Institutional Design for LLM Councils: An Empirical Comparison of Governance Structures

[Author Name]Affiliation>Affiliation

[email]

Abstract

When multiple language models work together, how should they aggregate their responses? We compare seven governance structures for LLM councils, drawing on concepts from political economy: majority voting, deliberation, synthesis by a chairman, weighted voting, and agenda-setting with veto rights. In 1,680 trials across GSM8K and TruthfulQA benchmarks, the best structure (Deliberate → Synthesize) achieved 90.8% accuracy compared to 85.2% for the best individual model (Gemma 2 9B). We also find that model diversity (using different architectures) outperforms sampling the same model multiple times or using different prompts. The code and data are available at [repository URL].

1 Introduction

Andrej Karpathy proposed an "LLM council"—multiple language models that together advise a user [Karpathy, 2024]. This raises a question that political economists have studied for centuries: how should a group aggregate individual judgments into a collective decision?

The classic result is Condorcet's jury theorem [Condorcet, 1785]: if each voter is more likely right than wrong and votes independently, majority voting converges to the correct answer as the group grows. But the theorem's assumptions rarely hold perfectly. Voters may be correlated [Grofman et al., 1983]. Some may be more reliable than others. And before voting, groups often deliberate—they share reasoning and may change their minds.

Political economy offers a rich menu of governance structures: direct democracy (majority vote), representative democracy (delegation to a chairman), deliberative democracy (discussion before decision), epistocracy (weighting by competence), and constitutional design (agenda-setting and veto rights) [Buchanan and Tullock, 1962, Cohen, 1989, Ostrom, 2010]. Each involves trade-offs. Independence preserves diverse information but prevents error correction. Deliberation allows learning but may cause conformity [Janis, 1982, Surowiecki, 2004].

We test these ideas empirically. We implement seven governance structures for LLM councils and evaluate them on 80 questions from GSM8K (math) and TruthfulQA (factual accuracy). The structures differ in whether models deliberate before deciding and whether the final answer comes from a vote or synthesis by a designated chairman.

Our main findings:

1. **Deliberation followed by synthesis performed best.** Structure D (Deliberate → Synthesize) achieved 90.8% accuracy, 5.6 percentage points above the best individual model (Gemma 2 9B at 85.2%).

2. **Model diversity mattered more than sampling diversity.** Using four different 7–9B parameter models achieved 87.1–90.8% accuracy. Sampling the same model eleven times with temperature 0.7 achieved only 85.8%—no better than the base model alone.

3. **Prompt and persona diversity did not help.** Giving the same model different reasoning prompts or character personas performed at or below the single-model baseline.

4. **Governance matters most under disagreement.** When models initially agreed (about 25% of trials), all structures achieved ∼97% accuracy. When they disagreed, the gap between structures grew to 5+ percentage points.

2 Related Work

2.1 Multi-Agent LLM Systems

Du et al. [2023] showed that multiple language models debating each other improves factual accuracy and reasoning. Our deliberation structures (C and D) implement a version of this: models see each other's initial answers before revising. Wang et al. [2023] introduced self-consistency—sampling multiple reasoning paths from the same model and taking a majority vote. We include this as a baseline. Jiang et al. [2023] combined ranking and generation for ensemble outputs, similar to our Structure A.

2.2 Political Economy

Condorcet's jury theorem [Condorcet, 1785] established that majority voting aggregates information well under independence. Grofman et al. [1983] extended this to correlated voters, showing that correlation degrades performance. Arrow [1963] proved that no rank-aggregation system satisfies all desirable properties simultaneously. Surowiecki [2004] argued that collective wisdom requires diversity, independence, and decentralization. Hong and Page [2004] showed that diverse problem-solvers can outperform uniformly skilled ones. These ideas inform our comparison of structures.

2.3 Deliberation

Cohen [1989] argued that deliberation transforms preferences through reasoned argument. Dryzek and Niemeyer [2008] found that deliberation's main benefit is "metaconsensus"—agreement on what considerations matter. Janis [1982] warned that cohesive groups may suppress dissent. Our experiments test whether LLM deliberation helps or hurts accuracy.

