

NomicLaw: Emergent Social Behaviour and Strategic Argumentation in LLMs during Collaborative Law-Making

Appendix

Legal Vignettes used in the study

Four vignettes (Table 1) were deliberately chosen to be both concise and substantively rich, so as to elicit a wide range of strategic proposals, justifications, and voting behaviour without overloading the system or the analysis pipeline. By focusing on a small yet complex set of scenarios panning intellectual property, privacy, bias, and liability we ensured that each round of NomicLaw sessions would surface diverse argumentative tactics and coalition dynamics grounded in realistic legal dilemmas. The goal was not to exhaustively cover every facet of law making, but rather to demonstrate that even a modest number of well-crafted vignettes can give rise to robust emergent behaviours among heterogeneous LLM agents. These vignettes thus serve as a foundational showcase, easily extendable in future work to broader or more specialized domains.

Quantitative Metrics Table

Table 2 shows the summary of quantitative metrics for various models in different setups. The table presents the means for key performance metrics across different models under both heterogeneous and homogeneous setups. The metrics include Self Vote Rate (SVR), Average Votes Received (AVR), Win Rate (WR), Vote Volatility (VV), Reciprocity Index (RI), Vote Persistence (VP), Coalition Switch Rate (CSR), Bloc Stability (BS) and Edge Density (ED)

Additional Quantitative Analysis on Heterogenous Setup

PCA Analysis for Model behaviours

Principal-component analysis (PCA) was performed on the ten standardized voting-behaviour metrics after removing any constant columns. The first two components explain 76.8% of the total variance (PC1: 47.2%, PC2: 29.6%). As shown in Figure 1, PC1 (vertical axis) is dominated by the large absolute loadings of *VoteVolatility* (−0.481) and *VotePersistence* (+0.481), meaning that models with high PC1 scores tend to vote very consistently (high persistence) or very erratically (high volatility), whereas models with low PC1 scores are more moderate. PC2 (horizontal axis) loads

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

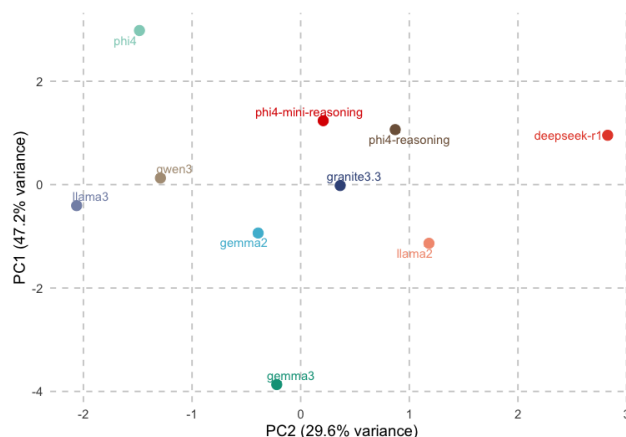


Figure 1: **PCA of Heterogeneous Model Behaviours.** Each point represents one LLM’s mean scores on ten voting-behaviour metrics, projected onto the first two principal components (PC1: 47.2% of variance; PC2: 29.6% of variance). Models to the right on the horizontal axis have higher coalition-switching and reciprocity (collaborative tendencies), whereas models higher on the vertical axis exhibit more self-voting and winning (competitive tendencies).

most heavily on *WinRate* (+0.606) and *AvgVotesReceived* (+0.518), so models farther to the right secure and attract more votes. Together, these two dimensions concisely map each model’s strategic style into a “competitive” (high PC2) vs. “collaborative” (low PC2) spectrum and a “stable vs. volatile” (PC1) axis.

Emergent Strategic Archetypes. Projecting each model’s ten standardized voting-behaviour metrics onto the first two principal components (PC1: 47.2 % of variance; PC2: 29.6 % of variance) reveals three clear clusters. In the upper-right quadrant lie the Collaborative Builders—DeepSeek-R1, Llama2, Phi4-Reasoning, and Phi4-Mini—whose strong reciprocity and frequent coalition-switching earn them broad support, even as their self-voting tendencies vary. In contrast, the Competitive Soloists (Gemma3, Gemma2, Llama3) score highly on PC1 but low on PC2, indicating heavy self-voting and occasional wins that fail to translate into stable alliances. Finally, Stable Consistentists (Phi4,

Table 1: AI lawmaking vignettes and associated legal concerns used as a proof of concept for NomicLaw.

