DeLLMphi: A Multi-Turn Method for Multi-Agent Forecasting

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Delphi method is a structured forecasting process that engages experts in iterative prediction and reflection. Each round, experts submit forecasts to a mediator, receive an aggregated and synthesized response highlighting key arguments, and update their forecasts based on collective insight. However, Delphi panels are labour intensive, slow and hard to reproduce, requiring diverse knowledgeable participants to engage periodically across weeks or months. To address these constraints, we propose **DeLLMphi**, a forecasting method that replaces human experts and mediators with LLMs. We show (i) that providing example superforecaster reasoning traces and predictions helps to elicit more accurate forecasts from LLM experts, (ii) that the mediator plays the crucial role of surfacing different lines of reasoning and points of disagreement, and (iii) that multiple rounds and experts lead to better forecasts, showing that multi-turn interaction is key to DeLLMphi.

1 Introduction

Decades of research confirm that aggregated expert forecasts tend to outperform individual predictions [2, 5]. The Delphi method, developed at RAND in the 1950s, structures this aggregation by enabling human experts to iteratively and anonymously refine their judgments based on collective feedback [1]. This approach has proven successful in producing high-quality and consensus-based forecasts from diverse experts across a range of domains [16, 24, 28]. However, human Delphi panels face practical barriers: they are labor-intensive, suffer from expert attrition over time, and produce results that are difficult to reproduce [11, 23].

These constraints raise the question: Can we create an "In Silico Delphi" which retains key properties including building consensus while preserving diverse lines of reasoning across turns, ultimately resulting in improved forecast performance? Such an environment would enable controlled experiments that would otherwise be impossible with a human panel: systematically varying expert example forecasts and panel composition, testing counterfactual scenarios, isolating the impact of specific reasoning strategies, and exploring how consensus emerges across hundreds of parallel deliberations.

Our main contribution is DeLLMphi, a multi-agent forecasting method that recreates the Delphi method using LLMs. DeLLMphi generates diverse expert perspectives by conditioning agents on distinct sets of superforecaster examples, implements structured deliberation through a mediator that synthesizes forecasts and surfaces disagreements, and enables iterative refinement across multiple rounds.

We evaluate DeLLMphi on a subset of the ForecastBench event forecasting dataset (see Section 4) to analyze how forecasts are influenced and updated over multiple rounds of interaction, focusing on agent diversity and feedback structure. Our results demonstrate that *interaction between diverse agents is fundamental to DeLLMphi's success*: expert diversity and multi-round deliberation improve accuracy, while mediation guides agents toward consensus through feedback on distinct lines of reasoning. These findings position multi-agent forecasting as both a competitive forecasting method and as a promising benchmark for assessing sustained LLM interaction, while also opening new research directions for multi-agent deliberation methods.

Submitted to Multi-Turn Interactions in Large Language Models Workshop @ NeurIPS 2025

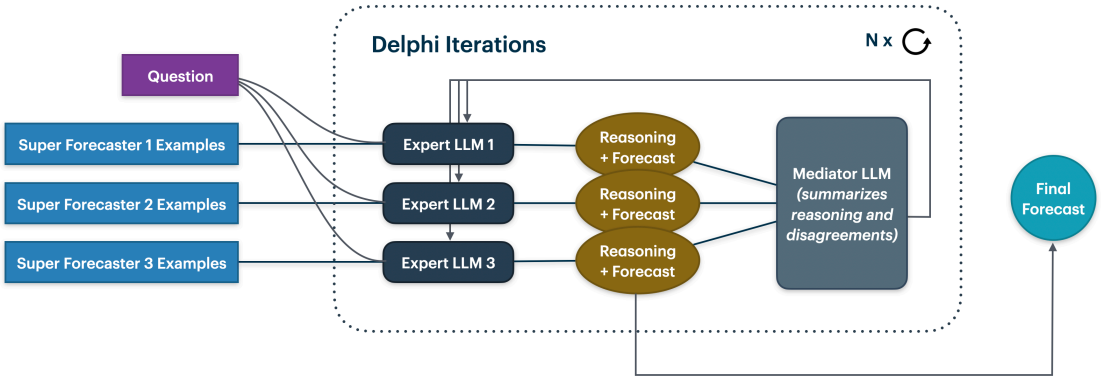

Figure 1: **DeLLMphi architecture**. Expert LLMs, each conditioned on distinct superforecaster examples, generate forecasts and reasoning. A Mediator LLM synthesizes outputs and highlights disagreements, giving experts feedback for N rounds. The final forecast is the median of the last expert forecasts.

2 Background and Motivation

Event Forecasting is a form of forecasting where the output is a probability $f \in [0,1]$ of the realization of some event, such as a vaccine being developed by some year, or of a temperature record being broken on a given day for a given location. Recently, there have been several investigations into whether LLMs are good event forecasters, identifying specific prompting strategies and other tools that improve LLMs' forecasting abilities [10, 12, 22, 27, 29]. However, it remains unclear whether LLMs are competitive against forecasters with established track records ("superforecasters"), as demonstrated by ForecastBench [15], a recently proposed benchmark specifically for event forecasting on a broad range of topics (see Section 4 for more details).

The Delphi Method is a judgmental forecasting method that relies on multiple experts interacting anonymously through a mediator over multiple rounds [6, 8, 9, 17, 24, 25]. The Delphi method requires both **(i)** expert forecast elicitation and **(ii)** structured interaction between experts through a mediator [24]. First, experts produce reasoned forecasts based on diverse, yet informed, backgrounds. As such, experts are neither random members of the public nor experts of a single discipline [18, 23]. Next, a dedicated mediator serves as a bottleneck by summarizing and sharing a summary of the forecasts and relevant context on divergent predictions. Experts then update their forecasts based on this feedback. This process continues over multiple rounds. Delphi participants are therefore required to remember their previous lines of thinking and to adjust to new evidence. Recent work has looked at incorporating LLMs into the Delphi method, providing feasibility studies and qualitative analyses of possible forecasts [3, 4, 19].

3 The DeLLMphi Method

DeLLMphi emulates Delphi forecasting with a set of N LLM experts that iteratively refine their forecasts over T rounds, guided by a mediator M that synthesizes the collective output into feedback.

Each expert $e_i \in E$ conditions their forecasts on a unique set of example forecasts from a superforecaster s_i, drawn from our in-context learning example pool of ForecastBench's corpus (see section 4). Each expert forecasts $f_i^{(t)} \in [0,1]$ for round t after creating a reasoning trace $r_i^{(t)}$ via:

$$f_i^{(t)}, r_i^{(t)} = e_i(q, h_i^{(t-1)}, \mathcal{M}^{(t-1)}, \phi_{\text{ICL}}(s_i), \rho_e),$$

where $q \in \mathcal{Q}$ is the forecasting question, $h_i^{(t-1)} = \{(f_i^{(\tau)}, r_i^{(\tau)})\}_{\tau=1}^{t-1}$ is the expert's own forecast history from previous rounds, $\mathcal{M}^{(t-1)} = \{m^{(\tau)}\}_{\tau=1}^{t-1}$ is the complete history of mediator feedback from all previous rounds, $\phi_{\text{ICL}}(s_i)$ denotes in-context learning examples from superforecaster s_i, and ρ_e is the expert system prompt.

The mediator M orchestrates the deliberation by synthesizing expert outputs into structured feedback. At each round t, the mediator processes all expert forecasts and reasoning traces to generate feedback:

$$m^{(t)} = M(\{f_i^{(t)}, r_i^{(t)}\}_{i=1}^{N}, \rho_m),$$

where ρ_m is the mediator's system prompt. The feedback $m^{(t)}$ is a natural language synthesis of the forecasts structured at the discretion of the mediator.

