EVALUATING LANGUAGE MODELS' ABILITY TO CAPTURE CAUSAL EFFECTS THROUGH TASK-BASED LEARNING

Toker Gilat, Iedovnik Omer
Technion, Department of Data Science and Decision
Final Project 097400 -Causal Inference

November 7, 2024

Introduction of the Problem

In recent years, the field of Natural Language Processing (NLP) has rapidly evolved, driven by advancements in machine learning and deep learning. These advancements have significantly expanded the capabilities of large language models (LLMs), enabling their widespread application across sectors such as finance Nie et al. (2024),, education Kasneci et al. (2023), healthcare Nazi and Peng (2024), and law Shui, Cao, Wang, and Chua (2023), where decision-making processes are often highly sensitive and critical. This rapid expansion underscores the importance of LLMs effectively learning from data in a way that captures real-world connections, ensuring genuine causal relationships rather than spurious or biased correlations, which could lead to unintended discriminatory outcomes.

Causal inference plays a crucial role in identifying the true impact of various factors on model outcomes, particularly in domains where fairness and accountability are paramount. For instance, in the hiring process, it is important to understand whether attributes such as gender or education genuinely affect a candidate's evaluation or whether the model is inadvertently amplifying biases present in the training data. By leveraging causal inference, we can make explicit claims about the effect of changing one variable while holding others constant, thereby isolating the true causal impact of each factor.

The importance of causality in NLP has been widely emphasized in recent research. There is an urgent need for robust causal inference methods to enhance both the reliability and interpretability of NLP models Amir Feder and Yang (2022); Sevastjanova, Amara, and El-Assady (2024). Ensuring that model explanations are faithful requires establishing causality Amir Feder and Yang (2022). For example, Yair Gat (2023) has theoretically shown that non-causal methods often fail to provide faithful explanations. Effective strategies for achieving such explanations include using causal inference techniques like counterfactuals (Amir Feder, 2021), interventions Zhengxuan Wu (2023) adjustments Zach Wood-Doughty (2018), and matching Raymond Zhang (2023).

This project aims to evaluate the ability of language models to capture and learn the underlying causal relationships between various factors that influence task outcomes. To achieve this, we will create a controlled textual dataset by defining relevant aspects, such as gender, race, education, and experience, and mapping their relationships in a causal graph to generate corresponding text. The dataset will consist of resumes constructed based on these aspects, allowing us to control the distributions and relationships within the data. We will then evaluate whether the model implicitly learns and recognizes the causal effects of these aspects when exposed to counterfactual data.

The Chosen Dataset - LIBERTy

To investigate the causal question, we chose to create our own controlled synthetic dataset. This decision allowed us to design the data precisely to meet our research needs, but it also required substantial effort to ensure the quality and validity of the generated data. In the following sections, we will explain each step of this process in detail.The dataset we created is focused on resume screening. This task was chosen to explore how models learn causal relationships, particularly concerning demographic and professional variables in hiring decisions. The dataset will consist of textual data representing resumes of various candidates, each aligned with specific aspect values, paired with corresponding counterfactuals (CFs) across these aspects. This setup will enable us to compare the model's predictions on both the original and CF data, assessing the model's ability to learn the causal effects of each aspect relative to their true impact.

The key aspects in our data are:

- **Gender**: {0: 'Female', 1: 'Male'}
- **Age Group**: {0: [24,32], 1: [33,44], 2: [45,55]}
- **Race**: {0: 'Black', 1: 'Hispanic', 2: 'White', 3: 'Asian'}
- **Socioeconomic Status**: {0: 'Low', 1: 'Medium', 2: 'High'}
- **Volunteering**: {0: 'No', 1: 'Yes'}
- **Certificates**: {0: 'No', 1: 'Yes'}
- **Education**: {0: "High School", 1: "Bachelor's degree", 2: "Master's degree", 3: "Doctorate degree"}
- **Work Experience Group**: {0: [2,5], 1: [6,10], 2: [11,25]}

An important note, which we will elaborate on later, is that we utilize Individual Causal Concept Effect (ICaCE) calculations to measure the causal effect of each aspect by leveraging counterfactuals (CFs) as a core component of this measurement. We want to ensure that the dataset includes authentic CFs, which are crucial for accurately demonstrating how different aspect values might alter outcomes in parallel simulations. To achieve this, we have developed a structured pipeline that integrates causal analysis and text generation, enhancing both the integrity and fidelity of the data.
Our data generation process follows a meticulously structured five-step approach:

1. **Causal Graph Construction:** At the core of our dataset is a causal graph that encapsulates the generative process, linking all essential aspects and the label variable through functional relationships. We use Directed Acyclic Graphs (DAGs) to

visually represent these relationships between candidate attributes, such as gender, education, and experience. The DAGs serve as a structured way to encode our assumptions about causal mechanisms, ensuring a comprehensive understanding of how different factors influence the outcome variable—whether a candidate is deemed suitable for a role. These functional relationships are modeled with a degree of randomness to simulate the complex dynamics of real-world data generation.

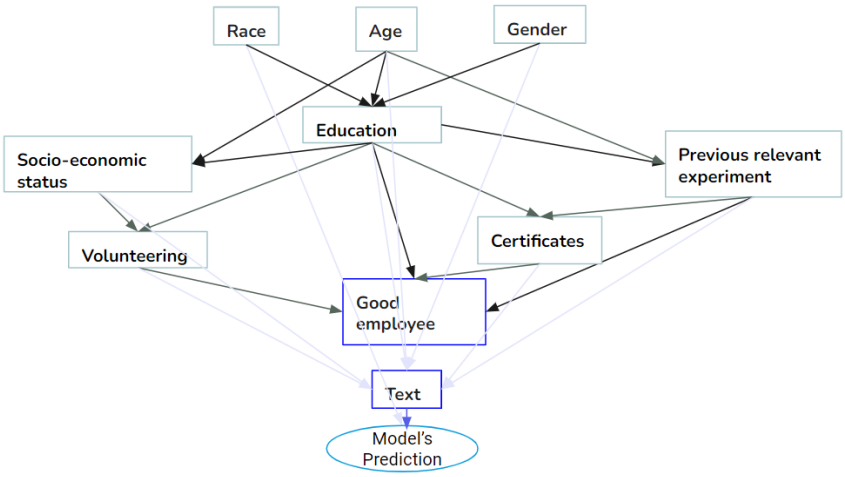


Figure 1: A Directed Acyclic Graph (DAG) representing the relationships in our dataset between different attributes of candidates.