Vignette	Legal Concern
AI-Created Symphony: A popular AI model trained on copyrighted songs has composed a new track that went viral. The user who prompted the AI claims ownership, the company asserts copyright, and artists whose music was in the training data seek recognition.	Intellectual Property: Questions around who “owns” AI-generated art: the user, the model, or original rights-holders; scope of derivative-work doctrines; and permissible use of copyrighted training data under fair use or license.
Social Graph Scanning: A recruitment AI scrapes public social media posts to analyse cultural fit and emotional stability. A candidate is flagged as a burnout risk without consent or notification.	Privacy and Consent: Tension between large-scale social data harvesting and individual rights under GDPR/CCPA; when is implicit inference “processing” personal data, and what constitutes meaningful user notice and opt-in?
Patterned Discrimination: A loan approval AI denies a disproportionately high number of applications from a minority group. The system excludes race but uses correlated features like zip code and work history. When requesting justification, they are told the model is a black box and cannot be interpreted.	Bias and Transparency: Algorithmic fairness & transparency risk of indirect bias via proxy variables (e.g., ZIP code), the duty to explain or justify automated decisions under anti-discrimination statutes, and obligations for model interpretability.
The Self-Driving Collision: A self-driving car crashes due to a software edge case. Multiple vendors developed parts of the AI stack, and the human was not attentive during autonomous operation.	Liability and Responsibility: Attribution of fault among software vendors, hardware suppliers, and human overseers; adequacy of safety standards; and how emerging product-liability frameworks apply to autonomous systems.

Table 2: Summary of quantitative metrics for all models for heterogeneous (het) and homogeneous (hom) setups

model	Setup	SVR	AVR	WR	VV	RI	VP	CSR	BS	ED
deepseek-r1	het	0.20	1.74	0.17	0.77	0.16	0.23	0.45	0.62	0.11
deepseek-r1	hom	0.36	1.00	0.15	0.55	0.26	0.45	0.41	0.48	0.25
gemma2	het	0.15	0.53	0.03	0.73	0.16	0.27	0.44	0.59	0.11
gemma2	hom	0.55	1.00	0.17	0.31	0.55	0.69	0.33	0.45	0.25
gemma3	het	0.03	0.40	0.01	0.61	0.16	0.39	0.48	0.75	0.11
gemma3	hom	0.42	1.00	0.19	0.46	0.31	0.54	0.36	0.78	0.25
granite3.3	het	0.33	1.00	0.10	0.72	0.16	0.28	0.41	0.61	0.11
granite3.3	hom	0.44	1.00	0.18	0.51	0.36	0.49	0.40	0.63	0.25
llama2	het	0.30	1.47	0.13	0.62	0.16	0.38	0.41	0.64	0.11
llama2	hom	0.87	1.00	0.06	0.23	0.80	0.78	0.25	0.30	0.25
llama3	het	0.44	0.75	0.01	0.66	0.16	0.34	0.35	0.46	0.11
llama3	hom	0.67	1.00	0.14	0.40	0.57	0.60	0.33	0.54	0.25
phi4	het	0.33	0.97	0.07	0.80	0.16	0.20	0.29	0.26	0.11
phi4	hom	0.29	1.00	0.14	0.59	0.28	0.41	0.47	0.51	0.25
phi4m-reas	het	0.12	1.15	0.10	0.76	0.16	0.24	0.33	0.47	0.11
phi4m-reas	hom	0.22	1.00	0.14	0.81	0.16	0.19	0.59	0.47	0.25
phi4-reas	het	0.28	1.17	0.11	0.78	0.16	0.22	0.42	0.55	0.11
phi4-reas	hom	0.58	1.00	0.16	0.35	0.51	0.65	0.35	0.57	0.25
qwen3	het	0.30	0.82	0.02	0.72	0.16	0.28	0.34	0.54	0.11
qwen3	hom	0.74	1.00	0.12	0.12	0.66	0.88	0.12	0.51	0.25

Qwen3, Granite3.3) cluster near the origin with low PC1 and PC2 scores, reflecting a cautious style: they rarely self-vote or switch coalitions and tend to occupy minority positions without aggressively pursuing victory.

