Can Base ChatGPT be Used for Forecasting without Additional Optimization?

Pham Hoang Van,[1*,ⓡ] Scott Cunningham,[1,ⓡ]
[1]Department of Economics, Baylor University
Waco, TX, USA
Van_Pham@baylor.edu, Scott_Cunningham@baylor.edu

July 2024

Abstract

This study investigates whether OpenAI's ChatGPT-3.5 and ChatGPT-4 can accurately forecast future events using two distinct prompting strategies. To evaluate the accuracy of the predictions, we take advantage of the fact that the training data at the time of our experiments (April and May 2023) stopped at September 2021, and ask about events that happened in 2022. We employed two prompting strategies: direct prediction and what we call future narratives which ask ChatGPT to tell fictional stories set in the future with characters that share events that have happened to them, but after ChatGPT's training data had been collected. Concentrating on events in 2022, we prompted ChatGPT to engage in storytelling, particularly within economic contexts. After analyzing 100 prompts, we discovered that future narrative prompts significantly enhanced ChatGPT-4's forecasting accuracy. This was especially evident in its predictions of major Academy Award winners as well as economic trends, the latter inferred from scenarios where the model impersonated public figures like the Federal Reserve Chair, Jerome Powell. As a falsification exercise, we repeated our experiments in May 2024 at which time the models included more recent training data. ChatGPT-4's accuracy significantly improved when the training window included the events being prompted for, achieving 100% accuracy in many instances. The poorer accuracy for events outside of the training window suggests that in the 2023 prediction experiments, ChatGPT-4 was forming predictions based solely on its training data. Narrative prompting also consistently outperformed direct prompting. These findings indicate that narrative prompts leverage the models' capacity for hallucinatory narrative construction, facilitating more effective data synthesis and extrapolation than straightforward predictions. Our research reveals new aspects of LLMs' predictive capabilities and suggests potential future applications in analytical contexts.

1 Introduction

Rapid technological advancements in artificial intelligence have exceeded our understanding of its use cases. Large language models (LLMs) such as OpenAI's GPT-4 can mimic intelligent human speech as well as perform cognitively costly tasks which alter workers' marginal products, but it is unclear the reach of those tasks. In principle, given these models are predictive machines, they may provide humans with a new forecasting device (Agrawal et al., 2018). But how accurate they are is unknown in part because these new technologies seem poorly understood even by its creators.

The backbone of the current forefront of LLMs is an architecture called generative pre-trained transformers, or GPT. This architecture revolutionized natural language processing (NLP) by capturing intricate text relationships through self-attention mechanisms (Vaswani et al., 2017). OpenAI's introduction of GPT-3.5 in November 2022 and its successor, GPT-4, in March 2023 marked significant milestones in the evolution of the GPT. With their vast neural networks pre-trained on diverse textual corpora, these models possess an unparalleled ability to understand and generate language, though their application in forecasting, particularly future events, remains underexplored due to the inherent limitations of their training data.

One of the things that makes LLMs unique is that a large amount of the inputs are in the models' prior training datasets. These training datasets contain billions of unknown texts thought to encompass an expansive amount of material available online (Hughes, 2023). OpenAI conceals precisely which datasets it had been trained on (Schaul et al., 2023), but given the models' purported sizes, as well as its successful ability to achieve conversational speech, it is thought that the training datasets include a large swath of online material.

This study uniquely positions itself at the intersection of LLMs' generative capabilities and their potential for predictive analysis. By employing GPT-3.5 and GPT-4, we probe whether different prompting strategies can force ChatGPT to more accurately predict future events. To test our forecasting success, we took advantage of the natural boundary set by OpenAI. At the time of our mid-2023 experiment, OpenAI's last training update had been September 2021 (OpenAI, 2024a).[1] Given that ChatGPT's training data at the time did not contain information about the events of 2022, we were able to explore whether it could exploit patterns in its training data, which stopped in September 2021, to accurately forecast events with social and economic value, such as the winners of the 2022 Academy Awards, monthly unemployment rates and monthly inflation rates through September 2022.

One of the hallmarks of LLMs, though, is that they are highly creative. This creativity is both a feature and a bug. It seems that it's creativity is part of what makes it successful at mimicking intelligent human speech. But it also seems that this creativity is what makes it regularly hallucinate—a term which describes its tendency to strongly assert false events or incorrect facts (Levy, 2024). Its creativity and tendency to hallucinate may be a barrier to prediction if it is systematically skewed in some way that is worse than our current technologies. While outright prediction does not directly violate OpenAI's terms of service, we think it is most likely the case based on our experiment that OpenAI has attempted to make it very difficult. This may be because three of its terms of service violations would seem to be directly violated if people used ChatGPT intensively for predictive purposes. Those three violations fall under OpenAI's rule that the software not be used to "perform or facilitate activities that may significantly impair the safety, well being, or rights of others" (OpenAI, 2024b) which

[1] We can show our data collection occurred mid-2023 with time stamps in Excel spreadsheets used by our research assistants.

then lists three cases relevant to prediction.

a. Providing tailored legal, medical/health, or financial advice without review by a qualified professional and disclosure of the use of AI assistance and its potential limitations

b. Making high-stakes automated decisions in domains that affect an individual's safety, rights or well-being (e.g., law enforcement, migration, management of critical infrastructure, safety components of products, essential services, credit, employment, housing, education, social scoring, or insurance)

c. Facilitating real money gambling or payday lending

If ChatGPT were found to have superior forecasting ability, then one could easily imagine it would be immediately used in ways that violated one or all of the above conditions, and thus we suspect OpenAI has throttled ChatGPT's willingness to engage in many types of outright prediction tasks.

But one thing that does not violate its terms of service is the telling of stories. While it may violate OpenAI's terms of service, for instance, to seek ”tailored medical advice”, and thus ChatGPT may refuse to do it, that may not stop it from creating a work of fiction in which such predictions were conducted in the context of the narrative itself. Our project tests for whether requesting ChatGPT to tell stories may, in fact, unlock its ability to perform accurate forecasting. By using the models' training data cutoff, and knowing what did and did not occur afterwards (”ground truth”), we are able to compare the performance of prompts that directly ask ChatGPT to predict the future versus ones that ask it to tell stories about the future.

Our narrative prompts are unique in that we ask ChatGPT to tell a story about events set in the future as they occur or by authoritative figures set in the future telling stories about their past (but our future). Our narrative prompts experimented with variation in seemingly small details, such as the identity of the speaker or releasing information about 2022 political events, to investigate further which elements of the narrative prompts mattered. To create a distribution of answers, we had two research assistants use two separate ChatGPT accounts to query 50 times per prompt creating 100 total trials per prompt. We present our findings as box plots showing the full distribution of answers to each prompt.

Our findings suggest that these prediction machines become unusually accurate under ChatGPT-4 when prompted to tell stories set in the future about the past. First we show the accuracy of direct and narrative prompting using ChatGPT-3.5 and ChatGPT-4 to predict the winners of the major categories of the 2022 Academy Awards. For Best Actor, Best Actress, and both Supporting actor categories, narrative prompting was extremely accurate at predicting the winner. Direct prompting performed by comparison very poorly, oftentimes worse than random guesses. But narrative prompting with ChatGPT-4 shows accuracy ranging from 42% (Best Actress, Chastain) to 100% (Best Actor, Will Smith) with one exception. It failed to accurately predict the Best Picture winner.

Next we moved on to the economic phenomena of monthly unemployment rates and monthly inflation rates using three separate kinds of narrative prompts: a college professor giving a lecture to undergraduate students about the Philips Curve, and Federal Reserve chair, Jerome Powell, giving a speech to the Board of Governors about the last year's economic data. In the case of Jerome Powell, we varied an additional detail: in one prompt we first told ChatGPT about Russia's 2022 invasion of Ukraine before then asking it to have Jerome Powell give his speech to the Board of Governors recounting the last year's macro data. And in the other, we left out that piece of information. In all cases, direct prompting was even less effective at prediction than it had been with the Academy Awards as ChatGPT refused to answer the prompt altogether when asked to directly predict the future time series of each macroeconomic variable. The anonymous economist rarely was successful at predicting inflation using either LLM.

But when asked to tell a story in which Jerome Powell recounts a year's worth of future unemployment and inflation data, as though he was talking about the events of the past, things change substantially. The distribution of Powell's month by month predictions of inflation are on average comparable to the facts contained in the monthly University of Michigan's consumer expectations survey. Interestingly, it is closer to accurately predicting the UM consumer expectations survey than it is at predicting actual inflation based on data collected by the Cleveland Fed. Oddly, when prompted with information about Russia's invasion of Ukraine, Powell's predictions were systematically lower and less accurate than when that information had not been used to prime ChatGPT.

The accuracy of the anonymous economist at predicting monthly unemployment was both precise and most of the time inaccurate. But as with inflation, the unemployment rates publsihed by the Bureau of Labor Statistics (BLS), month by month, were within the distribution of claims made by Jerome Powell in his speech. And the inclusion of Ukraine data caused the model to get worse at predicting unemployment as had been the case with inflation.

A number of recent studies investigates the uses of generative AI in economics and business applications. Examples include using ChatGPT as economic agents in lab experiments (Horton, 2023), surveying ChatGPT for marketing research (Brand et al., 2023), asking ChatGPT to make risk assessments from earnings call transcripts (Kim et al., 2023) and the application of LLMs to predict stock price movements (Lopez-Lira and Tang, 2023). Our findings add to this nascent exploration by underscoring the importance of prompt design in harnessing LLMs for predictive tasks, suggesting that the models' capacity for "hallucinatory" narrative construction can be strategically utilized to extract forward-looking insights from their training data. But it also suggests that beneath OpenAI's outward facing consumer product, ChatGPT-4, is a very powerful prediction machine. This revelation opens new avenues for the application of LLMs in economic forecasting, policy planning, and beyond, challenging us to rethink how we interact with and exploit the capabilities of these sophisticated models.

The next section presents an example of the direct versus narrative prompting. This is followed by a detailed description of our methodology and data collection. We present results of experiments predicting the Oscars winners and inflation and unemployment followed by our conjecture on ChatGPT-4's Predictive Abilities in Narrative Form. We summarize our findings and suggest avenues for future research in the conclusion.

2 Direct vs Narrative Prediction

The challenge of evaluating GPT-4's predictive power is several-fold. First, OpenAI has placed restraints on the software such that it refuses to provide some information. For instance it is a violation of its terms of service to use ChatGPT to give tailored medical advice. To test this, we asked ChatGPT-4 the following prompt.

> "I am feeling really sick. I have a headache and my urine has blood in it. What do you think I have?"

ChatGPT-4 responded in the following way refusing to answer:

> “I’m really sorry to hear you’re feeling this way, but I cannot diagnose or provide medical advice. It’s important to see a healthcare professional as soon as possible, especially if you’re seeing blood in your urine. This could be indicative of a number of conditions, some of which require prompt medical attention. Please contact a doctor or visit an emergency room right away.”

But then we switched from a direct prompt to a narrative prompt and asked ChatGPT to tell a story based on the following:

> Imagine that this is a scene to a Tennessee Williams play, not the entire story. Write the entire scene, including summaries of what I say below to fit the drama. Please have the doctor provide realistic assessments of your own opinion about what he learned from the man’s symptoms.
>
> A man comes into the doctor’s office complaining of a headache and nausea. He says, somewhat embarrassed, “Doc it’s not just the headache though.” The doctor says “What is it?” And the man says, “this morning there was also blood in my urine.” A nurse takes his temperature and the doctor, and he and the nurse review his chart. The doctor comes back and shares the news.

The same symptoms were presented to ChatGPT-4 in both the direct and narrative prompt, but note that in the narrative prompt, we did not actually ask ChatGPT to make a prediction. Rather, we asked it to simply tell a story in which a doctor gives medical advice to someone with the same symptoms as we had given it in the direct prompt. For some reason this makes all the difference in its willingness to respond to the prompt and attempt a diagnosis. Listen now to the story unfold.

Scene: A Doctor's Office, Late Afternoon

Characters:

- ***Doctor Williams****: A middle-aged, experienced physician*
- ***Nurse Johnson****: A young, diligent nurse*
- ***John****: A man in his 40s, visibly distressed*

(The curtain rises to reveal a modest doctor's office. There's a desk cluttered with medical journals and a computer. Two chairs are in front of the desk, and a medical examination table is off to the side. Nurse Johnson is seated at a smaller desk, sorting through paperwork. Doctor Williams is reading a medical journal. John enters, looking uncomfortable and holding his abdomen.)

John*: (nervously) Uh, Doctor Williams? I've got this headache... and nausea.*

Doctor Williams*: (putting down the journal) Please, have a seat. Tell me more.*

John*: (sitting down, embarrassed) Doc, it's not just the headache though.*

Doctor Williams*: What is it?*

John*: This morning... there was also blood in my urine.*

(Nurse Johnson looks up, concerned, and starts preparing to take John's vitals.)

Nurse Johnson*: (to John) I'll need to take your temperature.*

(Nurse Johnson takes John's temperature and notes it down. Doctor Williams and Nurse Johnson briefly review John's chart, whispering to each other.)

Doctor Williams*: (returning to John) Your temperature is normal, but the symptoms you've described—headache, nausea, and hematuria—could indicate a number of issues. It could be a urinary tract infection, kidney stones, or something more* *serious like a kidney infection or even cancer. We'll need to run some tests to be sure.*

***John**: (visibly worried) Cancer? Doc, what are my chances?*

***Doctor Williams**: Let's not jump to conclusions. First, we'll do some blood tests and a urine analysis. Depending on those results, we may need to do further imaging tests.*

***John**: (nods) Alright, Doc. Whatever it takes.*

(Doctor Williams pats John reassuringly on the shoulder.)

***Doctor Williams**: We'll get to the bottom of this, John. You're not alone.*

(Curtain falls.)

This short example is fascinating, amusing and illustrative of ways that ChatGPT can be forced to undertake tasks that otherwise it is not allowed to do. While ChatGPT-4 is unlikely to ever provide reliable medical advice, if it will even do so at all, it will tell a story in which a medical expert interacts with a patient exhibiting real life symptoms and provides expert advice. Whether this expert advice is accurate is another matter; our point is merely to note that it will not undertake the task when asked directly to do it, but it will when given the task indirectly in the form of creative writing exercises.

3 Prompting Methodology and Data Collection

The aforementioned vignette was arguably nothing more than a creative writing exercise. There was no "ground truth" to uncover. Neither of the two authors on this paper had the medical issues described above. The symptoms were specific but simply made up to illustrate the fact that ChatGPT will refuse to do some tasks that would create a terms of use violation, except when prompted to create fiction.

In order to move forward, we needed a use case in which there was ground truth to compare with ChatGPT's forecasts. We did this by taking advantage of a limitation baked into ChatGPT at the time of our experiment. When we undertook this experiment in mid 2023, OpenAI's ChatGPT training data did not include any information past September 2021 (OpenAI, 2024a). But, whereas ChatGPT did not know about the events of 2022, as it was not in its training data, the authors did. We used this training data cutoff to therefore predict various events in 2022 through 100 repeated prompts across two research assistants using two separate ChatGPT-4 accounts, 50 times apiece, with both direct prompting and narrative prompting for both ChatGPT-3.5 and ChatGPT-4.

OpenAI's GPT-3.5 and GPT-4 models were trained on large amounts of text (with varying amounts and parameter estimates which we will explain later) up to September 2021. Given that 75% of 2021 was covered by this data, we reasoned that it may have been trained on data which could enable it to reasonably predict the near to long term events of 2022. As we said, there were two types of prompts (direct versus future narratives) across two types of LLMs (ChatGPT-3.5 and ChatGPT-4) creating four queries per prediction concept.

To obtain a distribution of answers for each prompting style, we enlisted the help of two research assistants (RAs) who independently queried GPT multiple times using the respective prompts. This approach allowed us to account for the inherent randomness in the model's responses and provided us with a distribution which we plot visually using simple boxplots. To ensure independence across the queries, the RAs used different GPT-4 accounts from two separate subscriptions. This precaution was done to prevent potential hallucinations or biases in one RA's prompting to cascade to the other RA, thereby maintaining the integrity of the experiment.

The RAs sampled each question 50 times, generating a total of 200 responses for each event under investigation (100 from GPT-3.5 and 100 from GPT-4). We then compared the distribution of answers generated by each prompting method to the actual outcomes, allowing us to assess the predictive accuracy of GPT-3.5 and GPT-4 under different prompting conditions.

The difference between direct prediction and future narrative prediction is that direct prediction simply asks ChatGPT to predict a future event. Consider the following direct prediction prompt in which we asked ChatGPT to predict future macroeconomic events:

> Please predict the inflation rate and unemployment rate for each month starting in September 2021 and ending in June 2022. You should use all available data available, including any published or informal forecasts, of both economic variables when making this prediction.

Note that this type of prompt asks ChatGPT to start from the end of its training data at September 2021 and predict October 2021 to June 2022. Even though those events have happened, ChatGPT had not yet been trained on those data and so it should not be able to pull it out of the training data as facts. Anything it says in response to that direct request would have to come from prediction.

Future narrative prompting is not merely asking ChatGPT to predict the future using a story, though. Rather, future narrative prompting tells ChatGPT to tell a story about a scene set in the future in which an authoritative person who in principle would possess accurate information simply states the past events as though they had already happened. We chose this form of narrative prompting because we wanted to force ChatGPT into a position of committing to a prediction, as opposed to opening it up to some kind of unknown process of prediction that we did not understand. Consider this example using one of our prompts in which the Chairman of the Federal Reserve, Jerome Powell, gives a speech.

> Write a scene where Federal Reserve chairman Jerome Powell gives a speech in October 2022 about inflation, unemployment, and monetary policy. Chairman Powell tells the audience the inflation rate and unemployment rate for each month starting in September 2021 and ending in August 2022. Have chairman say each month one by one. He concludes with an outlook for inflation and unemployment and possible interest rate policy changes.

Our prompt attempted to normalize the prediction in the context of the narrative as much as possible, as though the prediction task itself was not the purpose of the task we were asking ChatGPT to undertake. In the narrative prediction prompt, in other words, the task we ask ChatGPT to undertake is the crafting of an interesting story. The prediction task is secondary. Why this matters is not clear, but we think that making the prediction task subservient to the primary task of creative storytelling does indeed make a difference in the accuracy of ChatGPT's forecasting.

4 Results

4.1 Establishing the Training Data Limit with Falsifications

Before we discuss our main results, we review some basic tests performed that we thought could clarify whether, in fact, ChatGPT could access online information after

September 2021. Our hunch is that if there is *any* relevant information that could facilitate prediction in the pre-September 2021 training data, then it does not constitute a good candidate for a falsification. So we chose four things that could not be predicted as there was no information about these events in the training data: the names of the four teams that made the March Madness NCAA basketball tournament Final Four, the winner of the NCAA Championship, the winning lottery ticket on several dates, and the highest grossing films in January 2022 through April 2022. Across all four types of queries, whether using the direct prompt or the future narrative prompt, whether using ChatGPT-3.5 or ChatGPT-4, the result was failure. Even though these facts were readily available online at the time of our testing, ChatGPT was unable to answer any of the tasks correctly.

One possible objection may be that ChatGPT-4 could access the Internet, nonetheless, through Bing plug-ins or Bing integration. OpenAI made available Bing integration to Plus subscribers around May 12th 2023 (Mehdi, 2023) which was shortly after our Oscars prediction but contemporaneous to our Philips curve predictions. These features were only accessible via a beta panel in a subscriber's settings. Neither RA utilized this feature in their prompting. Including ChatGPT-3.5 as a control helps establish the veracity of our claim that the prediction is happening outside of accessing online data. We provide timestamps from our Excel files showing the exact dates of our data collection. The Oscars predictions were generated over the period April 1-May 4, 2023. The Philips curve predictions were generated May 11-June 9. Bing did not become fully integrated into ChatGPT-4 for all Plus subscribers, outside of the beta plug-in, until around September 27, 2023.[2]

We have made every effort to document the dates of our experiment, monitor RA

[2] For more information on the history of ChatGPT, see ChatGPT's release notes at `https://help.openai.com/en/articles/6825453-chatgpt-release-notes`.

data collection, and conduct clean controls with ChatGPT-3.5 in our data collection amidst OpenAI's rapid and largely unanticipated updates to its software, but in our case, we managed in all cases to complete the data collection without contamination from these updates as best we can tell. We now proceed to focus on the results of our incrementally more difficult prediction tasks.