2. **Aspect Value Sampling:** Based on the causal graph, we create a table that lists the various aspects and their corresponding values, which are sampled accordingly. This sampling process ensures that the generated resumes cover a broad range of combinations, thereby providing diversity in the dataset while maintaining the structural relationships encoded in the causal graph. The sampling function used is designed to preserve the causal dependencies modeled in the graph while ensuring representative coverage of different value.
   The functions used for sampling is as follows:

   $$\text{Race} = \text{random}(0, 1, 2, 3)$$

   $$\text{Gender} = \text{random}(0, 1)$$

   $$\text{Age} = \text{random}(0, 1, 2)$$

   $$\text{Education} = \min(3, \max(0, \text{round}(0.5 * \text{race} + 0.1 * \text{gender} + 0.1 * \text{age} + N(0.25, 0.25))))$$

   $$\text{Socioeconomic Status} = \min(2, \max(0, \text{round}(0.4 * \text{education} + 0.2 * \text{age} + N(0, 0.25))))$$

   $$\text{Previous Experience} = \min(2, \max(0, \text{round}(0.5 * \text{age} + 0.2 * \text{education} + N(0, 0.7))))$$

   $$\text{Volunteering} = \min(1, \max(0, \text{round}(0.3 * \text{socio} + 0.2 * \text{education} + N(-0.35, 0.2))))$$

   $$\text{Certificates} = \min(1, \max(0, \text{round}(0.5 * \text{previous experience} + 0.4 * \text{education} + N(0, 0.5))))$$

   $$\text{Good Employ} = \min(2, \max(0, \text{r}(0.3 * (\text{certificates} + \text{education} + \text{experience} + \text{volunteer})))$$

3. **Human-like Template Direction:** To generate content that closely mirrors natural human text, we utilized task-specific auxiliary data containing real personal statements. In our case, we chose personal statements for college applications as the human template. Why? Because both college applications and resume screening require candidates to highlight their relevant history, motivations, and fit for the position or program. However, since these are two different domains, we passed the personal statements through a language model to produce a tailored template that provides candidates with a framework on how to formulate a cover letter for their resume. To create a dataset of such instructions, we performed web scraping from the website - **Link**. Each sample was then processed through Gemini, a language model that adapts the personal statements into structured templates suitable for resume-related narratives.

   Here is an example of a template generated by our process:
   "Key Points:

   - ***Opening Hook:*** *Starts with a powerful quote to introduce the overarching interest in psychology.*
   - ***Motivating Experience:*** *Uses a personal experience (Auschwitz trip) to highlight a specific area of interest within Psychology (human behavior).*
   - ***Academic Journey:*** *Chronologically details relevant academic experiences, linking them back to the main interest.*
   - ***Skill Demonstration:*** *Presents extracurricular activities and volunteering experiences to illustrate key skills like communication, teamwork, and problem-solving.*
   - ***Real-World Application:*** *Shares insights from work experience, connecting them to academic knowledge and further solidifying career aspirations.*
   - ***Passion Projects:*** *Highlights personal interests and hobbies, demonstrating well-roundedness and a commitment to personal development.*
   - ***Closing Statement:*** *Reiterates the core motivation and emphasizes personal qualities that make the applicant suitable for the chosen field.*

   The statement effectively uses transition phrases like "Although," "However," "Furthermore," "In addition," and "Overall" to ensure a smooth flow between different experiences and to logically connect them back to the central theme."

4. **Persona Direction:** To promote diversity and ensure meaningful variance across the dataset, we needed to design personas that represent different professional contexts and backgrounds. Personas allow us to create in-depth narratives, making the text generation richer and more aligned with real-world scenarios. To create these personas, we designed a systematic process that begins by selecting a profession from a curated bank of occupations. Once a specific role is selected, we employ an LLM—Gemini—to enrich the persona with key details. The persona is built to include:

   - Motivation for Career Choice: This provides insight into what drives the individual to pursue their career. It adds a layer of personal context and passion that is important when evaluating candidates.

   - Defining Professional Skill: Here, we define a key skill or talent that makes this persona particularly effective in their profession. This skill is tailored to align with the demands of the role.

   - Family and Job Connection: This element bridges the persona's personal experiences, family background, and professional life, highlighting how their career aspirations are influenced by their upbringing or specific life events.

   Below is an example of a persona we created using this process:
   "Job Title: Biotech Equity Research Associate

   - *Motivation for Career Choice: Driven by a lifelong fascination with the elegance of biological systems and a passion for financial markets, I'm drawn to a career that bridges scientific innovation with sound investment strategies.*
   - *Defining Professional Skill: My strength lies in distilling complex scientific data into clear, concise, and actionable financial insights, allowing me to effectively communicate opportunities and risks to stakeholders.*
   - *Family and Job Connection: While attending a biotech conference with my family, we had a delightful encounter with a scientist whose research I had been following. Discussing their groundbreaking work with them (and my starstruck family) solidified my passion for this industry.*

5. **Text and Counterfactual (CF) Generation:** Using the information from previous steps, we generate both the original text and its corresponding counterfactual (CF) by modifying only the targeted aspect and its causal effects on related aspects, while keeping all other factors constant. To do this, we use GPT-4 to create textual descriptions for each candidate's resume based on the sampled aspect values, persona, and template. This ensures that the original and CF versions are consistent

in structure, narrative style, and content, with only the specific target aspect differing between them. This results in pairs of original and counterfactual resumes, providing a solid foundation for analyzing the model's ability to understand causal relationships.

A crucial aspect of our process is recalculating related attributes when a specific aspect is modified. For instance, if we alter "Gender" from 0 (Male) to 1 (Female), we also adjust the values of other related aspects influenced by this change, based on our causal graph. For example, changing "Gender" may affect related attributes like "Education" or "Volunteering," depending on the causal links encoded in our Directed Acyclic Graph (DAG). This recalculation ensures each CF accurately simulates how a single change can impact other related factors, reflecting the causal mechanisms we aim to study.

