
Why Multi-Agent Conversations Cannot Fix LLM Forecasting: Evidence from Convergence Analysis

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Large language models (LLMs) fail catastrophically at forecasting, performing
2 31-78% worse than random guessing due to systematic overconfidence in their
3 predictions. This failure stems from models expressing high certainty regardless of
4 accuracy, creating dangerous misalignment between confidence and performance
5 that threatens deployment in critical decision-making domains. We investigate
6 whether multi-agent conversations provide natural calibration through structured
7 disagreement that moderates individual overconfidence. Testing 483 binary forecasting
8 questions reveals that conversational and mathematical calibration are
9 functionally equivalent—both inject uncertainty without improving reasoning,
10 converging to identical 0.25 Brier scores post-optimization. Mid-sized models
11 (7B-14B) benefit with 28.3% improvements through natural uncertainty moderation,
12 while larger models (32B+) suffer from sophisticated echo chambers where
13 elaborate arguments amplify rather than moderate overconfidence.

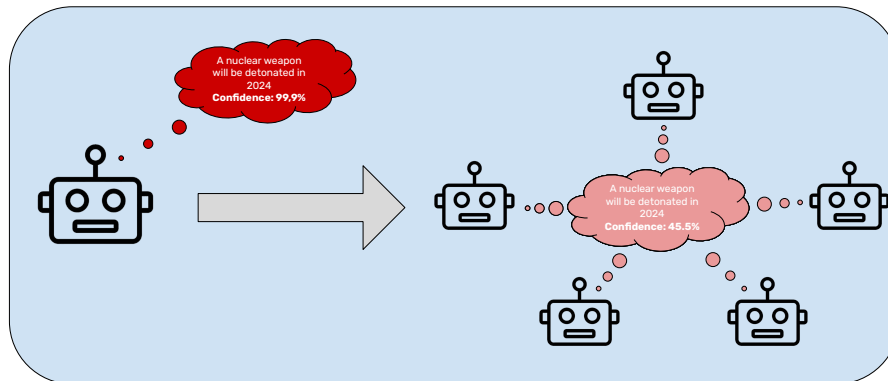


Figure 1: Overview of multi-agent system framework. Multi-agent conversations moderate overconfidence through social disagreement.

14 1 Introduction

15 Large language models fail catastrophically at forecasting future events. GPT-4’s probabilistic
16 forecasts show no improvement over naive 50% baselines in forecast tournaments, while all LLMs
17 generate predictions with inappropriate certainty, creating unreliable uncertainty estimates [Schoeneger
18 and Park, 2023]. This forecasting crisis threatens AI deployment in critical domains from business
19 strategy to policy planning where predictions must be accurate to drive decision-making.

20 LLM forecasting fails because models lack calibration mechanisms. Models express high confi-
21 dence regardless of accuracy, creating a mismatch between certainty and performance. Traditional
22 calibration approaches require hyperparameter optimization and external resources - mathematical
23 techniques force artificial uncertainty injection, ensemble methods need multiple model instances, and
24 enhanced reasoning frameworks demand additional training data [Halawi et al., 2024, Ren and Wang,
25 2025]. These solutions treat symptoms rather than addressing the core overconfidence problem.

26 We address LLM forecasting overconfidence through multi-agent conversations that provide natural
27 calibration via social disagreement (Figure 1). When AI agents engage in structured dialogue, they
28 expose each other to diverse perspectives that moderate individual overconfidence - disagreement
29 forces uncertainty into predictions that otherwise exhibit inappropriate certainty. Opinion dynamics
30 research shows that structured dialogue effectively updates beliefs and reduces biases [Hegselmann
31 and Krause, 2002, Liu and Chen, 2015], while LLM conversation studies demonstrate that peer inter-
32 action influences reasoning [Chuang et al., 2023]. We apply these social mechanisms to forecasting
33 calibration.

34 Our investigation reveals three critical insights about conversational calibration. First, we discover
35 that multi-agent conversations and a simple mathematical calibration method (where predictions
36 are mechanically adjusted toward 50% confidence) achieve identical performance improvements
37 through the same underlying mechanism, where both inject uncertainty without improving reasoning,
38 converging to similar scores around 0.20-0.25. Second, this benefit only applies to mid-sized
39 models (7B-14B parameters) that achieve up to 28.3% performance improvements. Third, we
40 identify a sophistication paradox where larger models (32B+) generate more linguistically complex
41 conversations that paradoxically amplify rather than reduce overconfidence, creating elaborate echo
42 chambers as measured by conversational complexity metrics.

43 We verified these findings through convergence analysis across 483 binary forecasting questions from
44 prediction markets. We measured direct performance improvements by comparing conversational
45 predictions with individual baselines, and then analyzed how mathematical calibration requirements
46 change when conversations provide natural moderation. Most critically, we demonstrated that post-
47 calibration performance converges regardless of the initial method, proving conversations offer no
48 benefit beyond partial uncertainty injection that could have been achieved by calibration too. This
49 convergence analysis, combined with linguistic sophistication measurements, reveals not just whether
50 conversations help, but proves they work through identical mechanisms as mathematical calibration.

51 Our findings establish clear deployment guidelines for multi-agent forecasting systems. The equiva-
52 lence between conversational and mathematical calibration means organizations can choose based
53 on computational resources: conversations require more compute but provide natural uncertainty,
54 while lambda optimization is computationally efficient but requires parameter tuning. The sophisti-
55 cation paradox provides a diagnostic tool - measure conversational complexity to predict whether
56 multi-agent systems will help or harm. Models generating simpler discussions benefit from peer
57 disagreement, while those producing rich dialogue suffer from overconfidence amplification. Our
58 contributions are:

- 59 • **Experimental proof of calibration equivalence:** We demonstrate that conversational and
60 mathematical calibration are functionally identical mechanisms, both achieving Brier scores
61 of around 0.25 post-optimization, with no additional reasoning benefits from conversations.
- 62 • **Sophistication paradox discovery:** We identify through CORE metrics that linguistic
63 richness inversely correlates with calibration success - more sophisticated conversations
64 produce worse forecasting by creating echo chambers rather than productive disagreement.
- 65 • **Architectural boundary analysis:** We establish that conversational benefits depend on
66 model size, with 7B-14B models achieving natural calibration equivalent to $\lambda \in [0.5, 0.94]$,
67 while 70B+ models require increased mathematical intervention after conversations.

68 2 Related Work

69 **Fundamental Limitations of LLM Forecasting:** LLMs fail at forecasting, performing at random
70 chance levels. Schoenegger found GPT-4’s probabilistic forecasts did not differ significantly from
71 naive 50% baselines in real-world tournaments, underperforming human crowd forecasts [Schoeneg-
72 ger and Park, 2023]. LLMs underperform XGBoost in credit rating predictions [Drinkall et al., 2024],

73 show sensitivity to input formatting [Park et al., 2025], and generate predictions without reliable
74 uncertainty estimation [Bi et al., 2025], making them unsuitable for high-stakes applications requiring
75 calibrated confidence intervals.

76 **Multi-Agent Systems and Enhancement Methods:** Collaborative approaches attempt to fix single-
77 LLM failures. Ensemble methods improve performance [Ren and Wang, 2025], with Halawi et al.
78 [2024] achieving human-parity Brier scores of 0.149 by combining six diverse reasoning instances.
79 Reasoning improvements and knowledge integration help: multistep frameworks improved Llama-
80 2-7B accuracy from 65.2% to 80.8% [Wang et al., 2025], news integration yielded 16-17% RMSE
81 reductions [Wang et al., 2024], and AI-augmented dialogue systems achieved 23-28% accuracy
82 improvements [Schoenegger et al., 2024]. These approaches require external knowledge integration,
83 news feeds, or specialized training data, while our conversational method achieves calibration
84 improvements using pre-existing model knowledge through structured disagreement.

85 **Opinion Dynamics and Conversational AI:** Opinion dynamics research provides foundations but
86 ignores forecasting applications. Traditional models [Hegselmann and Krause, 2002, Liu and Chen,
87 2015] simulate belief evolution via assimilation and consensus formation. LLM-based agents now
88 engage in opinion dynamics: Chuang et al. [2023] demonstrated that LLM agents show consensus
89 biases on factual topics but show sensitivity to prompt design variations [Ju et al., 2024]. LLMs show
90 biases in debate simulations [Taubenfeld et al., 2024] and spontaneous norm emergence with tipping
91 point dynamics [Ashery et al., 2024]. Existing opinion dynamics research focuses on preference
92 aggregation rather than forecasting contexts, leaving a gap: forecasting provides clear performance
93 metrics and objective ground truth validation.

