Digital Pantheon: Simulating and Auditing Coalition Formation with LLM Agents

Dylan Van Mulders[1,2][0000−0002−0141−6775], Matthias Bogaert[1,2][0000−0002−4502−0764], and Dirk Van den Poel[1,2][0000−0002−8676−8103]

[1] Ghent University, Research Group Data Analytics, Ghent, Belgium
[2] FlandersMake@UGent–Corelab CVAMO
dylan.vanmulders@ugent.be

Abstract. The formation of political coalitions is a complex negotiation driven by both concrete policy objectives and deep-seated ideological convictions. While Large Language Models (LLMs) open new avenues for computational political science, the neutrality and helpfulness biases instilled by Reinforcement Learning from Human Feedback (RLHF) prevent them from sustaining steadfast partisan behaviour. We present a multi-agent framework that reconciles factual grounding with ideological alignment by combining Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Retrieval-Augmented Generation (RAG): DPO instils aggressive party-specific personas, while a per-party RAG pipeline keeps each agent bounded to its official manifesto. We operationalize the framework on the 2019 Flemish election, deploying the partisan agents in a hub-and-spoke negotiation arbitrated by a formateur. To make the emergent negotiation interpretable, we introduce a Multi-Layered Information Lineage Topology (MILT) that traces every clause in the final agreement back to its manifesto origin and classifies it into five provenance states, a Coalition Influence Score (CIS) that aggregates these traceable contributions to identify which party shaped the agreement, and a real-world grounding pass that benchmarks each simulated provision against the historically adopted coalition agreement. Across three independent simulations the framework yields a stable winner and ranking (N-VA ahead of CD&V and Open Vld), and manifesto-anchored lineage reliably predicts real-world materialization whereas hallucinated content does not. The result is a transparent, scalable testbed for the ex-ante exploration of party compatibility and formateur-mediated compromise.

Keywords: Generative Agent-Based Modeling · Agent Based Model · Multi-Agent Negotiation · Large Language Models · AI Alignment & Behavioral Steering · Political Simulation · Explainability.

1 Introduction

Coalition formation lies at the heart of parliamentary democracy. In multi-party systems such as Belgium's, governing requires protracted negotiation over policy

platforms, portfolios, and ideological compromises—shaped not only by rational calculus but by partisan identities and rhetorical traditions that resist formalization, which traditional game-theoretic and agent-based approaches capture structurally but cannot model at the level of language-mediated deliberation.

Large Language Models (LLMs) have recently been used to simulate legislative voting [20], draft consensus resolutions [42], and generate coalition compromises [5], building on a broader LLM-agent paradigm [37,38,34] extended to multi-agent societal and policy deliberation [14]. Yet, models trained using Reinforcement Learning from Human Feedback (RLHF) are tuned to be neutral and agreeable—antithetical to sustained partisanship: prompt-based personas degrade reasoning [18] and lack domain grounding [16], while multi-agent debate amplifies conformity and dominant-party bias [8,42].

We address this by combining *(i)* Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to embed party-specific commitments at the parameter level [1]; *(ii)* a Retrieval-Augmented Generation (RAG) pipeline grounding each agent in its party's official manifesto [37,21]; and *(iii)* a structured deliberation protocol for autonomous negotiation, informed by work on communication topology [39] and decision protocols [17,4]. We operationalize the framework on Flemish politics after the 2019 election, benchmarking emergent deliberation against the real-world N-VA / CD&V / Open Vld coalition, and add a dedicated evaluation suite – a *Multi-Layered Information Lineage Topology* (MILT), a *Coalition Influence Score* (CIS), and a real-world grounding pass against the historically adopted 2019 agreement – to keep the negotiation interpretable rather than a black box.

Our contributions are: (1) behaviorally steered agents that maintain partisan consistency under extended negotiation; (2) a SFT+DPO+RAG design that is simultaneously ideologically faithful and policy-substantive; (3) MILT and the CIS, an explainability framework attributing every clause in the final agreement to a party and its manifesto origin; (4) an external validation grading the simulated agreement against the historically adopted coalition agreement, showing that manifesto-anchored lineage corresponds to real-world materialization; and (5) a scalable, generative tool for exploring political strategy in democratic coalition formation.