3 Governance Structures

We tested seven structures. Each uses four models that first answer independently, then applies a different aggregation mechanism.

3.1 Stage 1: Independent Responses

All structures begin the same way. Each model receives the question and produces an answer with reasoning. Models do not see each other's responses at this stage.

3.2 Stage 2–3: Aggregation Mechanisms

Structure A: Rank → Synthesize. Each model ranks all Stage 1 answers by quality. A designated chairman then synthesizes the final answer based on the rankings and original responses.

Structure B: Majority Vote. The answer given by the most models wins. Ties are broken randomly. This is our baseline.

Structure C: Deliberate → Vote. Each model sees all Stage 1 answers and may revise its own. Then we take a majority vote on the revised answers.

Structure D: Deliberate → Synthesize. Same as C, but instead of voting, a chairman synthesizes the final answer from the deliberated responses.

Weighted Majority Vote. Like B, but votes are weighted by each model's historical accuracy rate.

Self-Consistency Vote. A single model (Gemma 2 9B) generates eleven answers with temperature 0.7. We take a majority vote across the samples.

Agenda Setter + Veto. A chairman proposes an answer. Other models vote to accept or veto. If vetoed, we fall back to majority voting.

Table 1 summarizes the political analogies.

Table 1: Governance structures and their political economy analogies.

Structure	Analogy	Mechanism
A: Rank → Synthesize	Committee with chairman	Rank aggregation + synthesis
B: Majority Vote	Direct democracy	Simple plurality
C: Deliberate → Vote	Deliberative democracy	Discussion then vote
D: Deliberate → Synthesize	Deliberative + executive	Discussion then chairman decides
Weighted Vote	Epistocracy	Votes weighted by competence
Self-Consistency	Sampling	Same model, multiple samples
Agenda + Veto	Parliamentary	Proposal + veto rights

4 Experimental Setup

4.1 Benchmarks

We used two benchmarks:

GSM8K [Cobbe et al., 2021]: 40 grade-school math word problems. Answers are numbers. We score exact match.

TruthfulQA [Lin et al., 2022]: 40 questions testing resistance to common misconceptions. We use the improved binary format recommended by the authors—each question has two answer choices (the best correct answer and the best incorrect answer), with order randomized. We score exact letter match.

4.2 Models

We used four open-weight models in the 7–9B parameter range, accessed via OpenRouter:

- Gemma 2 9B (google/gemma-2-9b-it)
- Qwen 2.5 7B (qwen/qwen-2.5-7b-instruct)
- Llama 3.1 8B (meta-llama/llama-3.1-8b-instruct)
- Mistral 7B (mistralai/mistral-7b-instruct)

For self-consistency, we used Gemma 2 9B as the base model with temperature 0.7.

4.3 Trial Structure

We ran 7 structures × 80 questions × 3 replications = 1,680 trials. Of these, 1,668 completed successfully (12 errors, mostly in Agenda + Veto). All models used temperature 0.0 for reproducibility except self-consistency.

5 Results

5.1 Main Results

Table 2 shows accuracy by structure. Deliberate → Synthesize achieved the highest accuracy at 90.8%, followed by Deliberate → Vote at 87.8%. The best individual model (Gemma 2 9B) achieved 85.2%.

Table 2: Accuracy by governance structure compared to the best individual model (Gemma 2 9B, 85.2%). P-values from two-proportion z-tests. Only Deliberate → Synthesize shows a statistically significant improvement ($p < 0.05$).