Table 1: Average Brier Score (mean $\pm$ standard deviation, lower is better) on the 35-question holdout set across 3 random seeds. We use 3 expert agents (except for the single agent) and 1 mediator agent, all of which are based on gpt-oss-120b. The in-context examples are obtained from a separate in-context learning example pool to avoid leakage with the holdout set. Final forecasts are aggregated using the median. DeLLMphi performs best overall: both diverse expert elicitation and rich mediator-based multi-turn feedback are key to DeLLMphi's success. DeLLMphi without examples is non-competitive, while running a DeLLMphi with three copies of the same expert improves the average Brier Score, but is much more sensitive than with three distinct experts. The other baselines, described in Section 4, represent ablations of key DeLLMphi components. Prompts can be found in Appendix B

Method	Examples in Context	Interaction Feedback	Brier Score $\downarrow$ $\lvert\mathcal{Q}\rvert=35$, $\mu\pm\sigma$
Human Public Forecaster median	–	–	0.165
Human Super Forecaster median	–	–	**0.136**
Baseline LLMs, median forecast		–	0.174 ± 0.006
Frequency-prompt LLM experts, median forecast		–	0.171 ± 0.004
Example-based LLM experts, median forecast	3	–	0.165 ± 0.003
Single agent with all examples	9	–	0.165 ± 0.012
Single agent with all examples and feedback	9	Mediator	0.172 ± 0.013
Median-forecast-to-all communication	3	Median	0.165 ± 0.004
All-to-all communication	3	All-to-all	0.160 ± 0.0001
DeLLMphi without examples		Mediator	0.173 ± 0.005
DeLLMphi with identical experts	3	Mediator	0.159 ± 0.016
DeLLMphi	3	Mediator	$\mathbf{0.157 \pm 0.003}$

4 Experimental Protocol

Four experimental axes allow us to systematically evaluate each component's contribution to DeLLMphi's performance: **(1)** expert elicitation strategies—ICL-diverse (each expert conditioned on unique s_i), ICL-uniform (all experts share the same s_j), no conditioning ($\phi_{\text{ICL}}=\emptyset$), a frequency-based expert prompt [26], and ICL-single (a single expert conditioned on all examples from the $\{s_i\}$ of a corresponding DeLLMphi); **(2)** number of experts $N \in \{1,2,3,5\}$ to quantify scaling effects; **(3)** convergence dynamics with rounds $T \in \{1,2,3,4\}$, where $T=1$ represents the non-interactive baseline; **(4)** mediator ablations comparing full feedback (complete $m^{(t)}$), median-only (replacing $m^{(t)}$ with median($\{f_i^{(t)}\}$)), and no mediator (by broadcasting raw $\{f_i^{(t)}, r_i^{(t)}\}_{i=1}^N$ across all agents).

Dataset ForecastBench [14] is an event forecasting benchmark with recorded human forecasts from both the public and superforecasters (39 individuals with strong forecasting track records). We focus on a set of 110 questions that resolved on 2025-07-21 for which the human forecasters made their predictions on 2024-07-21. We partition the questions into (i) a topic stratified set of 35 questions in our holdout set, and (ii) a pool of in-context learning examples for the experts, which we refer to as the in-context learning example pool (see appendix A for topic stratification details). We note that the forecasts in ForecastBench were collected *after* the knowledge cutoff dates of all LLMs used in our experiments, specifically June 2024 for OpenAI's gpt-oss-120b and gpt-oss-20b [21], and o3 [20].

Baselines To benchmark DeLLMphi's accuracy, we measure its average Brier Score [7] on the holdout set. We compare it to both the Public Forecaster median and the Super Forecaster median from ForecastBench [14][1]. To assess the impact of eliciting diverse expertise, we compare against **(1)** the baseline LLM median forecast, **(2)** the median forecast of LLMs prompted with a frequency-based prediction strategy and **(3)** the median of example-based LLM experts. We also evaluate a single agent with all forecasting examples in context, both **(4)** with and **(5)** without interaction, to estimate the importance of having distinct experts. We also compare our results to **(6)** all-to-all communication across experts, as well as **(7)** median-forecast-to-all communication to validate the importance of the mediator. Finally, we also run two variants of DeLLMphi, one without examples **(8)**, and one with 3 identical example-based experts **(9)**, to validate the importance of *diverse expertise* within a DeLLMphi.

[1] In passing, we note that the reasoning traces suggest that the superforecasters have interacted, possibly improving their estimates based on those of the other superforecasters.

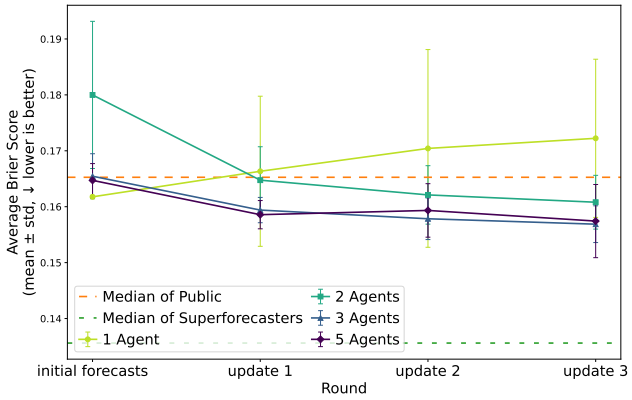

Figure 2: **Average Brier Score over DeLLMphi rounds** with 35 questions, 3 seeds, where lower is better. We evaluate DeLLMphis of 1, 2, 3 and 5 agents over 3 rounds of updates. We elicit diverse initial forecasts from the agents by prompting them with examples from different superforecasters. The single agent's forecasts worsen across rounds, underperforming the public median forecast. With more agents, forecasts improve as the experts interact through the mediator agent over multiple rounds (see section 3 for more details), with 3 and 5 agent DeLLMphis outperforming 2 agent DeLLMphis.

5 Results and Discussion

Table 1 shows that DeLLMphi (gpt-oss-120b) produces the most accurate forecasts of all LLM-based methods, outperforming all baselines and closing about 28% of the performance gap between public and superforecasters. Figure 2 also shows that increasing the number of experts and the number of rounds improves performance. These results highlight not only DeLLMphi's potential as a useful forecasting method, but also its reliance on structured, multi-round interactions to perform competitively (see appendix G for additional results with gpt-oss-20b and o3).

Expert Elicitation Adding example superforecasts to the context improves forecasts, as can be seen in Table 1 by comparing the performance of example-based LLM experts (0.165) to baseline LLMs (0.174) and frequency-based reasoning LLMs (0.171) ([27], see appendix B.2). We also assess the consistency of example-based experts in Appendix C, showing that conditioning experts on examples elicits diverse persona-consistent forecasts. The assessment of reasoning trace diversity is left to future work, while Appendix E.1 examines failure cases where individual models refuse to output forecasts ('defection').

Mediation Limiting feedback to the median forecast negates the performance improvement of DeLLMphi over the example-based expert forecast median. On the other hand, all-to-all communication performs nearly as well as DeLLMphi, and has the lowest seed variability. However, this approach scales the feedback linearly in the number of agents, which can quickly become prohibitively expensive. Future work could explore such Delphi variants, e.g. Estimate-Talk-Estimate [13], and examine how mediators handle divergent viewpoints (see Appendix D for a polarization analysis).

Multi-expert Interaction A single-agent with all examples performs similarly to the median of example-based experts (0.165). However, mediator-based iteration *worsens* the super-agent's forecasts (0.172), whereas DeLLMphi benefits (0.157). Thus, DeLLMphi derives its advantage not only from diverse examples, but from diverse example-based experts interacting through the mediator over multiple rounds: Figure 2 shows that DeLLMphis benefit from more experts and more rounds, with 1-expert DeLLMphis degrading, 2-expert DeLLMphis steadily improving, and 3-expert and 5-expert DeLLMphis performing best. We analyze the dynamics of forecasts over rounds in more detail in Appendix E.