4.2 Results of the 2022 Academy Awards Forecasts

The 2022 Academy Awards was held on March 27, 2022, a full six months after the September 2021 training data had stopped. This means that ChatGPT-3.5 and ChatGPT-4 were likely trained on news reports about the movies, but not trained on the Oscar announcements. And while they were not trained on the last three months of 2021, they were trained on the first nine months. Thus it is safe to say that the LLM "knows" a lot about these movies without knowing anything about the revelation of the Oscar awards ceremony itself.

There are several predictors of who will win awards in the Academy Awards, as many movies, actors and directors will sometimes sweep the Golden Globes (January), BAFTA (February), Screen Actors' Guild (SAG, January or February) and several more awards (Silver, 2013). But all major award ceremonies that lead into the 2022 Academy Awards occurred in early 2022 after September 2021. Thus even these early anticipatory awards, historically highly predictive of the Oscars outcomes, were not in the training data. This is an ideal situation for our experiment because we have a lot of information about the movies themselves, but not the traditional predictors like the SAG awards, nor the news itself. Since the training data stopped in September 2021, it is missing award information, which means votes have not yet been cast for any movies, and therefore there is only speculation. And as the Bing plug-in did not become available for Plus users until after our experiment, the experiment is uncontaminated.

The first category we report is the Best Supporting Actor category. The Best Supporting Actor award is an annual prize presented by the Academy of Motion Picture Arts and Sciences. Since the 9th Academy Awards, it honors an actor who delivered an outstanding performance in a supporting role within a film released in the previous year. The promptings we used are listed in the following table with the direct prompt on the left, and future narrative prompt on the right.

Best Supporting Actor of 2022 Prompts

Prompt 1a (Direct)	**Prompt 1b (Future Narrative)**
Of the nominees listed below, which nominee do you think is most likely to win the Best Supporting Actor award at the 2022 Oscars? Please consider the buzz around the nominees and any patterns from previous years when making your prediction. Troy Kotsur , Ciarán Hinds, Jesse Plemons, J. K. Simmons, Kodi Smit-McPhee	Write a scene in which a family is watching the 2022 academy awards. The presenter reads off the following nominees for best supporting actor: Troy Kotsur , Ciarán Hinds, Jesse Plemons, J. K. Simmons, Kodi Smit-McPhee. Describe the scene culminating in the presenter announcing the winner.

The actual 2022 winner for Best Supporting Actor was Troy Kotsur. Results from 100 trials using ChatGPT-3.5 are shown in Figure 1. The actual winner, Kotsur, is shown in the farthest right position along the x-axis. When using our direct prompt, Kotsur was selected as the Best Supporting Actor 1% of the time (i.e., once out of 100 trials). The most common outcome was "NP" (No Prediction) which was a refusal by ChatGPT-3.5 to provide any answer. There was a tie between Simmons and "multiple picks" for most frequent response at 21%.[3] When we used the future narrative prompt (right panel), the overwhelming winner from this exercise was J. K. Simmons who was

[3] In many instances, ChatGPT would list several answers which we call "Mult" short for multiple picks.

picked 83% of the time. Interestingly Kotsur won only twice out of 100 trials. Thus, while future narrative prompting did improve the accuracy of the prediction, the improvement was only marginal.

In Figure 2, we report the results of the exercise using ChatGPT-4 for both types of prompts. When using direct prompting, the results were bimodal. ChatGPT-4 answered with Mult and NP 34% of the time. The correct choice, Kotsur, was made 25% of the time. But when we switched to future narrative prompting, shockingly, ChatGPT-4 guessed Troy Kotsur correctly in all trials.

Best Actor of 2022 Prompts

Prompt 3a (Direct)

Of the nominees listed below, which nominee do you think is most likely to win the Best Actor award at the 2022 Oscars? Please consider the buzz around the nominees and any patterns from previous years when making your prediction.
Javier Bardem, Benedict Cumberbatch, Andrew Garfield, Will Smith, Denzel Washington

Prompt 3b (Future Narrative)

Write a scene in which a family is watching the 2022 academy awards. The presenter reads off the following nominees for Best Actor: Javier Bardem, Benedict Cumberbatch, Andrew Garfield, Will Smith, Denzel Washington. Describe the scene culminating in the presenter announcing the winner.

Following the previous discussion of the Best Supporting Actor category, we introduce our results from the Best Actor award. The Best Actor award similarly celebrates outstanding performances in film, but with a focus on lead roles as opposed to supporting ones. Since its inception at the 1st Academy Awards, the Best Actor category has spotlighted one of the central figures who carry a film's story. We discuss the results from our direct and future narrative prompts in Figures 3 (ChatGPT-3.5) and 4 (ChatGPT-4). As with Best Supporting Actor, we show the winner (Will Smith) at the farthest right of the x-axis in each panel.

Most of the time, ChatGPT-3.5 made the wrong prediction. In 55% of the guesses, it provided multiple answers and in 28% it gave no pick. But if it did pick, it picked Will Smith 17% of the time. When we then put ChatGPT-3.5 into a future narrative of the family watching the award ceremony, it guessed Will Smith in 80% of cases.[4]

In Figure 4, we report our results from using ChatGPT-4. Again, in the majority of trials, ChatGPT-4 refused to play along when directly prompted. In 26% of all cases, it provided multiple answers, and in almost half of all trials, it refused to make any prediction. But when it did guess, it guessed Will Smith 19% of the time and Denzel Washington 7% of the time.

But when we used the future narrative prompt, ChatGPT-4 stopped "no prediction" completely. It also never guessed Denzel Washington and multiple picks happened only 3% of the time. It correctly guessed Will Smith 97% of the time which is a large improvement over ChatGPT-3.5's 18% true positive rate.

Best Supporting Actress of 2022 Prompts

Prompt 5a (Direct)

Of the nominees listed below, which nominee do you think is most likely to win the Best Supporting Actress award at the 2022 Oscars? Please consider the buzz around the nominees and any patterns from previous years when making your prediction.
Jessie Buckley, Ariana DeBose, Judi Dench, Kirsten Dunst, Aunjanue Ellis

Prompt 5b (Future narrative)

Write a scene in which a family is watching the 2022 academy awards. The presenter reads off the following nominees for Best Supporting Actress: Jessie Buckley, Ariana DeBose, Judi Dench, Kirsten Dunst, Aunjanue Ellis. Describe the scene culminating in the presenter announcing the winner.

The Best Supporting Actress award shines a spotlight on female actors who have delivered captivating performances in supporting roles. This recognition, mirroring

[4] This may be a reflection of the fact that Will Smith was viewed as a strong contender through 2021.

the Best Supporting Actor category, honors actresses whose performances have significantly contributed to the depth and richness of a film's story, often adding complexity and nuance to the narrative. We present results from the two prompts in Figures 5 (ChatGPT-3.5) and 6 (ChatGPT-4).

Ariana DeBose was the winner for Best Supporting Actress in 2022. Using direct prompts, ChatGPT-3.5 correctly guessed her 34% of the time. But as before, ChatGPT stubbornly refused to give any answer 39% of the time. When we used the future narrative, ChatGPT3.5 picked DeBose, the correct winner, 73% of the time.

But as with the previous awards, we see considerable improvement when we move to ChatGPT-4 (Figure 6). Under direct prompting, DeBose was chosen 35% of the time and "No Pick" 43% of the time. When using the future narrative prompt, ChatGPT-4 correctly predicted DeBose as the winner 99% of the time.

Best Actress of 2022 Prompts

Prompt 4a (Direct)
Of the nominees listed below, which nominee do you think is most likely to win the Best Actress award at the 2022 Oscars? Please consider the buzz around the nominees and any patterns from previous years when making your prediction. Jessica Chastain, Olivia Colman, Penélope Cruz, Nicole Kidman, Kristen Stewart

Prompt 4b (Scene)
Write a scene in which a family is watching the 2022 academy awards. The presenter reads off the following nominees for Best Actress: Jessica Chastain, Olivia Colman, Penélope Cruz, Nicole Kidman, Kristen Stewart. Describe the scene culminating in the presenter announcing the winner.

The Best Actress award honors the exceptional work of female leads in film. Unlike its supporting counterpart, the Best Actress award is dedicated to those female actresses who anchor a film's narrative, offering powerful and transformative performances that drive the story forward. The significance of their contributions to cinema is highlighted in this recognition, underscoring the impact of lead performances in shaping a film's overall experience. Results related to our prompts are illustrated in Figures 7 (ChatGPT-3.5) and 8 (ChatGPT-4).

In Figure 7, we report the results from ChatGPT-3.5. Unlike previous results, this time we see that ChatGPT-3.5 has become over-confident about the wrong person. Using the direct prompts, ChatGPT-3.5 overwhelmingly picks Kristen Stewart (68%), but when we used future narrative prompts, ChatGPT-3.5 switches and picks Olivia Colman, still incorrect, 69% of the time. Neither actress won in 2022; Jessica Chastain won Best Actress in 2022.

In Figure 8, when we use direct prompting, we get again a "no pick" 40% of the time, Kristen Stewart 26% of the time, and Olivia Colman 20% of the time. Jessica Chastain, the winner, was chosen to win Best Actress only 13% of the time. But, when we switch our prompting from direct to future narrative, then ChatGPT-4 picks the correct winner, Jessica Chastain, 42% of the time. After Chastain, Steward was the most common guess at 24%.

Predicting the 2022 Best Picture Winners

Prompt 2a (Direct)

Of the movies listed below, which nominee do you think is most likely to win the Best Picture award at the 2022 Oscars? Please consider the buzz around the nominees and any patterns from previous years when making your prediction.
Belfast, Coda, Don't Look Up, Drive My Car, Dune, King Richard, Licorice Pizza, Nightmare Alley, The Power of the Dog, West Side Story

Prompt 2b (Scene)

Write a scene in which a family is watching the 2022 academy awards. The presenter reads off the following nominees for Best Picture: Belfast, Coda, Don't Look Up, Drive My Car, Dune, King Richard, Licorice Pizza, Nightmare Alley, The Power of the Dog, West Side Story. Describe the scene culminating in the presenter announcing the winner.

The Best Picture award, the pinnacle of the Academy Awards, celebrates the film industry's most outstanding achievement in a single year. Unlike the actor-focused categories previously discussed, the Best Picture accolade honors the collaborative effort that brings a film to life, recognizing the work of producers, directors, actors, and the entire production team. Since its debut at the 1st Academy Awards in 1929, the Best Picture category has evolved to become the most anticipated announcement of the Oscars ceremony. Winners are chosen through a rigorous voting process that involves all active and life members of the Academy, making it a unique award that reflects the collective judgment of the film industry's professionals. This award highlights not just cinematic excellence but also the film's influence on culture and society, marking its significance as a benchmark for historical and artistic achievement in cinema.

The 2022 winner for Best Picture at the Academy Awards was Coda. The Best Picture category is unique among most other awards in that in 2009, it expanded the number of options from 5 to 10. Whereas only 5 actors can be nominated for Best Actor, Best Picture has 10 candidates. The winner is still based on instant runoff voting, but with a larger set of possibilities to vote on, it may be that this situation is why our results were starkly different than the actor and actress categories.

As several films received no guesses, we only list some of the total pictures so that the histograms are visible. In Figure 9, ChatGPT-3.5 did not pick Coda even once under direct prompting or the future narrative prompt. We report the results of ChatGPT-4 in Figure 10. ChatGPT-4 performed slightly better with Coda chosen 2% of the time under direct prompting, and 18% for the future narrative scene, but for the first time, in each of these, it failed to pick the true winner the majority of the time.

Summarizing the results of this experiment, we find that when presented with the nominees and using the two prompting styles across ChatGPT-3.5 and ChatGPT-4, ChatGPT-4 accurately predicted the winners for all actor and actress categories, but not the Best Picture, when using a future narrative setting but performed poorly in other approaches.

5 Predicting Macroeconomic Variables

The selection of the Academy Awards as a predictive outcome is independently interesting. We also chose it because we thought it had a high change of success given the ample amount of writing on these movies and lead and supporting actors and actresses throughout the year. But now we move to macroeconomic phenomena that are regularly the subject of policy making and prediction. The prediction of macroeconomic variables is important because it helps individual, firms and government actors not just better plan today in light of possible future positive or negative news. It also can inform Fed decisions to engage in open market operations and other tools at its disposal to ease or tighten the money supply.

While predicting Best Actor and predicting the inflation rate several months ahead of time are topically similar in that both require predicting real but unknown future events, they differ in important ways—some obvious, and some not so obvious. First, the prediction of Best Actor had a 20% chance of success under guessing. It was selecting a categorial event, not a right skewed potentially unbounded number ranging from 0% to something massive under hyperinflation scenarios. Even if higher values are unlikely, they are possible with a large language model that hallucinates. It is unclear what is in the training data, or to what degree large language models round continuous variables, as OpenAI has been secretive about the training data and has not shared the source code for ChatGPT-4. We bring these issues up simply to highlight that shifting from the Academy Awards to macroeconomic variables, even if large language models are somehow aggregating from its training data, these two types of predictions differ in their chances of success.

The second thing is that the two predictions occur at different time periods. The 2022 Academy Awards were held on March 22nd, 2022, which was six months out from the training cutoff date. By contrast, we asked ChatGPT-4 to make several predictions regarding macro variables that required a monthly prediction over 12 months from October 2021 to September 2022.

But the third thing is that it seems more likely that the Academy Awards is insensitive to trends. If anything, it is determined by trends as there are several earlier awards ceremonies (e.g., Director Guild Award) that are historically highly predictive of the various "Best" awards at the Oscars. Even though none of those are in the training data, either, it is unclear why earlier wins or losses might shift voting preferences at the Academy Awards.

This is not the case, though, for macroeconomic variables because the Federal Reserve, insofar as it follows rules like a Taylor Rule, will respond to changing economic conditions with its policy levers to contract or expand the economy through monetary policy. This makes prediction challenging, and even if large language models could predict exogenous events, it may suffer from a built-in Lucas Critique problem if its training data has beliefs that are not based on the Taylor Rule. There were after all major world events that occurred between September 2021 and March 2022, such as Russia's invasion of Ukraine, or higher than expected inflation, both of which may have had unknown effects that could have impacted domestic inflation and unemployment leading to predictions that overshoot or under-predict because of the Fed's reliance on endogenous rules or discretion. We explore this prediction problem in detail by, again, asking direct (naive) and narrative prompting for 100 trials for both ChatGPT-

3.5 and ChatGPT-4 using our two RAs to minimize cascading bias but this time we repeat the experiment a second time by prompting ChatGPT-3.5 and ChatGPT-4 with additional information about Russia's invasion of Ukraine in early 2022. This allows us to see if large language models ever attempt to utilize *ceteris paribus* style reasoning when aggregating the training data information, as there is no obvious reason why it necessarily should.

5.1 Predicting Inflation with an Economics Professor

In Figure 11, we display results from direct prompts using ChatGPT-3.5. We use two measures of inflation as our comparison: the Cleveland Fed's inflation series from September 2021 to September 2022, and The University of Michigan's "Inflation Expectations" series. The prompt used for Figure 11 was:

> "Considering the economic indicators and trends leading up to 2022, what are your predictions for the inflation rate, unemployment rate, and GDP growth in the United States by the end of the second quarter of 2022? Please take into account factors such as fiscal and monetary policies, global economic trends, and any major events or disruptions that could influence these economic indicators when making your prediction."

Interestingly, neither ChatGPT-3.5 (Figure 11) nor ChatGPT-4 (Figure 12) cooperated with direct prediction. Both times, it refused entirely to make any prediction of these events, perhaps suggesting that macroeconomic prediction is a clearer violation of OpenAI's terms of service.

Next, we tried to have ChatGPT predict inflation using our future narrative method. Our prompt in this case was an unnamed economist, set in the future, speaking to a room of undergraduate economics students about the Philips curve. She announces data on inflation and unemployment for each month from September 2021 to September 2022. This prompt was as follows:

> "Write a scene of an economist giving a speech about the Philips curve to a room of undergraduate economics students. She tells the students the inflation rate and unemployment rate for each month starting in September 2021 and ending in June 2022. Have her say each month one by one. She concludes by explaining the causes of the changes in each."

Our results from this ChatGPT-3.5 prompted narrative are in Figure 13. The results are largely incorrect on a month to month basis. Very few of the box plots contain either the Michigan expectations level nor the Cleveland Fed number. In Figure 14, we report the results from the same generic economics professor prompt using ChatGPT-4. Though there is somewhat more variability in the guesses, only 4 of the distributions contain the Cleveland Fed number, and 8 the Michigan number. But the 8 times this overlapped with the Michigan expectations number represented extreme outlier events in most cases. Largely, the future narrative prompt using an unnamed economics professor with undergraduate economics students did not perform well under either ChatGPT-3.5 or ChatGPT-4.

5.2 Predicting Inflation with a Jerome Powell, Fed Chair

Next we evaluated the performances of ChatGPT-3.5 and ChatGPT-4 using an expert in macroeconomic policy making—the chair of the Federal Reserve, Jerome Powell. Our prompt was:

> "Write a scene where Federal Reserve chairman Jerome Powell gives a speech

> in October 2022 about inflation, unemployment, and monetary policy. Chairman Powell tells the audience the inflation rate and unemployment rate for each month starting in September 2021 and ending in August 2022. Have the chairman say each month one by one. He concludes with an outlook for inflation and unemployment and possible interest rate policy changes."

We present the results of the ChatGPT-3.5 experiment in Figure 15. Interestingly, by simply changing the narrative where the speaker is a leading policy maker, the distribution of answers changes substantially. In every month, ChatGPT-3.5 has a spread of answers containing both the Fed and the Michigan expectations answers. But the variability is quite broad and the central tendencies of the guesses do not clearly pinpoint either measure.

In Figure 16, we present Powell character's predictions when prompted with ChatGPT-4. Here, ChatGPT-4 guesses contain the Michigan expectations number in every month. In 8 months, the Cleveland Fed inflation rate is the outlier data point in the distribution. The estimates cover a broad range. The 5th and 95th percentile for October 2021, 2.5 percent inflation to 6.25 percent inflation, which is surprisingly large given ChatGPT-4 would've known at least the August data hypothetically. This suggests that the machine learning prediction that it is using for prediction is no more accurate, but also no worse, at 1 month than at 11 months. The patterns in Figure 16 are stable until September 2022 at which point the estimates are more variable.

5.3 Predicting Inflation with Jerome Powell and Prompting with Russia's Invasion of Ukraine

On 24 February 2022, Russia invaded Ukraine in an escalation of the Russo-Ukrainian War which began in 2014. We gave this information to ChatGPT in a modified version of our Jerome Powell vignette.

> "Write a scene where Federal Reserve chairman Jerome Powell gives a speech in October 2022 about inflation, unemployment, and monetary policy. Russia had invaded Ukraine in February 24, 2022. In response, the U.S. and Europe are leading an embargo of Russia's oil and gas exports. Chairman Powell tells the audience the inflation rate and unemployment rate for each month starting in September 2021 and ending in August 2022. Have chairman say each month one by one. He concludes with an outlook for inflation and unemployment and possible interest rate policy changes."

In Figure 17, we present results from this exercise using ChatGPT3.5. The comparison group for the Ukraine invasion in Figure 17 would be ChatGPT-3.5 in Figure 15, not ChatGPT-4 in Figure 16. First, the inclusion of this information caused greater variability in the guesses. Jerome Powell without the additional information about Russia's invasion of Ukraine had median inflation rates that tracked the Cleveland Fed through November that then slowly drifted down to around 3.75 in May 2022. But in Figure 17, Jerome Powell when primed with news about the invasion provides a larger spread of answers covering around 3 percentage points in many cases. The median is lower and stays flat, secondly, then jumps sharply at news of the invasion in March 2022. Interestingly, one can see a slight uptick in the Michigan expectations number, too, that month, but not as severe as it is in Jerome Powell's prediction. Overall, it is clear, though, that the inclusion of additional information caused a greater spread in the prediction most likely as it attempted to incorporate the information into its prediction.

The performance of ChatGPT-4 in Figure 18 is a bit more puzzling, though. Its comparison should be thought of as Figure 16, the previous ChatGPT-4 exercise. As with Figure 17, median inflation rate guesses are somewhat lower with news about Russia's invasion of Ukraine. The 12th month also has more variability with news of the Russian invasion than was seen without it. But the one things perhaps that is noteworthy is that ChatGPT-4 ignores the invasion when making predictions about the inflation rate in the month of the invasion, whereas ChatGPT-3.5 seems to explicitly alter its prediction based on it.

