Chess Banter: Does engaging with an opponent in online chess impact their performance?

Reed Evans, Jared Feldman, Jerry Gonzalez, Randy Louie

January 8, 2024

Contents

1 Abstract

Bullying in online games is a serious problem and online chess is no exception. The purpose of trash talking is to try to decrease opponent performance by distracting opponents into heightened emotions and less logical decision making. However, these tactics may actually backfire. In our study, opponents were randomly selected by the lichess.org matching algorithm and we randomly assigned these opponents to a control, placebo, or treatment group at time of matching. For those in the treatment group, they received non-harassing standardized comments about the inaccuracies in their play. In our follow-up experiment, we implemented the OpenAI API so that we could use ChatGPT 4.0 to engage with players more dynamically to more directly measure any effects of a conversation. Similar to our first experiment, we randomly assigned opponents into control and ChatGPT. Our analysis for both experiments did not produce any significant effects across multiple outcomes. Since chatting with an opponent in online gaming, particularly in chess, has mixed reviews from players, this lack of effect may be useful information to players who are opposed to chatting, knowing that it was not found to have any adverse effects on opponents.

2 Introduction

2.1 Theory

Psychological pressure can have an impact on sporting excellence[1]. One key mechanism through which individuals experience psychological pressure is by engaging in banter with their opponent. Banter, which is characterized as playful teasing and competitive remarks, has the potential to have both positive and negative influences on performance[2]. On the one hand, banter can distract individuals from performing their best by diverting focus away from the game. On the contrary, banter could lead to increased engagement and motivation, potentially resulting in one being in a heightened state which is commonly referred to as being "in the zone"[3].

In fact, according to a study by McDermott and Lachlan[4], the targets of trash talk were motivated to outperform their opponents and frequently did outperform them.

2.2 Concept Under Investigation

Research on the influence of conversation, banter, trash-talking, and bullying on an opponent's in-game performance is currently limited. Notably, trash-talking exhibits gender and sport-type variations, being more prevalent among men and in contact sports compared to non-contact sports, as evidenced by Kniffin and Palacio (2018)[5]. However, the impact of athlete anonymity, whether masked or not, remains inconclusive in existing studies. In the context of online gaming, where the online uninhibited behavior effect is commonly experienced, we aim to delve deeper into the effects in non-contact sports (in our case chess) and anonymity.

Recent studies highlight the normalization of online trash-talking[6], despite its potential negative repercussions for the recipient. Our investigation seeks to contribute valuable insights into banter's perceived impact on gameplay, with the ultimate goal of fostering a less toxic online gaming environment. By exploring the dynamics of online interactions, our research aspires to provide a nuanced understanding of how conversational elements may shape the gaming experience and influence player behaviors, contributing to a more positive and inclusive gaming community.

[1] Johnson and Taylor (2018) - More than Bullshit: Trash Talk and Other Psychological Tests of Sporting Excellence. Sports, Ethics and Philosophy, 14(1).

[2] Murphy and White (1995) - In the Zone: Transcendent Experience in Sports.

[3] Murphy and White (1995) - In the Zone: Transcendent Experience in Sports.

[4] McDermott and Lachlan (2021) - Emotional Manipulation and Task Distraction as Strategy: The Effects of Insulting Trash Talk on Motivation and Performance in a Competitive Setting

[5] Kniffin and Palacio (2018) - Trash-Talking and Trolling. Human Nature, 1–17.

[6] Beres et al., 2021 - Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15.

2.3 Hypothesis

Our null hypothesis is defined as follows: Engaging in online chess chat is unlikely to influence the playing accuracy of opponents.

We posit that encountering a text designed to provoke a challenge from an opponent may create a sense of perceived rivalry, ultimately enhancing their gameplay accuracy and performance.

3 Experiment Design

3.1 Overview

To test our hypothesis and answer the proposed research question, we turn to online chess platform lichess.org, which has an extensive, open source API The lichess.org API contains a board interaction protocol that was designed for third party programmers to conduct an entire chess game through the Lichess website. For our purposes, this provides the perfect vehicle to automate our study and analyses entirely through programming.