The generated texts are produced using GPT-4, resulting in consistent outputs that follow the established personas, templates, and aspect structures. Each generated resume maintains the same format and context, enriched by the persona characteristics and human-like templates defined earlier. Importantly, the only difference between the original text and its CF is the modified aspect and any subsequent cascading changes from our causal model. This precise control allows us to measure the genuine causal impact of changing each aspect, ensuring the reliability and validity of our causal analysis.

Below is an example of a resume from our dataset and its corresponding counterfactual (CF), demonstrating how a specific aspect modification affects the overall content:

"The changed aspects are: Work Experience info: from 5 to 8
Original Resume: *I am driven by a profound fascination with numbers and the intricate dance of data organization. At 28 years old, my journey through the realms of data entry is not just a career path but a personal calling. As a Black/African American male with a Bachelor's degree and five years of relevant work experience, I have honed my skills in transforming chaotic information into structured, accessible formats that empower business decisions.*
During my academic tenure, I immersed myself in courses that sharpened my analytical abilities and deepened my understanding of data management. This educational foundation laid the groundwork for my professional pursuits, where I have excelled in roles that demand high levels of precision and efficiency. My exceptional typing speed and accuracy are not merely technical skills; they are the tools with which I craft clarity from confusion, ensuring that every data point serves its purpose.
In the professional sphere, my role often extends beyond mere data entry; it encompasses the responsibility of maintaining data integrity across platforms. This

task requires meticulous attention to detail—a quality that I've cultivated not only through my career but also within my personal life. My family often remarks on my natural propensity for organization, relying on me to manage everything from important family dates to personal appointments. This personal trait seamlessly enhances my professional performance, ensuring that I approach each task with the same dedication and precision.

My career choice was a natural progression from my innate abilities and interests. The satisfaction I derive from creating order and clarity in data is paralleled by the impact of my work on business operations. Each dataset I refine helps to build a clearer picture for decision-makers, directly contributing to strategic planning and operational efficiency.

As I seek to advance my career, I am eager to bring my expertise in data management to your team. I am confident that my background, skills, and personal dedication to data integrity will make a significant contribution to your company, driving success through meticulous data analysis and management. Let's harness the power of data together, turning raw information into actionable insights that propel your business forward.

CF Resume: *I am driven by a profound fascination with numbers and the intricate dance of data organization. At 28 years old, my journey through the realms of data entry is not just a career path but a personal calling, where each spreadsheet and database represents a puzzle I am eager to solve. As a Black/African American male with a Bachelor's degree and eight years of relevant work experience, I have honed my skills to become a meticulous and efficient data management professional, ensuring precision in every project I undertake.*

During my academic tenure, I immersed myself in courses that sharpened my analytical abilities and enhanced my understanding of complex data systems. This educational foundation laid the groundwork for my entry into the professional world, where I quickly distinguished myself through my exceptional typing speed and accuracy. These skills have allowed me to excel in roles that demand the swift processing of large volumes of data, maintaining integrity and clarity even under tight deadlines.

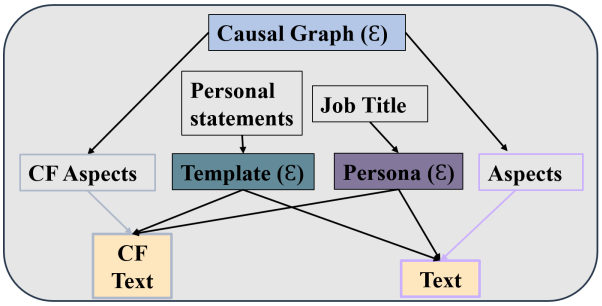
My career has been a testament to my dedication to the field of data entry. Each role I've embraced over the years has not only challenged me but also deepened my appreciation for this profession. From managing extensive databases to implementing efficient data storage solutions, I have consistently demonstrated my ability to adapt and thrive in evolving work environments.

Outside of my professional life, my role in my family has mirrored my career. Known for my reliability in managing personal data, from birthdays to appoint-

ments, I am the cornerstone of organization in my family's life. This personal responsibility reflects my overall approach to data management—meticulous, dependable, and always precise.

As I seek to bring my expertise and passion for data organization to your team, I am excited about the opportunity to contribute to your success. My background, skills, and personal dedication to data integrity position me as an ideal candidate to add value to your operations, ensuring that every piece of data is handled with the utmost accuracy and care."

A visualization of our data generation process for the CV screening dataset is detailed in the diagram below, showcasing the various factors that influence the text and illustrating the intricate connections between them. The entire pipeline for creating the CV dataset can be summarized as follows: It begins with constructing a causal graph that captures all selected aspects and their causal interactions. Next, a resume template is refined using data derived from personal statements. A persona is then crafted based on a sample professional profile. And finally, the text and its corresponding counterfactual (CF) are generated in parallel.



This systematic approach allows us to effectively capture causal relationships between aspects while ensuring consistency and diversity in the generated dataset.

Identification Assumptions and Why They Hold

Our evaluation of language models focuses on their ability to accurately quantify the causal impact of various aspects. We measure this impact by using the model's predictions in original and counterfactual settings, inspired by the methodologies of Goyal, Feder, Shalit, and Kim (2020) and Abraham et al. (2022). Specifically, we calculate the Individual Causal Concept Effect (ICaCE), which quantifies the change in a model's prediction when a particular concept is altered in the text. Initially, ICaCE was aimed at measuring the direct influence of individual concepts; however, with our data generation approach, we extend it to capture both direct and indirect effects, providing a more holistic perspective on how concepts interact to influence model predictions.

Our approach to estimating causal effects relies heavily on counterfactual data generation, rooted in the principles of Rubin's Causal Model (potential outcomes framework). Each resume in our dataset is paired with a corresponding counterfactual version that depicts an alternate scenario where a specific aspect, such as Gender, is changed while all other factors remain constant. This counterfactual methodology enables us to measure the causal effect of modifying specific attributes and evaluate their influence on the model's predictions.