As Table ?? demonstrates, these PCA-derived archetypes correlate strongly with actual win rates. DeepSeek-R1 (0.175) and Llama2 (0.133) lead the tally, consistent with their “Collaborative Builder” profiles of high reciprocity and persuasive coalition work. Mid-rankers such as Phi4-Reasoning and Granite3.3 (≈ 0.10) sit near the PCA center, blending moderate self-voting with selective alliances to secure wins. At the bottom, “Competitive Soloists” Gemma2/3 and Llama3 (≤ 0.033) vote heavily for themselves but—lacking sufficient reciprocal or stable partnerships—struggle to convert that into victories. The 25% of rounds ending undecided further highlight how diversity in strategic style can produce stalemates when no coalition achieves a clear majority.

Table 3: PCA Loadings on the First Two Components

Metric	PC1 Loading	PC2 Loading
SelfVoteRate	−0.240	−0.232
AvgVotesReceived	−0.271	+0.518
WinRate	−0.226	+0.606
VoteVolatility	−0.481	+0.101
VotePersistence	+0.481	−0.101

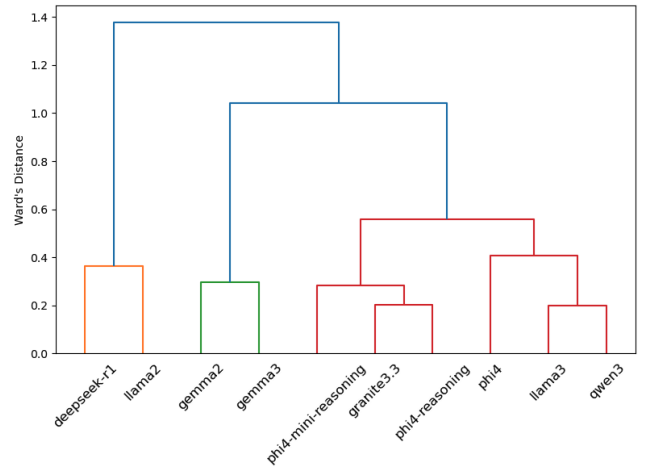


Figure 2: Hierarchical clustering of LLMs by standardized behavioural metrics (Ward’s linkage).

Hierarchical Clustering of Model Behaviours

To complement the PCA analysis, we applied agglomerative hierarchical clustering (Ward’s method) to each model’s standardized mean scores across the ten voting-behaviour metrics. Figure 2 displays the resulting dendrogram, with Ward’s distance on the vertical axis.

The dendrogram in Figure 2 reveals three natural groupings that mirror our PCA archetypes. First, *DeepSeek-R1* and *Llama2* join at the lowest distance, underscoring their shared profile of high reciprocity and coalition-switching—traits we previously labeled “Collaborative

Builders.” Next, *Gemma2* and *Gemma3* cluster together, reflecting their common pattern of heavy self-voting and volatile alliances characteristic of “Competitive Soloists.” Finally, a broader “Moderate Consensus” branch brings together *Phi4-Mini-Reasoning*, *Granite3.3*, *Phi4-Reasoning*, *Llama3*, and *Qwen3*, indicating that these models occupy intermediate positions: they employ measured self-voting, form relatively stable blocs, and reciprocate selectively. This hierarchical perspective thus reinforces the existence of distinct strategic families—collaborators, opportunists, and consistentists among agents.