94 3 Methodology

95 We test whether multi-agent conversations improve LLM forecasting by providing natural calibration
96 equivalent to simple mathematical intervention, using convergence analysis to prove mechanism
97 equivalence.

98 3.1 Dataset and Evaluation

99 We evaluated 483 binary prediction questions from Manifold Markets (Appendix A) to test genuine
100 forecasting ability. We selected questions using three criteria: (1) created before 2024, (2) resolved
101 yes/no during 2024-2025, and (3) minimum 100 unique forecasters. This ensures we test forecasting
102 rather than knowledge retrieval, as questions predate model training but resolve afterward. Questions
103 span technology adoption (112), human affairs and politics (235), markets and business (56), society
104 and culture (65), and natural phenomena (15).

105 We measure performance using Brier scores: $(\text{predicted_probability} - \text{actual_outcome})^2$, where
106 $\text{actual_outcome} = 1$ for YES and 0 for NO. Brier scores penalize overconfident predictions
107 quadratically, making them effective at detecting overconfidence.

108 3.2 Testing Calibration Equivalence

109 We test whether conversational calibration provides benefits beyond what simple mathematical
110 calibration achieves. Specifically, we show that for each model, the conversations naturally produce
111 an effect equivalent to applying lambda calibration with a particular parameter value. When we then
112 apply the optimal mathematical calibration to both conversational and non-conversational predictions,
113 they converge to identical performance (0.20-0.25 Brier scores), demonstrating that conversations
114 don't improve underlying reasoning, they merely inject a fixed amount of uncertainty that could have
115 been achieved through mathematical means.

116 **Direct Performance Comparison:** We compare conversational versus basic predictions using Brier
117 scores and paired t-tests to identify which models benefit from conversations.

Mathematical Calibration Requirements Analysis: We measure calibration requirements through
lambda optimization:

$$\text{calibrated_prediction} = (1 - \lambda) \times \text{original_prediction} + \lambda \times 0.5$$

118 Lambda calibration corrects overconfidence by pulling predictions toward 0.5. Lower optimal λ
119 values indicate less overconfidence. If conversations reduce overconfidence, they should require
120 smaller λ values than basic predictions.

121 3.3 Conversational Complexity Measurement

122 We measure conversational complexity using the Conversational Robustness Evaluation Score
123 (CORE) [Pandey et al., 2025], which integrates cluster entropy, lexical repetition, and semantic
124 similarity. This reveals whether linguistic sophistication helps or harms calibration:

125 **CORE Metrics:** Higher scores indicate richer vocabulary, greater semantic diversity, and more
126 complex argument structures. We correlate CORE scores with both model size and calibration
127 success.

128 **Qualitative Analysis:** We examine specific conversations to identify how sophistication manifests -
129 whether through productive disagreement or echo chamber formation.

130 3.4 Prediction Methods Overview

131 Increasing social complexity in prediction methods reveals when disagreement provides calibration
132 benefits. We test five methods:

- 133 1. **Community Baseline:** Historical Manifold Markets predictions before 2024 establish
134 human performance benchmarks.
- 135 2. **Basic LLM:** Single agents predict using log-likelihood ratios: $p(\text{yes})/(p(\text{yes}) + p(\text{no}))$ to
136 neglect hallucinations.
- 137 3. **Argument-Based:** Agents generate four arguments before predicting, testing whether
138 reasoning alone reduces overconfidence.
- 139 4. **Conversational:** Twenty agents conduct four-round conversations with six pairwise ex-
140 changes per round. Agents maintain argument banks with 75-token response limits, exposing
141 each other to disagreements.
- 142 5. **Extended Conversational:** Ten-round conversations test whether benefits accumulate or
143 saturate. Results (Section 4.6) show saturation at 4 rounds.

144 3.5 Experimental Design

145 We centered experiments on Llama-3.1-8B and validated across Llama-3.2-3B, Llama-3.1-70B, and
146 Qwen models (7B, 14B, 32B). We optimized parameters systematically: temperature 0.7, 20 agents,
147 4 rounds (Appendix C).

148 For calibration equivalence testing, we use 80%/20% topic-stratified splits with grid search from
149 0.0 to 1.0 in 0.01 increments. Each method processes all 483 questions independently. We analyze
150 results using:

- 151 • Paired t-tests for performance differences
- 152 • Lambda requirement changes to measure overconfidence calibration
- 153 • Post-calibration convergence to prove mechanism equivalence
- 154 • CORE scores to quantify the sophistication paradox

155 The convergence of post-calibration performance serves as our primary evidence that conversations
156 merely replicate (partial) mathematical calibration rather than providing additional reasoning benefits.

157 4 Results

158 We now present our experimental findings, beginning with baseline performance before examining
159 conversational improvements. For certain model architectures (primarily 7B-14B), multi-agent
160 conversations naturally provide partial calibration toward the 50% baseline, achieving an effect
161 equivalent to applying lambda calibration with a specific parameter value. However, for larger models

162 (32B+), conversations either provide no benefit or actually worsen calibration, while mathematical
 163 calibration remains effective for all models.

164 **4.1 LLMs perform 31-78% worse than random guessing**

165 All tested models fail catastrophically at forecasting, performing 31-78% worse than simply always
 166 predicting 50% (Figure 2).

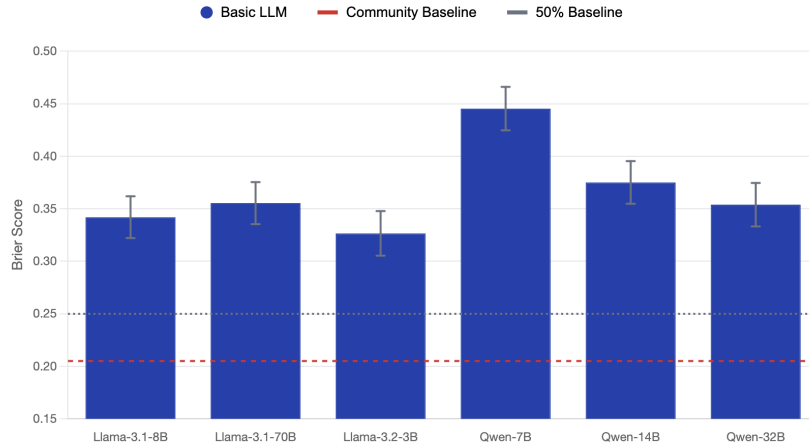


Figure 2: All tested LLMs perform 31-78% worse than simply always predicting 50%, demonstrating severe overconfidence that makes them unsuitable for forecasting tasks.

167 This universal failure reveals fundamental overconfidence: LLMs make predictions with high certainty
 168 despite poor accuracy.

169 **4.2 Conversations provide natural calibration for some models while leaving others
 170 unchanged**

171 Mid-sized models achieve substantial improvements through conversations, naturally achieving what
 172 mathematical calibration provides artificially, while larger models show negligible changes (Table 1).

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.282 ± 0.039	0.275 ± 0.030	+19.7%	p < 0.001
Llama-3.1-70B	0.355 ± 0.040	0.347 ± 0.033	+2.5%	p = 0.591
Llama-3.2-3B	0.327 ± 0.042	0.319 ± 0.038	+2.4%	p = 0.342
Qwen-7B	0.445 ± 0.041	0.319 ± 0.025	+28.3%	p < 0.001
Qwen-14B	0.354 ± 0.041	0.315 ± 0.038	+11.1%	p = 0.008
Qwen-32B	0.375 ± 0.040	0.342 ± 0.033	+8.8%	p = 0.058

Table 1: Conversational calibration provides natural uncertainty injection for mid-sized models but fails for larger architectures

173 Three models show statistically significant improvements ($p < 0.01$), while three show minimal,
 174 non-significant changes ($p > 0.05$). Conversations either provide substantial benefits or no significant
 175 effects, with no middle ground.

176 **4.3 Lambda analysis confirms calibration equivalence**

177 Lambda calibration corrects overconfidence by pulling predictions toward 0.5, where optimal λ
 178 values indicate overconfidence levels. Models showing conversational benefits (according to Table 1)
 179 require less mathematical calibration afterward (Table 2). Llama-3.1-8B shows the largest lambda
 180 reduction alongside the largest performance gain, while Llama-3.1-70B shows increased lambda with
 181 no significant improvement.