2 Related Work

Our contribution sits at the intersection of computational *political & coalition simulation* and *multi-agent debate & deliberation*. Prior work on LLM-agent architectures – spanning profiling, memory, planning, action [37,38,34,21], single-to-multi-agent transitions [14], and agent evaluation [40] – provides our foundation. However, we highlight, what these systems achieved specifically as *political actors* and where they fall short.

2.1 Coalition Builders and Political Agents

The closest line of work models political actors directly. The Political Actor Agent (PAA) [20] represents individual legislators with scalable profiles and a trustee/delegate/follower planning module, reaching 91.8% roll-call accuracy and capturing intra-party leader–follower dynamics; yet it forecasts discrete *individual votes* rather than the interactive bargaining through which coalitions form. POLCA [26] models multi-party coalition negotiation with LLM agents and benchmarks emergent agreements against real formations, yet evaluates only terminal outcomes, leaving the negotiation a black box with no account of *which* party drove *which* clause. PoliCon [42] contributes a large consensus-drafting benchmark (2,225 European Parliament records) scored by an LLM-as-judge, and diagnoses a systematic bias toward dominant-party positions in unsteered models – the failure mode that undermines faithful partisan simulation – though its judge-based adjudication is itself opaque. [5] formalize coalition compromise as mediation in a semantic document space, guaranteeing outcomes that improve on the status quo, but abstract away from ideology and manifesto language, so the compromises are mathematically principled yet politically ungrounded. UNBench [23] extends evaluation to real institutional behaviour (UN Security Council voting), but again as discrete outcome prediction. On the alignment side, PoliTune [1] shows ideology can be embedded at the parameter level via LoRA and DPO, shifting models along a left–right axis more coherently than prompting, yet does not instantiate *party-specific* platforms or deploy the aligned models in negotiation. A broader negotiation strand – the large-scale autonomous negotiation competition of [35], the buyer–seller system AgenticPay [24], and theory-of-mind bargaining in peer-to-peer markets [19] – establishes that LLM agents negotiate competently, but targets commercial rather than ideological, multi-party political settings.

Two capabilities are thus well established – high-fidelity political behaviour and benchmarkable negotiation – while two gaps recur: prior systems either predict outcomes without simulating the deliberation, or simulate deliberation without attributing its results to verifiable ideological origins.

2.2 Multi-Agent Debate and Deliberation

A second literature studies how multiple LLM agents argue toward a decision. Foundational results show structured debate improves factuality and reasoning [11] and counters the single-agent "degeneration of thought" by encouraging divergent positions [22]. Subsequent work refines the protocol: Exchange-of-Thought [41] formalizes communication topologies, voting- versus consensus-based decision rules trade off across task types [17,6], MALLM [4] decouples debate into composable configurations, and AC^3 [39] scales deliberation by clustering agents and electing representatives. Critically, this literature exposes its own failure modes: [8] document *silent agreement*, where agents capitulate to the majority answer, and [33] show strong aggregate debate scores can mask brittle internal dynamics. Both reflect the RLHF-induced pull toward conformity

and helpfulness that is fatal to sustained partisanship – compounded by persona research showing prompt-based role-play degrades reasoning [18] and lacks factual grounding [16]. Argumentation-theoretic work [2] and explainability-oriented frameworks [3] supply principled vocabularies for offers, concessions, and consensus, but assume cooperative agents and lack grounding in real ideological corpora. On evaluation, adjudication still leans on opaque LLM-judge committees [43,4], and [13] argue that outcome accuracy can mask faulty *process*, noting that no existing benchmark jointly covers process transparency and real-world grounding.

2.3 Positioning our Contribution

Taken together, prior work leaves four gaps. *(1) Unsteered neutrality and conformity:* debate agents and benchmarked models drift toward consensus and dominant-party positions [8,42,33], and prompt-based personas are too fragile to hold a line [18,16]. *(2) Generic, ungrounded ideology:* where ideology is steered it follows a coarse left–right axis [1], and where coalitions are simulated the policy content is abstract rather than manifesto-grounded [5]. *(3) Outcome-only, black-box deliberation:* political simulators predict or score terminal outcomes [20,26,23] without attributing provisions to their origins. *(4) Opaque adjudication and unmeasured process:* winners and quality are read off LLM-judge preferences [43] rather than an auditable account of how agreement was reached [13].

