, the paper could be significantly strengthened by addressing the following points:

1. Detailed Dataset Descriptions: More comprehensive descriptions of the datasets, including job category distributions and career sequence characteristics, are crucial.

2. Model Analysis Depth: There is a need for a detailed analysis of how the model's embeddings capture career trajectories and the impact of covariates on predictions.

3. Experimental Rigor: Including additional baseline models and model ablations would substantiate the claims about CAREER's effectiveness more robustly.

4. Empirical Validation: Enhancing empirical validation through varied performance metrics and a detailed error analysis would solidify the model's claimed advantages.

By addressing these areas, the paper would not only clarify its contributions but also enhance its impact and applicability to both economic research and practical policy formulation.

Prompt 4: The input to the GPT-4 above is a pdf version of my selected research article. Please read the appendix and next write a response essay sorely on the appendix. The critique should summarize every section that is mentioned in this the appendix. After that, give your evaluation / opinions on the appendix also section by section. Include some details e.g. numbers if possible.

GPT-4: I have read the appendix of your provided article which contains detailed descriptions and evaluations of the CAREER model, a machine learning approach developed to predict job sequences using large-scale resume data and fine-tuned with smaller survey datasets. Here's a structured critique of the appendix, summarizing and evaluating each section:

Summary and Critique of Each Section

CAREER Model Overview Summary: CAREER is a transformer-based model designed to handle large volumes of job sequence data from online resumes and fine-tuned with survey data. This approach aims to improve predictions of job sequences over traditional econometric models.

Critique: The model leverages modern neural network architectures for a novel application, bridging a gap in labor economics analysis. However, the reliance on large, passively collected datasets might introduce biases that are not fully corrected in the fine-tuning phase, potentially limiting the generalizability of the model.

Data Description

Summary: The appendix details the datasets used: 24 million job sequences from resumes and several smaller, curated survey datasets. It explains the preprocessing steps, such as how job sequences were encoded and how covariates were handled.

Critique: The description is thorough, providing a clear picture of the data's scope and preparation. The treatment of missing data and the transformation of raw resumes into structured inputs are well-justified. However, the appendix could benefit from a deeper discussion on the implications of the data's passively collected nature.

Model Architecture

Summary: The CAREER model uses a transformer architecture, originally designed for text data, adapted here for job sequences. This includes modifications like a two-stage prediction process to enhance job transition predictions.

Critique: The adaptation of transformer technology is innovative, particularly the two-stage approach to predict job changes. Nonetheless, the complexity of the model may pose challenges in terms of computational demands and transparency in how decisions are made.

Training Procedure

Summary: The appendix details the training process, including pretraining on resume data and fine-tuning on survey data. It mentions the use of stochastic gradient descent and hyperparameter settings.

Critique: The training methodology is robust, using established techniques in machine learning. However, the transition from pretraining to fine-tuning is critical and could be better addressed, especially how the model adjusts from the noisy resume data to the more reliable survey data.

Evaluation and Results

Summary: CAREER is evaluated against traditional econometric models and other baseline machine learning models. The results show that CAREER outperforms these models on job sequence predictions.

Critique: The comprehensive evaluation illustrates CAREER's effectiveness. However, the results section would benefit from more detailed statistical analysis and discussion on the practical significance of the improvements in prediction accuracy.

Discussion and Future Work

Summary: The appendix discusses the implications of CAREER's performance and suggests areas for future research, including exploring other datasets and refining the model architecture.

Critique: The discussion is insightful, outlining potential impacts on labor economics and policy making. However, it might also be beneficial to address more directly the limitations of the current approach, such as potential overfitting to specific types of job sequences or demographic groups.

Overall Evaluation

The appendix provides a detailed and technically sound description of the CAREER model. It effectively demonstrates the model's capability to improve job sequence predictions, which could have significant implications for labor economics. However, some areas, particularly around the potential biases introduced by the initial large-scale data collection and the complexity of the model, could be further elaborated to strengthen the understanding and applicability of the research.

2 Critique

Notice: How you processed the paper PDF to input to the LLM. Your final prompts, and a detailed explanation about how you came up with them. A detailed critique of the LLM-produced essay, identifying where it made good points about the research paper, made errors, or missed important problems with the paper.

I used GPT-4 in priority because GPT-4 enabled the full context as input and use OCR to firstly transform the PDF to text and then perform inference.

Critique on Answer 1

My first prompt is that "The input to the GPT-4 above is a pdf version of my selected research article. Please first read the whole document and next write a response essay. The critique should summarize every section that is mentioned in this paper (you must contain factual words if possible). After that, give your evaluation / opinions on this paper also section by section. Include some details e.g. numbers if possible."

The first answer returned by GPT-4 has shown its strength in summarizing paragraphs. It successfully mentioned the transfer learning and transformer architecture which are the core ideas of this paper. It has pointed out that the transformer model is innovative under this new circumstance. However, it lacks some in-depth analysis about the model architecture. Only mentioning the name of transformer model didn't bring information of inner architecture design, as well as the prediction target. It did not mention the training framework including which datasets they used, the baseline models that CAREER is compared with, and the important hyperparameters that they used. GPT-4 showed its concern on private and ethical aspects of this research, and recommend the authors to provide a call to policymakers in labor economics in order to test its feasibility in real-time prediction. I agree with this point and I am also concerned since for me it is similar to stock price prediction which might be unpredictable in real life.