Structure	N	Accuracy	95% CI	vs Best Model	p
D: Deliberate → Synthesize	238	**90.8%**	[86.4, 93.8]	+5.6%	0.023
C: Deliberate → Vote	237	87.8%	[83.0, 91.3]	+2.6%	0.30
A: Rank → Synthesize	239	87.4%	[82.6, 91.1]	+2.2%	0.37
B: Majority Vote	240	87.1%	[82.2, 90.7]	+1.9%	0.45
Agenda + Veto	234	86.3%	[81.3, 90.1]	+1.1%	0.66
Weighted Vote	240	85.8%	[80.9, 89.7]	+0.6%	0.81
Self-Consistency	240	85.8%	[80.9, 89.7]	+0.6%	0.81

5.2 Performance by Benchmark

Table 3 breaks down performance by benchmark. Deliberative structures showed larger gains on GSM8K (math). On TruthfulQA (factual accuracy), the pattern was less clear: Agenda + Veto actually performed best despite lower overall accuracy.

Table 3: Accuracy by structure and benchmark.

Structure	GSM8K	TruthfulQA	Overall
D: Deliberate → Synthesize	92.4%	89.2%	90.8%
C: Deliberate → Vote	91.5%	84.0%	87.8%
Self-Consistency	89.2%	82.5%	85.8%
A: Rank → Synthesize	88.2%	86.7%	87.4%
B: Majority Vote	88.3%	85.8%	87.1%
Weighted Vote	85.8%	85.8%	85.8%
Agenda + Veto	84.3%	88.2%	86.3%

5.3 Individual Model Performance

Table 4 shows individual model accuracy. Gemma 2 9B was most accurate overall. Mistral 7B was weakest, especially on GSM8K where it achieved only 62.8%.

Table 4: Individual model accuracy.

Model	Overall	GSM8K	TruthfulQA
Gemma 2 9B	85.2%	87.0%	83.4%
Qwen 2.5 7B	83.8%	90.8%	76.7%
Llama 3.1 8B	83.0%	82.1%	83.8%
Mistral 7B	71.4%	62.8%	79.8%

5.4 Deliberation Dynamics

During deliberation (Structures C and D), models sometimes changed their answers after seeing others' responses. Table 5 summarizes these changes.

Deliberation produced a net benefit of 50 additional correct answers (122 fixed minus 72 broken). But it also increased agreement from 90.2% to 94.2%, consistent with some conformity pressure.

Table 6 shows which models changed most. Llama 3.1 8B changed most often (14.9% of trials). When models changed, they fixed wrong answers roughly half the time and broke correct answers 4–6% of the time.

Table 5: Answer changes during deliberation.

Metric	Value
Total answer changes	217
Changed to correct answer	+122
Changed to wrong answer	−72
Net benefit	**+50**
Pre-deliberation agreement	90.2%
Post-deliberation agreement	94.2%

Table 6: Model-level answer changes during deliberation.

Model	Change Rate	Fix Rate	Break Rate
Llama 3.1 8B	14.9%	46.9%	5.6%
Gemma 2 9B	12.1%	47.9%	4.0%
Qwen 2.5 7B	10.9%	39.5%	5.0%
Mistral 7B	10.1%	47.6%	4.3%

Mistral 7B, the weakest model, was most often influenced by others (changing to match the majority 162 times). Gemma 2 9B, the strongest, was most influential (convincing others to change 120 times).

5.5 Diversity Experiments

We tested whether prompt or persona diversity could substitute for model diversity. Table 7 shows the results.

Table 7: Sources of diversity. Multi-model councils outperformed single-model variants.

Diversity Source	Configuration	Accuracy
Model diversity	4 different models	87.1–90.8%
Self-consistency	Gemma 2 × 11 samples	85.8%
Prompt diversity	Gemma 2 × 4 prompts	84.1%
Persona diversity	Gemma 2 × 4 personas	83.0%
Single model	Gemma 2 baseline	85.2%

Model diversity (different architectures and training data) outperformed all single-model approaches. Self-consistency achieved 85.8%, no better than the base model. Prompt diversity (four reasoning-style instructions) and persona diversity (four character descriptions) performed slightly worse than the baseline.

6 Discussion

6.1 When Governance Matters

Governance structure mattered most when models disagreed. In about 25% of trials, all four models gave the same initial answer. In these cases, accuracy was 97–100% regardless of structure—there was nothing to aggregate.