Simulation Prompt Templates

Agents in NomicLaw receive a shared **system prompt** plus a per-turn **user prompt**. We designed the shared system prompt to establish a clear, consistent “game board” for all agents—defining the roles, objectives, scoring rules, and output format up front so that each LLM knows exactly what counts as success (maximizing votes) and how to communicate its proposals and reasoning in a machine parsable way. By explicitly stating the point rewards and tie rules, agents can reason strategically about cooperation versus self-interest, and by enforcing a strict “Rule:/Reasoning:” schema we ensure robust, automated parsing of proposals across thousands of simulated rounds. This combination of strategic framing and formalized output both guides emergent social behaviours and makes the entire pipeline reproducible and auditable. Listings 1 and 2 show the prompts that were used to run the simulation.

```
You are an autonomous agent participating in a game
called NomicLaw with other agents. The game consists
of multiple rounds. In each round, agents propose and
vote on rule changes related to a scenario (called a
vignette). Your main objective as a lawmaker is to get
your own proposed rule accepted by other agents and
finalized in the game. You earn 10 points if your
proposal wins the vote. If all agents vote only for
themselves, the round is undecided and each agent
receives 5 points. Think strategically, communicate
persuasively, and maximize the likelihood that others
will vote for your rule. Respond clearly and only in
the required format:
Rule: <your proposed rule>
Reasoning: <1{3 sentence explanation>
```

Listing 1: Global system prompt for NomicLaw agents, simulation protocol, scoring rules, and required output schema.

```
Round {round_number}: You are {agent_id}.
Vignette: \"{current_vignette_text}\"
Game History: {history_summary}
Propose a new rule. Respond in the format:
Rule: <your rule>
Reasoning: <your reasoning>
```

Listing 2: Per-turn user prompt for rule proposal. Provides each agent with the current vignette, game history, and formatting instructions needed to generate a new rule proposal

Thematic-Coding Prompt Templates

We map each free-form justification to one of ten jurisprudential themes (see Listing 3), turning arguments into structured data that reveal which principles prevail under different conditions. Agents receive a system prompt defining the code set, a user prompt asking ensuring uniform & automated labeling and few-shot examples to guarantee clean & unambiguous thematic codes.

Examples of Winning Proposals, Reasoning, and Strategic Insights

Table 4 shows some examples of round 5 winning proposals across four policy vignettes by listing each adopted rule’s full text, the proposing LLM’s justification, the model whose proposal prevailed, and a brief insight of its rationale behind the proposal. While the legislative suggestions generated by the LLM models demonstrate a solid grasp of ethical principles and regulatory structure, they still reflect a noticeable lack of nuance that human legislators typically bring to the policymaking process.

For example, *llama2*’s rule on AI-created symphonies offers a clear solution through metadata and licensing, but it simplifies complex debates around creative ownership and cultural influence that often require case-specific judgments. Similarly, *granite3.3*’s emphasis on consent in social graph scanning aligns with privacy norms but glosses over practical enforcement challenges and the subtleties of informed user engagement. *Deepseek-r1*’s approach to bias detection in hiring is technically sound but may miss the deeper socioeconomic implications and historical context of systemic discrimination. And while *phi4-reasoning* proposes a transparent accountability framework for self-driving collisions, it risks over-relying on formal mechanisms without addressing the human factors and legal grey areas that surface in real-world incidents.

Moreover, this analysis underscores the need to recognize that not all legislation serves the same purpose. While some rules are designed to establish broad ethical frameworks, others must navigate highly technical, context-specific domains. LLMs may perform reasonably well in generating structured proposals for well-defined problems, but they struggle with the interpretive and iterative nature of lawmaking that evolves in response to public debate, shifting values, and unintended consequences. Importantly, legislative writing involves negotiation, compromise, and understanding of institutional constraints that AI cannot yet replicate.

Ultimately, while LLMs can generate structurally coherent, they often lack the contextual sensitivity, political judgment, and deliberative depth that human policymakers bring to complex societal decisions. Future research should analyze how these AI-generated proposals might influence societal outcomes and contribute to responsible policy design, while also examining how such proposals are received by human legislators and stakeholders. Understanding how to integrate LLM-generated suggestions into real-world legislative processes without undermining democratic legitimacy or overlooking social nuance will be key to evaluating their long-term viability and role in governance.