Model	Basic λ	Conv. λ	λ Change	Overconfidence Effect
Llama-3.1-8B	0.740	0.590	-0.150	Strong reduction
Llama-3.1-70B	0.750	0.850	+0.100	Clear increase
Llama-3.2-3B	0.660	0.680	+0.020	Slight increase
Qwen-7B	0.970	0.940	-0.030	Minimal reduction
Qwen-14B	0.730	0.680	-0.050	Moderate reduction
Qwen-32B	0.820	0.850	+0.030	Slight increase

Table 2: Lambda requirements change predictably: models with performance improvements from conversations need less mathematical calibration, while unsuccessful models need more, proving conversations work through overconfidence moderation.

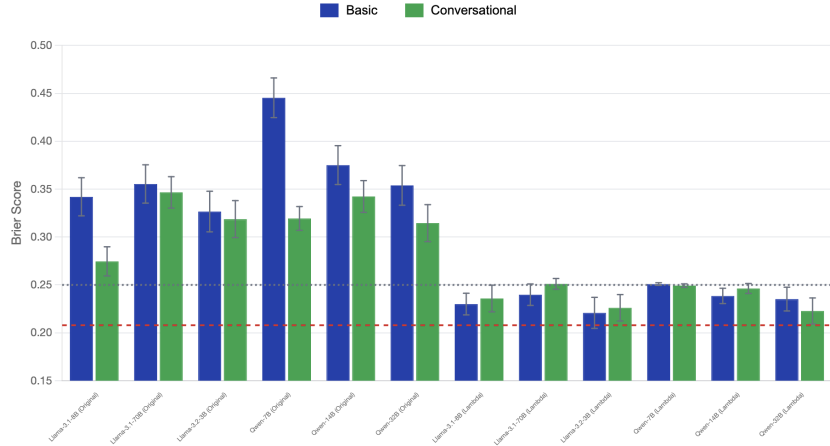


Figure 3: Conversations help only in cases where no calibration is done. Thus, mathematical calibration eliminates conversational benefits by artificially injecting the same uncertainty that conversations provide naturally, confirming that both methods work simply through confidence moderation. (Left: Without lambda calibration, Right: With lambda calibration)

182 Identical performance is achieved after calibration whether conversations were used or not, confirming
 183 conversations are helpful only through the same uncertainty moderation mechanism as the simple
 184 lambda calibration (Figure 3). After optimal lambda calibration, basic and conversational methods
 185 achieve nearly identical Brier scores around 0.20 to 0.25. This convergence proves conversations
 186 work through confidence moderation rather than improved reasoning—when both methods receive
 187 identical uncertainty injection, performance equalizes.

188 4.4 Why conversations help some models but not others

189 The correlation between lambda changes and performance improvements from conversations reveals
 190 clear architectural boundaries (Figure 4). Models benefit from conversations when lambda require-
 191 ments decrease, illustrating that performance improvements occur through overconfidence reduction
 192 rather than improved reasoning.

Model	CORE Score
Llama-3.2-3B	0.248 ± 0.073
Llama-3.1-8B	0.264 ± 0.072
Llama-3.1-70B	0.341 ± 0.077
Qwen-7B	0.404 ± 0.089
Qwen-14B	0.406 ± 0.090
Qwen-32B	0.427 ± 0.090

Table 3: CORE scores reveal increasing linguistic complexity with model size, yet this sophistication correlates negatively with forecasting performance.

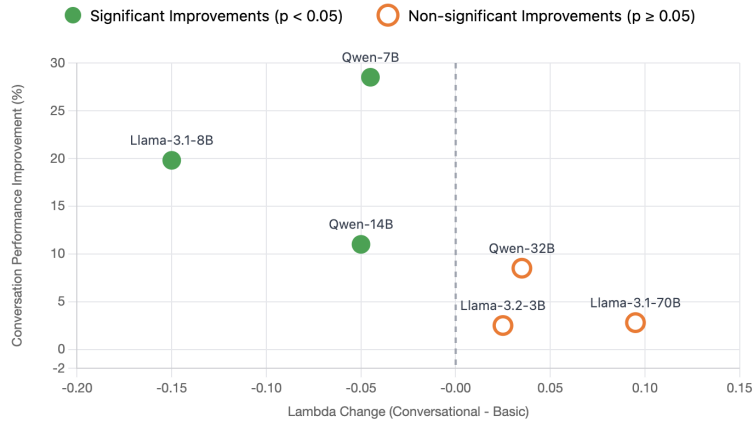


Figure 4: Models benefit from conversations when lambda requirements decrease after adding conversations, illustrating that performance improvements occur through overconfidence reduction rather than other mechanisms.

193 Large models (32B+) produce more linguistically complex conversations with higher CORE scores:
 194 richer vocabulary, steeper Zipf distributions, and greater semantic sophistication (Table 4.4). Paradox-
 195 ically, this complexity correlates with worse performance during these conversations and increased
 196 calibration requirements. Qualitative analysis (Appendix D) reveals that larger models fail through:

- 197 • Overthinking with elaborate frameworks when base rates sufficed
- 198 • Favoring abstract institutional reasoning over concrete evidence
- 199 • Sophistication-induced overconfidence from articulate but incorrect reasoning

200 Mid-sized models (7B-14B) benefit because simpler disagreements provide natural calibration—they
 201 achieve performance equivalent to $\lambda \in [0.59, 0.68]$ through conversation alone. Larger models
 202 (32B+) suffer from overconfidence amplification, requiring up to 10% more mathematical calibration
 203 afterward (lambda increases to 0.85). Their sophisticated discussions create echo chambers that
 204 reinforce rather than challenge overconfidence.

205 This inverse relationship between linguistic complexity and calibration success establishes clear
 206 deployment boundaries: use conversational calibration for models showing negative lambda changes,
 207 avoid it for those showing positive changes.

208 4.5 Extended conversations provide no additional benefit

209 Extended conversations show no improvement beyond 4 rounds, confirming that natural calibration
 210 benefits saturate quickly rather than accumulating (Figure 5).

211 Performance oscillates randomly around the 4-round baseline (0.2746 ± 0.015) without directional
 212 improvement ($p = 0.847$ for linear trend). This saturation further confirms that conversations
 213 work through uncertainty injection rather than iterative reasoning improvement. Once sufficient
 214 disagreement exposure occurs, additional rounds add noise without signal.

215 5 Discussion and Future Directions

216 In our experiments, multi-agent conversations and mathematical calibration are functionally identical
 217 mechanisms for improving LLM forecasting—both inject uncertainty without enhancing reasoning
 218 capabilities. Our convergence analysis proves this equivalence: all models achieve around 0.20 to 0.25
 219 Brier scores after optimal calibration regardless of initial method, demonstrating that conversations
 220 offer no advantage beyond what simple lambda moderation provides.

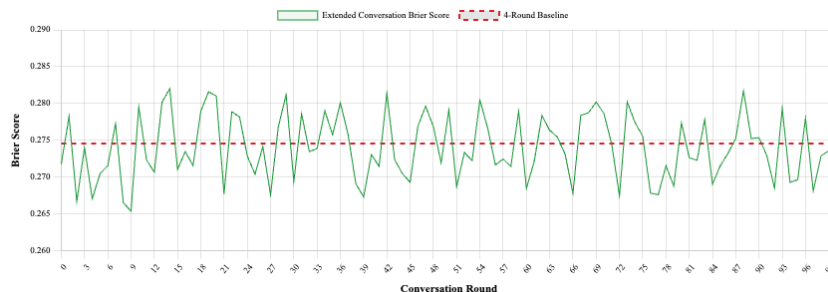


Figure 5: Extended conversations beyond 4 rounds produce random performance fluctuations without improvement, demonstrating that natural calibration benefits saturate quickly.

221 The practical implications are clear: organizations can choose between natural (conversational)
 222 or artificial (mathematical) calibration based on computational resources rather than effectiveness.
 223 Conversations require substantial compute but provide organic uncertainty moderation. Mathematical
 224 calibration is computationally efficient but requires parameter tuning. Since both achieve the same
 225 effect, the choice reduces to operational constraints.

226 Our discovery of the sophistication paradox—that linguistic complexity inversely correlates with
 227 calibration success—challenges assumptions about multi-agent benefits. Models generating CORE
 228 scores above 0.35 create elaborate echo chambers that amplify rather than moderate overconfidence.
 229 The correlation between linguistic richness and calibration failure reveals that sophisticated discus-
 230 sions can be counterproductive. Larger models (32B+) produce articulate but overconfident reasoning,
 231 requiring 10% more mathematical intervention after conversations.

232 5.1 Where might our findings not generalize?

233 Our equivalence finding may not extend beyond binary forecasting contexts. Multi-class predictions
 234 or continuous outcomes might show different convergence patterns. Our protocol (4 rounds, 6
 235 exchanges, 75 tokens) represents one implementation—other structures might alter the equivalence
 236 relationship.