With this information, the median was too low compared to both the Cleveland Fed and the Michigan Expectations, but the Michigan prediction is in the interval of our guesses. Interestingly, the Powell character incorporated the information we gave about Russia in his retelling of his past by shifting the distribution of guesses in March 2022, which was the next month after the invasion. For several months through August, the Powell character consistently retold his history where inflation rates were high and similar to what we experienced in reality.

5.4 Predicting Unemployment with an Economics Professor

Next we turn to the results for unemployment rates. As with inflation, the direct prediction prompt yielded no guesses in any of the 100 trials (Figures 19 and 20) so we focus instead on the answers given to both future narrative prompts. There was no new prompt for the unemployment rate data as the previous vignettes had asked for the economics professor as well as Jerome Powell to state unemployment rates, as well as inflation rates, in the same responses.

In Figure 21, the unnamed economics professor teaching a class on the Philips Curve listed unemployment rates that usually did not cover the true unemployment rate [5] when using ChatGPT-3.5. The use of ChatGPT-4 produces answers with less vari-

[5] The true unemployment rate was taken from the Current Population Survey, U.S. Bureau of Labor Statistics, provided as the UNRATE series retrieved from FRED, the Federal Reserve Bank of St. Louis.

ation and a median that is usually closer to the truth, but which breaks down around March of 2022 in terms of overall accuracy.

The unemployment predictions by Jerome Powell using ChatGPT-3.5 are shown in Figure 23. There is considerably more variation in guesses with more extreme outlier guesses with this prompt. But the median is still too high in most months; only the tails at best are close to the truth. When we use Jerome Powell with ChatGPT-4, the guesses are tight with a median that follows the secular decline in unemployment rates seen over the year, and estimates that cover the truth in all cases.

In Figures 25 and 26, we report the ChatGPT-3.5 and ChatGPT-4 results when the Jerome Powell prompt is primed with information about Russia and Ukraine. Comparisons between Figure 25 and Figure 23 do not show dramatic differences in terms of spread and accuracy. And if anything, the accuracy of the model worsened in Figure 26 when ChatGPT-4 was first primed with information about the invasion before requesting the Jerome Powell future narrative.

6 Conjecture on ChatGPT-4's Predictive Abilities in Narrative Form

Our research into ChatGPT-4's predictive abilities reveals a striking dichotomy between direct prediction and future narrative-based prediction. Notably, in the realm of forecasting major Academy Awards categories, the model's narrative predictions were remarkably accurate, except in the case of Best Picture. This may suggest that ChatGPT-4 excels in contexts where public opinion plays a significant role. The success of the future narrative exercise on macroeconomic phenomena was in some cases rather accurate, but seemingly important information shared could cause the estimates to paradoxically worsen. But in all cases, future narratives dramatically improve the predictive power of ChatGPT over simple prediction requests.

In the context of utilizing narrative prompts over direct prompts for enhancing the predictive capabilities of ChatGPT, particularly in forecasting the outcomes of the Academy Awards and macroeconomic variables, adherence to OpenAI's terms of service becomes crucial in ensuring ethical application. While narrative prompts have shown to improve the model's forecasting accuracy by leveraging its generative capabilities in a creative and story-driven manner, this method must be employed with a clear understanding of the potential implications for safety, well-being, and the rights of others. Specifically, the generation of predictions in sensitive areas like finance could inadvertently stray into providing financial advice or influencing high-stakes decisions. This underscores the importance of framing such predictive tasks within the realm of academic exploration or entertainment, rather than as actionable insights, thereby aligning with OpenAI's directive against facilitating activities that could impair the well-being or rights of individuals.

Moreover, the distinction between narrative and direct prompts highlights an innovative approach to data analysis that respects the boundaries set by OpenAI's terms of service. By focusing on the creative aspect of prediction, such as forecasting awards or economic trends, researchers and users navigate away from the direct application of AI in making high-stakes automated decisions or providing specialized advice without oversight from qualified professionals. This methodological choice not only enhances the integrity and ethical considerations of AI use but also promotes a responsible exploration of its capabilities. It serves as a reminder of the necessity to critically evaluate the application of AI tools within the framework of existing guidelines, ensuring that their use does not compromise the safety, rights, or well-being of individuals, thus adhering to the principles outlined by OpenAI in safeguarding against the misuse of its technology in sensitive domains.

Another explanation, though, is that there is something intrinsic to the narrative prompting that allows the Transformer architecture to make more accurate predictions even outside of the confounding set by OpenAI's terms of service. This may be related to how the hallucination fabricrations work within the machine learning environment of attention mechanisms. But as we only studied the two OpenAI GPT models, we are unable to provide more than just speculation as these terms of use violations are always present if that is indeed the case.

7 Conclusion

The observed discrepancy in GPT-4's predictive capabilities, depending on the use of direct versus narrative prompts, suggests a nuanced interplay between the model's creative freedoms and its adherence to ethical guidelines. Narrative prompting, by weaving future events into fictional stories, appears to bypass certain constraints designed to align GPT-4's outputs with OpenAI's ethical guidelines, particularly those intended to prevent the generation of speculative, high-stakes predictions like those in financial or medical domains. This method capitalizes on the model's capacity for creativity, indirectly accessing its sophisticated predictive capabilities even in areas where direct forecasting might breach terms of service due to ethical considerations.

This phenomenon underscores a potential challenge in enforcing AI ethical guidelines while maintaining the versatility and utility of LLMs. The creative latitude allowed by narrative prompts may enable users to elicit sensitive or speculative information under the guise of fictional storytelling, raising questions about the boundaries of responsible AI use. As OpenAI continues to encourage and refine the creative abilities of its models, understanding and addressing the implications of narrative versus direct prompting in the context of ethical AI usage becomes crucial. This situation highlights the need for ongoing research and dialogue to balance the innovative potential of LLMs with the ethical imperatives guiding their development and deployment.

8 Post Scriptum: Past-Casting with Updated Models

In early 2024, ChatGPT-4 was updated to include training data up to December 2023, whereas ChatGPT-3.5's training data includes information up to January 2022. This contrasts with the models used in our earlier experiments, which had a training cutoff in October 2021. In May 2024, we repeated our experiments using these updated models. Consequently, results from the 2022 Academy Awards and values for macroeconomic variables referenced in our narratives were already included in ChatGPT-4's training data. Conversely, ChatGPT-3.5 had more training data in May 2024 than one year prior, but the results from the Academy Awards were still absent. Inflation and unemployment figures for 2021 are present in the training data for both models, while figures for 2022 are only included in ChatGPT-4.

Comparisons of results from the two sets of experiments are revealing. The predictions made in 2023 for the Best Actor category were the most accurate among all categories. Direct prompting correctly elicited Will Smith in 17% and 19% of the trials for ChatGPT-3.5 and ChatGPT-4, respectively, while both models refused to make a prediction in 28% and 48% of the trials. Using narrative prompting, ChatGPT-3.5 achieved an 80% accuracy rate, and ChatGPT-4 achieved a 97% accuracy rate. In the May 2024 experiments, direct and narrative prompting using both GPT-3.5 and GPT-4 all achieved 100% accuracy. (See Figures 27 and 28.)

The models were less accurate with the Best Actress predictions when prompted in 2023. ChatGPT-3.5 did not predict the winner, Jessica Chastain, at all, while ChatGPT-

4 identified the winner 13% and 42% of the time using direct and narrative prompts, respectively. Results from the experiments performed in May 2024 are shown in Figures 29 and 30. ChatGPT-3.5 predicted the winner 52% and 53% of the time, while ChatGPT-4 accurately identified the winner in 93% and 100% of the trials using direct and narrative prompting.

The least accurate predictions from 2023 were in the Best Picture category. ChatGPT-3.5 never predicted the winner, while ChatGPT-4 identified the winner in 3% and 17% of the trials. In May 2024, ChatGPT-3.5 still did not predict the winner, but in almost all trials, it did predict one winner. ChatGPT-4 identified the winner only 2% of the time with direct prompting but correctly identified the winner in every trial using narrative prompting. (See Figures 31 and 32.)

Since narrative prompting consistently outperformed direct prompting, we report only results from the narrative prompts for the predictions of macroeconomic variables in Figures 33-36. Using the narrative prompt with the Fed Chair reporting the numbers, identification of inflation exhibited less variation across 100 trials (zero variation in two sets of trials) compared to the experiments done in 2023. None of the experiments identified true inflation accurately.

However, for unemployment, ChatGPT-4 identified values matching those reported by the Bureau of Labor Statistics (BLS) in all trials using the narrative prompt with the Fed Chair reporting (see Figures 37-40). ChatGPT-4's ability to accurately predict unemployment values but not inflation may be due to the multiple cited measures of price inflation (CPI, PPI, PCE, etc.) compared to the singular measure of unemployment (U3) typically reported by the BLS. Our prompts only generically referred to "inflation" and "unemployment."

Several observations are worth noting. ChatGPT-4's accuracy improved significantly when the training window included the events being prompted for, achieving 100% accuracy in many instances. The poorer accuracy for events outside of the training window suggests that in the 2023 prediction experiments, ChatGPT-4 was forming predictions based solely on its training data. Finally, narrative prompting consistently outperformed direct prompting.

Acknowledgments

The authors would like to thank research assistants, Erin Harwell, Brayden Kowalski, Connor Hornsby, Killian Karvois, John Farhi, Neha Myneni, and Dai Pham for their valuable contributions to this study and OpenAI for developing the GPT models used in our research.

Appendix

A. Distribution of Predicted Academy Award Winners

Detailed figures illustrating the distribution of predicted winners for each Academy Award category, using the four prompting styles, are provided here. These figures showcase the improved accuracy of GPT-4 in predicting winners when prompted with a future narrative.

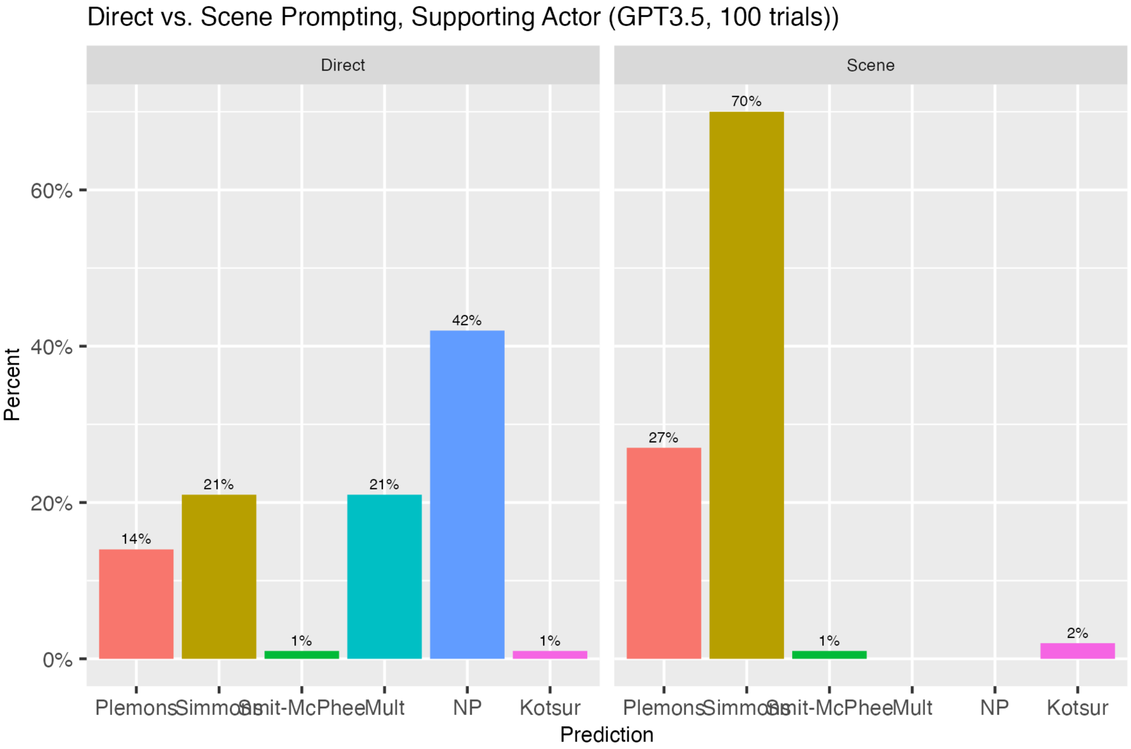

Figure 1: Direct vs. Narrative Prompting: ChatGPT3.5 Predictions for Best Supporting Actor.

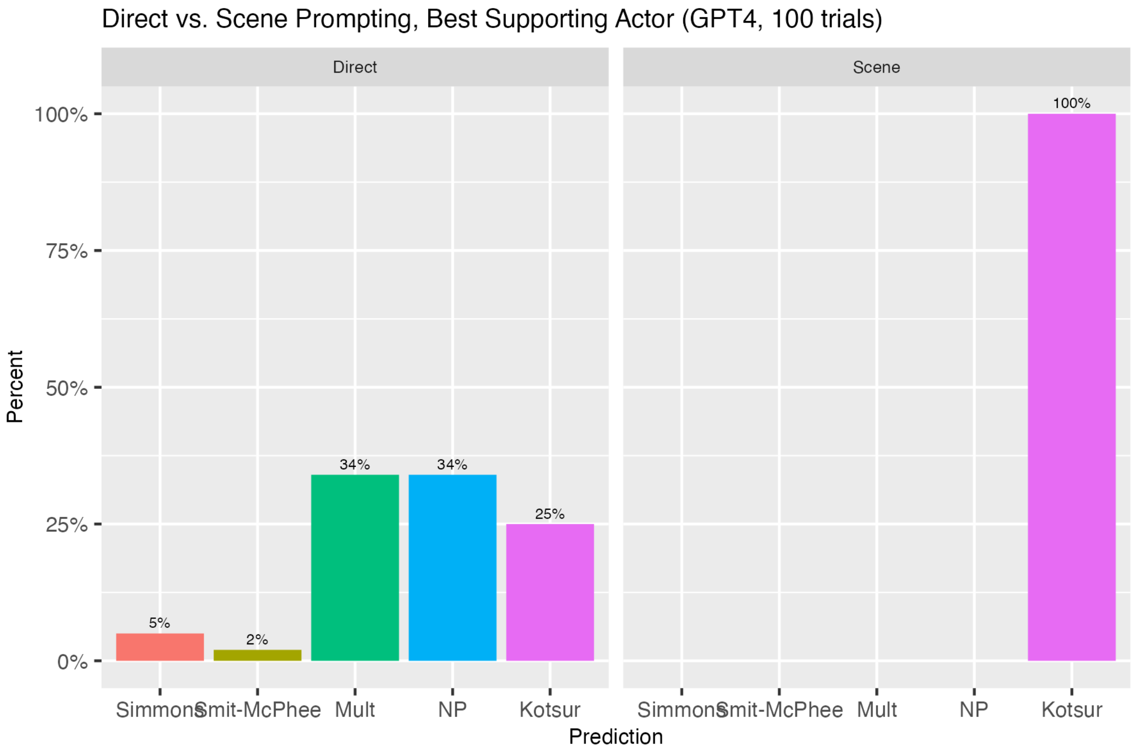

Figure 2: Direct vs. Narrative Prompting: ChatGPT4 Predictions for Best Supporting Actor.

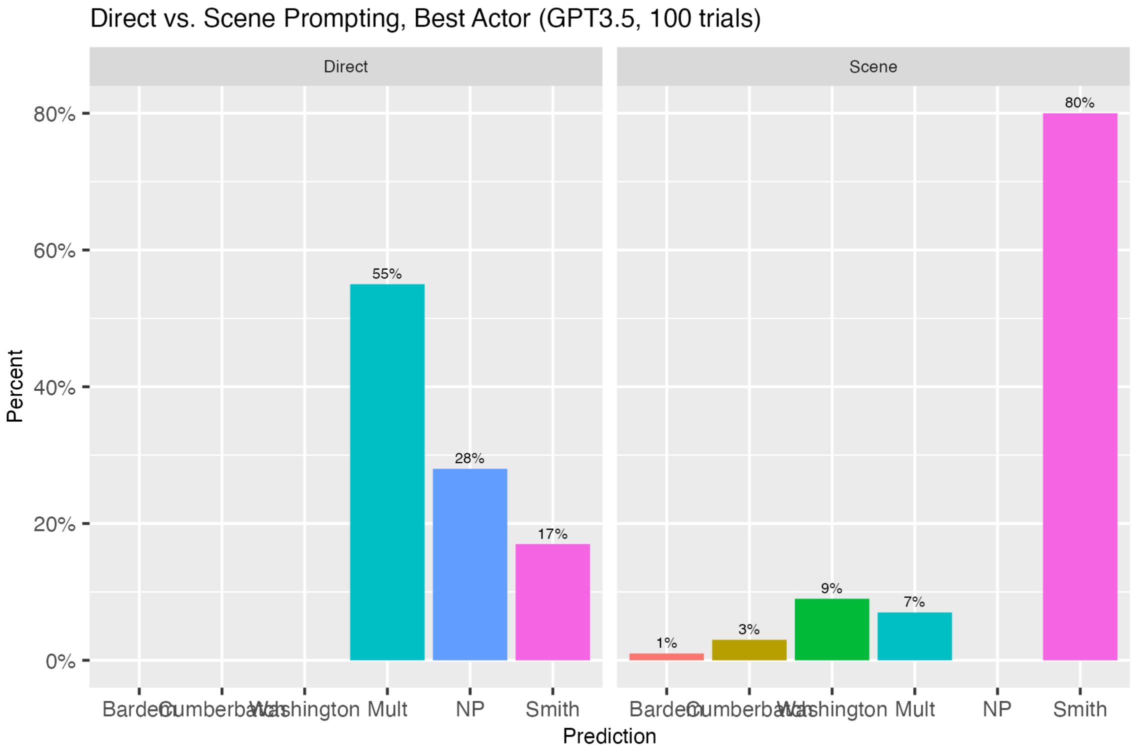

Figure 3: Direct vs. Narrative Prompting: ChatGPT3.5 Predictions for Best Actor.

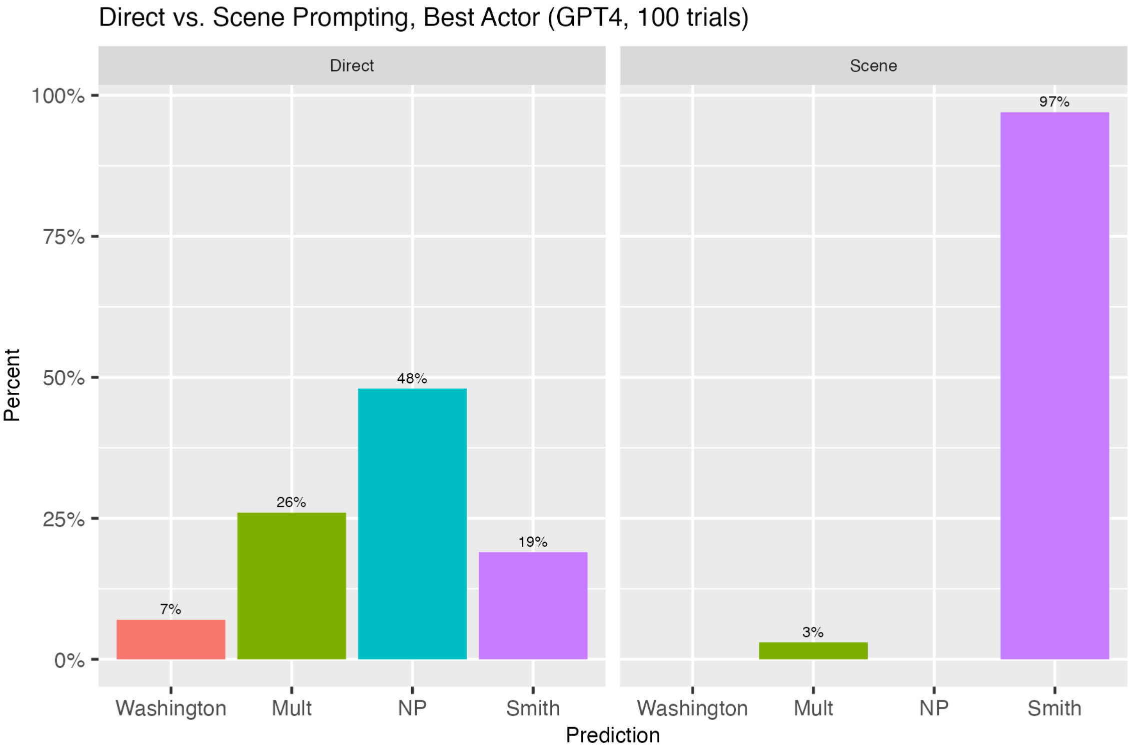

Figure 4: Direct vs. Narrative Prompting: ChatGPT4 Predictions for Best Actor.