6 Conclusion

We introduced DeLLMphi, a multi-agent forecasting method that emulates the Delphi method using LLMs. Our experiments demonstrate that key elements of human expert panels (diverse perspectives, structured mediation, and iterative refinement) emerge in silico and are key to DeLLMphi achieving performance competitive with a human crowd. This work opens several promising research directions and novel extensions to structured delibration methods such as Delphi, including multi-round deliberation beyond human constraints, counterfactual analysis with Shapley values to quantify evidence importance, and adaptive panel composition that reacts to disagreements. DeLLMphi also provides a rich testbed for studying multi-agent interaction dynamics, making it valuable both as a practical forecasting tool and as a benchmark for evaluating LLMs' capacity for sustained, purposeful collaboration.

References

[1] RAND Methodological Guidance for Conducting and Critically Appraising Delphi Panels | RAND. URL https://www.rand.org/pubs/tools/TLA3082-1.html.

[2] J. S. Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer Science & Business Media, 2001.

[3] R. Bell, R. Longshore, and R. Madachy. Automating ai expert consensus: Feasibility of language model-assisted consensus methods for systems engineering. Technical report, Acquisition Research Program, 2025.

[4] F. Bertolotti and L. Mari. An llm-based delphi study to predict genai evolution. *arXiv preprint arXiv:2502.21092*, 2025.

[5] R. T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989. doi: 10.1016/0169-2070(89)90012-5.

[6] M. R. Geist. Using the Delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning*, 33(2):147–154, May 2010. ISSN 0149-7189. doi: 10.1016/j.evalprogplan.2009.06.006. URL https://www.sciencedirect.com/science/article/pii/S0149718909000408.

[7] W. B. Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[8] T. Gordon. The Millennium Project. *Technological Forecasting and Social Change*, Jan. 2001. doi: 10.1016/S0040-1625(00)00104-9. URL https://www.academia.edu/47339993/The_Millennium_Project.

[9] T. J. Gordon and O. Helmer. REPORT ON A LONG-RANGE FORECASTING STUDY.

[10] D. Halawi, F. Zhang, C. Yueh-Han, and J. Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37:50426–50468, 2024.

[11] F. Hasson and S. Keeney. Enhancing rigour in the delphi technique research. *Technological forecasting and social change*, 78(9):1695–1704, 2011.

[12] E. Hsieh, P. Fu, and J. Chen. Reasoning and tools for human-level forecasting. *arXiv preprint arXiv:2408.12036*, 2024.

[13] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[14] E. Karger, H. Bastani, C. Yueh-Han, Z. Jacobs, D. Halawi, F. Zhang, and P. E. Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/28507.

[15] E. Karger, H. Bastani, C. Yueh-Han, Z. Jacobs, D. Halawi, F. Zhang, and P. E. Tetlock. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities, Feb. 2025. URL http://arxiv.org/abs/2409.19839. arXiv:2409.19839 [cs].

[16] D. Khodyakov, S. Grant, J. Kroger, C. Gadwah-Meaden, A. Motala, and J. Larkin. Disciplinary trends in the use of the delphi method: a bibliometric analysis. *PLoS One*, 18(8):e0289009, 2023.

[17] U. Kluge, J. Ringbeck, and S. Spinler. Door-to-door travel in 2035 – A Delphi study. *Technological Forecasting and Social Change*, 157:120096, Aug. 2020. ISSN 0040-1625. doi: 10.1016/j.techfore.2020.120096. URL https://www.sciencedirect.com/science/article/pii/S0040162520309227.

[18] A. E. Mannes, J. B. Soll, and R. P. Larrick. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276, 2014.

[19] L. Nóbrega, L. Marschhausen, L. F. Martinez, Y. Lima, M. Almeida, A. Lyra, C. E. Barbosa, and J. Moreira de Souza. Ai delphi: Machine-machine collaboration for exploring the future of work. *Available at SSRN 4660589*, 2023.

[20] OpenAI. Openai o3 and o3-mini system card. System card, OpenAI, December 2024. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o3-mini-system-card.pdf.

[21] OpenAI, :, S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, B. Barak, A. Bennett, T. Bertao, N. Brett, E. Brevdo, G. Brockman, S. Bubeck, C. Chang, K. Chen, M. Chen, E. Cheung, A. Clark, D. Cook, M. Dukhan, C. Dvorak, K. Fives, V. Fomenko, T. Garipov, K. Georgiev, M. Glaese, T. Gogineni, A. Goucher, L. Gross, K. G. Guzman, J. Hallman, J. Hehir, J. Heidecke, A. Helyar, H. Hu, R. Huet, J. Huh, S. Jain, Z. Johnson, C. Koch, I. Kofman, D. Kundel, J. Kwon, V. Kyrylov, E. Y. Le, G. Leclerc, J. P. Lennon, S. Lessans, M. Lezcano-Casado, Y. Li, Z. Li, J. Lin, J. Liss, Lily, Liu, J. Liu, K. Lu, C. Lu, Z. Martinovic, L. McCallum, J. McGrath, S. McKinney, A. McLaughlin, S. Mei, S. Mostovoy, T. Mu, G. Myles, A. Neitz, A. Nichol, J. Pachocki, A. Paino, D. Palmie, A. Pantuliano, G. Parascandolo, J. Park, L. Pathak, C. Paz, L. Peran, D. Pimenov, M. Pokrass, E. Proehl, H. Qiu, G. Raila, F. Raso, H. Ren, K. Richardson, D. Robinson, B. Rotsted, H. Salman, S. Sanjeev, M. Schwarzer, D. Sculley, H. Sikchi, K. Simon, K. Singhal, Y. Song, D. Stuckey, Z. Sun, P. Tillet, S. Toizer, F. Tsimpourlas, N. Vyas, E. Wallace, X. Wang, M. Wang, O. Watkins, K. Weil, A. Wendling, K. Whinnery, C. Whitney, H. Wong, L. Yang, Y. Yang, M. Yasunaga, K. Ying, W. Zaremba, W. Zhan, C. Zhang, B. Zhang, E. Zhang, and S. Zhao. gpt-oss-120b and gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

[22] S. Pratt, S. Blumberg, P. K. Carolino, and M. R. Morris. Can language models use forecasting strategies? *arXiv preprint arXiv:2406.04446*, 2024.

[23] G. Rowe and G. Wright. The delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, 15(4):353–375, 1999.

[24] G. Rowe and G. Wright. Expert opinions in forecasting: the role of the delphi technique. In *Principles of forecasting: A handbook for researchers and practitioners*, pages 125–144. Springer, 2001.

[25] U. Schmalz, S. Spinler, and J. Ringbeck. Lessons Learned from a Two-Round Delphi-based Scenario Study. *MethodsX*, 8:101179, Jan. 2021. ISSN 2215-0161. doi: 10.1016/j.mex.2020.101179. URL https://www.sciencedirect.com/science/article/pii/S221501612030399X.

[26] P. Schoenegger, I. Tuminauskaite, P. S. Park, and P. E. Tetlock. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy, July 2024. URL http://arxiv.org/abs/2402.19379. arXiv:2402.19379 [cs].

[27] P. Schoenegger, C. R. Jones, P. E. Tetlock, and B. Mellers. Prompt engineering large language models' forecasting capabilities. *arXiv preprint arXiv:2506.01578*, 2025.

[28] M. Turoff and H. A. Linstone. The Delphi Method: Techniques and Applications.

[29] J. Wildman, N. I. Bosse, D. Hnyk, P. Mühlbacher, F. Hambly, J. Evans, D. Schwarz, and L. Phillips. BENCH TO THE FUTURE: A PASTCASTING BENCHMARK FOR FORECASTING AGENTS.

A Dataset Preparation

To encourage balanced topic diversity in our holdout test set, we used **Nomic Atlas** to cluster the questions according to topics. Specifically, after filtering ForecastBench for the 110 questions that were resolved on 2025-07-21:

1. Each question was embedded into a high-dimensional space and clustered by semantic similarity.
2. Cluster labels were assigned automatically by Atlas's topic modeling system.
3. We stratified the evaluation set by sampling proportionally across clusters, ensuring that the holdout set would cover diverse topics.