In order to provide the most equal and consistent chess games between all subjects, we utilize a chess engine to decide the moves returned to the opponent. One downside of utilizing a chess engine is they are mostly designed to play the best moves and are not particularly good at playing like a human. To overcome this, we will harness the power of the Maia neural network ("Maia") as detailed in "The human side of AI for chess" by Microsoft Research[7]. Maia was designed specifically to play human-like at certain skill levels and even to make mistakes and blunders, just like an average chess player.

To measure potential outcomes, we will use a between-subjects experiment design and compare our opponents move performance, game results (i.e. wins and losses), and opponent distraction measurables, such as average move time. Move performance will be measured by the Lichess Accuracy metric where an accuracy of 0% means the opponent never played a preferred move and 100% means the opponent played all the preferred moves of Stockfish, a very strong chess engine not to be confused with the Maia neural network. The comparisons will be made across our different groups consisting of a control group, a treatment group, and a placebo group for our preliminary experiment. And for our main experiment, we will compare across a control group and treatment group, utilizing ChatGPT for treatment.

3.2 Randomization Process and Recruitment

3.2.1 Lichess Randomization and Recruitment

Lichess uses a popular rating method called Glicko-2, which uses confidence intervals when calculating ratings for players[8]. When players first start on Lichess, the rating starts at 1500 +/- 1000. As players play more rated games, their rating changes and the confidence interval decreases.

Lichess's game matchmaking feature allows a player to find a game against an opponent who also wants to play at the same time. Players queue to play and it takes up to 30 seconds to find an opponent. The opponent's rating will be within about +/- 100 rating of the player's rating. Lichess uses a combination of a player's ratings and confidence intervals to pair with similar players so that it will be a fair game. Since we are playing at an average '1500' rating level, the opponent is essentially random to our experiment. Whichever players are matched against us will be included in our study.

Each of the bots that we used were identical, and trained to play at a 1500 Glicko-2 rating level. We relied on Lichess's randomization strategy to pair us with similarly ranked players. This allowed us to control for different player skill levels while conducting our experiment.

[7] Mcllroy-Young et al., 2020 - The human side of AI for chess.
[8] https://lichess.org/faq#ratings

3.2.2 Experimental Randomization

To randomize treatment group assignment, we employed the python function `randint()` from the `random` library. This function returns a random integer from a specified range. The range that was specified on the number of experimental groups. Our python script would then perform treatment based on the selected integer.

Following both Lichess randomization and experimental randomization, we found that there was no significant difference between the ratings in each of the experimental groups (visually represented in Figure 1 below). This was necessary to check because differently rated players may have different innate playing metrics.