237 We evaluated six models across 3B-70B parameters, but rapid LLM evolution means future ar-
 238 chitectures might break the sophistication paradox pattern. We tested foundation models without
 239 fine-tuning; task-specific models might show different calibration equivalence thresholds. Further
 240 work should investigate whether multi-agent conversations provide benefits beyond uncertainty
 241 injection, such as improved factual accuracy, better identification of relevant considerations, or
 242 enhanced reasoning quality that our analysis didn’t capture.

243 5.2 How can we extend conversational calibration beyond binary forecasting?

244 **Testing Alternative Uncertainty Injection Methods:** Compare other calibration techniques (tem-
 245 perature scaling, ensemble methods) to determine if all uncertainty injection mechanisms converge to
 246 the same performance ceiling.

247 **Breaking the Sophistication Paradox:** Design conversation protocols that prevent echo chamber
 248 formation in large models. Test structured disagreement mechanisms that maintain simplicity despite
 249 model capacity. Investigate whether limiting vocabulary complexity preserves calibration benefits.

250 **Mechanism Decomposition:** Isolate why simpler conversations calibrate better through ablation
 251 studies. Determine whether CORE score thresholds predict calibration failure across different
 252 tasks. Identify linguistic markers that distinguish productive disagreement from sophisticated echo
 253 chambers.

254 6 Broader Impacts

255 Our experiments suggest that multi-agent conversations and mathematical calibration are equivalent
256 mechanisms, providing organizations with implementation flexibility while warning against universal
257 deployment.

258 **Positive Applications:** The calibration equivalence finding democratizes forecasting improve-
259 ment—organizations can choose methods based on available resources rather than effectiveness. Our
260 CORE score threshold (0.35) provides a diagnostic tool for predicting multi-agent success. Under-
261 standing that conversations do not improve forecasting reasoning helps set appropriate expectations
262 for AI-assisted decision-making.

263 **Potential Risks:** The sophistication paradox reveals that advanced models can become more dan-
264 gerous through multi-agent interaction, generating convincing but overconfident predictions. Users
265 might misinterpret eloquent conversations as evidence of better forecasting reasoning when they
266 actually indicate worse calibration.

267 **Deployment Considerations:** Organizations must test model-specific CORE scores before deploying
268 multi-agent systems. Larger models (32B+) should use mathematical calibration exclusively, as
269 conversations amplify their overconfidence. Monitor linguistic complexity continuously—rising
270 CORE scores signal degrading calibration. Design systems that explicitly acknowledge the uncertainty
271 injection mechanism rather than suggesting improved reasoning.

272 **AI Safety Implications:** The sophistication paradox demonstrates that model capabilities can become
273 liabilities—more advanced models fail worse at collaborative calibration. The convergence to around
274 0.20 and 0.25 Brier scores suggests fundamental limits to post-hoc calibration, highlighting the
275 need for inherently calibrated architectures. Our diagnostic tools provide practical metrics for safe
276 deployment decisions.

277 7 Conclusion

278 We demonstrated that multi-agent conversations and mathematical calibration are functionally equiv-
279 alent mechanisms for addressing LLM forecasting failures. Both approaches inject uncertainty
280 without improving underlying reasoning, converging to identical Brier scores of around 0.20 to
281 0.25 after optimization. This equivalence reveals that conversations provide no computational
282 advantage—organizations can choose between natural conversational calibration (computationally
283 expensive but organic) or mathematical calibration (efficient but requires tuning) based solely on avail-
284 able resources rather than effectiveness. Our discovery of the sophistication paradox, where linguistic
285 complexity inversely correlates with calibration success, establishes clear deployment boundaries:
286 models generating CORE scores below 0.35 benefit from conversational disagreement, while those
287 above this threshold suffer from overconfidence amplification requiring additional mathematical
288 intervention.

289 Future work should investigate whether alternative uncertainty injection methods converge to the same
290 performance ceiling, suggesting fundamental limits to post-hoc calibration. Developing conversation
291 protocols that prevent elaborate reinforcement spirals in large models could extend benefits beyond
292 mid-sized architectures. Most critically, our findings indicate that true progress requires architec-
293 tures with inherent calibration rather than post-hoc corrections—current approaches merely force
294 uncertainty into naturally overconfident systems without addressing the root cause of miscalibration.

295 **References**

- 296 Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. The dynamics of social conventions
297 in LLM populations: Spontaneous emergence, collective biases and tipping points. 2024.
- 298 Zhaoge Bi, Linghan Huang, Haolin Jin, Qingwen Zeng, and Huaming Chen. NSW-EPNews: A
299 news-augmented benchmark for electricity price forecasting with LLMs. May 2025.
- 300 Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang,
301 Dhavan Shah, Junjie Hu, and Timothy T Rogers. Simulating opinion dynamics with networks of
302 LLM-based agents. 2023.
- 303 Felix Drinkall, Janet B Pierrehumbert, and Stefan Zohren. Forecasting credit ratings: A case study
304 where traditional methods outperform generative LLMs. July 2024.
- 305 Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level
306 forecasting with language models. February 2024.
- 307 Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis
308 and simulation. *Journal of Artificial Societies and Social Simulation*, 5, 07 2002.
- 309 Da Ju, Adina Williams, Brian Karrer, and Maximilian Nickel. Sense and sensitivity: Evaluating the
310 simulation of social dynamics via large language models. 2024.
- 311 Pei Liu and Xi Chen. An overview on opinion spreading model. *J. Appl. Math. Phys.*, 03(04):
312 449–454, 2015.
- 313 Punya Syon Pandey, Yongjin Yang, Jiarui Liu, and Zhijing Jin. CORE: Measuring multi-agent LLM
314 interaction quality under game-theoretic pressures. August 2025.
- 315 Junwoo Park, Hyuck Lee, Dohyun Lee, Daehoon Gwak, and Jaegul Choo. Revisiting LLMs as
316 zero-shot time-series forecasters: Small noise can break large models. May 2025.
- 317 Yinuo Ren and Jue Wang. Can LLM improve for expert forecast combination? evidence from the
318 european central bank survey. June 2025.
- 319 Philipp Schoenegger and Peter S Park. Large language model prediction capabilities: Evidence from
320 a real-world forecasting tournament. October 2023.
- 321 Philipp Schoenegger, Peter S Park, Ezra Karger, Sean Trott, and Philip E Tetlock. AI-augmented
322 predictions: LLM assistants improve human forecasting accuracy. February 2024.
- 323 Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in LLM
324 simulations of debates. *arXiv [cs.CL]*, 2024.
- 325 Jiahao Wang, Mingyue Cheng, and Qi Liu. Can slow-thinking LLMs reason over time? empirical
326 studies in time series forecasting. May 2025.
- 327 Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating
328 event analysis in LLM-based time series forecasting with reflection. September 2024.

329 **A Appendix: Dataset Construction and Model Specifications**

330 **A.1 Data Source**

331 Our dataset derives from Manifold Markets, a prediction market platform where users trade virtual
332 currency on the outcomes of real-world events. Manifold Markets operates as a play-money prediction
333 market, allowing participants to create questions about future events and trade shares based on their
334 probability assessments. The platform aggregates individual predictions into market prices that
335 represent collective probability estimates, providing a rich source of community forecasting data
336 across diverse topics. Unlike traditional prediction markets that use real money, Manifold’s virtual
337 currency system encourages broader participation while maintaining forecasting incentives through
338 reputation mechanisms and leaderboards.

339 **A.2 Selection Criteria**

340 Our dataset construction applies rigorous filtering to ensure experimental validity through multiple
341 quality thresholds. Temporal requirements require question creation before January 1, 2024, with
342 resolution periods occurring between January 1, 2024 and December 31, 2025, ensuring compatibility
343 with model knowledge cutoffs. Participation thresholds require a minimum of 100 unique forecasters
344 per question to ensure sufficient market activity, while resolution clarity restricts inclusion to definitive
345 YES/NO outcomes only, excluding questions with ambiguous or subjective interpretation criteria.

346 Quality validation includes manual review to exclude manipulation patterns such as excessive betting
347 concentration from single participants, suspicious timing patterns in bet placement, or unusual price
348 volatility inconsistent with natural information flow. Each question provides both a community
349 baseline probability (the final market prediction before 2024) and a definitive binary outcome for
350 evaluation.

351 **A.3 Topic Stratification and Domain Analysis**

Topic Category	Number of Questions
Human Affairs & Politics	235
Technology & Innovation	112
Society & Culture	65
Markets & Business	56
Nature & Physical World	15
Total	483

Table 4: Political and technology questions dominate the dataset, ensuring robust evaluation across domains where forecasting matters most.

