Natural Language Processing Final Project Report Figures Spring 2024

Sheyda Nazarian, Yingyin Yu, Jedrick Regala Zablan

DATA 200S

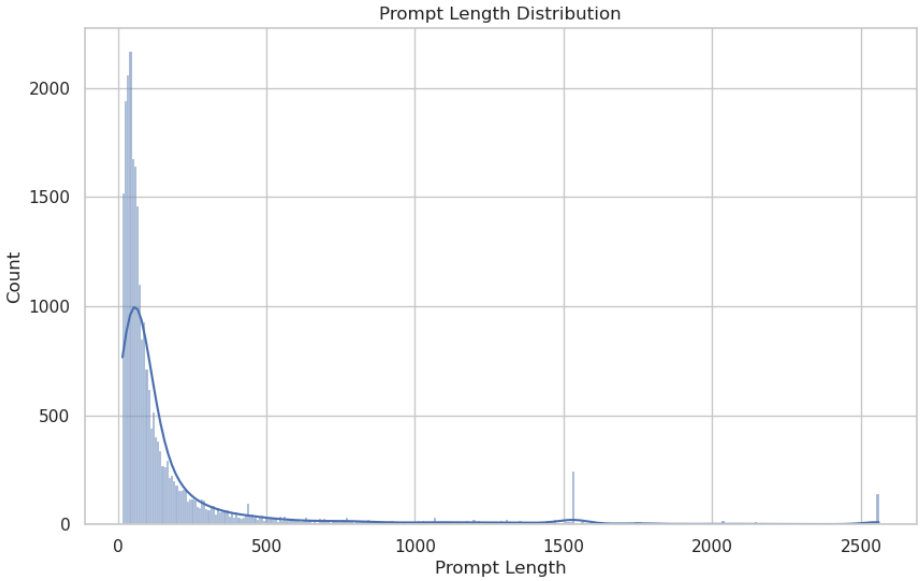

Figure 1: Prompt Length Distribution Before Filtering Outliers

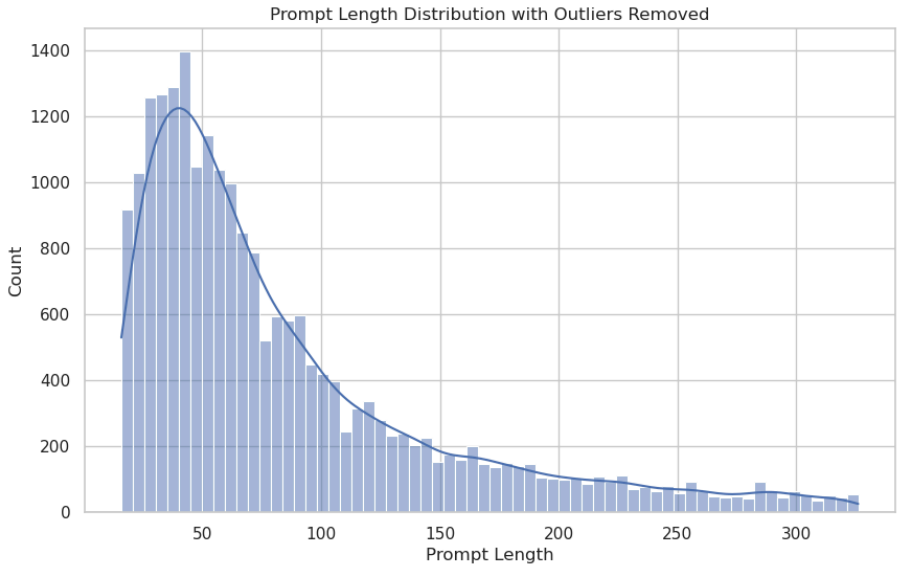

Figure 2: Prompt Length Distribution After Filtering Outliers

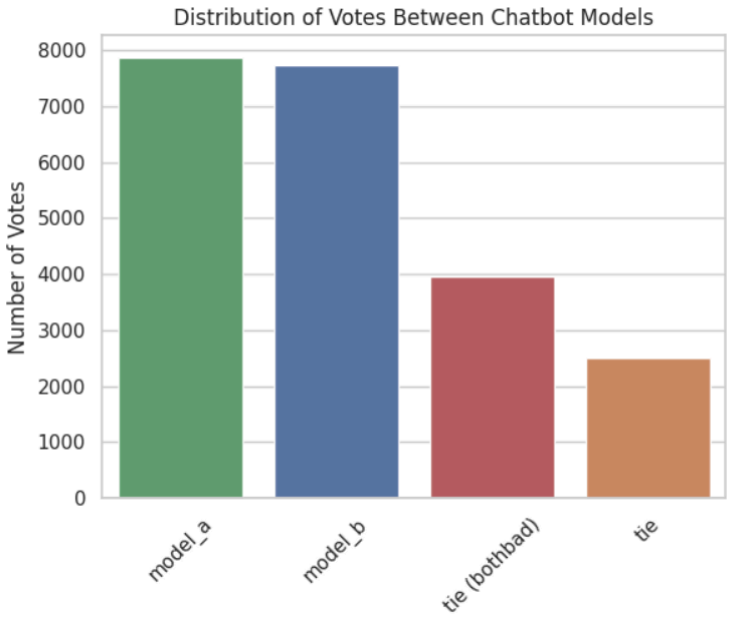

Figure 3: Distribution of Votes Between Chatbot Models