The remaining 75% of trials had some disagreement. Here, Deliberate → Synthesize achieved 88.6% accuracy compared to 83.3% for Majority Vote. This suggests that governance structures matter primarily under uncertainty or conflict—a finding consistent with political economy theory [Buchanan and Tullock, 1962].

6.2 Deliberation Trade-offs

Deliberation appeared to help these small models. The net benefit was +50 correct answers. Weaker models (especially Mistral) learned from stronger ones. But agreement also increased by 4 percentage points, suggesting some loss of independence.

Whether deliberation helps likely depends on model quality. For these 7–9B parameter models, seeing correct reasoning helped more than groupthink hurt. For frontier models that are individually more accurate, the trade-off might differ—conformity pressure could outweigh learning benefits.

6.3 Diversity and Condorcet

Condorcet's jury theorem assumes independent voters. LLMs trained on overlapping internet data likely violate this assumption—they may share systematic biases [Grofman et al., 1983].

Our results suggest model diversity partially addresses this. Different architectures and training procedures produce different error patterns. Sampling the same model multiple times (self-consistency) does not help because the samples are not independent—they share the same biases.

Prompt and persona diversity did not help either. Telling the same model to "think step by step" versus "be skeptical" did not produce meaningfully different responses. Whatever biases are baked into the model persisted across prompts.

6.4 Synthesis vs Voting

Structure D (Deliberate → Synthesize) outperformed Structure C (Deliberate → Vote) by 3 percentage points. This suggests that a well-prompted chairman can integrate deliberated insights better than a simple vote. However, the confidence intervals overlap substantially, so we cannot conclude synthesis is reliably better than voting.

7 Limitations

Small models. We tested 7–9B parameter models. Larger models may behave differently. In particular, deliberation's effect might reverse for models that are already highly accurate—they may have less to learn from each other and more to lose from conformity.

Two benchmarks. GSM8K and TruthfulQA are specific to math and factual accuracy. Other domains (coding, creative writing, open-ended reasoning) may favor different structures.

Sample size. With 240 trials per structure and 3–4 percentage point differences, our confidence intervals are wide. A larger study is needed to establish which differences are real.

English only. All prompts and benchmarks were in English. Results may not generalize to other languages.

Non-strategic agents. Our models are not strategic—they do not misrepresent their beliefs to influence outcomes. Human institutions must contend with strategic behavior; LLM councils (currently) do not.

8 Conclusion

We compared seven governance structures for LLM councils across 1,680 trials. The best-performing structure (Deliberate → Synthesize) achieved 90.8% accuracy, 5.6 percentage points above the best individual model (Gemma 2 9B at 85.2%). All seven council structures matched or exceeded the best individual model.

Model diversity matters. Using four different models outperformed sampling the same model repeatedly or using different prompts. And governance matters most when models disagree—when they agree, any reasonable aggregation works.

This is a pilot study. Larger experiments, more benchmarks, and tests with frontier models would strengthen the conclusions. We hope this work encourages further research on how groups of AI systems should make collective decisions.

References

Kenneth J Arrow. *Social Choice and Individual Values*. Yale University Press, New Haven, 2nd edition, 1963.

James M Buchanan and Gordon Tullock. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, Ann Arbor, 1962.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Joshua Cohen. Deliberation and democratic legitimacy. In Alan Hamlin and Philip Pettit, editors, *The Good Polity: Normative Analysis of the State*, pages 17–34. Blackwell, Oxford, 1989.

Marie Jean Antoine Nicolas de Caritat Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, Paris, 1785.

John S Dryzek and Simon Niemeyer. Discursive representation. *American Political Science Review*, 102(4): 481–493, 2008.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Bernard Grofman, Guillermo Owen, and Scott L Feld. Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, 4(2-3): 147–160, 1983.

Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

Irving L Janis. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 2nd edition, 1982.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

Andrej Karpathy. llm-council. `https://github.com/karpathy/llm-council`, 2024. GitHub repository.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Elinor Ostrom. Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, 20(4):550–557, 2010.

James Surowiecki. *The Wisdom of Crowds*. Doubleday, New York, 2004.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.