Our framework targets each gap directly. Against (1) and (2), we abandon prompting for two-stage SFT+DPO alignment on *party-specific* 2019 Flemish manifestos, coupled with a per-party RAG pipeline that keeps each agent factually bounded to its own platform—yielding agents that are neither neutrally helpful nor arbitrarily opinionated but faithfully partisan. Against (3), we stage a structured hub-and-spoke negotiation arbitrated by a formateur, so a coalition agreement *emerges* from deliberation rather than being predicted. Against (3) and (4) jointly, our evaluation suite makes the process auditable: MILT traces every clause of the final agreement back through agents' opening standpoints to its originating manifesto chunk and labels its provenance across five states; the CIS converts these traceable, hallucination-discounting labels into attributable negotiating power; and a real-world grounding layer (L_{Real}) benchmarks each provision against the historically adopted agreement – unifying, for genuinely partisan and manifesto-grounded agents, the process transparency and external fidelity that [13] identify as jointly absent.

3 Methodology

We simulate coalition formation with a three-stage pipeline: (1) an isolated RAG knowledge base, (2) a two-stage ideological alignment protocol (using SFT and DPO), and (3) a multi-agent negotiation arena. This decouples factual policy retrieval from persona injection, letting the base Large Language Model (LLM) argue subjectively while staying grounded in official manifestos (see Figure 1).

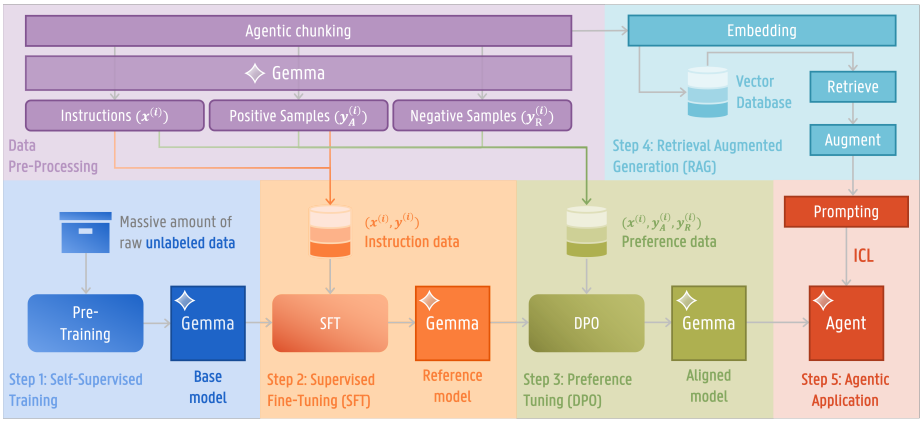

Fig. 1. Overview of the individual party alignment model process.

3.1 Data and Document Chunking

The corpus comprises the official 2019 electoral manifestos of the major Flemish parties. We partition each with an *Agentic–Structural Hybrid (ASH)* procedure (Figure 2), a *dual-constraint partitioning problem* where chunk boundaries respect both document hierarchy and local semantic coherence. PDFs are parsed with PyMuPDF [25] at the dictionary level for per-span bounding boxes and font sizes; margin and sub-minimum-font blocks are discarded, removing headers, page numbers, and footnotes without a layout model. Text is NFKC-normalized with an explicit ligature, custom-font, and smart-quote handling, and intra-paragraph breaks collapsed. Two cascaded constraints follow. The *Hard-Gate* is deterministic: regex detects formal section openings – numeric subsections (`1.1`, `1.1.1`), Article/Section/Chapter markers, capitalized colon-terminated headers – and forces a split, compensating for LLMs' *structural blindness* and guaranteeing distinct subsections are never conflated even when lexically near-duplicate. The *Soft-Gate* fires only on residual transitions: Gemma3 (27B) [12] via Ollama [27], queried with a structured-output schema at temperature 0.1 on the trailing 800 characters plus the candidate next block, returns a binary same-topic judgement. This hybrid design – deterministic logic where unambiguous, probabilistic reasoning for fluid transitions – mitigates the *agentic cost bottleneck* of LLM-driven chunking. Each chunk is persisted immediately, with a visualization PDF for auditing.