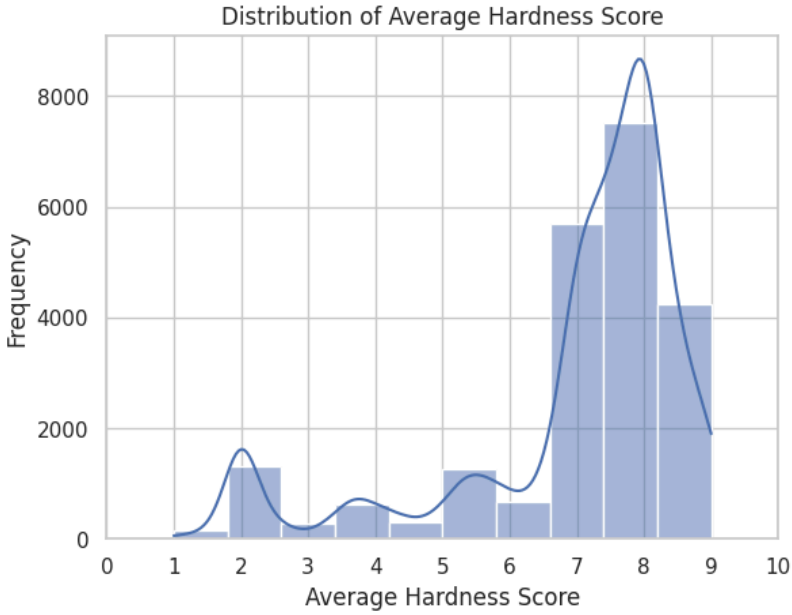

Figure 4: Distribution of Average Hardness Score Values

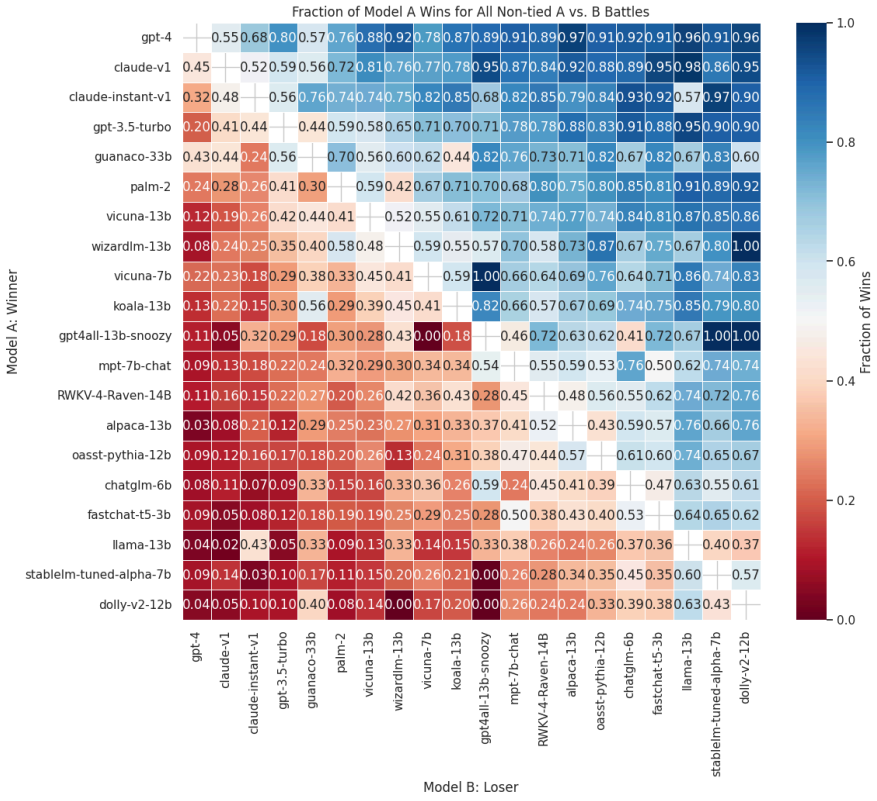

Figure 5: Heat Map of Model A Wins for Non-Tied Duels

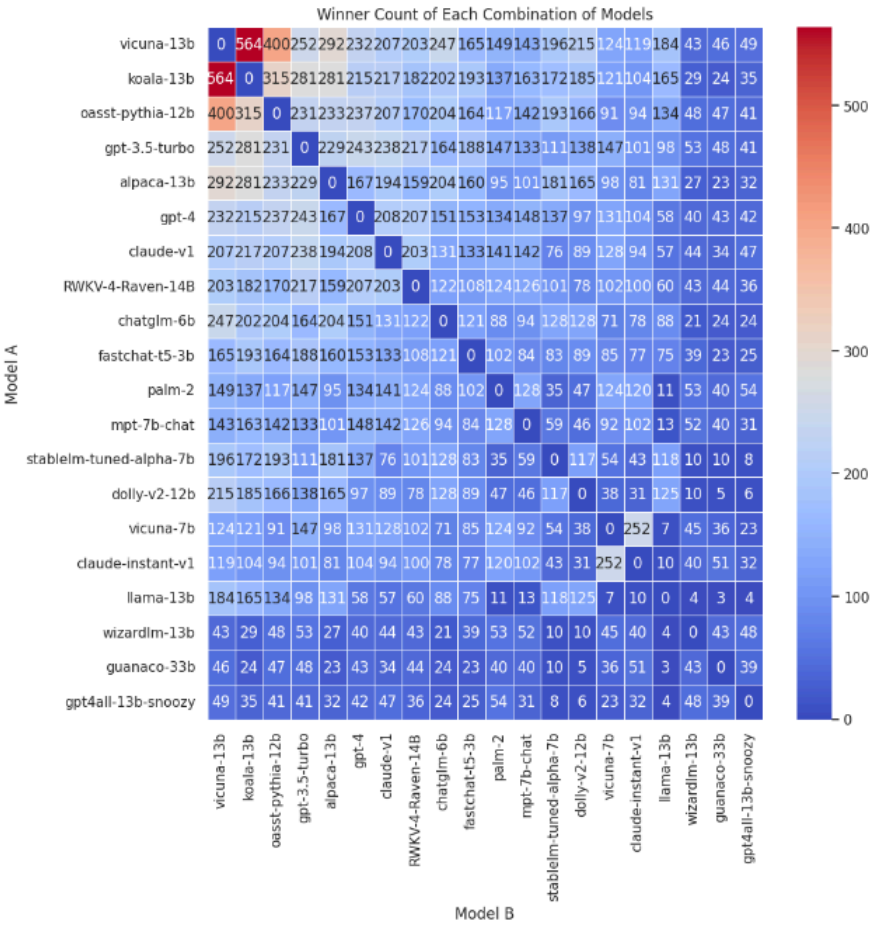

Figure 6: Heat Map of Winner Count for Specific Chatbot Models

	Model	Elo rating
1	gpt-4	1146.470637
2	claude-v1	1128.968883
3	claude-instant-v1	1100.264798
4	gpt-3.5-turbo	1046.514677
5	guanaco-33b	1037.662585
6	palm-2	1008.338058
7	vicuna-13b	985.755986
8	vicuna-7b	985.656296
9	wizardlm-13b	980.489436
10	koala-13b	968.981895
11	RWKV-4-Raven-14B	908.201345
12	mpt-7b-chat	905.415857
13	gpt4all-13b-snoozy	899.827806
14	chatglm-6b	894.155997
15	alpaca-13b	882.776758
16	fastchat-t5-3b	848.375863
17	oasst-pythia-12b	847.978710
18	stablelm-tuned-alpha-7b	836.120650
19	dolly-v2-12b	801.005663
20	llama-13b	800.000000