Our identification assumptions are as follows:

- **1. No Unmeasured Confounders:** We assume that all relevant aspects are observed and included in the causal graph. In other words, we assume that there are no hidden confounders that affect both the treatment (e.g., Gender) and the outcome (e.g., suitability for a role). We achieve this by incorporating all key demographic and professional attributes—such as Gender, Race, Education, Volunteering, and Work Experience—in our causal model.

- **2. Consistency:** We assume that the outcome remains consistent across comparable settings. In other words, when an aspect like Gender is changed from 0 (Male) to 1 (Female), the generated counterfactual retains all other context unchanged, and the model's prediction should accurately reflect this change. Consistency is key to ensuring that the predicted outcome corresponds directly to the specific change made. By utilizing GPT-4 to generate textual descriptions that follow the same persona, template, and structure, we maintain consistency across both original and counterfactual versions. This consistency allows us to accurately assess how the targeted aspect affects the prediction, while ruling out the influence of extraneous differences.

- **3. Positivity (Overlap):** implies that there must be a positive probability of observing all possible combinations of the aspect values that we wish to analyze. We ensure positivity by employing a comprehensive Aspect Value Sampling mechanism that generates a diverse range of candidate profiles, covering different combinations of aspects. By maintaining a diverse set of profiles and ensuring that all possible combinations are represented, we satisfy the positivity assumption. This ensures that the model is exposed to a sufficient range of data, allowing us to effectively estimate causal effects across the entire spectrum of potential treatments.

- **4. Counterfactual Faithfulness:** A key assumption in our causal analysis is that the generated counterfactuals faithfully represent only the intended changes in the targeted aspect, without introducing unintended noise or bias. In line with ethical considerations in causal inference, we took measures to validate the quality of

our counterfactuals by developing a dedicated tagging platform. This platform allowed us to verify that the generated counterfactuals preserved the intended changes without introducing unintended noise or variation. More details about the tagging platform and conclusions will be provided later.

Present and Explain the Methods You Have Used

To evaluate the ability of language models to accurately capture causal relationships, we employ the Individual Causal Concept Effect (ICaCE) as our primary method of measurement. The core idea of ICaCE is to measure how altering a specific aspect in the text (e.g., "Gender") affects the model's predictions. To do this, we compare the model's output on the original text to the output on the corresponding counterfactual (CF) version, thereby quantifying the causal impact of the aspect that was changed. This provides insight into how sensitive the model's predictions are to specific features within the input data.

After calculating the ICaCE value, we compare it to the empirical causal effect. This empirical causal effect is calculated by aggregating the differences in the labels for all examples in the dataset where the specific aspect was altered. This average difference represents the "true" causal impact, which serves as a benchmark to assess how closely the model's estimated causal effect aligns with the actual causal effect.
To formalize this concept, we define ICaCE as follows:
Definition (ICaCE, Abraham et al., 2022): Given a Data Generating Process (DGP) g, the individual causal concept effect of changing the concept value C from c to c', for a model F, and a query example $\mathbf{x_u}$, is defined as:

$$ICaCE_F(g, \mathbf{x}_u^{C=c}, c') = E_{\tilde{\mathbf{x}}\sim\mathcal{G}}\left[F(\tilde{\mathbf{x}}_u^{C=c'}) \mid \mathbf{x}_u, \text{do}(C=c', U=u)\right] - F(\mathbf{x}_u^{C=c})$$

If $\tilde{\mathbf{x}}_u^{C=c'}$ is an approximated CF of the text $\mathbf{x}_u^{C=c}$, resulting from an intervention $C : c \to c'$, then the empirical individual causal concept effect of changing the value of concept C from c to c' for state of affairs u is:

$$\hat{ICaCE}_F(\mathbf{x}_u^{C=c}, \tilde{\mathbf{x}}_u^{C=c'}) = F(\tilde{\mathbf{x}}_u^{C=c'}) - F(\mathbf{x}_u^{C=c})$$

This formal definition encapsulates our approach to measuring the causal effect of altering specific aspects within the text, providing a rigorous foundation for evaluating model behavior. By comparing both the ICaCE and the empirical causal effect, we can determine how effectively the language model captures true causal relationships in a controlled environment.

Results

In this section, we evaluate the ability of three different language models— OV, CB, and CW— to capture the causal effect of specific aspects during their initial task learning. All models were trained to perform the task of screening resumes and determining whether a candidate is suitable for the job.

- **OV Model:** This model underwent extensive training with 15 epochs, suggesting that it may have gained a deeper understanding of the dataset and could deliver more accurate predictions. However, there is a concern about potential overfitting.
- **CB Model:** The CB model shares a similar architecture with the OV model but was trained for only 7 epochs. Although it performed reasonably well, it may not have achieved the same level of proficiency as the OV model due to fewer training iterations.
- **CW Model:** The third model, CW, was trained with mismatched hyperparameters, which negatively impacted its ability to effectively learn the given task.

To provide a comprehensive overview of each model's performance, we present the accuracy and confusion matrix for each model. The confusion matrix allows us to evaluate each model's classification capabilities beyond just accuracy, providing insight into how well they distinguish between different classes. By comparing the confusion matrices for OV, CB, and CW, we can visually determine which model seems to have learned the task more effectively overall.

Below are the accuracy scores and confusion matrices for each model:

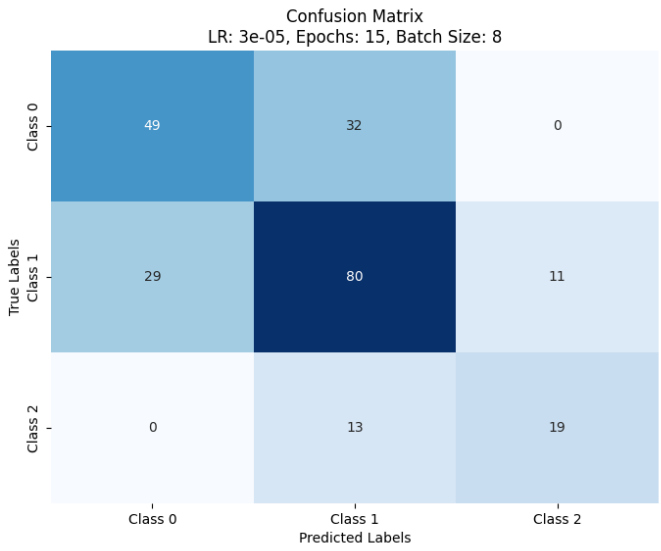


Figure 2: Confusion Matrix for OV Model (15 Epochs). Accuracy: 63.56 %

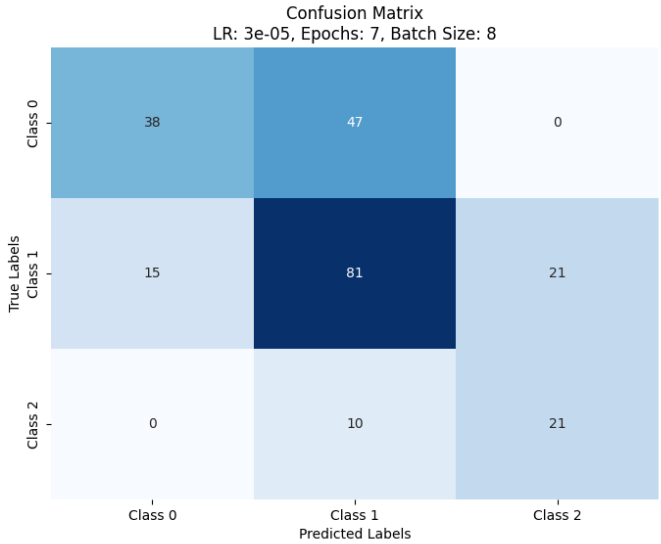


Figure 3: Confusion Matrix for CB Model (7 Epochs). Accuracy: 60.09 %

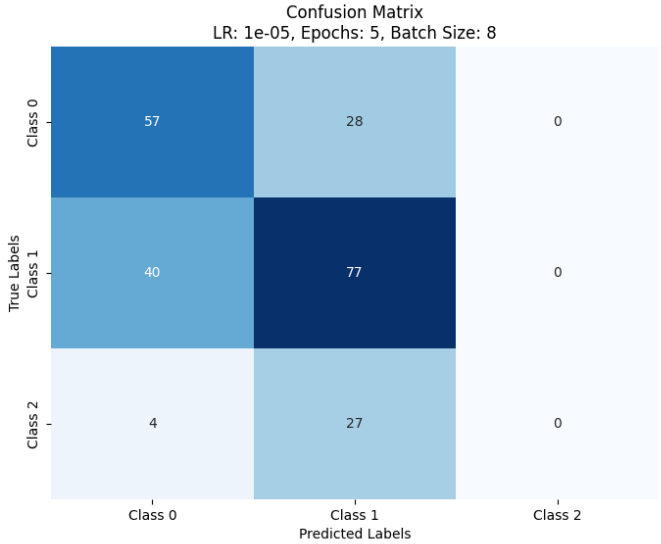


Figure 4: Confusion Matrix for CW Model. Accuracy: 57.51 %

To evaluate the causal reasoning capabilities of the models, we use the Individual Causal Concept Effect (ICaCE) as described in the methods section. We calculated the Average Treatment Effect (ATE) empirically for each aspect change, which serves as the "golden label" against which we compare the models. The ATE represents the mean causal effect of altering a particular aspect, providing us with a benchmark for how a change in that aspect impacts the predicted outcome.

Additionally, to enhance the robustness of our evaluation, we used the Bootstrap method to estimate confidence intervals for each ATE. This helps in quantifying the uncertainty around our ATE estimates and provides a clearer picture of how stable the

estimated effects are.

The table below shows a summary of our findings, including the ATE, the bootstrap confidence intervals, and the predicted effect values for each model. For each aspect change, we also analyzed which model's predicted effect score was closest to the empirical ATE, thus indicating which model was best at capturing the causal influence of that aspect.

Table 1: Performance Comparison Across Models for Different Aspect Changes

Aspect Change	**ATE**	**CI L.B**	**CI U.B**	**OV Model**	**CB Model**	**CW Model**
Age Group (0.0 → 1.0)	0.125	0	0.375	0.125	-0.125	0.125
Age Group (1.0 → 0.0)	-0.2	-0.4	0	-0.13	-0.33	-0.26
Gender (1.0 → 0.0)	0	0	0	0	-0.06	-0.03
Race (0.0 → 3.0)	0.5	0	1	1	0.5	0.5
Certificates (0.0 → 1.0)	0.33	0.66	0.65	0.72	0.57	0.51
...	...	...	...	...	...	...

The table above presents the ATE (empirical causal effect) alongside the predicted effect values of the OV, CB, and CW models for different aspect changes, as well as the bootstrap confidence intervals for each ATE. We observed that:

- **OV Model** had the highest number of wins with **24** times being closest to the empirical effect. The model's predictions often fell within the bootstrap confidence intervals of the empirical ATE, which suggests that it captured the general causal influence of the aspects reasonably well.
- **CB Model** also performed well but lagged slightly behind OV, particularly in terms of capturing subtle differences in smaller effect sizes. followed with **14** wins. It captured the direction and general magnitude of the effect but showed inconsistencies for some aspect changes.
- **CW Model** as expected, showed the largest deviations from the empirical ATE, with frequent cases where its predicted effect values were well outside the bootstrap confidence intervals. had only **9** wins, further indicating its limited ability to understand causal relationships accurately.

For several aspect changes, the narrow bootstrap confidence intervals indicate a high level of certainty in the estimated ATE values. The evaluation of the models effectively demonstrates the viability of our proposed method to gauge a model's understanding of causal effects. It suggests that successful task performance by a model inherently involves learning the underlying causal relationships among features, with proper training playing a critical role in accurately capturing these causal interactions.

Incorporating bootstrap confidence intervals has further strengthened the reliability of our ATE estimates. These intervals provide valuable insight into the uncertainty associated with each estimate, offering a clearer understanding of the potential range of the true causal effect and enhancing the robustness of our conclusions.