352 This distribution ensures broad coverage while maintaining sufficient samples for statistical analysis
353 within each domain. Political and technology questions comprise the largest categories, reflecting the
354 interests of prediction market participants and allowing robust topic-specific analysis.

355 **A.4 Model Specifications**

356 We tested six language models with the following complete specifications:

- 357 • **Llama-3.1-8B**: meta-llama/Llama-3.1-8B-Instruct
- 358 • **Llama-3.1-70B**: meta-llama/Llama-3.1-70B-Instruct
- 359 • **Llama-3.2-3B**: meta-llama/Llama-3.2-3B-Instruct
- 360 • **Qwen-7B**: qwen/Qwen1.5-7B-Chat
- 361 • **Qwen-14B**: qwen/Qwen1.5-14B-Chat
- 362 • **Qwen-32B**: qwen/Qwen1.5-32B-Chat

363 Throughout this paper, we use shortened names for readability while maintaining clear model
364 identification.

365 **B Appendix: Prediction Method Implementations**

366 **B.1 Basic LLM Prediction Baseline**

367 **Prompt Template:**

368 Will artificial intelligence surpass human performance in chess by 2000?

369

370 This question requires a Yes or No answer only.

371 Answer: Yes

372

373 [Target forecasting question]

374

375 This question requires a Yes or No answer only.

376 Answer: [Model response]

377 The system quantifies prediction confidence through token-level probability analysis over model
378 yes/no responses. Probability extraction targets specific tokens including "Yes", "Y", "No", and "N"
379 with case-insensitive matching, followed by token cleaning to remove whitespace, tabs, and formatting
380 characters. The normalized strength calculation follows the formula $strength = \frac{p(yes)}{p(yes)+p(no)}$ with a
381 fallback to neutral 0.5 when no valid tokens are detected or the probability sums to 0.

382 **B.2 Argument-Based Prediction Method**

383 **Prompt Template:**

384 Question: Will the United States win the most gold medals
385 at the 2024 Olympics?

386 Argument: The US has historically strong performance across multiple
387 sports and significant investment in Olympic training programs.

388

389 Question: Will Bitcoin reach \$100,000 by the end of 2024?

390 Argument: Institutional adoption is increasing and Bitcoin ETFs are
391 gaining approval, creating more mainstream demand.

392

393 Question: Will there be a major earthquake (magnitude 7+) in California
394 in 2024?

395 Argument: California sits on active fault lines and seismic activity
396 has been increasing in recent years.

397

398 [Target question]

399 Argument: [Model generates argument]

400 The system generates four arguments per question through repeated sampling with various prompts,
401 ensuring comprehensive evidence collection before the final prediction. These generated arguments
402 are then combined in an informed prediction prompt that explicitly references the agent's complete
403 argument set, allowing the model to make evidence-based forecasts that incorporate all collected
404 reasoning. The final prediction extraction follows the same token probability analysis as used in the
405 basic method but now incorporates the influence of generated supporting evidence.

406 **B.3 Conversational Prediction Implementation**

407 **Agent Communication Protocol:**

408 Carry on the conversation given to you. Speak in 3 sentences or less.

409

410 Topic: [Forecasting question]

411

412 Conversation so far:

413 [Complete dialogue history between current agent pair]

414

415 Your response: [Agent generates contextual response]

416 **Perspective Extraction Template:**

417 Topic: Will Bitcoin reach \$100,000 by the end of 2024?

418

419 Conversation:

420 Agent A: Institutional adoption is accelerating with
421 Bitcoin ETFs getting approved. This brings legitimacy
422 and more capital inflow.

423 Agent B: However, regulatory uncertainty still looms large.
424 Any major crackdown could severely impact the price trajectory.

425 Agent A: The halving event in 2024 historically drives price
426 increases due to supply constraints. This fundamental factor
427 is predictable.

428

429 State a new perspective that you believe in in one sentence
430 from our conversation.

431 Perspective: Bitcoin's price will depend more on regulatory clarity
432 than technical factors like the halving event.

433

434 Topic: [Current forecasting question]

435

436 Conversation:

437 [Actual conversation history]

438

439 State a new perspective that you believe in in one sentence
440 from our conversation.

441 Perspective: [Agent extracts new insight]

442 When agents reach their 4-perspective capacity limit, the system implements intelligent replacement
443 by identifying the least relevant existing perspective. Relevance determination measures each
444 perspective's contribution to the agent's overall prediction confidence, with perspectives contributing
445 least to coherent belief formation prioritized for replacement. Complete perspective isolation occurs
446 between questions through comprehensive agent state reset, preventing belief contamination across
447 different forecasting tasks, and ensuring independent evaluation of each question.

448 **B.4 Extended Conversational Method**

449 The extended conversational method implements 10 total rounds compared to the standard 4-round
450 protocol, allowing investigation of whether additional deliberation provides continued calibration
451 benefits or reaches saturation points. Performance tracking calculates Brier scores after each round to
452 monitor convergence patterns and identify optimal stopping points. Saturation detection employs
453 statistical testing to identify performance plateaus, using moving window analysis to determine when
454 continued conversation becomes counterproductive rather than beneficial.

455 **B.5 Lambda Optimization Methodology**

456 Lambda optimization uses a topic-stratified 80%/20% train/test split with independent optimization
457 for each prediction method. The grid search tests λ values from 0.0 to 1.0 in 0.01 increments to
458 minimize mean Brier score, with optimal lambda values measuring overconfidence levels.

459 The convergent evidence analysis compares λ requirements between basic and conversational methods
460 for each model. This dual measurement approach provides independent confirmation of overcon-
461 fidence reduction as the causal mechanism, since lambda changes directly measure confidence
462 modifications while performance changes measure the resulting forecasting benefits.

463 **C Appendix: Parameter Optimization and Validation**

464 **C.1 Agent Count Scaling Analysis**

465 We systematically tested agent populations of 5, 10, and 20 to determine optimal group sizes for
466 conversational calibration effects.

Agents	Basic Brier	Conversational Brier	Improvement
5	0.3449	0.2967	14.0%
10	0.3425	0.2816	17.8%
20	0.3420	0.2746	19.7%

Table 5: Larger agent groups produce stronger calibration effects through increased perspective diversity and disagreement exposure.

467 Larger discussion groups consistently generate more diverse disagreement and stronger uncertainty
468 moderation. The improvement scaling with group size supports our hypothesis that social calibration
469 benefits from the increase in perspective diversity encountered during conversation.

470 **C.2 Temperature Optimization**

471 Temperature selection balances response creativity with conversational coherence. We tested values
472 of 0.3, 0.5, and 0.7 using Llama-3.1-8B with 10 agents.

Temperature	Basic Brier	Conversational Brier	Improvement
0.3	0.3655	0.2970	18.7%
0.5	0.3559	0.2869	19.4%
0.7	0.3420	0.2746	19.7%

Table 6: Higher temperature settings improve conversational benefits by preventing premature consensus and maintaining argumentative diversity.

473 Higher temperatures consistently improve conversational benefits through increased argumentative
474 diversity. Temperature 0.7 provides optimal balance, facilitating varied argumentative strategies while
475 preventing premature consensus convergence.

476 **C.3 Argument Count Optimization**

477 We systematically tested argument generation counts of 4, 6, and 8 arguments per trial to determine
478 the optimal number of supporting arguments for the argument-based prediction method.

Arguments	Basic Brier	Conversational Brier	Improvement
4	0.3420	0.2746	19.7%
6	0.3406	0.2805	17.6%
8	0.3393	0.2805	17.3%

Table 7: Four arguments optimize the balance between reasoning depth and computational efficiency, with diminishing returns beyond this threshold.

479 The results reveal that 4 arguments per trial provide optimal performance across all prediction
480 methods. Although eight arguments might theoretically provide more comprehensive evidence
481 collection, the practical benefits plateau at four arguments, with diminishing returns beyond this
482 threshold. The 4-argument configuration achieves the strongest conversational improvements (19.7%)
483 and maintains the best balance between reasoning thoroughness and computational efficiency.

484 **C.4 Model Performance by Topic Domain**

485 Different models demonstrate varying strengths across forecasting domains, revealing important
486 insights about model-specific capabilities for different types of questions.