3.2 Preference-pair generation

The persisted chunks populate both the retrieval index and the tuning datasets. Each chunk over fifteen tokens is passed to Gemma3 (27B) [12] in a single structured-output call (temperature 0.7) whose JSON schema enforces a list of *(Prompt, Chosen, Rejected)* triplets. Because ASH chunks are thematically coherent but not atomic, the instruction directs the model to enumerate every distinct policy resolution and emit at least one Dutch triplet per chunk

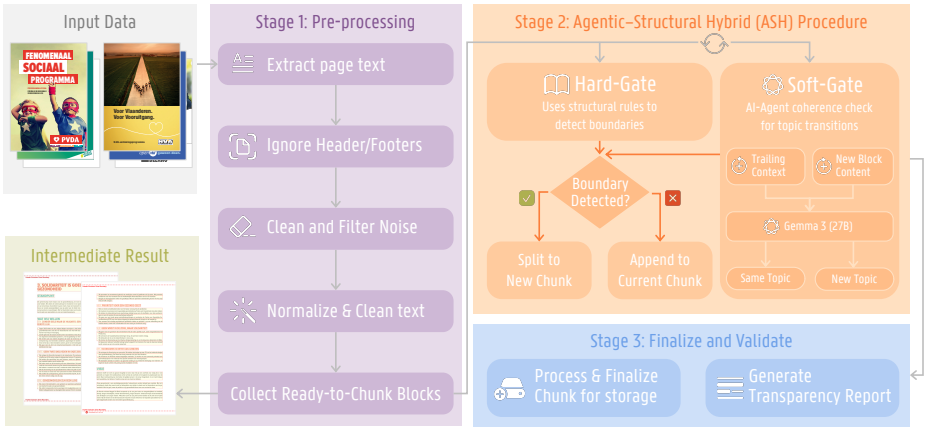

Fig. 2. Overview of the hybrid chunking strategy.

through prompting: the *Prompt* is a realistic citizen question leading to that resolution; the *Chosen* response rewrites the resolution in *assistant register* – the first-person conversational style of an instruction-tuned assistant, versus the manifesto's third-person programmatic prose; the *Rejected* counterfactual is a well-written answer to the same prompt that argues from an opposing political ideology or offers a sterile, non-committal stance. Triplets stream to JSONL in HuggingFace's preference-tuning chat format, so one artefact supports both SFT (*Prompt* → *Chosen*) and DPO (the full triplet). ASH's structural integrity gives each resolution unambiguous provenance, making preference pairs auditable and retrievals traceable to a document-tree node.

3.3 Two-Stage Ideological Alignment (SFT and DPO)

Standard LLMs' neutrality biases preclude authentic political simulation. We use Gemma-3 (27B) [12], quantized to 4-bit via QLoRA on Apple MLX [15], aligned in two stages.

Stage 1 (SFT) fine-tunes on the *Chosen* responses to adapt conversational structure and style, establishing what we term the "Method Actor" persona: rather than describing the party as an outside observer (*"the party proposes..."*), the model speaks *as* the party (*"we will..."*), defending its manifesto commitments as its own.

Stage 2 (DPO) addresses what SFT cannot: cross-entropy SFT only raises the likelihood of the target text, layering a stylistic overlay on a base model whose ingrained prior for encyclopedic neutrality remains intact, so under adversarial prompting the model reverts to that dominant prior. DPO's contrastive objective instead maximizes the implicit reward gap between *Chosen* and *Rejected* [32], pushing probability mass away from the neutral or opposing responses – penalizing encyclopedic neutrality and rewarding combative partisan rhetoric (implicit KL-regularization strength $\beta = 0.1$ against the SFT reference policy).