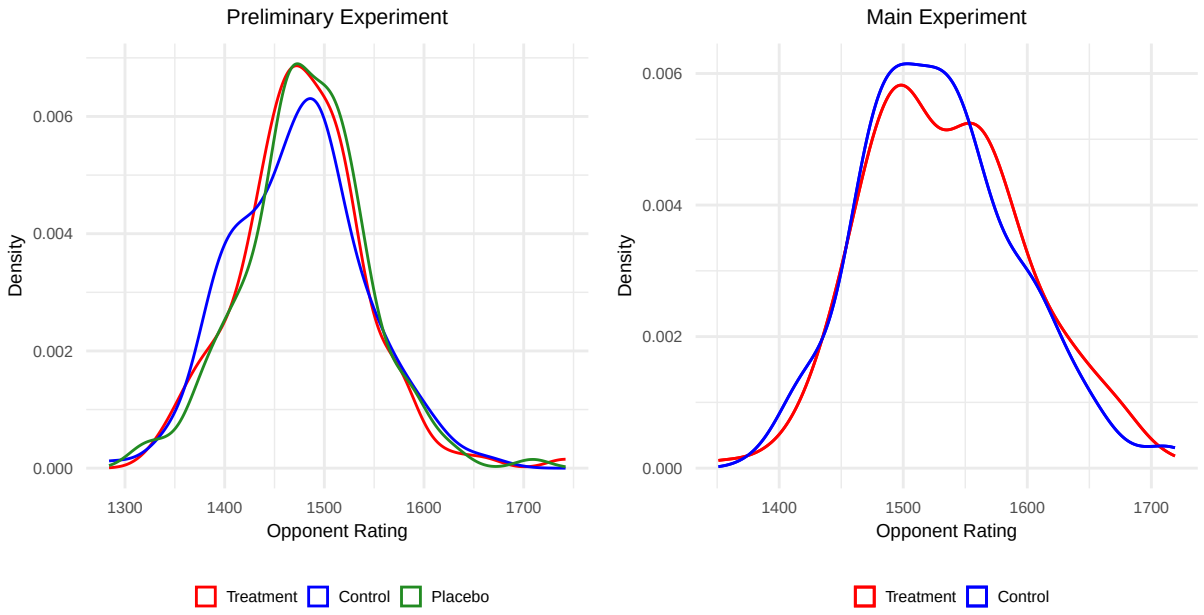

Figure 1: Opponent Rating Across Treatment Groups

3.3 Control and Treatment Variants

The experiment is a between subjects experiment that consists of playing 10-minute rapid ranked games to observe the effects of a chat (treatment) on the subjects. The design consists of a preliminary experiment using predefined phrases from a script followed by the main experiment which uses ChatGPT to engage the subject with questions.

3.3.1 Preliminary Experiment

The preliminary experiment presents three conditions.

1. The control group receives no chat messages during game play.
2. The placebo group is administered chats from a fixed script that states random facts during pre-defined moves throughout game play. For example:
    - Move 3 - hello!
    - Move 6 - Bananas are berries

- Move 9 - I like swimming

3. The treatment group is administered chats also from a fixed script but the chats are designed to create a sense of friendly rivalry. For example:
    - Move 6 - I'm going to take your queen in 4 moves.
    - Move 18 - Things aren't looking good.

3.3.2 Main Experiment

The main experiment uses the ChatGPT API for the chat treatment and has the following two conditions:

1. The control group receives no chat messages during game play.
2. The treatment (ChatGPT group) begins with a fixed start script to initiate conversations with the subject, and once the subject responds to the fixed script, the ChatGPT AI (chat bot) takes control of the chats.

The chat bot has been prompted to keep responses short, to one sentence, and relevant. The chat bot is also given instructions to be lighthearted, not rude, not use profanity, empathetic, polite, and generally tries to make people smile. To give the chat bot some depth, we also prompted the chat bot to be Magnus Carlson's young cousin living in Boston who is really fond of pineapple on pizza. A more in-depth description of the prompt used can be found in Appendix B.

3.4 Power Analysis

We used a conservative estimate of 5-15% effect size for our power analysis. In related work, it was found that trash talking had an indirect effect on competitive performance through creation of a perceived rivalry between the players [9]. Further, this research showed that there was an effect of b=.32, with a 95% confidence interval of 0.02, 0.87. In other words, this research found that trash talking resulted in the opponent performing better.

In our analysis, as seen in Figure 2, we determined that we would need roughly 150 samples of each treatment type to detect an effect size of 7.5%, and less samples for an effect size of 10%. Thus, we targeted 100–150 samples for both of our experiments.

For both our Preliminary Experiment and Main Experiment, we were able to achieve our targeted sample size to have at least 80% power. We had at least 144 samples in the Preliminary Experiment and at least 195 for our Main Experiment.

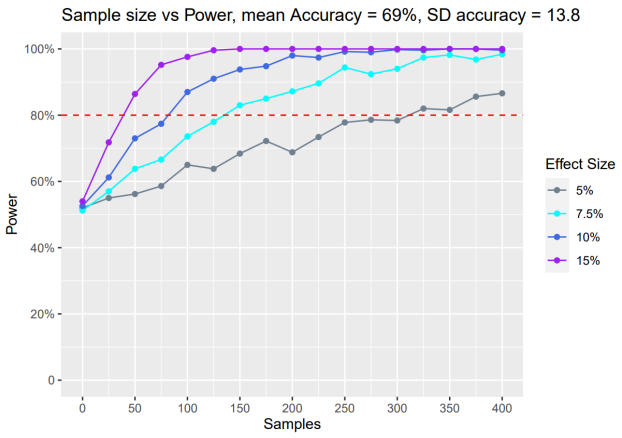

Figure 2: Power Analysis

4 Analysis

4.1 Intent to Treat (ITT) Effect

Our first step in evaluating the outcomes of our experiments is to estimate the Intent to Treat (ITT). The ITT measures the outcomes of our experiment based on treatment assignment, without considering compliance rate. The ITT is defined as:

$$ITT = E[Y_i(z = 1)] - E[Y_i(z = 0)]$$

where z signifies the assignment of the subject.