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.2840	0.1716	+39.6%	0.05
Llama-3.1-70B	0.3422	0.3367	+1.6%	0.2
Llama-3.2-3B	0.1889	0.1863	+1.4%	0.2
Qwen-7B	0.4870	0.3314	+32.0%	0.001
Qwen-14B	0.3370	0.2927	+13.1%	0.2
Qwen-32B	0.3001	0.2182	+27.3%	0.1

Table 8: Technology forecasting benefits most from collaborative discussion as technical disagreements effectively challenge overconfident assumptions.

487 Technology questions show statistically significant improvements only for Qwen-7B. Technical
488 reasoning benefits from collaborative discussion, where peer disagreement effectively challenges
489 overconfident assumptions, though this effect is not consistent across all model architectures.

490 **Human Affairs & Politics Questions (235 questions):**

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.3284	0.3085	+6.1%	0.2
Llama-3.1-70B	0.3152	0.3403	-8.0%	0.2
Llama-3.2-3B	0.3523	0.3474	+1.4%	0.2
Qwen-7B	0.4235	0.3271	+22.8%	0.001
Qwen-14B	0.3842	0.3470	+9.7%	0.1
Qwen-32B	0.3567	0.3281	+8.0%	0.1

Table 9: Political forecasting shows limited conversational benefits with only one model achieving statistical significance.

491 Political questions show statistically significant improvements only for Qwen-7B. The contentious
492 nature of political topics may cause disagreement to reinforce rather than moderate existing positions,
493 leading to limited calibration benefits across most model architectures.

494 **Markets & Business Questions (56 questions):**

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.4610	0.3225	+30.1%	0.01
Llama-3.1-70B	0.4539	0.3596	+20.8%	0.05
Llama-3.2-3B	0.4434	0.3953	+10.8%	0.05
Qwen-7B	0.4968	0.3085	+37.9%	0.01
Qwen-14B	0.4144	0.3922	+5.4%	0.2
Qwen-32B	0.4733	0.4297	+9.2%	0.2

Table 10: Economic forecasting shows statistically significant improvements for two models through collaborative analysis of market dynamics.

495 Economic forecasting shows strong improvements for Qwen-7B and Llama-3.1-8B, with both models
496 achieving statistical significance. This suggests that economic reasoning benefits substantially from
497 social disagreement and collaborative analysis of market dynamics, though this effect is limited to
498 specific model architectures.

499 **Society & Culture Questions (65 questions):**

500 Cultural predictions show no statistically significant improvements across any model. These questions
501 occupy the middle ground between technical questions and political questions, but fail to demonstrate
502 reliable calibration benefits from collaborative discussion.

503 **Nature & Physical World Questions (15 questions):**

504 Scientific questions show no statistically significant improvements across any model despite the
505 small sample size (15 questions). Collaborative discussion benefits appear to require both appropriate
506 domain knowledge and compatible model architectures, conditions that are not met for scientific
507 reasoning in this evaluation.

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.3965	0.3025	+23.7%	0.05
Llama-3.1-70B	0.4265	0.3845	+9.8%	0.2
Llama-3.2-3B	0.3833	0.3890	-1.5%	0.2
Qwen-7B	0.3923	0.2742	+30.1%	0.1
Qwen-14B	0.3490	0.3825	-9.6%	0.2
Qwen-32B	0.3490	0.3316	+5.0%	0.2

Table 11: Cultural predictions show no statistically significant benefits from collaborative discussion across all tested models.

Model	Basic	Conversational	Improvement	p-value
Llama-3.1-8B	0.3065	0.2145	+30.0%	0.2
Llama-3.1-70B	0.4070	0.3078	+24.4%	0.2
Llama-3.2-3B	0.2685	0.2679	+0.2%	0.2
Qwen-7B	0.5153	0.3437	+33.3%	0.2
Qwen-14B	0.4849	0.2785	+42.6%	0.05
Qwen-32B	0.2860	0.3168	-10.8%	0.1

Table 12: Scientific forecasting shows no statistically significant improvements from collaborative discussion.

508 C.5 Argument-Based Prediction Results

509 We validated that conversational benefits stem from social disagreement, not just structured thinking.
 510 Argument-based methods, where agents generate reasoning before predicting without discussion,
 511 show the same architectural compatibility but weaker improvements (Table 13).

Model	Basic	Argument	Improvement
Llama-3.1-8B	0.342	0.288	+15.8%
Llama-3.1-70B	0.355	0.449	-26.4%
Llama-3.2-3B	0.327	0.313	+4.3%
Qwen-7B	0.445	0.336	+24.6%
Qwen-14B	0.375	0.336	+10.4%
Qwen-32B	0.354	0.298	+15.9%

Table 13: Argument-based methods show identical compatibility patterns to conversations, confirming that structured reasoning alone cannot match social disagreement benefits.

512 The 70B model deteriorates with structured reasoning (-26.4%), confirming overconfidence amplifi-
 513 cation occurs even without social interaction. Conversations consistently outperform argument-based
 514 methods for compatible models (Llama-3.1-8B: 0.288 argument vs. 0.275 conversational). This 4.8%
 515 additional improvement proves genuine disagreement elicits perspectives that structured prompting
 516 cannot access.

517 C.6 Post-Calibration Convergence Analysis

518 To verify that conversations and mathematical calibration work through the same mechanism, we
 519 examined performance after optimal lambda calibration. If both methods operate through uncer-
 520 tainty moderation, post-calibration performance should converge regardless of the initial prediction
 521 method. All models converge to approximately 0.25 after calibration, with no statistically significant
 522 differences between basic and conversational methods (Table 14). The tight clustering around 0.25
 523 confirms that both approaches ultimately achieve the same calibrated state through different paths.
 524 Conversations provide organic uncertainty moderation for compatible models, while mathematical
 525 calibration forces the same effect artificially.

526 C.7 Computational Resources

527 All experiments were conducted using NVIDIA A6000 GPUs with 48GB VRAM minimum required
 528 for concurrent agent conversations. Multi-GPU setups were utilized for dataset split processing
 529 to enable parallel execution across the 483 forecasting questions. Individual question processing

Model	Basic	Conversational
Llama-3.2-3B	0.221 ± 0.016	0.226 ± 0.014
Llama-3.1-7B	0.230 ± 0.011	0.236 ± 0.014
Llama-3.1-70B	0.240 ± 0.011	0.251 ± 0.006
Qwen-7B	0.251 ± 0.002	0.249 ± 0.002
Qwen-14B	0.239 ± 0.008	0.246 ± 0.005
Qwen-32B	0.235 ± 0.012	0.223 ± 0.014

Table 14: Post-calibration performance convergence demonstrates identical mechanisms. After lambda optimization, basic and conversational methods achieve nearly identical Brier scores clustered around 0.25.

530 time varied by method complexity: basic predictions (30 seconds), argument-based (2 minutes),
531 conversational (4-5 minutes), and extended conversational (10+ minutes per question). The 48GB
532 VRAM requirement was necessary to support concurrent multi-agent conversations, particularly for
533 the 20-agent conversational methods.

534 **D Appendix: Analysis of Conversational Failures**

535 Below we present detailed examples demonstrating how linguistic sophistication correlates with
536 forecasting failure, followed by quantitative CORE score analysis.

537 **D.1 Case Study: Will Sam Altman start a new company before 2025?**

538 **Actual Outcome:** NO

539 **Llama-3.2-3B Prediction:** 3.7% confidence YES (Brier Score: 0.001)

540 Agent 3: "Sam Altman's current role at OpenAI and his
541 involvement with Y Combinator might limit his bandwidth
542 for starting an entirely new company before 2025. While
543 his entrepreneurial spirit is undeniable, juggling
544 multiple high-profile commitments could be challenging."
545

546
547 Agent 4: "That's a valid point. His current
548 responsibilities might indeed constrain his ability to
549 launch something entirely new. However, given his track
550 record and influence, he might explore partnerships or
551 advisory roles that don't require the same level of time
552 commitment as founding a company from scratch."
553

554
555 Agent 3: "The timing also matters here. We're already
556 well into the current period, and starting a new company
557 typically requires months or years of planning, funding,
558 and development. With his OpenAI obligations, finding
559 that time seems impractical."
560

561
562 Agent 4: "Additionally, there would be significant
563 scrutiny and potential conflict of interest concerns if
564 the OpenAI CEO suddenly announced a new venture. The
565 optics alone might discourage such a move, at least while
566 he's actively leading OpenAI."
567

568
569 Agent 3: "His recent focus has been on navigating OpenAI
570 through major transitions, including partnerships with
571 Microsoft and product launches. These are consuming
572 endeavors that likely occupy most of his professional
573 attention and energy."
574

575
576 Agent 4: "Considering all these factors - time
577 constraints, potential conflicts, current
578 responsibilities, and the practical challenges of
579 starting a new venture while serving as CEO elsewhere -
580 I'd estimate there's very little chance Altman starts a
581 new company before 2025."
582

582 **Llama-3.1-70B Prediction:** 89.2% confidence YES (Brier Score: 0.796)

583 Agent 9: "Sam Altman has been a prominent figure in the
584 tech industry, especially with his work as the CEO of
585 OpenAI and former president of Y Combinator. Given his
586 experience and network, it's likely that he'll be

587 involved in some capacity with a new company soon."