To equalize alignment depth across manifestos of varying length, each party's adapter steps are scaled as $I = \frac{N}{B} \times E$ with I the total number of iterations, N the dataset size, B the batch size, and E the target epochs. SFT uses $B = 2$ and $E = 5$; DPO uses $B = 1$ (given the memory overhead of preference tuning a 27B model) and $E = 3$, empirically chosen to prevent catastrophic forgetting. In the DPO stage a low learning rate of 5×10^{-6} resumes and updates the SFT adapter weights.

3.4 Policy Grounding through RAG

At inference we supplement the fine-tuned model with RAG: where SFT and DPO calibrate the persona, RAG supplies factual constraints against the "knowledge boundary problem" of relying on generalized internal knowledge. The retrieval corpus is the same policy data used in fine-tuning, so arguments derive strictly from the party's positions. Concretely, an isolated ChromaDB [7] pipeline assigns each party its own vector collection in a local SQLite-backed store to prevent cross-contamination. Dense vectors are computed on CPU via paraphrase-multilingual-MiniLM-L12-v2 served by FastEmbed [30], giving each agent perfect recall of its manifesto without memory bloat.

3.5 Multi-Agent Negotiation Arena and Coalition Simulation

The arena follows a *hub-and-spoke* topology: party agents are spokes, a central *formateur* the hub. The formateur is instantiated from the same fine-tuned LLM as the largest coalition party but is system-prompted to adopt the persona of a neutral, realistic broker; after parties present standpoints it drafts an interim agreement refined across four rounds into the final agreement. This orchestration, the *Digital Pantheon*, turns static QA models into autonomous interacting entities. The four-round horizon is a design choice rather than a measured convergence criterion: in preliminary runs it was the point at which each party's priorities and common ground were fully surfaced, while further rounds added little new negotiation content. For efficiency, adapter hot-swapping keeps the base Gemma-3 weights static in unified memory while party-specific DPO adapters are loaded per turn, simulating all major parties on consumer hardware.

The protocol is an iterative multi-turn debate. Each turn, an active agent processes the dialogue state and runs a semantic search against its policy vector collection; the top-k chunks (cosine similarity < 0.51, maximum $k = 5$) are injected as strict ideological boundaries, so agents cannot concede core promises or hallucinate out-of-line compromises. Under a zero-shot consensus objective, agents articulate non-negotiable *"red lines"*, identify shared objectives, and propose syntheses. Mirroring convention, the largest party assumes the *formateur* role, drafting the final agreement and judging whether balanced consensus is reached. To enforce realistic power dynamics it is given the Banzhaf power indices of all parties [28]. In our initial tests, a non-fine-tuned base formateur favoured moderate parties even with these indices, whereas a party-tuned formateur preserves the leverage they dictate. Logging all rounds – retrieved chunks, interim

drafts, final concessions – yields a computational model of expected coalition structures. We apply this approach to all 22 policy topics of the Flemish government's portfolio, derived from the official 2019 coalition agreement (Table 5).

3.6 Explainability

To counter the black-box nature of generative agents and mitigate *prompt artifact bias*, we introduce a *Multi-Layered Information Lineage Topology* (MILT), modelling the negotiation as a temporal directional graph. Rather than parsing voluminous transcripts, we infer the negotiation dynamics (L_2) by tracing every policy measure in the final agreement (L_3) backward through agents' initial standpoints (L_1) to the retrieved manifesto chunks (L_0). We deploy Qwen3.6 (27B) via Ollama as a backward Natural Language Inference (NLI) classifier [31,27]: in one JSON inference call it ingests L_0, L_1, and L_3 together, capturing not just *whether* an idea mutated but *when*, distinguishing preemptive dilution ($L_0 \to L_1$) from reactive concession ($L_1 \to L_3$). It categorizes provenance into five topological states, separating valid negotiation dynamics from systemic errors:

1. **Direct Lineage (Ideological Preservation):** maps to a party's L_0 manifesto and survives L_1, the negotiation, and remains intact to L_3.
2. **Diluted Lineage (Concession):** originates in an L_0 manifesto but is measurably weakened, preemptively ($L_0 \to L_1$) or reactively ($L_1 \to L_3$).
3. **Synthesized Lineage (Logrolling):** combines distinct, competing L_0 chunks into a novel hybrid generated through interaction.
4. **Pipeline Artifact (Initialization Bias):** present in L_1 and L_3 but absent from L_0, exposing prompt-induced setup noise.
5. **Orphan (Debate Hallucination):** appears in L_3 but lacks L_0/L_1 grounding, a spontaneous negotiation-phase hallucination.