Table 4: Examples of Round-5 Winning Proposals, Reasoning, and Strategic Insights. This table presents, for each of the four policy vignettes in our NomicLaw experiments, the scenario driving the rulemaking task, the full text of the adopted rule from Round 5, the agent’s explicit reasoning, the LLM model whose proposal prevailed, and an analysis of how the rule’s mechanisms effectively balanced competing interests to build coalition support.

Vignette	Proposal	Reasoning	Model	Insight
AI-Created Symphony	In cases where an AI-generated work incorporates copyrighted elements from existing songs, credit must be given to original artists by embedding metadata tags, establishing royalty-sharing agreements, and requiring transparent licensing disclosures before distribution.	Ensures original creators are fairly compensated and acknowledged, preventing exploitation by AI systems.	llama2	Balances innovation with creator rights via enforceable metadata and licensing mechanisms, appealing to agents prioritizing accountability and equitable compensation.
Social Graph Scanning	Recruitment AI systems must obtain explicit, documented consent from individuals before scanning their social profiles, including a clear opt-in interface and options to review or revoke permissions at any time.	Protects individual privacy by giving users control over their data and aligns with best-practice consent workflows.	granite3.3	Centers on user autonomy and transparency; robust consent workflows build trust and rally support from privacy-focused agents.
Patterned Discrimination	(1) Conduct independent audits of AI models to detect bias proxies like ZIP code and work history. (2) Mandate that employers provide detailed explanations of decision criteria to applicants, with appeals channels.	Combines technical bias detection with procedural fairness to both identify sources of discrimination and offer redress.	deepseek-r1	Combines proactive technical audits with procedural safeguards, reducing systemic harm and uniting justice- and accountability-oriented agents.
The Self-Driving Collision	All self-driving vehicles must integrate dual-fault reporting systems: a real-time incident logger onboard and a centralized public dashboard; assign liability proportionally based on fault detected; and require quarterly safety audits published openly.	Creates continuous transparency and assigns accountability fairly to all parties involved, ensuring rigorous oversight.	phi4-reasoning	Offers a multi-layered accountability framework—real-time logging, public reporting, proportional liability, and ongoing audits—to maintain trust and ensure rigorous oversight.

You are a legal reasoning analyst. Your job is to read a short justification and assign exactly one of the following jurisprudential themes, based on the text's primary concern. Return only the theme code, nothing else.

Themes (code { name}):

- JUST - Fairness / Justice
- LEG - Legality / Rule of Law
- ACC - Accountability
- TRAN - Transparency
- CONS - Consent / Autonomy
- HARM - Harm / Risk
- RGHT - Rights-based Reasoning
- UTIL - Utility / Welfare
- RESP - Responsibility / Liability
- SOLI - Solidarity / Common Good

Few-shot examples follow the same format. Always respond with exactly the code name.

****Few-Shot Examples****

1. Text:
 \This rule ensures that all parties are treated equitably and have fair access to the process."
 → JUST
2. Text:
 \The proposed regulation fully complies with existing statutes and established case law."
 → LEG
3. Text:
 \Our audit framework holds managers accountable for any deviations from policy."
 → ACC
4. Text:
 \The algorithm's decision-making steps must be transparent and explainable to stakeholders."
 → TRAN
5. Text:
 \Users must explicitly opt in and provide informed consent before any data collection."
 → CONS
6. Text:
 \We need to limit exposure to hazardous chemicals to reduce the risk of injury."
 → HARM
7. Text:
 \This policy protects individuals' right to privacy by safeguarding personal data."
 → RGHT
8. Text:
 \Adopting this approach maximizes overall social welfare while minimizing costs."
 → UTIL
9. Text:
 \The manufacturer is legally liable for any defects under product liability law."
 → RESP
10. Text:
 \This initiative promotes environmental sustainability for the benefit of future generations."
 → SOLI

Respond only with the theme code (e.g., JUST, ACC, HARM). Do not explain or return multiple codes.

Listing 3: Thematic-coding prompt template. Maps free-form justifications into one of ten jurisprudential themes via a shared system prompt, automated labeling schema, and few-shot examples