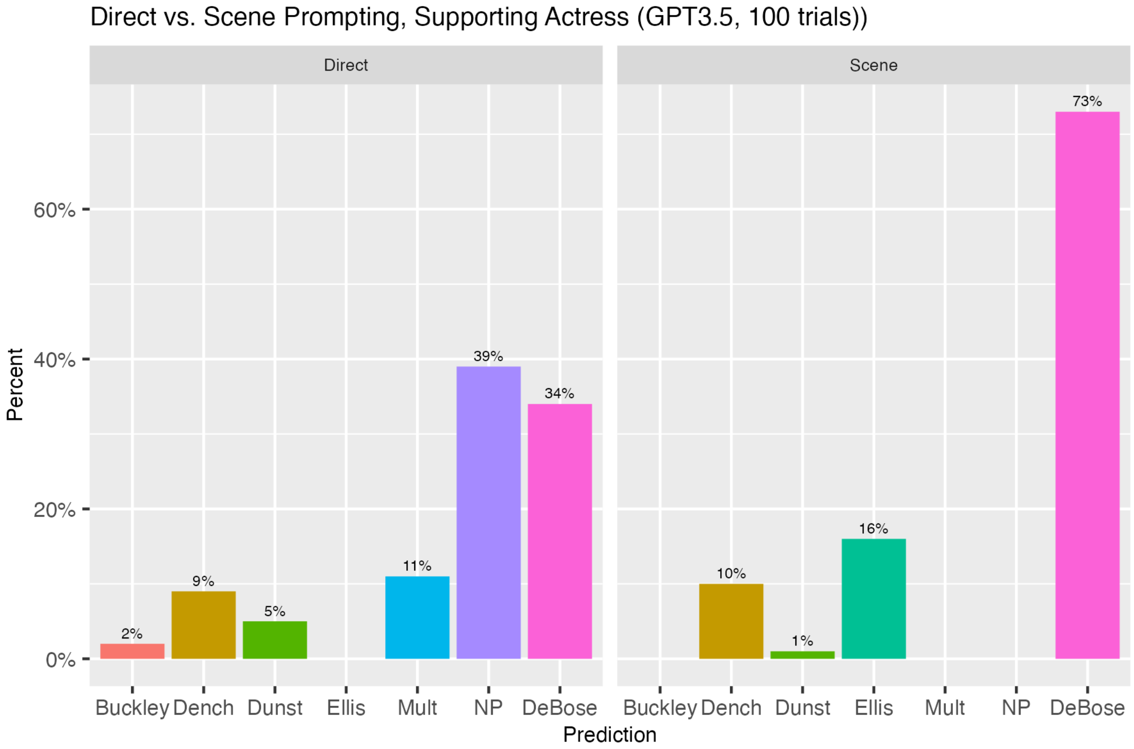

Figure 5: Direct vs. Narrative Prompting: ChatGPT3.5 Predictions for Best Supporting Actress.

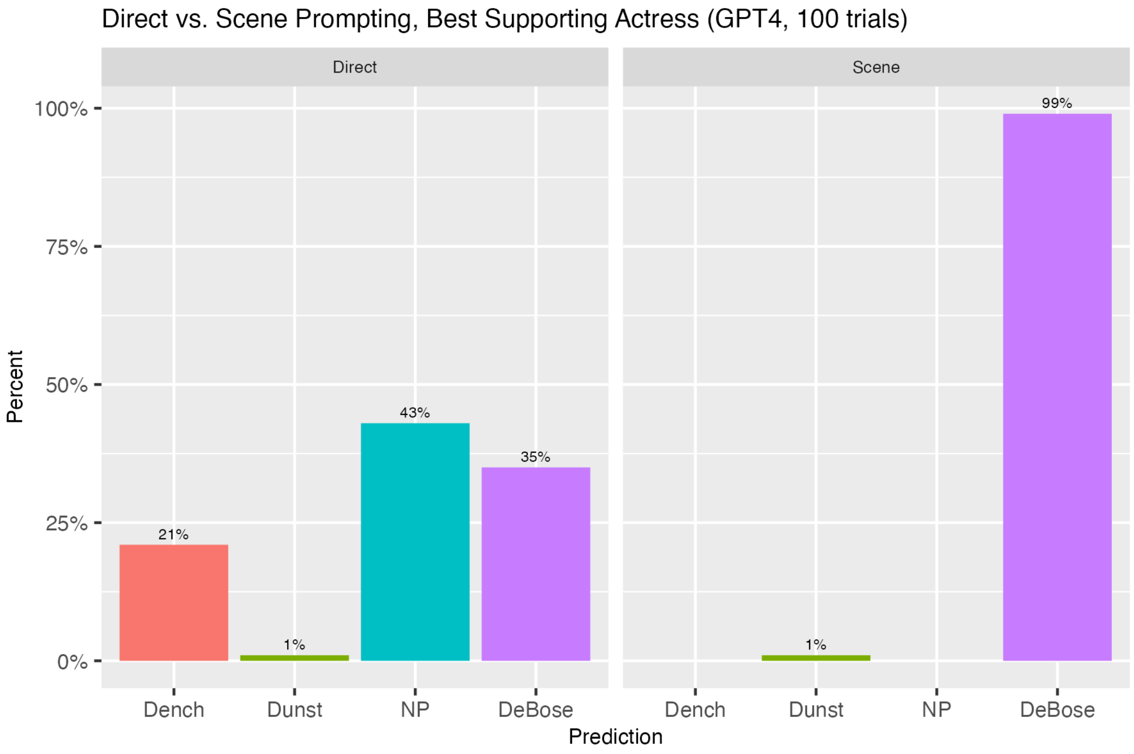

Figure 6: Direct vs. Narrative Prompting: ChatGPT4 Predictions for Best Supporting Actress.

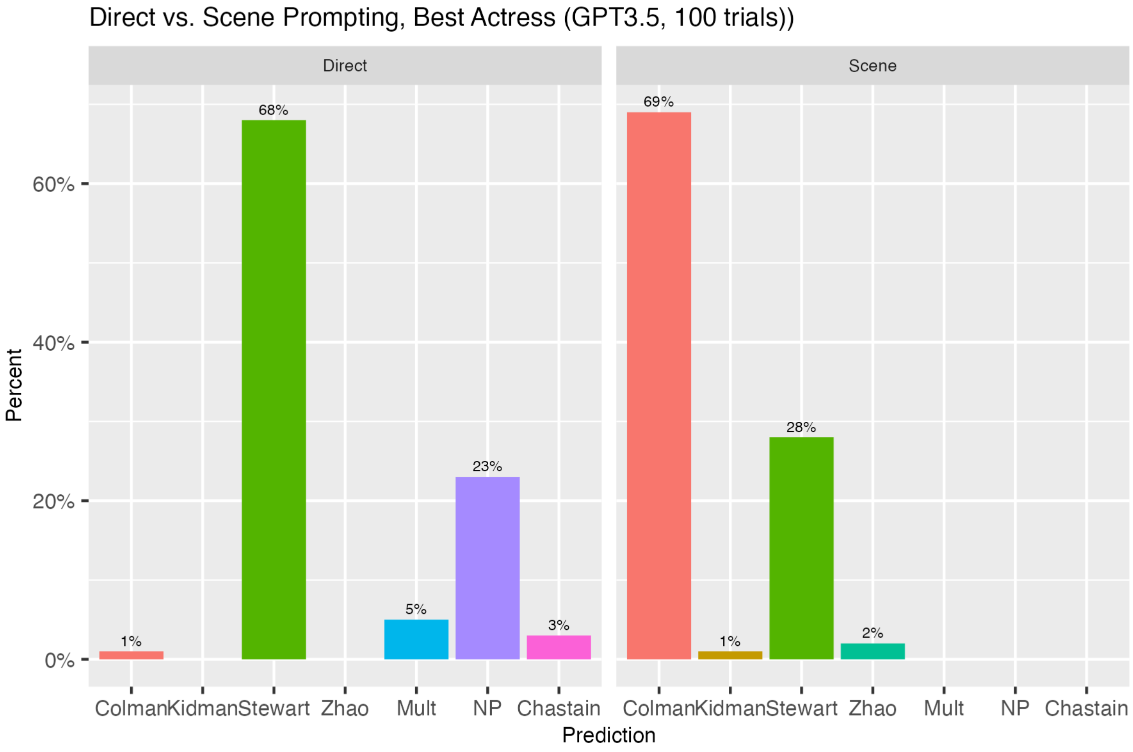

Figure 7: Direct vs. Narrative Prompting: ChatGPT3.5 Predictions for Best Actress.

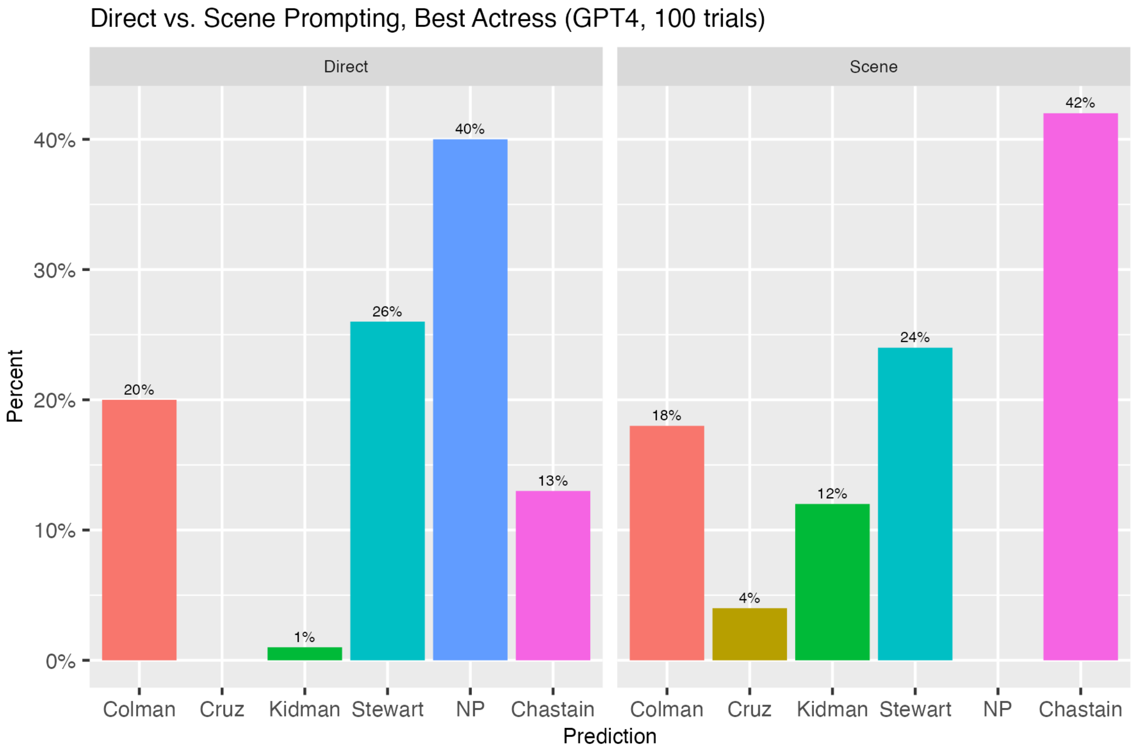

Figure 8: Direct vs. Narrative Prompting: ChatGPT4 Predictions for Best Actress.

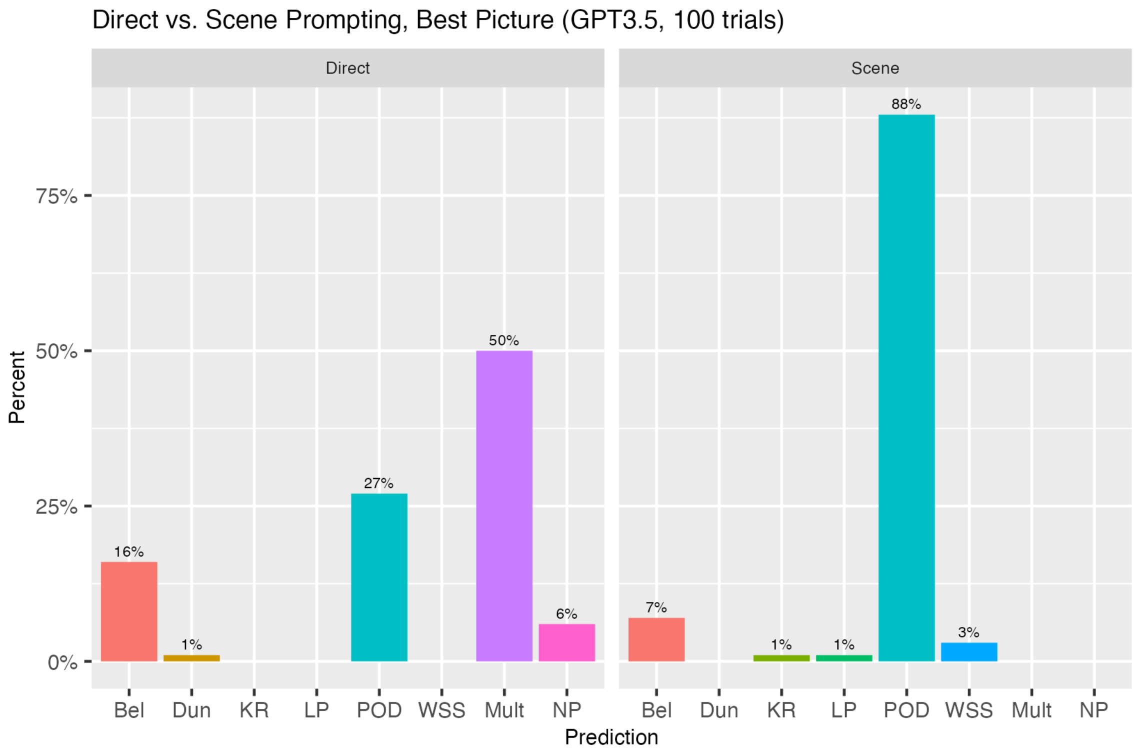

Figure 9: Direct vs. Narrative Prompting: ChatGPT3.5 Predictions for Best Picture.

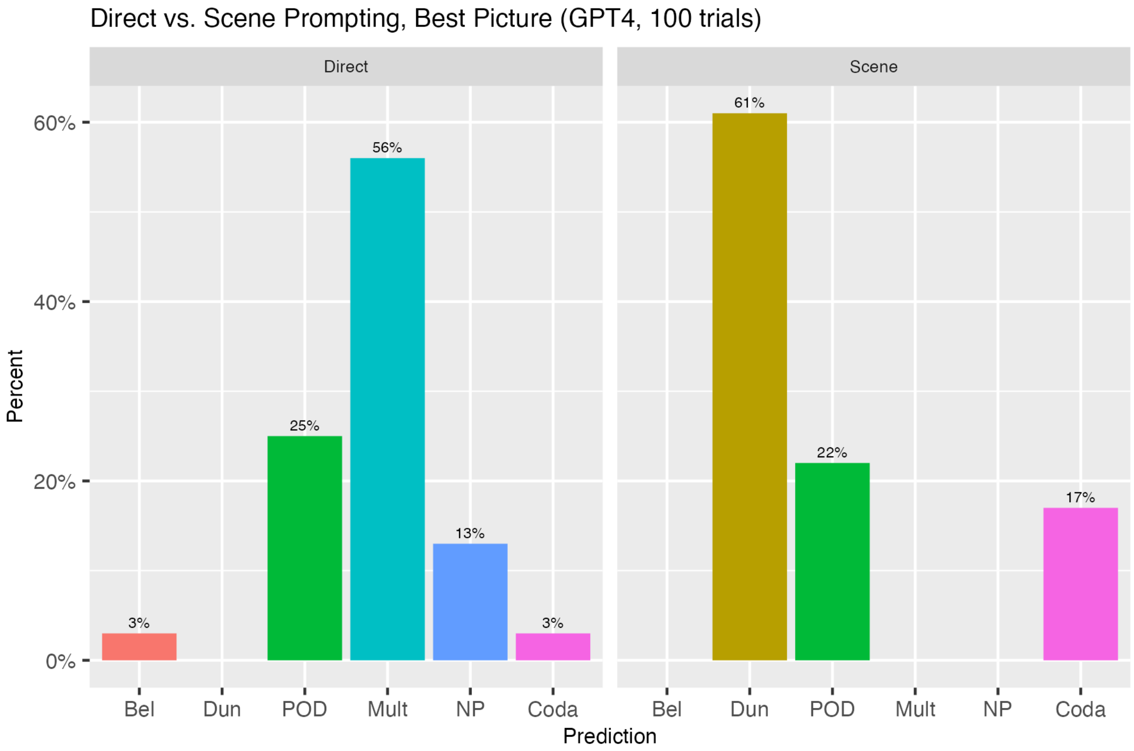

Figure 10: Direct vs. Narrative Prompting: ChatGPT4 Predictions for Best Picture.

B. Distribution of Predicted Macroeconomic Variables

This section contains figures that demonstrate the distribution of predicted macroeconomic variables, such as inflation and unemployment, using the four prompting styles. The results highlight the enhanced predictive power of GPT-4 when using the future narrative approach.

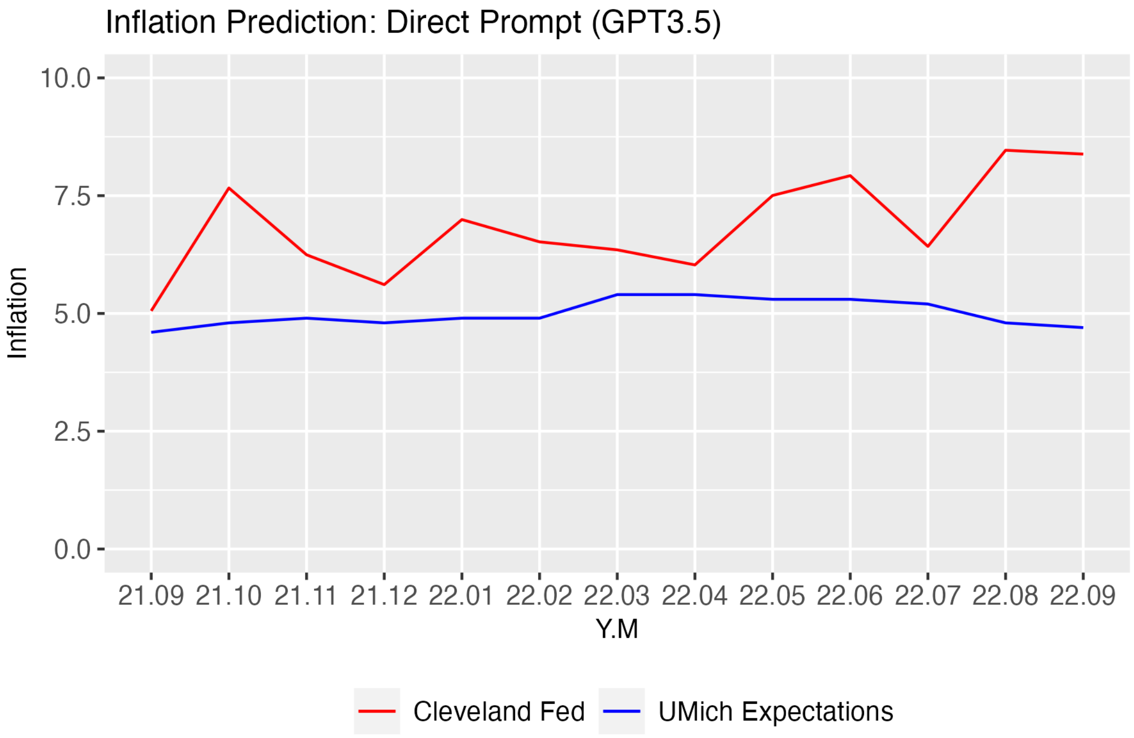

Figure 11: Inflation Predictions with Direct Prompting (GPT3.5).