Critique on Answer 2

My second prompt is that " I think the generated introduction part is appropriate. However, I think the generated methodology summary part might miss some technical details. For example it utilized a two-stage occupation model, and the description of the computation graph (parameterization). Besides, you also didnt comment on the computational complexity and transfer learning capability of CAREER model. Thirdly, you didnt include the related work part where you should comment on the comparison between CAREER and previous models. In the results part, you miss the description of 1) pretraining process 2) dataset 3) model comparison 4) parameter settings 5) evaluation 6) extension of CAREER to wage prediction. In the conclusion part, your summary is too short and should include all important parts above. Based on these preliminary feedback, regenerate the essay."

So the prompt is given by the feedback from prompt 1. I asked the GPT-4 to improve its generation by adding comments on algorithmic side of the model, the model training process and its evaluation. This time the methodology part has improved a lot by including the description of computation graph and likelihood. It finds the drawbacks of the original paper with few discussion on computation complexity and scalability. At the first glance, it might be wrong to claim that hyper-parameters should be clarified during training. However, when I looked into the appendix, actually in part C, F and G, they have mentioned their settings for the experiments. This is not a mistake therefore later in prompt 4 I asked it to generate a summary based on the appendix content. In results part, again it wrongly stated that the paper didn not mention the performance metrics, which shows the limitation of GPT-4 to read tables in the PDF file. And it said the extension to wage prediction and the model's implication for economic forecasting is not mentioned, however it actually appears in part 4. The reason might be that GPT-4 didn't re-read the paper again but just generate the contents that I mentioned (copy and paste work).

Critique on Answer 3

My third prompt is that "I won't tell you where you are wrong when you are generating this essay. But I will give you an official review from OpenReview where it is accepted to TMLR, where this reviwer gave most of the correct opinions on this paper. Based on this and previous essay, regenerate essay again." together with the review from OpenReview.

The review for this paper is valuable. I found that this paper is accepted by TMLR where its review is public and is on the OpenReview website. I passed the first reviewer's opinion and asked the GPT-4 to regenerate the essay. For the methodology part it additionally said how the binary job change indicator influences the input needs to be explored more explicitly. In the related work, I think it mistakenly took the results as the related works so it asked to include comparison of different models. Importantly, it points out different metrics such as F1, AUROC could be used apart from perplexity. It also concerned how the data distribution could affect the final results. This is true because sometimes out-of-distribution (OOD) data could make the results worse. Again it didn't include many detailed as I wanted so in my opinion GPT-4 attempted to generate succinct, abstract and general texts other than complex, concrete and specific summary, even if I give it more detailed reviews and my personal detailed suggestions.

Critique on Answer 4

My fourth prompt is that "The input to the GPT-4 above is a pdf version of my selected research article. Please read the appendix and next write a response essay sorely on the appendix. The critique should summarize every section that is mentioned in this the appendix. After that, give your evaluation / opinions on the appendix also section by section. Include some details e.g. numbers if possible."

We are asked to generate a summary for the appendix. So I pass this prompt to GPT-4. It made good points on stating the bias caused by the reliance on large, passively collected datasets. Overall its great in summarizing all viewpoints from each paragraph. It didn't talk too much about how transformer propagates in the model architecture part. But is successfully mentioned the computational complexity which lacks in the analysis for main contents. I agree with that CAREER is robust since it could be extended to wage prediction. It suggests to explore more datasets and refine the model architecture, and solve the potential over-fitting problem. Overall, the generated summary based on prompt 4 is somehow the best one.

3 Reflection

Learning from Prompting

So firstly we need to start with some simple prompt words like "summary" or "essay". Next, based on the first generation, we refine the prompts by something which is missing potentially in its analysis. Thirdly, we have to include the appendix as additional input since it includes more details about model architecture, experiment settings, explanations about the experiments, and future works as well.

Functionality of LLMs

So from my previous research experience, it could translate one language to another language. From this second essay I learned that it could summarize the paper. It might also generate images if LLM is ChatGPT since it's integrated with DALLE. Maybe it could also correct the grammar or mis-spelling words. But the quality of the generated content would be related to the LLM that we used. So future work might be changing from GPT-4 to PaLM or Llama for testing the robustness of LLMs.

Process of scientific review

I will first state my understanding and writing style of scientific review. Very often I would first state one fact about the paper. And I would stand or oppose this fact with some arguments. I have experienced reviewing machine learning conference papers before, and as a reviewer I am required to give a summary first, then its advantages and disadvantages with facts, and then point out some minor concerns with some scoring. This review would be passed to the authors for rebuttal or a giving opportunity to revise the paper. So its very important and overall I do not recommend using GPT-4 to generate review as it's irresponsible at all. If you have seen the generation quality above, you would also not recommend it at all.