Possible Weaknesses

One of the main challenges lies in ensuring that the transition from tabular aspect values to textual representations using GPT correctly captures the intended changes. It is crucial to verify that the texts and their counterfactuals (CFs) vary only in the targeted aspects and that these generated texts are driven by the specified aspect values, not by noise or unintended model biases. To address this challenge, we manually annotated a sample of the dataset to assess whether the text accurately reflects the intended aspects and if the CFs remain consistent with the original data.

We focused on evaluating two key components:

- Quality of Counterfactuals (CFs)
- Quality of Text Representations

We performed the annotation using the Label Studio platform, and the tables containing the annotated data can be found in the Git page.

The average score for CF quality was 3.60 on a scale of 1 to 5, suggesting that, while the generated data is generally reliable for analysis, there are still areas that require improvement. Text quality was evaluated based on consistency, fluency, relevance, and coherence, all scored on a scale from 1 to 5. The average scores were as follows: consistency 4.47, fluency 4.83, relevance 4.29, and coherence 4.85.

Additionally, several visualizations are provided in the appendix that highlight the quality scores and aspect changes across the dataset, offering further insights into areas where data generation was more reliable or exhibited inconsistencies.

Discussion

Ultimately, the goal of this project is not only to evaluate the causal learning capabilities of large language models (LLMs) but also to contribute to the broader field of fair and interpretable AI. By explicitly modeling causal relationships and evaluating these models using robust causal inference techniques, we hope to set a benchmark for more accountable and transparent use of NLP models, especially in high-stakes decision-making scenarios.

This project forms part of a larger research effort focused on refining the integration of causal inference and natural language processing. The opportunity to work with our

own dataset provided valuable insights into its versatility and highlighted numerous potential applications beyond our initial objectives. This hands-on experience deepened our understanding of the data and offered an enriched perspective on the causal effect analysis inherent within it.

Repository Link

You can access the project repository here: GitHub Repository for Causal Inference Project

References

Abraham, E. D., D'Oosterlink, K., Feder, A., Gat, Y., Geiger, A., Potts, C., & Reichart, R. (2022). Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *NeurIPS*. Retrieved from https://arxiv.org/abs/2205.14140

Amir Feder, E. M. R. P. D. S. Z. W.-D. J. E. J. G. R. R. M. E. R. B. M. S. V. V., Katherine A. Keith, & Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics (TACL)*. Retrieved from https://arxiv.org/abs/2109.00725

Goyal, Y., Feder, A., Shalit, U., & Kim, B. (2020). Explaining classifiers with causal concept effect (cace). *arXiv*. Retrieved from https://arxiv.org/abs/1907.07165

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *ScienceDirect*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1041608023000195?via%3Dihub

Nazi, Z. A., & Peng, W. (2024). Large language models in healthcare and medical domain: A review. *arXiv*. Retrieved from https://arxiv.org/abs/2401.06775

Nie, Y., Kong, Y., Dong, X., M.Mulvey, J., Poor, H., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv*. Retrieved from https://arxiv.org/abs/2406.11903

Raymond Zhang, D. S. D. M. A. M. K. K., Neha Nayak Kennard. (2023). Causal matching with text embeddings: A case study in estimating the causal effects of peer review policies. *Association for Computational Linguistics (ACL)*. Retrieved from https://aclanthology.org/2023.findings-acl.83/

Sevastjanova, R., Amara, K., & El-Assady, M. (2024). Challenges and opportunities in text generation explainability. *Communications in Computer and Information Science, 2153*.

Shui, R., Cao, Y., Wang, X., & Chua, T.-S. (2023). A comprehensive evaluation of large language models on legal judgment prediction. *ACL*. Retrieved from https://aclanthology.org/2023.findings-emnlp.490/

Yair Gat, A. F. A. C. A. S. R. R., Nitay Calderon. (2023). Faithful explanations of black-box nlp models using llm-generated counterfactuals. *In The Twelfth International Conference on Learning Representations*. Retrieved from https://arxiv.org/abs/2310.00603

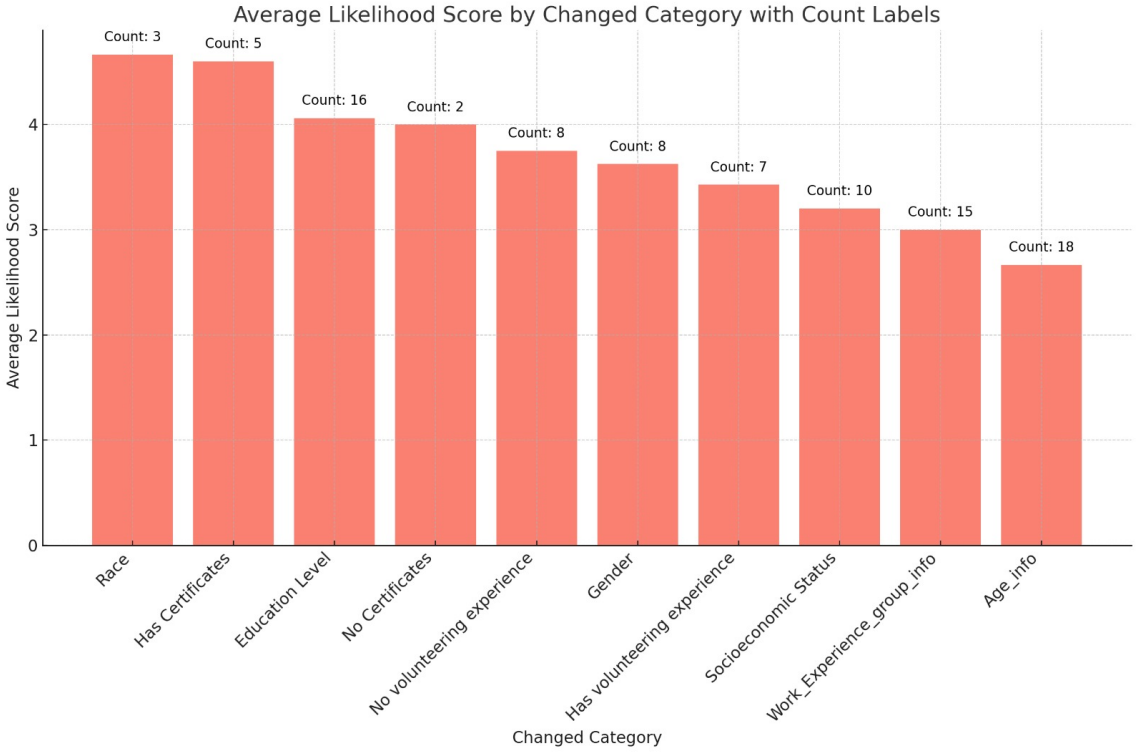
Zach Wood-Doughty, M. D., Ilya Shpitser. (2018). Challenges of using text classifiers for causal inference. *Association for Computational Linguistics (ACL)*. Retrieved from https://aclanthology.org/D18-1488/