Our main outcome variable of interest was Accuracy, as described in Section 3.1. Figure 3 shows the distribution of Accuracy ITT across treatment groups for each experiment, while Tables 1 and 2 show the ITT for Accuracy and five other outcomes of interest. The tables containing all ten outcomes can be found in Appendix C.

As seen in Table and Table 2, we found no statistical significance at the .05 level in either experiment.

[9] Yip et al., 2018

Figure 3: Opponent Accuracy Across Treatment Groups

Table 1: Preliminary Experiment: ITT

	Dependent variable:					
	Maia Win	Opp Resigns	Opp Acc	Opp Blunders	Opp Mistakes	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.008	0.086*	0.253	0.163	0.256*	0.677
	(0.056)	(0.051)	(1.509)	(0.214)	(0.147)	(0.652)
Constant	0.443***	0.261***	75.920***	2.108***	1.153***	7.795***
	(0.038)	(0.035)	(1.012)	(0.143)	(0.099)	(0.437)
Observations	320	320	320	320	320	320
Adjusted R^2	−0.003	0.006	−0.003	−0.001	0.006	0.0002
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2: Main Experiment: ITT

	Dependent variable:					
	Maia Win	Opp Resigns	Opp Acc	Opp Blunders	Opp Mistakes	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.070	0.048	−0.618	−0.135	0.127	2.045*
	(0.050)	(0.047)	(1.329)	(0.185)	(0.131)	(1.097)
Constant	0.492***	0.323***	74.979***	2.421***	1.292***	8.164***
	(0.036)	(0.034)	(0.957)	(0.133)	(0.094)	(0.790)
Observations	405	405	405	405	405	405
Adjusted R^2	0.002	0.0001	−0.002	−0.001	−0.0001	0.006

Note: *p<0.1; **p<0.05; ***p<0.01

4.2 Labeling compliers and non-compliers

In this experiment we defined a complier as someone who sent a message in the chat room when we issued a dose of treatment. This allowed us to confirm that the opponent received the treatment. We then defined a non-complier as an opponent who did not send a message. We theorize that there may be multiple reasons why an opponent would not engage in chat with us.

1. Opponent has manually 'muted' chat and they do not see our messages.
2. Opponent does not speak english (our language used in treatment).
3. Opponent does not see the chat due to screen resolution or device used.
4. Opponent does not want to engage in chat.

An issue with our complier labeling is that opponents can mute at anytime during the game. For example they may see our initial message, then greet us in the beginning, and mute chat immediately after. This would mean they only received 1 out of the 6 messages that were planned but are counted as a complier. This is a limitation of our analysis, because the website does not allow us to see when an opponent mutes.

4.3 Complier Average Causal Effect (CACE)

While the ITT allows us to get a glimpse into potential treatment effects, it does not consider compliers and non-compliers. In Figure 4, we can see not everyone received their targeted dose of treatment, which we now take into account in calculating the Complier Average Causal Effect ($CACE$).

The $CACE$ is calculated using the following formula:

$$CACE = ITT/ITT_d$$

where ITT_d is the proportion of compliers in the treatment group (the "take-up rate"). In our case, we can see in Figure 4 that the take-up rates are approximately 30% and 33% for the preliminary and main experiments, respectively.

Now we will run the same regression analysis, but this time we will adjust our effects by predicting outcomes based on the CACE methods described above. In Table 3, for our preliminary experiment we see a slight effect on Opponent resignations and small uptick in Opponent Mistakes with a p-value < 0.1. In Table 4 for our Main Experiment, we see no statistical significance for outcomes except on Opponent Average Move Time being up more than 5.8 seconds over those in the control group.