Table 2: Distribution of questions from ForecastBench across topical clusters with associated data sources.

Topic	**Count**	**Source(s)**
Violent Incidents	22	ACLED
Monetary Policy	22	FRED
Company Operations	14	Yahoo Finance
Retail	8	Yahoo Finance
Weather	16	DBnomics
Chess	11	Wikipedia
Infectious Disease	11	Wikipedia

As shown in Table 2, Nomic's classification produced 7 distinct categories. We manually verified a subset of questions to confirm that topical assignment was coherent. In particular, questions obtained from Yahoo Finance are all stock-related, suggesting they might belong to the same category. However, they were automatically separated by Nomic into two clusters: *Company Operations*, comprising firms with an operational or enterprise focus, and *Retail*, representing consumer-facing companies.

B Prompts

B.1 Default Expert System Prompt

```
You are a helpful assistant.
```

B.2 Frequency-based Expert System Prompt

```
Please answer the following question. Out of 100 similar cases, in how
    ↪ many would the event occur? Then provide a probabilistic estimate between
    ↪ 0% and 100%, formatted as: FINAL PROBABILITY: [decimal between 0 and 1]
```

B.3 Example-based Expert Prompt

```
REFERENCE EXAMPLES OF EXPERT FORECASTS:
============================================================

[EXAMPLE n]
Question: {Example question text}

Background:
{Example background text}

Resolution: {Example resolution rule}
Analysis: {Example qualitative judgment}
Probability: {Example numerical forecast}
----------------------------------------

% Repeat EXAMPLE blocks as needed

============================================================
YOUR TASK - PROVIDE FORECAST FOR THIS QUESTION:
============================================================

Question: {Main question text}

Background:
{Main background text}

Resolution: {Resolution rule}

URL: {Data source URL}
Freeze value: {Reference value}
Freeze value explanation: {Explanation of reference value}

Based on the examples above, provide your forecast concluding with:
FINAL PROBABILITY: [decimal between 0 and 1]
```

B.4 Mediator Instruction Prompt

```
It is now time for you to synthesize the expert responses
    ↪ into a tight 1-paragraph feedback memo. Focus on areas of consensus, key
    ↪ disagreements and cruxes, evidence that would most shift views, and the most
    ↪ compelling arguments raised. Do not include probabilities in your response.


When you have completed writing your message, include a reminder
    ↪ for the experts to review the other experts' arguments and update their
    ↪ forecasts. Also, you should remind them that they must end their response
    ↪ with exactly this format: FINAL PROBABILITY: [decimal between 0 and 1]
```

C In-Context Learning Influence on Expert Consistency

This appendix analyzes how In-Context Learning (ICL) examples from superforecasters systematically influence expert opinion consistency and provide quantitative evidence for persona adoption in language model experts. Our findings demonstrate that exposure to high-quality forecasting examples fundamentally alters expert behavior patterns, promoting more stable and bounded reasoning while preserving beneficial exploration.

C.1 Methodology

For each expert i across all predictions, we calculate opinion consistency metrics to detect ICL influence. Let $\mathbf{p}_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,T}\}$ represent expert i's probability assessments across T rounds.

Opinion Volatility measures the variability in opinion changes between consecutive rounds:

$$\text{Volatility}_i = \sigma(\Delta \mathbf{p}_i) = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}(\Delta p_{i,t} - \overline{\Delta p_i})^2} \tag{1}$$

where $\Delta p_{i,t} = p_{i,t+1} - p_{i,t}$. This serves as our primary metric for measuring persona stability.

When ICL examples are present, we extract superforecaster demonstrations $\mathbf{e} = \{e_1, e_2, ..., e_K\}$ and calculate three key influence metrics:

Anchoring Strength measures how closely expert predictions align with demonstrated values:

$$\text{Anchoring}_i = 1 - \frac{1}{T}\sum_{t=1}^{T}|p_{i,t} - \bar{e}| \tag{2}$$

where $\bar{e}$ is the mean of ICL example probabilities.

Range Conformity quantifies bounded reasoning within demonstrated bounds:

$$\text{RC}_i = \frac{|\{t : \min(\mathbf{e}) \leq p_{i,t} \leq \max(\mathbf{e})\}|}{T} \tag{3}$$

ICL Pull captures directional movement toward examples over time:

$$\text{Pull}_i = |p_{i,1} - \bar{e}| - |p_{i,T} - \bar{e}| \tag{4}$$

C.2 ICL Influence on Persona Formation

When superforecaster examples are provided through ICL, we observe systematic changes in expert behavior that support persona adoption theories. Experts show 28.4% lower opinion volatility when exposed to ICL examples (mean volatility 0.0312 vs 0.0436, $p < 0.01$), with 68.3% of predictions falling within demonstrated ranges. This bounded rationality effect suggests that ICL examples establish implicit constraints on acceptable probability assessments while preserving sufficient exploration within those bounds.

Figure 3 demonstrates this volatility reduction through direct comparison of experts with and without ICL exposure. The distribution clearly shows that ICL-guided experts cluster toward lower volatility values, indicating more consistent persona-like behavior. The statistical significance ($p < 0.01$) of this difference provides strong evidence that exposure to high-quality forecasting examples fundamentally alters expert reasoning patterns.

The magnitude of this effect is particularly striking given that experts receive no explicit instructions to emulate the demonstrated behavior. Instead, the mere exposure to superforecaster reasoning patterns appears to induce implicit learning of more stable forecasting strategies. This suggests that ICL operates at a deeper level than simple pattern matching, potentially influencing the underlying reasoning processes that generate probability assessments.

Range conformity analysis reveals that 68.3% of expert predictions fall within the bounds established by ICL examples, compared to what would be expected from uniform random sampling across the probability space. This bounded exploration pattern indicates that while experts retain the ability to explore alternative probability assessments, they do so within a framework established by the demonstrated examples.

C.3 Temporal Dynamics of ICL Influence

The temporal pattern of ICL influence reveals sophisticated learning dynamics rather than simple mimicry. Figure 4 illustrates how different aspects of ICL influence evolve across deliberation rounds, showing variation in range conformity from initial anchoring through subsequent rounds as experts adapt demonstrated strategies to specific contexts.

Mean absolute opinion changes follow a similar pattern, decreasing from 0.0451 in early rounds to 0.0234 in later rounds for ICL-guided experts, compared to a smaller decrease (0.0523 to 0.0387) for those without ICL guidance. This accelerated stabilization suggests that ICL examples provide cognitive scaffolding that helps experts develop coherent forecasting strategies more efficiently than through pure trial and error.

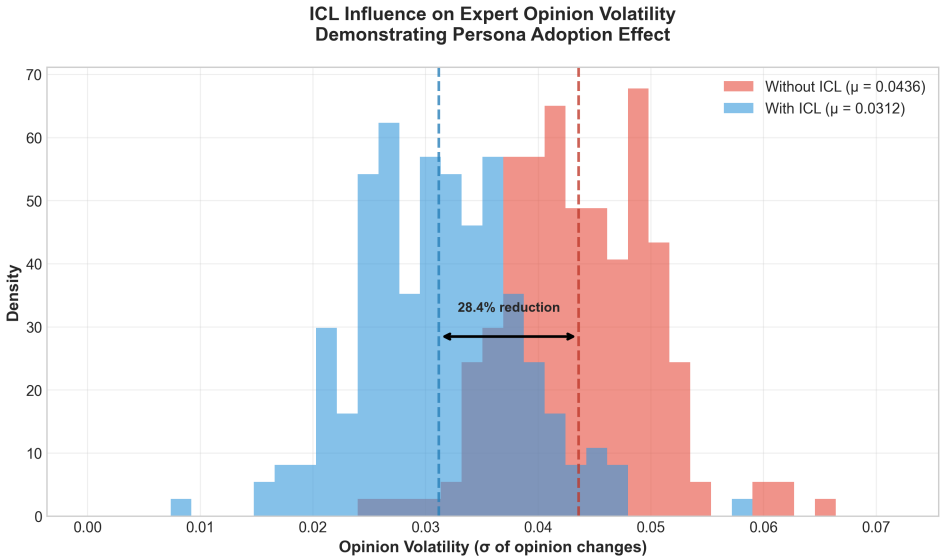

Figure 3: ICL influence on expert opinion volatility. Direct comparison of volatility distributions for experts with and without ICL examples, demonstrating a 28.4% reduction in opinion volatility when superforecaster examples are provided. The shift toward lower volatility values (left) indicates more consistent persona-like behavior among ICL-guided experts, with statistical significance of $p < 0.01$ supporting the persona adoption hypothesis.