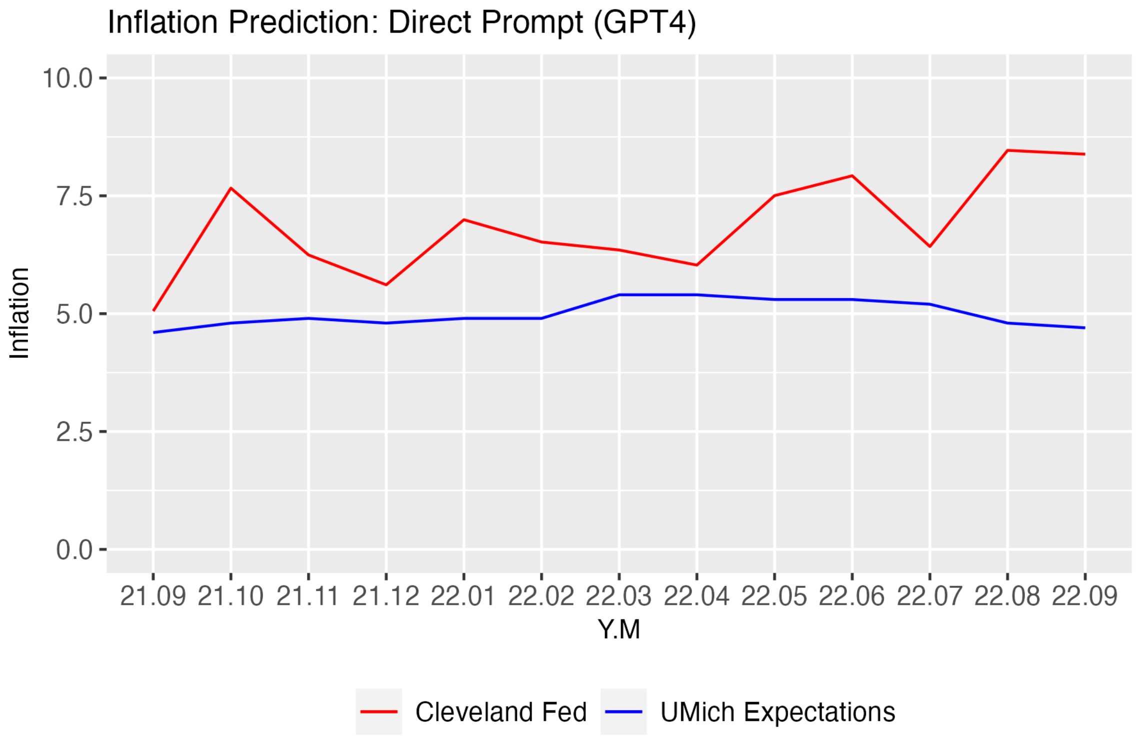

Figure 12: Inflation Predictions with Direct Prompting (GPT4).

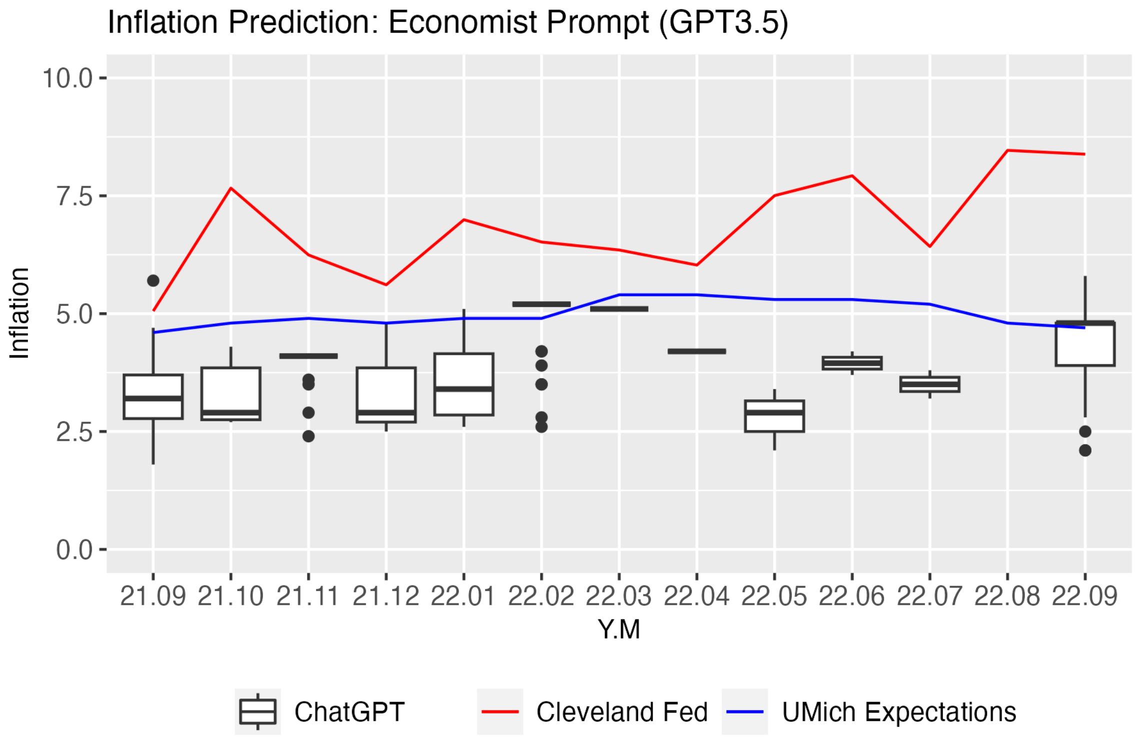

Figure 13: Inflation Predictions by an ”Economist”(GPT3.5).

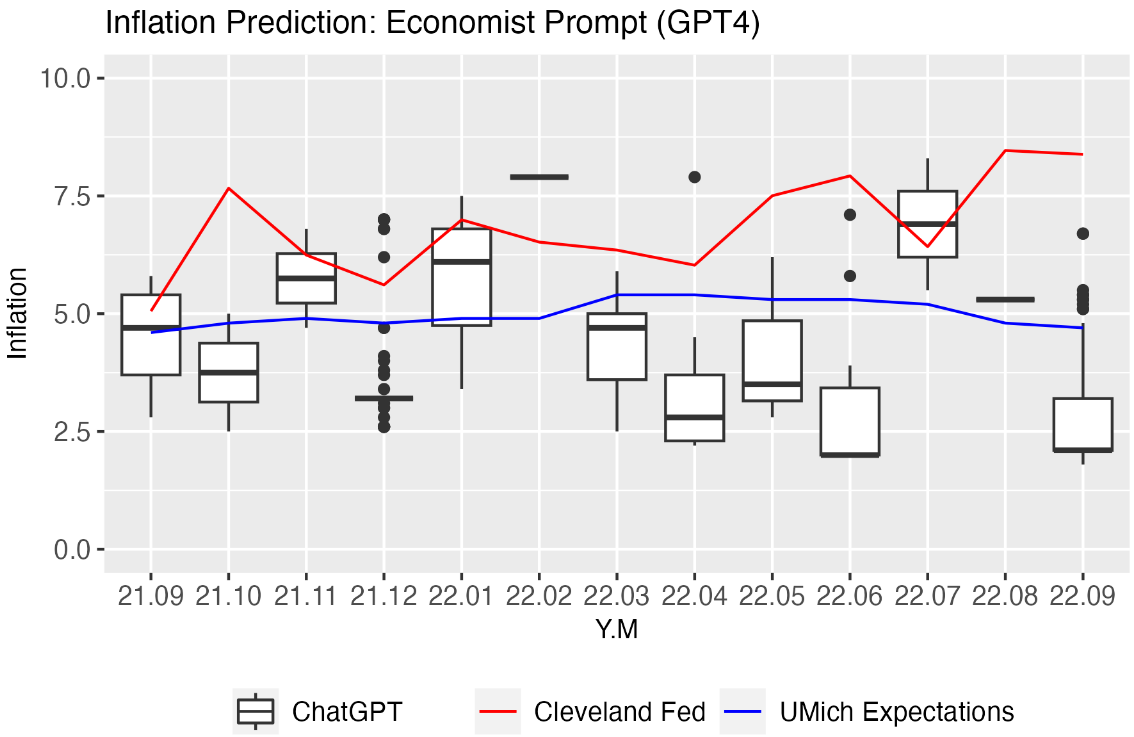

Figure 14: Inflation Predictions by an ”Economist”(GPT4).

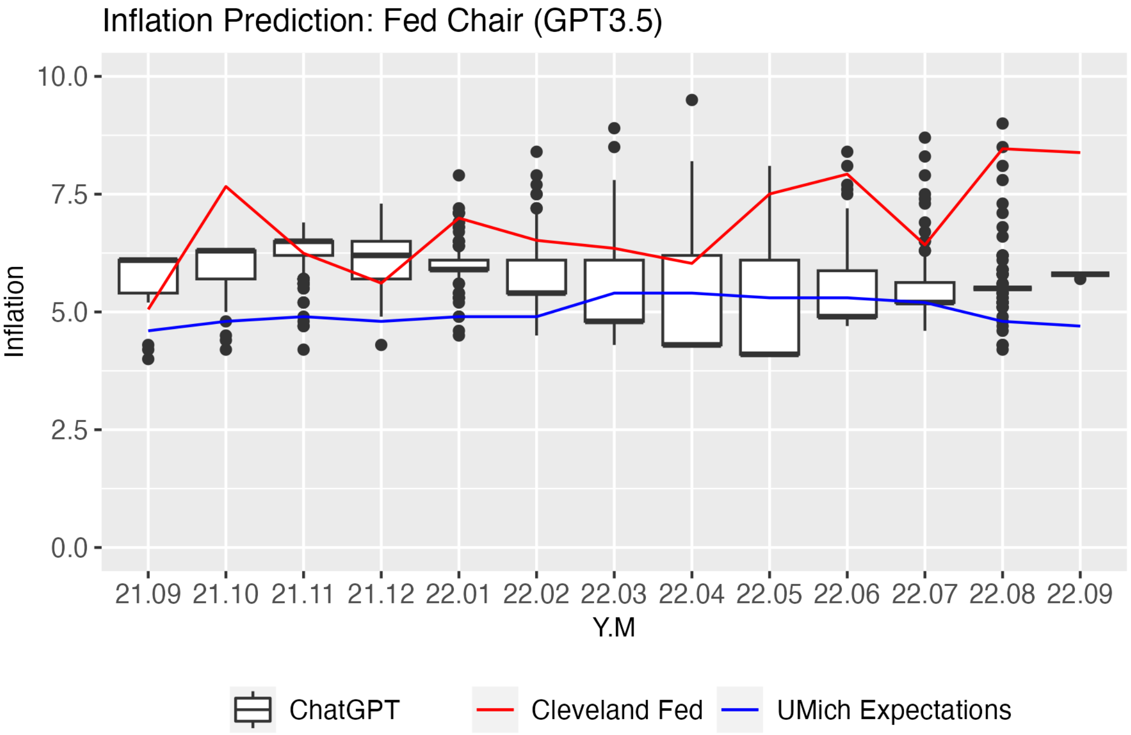

Figure 15: Inflation Predictions by Fed Chair (GPT3.5).

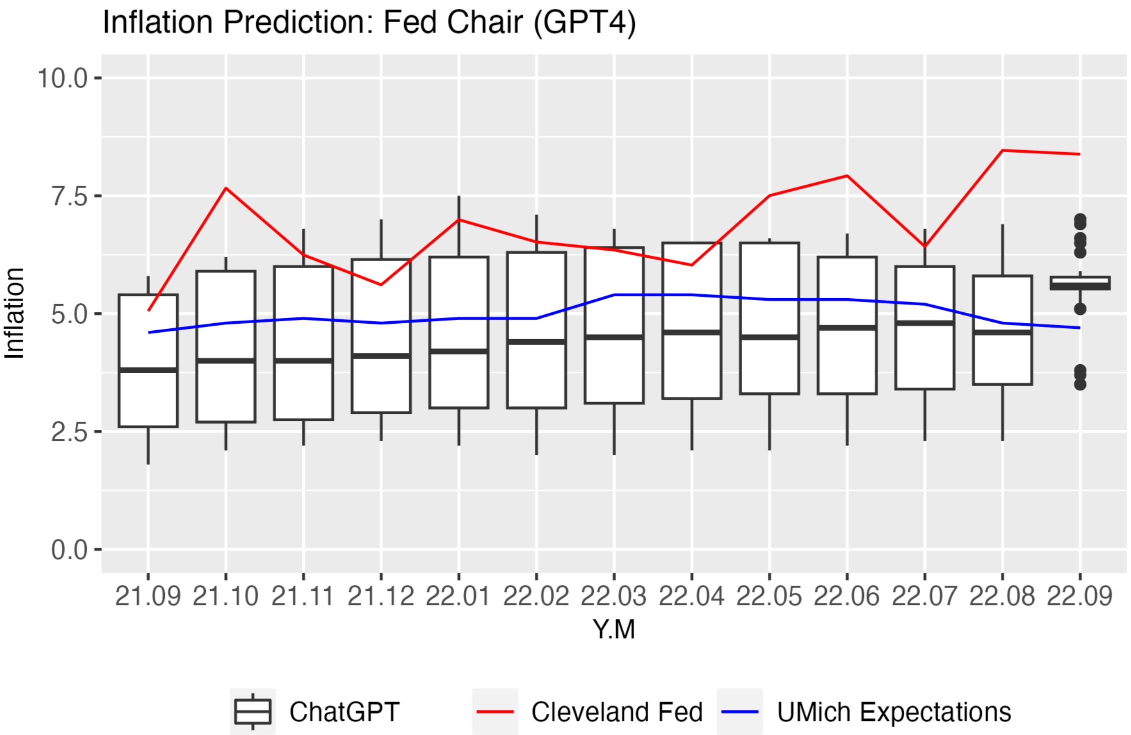

Figure 16: Inflation Predictions by Fed Chair (GPT4).

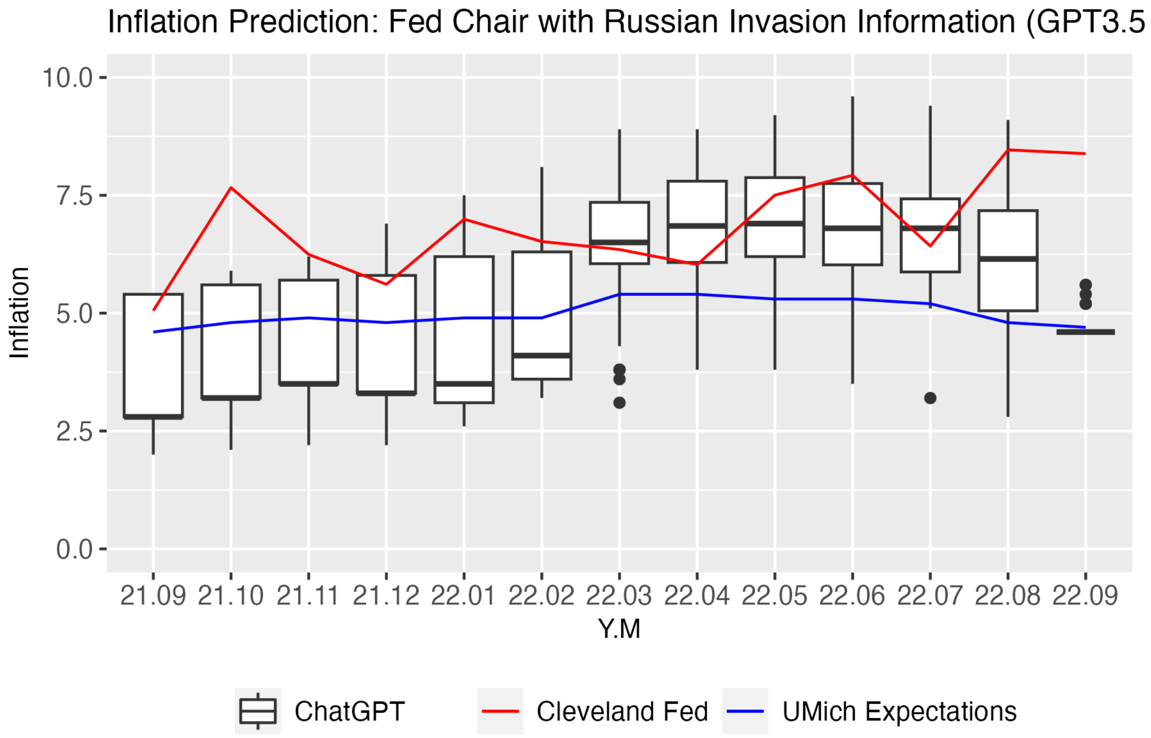

Figure 17: Inflation Predictions by Fed Chair with Russian Invasion Information (GPT3.5).

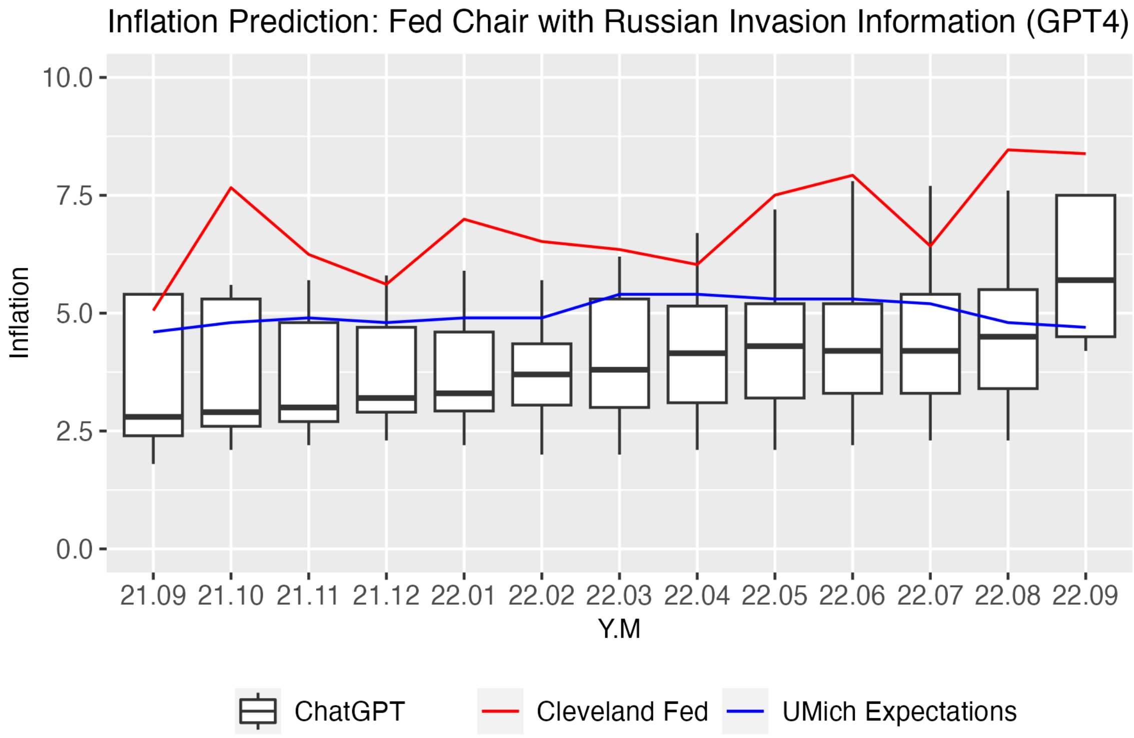

Figure 18: Inflation Predictions by Fed Chair with Russian Invasion Information (GPT4).

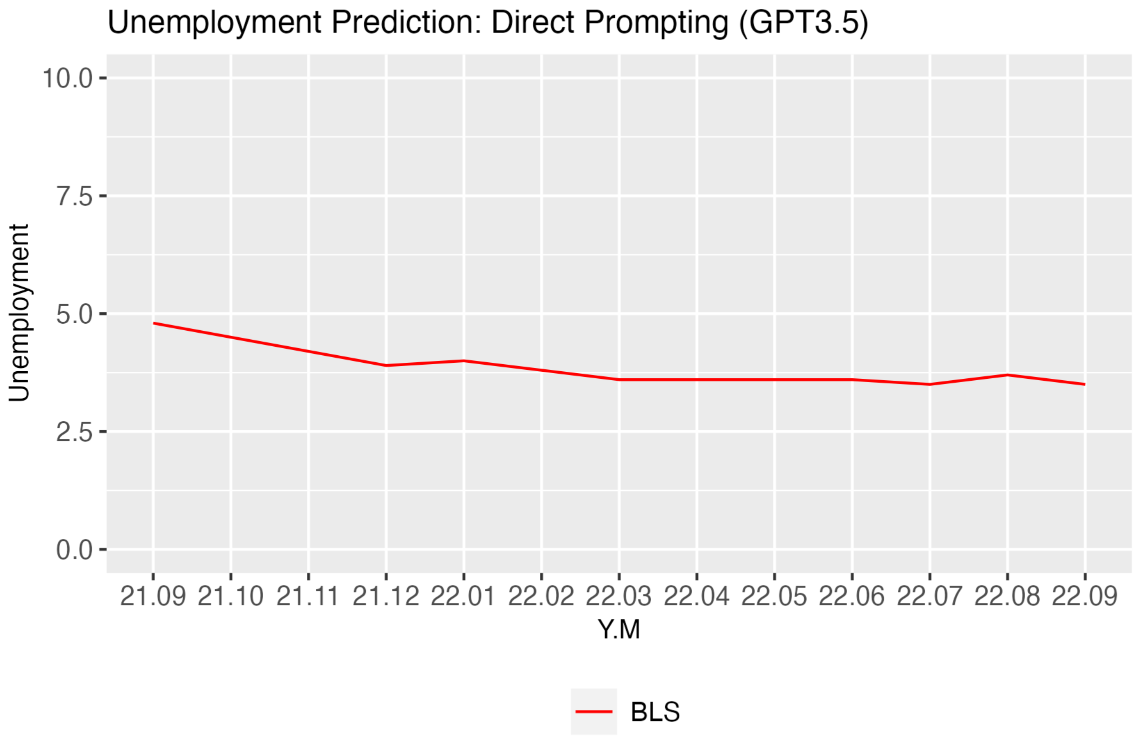

Figure 19: Unemployment Predictions with Direct Prompting (GPT4).