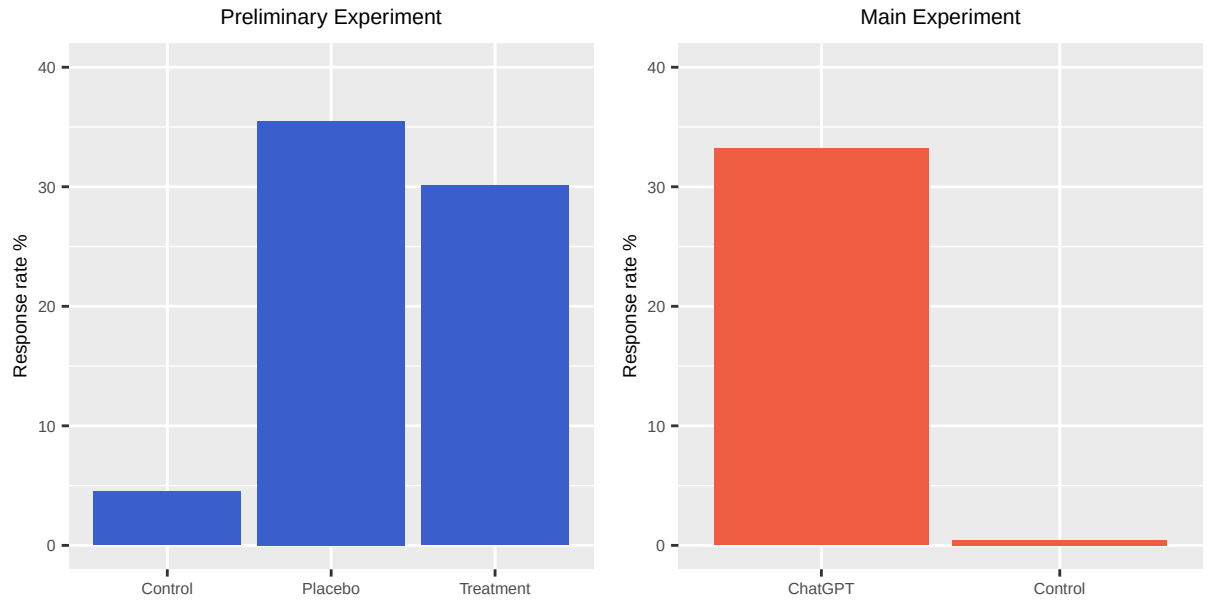

Figure 4: Response Rates by Experiment

Table 3: Preliminary Experiment: CACE

	Dependent variable:					
	Maia Win	Opp Resigns	Opp Acc	Opp Blunders	Opp Mistakes	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.032	0.330*	0.973	0.626	0.985*	2.603
	(0.215)	(0.198)	(5.801)	(0.821)	(0.567)	(2.506)
Constant	0.442***	0.246***	75.876***	2.079***	1.109***	7.677***
	(0.045)	(0.041)	(1.205)	(0.171)	(0.118)	(0.521)
Observations	320	320	320	320	320	320
Adjusted R^2	−0.003	0.006	−0.003	−0.001	0.006	0.0002

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Main Experiment: CACE

	Dependent variable:					
	Maia Win	Opp Resigns	Opp Acc	Opp Blunders	Opp Mistakes	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.200	0.139	−1.777	−0.388	0.365	5.882*
	(0.143)	(0.136)	(3.824)	(0.533)	(0.376)	(3.155)
Constant	0.492***	0.323***	74.979***	2.421***	1.292***	8.164***
	(0.036)	(0.034)	(0.957)	(0.133)	(0.094)	(0.790)
Observations	405	405	405	405	405	405
Adjusted R^2	0.002	0.0001	−0.002	−0.001	−0.0001	0.006

Note: *p<0.1; **p<0.05; ***p<0.01

4.4 Benjamini-Hochberg (BH) Procedure

While some test results in the tables above show statistical significance at the 0.1 p-value threshold, we need to adjust due to the number of tests that we conducted. Since we performed ten statistical tests, we have an increased chance of finding a p-value less than or equal to 0.1 by chance, compared to running one or two tests. The Benjamini-Hochberg (BH) procedure is one way to try to control for the false discovery rate (FDR) between all of our tests. Unlike the False Positive Rate, which is used in the Bonferroni method, the FDR is the expected proportion of false positives among *all* positives which rejected the null hypothesis.