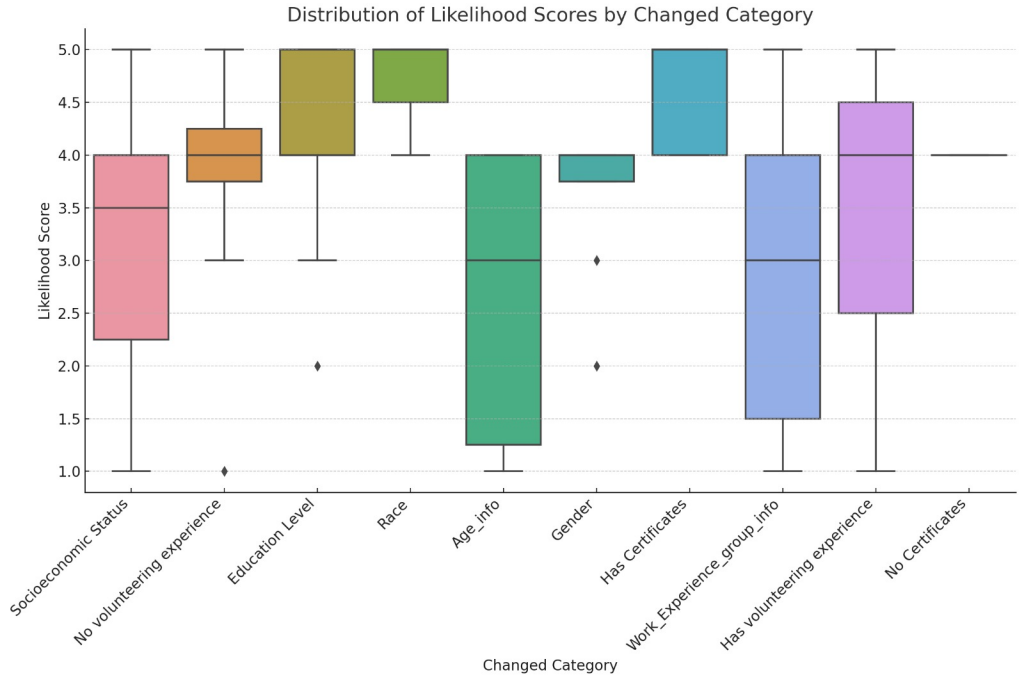
Zhengxuan Wu, T. I. C. P. N. G., Atticus Geiger. (2023). Interpretability at scale: Identifying causal mechanisms in alpaca. *NeurIPS*. Retrieved from https://papers.nips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html