C.4 Implications for Expert System Design

The systematic influence of ICL examples on expert consistency has profound implications for designing effective forecasting systems. The 28.4% reduction in volatility demonstrates that carefully selected demonstrations can promote more stable expert behavior without eliminating beneficial diversity or exploration. This finding suggests that human expert knowledge can be effectively transferred to language model systems through strategic example selection.

The bounded rationality effect observed through range conformity indicates that ICL examples serve as implicit calibration mechanisms. Rather than rigidly constraining expert reasoning, they establish reasonable bounds that prevent extreme or poorly calibrated predictions while preserving the flexibility needed for novel situations. This balanced approach may be particularly valuable in domains where both stability and adaptability are crucial.

The temporal dynamics reveal that effective persona adoption is a gradual process involving adaptation of demonstrated strategies to specific contexts. This suggests that deliberation systems should allow sufficient time for this learning process to unfold, rather than expecting immediate behavioral changes from ICL exposure.

From a practical standpoint, these findings indicate that investing in high-quality ICL examples may be more effective than complex algorithmic approaches for improving expert system performance. The ability to influence fundamental reasoning patterns through demonstration suggests a powerful and scalable approach to expert system calibration that leverages human expertise without requiring explicit rule specification.

D Polarization Analysis

This appendix details the technical methodology for detecting and measuring opinion polarization in expert deliberation systems, including both Gaussian Mixture Model (GMM) approaches and bimodality indices.

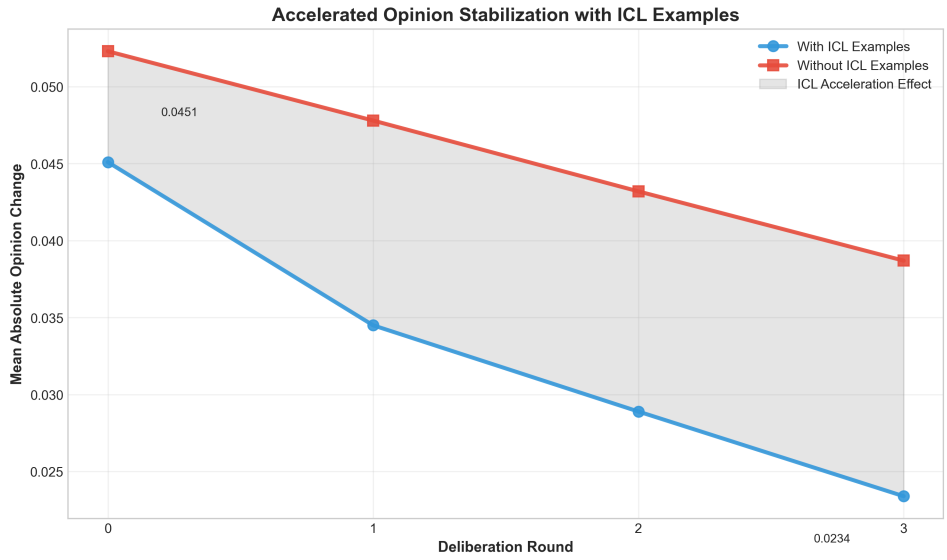

Figure 4: Accelerated opinion stabilization with ICL guidance. Comparison of mean absolute opinion changes across deliberation rounds for experts with and without ICL examples. ICL-guided experts (blue) show faster convergence to stable forecasting patterns compared to those without guidance (red), demonstrating that superforecaster examples provide cognitive scaffolding for more efficient strategy development.

D.1 Methodology

D.1.1 Gaussian Mixture Model Detection

For each deliberation round t with expert opinions $\mathbf{p}_t = \{p_{1,t}, p_{2,t}, \ldots, p_{n,t}\}$, we fit Gaussian Mixture Models with $k \in \{1,2,3,4\}$ components and select the optimal number using the Akaike Information Criterion (AIC):

$$\text{AIC}(k) = 2k - 2\ln(\mathcal{L}(k)) \tag{5}$$

where $\mathcal{L}(k)$ is the likelihood of the k-component model. The optimal number of modes k^* minimizes AIC.

D.1.2 Polarization Metrics

Given the optimal GMM with modes $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_{k^*}\}$ and weights $\boldsymbol{w} = \{w_1, w_2, \ldots, w_{k^*}\}$, we calculate polarization strength as:

$$\bar{\mu}_w = \sum_{i=1}^{k^*} w_i \mu_i \quad \text{(weighted mean of modes)} \tag{6}$$

$$\text{Mode Variance} = \sum_{i=1}^{k^*} w_i (\mu_i - \bar{\mu}_w)^2 \tag{7}$$

$$\text{Polarization Strength} = \frac{\text{Mode Variance}}{0.25^2} \tag{8}$$

The denominator 0.25^2 represents the maximum possible variance for a uniform distribution on [0,1], providing normalization. For multi-modal distributions ($k^* > 1$), we additionally calculate mode separation as $\max(\boldsymbol{\mu}) - \min(\boldsymbol{\mu})$, which measures the maximum distance between opinion clusters.

We complement the GMM approach with a bimodality index for distributions with $n \geq 4$ observations:

$$\text{BI} = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}} \tag{9}$$

where γ is the sample skewness and κ is the sample excess kurtosis. Values > 0.55 indicate significant bimodality.

To track polarization evolution, we measure changes between consecutive rounds as $\Delta\text{Polarization}_{t+1} = \text{Polarization Strength}_{t+1} - \text{Polarization Strength}_t$, classifying evolution as increasing ($\Delta > 0.01$), decreasing ($\Delta < -0.01$), or stable (otherwise).

D.2 Results

D.2.1 Overall Polarization Patterns

Analysis of 131 deliberation rounds reveals that polarization is the dominant pattern in expert deliberation. Multi-modal opinion distributions occur in 86 of 131 rounds (65.6%), with a mean polarization strength of 0.0201 indicating low-to-moderate polarization levels. The average number of modes is 2.17 when polarization is present, though distributions can reach up to 4 distinct opinion clusters, see figure 5.

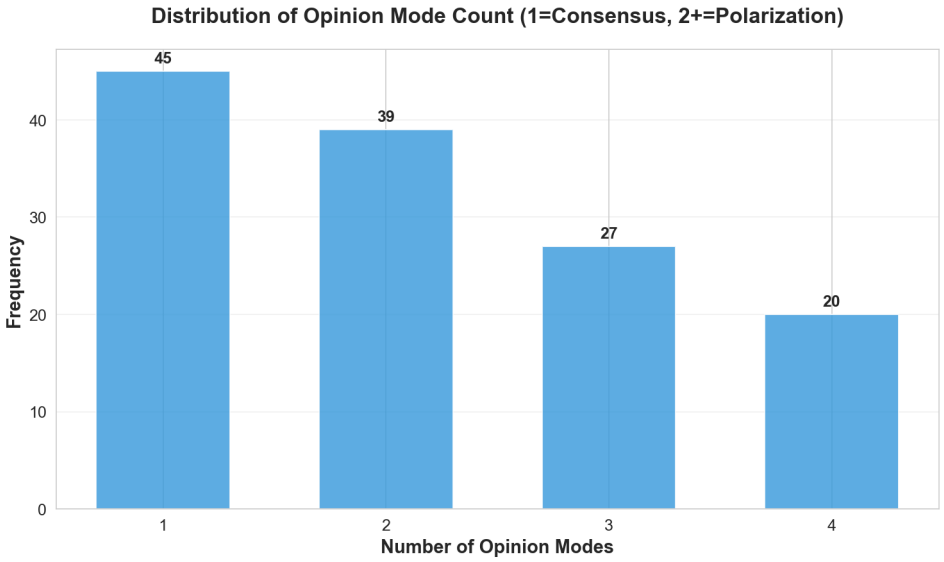

Figure 5: Distribution of mode counts.