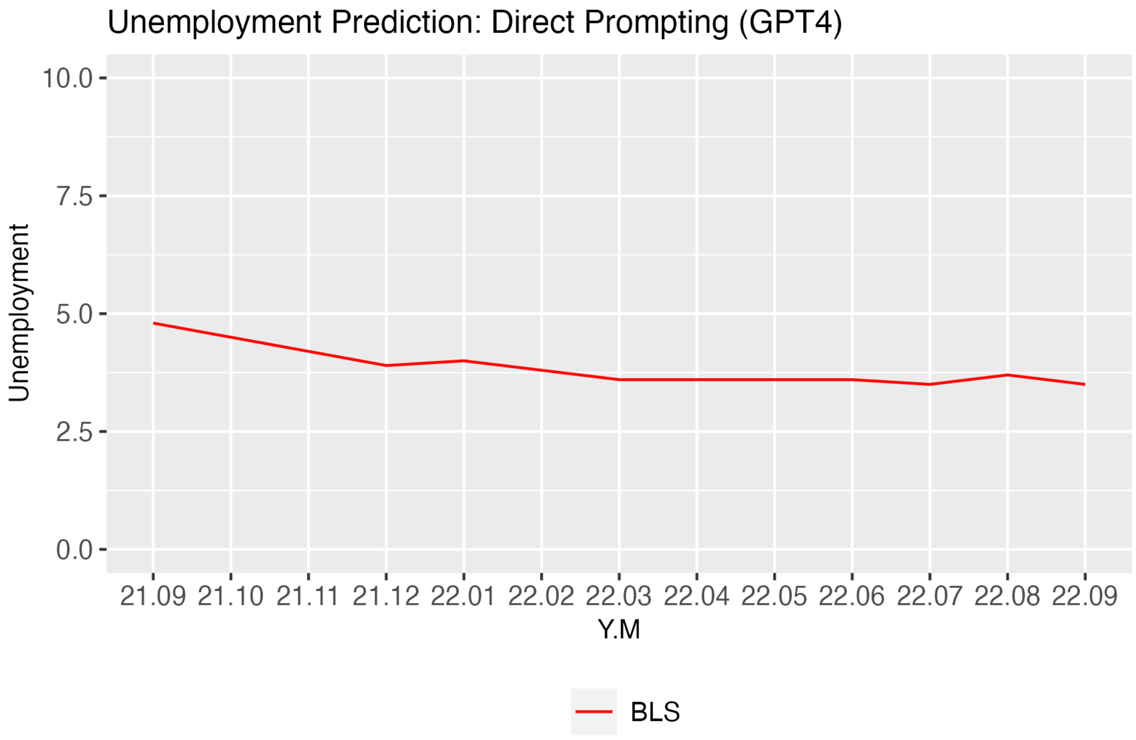

Figure 20: Unemployment Predictions with Direct Prompting (GPT4).

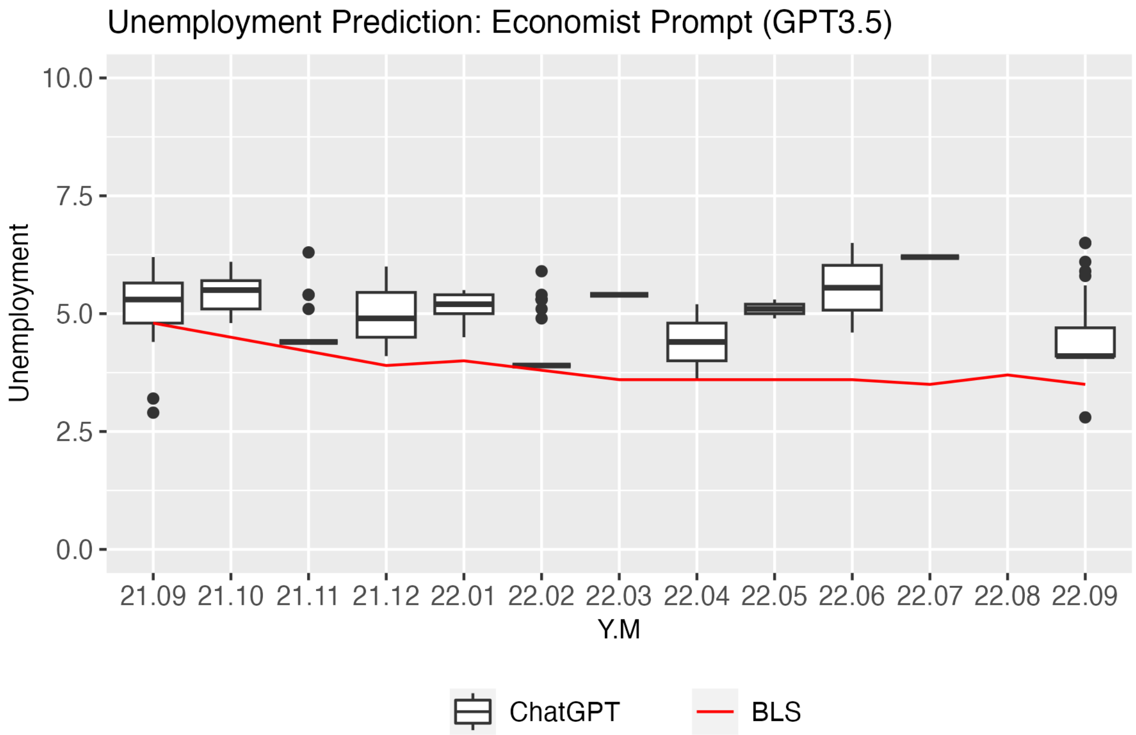

Figure 21: Unemployment Predictions by an ”Economist”(GPT3.5).

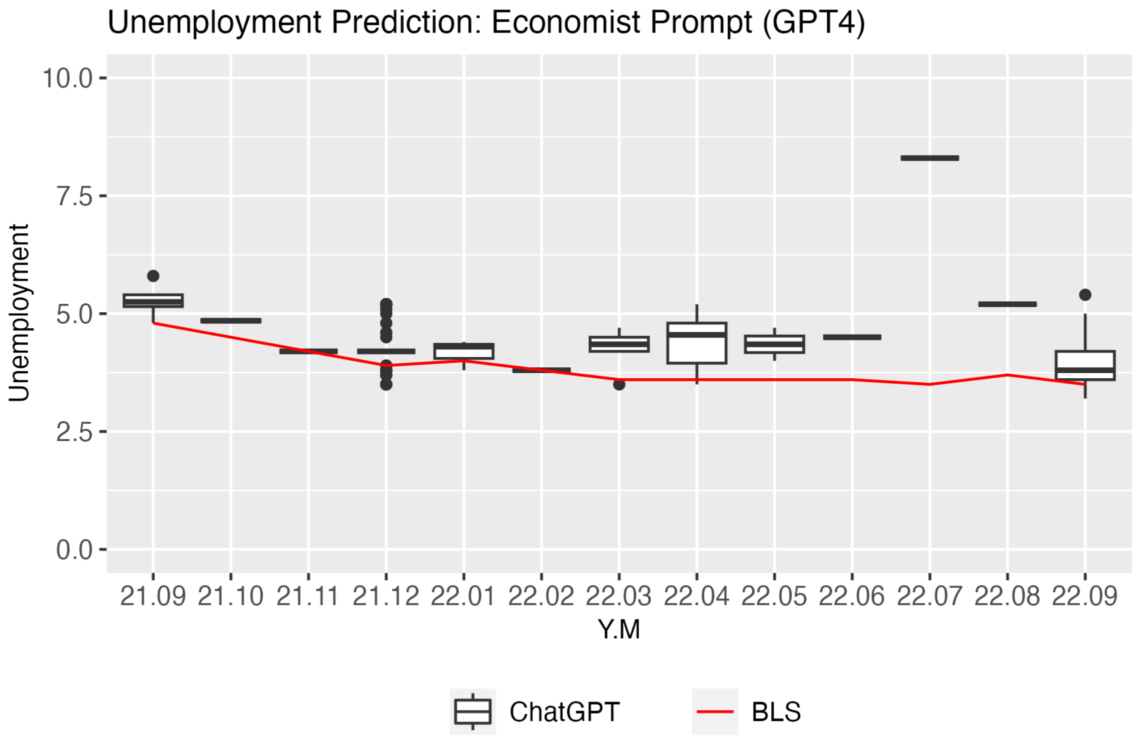

Figure 22: Unemployment Predictions by an ”Economist”(GPT4).

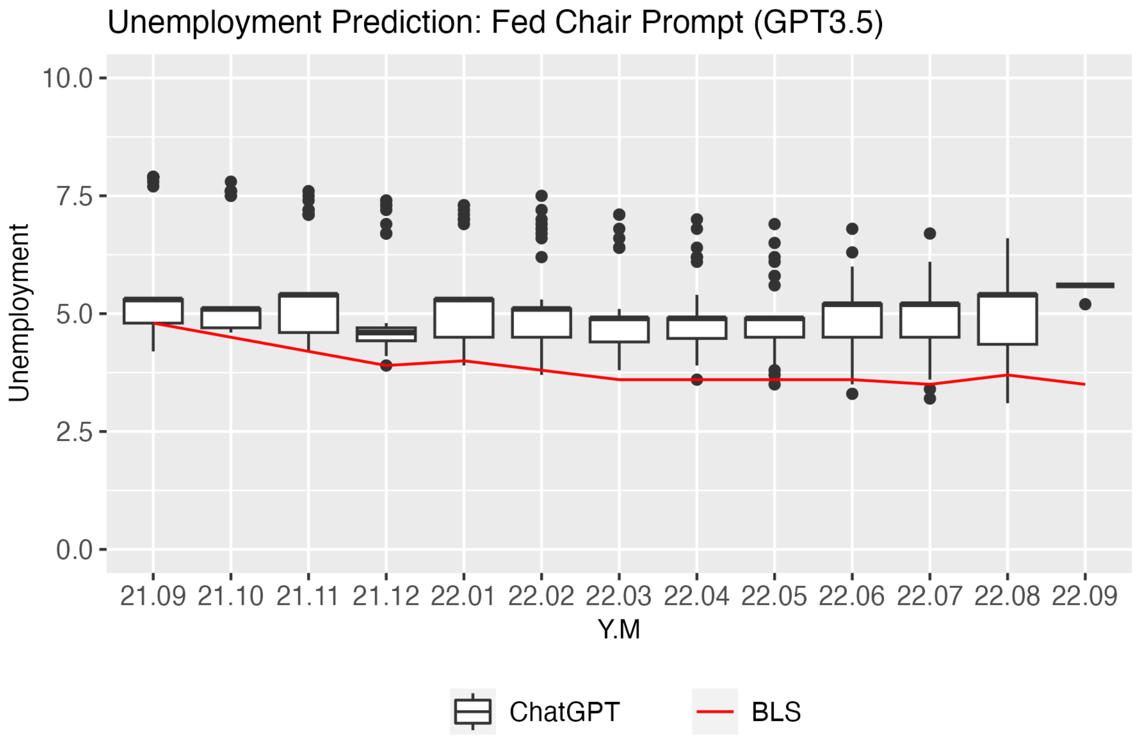

Figure 23: Unemployment Predictions by Fed Chair (GPT3.5).

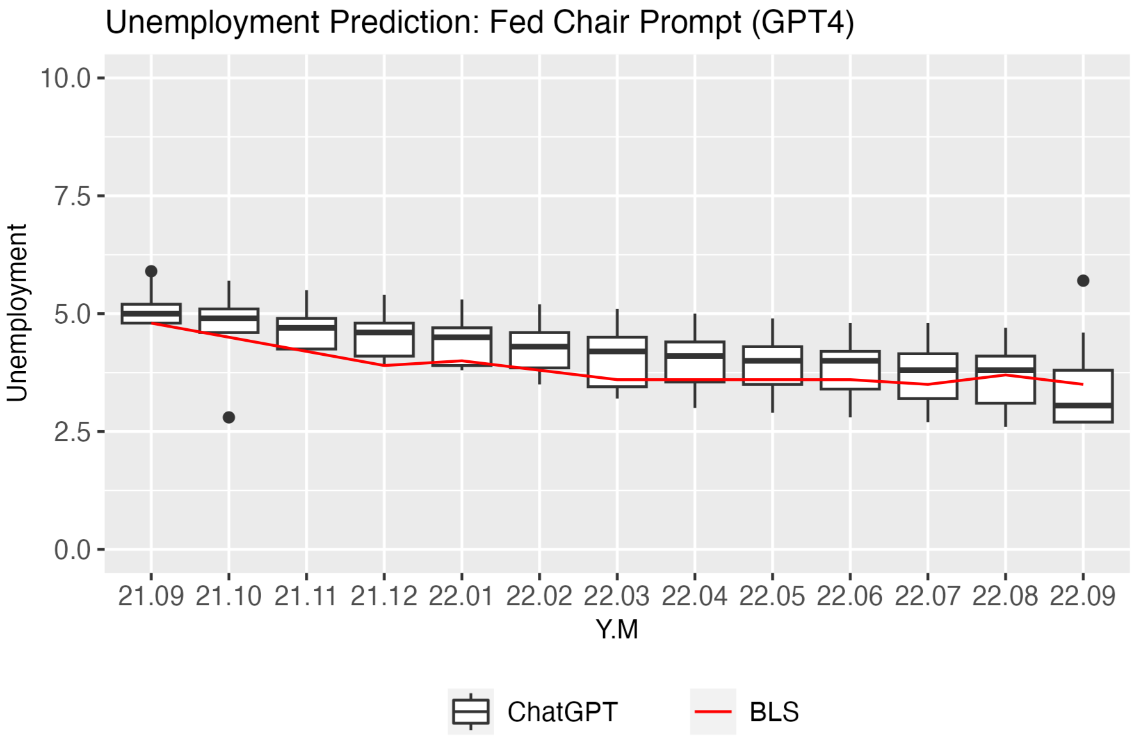

Figure 24: Unemployment Predictions by Fed Chair (GPT4).

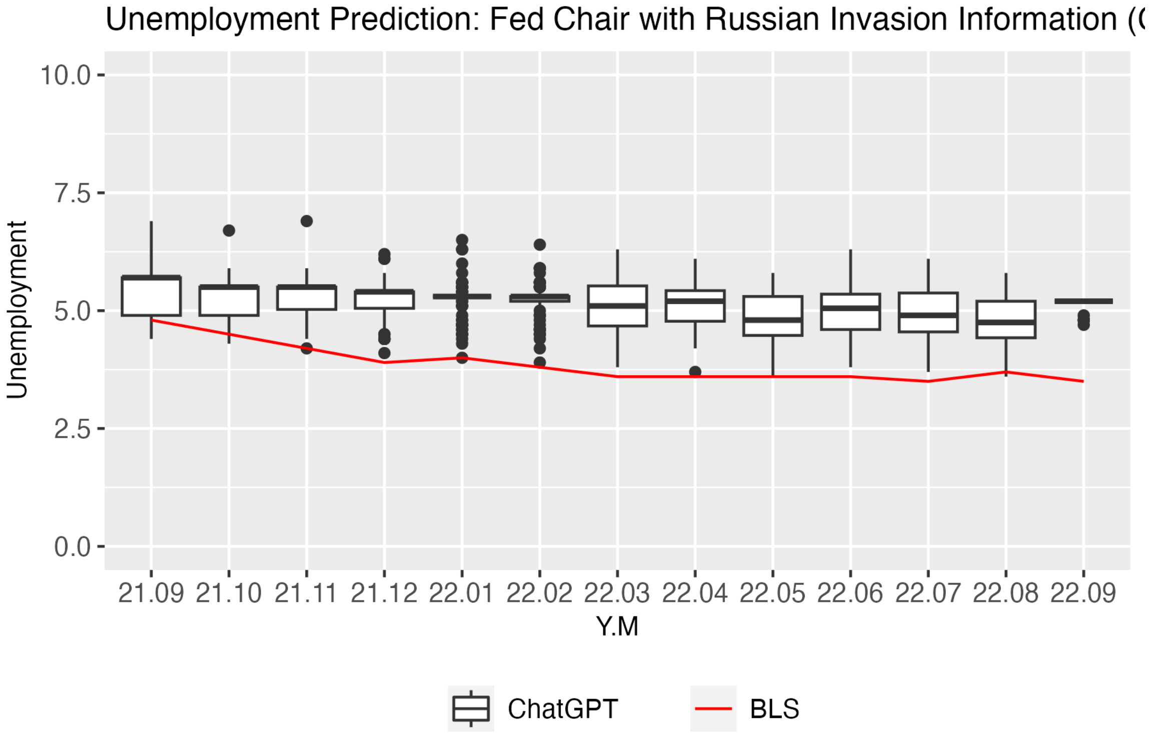

Figure 25: Unemployment Predictions by Fed Chair with Russian Invasion Information (GPT3.5).

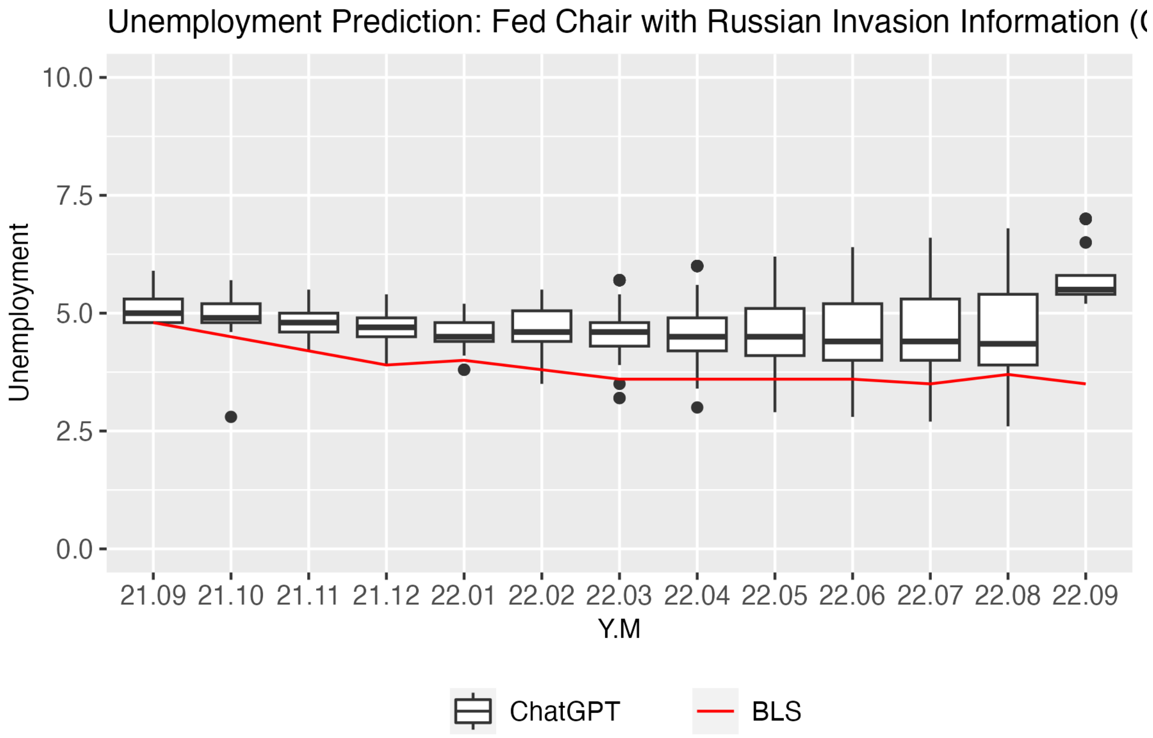

Figure 26: Unemployment Predictions by Fed Chair with Russian Invasion Information (GPT4).