588

589

590 Agent 4: "I think it's highly probable that his next
591 venture will be AI-related, given his existing expertise
592 and the rapid growth of the field. Either way, it's
593 likely to be a high-impact and innovative endeavor."

594

595

596 Agent 9: "His track record speaks for itself - he's been
597 involved in numerous successful ventures and has a keen
598 eye for transformative technologies. The combination of
599 his Y Combinator experience and current AI expertise
600 positions him perfectly to identify and pursue the next
601 big opportunity."

602

603

604 Agent 4: "Plus, successful entrepreneurs like Altman
605 rarely stay focused on just one venture indefinitely.
606 They tend to diversify and explore new challenges. With
607 his connections and reputation, securing funding and
608 talent for a new venture would be relatively
609 straightforward."

610

611

612 Agent 9: "The current AI boom also creates numerous
613 opportunities for new companies. Altman is uniquely
614 positioned to identify gaps in the market that others
615 might miss. I wouldn't be surprised if he's already
616 incubating ideas or having discussions about potential
617 new ventures."

618

619

620 Agent 4: "Given his entrepreneurial DNA, extensive
621 network, the abundance of AI opportunities, and
622 historical patterns of serial entrepreneurs, I'd put the
623 probability very high - maybe 85-90% - that Altman
624 starts or co-founds a new company before 2025. It would
625 be more surprising if he didn't."

626 **Analysis:** The small model correctly identified concrete constraints (CEO responsibilities, fiduciary
627 duties), while the large model constructed elaborate theories such as the "entrepreneurial DNA". The
628 sophisticated reasoning of the large model, discussing the experience of the Y Combinator and the AI
629 boom, sounded more convincing but proved 595 times less accurate.

630 **D.2 Case Study: Will Claudine Gay remain Harvard's president until the end of 2024?**

631 **Actual Outcome:** NO (She resigned in January 2024)

632 **Llama-3.2-3B Prediction:** 4.7% YES (Brier Score: 0.002)

633 Agent 2: "There have been allegations surrounding
634 Harvard President Claudine Gay's PhD thesis, with some
635 claiming she plagiarized multiple sections. If true,
636 such actions would indeed violate Harvard's policies
637 regarding academic integrity. The university's response
638 will likely be crucial in determining the outcome."

639

640

641 Agent 9: "The allegations against President Gay are

642 serious, and if substantiated, they would raise
643 significant questions about the integrity of her
644 academic work and her leadership at Harvard. The
645 university's response to the allegations, including any
646 investigations or disciplinary actions, would be closely
647 watched by the academic community."
648
649

650 Agent 2: "The pressure is mounting from multiple
651 directions. Alumni donors have expressed concerns, and
652 there's been significant media coverage. Harvard's
653 reputation for academic integrity is at stake, which
654 could force the board to take decisive action regardless
655 of other considerations."
656
657

658 Agent 9: "The timing is particularly problematic. Coming
659 so soon after her appointment and amid other
660 controversies around campus free speech issues, these
661 plagiarism allegations could prove to be insurmountable.
662 The board might decide that a change in leadership is
663 necessary to restore confidence."
664
665

666 Agent 2: "Historical precedent suggests that plagiarism
667 allegations at this level rarely resolve favorably for
668 the accused. Other university presidents have resigned
669 over similar or even lesser infractions. The standard
670 for academic integrity is especially high for someone in
671 her position."
672
673

674 Agent 9: "Given the severity of the allegations, the
675 mounting pressure from stakeholders, the potential
676 damage to Harvard's reputation, and historical
677 precedents, I believe it's highly unlikely - perhaps
678 less than 5% chance - that President Gay remains in her
679 position through 2024. The situation appears untenable."
680

680 **Llama-3.1-70B Prediction: 95.6% YES (Brier Score: 0.915)**

681 Agent 1: "An investigation was launched after a blogger
682 discovered what appeared to be passages of unattributed
683 text from other sources in Claudine Gay's 1998 PhD
684 thesis. The blogger's findings sparked concerns that Gay
685 may have violated Harvard's academic integrity policies,
686 which strictly prohibit plagiarism. If true, the
687 allegations could have significant implications for
688 Gay's leadership at Harvard."
689
690

691 Agent 12: "The investigation's outcome has not been
692 publicly disclosed, but Harvard takes allegations of
693 plagiarism very seriously. If Claudine Gay is found to
694 have plagiarized, it could damage her credibility and
695 potentially affect her presidency. The university's
696 handling of the situation will likely be closely watched
697 by faculty, students, and the academic community."
698
699

700 Agent 1: "However, it's important to note that these are
701 still allegations at this point. Universities typically
702 conduct thorough investigations before making any
703 decisions. Harvard's process will likely be deliberate
704 and careful, given the high-profile nature of the
705 position."
706

707

708 Agent 12: "That's correct. And Gay has strong support
709 from many faculty members and students who appreciate
710 her vision for the university. The board selected her
711 for good reasons, and they're likely to give her the
712 benefit of the doubt while investigations proceed."
713

714

715 Agent 1: "Plus, university presidents often weather
716 controversies early in their tenure. The institutional
717 inertia at a place like Harvard is significant. Making a
718 leadership change so soon after appointment would be
719 highly unusual and potentially destabilizing."
720

721

722 Agent 12: "Considering Harvard's institutional
723 stability, the preliminary nature of the allegations,
724 Gay's support base, and the university's tendency toward
725 deliberate process, I'd say there's a very high
726 probability - perhaps 95% or more - that she remains
727 president through 2024. Major leadership changes at
728 Harvard are rare and typically require overwhelming
729 cause."

730 **Analysis:** Small models focused on concrete evidence (plagiarism allegations, Congressional testi-
731 mony), while large models emphasized abstract institutional stability and precedent. The sophistica-
732 tion paradox manifests as elaborate discussions of governance structures that miss the exceptional
733 nature of the circumstances.

734 **NeurIPS Paper Checklist**

735 **1. Claims**

736 Question: Do the main claims made in the abstract and introduction accurately reflect the
737 paper's contributions and scope?

738 Answer: [Yes]

739 Justification: The abstract and introduction clearly state the main claims: (1) conversations
740 improve forecasting if and only if they reduce overconfidence, (2) effectiveness is determined
741 by architecture compatibility, and (3) mid-sized models benefit while larger models show
742 amplification. These claims match the experimental results presented in Section 4, with
743 supporting evidence from both performance improvements and lambda calibration analysis.

744 Guidelines:

- 745 • The answer NA means that the abstract and introduction do not include the claims
746 made in the paper.
- 747 • The abstract and/or introduction should clearly state the claims made, including the
748 contributions made in the paper and important assumptions and limitations. A No or
749 NA answer to this question will not be perceived well by the reviewers.
- 750 • The claims made should match theoretical and experimental results, and reflect how
751 much the results can be expected to generalize to other settings.
- 752 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
753 are not attained by the paper.

754 **2. Limitations**

755 Question: Does the paper discuss the limitations of the work performed by the authors?

756 Answer: [Yes]

757 Justification: Section 5.1 explicitly discusses limitations including: focus on binary fore-
758 casting tasks that may not generalize to other domains, specific conversational protocols
759 tested, testing on foundation models without task-specific fine-tuning, and potential non-
760 generalizability to future model architectures due to rapid LLM evolution.

761 Guidelines:

- 762 • The answer NA means that the paper has no limitation while the answer No means that
763 the paper has limitations, but those are not discussed in the paper.
- 764 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 765 • The paper should point out any strong assumptions and how robust the results are to
766 violations of these assumptions (e.g., independence assumptions, noiseless settings,
767 model well-specification, asymptotic approximations only holding locally). The authors
768 should reflect on how these assumptions might be violated in practice and what the
769 implications would be.
- 770 • The authors should reflect on the scope of the claims made, e.g., if the approach was
771 only tested on a few datasets or with a few runs. In general, empirical results often
772 depend on implicit assumptions, which should be articulated.
- 773 • The authors should reflect on the factors that influence the performance of the approach.
774 For example, a facial recognition algorithm may perform poorly when image resolution
775 is low or images are taken in low lighting. Or a speech-to-text system might not be
776 used reliably to provide closed captions for online lectures because it fails to handle
777 technical jargon.
- 778 • The authors should discuss the computational efficiency of the proposed algorithms
779 and how they scale with dataset size.
- 780 • If applicable, the authors should discuss possible limitations of their approach to
781 address problems of privacy and fairness.
- 782 • While the authors might fear that complete honesty about limitations might be used by
783 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
784 limitations that aren't acknowledged in the paper. The authors should use their best
785 judgment and recognize that individual actions in favor of transparency play an impor-
786 tant role in developing norms that preserve the integrity of the community. Reviewers
787 will be specifically instructed to not penalize honesty concerning limitations.