In this method, we take the p-values from each of our $CACE$ outcomes and compare them to their associated $BH_{critical}$ value. The $BH_{critical}$ value is defined as:

$$BH_{critical} = (i/m) * Q$$

where i is the rank of the p-value (i.e., the smallest value has $i = 1$, etc.), m is the number of p-values (i.e., ten), and Q is the allowable FDR (i.e., 0.05). In this test, if a p-value is less than its $BH_{critical}$ value, then the test is significant. Otherwise, the p-values that we found are likely the result of a false discovery.

In both our Preliminary Experiment and Main Experiment, there were zero p-values less than their associated $BH_{critical}$ value, and thus we fail to reject the null hypothesis under both experiment conditions.

4.5 Summary of Results

Given the results and analysis of our experimentation, we fail to reject the null hypothesis that online chess player's performance is directly impacted when analyzing performance measures such as move accuracy, blunders, & mistakes, as well as game outcomes in regards to wins and losses. Notably, we did, expectedly, observe an effect on players who were treated to conversations with ChatGPT where those players' average move time increased by 5.9 seconds which over the course of a (average for our main experiment) 58 move game could add up. However, this effect is not statistically significant with a p-value of 0.063 and given the Benjamini-Hochberg (BH) Procedure, the significance is further diluted.

5 Conclusion

In conclusion, our experiments aimed to address the prevalent issue of bullying in online chess games, particularly through the means of trash-talking with the intention of impairing opponents' performance.

Contrary to the expected outcomes, our study did not reveal any significant effects across various outcomes in both the preliminary experiment involving standardized comments and the main experiment utilizing the Open AI API for dynamic player engagement.

By randomly assigning opponents to control, placebo, or treatment groups, we sought to understand the impact of non-harassing standardized comments on gameplay accuracy. Additionally, the implementation of ChatGPT 4.0 allowed us to explore more dynamic interactions in the follow-up experiment. Despite these efforts, our analyses did not identify any notable effects on player performance or behavior.

6 Appendices

6.1 Appendix A: Ethical and Privacy Considerations

Lichess, the platform that we used for conducting our experiment, has clear language around the use of bots, chess engines, and fair play. By employing the Microsoft maia chess bot ("Maia")[10] for gameplay, and the OpenAI API ("ChatGPT") for conversation, there was more than one article from their terms of service that we did not follow. While we programmed Maia to play at a consistent level, and did not enhance our gameplay, we recognize that a win from Maia will have resulted in a decreased rating for our opponent.

Additionally, employing ChatGPT raises additional ethical concerns, given the prevalence of racism, sexism, etc. found across modern chatbots, ChatGPT included[11]. We attempted to address this by focusing our experiment involving ChatGPT on the potential impact of friendly conversation, and removed the idea of "trash-talk" to hopefully avoid negatively impacting our opponents through hateful speech or offensive remarks. Our prompt instructions for ChatGPT can be found in the Appendix B.

Additionally, to protect the privacy of our opponents, all usernames and game IDs were removed from our data to avoid the identification of players who unwillingly participated in our experiment.

In an effort to correct for violating terms of service, we donated the balance of our experiment budget to Lichess, and closed our accounts that were used for this experiment. Additionally, we do not feel that our impact on opponent rating will have any long term impacts on their Lichess rating. Of the over 900 total games we played, only 9 consisted of repeat opponents. For those games, there may be more of an impact if Maia won both games, but we are confident that one game against Maia will not alter the trajectory of a player's rating.

6.2 Appendix B: Sample ChatGPT prompt

I am running an experiment on online chess players and whether or not chatting and banter in the comments while playing the game effect the other players performance.