Figure 7: Top Twenty Chatbot Models Based on Elo Rating

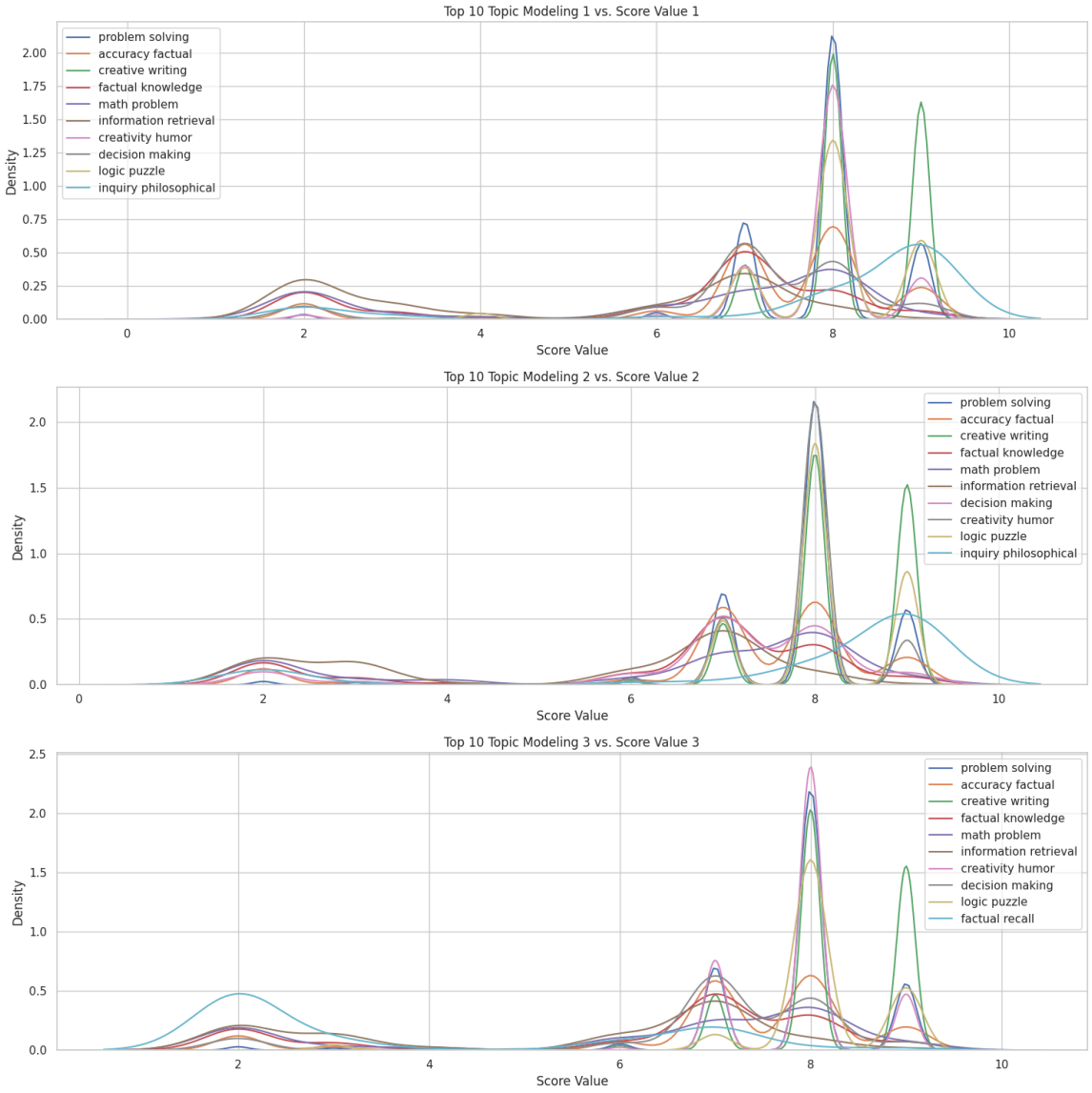

Figure 8: Three KDE Graphs Showing Density of Topics Across Score Values

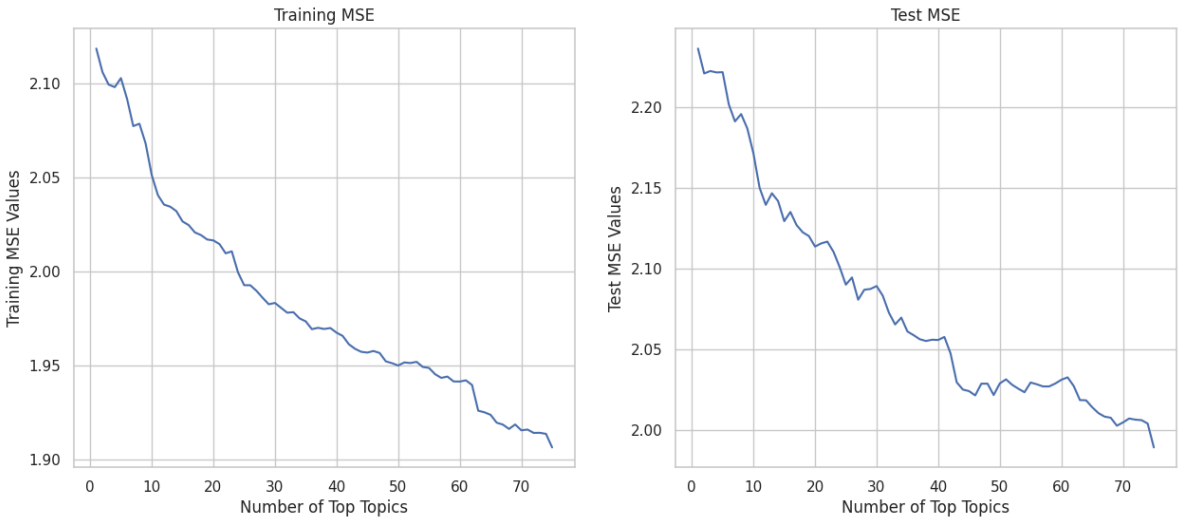