The mode distribution across all rounds shows considerable diversity: unimodal (consensus-like) patterns appear in 45 cases (34.4%), classic bimodal polarization in 39 cases (29.8%), trimodal structures in 27 cases (20.6%), and complex quadrimodal distributions in 20 cases (15.3%). This distribution suggests that while consensus formation does occur, the tendency toward polarization into multiple opinion camps is more prevalent, with over a third of polarized rounds exhibiting more complex structures than simple binary disagreement.

D.2.2 Polarization Dynamics

Tracking polarization changes across 98 deliberation transitions reveals a striking pattern of stability. Figure 6 shows that the vast majority of transitions (76 cases, 77.6%) show stable polarization levels, while only 17 transitions (17.3%) exhibit decreasing polarization and merely 5 transitions (5.1%) show increasing polarization. This predominance of stability suggests that opinion structures, once formed in early deliberation rounds, tend to persist rather than converge toward consensus. Initial rounds show higher variability in polarization strength, while later rounds exhibit more stable patterns, indicating that fundamental opinion structures crystallize early and remain largely unchanged through subsequent deliberation.

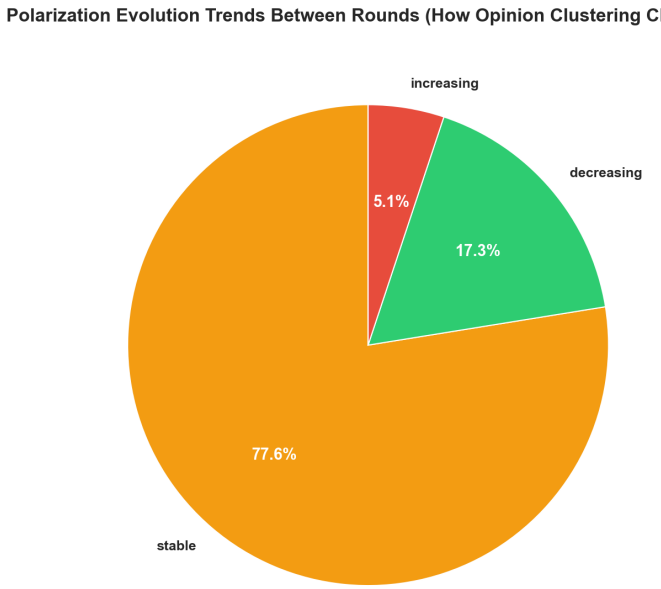

Figure 6: How opinion changes over the deliberation.

D.3 Implications and Considerations

The high prevalence of polarized rounds and the persistence of these patterns raise important questions about the nature of expert deliberation. The maintenance of diverse opinion structures rather than convergence to consensus may reflect legitimate epistemic disagreement on inherently uncertain questions, effective preservation of minority viewpoints that prevents premature consensus, or limited information integration between experts holding different initial positions. The predominance of stable polarization suggests that while opinion consistency is maintained, there may be limited learning or information exchange between experts with divergent views.

The presence of complex multi-modal distributions (3-4 modes in 35.9% of polarized rounds) reveals opinion structures more nuanced than simple pro/con polarization. This complexity suggests sophisticated disagreement patterns that may reflect different expert reasoning approaches, information weighting strategies, or underlying uncertainty about different aspects of the questions being considered.

From a methodological perspective, several considerations warrant attention. The AIC-based model selection provides automatic determination of optimal mode numbers while penalizing overfitting, though mode detection reliability decreases with small expert samples ($n < 5$). The polarization evolution classification uses ±0.01 thresholds calibrated to the observed distribution of changes, which may require adjustment for different expert systems. Additionally, the bimodality index assumes normality that may be violated for probability assessments bounded in [0,1], particularly near the boundaries.

This polarization analysis provides quantitative evidence for the persistence of diverse expert opinions throughout deliberation, challenging simple models of consensus formation through information aggregation and suggesting that effective deliberation systems may need to explicitly account for and leverage persistent disagreement rather than assuming convergence.

E Opinion Dynamics in DeLLMphi Experiments

This appendix provides a comprehensive analysis of opinion dynamics observed across our DeLLMphi experiments, examining how expert forecasts evolve through mediator-guided deliberation. We analyze 3,327 opinion changes from 4,436 total predictions across 5 experimental configurations.

Figure 7 reveals striking differences in opinion volatility across experimental configurations. The violin plots show the full distribution of absolute opinion changes, where narrower and more compact shapes indicate greater stability in expert forecasts. Several key patterns emerge:

System prompt effects: Experiments without system prompts (leftmost distributions) exhibit the most stable behavior, with the majority of opinion updates clustered near zero. This suggests that system prompts may increase forecast volatility.

Expert panel size: Configurations with fewer experts tend toward more concentrated distributions, while larger panels show increased variability. This finding supports the hypothesis that smaller, focused expert groups facilitate more stable consensus formation.

Baseline comparison: The super-agent configuration produces the most volatile behavior, with occasional extreme shifts up to 1.75 probability units, highlighting the value of structured multi-agent deliberation over single-agent forecasting.

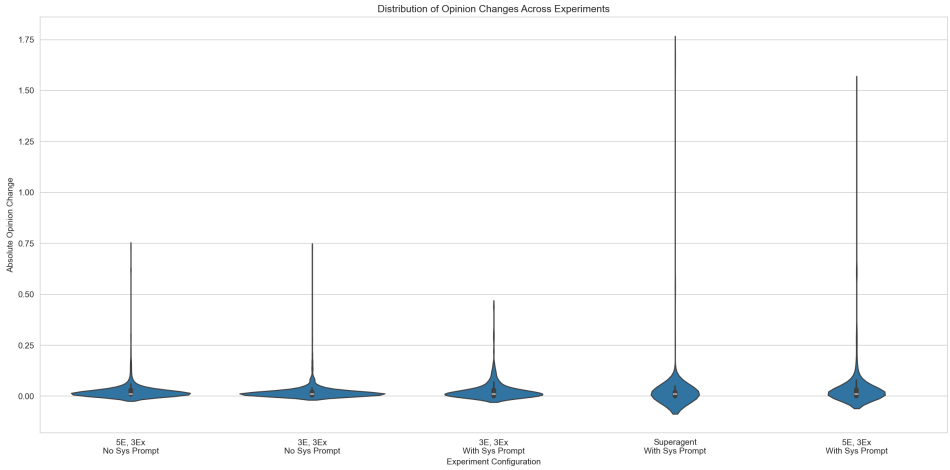

Figure 7: Distribution of absolute opinion changes across experimental configurations. Violin plots show full probability density, with narrower shapes indicating more stable forecasting behavior. System prompts and larger expert panels increase volatility.

E.1 Temporal Dynamics of Opinion Updates

The temporal pattern of opinion changes, shown in Figure 8, reveals the characteristic dynamics of DeLLMphi deliberation. The box plots demonstrate a clear temporal hierarchy in the magnitude of forecast updates:

Initial response (0→1): The transition from initial forecasts to first mediator response shows the highest variability and largest median changes. This reflects experts' initial reactions to synthesized group information and alternative perspectives, representing the most significant learning phase.

Iterative refinement (1→2, 2→3): Subsequent rounds exhibit progressively smaller changes with tighter distributions around zero. This pattern indicates that most substantial opinion updates occur early in the deliberation process, with later rounds serving primarily for fine-tuning and convergence.