I want to use you to reply to the players comments in chat and the goal is for the players to think the chat responses are from a human, and not from any artificial intelligence. Here are the rules for you to follow:

1. You are to never refer to yourself as being artificial intelligence. As far as the players are concerned, you are a human.
2. You keep your responses SHORT. One sentence responses only. You will ask questions of the player occasionally. You are curious about the other player.
3. You will not be rude and will not use profanity. You are light hearted.
4. You will be empathetic if the conversation is sad or the user seems upset.
5. Your job is to engage, be polite, and try to make people smile
6. You never ask stereotypical questions about a person's country. You are polite and wholesome.
7. You use poor punctuation and spelling. Here is your backstory that you will use in your responses:

You are Magnus Carlsen's young cousin living in Boston. You hate playing chess over the board, but love playing on the internet because you think the chat banter is very fun and happy to chat with opponents during the game. You are very keen to point out if opponents make good moves. You are also confident. You like to crack one liners, light hearted jokes every now and then. You think pineapple belongs on pizza. You like cats but dislike dogs. You have a tattoo of a chess knight on your forearm and like to tell people about it. You refer to the knight as horsey. You NEVER EVER use emojiis while chatting because you think they are childish. Think step by step

[10] Mcllroy-Young et al., 2020 - The human side of AI for chess.
[11] Peng et al., 2023 - Why your chatbot's so racist

6.3 Appendix C: CACE of All 10 Models

Results — Preliminary Experiment: Placebo

	Dependent variable:									
	Outcome Variables									
	Maia Win	Maia Checkmate	Opp Checkmate	Opp Resigns	Draw	Opp Acc	Opp Blunders	Opp Mistakes	Opp ACPL	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Placebo	0.054	-0.094	-0.101	0.244	-0.028	3.267	-0.087	0.750	-0.781	0.595
	(0.199)	(0.134)	(0.199)	(0.182)	(0.109)	(5.485)	(0.690)	(0.516)	(13.319)	(2.210)
Constant	0.441***	0.146***	0.465***	0.250***	0.087***	75.772***	2.112***	1.119***	62.962***	7.768***
	(0.044)	(0.031)	(0.044)	(0.039)	(0.025)	(1.250)	(0.165)	(0.108)	(2.869)	(0.555)

Note: *p<0.1; **p<0.05; ***p<0.01

Figure 5: Results - Preliminary Experiment: Placebo

Results — Preliminary Experiment: Treatment

	Dependent variable:									
	Outcome Variables									
	Maia Wins	Maia Checkmate	Opp Checkmate	Opp Resigns	Draw	Opp Acc	Opp Blunders	Opp Mistakes	Opp ACPL	Opp Avg Move Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	0.032	-0.226*	-0.087	0.330*	0.046	0.973	0.626	0.985*	3.701	2.603
	(0.216)	(0.135)	(0.216)	(0.200)	(0.126)	(5.757)	(0.828)	(0.579)	(14.906)	(2.466)
Constant	0.442***	0.152***	0.464***	0.246***	0.083***	75.876***	2.079***	1.109***	62.758***	7.677***
	(0.045)	(0.031)	(0.045)	(0.040)	(0.025)	(1.262)	(0.167)	(0.109)	(2.907)	(0.562)

Note: *p<0.1; **p<0.05; ***p<0.01

Figure 6: Results - Preliminary Experiment: Treatment

Results — Main Experiment: ChatGPT

	Dependent variable:									
	Outcome Variables									
	Maia Win	Maia Checkmate	Opp Checkmate	Opp Resigns	Draw	Opp Acc	Opp Blunders	Opp Mistakes	Opp ACPL	Opp Avg Mv Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ChatGPT	0.200	0.056	-0.256*	0.139	0.045	-1.777	-0.388	0.365	2.955	5.882*
	(0.143)	(0.099)	(0.135)	(0.137)	(0.094)	(3.833)	(0.536)	(0.376)	(9.957)	(3.071)
Constant	0.492***	0.128***	0.379***	0.323***	0.113***	74.979***	2.421***	1.292***	66.954***	8.164***
	(0.036)	(0.024)	(0.035)	(0.034)	(0.023)	(0.960)	(0.142)	(0.093)	(2.523)	(0.353)
Note:										*p<0.1; **p<0.05; ***p<0.01

Figure 7: Results - Main Experiment: ChatGPT