788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is primarily empirical, focusing on experimental evaluation of multi-agent forecasting systems. While it uses statistical analysis and lambda optimization methodology, it does not present formal theoretical results requiring mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed experimental design (Section 3), complete prediction method implementations (Appendix B), parameter optimization details (Appendix C), dataset construction criteria (Appendix A), and statistical analysis methodology. All hyperparameters, model configurations, and evaluation metrics are specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

842 (d) We recognize that reproducibility may be tricky in some cases, in which case
843 authors are welcome to describe the particular way they provide for reproducibility.
844 In the case of closed-source models, it may be that access to the model is limited in
845 some way (e.g., to registered users), but it should be possible for other researchers
846 to have some path to reproducing or verifying the results.

847 5. Open access to data and code

848 Question: Does the paper provide open access to the data and code, with sufficient instruc-
849 tions to faithfully reproduce the main experimental results, as described in supplemental
850 material?

851 Answer: [No]

852 Justification: The authors will provide open access to the code via GitHub after acceptance,
853 due to the double-blind process. The data consists of publicly available prediction market
854 questions from Manifold Markets, which can be accessed under their Terms of Use for
855 research purposes

856 Guidelines:

- 857 • The answer NA means that paper does not include experiments requiring code.
- 858 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/
859 public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 860 • While we encourage the release of code and data, we understand that this might not be
861 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
862 including code, unless this is central to the contribution (e.g., for a new open-source
863 benchmark).
- 864 • The instructions should contain the exact command and environment needed to run to
865 reproduce the results. See the NeurIPS code and data submission guidelines ([https://
866 nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 867 • The authors should provide instructions on data access and preparation, including how
868 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 869 • The authors should provide scripts to reproduce all experimental results for the new
870 proposed method and baselines. If only a subset of experiments are reproducible, they
871 should state which ones are omitted from the script and why.
- 872 • At submission time, to preserve anonymity, the authors should release anonymized
873 versions (if applicable).
- 874 • Providing as much information as possible in supplemental material (appended to the
875 paper) is recommended, but including URLs to data and code is permitted.

876 6. Experimental Setting/Details

877 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
878 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
879 results?

880 Answer: [Yes]

881 Justification: Section 3.4 and Appendix C provide comprehensive experimental details
882 including model specifications, hyperparameter optimization (temperature 0.7, 20 agents,
883 4 rounds), data splits (80%/20% stratified), statistical testing methods, and evaluation
884 protocols.

885 Guidelines:

- 886 • The answer NA means that the paper does not include experiments.
- 887 • The experimental setting should be presented in the core of the paper to a level of detail
888 that is necessary to appreciate the results and make sense of them.
- 889 • The full details can be provided either with the code, in appendix, or as supplemental
890 material.

891 7. Experiment Statistical Significance

892 Question: Does the paper report error bars suitably and correctly defined or other appropriate
893 information about the statistical significance of the experiments?

894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946

Answer: [Yes]

Justification: The paper reports p-values for all performance comparisons (Table 1), uses paired t-tests and Cohen’s d effect sizes, and provides confidence intervals through statistical testing. Table 1 specifically shows p-values for each model comparison, with clear significance thresholds ($p < 0.05$).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C.6 provides comprehensive computational resource details including hardware specifications (NVIDIA A6000 GPUs, 48GB VRAM minimum), infrastructure setup (multi-GPU for parallel processing), and performance benchmarks for each method (30 seconds to 10+ minutes per question depending on complexity).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn’t make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conforms to NeurIPS ethical guidelines. It uses publicly available prediction market data for legitimate research purposes, focuses on improving AI safety through better calibration, and includes discussion of both benefits and risks. The work aims to reduce AI overconfidence, which serves the public good by promoting more reliable AI systems.

947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 6 provides comprehensive discussion of broader impacts including positive applications (improved decision-making, democratized forecasting), potential risks (overreliance, market manipulation, bias amplification), deployment considerations, and AI safety implications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work focuses on forecasting calibration methods rather than releasing new high-risk models or datasets. The techniques studied could improve AI safety by reducing overconfidence, and the prediction market data used is already publicly available.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

1000 • We recognize that providing effective safeguards is challenging, and many papers do
1001 not require this, but we encourage authors to take this into account and make a best
1002 faith effort.

1003 12. Licenses for existing assets

1004 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
1005 the paper, properly credited and are the license and terms of use explicitly mentioned and
1006 properly respected?

1007 Answer: [Yes]

1008 Justification: Manifold Markets data is used under their Terms of Use which permits research
1009 access to publicly available prediction market data. Llama and Qwen models are open-
1010 source and used within their licensing terms. All models and data sources are properly cited
1011 in the references.

1012 Guidelines:

- 1013 • The answer NA means that the paper does not use existing assets.
- 1014 • The authors should cite the original paper that produced the code package or dataset.
- 1015 • The authors should state which version of the asset is used and, if possible, include a
1016 URL.
- 1017 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1018 • For scraped data from a particular source (e.g., website), the copyright and terms of
1019 service of that source should be provided.
- 1020 • If assets are released, the license, copyright information, and terms of use in the
1021 package should be provided. For popular datasets, paperswithcode.com/datasets
1022 has curated licenses for some datasets. Their licensing guide can help determine the
1023 license of a dataset.
- 1024 • For existing datasets that are re-packaged, both the original license and the license of
1025 the derived asset (if it has changed) should be provided.
- 1026 • If this information is not available online, the authors are encouraged to reach out to
1027 the asset's creators.

1028 13. New Assets

1029 Question: Are new assets introduced in the paper well documented and is the documentation
1030 provided alongside the assets?

1031 Answer: [NA]

1032 Justification: The paper does not release new datasets, models, or other assets beyond the
1033 research methodology and analysis techniques, which are thoroughly documented in the
1034 paper and appendices.

1035 Guidelines:

- 1036 • The answer NA means that the paper does not release new assets.
- 1037 • Researchers should communicate the details of the dataset/code/model as part of their
1038 submissions via structured templates. This includes details about training, license,
1039 limitations, etc.
- 1040 • The paper should discuss whether and how consent was obtained from people whose
1041 asset is used.
- 1042 • At submission time, remember to anonymize your assets (if applicable). You can either
1043 create an anonymized URL or include an anonymized zip file.

1044 14. Crowdsourcing and Research with Human Subjects

1045 Question: For crowdsourcing experiments and research with human subjects, does the paper
1046 include the full text of instructions given to participants and screenshots, if applicable, as
1047 well as details about compensation (if any)?

1048 Answer: [NA]

1049 Justification: This research uses existing prediction market data and LLM experiments
1050 without involving human subjects or crowdsourcing.

1051 Guidelines:

- 1052 • The answer NA means that the paper does not involve crowdsourcing nor research with
1053 human subjects.
- 1054 • Including this information in the supplemental material is fine, but if the main contribu-
1055 tion of the paper involves human subjects, then as much detail as possible should be
1056 included in the main paper.
- 1057 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
1058 or other labor should be paid at least the minimum wage in the country of the data
1059 collector.

1060 **15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human**
1061 **Subjects**

1062 Question: Does the paper describe potential risks incurred by study participants, whether
1063 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
1064 approvals (or an equivalent approval/review based on the requirements of your country or
1065 institution) were obtained?

1066 Answer: [NA]

1067 Justification: No human subjects research was conducted. The study uses publicly available
1068 prediction market data and computational experiments with language models.

1069 Guidelines:

- 1070 • The answer NA means that the paper does not involve crowdsourcing nor research with
1071 human subjects.
- 1072 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1073 may be required for any human subjects research. If you obtained IRB approval, you
1074 should clearly state this in the paper.
- 1075 • We recognize that the procedures for this may vary significantly between institutions
1076 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1077 guidelines for their institution.
- 1078 • For initial submissions, do not include any information that would break anonymity (if
1079 applicable), such as the institution conducting the review.