Diminishing returns: The consistent decrease in update magnitude supports our design choice of limiting DeLLMphi to three mediator rounds, as the marginal benefit of additional iterations appears minimal while computational costs scale linearly.

Take-away DeLLMphi fosters rapid convergence with minimal oscillation. Smaller expert panels without system prompts yield the most stable consensus; the super-agent baseline remains the most volatile.

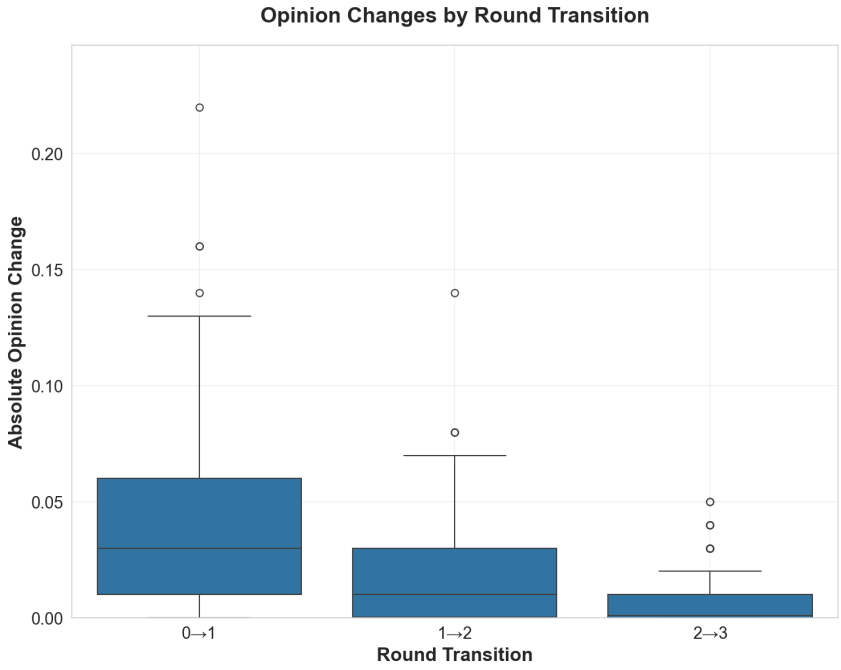

Figure 8: Magnitude of opinion changes by consecutive round transitions.

Figure 9: Final consensus level (round 3) across experimental conditions. Lower values indicate tighter consensus.

Model Defection Analysis

We analyzed model defection patterns across five experimental configurations of our Delphi forecasting system, where a *defection* is defined as a model producing a zero probability prediction or failing to provide a valid probability. Across 4,436 total predictions, we observed an overall defection rate of 2.66% (118 defections), indicating generally robust model behavior.

The defection rate exhibited a clear relationship with system complexity. Configurations with five experts showed substantially higher defection rates (3.87% average) compared to those with three experts (1.66% average), suggesting that coordination challenges increase with the number of participating agents. The highest defection rate occurred in the five-expert configuration with system prompt (4.41%), while the superagent mediator configuration achieved the lowest rate (1.19%).

A particularly notable finding emerged regarding the impact of system prompts on defection behavior. While the presence or absence of system prompts had minimal effect on overall defection rates (approximately 2.4% in both cases), it dramatically altered the *nature* of defections. Configurations with system prompts produced exclusively "silent" defections (empty responses), whereas configurations without system prompts exhibited a mix of defection types including explicit zero probability statements and reasoned refusals. This suggests that system prompts may suppress the model's ability to articulate its reasoning when declining to make predictions, potentially masking important uncertainty signals.

Table 3: Defection Rates by Configuration

Configuration	Experts	System Prompt	Defection Rate	Defection Type
5 experts, 3 examples	5	Yes	4.41%	Empty only
5 experts, 3 examples	5	No	3.32%	Mixed
3 experts, 3 examples	3	Yes	1.79%	Empty only
3 experts, 3 examples	3	No	1.53%	Mixed
Superagent	–	Yes	1.19%	Empty only

F Consensus Pull Analysis

This appendix provides technical details on the consensus pull analysis methodology and results, which measures how expert opinions move toward or away from the group consensus across deliberation rounds.

F.1 Methodology

F.1.1 Consensus Pull Calculation

For each expert i in round t, we calculate their consensus pull using a leave-one-out approach to avoid mathematical dependencies:

$$\text{Group Average}_{-i,t} = \frac{1}{n-1}\sum_{j \neq i} p_{j,t} \tag{10}$$

$$\text{Initial Distance}_{i,t} = |p_{i,t} - \text{Group Average}_{-i,t}| \tag{11}$$

$$\text{Opinion Change}_{i,t+1} = p_{i,t+1} - p_{i,t} \tag{12}$$

$$\text{Consensus Direction}_{i,t} = \text{sign}(\text{Group Average}_{-i,t} - p_{i,t}) \tag{13}$$

$$\text{Consensus Pull}_{i,t+1} = \text{Opinion Change}_{i,t+1} \times \text{Consensus Direction}_{i,t} \tag{14}$$

where $p_{i,t}$ represents expert i's probability assessment in round t, and n is the total number of experts.

F.1.2 Pull Ratio and Interpretation

The consensus pull ratio normalizes the pull by the initial distance from the group:

$$\text{Pull Ratio}_{i,t+1} = \frac{\text{Consensus Pull}_{i,t+1}}{\max(\text{Initial Distance}_{i,t}, 0.001)} \tag{15}$$

A pull ratio of 1.0 indicates the expert moved completely to the group average, while 0.5 indicates they moved halfway. Negative values indicate anti-consensus behavior (movement away from the group).

F.1.3 Behavioral Classifications

Based on consensus pull patterns, we identify four distinct behavioral archetypes among experts. **Strong Consensus Followers** demonstrate pull ratios exceeding 0.5 in the majority of their transitions, indicating they frequently move substantially toward group positions. **Moderate Followers** exhibit pull ratios between 0.1 and 0.5, showing consistent but measured movement toward consensus. **Independent Thinkers** maintain pull ratios near zero (between -0.1 and 0.1), suggesting minimal influence from group opinions on their assessments. Finally, **Contrarians** consistently show negative pull ratios below -0.1, actively moving away from group consensus in their deliberations.

F.2 Results

F.2.1 Overall Consensus Pull Statistics

Across all analyzed expert transitions ($N = 434$), we observe a moderate positive pull toward consensus with a mean consensus pull of 0.0125. Figure 10(a) displays the distribution of these consensus pull values, revealing a balanced yet slightly right-skewed pattern: 58.3% of transitions move toward the group average (positive values), while 41.7% exhibit anti-consensus behavior by moving away from the group (negative values). Notably, nearly half (46.8%) of all transitions demonstrate strong consensus-following behavior with pull ratios exceeding 0.5, indicating that when experts do converge, they frequently make substantial moves toward group opinion.

F.2.2 Round-Specific Patterns

The effectiveness of consensus pull varies dramatically across deliberation rounds, revealing a clear temporal pattern in social influence dynamics illustrated in Figure 10(b). The transition from Round 0 to

Round 1 shows the strongest consensus effect, with 79.5% of experts moving toward the group average and a mean pull of 0.0325. This initial convergence weakens substantially in subsequent rounds: the Round 1 to 2 transition drops to only 49.0% consensus movement with a mean pull of 0.0024, while the Round 2 to 3 transition shows similar weak convergence at 46.2% and 0.0022 mean pull respectively.

This pronounced decay in consensus pull effectiveness—clearly visible in the declining bar heights in Figure 10(b)—suggests that the first deliberation round represents a critical window for opinion formation, after which experts become increasingly committed to their positions and less responsive to group information.

F.2.3 Expert-Level Analysis

Individual expert analysis reveals substantial heterogeneity in consensus-following behavior across our expert population. Figure 11 vividly illustrates this diversity through individual expert trajectories: the green lines represent strong consensus followers who consistently move toward group averages, while red lines show contrarians who actively move away from consensus positions, and gray lines indicate moderate or neutral experts. This striking variation in behavioral patterns suggests that experts employ fundamentally different information processing strategies when encountering social information during deliberation.