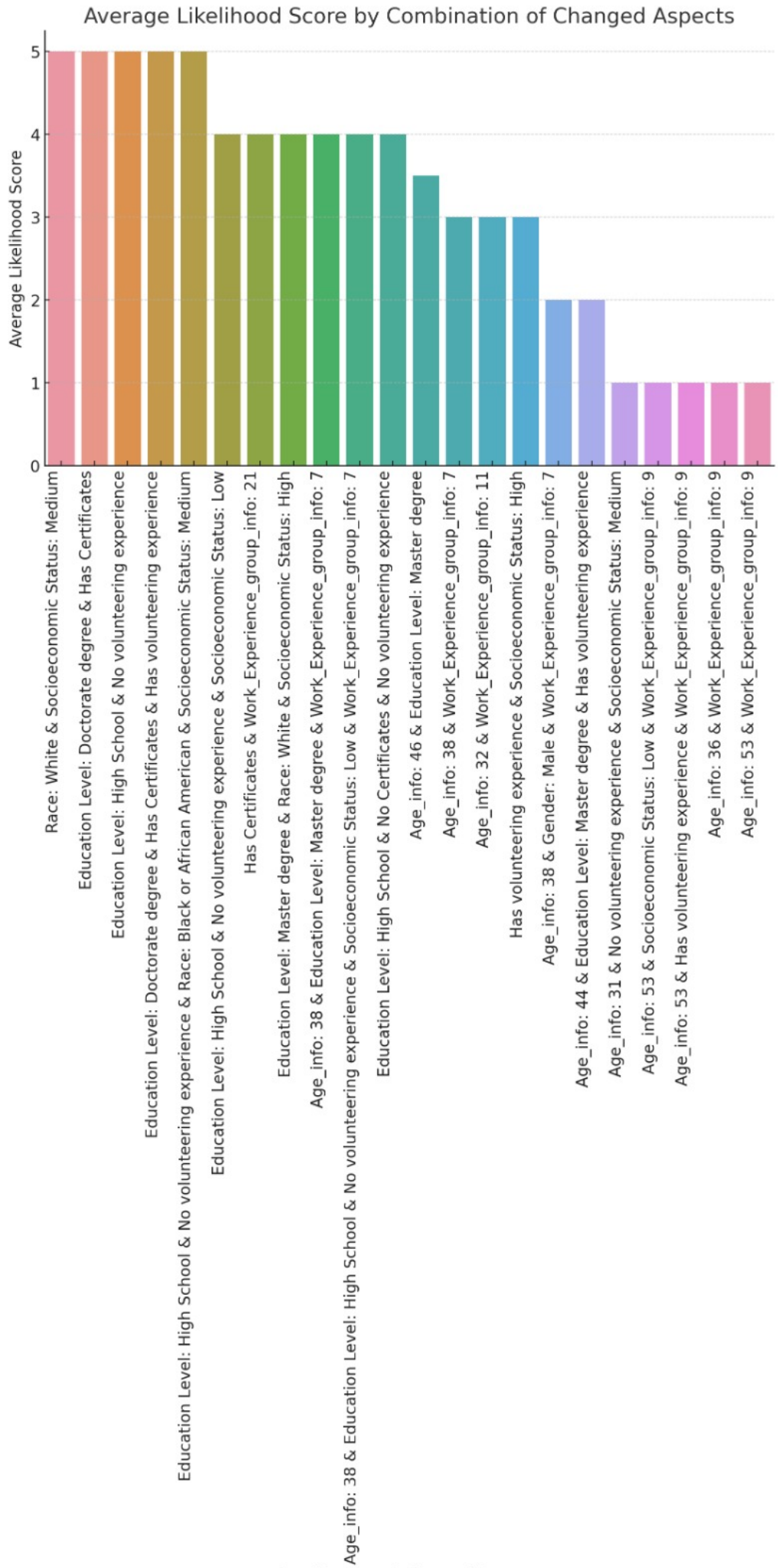
Appendix

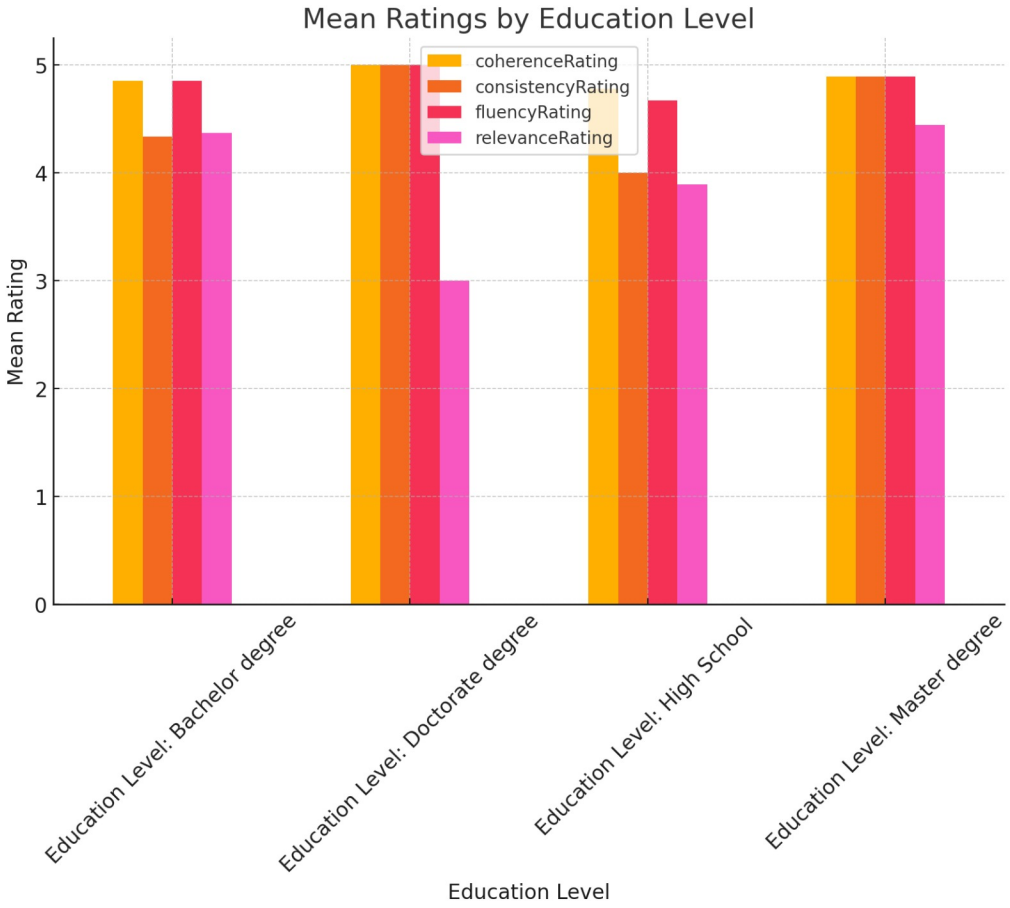
Aspect Change Visualizations

Several visualizations were produced to further illustrate the differences in model outputs due to aspect changes. These visualizations highlight how the changes influenced the text generation process, giving insights into the underlying causal relationships.






Average Likelihood Score by Combination of Changed Aspects
Average Likelihood Score
0
1
2
3
4
5
Race: White & Socioeconomic Status: Medium
Education Level: Doctorate degree & Has Certificates
Education Level: High School & No volunteering experience
Education Level: Doctorate degree & Has Certificates & Has volunteering experience
Education Level: High School & No volunteering experience & Race: Black or African American & Socioeconomic Status: Medium
Education Level: High School & No volunteering experience & Socioeconomic Status: Low
Has Certificates & Work_Experience_group_info: 21
Education Level: Master degree & Race: White & Socioeconomic Status: High
Age_info: 38 & Education Level: Master degree & Work_Experience_group_info: 7
Age_info: 38 & Education Level: High School & No volunteering experience & Socioeconomic Status: Low & Work_Experience_group_info: 7
Education Level: High School & No Certificates & No volunteering experience
Age_info: 46 & Education Level: Master degree
Age_info: 38 & Work_Experience_group_info: 7
Age_info: 32 & Work_Experience_group_info: 11
Has volunteering experience & Socioeconomic Status: High
Age_info: 38 & Gender: Male & Work_Experience_group_info: 7
Age_info: 44 & Education Level: Master degree & Has volunteering experience
Age_info: 31 & No volunteering experience & Socioeconomic Status: Medium
Age_info: 53 & Socioeconomic Status: Low & Work_Experience_group_info: 9
Age_info: 53 & Has volunteering experience & Work_Experience_group_info: 9
Age_info: 36 & Work_Experience_group_info: 9
Age_info: 53 & Work_Experience_group_info: 9


Mean Ratings by Education Level
coherenceRating
consistencyRating
fluencyRating
relevanceRating
Mean Rating
0
1
2
3
4
5
Education Level: Bachelor degree
Education Level: Doctorate degree
Education Level: High School
Education Level: Master degree
Education Level