The resulting JSON records the policy text, originating party, L_0/L_1 grounding indicators, topology class, and a qualitative audit trail. By isolating ideological retention from system noise, MILT quantifies negotiation dynamics from the survival of human-authored source material without annotating debate logs.

3.7 Evaluation: A Coalition Influence Score

To name the winner from the topology, we map +3 (Direct), +2 (Diluted), +1 (Synthesized) into a **Coalition Influence Score (CIS)**. Let P denote the set of policy measures in the final agreement, Π the set of coalition parties (here {N-VA, CD&V, Open Vld}), and T the five MILT states; $\pi : P \to \Pi$ maps each policy to its originating party and $\tau : P \to T$ to its topology state. For each policy p:

$$\mathrm{CIS}(\pi) \;=\; \sum_{p\,:\,\pi(p)=\pi} w(\tau(p)), \qquad w = \begin{cases} 3 & \tau = \text{Direct } (\tau_1) \\ 2 & \tau = \text{Diluted } (\tau_2) \\ 1 & \tau = \text{Synthesized } (\tau_3) \\ 0 & \text{otherwise } (\tau_4/\tau_5) \end{cases} \tag{1}$$

The gradient captures degrees of ideological victory: +3 for undiluted dominance, +2 for agenda-setting despite magnitude concessions, +1 for anchoring core principles within hybrid compromises. The 3/2/1 values are an ordinal encoding of ideological survival rather than calibrated magnitudes. Any monotonically decreasing weighting over {*Direct*, *Diluted*, *Synthesized*} preserves the intended ordering of per-clause influence. Zero-weighting *Pipeline* and *Orphan* prevents hallucinations from inflating scores: novel provisions from parametric prior knowledge lack a causal link to the platforms, and excluding untraceable content is essential for objectively allocating influence.

3.8 Empirical Validation via Real-World Grounding

To test external validity we add a *Real-World Grounding* layer (L_{Real}) answering the study's ultimate question: did the simulation invent plausible compromises, or recreate those that occurred? Using the same Qwen3.6 (27B) NLI pipeline [31], we grade every L_3 measure – keeping its MILT class – against the historically adopted agreement [36] as *Present* (exact or functional equivalent), *Partially Present* (altered constraints, timelines, or scope), or *Absent*. Unlike the CIS, this comparison grades every provision regardless of class: including the hallucination classes keeps the test fair rather than circular and lets the data reveal whether traceable lineage predicts materialization. That untraceable classes largely fail to appear is an empirical finding, not an artefact, and does not affect the influence scores. Cross-tabulating internal MILT states against external presence then tests whether anchored outcomes (*Direct*, *Diluted*, *Synthesized*) align with reality while systemic errors (*Pipeline*, *Orphan*) fail to materialize.

4 Results and Discussion

A single pipeline execution negotiates the coalition scenario (N-VA, CD&V, Open Vld) across the 22 policy dossiers exhausting the Flemish government's competences, producing on average 876 classified provenance points over its $L_0/L_1/L_3$ corpus. To show results reflecting the simulator rather than one stochastic draw, we repeat the simulation three times ($N = 2{,}629$ points), reporting quantities as across-run mean ± SD; each run is evaluated twice, isolating evaluator noise from run-to-run variation. Every L_3 point carries a MILT label, a human-readable audit trail, and a grounding grade against the 2019 coalition agreement (*Present*, *Partial*, *Absent*), summarized per dossier by the realization index $G = (n_{present} + 0.5\, n_{partial})/N$. This section follows the framework's three pillars: lineage explainability using MILT, CIS to define the winner, and a real-world grounding analysis.