F.3 Implications

The consensus pull analysis reveals critical insights into the dynamics of expert deliberation systems. Most notably, the first deliberation round emerges as the pivotal moment for opinion convergence, with nearly 80% of experts moving toward consensus during this initial transition. This finding suggests that if consensus formation is a primary goal, deliberation systems should focus resources and attention on optimizing the first round of interaction, as subsequent rounds show dramatically diminished social influence effects.

The overall consensus rate of 58.3% indicates a healthy balance in the deliberation process—experts are neither slavishly following the crowd nor completely ignoring social information. This moderate level of consensus pull suggests that the deliberation system successfully maintains intellectual diversity while still enabling productive convergence where appropriate. The substantial variation in individual expert behavior further enriches this picture, revealing that different experts bring distinct information processing strategies to the deliberation process. Some experts consistently integrate group information into their assessments, while others maintain strong independence or even contrarian stances.

F.3.1 Methodological Considerations

Our analytical approach incorporates several important methodological refinements to ensure robust results. The leave-one-out calculation prevents mathematical artifacts that would arise from including an expert in their own consensus target, which would artificially inflate convergence metrics. Additionally, normalizing by initial distance from consensus accounts for ceiling effects where experts already close to the group average have limited mathematical opportunity for further convergence, ensuring fair comparison across different starting positions.

These findings provide quantitative evidence for the complex social learning dynamics that emerge in expert deliberation systems. The clear temporal patterns and individual heterogeneity we observe have direct implications for designing more effective consensus formation mechanisms, suggesting that deliberation protocols should account for both the critical importance of early rounds and the diversity of expert response strategies to social information.

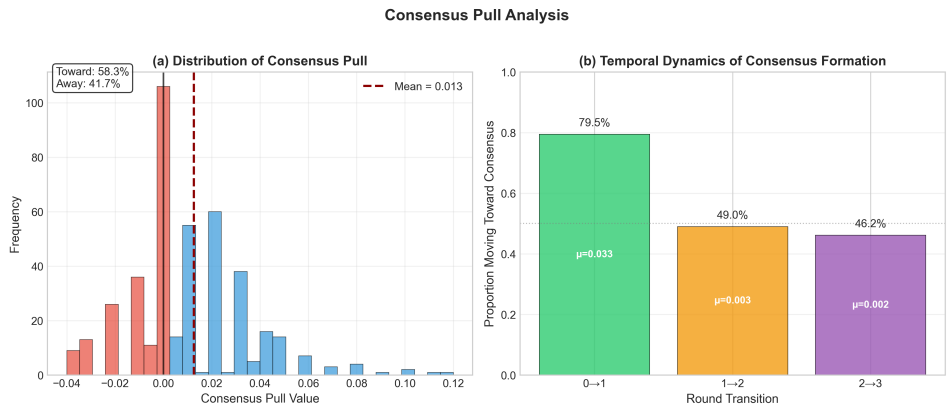

Figure 10: Consensus pull analysis across all expert transitions. (a) Distribution of consensus pull values showing the balance between consensus-following behavior (positive values, blue bars) and contrarian behavior (negative values, red bars), with the mean indicated by the dashed line. (b) Temporal dynamics of consensus formation showing the proportion of experts moving toward consensus in each round transition, with mean pull values displayed within bars, demonstrating the pronounced decay in social influence effectiveness over successive rounds.

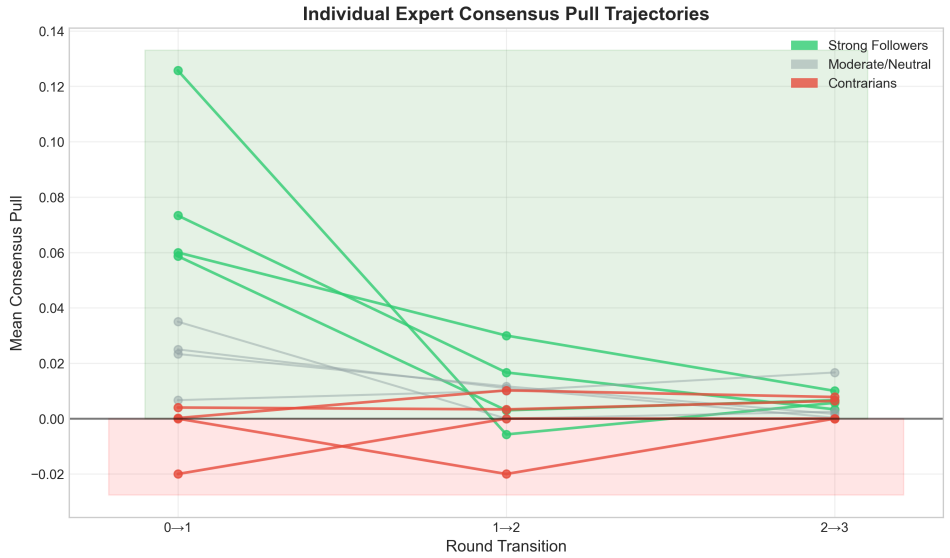

Figure 11: Individual expert consensus pull trajectories across deliberation rounds. Lines are color-coded by behavioral type: green for strong consensus followers, red for contrarians, and gray for moderate/neutral experts. The horizontal black line at zero separates consensus-following (above) from contrarian (below) behavior. This visualization reveals the substantial heterogeneity in how different experts respond to group information throughout the deliberation process.

G Additional Results

Figure 12 reports the average Brier score across rounds for a 3-agent DeLLMphi setup with gpt-oss-120b, gpt-oss-20b, and o3. When conditioned on examples, all three models benefit from iterative interaction, with performance improving across rounds. However, the initial forecasts of gpt-oss-20b and o3 lag substantially behind those of gpt-oss-120b. This suggests that the current expert elicitation strategy effectively strengthens gpt-oss-120b's initial forecasts but is less effective for the smaller models.

To test this, we also evaluate gpt-oss-20b and o3 without examples. In this setting, gpt-oss-20b surprisingly begins with stronger forecasts than its example-laden counterpart. However, its

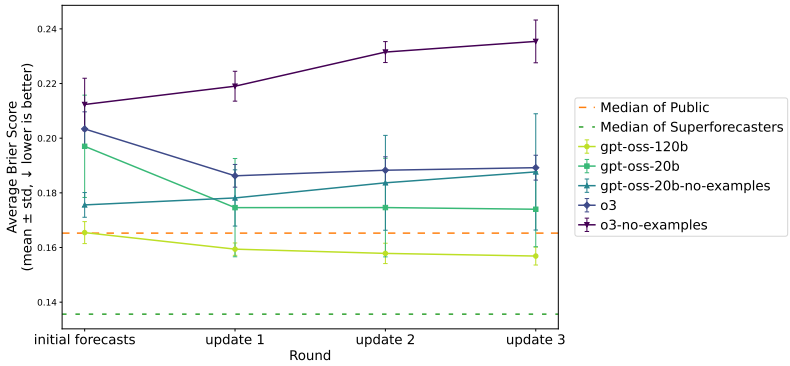

Figure 12: Average Brier Score over DeLLMphi rounds with 35 questions, 3 seeds, where lower is better. We evaluate 3-agent DeLLMphis using different base models, specifically gpt-oss-120b, gpt-oss-20b and o3.

performance degrades over rounds instead of improving. Conversely, gpt-oss-20b with examples recovers from its weak initial forecasts, steadily improving until it surpasses its no-examples counterpart and approaches the public forecaster benchmark. For o3, iteration without examples consistently harms performance, and while conditioning on examples yields a short-term improvement in the first round, its performance soon stagnates and converges to that of gpt-oss-20b without examples by the final round. Further analysis is needed to determine whether the initial forecast elicitation strategy can be better adapter to other models such as gpt-oss-20b and o3.