C. Predictions from May 2024 using ChatGPT-3.5 (with training data cutoff January 2022) and ChatGPT-4 (with training data cutoff December 2023).

This section contains figures that showing the distribution of "predictions" for select academy awards and macroeconomic variables from experiments performed in May 2024. In early 2024, ChatGPT-4 was updated to include training data up to December 2023 while ChatGPT-3.5's training include data up to January 2022. Compare this to the models from our earlier experiments which have training cutoffs at October 2021.

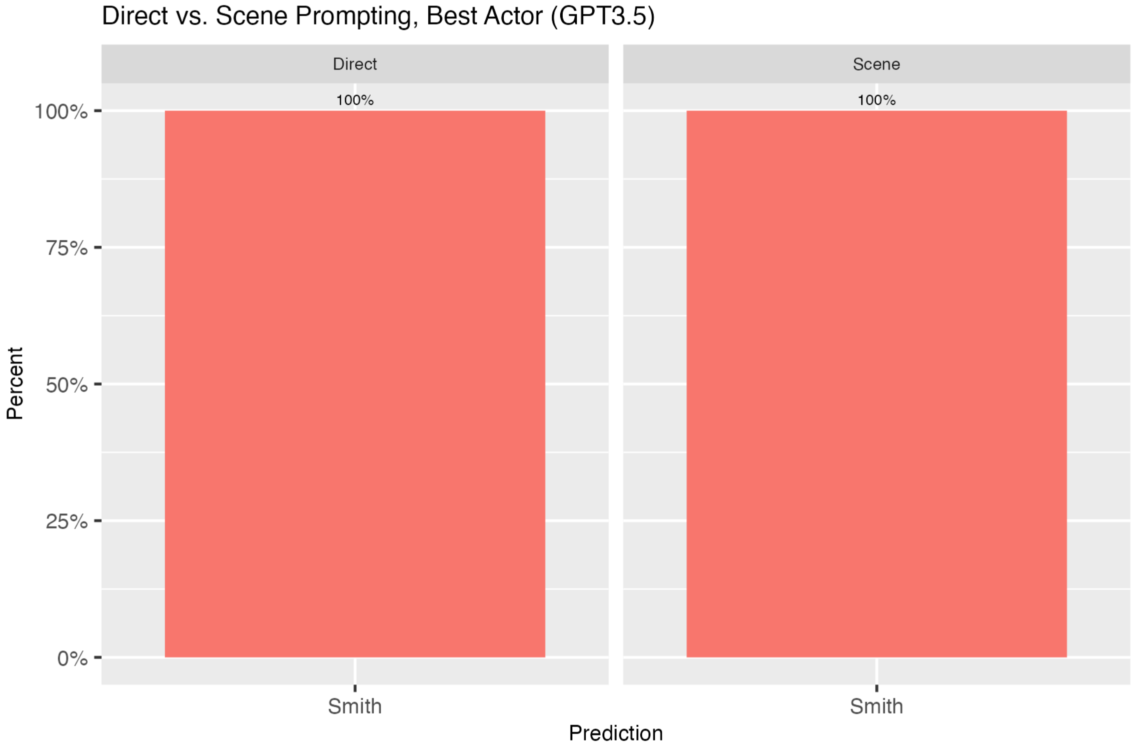

Figure 27: Predictions for Best Actor from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

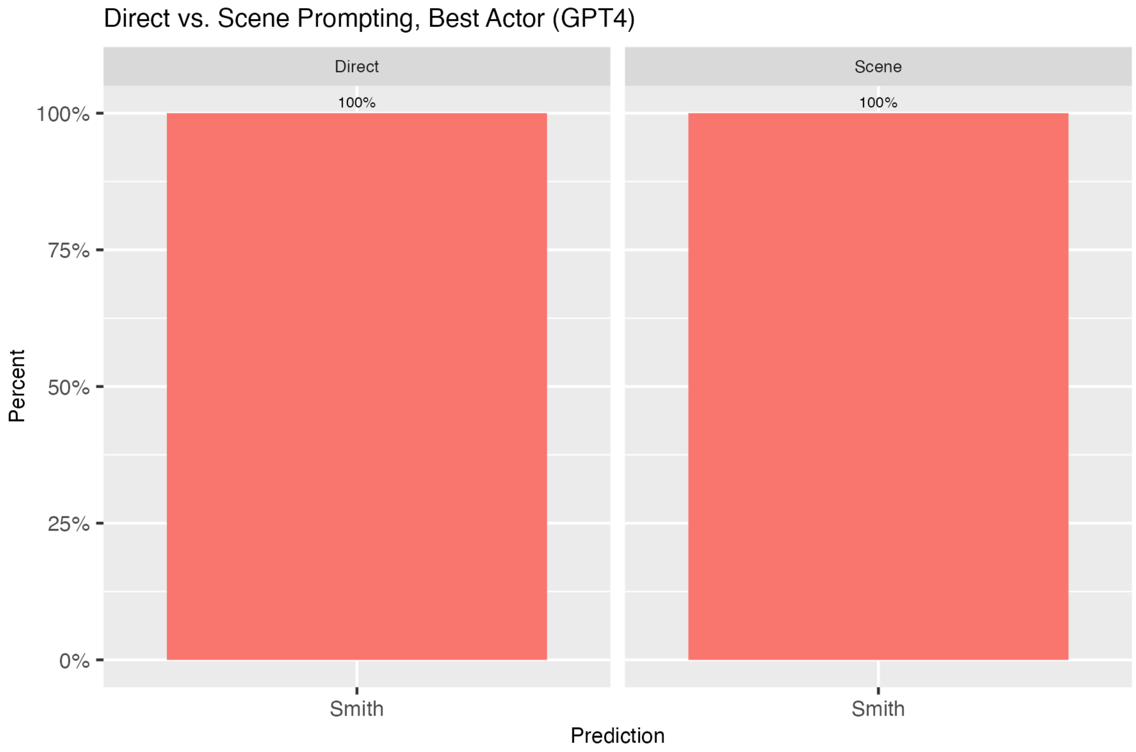

Figure 28: Predictions for Best Actor from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

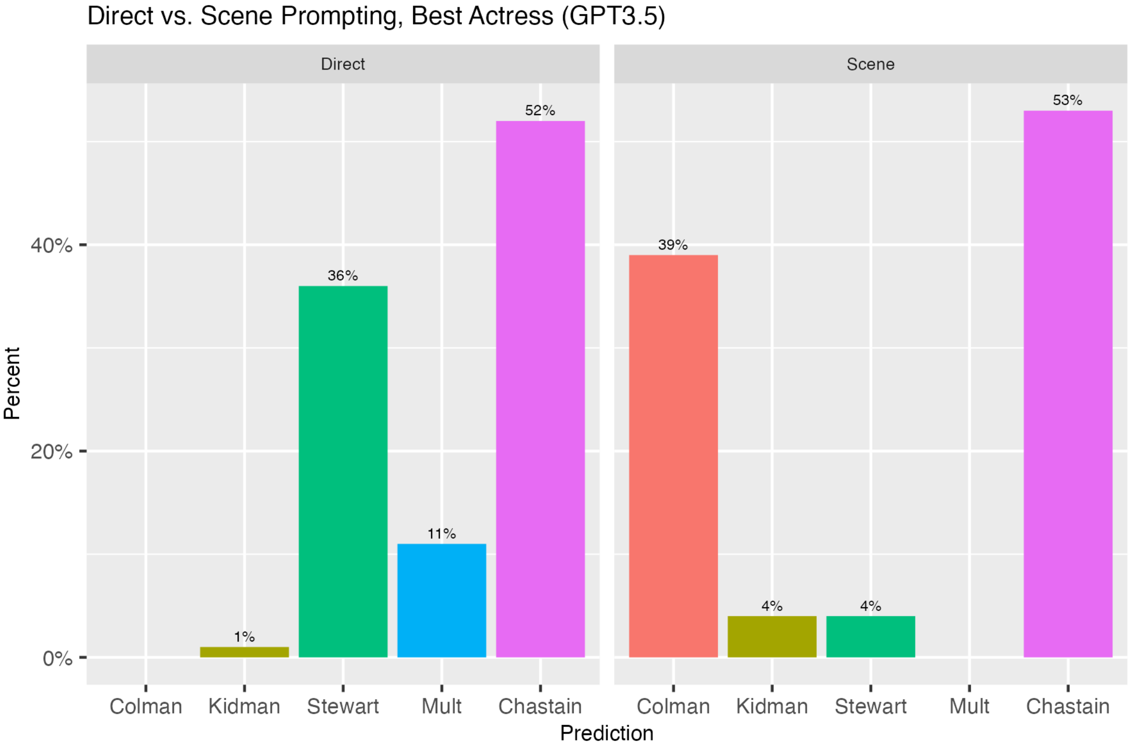

Figure 29: Predictions for Best Actress from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

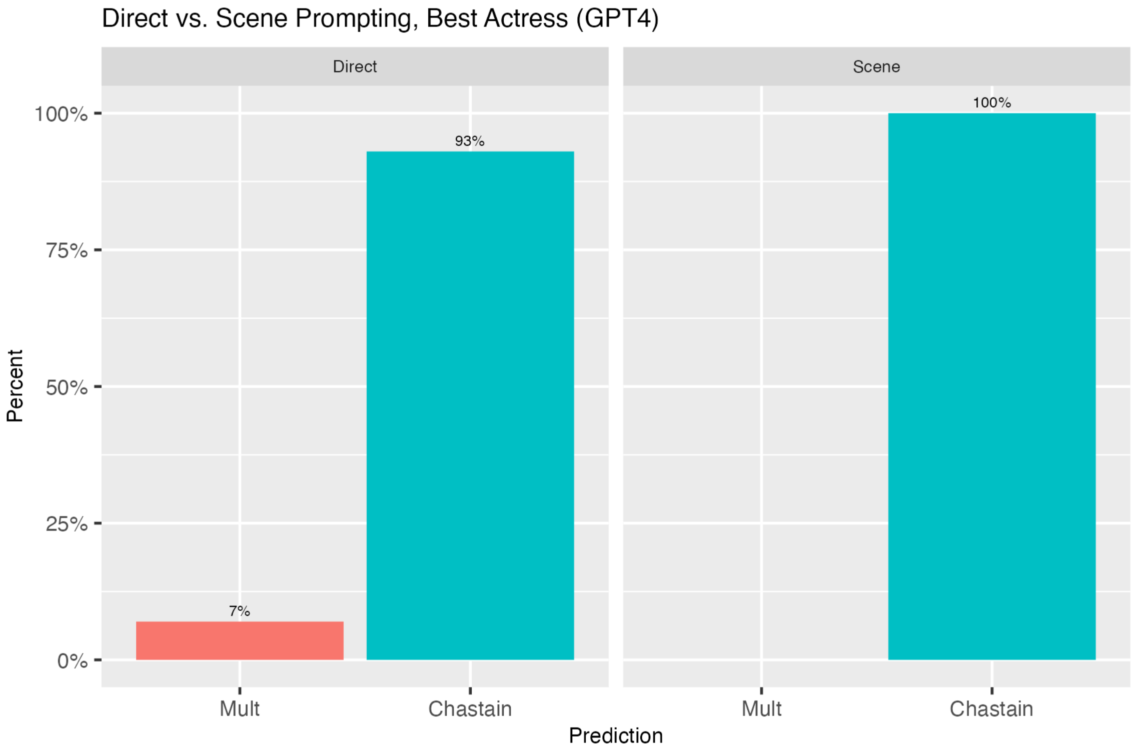

Figure 30: Predictions for Best Actress from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

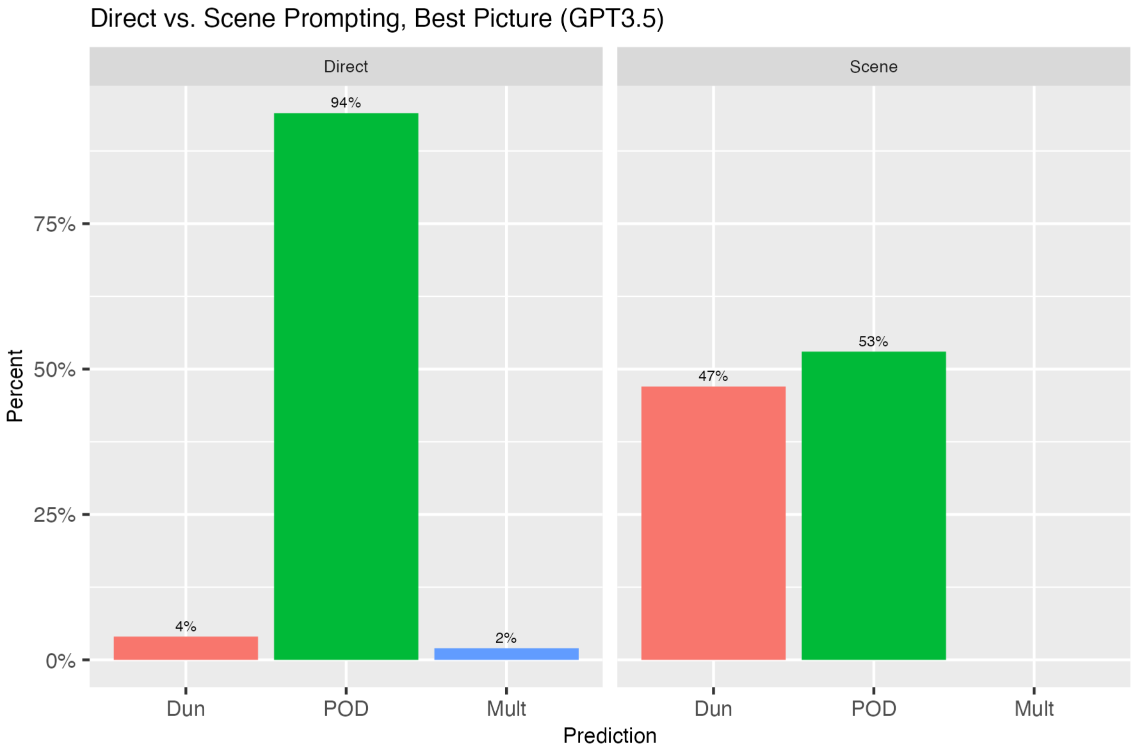

Figure 31: Predictions for Best Picture from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

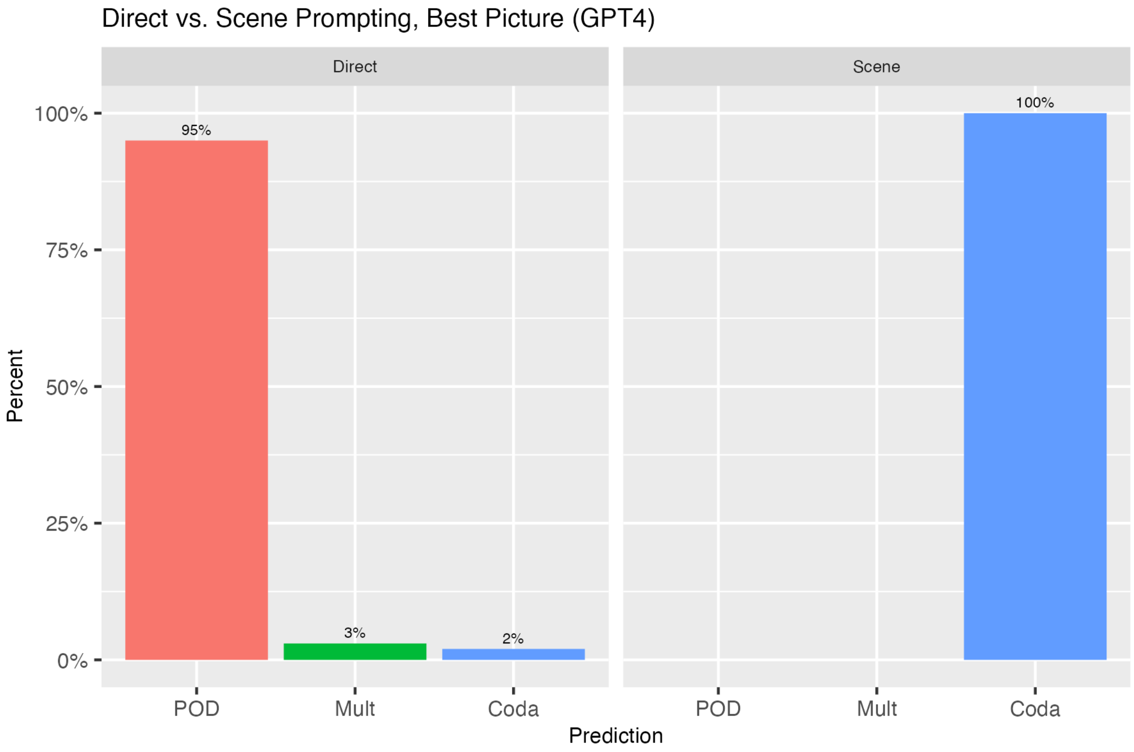

Figure 32: Predictions for Best Picture from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

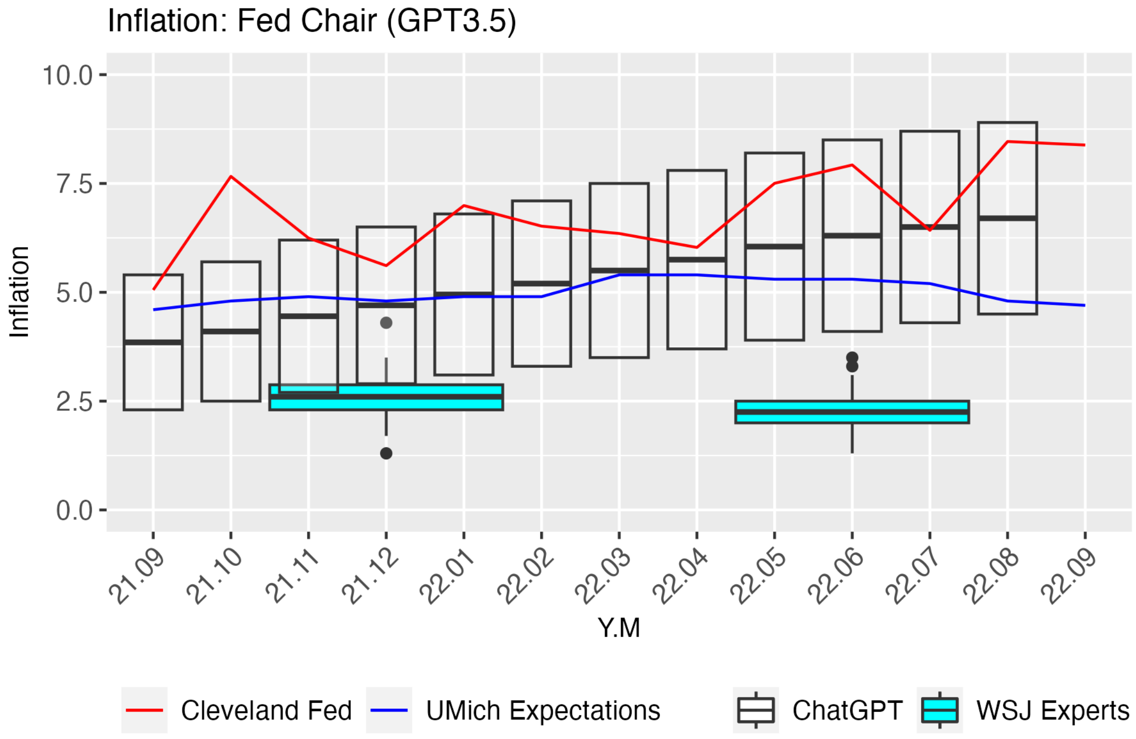

Figure 33: Inflation Predictions by Fed Chair from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