Figure 9: Line Graph of MSE Versus Number of Topics

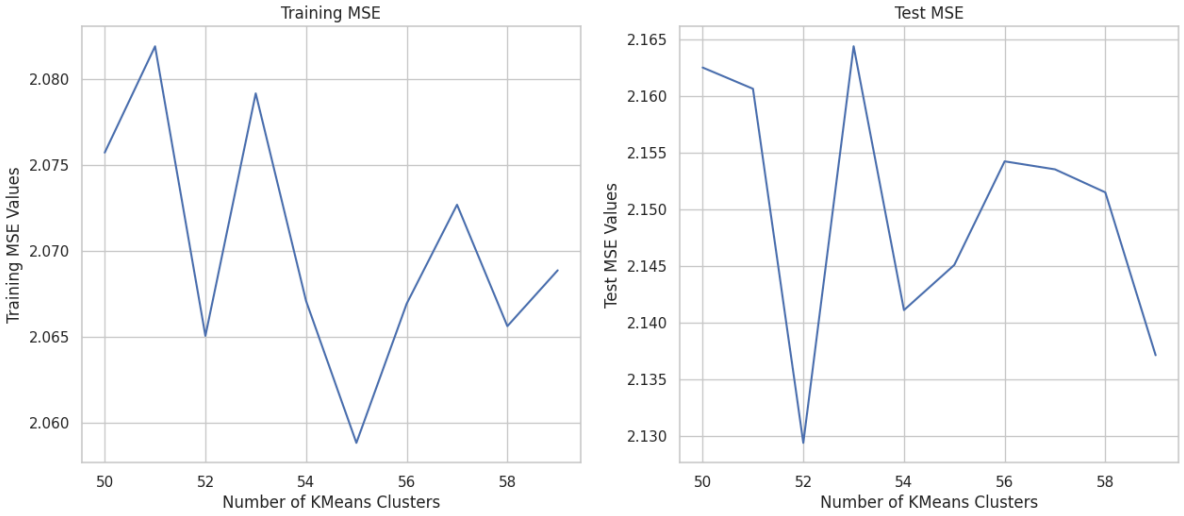

Figure 10: Line Graph of MSE Versus Number of KMeans Clusters

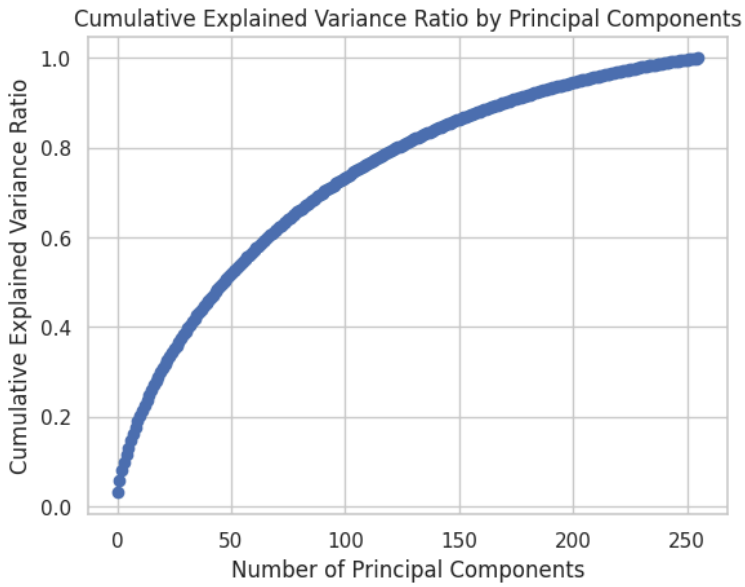

Figure 11: Cumulative Explained Variance Ratio Versus Number of Principal Components