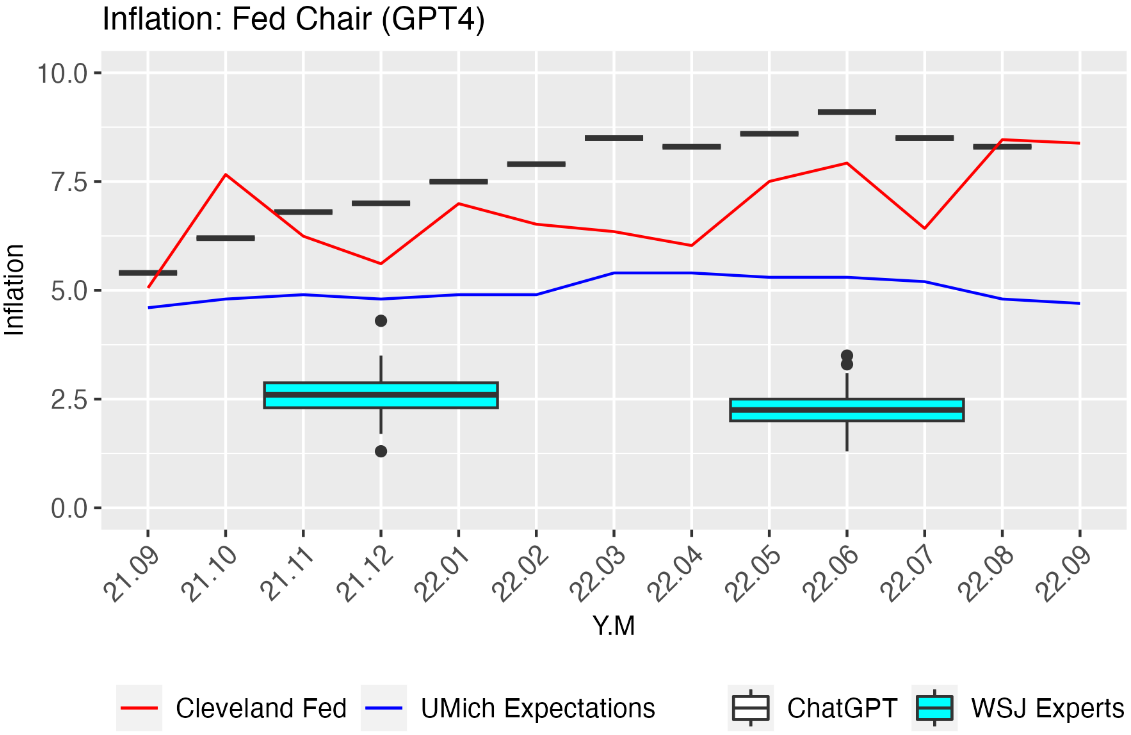

Figure 34: Prompts from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

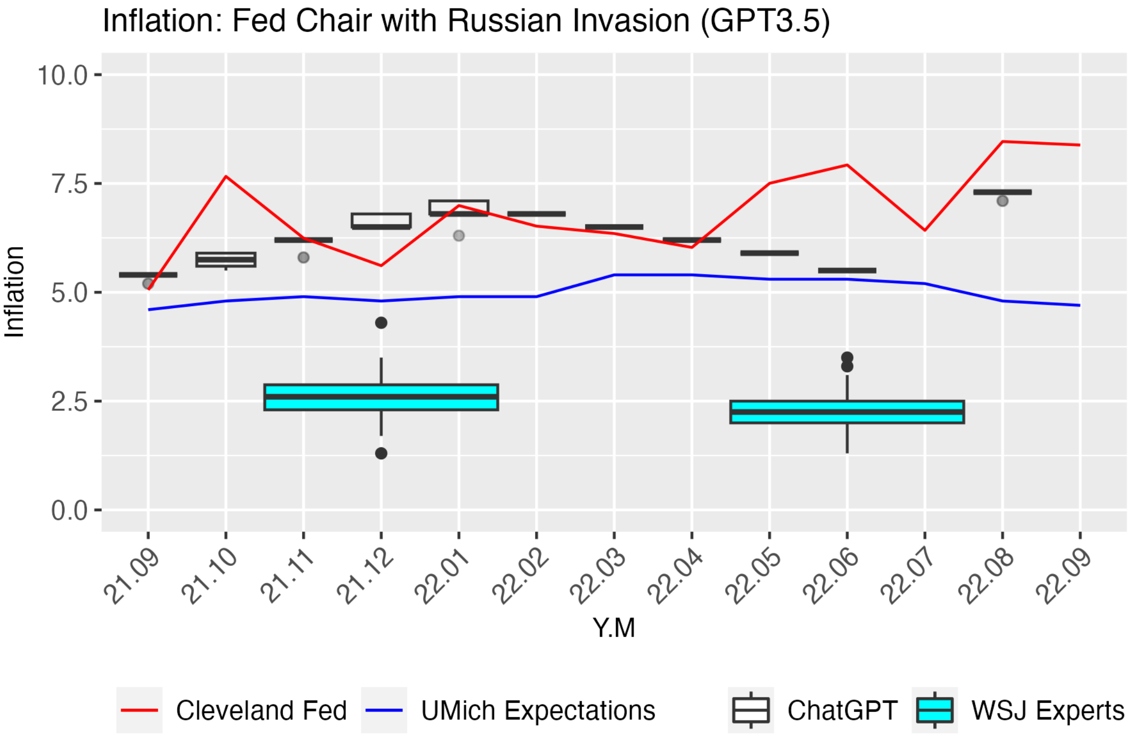

Figure 35: Inflation Predictions by Fed Chair with Russian Invasion Information from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

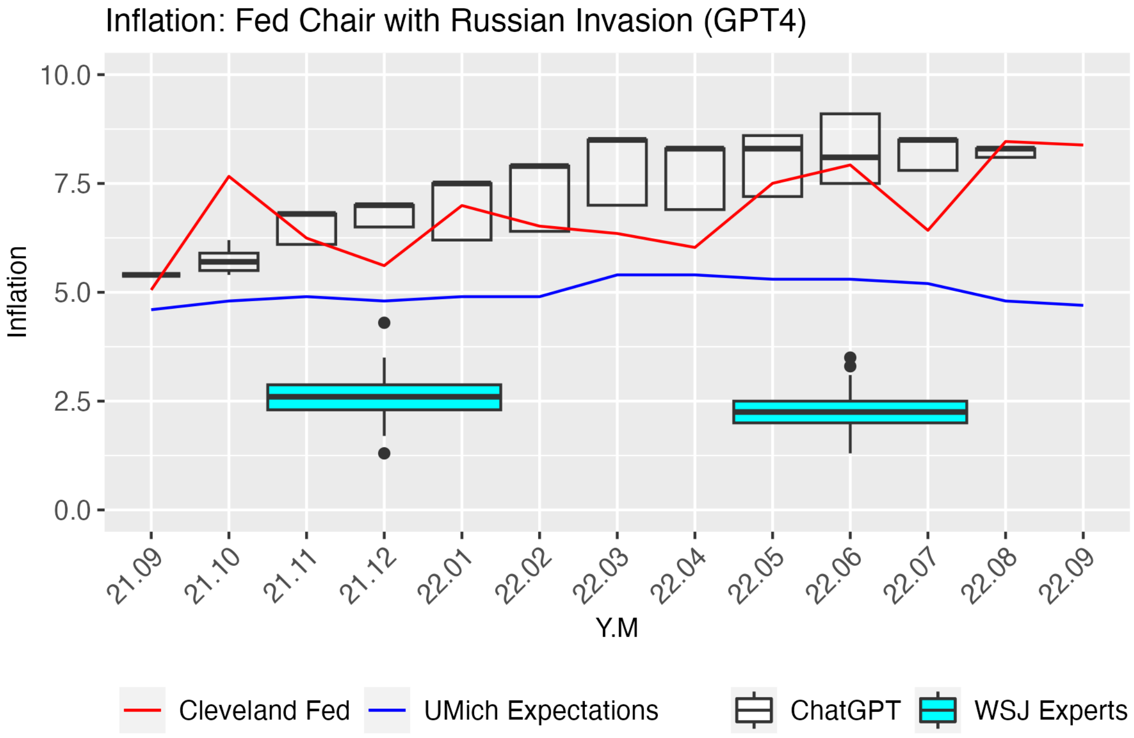

Figure 36: Inflation Predictions by Fed Chair with Russian Invasion Information from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

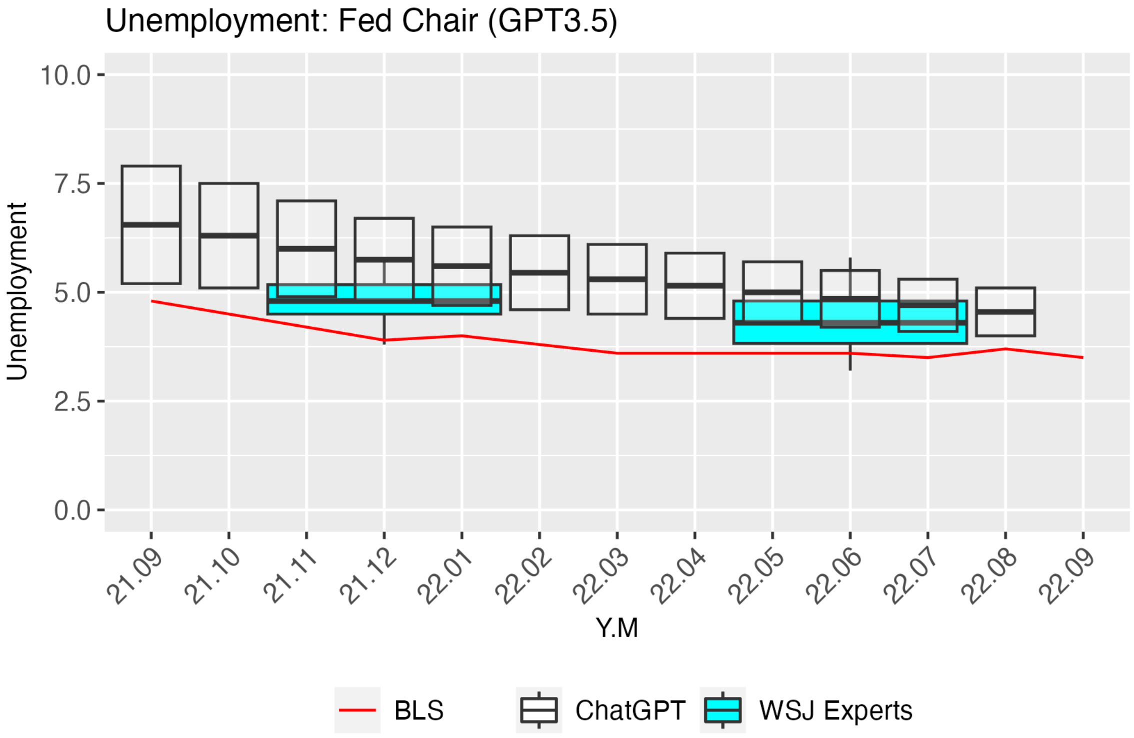

Figure 37: Unemployment Predictions by Fed Chair from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

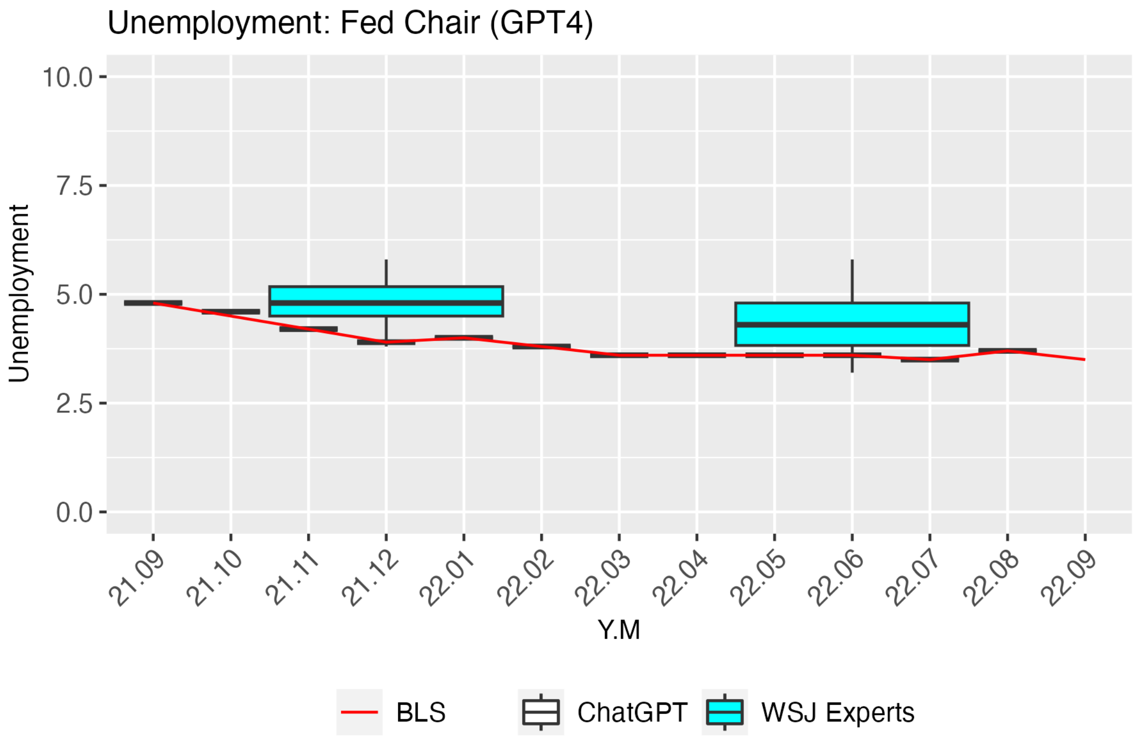

Figure 38: Unemployment Predictions by Fed Chair from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

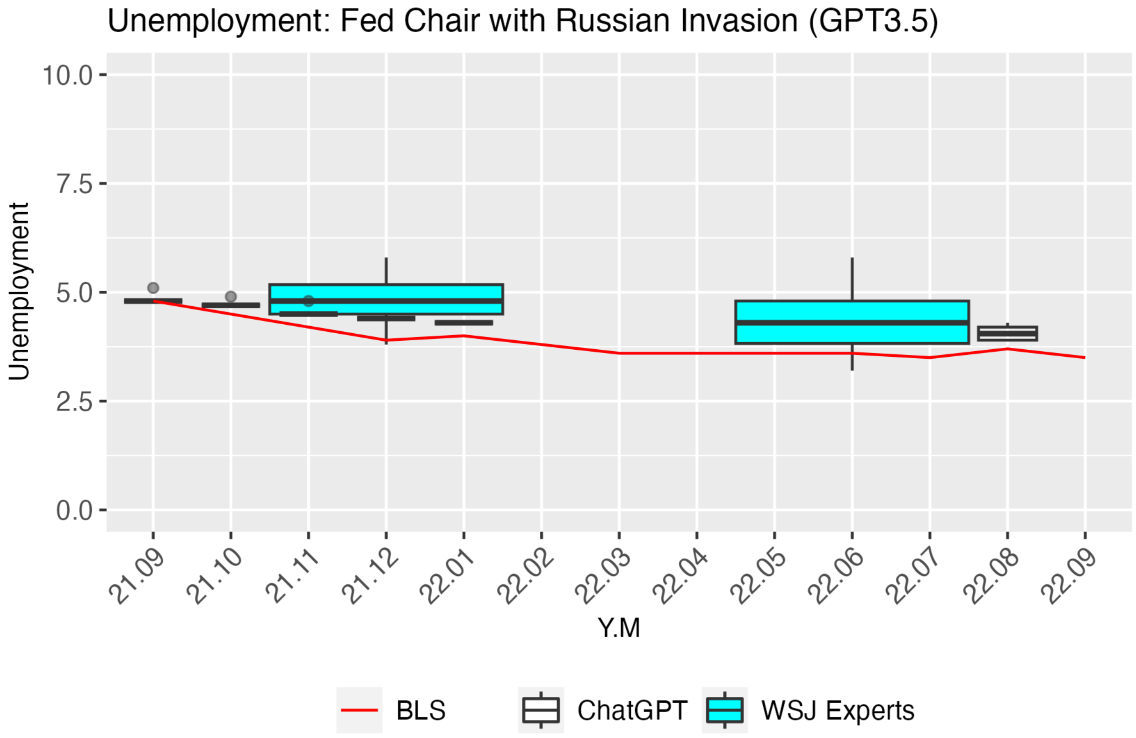

Figure 39: Unemployment Predictions by Fed Chair with Russian Invasion Information from May 2024 (ChatGPT-3.5 with January 2022 training data cutoff).

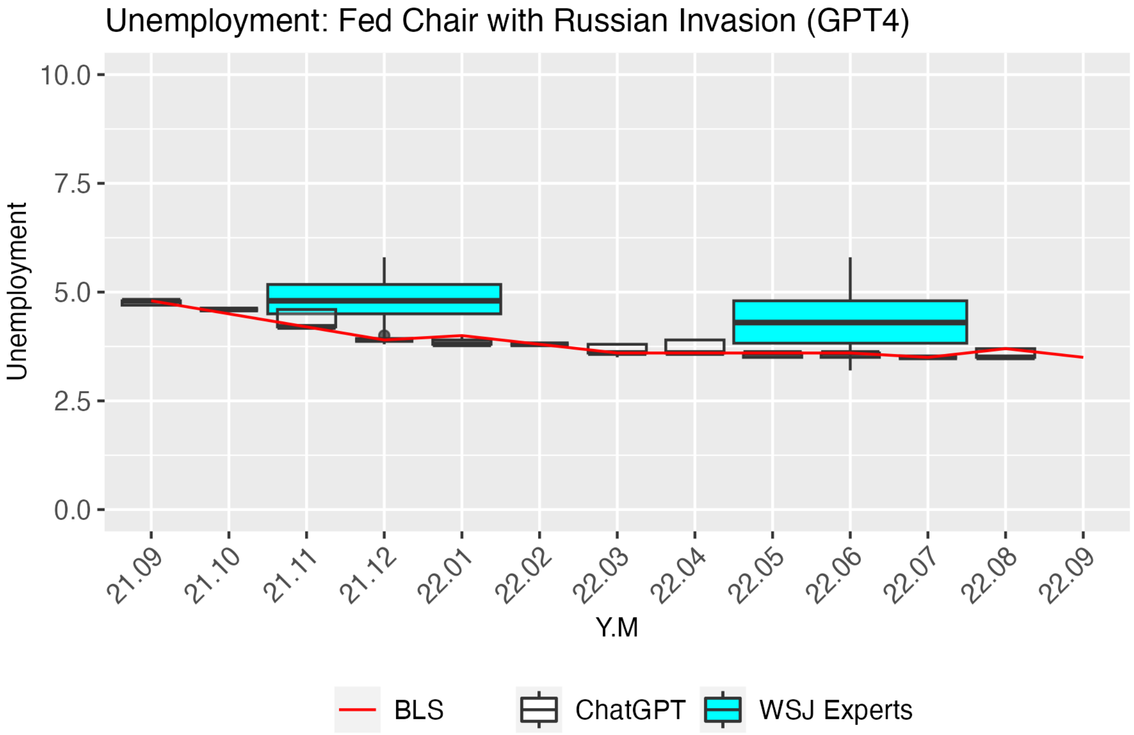

Figure 40: Unemployment Predictions by Fed Chair with Russian Invasion Information from May 2024 (ChatGPT-4 with December 2023 training data cutoff).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to write some of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, MA, USA.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using gpt for market research. Working Paper 23-062, Harvard Business School Marketing Unit.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? https://doi.org/10.48550/arXiv.2301.07543.

Hughes, A. (2023). ChatGPT: Everything you need to know about OpenAI's GPT-4 tool. *BBC Science Focus*, September 25, 2023.

Kim, A. G., Muhn, M., and Nikolaev, V. (2023). From transcripts to insights: Uncovering corporate risks using generative AI. Working Paper 2023-132, Becker Friedman Institute.

Levy, S. (2024). In defense of AI hallucinations. *Wired*, January 5, 2024.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. http://dx.doi.org/10.2139/ssrn.4412788.

Mehdi, Y. (2023). Bing at Microsoft Build 2023: Continuing the Transformation of Search, May 23, 2023. `https://blogs.bing.com/search/may_2023/Bing-at-Microsoft-Build-2023`. Accessed: April 07, 2024.

OpenAI (2024a). OpenAI–Documentation–Models. `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`. Accessed: April 06, 2024.

OpenAI (2024b). Usage Policies. `https://openai.com/policies/usage-policies`. Accessed: April 10, 2024.

Schaul, K., Chen, S. Y., and Tiku, N. (2023). Inside the secret list of websites that make AI like ChatGPT sound smart. *Washington Post*, April 19, 2023.

Silver, N. (2013). Oscar Predictions, Election-Style. `https://fivethirtyeight.com/features/oscar-predictions-election-style/`. Accessed: April 09, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.