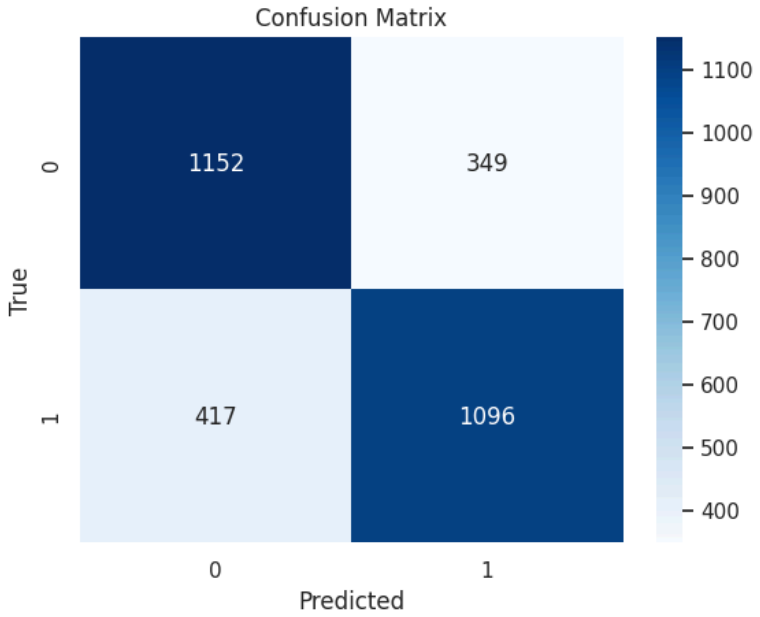

Figure 12: Confusion Matrix of Predicted and True Classifications

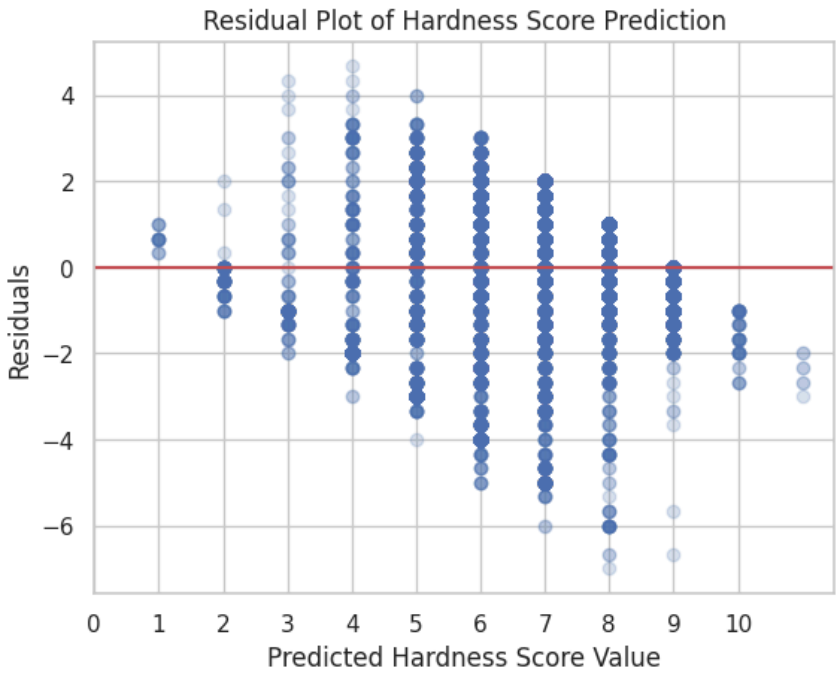

Figure 13: Residual Plot of Hardness Score Predictions