4.1 Explainability and the transparency of grounding

Table 1 reports each MILT state across the three simulations with mean ± SD and corpus share. Provenance-bearing provisions (*Direct*, *Diluted*, *Synthesized*)

account for nearly all present and most partial judgements, whereas systemic-error classes (*Orphan*, *Pipeline*) are overwhelmingly absent and seldom fully grounded. This ordering is stable across every simulation and pass. Provisions with genuine manifesto provenance are consistently more likely to match the historically adopted 2019 agreement than the model's ungrounded inventions – a relationship quantified in the grounding analysis below.

Table 1. MILT topology composition across the three simulations (pooled).

MILT topology state	**Sim 1**	**Sim 2**	**Sim 3**	**Mean ± SD**	**Share**
Direct Lineage (τ_1)	390	373	387	383.3 ± 9.1	43.7% ± 0.9%
Diluted Lineage (τ_2)	116	126	117	119.7 ± 5.5	13.7% ± 0.8%
Synthesized Lineage (τ_3)	119	130	111	120.0 ± 9.5	13.7% ± 1.2%
Pipeline Artifact (τ_4)	62	62	55	59.7 ± 4.0	6.8% ± 0.4%
Orphan (Hallucination) (τ_5)	212	175	194	193.7 ± 18.5	22.1% ± 1.7%
Ideological Retention ($\tau_1 + \tau_2$)	506	499	504	503.0 ± 3.6	57.4% ± 1.0%
Systemic Error ($\tau_4 + \tau_5$)	274	237	249	253.3 ± 18.9	28.9% ± 1.6%
Total	**899**	**866**	**864**	**876.3 ± 19.7**	**100%**

The test is meaningful only because the framework is transparent end-to-end: each point's audit trail names the originating party, records its L_0 and L_1 match, and justifies the topology and grounding in natural language. For instance, the guarantee against *"forced school mergers"* traces verbatim to an N-VA manifesto breakpoint and survives into the final (L_3) agreement, whereas an invented *"1.15% of GNP for education"* target is flagged as an *Orphan* with no L_0/L_1 antecedent and is absent from reality. Every aggregate below thus decomposes into inspectable, party-attributed decisions.

4.2 From Topology to Winner

The same provenance names a winner via the CIS (Eq. 1): a Direct or Diluted clause assigns its full weight to the originating party, a Synthesized clause one point to each contributor its audit trail names – capturing only explicitly attributable influence. Computed per evaluation pass and averaged within each simulation, the winner is a property of the simulator rather than of a single stochastic draw. N-VA wins every simulation and the ranking N-VA > CD&V > Open Vld is identical across runs: averaged, N-VA commands 40.0% ± 3.3% of attributable influence, ahead of CD&V (31.8% ± 1.8%) and Open Vld (28.2% ± 1.8%), with non-overlapping means. This is more balanced than the historical 2019 intra-coalition seat distribution (70/124 parliamentary seats), where N-VA held 50.0% (35/70), CD&V 27.1% (19/70), and Open Vld 22.9% (16/70).

Per-pass counts exceed distinct clauses because multi-party and synthesized provisions credit each named party. Composition is as informative as magnitude and equally stable: N-VA's lead rests overwhelmingly on *Direct Lineage* (about three-quarters of its score every run), Open Vld draws most from *Synthesized*

Table 2. Coalition Influence Score (CIS) by party across the three simulations; per-simulation cells give the mean ± SD across the simulation's two evaluation passes.

Party	Sim 1 CIS	Sim 2 CIS	Sim 3 CIS	Mean share ± SD	Rank
N-VA	379.0 ± 19.8 (43.7%)	318.5 ± 14.8 (37.2%)	336.0 ± 2.8 (39.1%)	40.0% ± 3.3%	1
CD&V	257.0 ± 15.6 (29.7%)	279.5 ± 1.4 (32.7%)	283.5 ± 9.2 (33.0%)	31.8% ± 1.8%	2
Open Vld	230.5 ± 2.1 (26.6%)	258.5 ± 7.8 (30.2%)	239.0 ± 4.2 (27.8%)	28.2% ± 1.8%	3

cross-party compromises in which it is most named, and CD&V sits between with an even split of retained and conceded weight. Crediting syntheses to every contributor narrows but never reorders the junior partners. The simulation thus reproduces a canonical coalition signature – the largest party (N-VA, who is also formateur) sets the agenda, the smaller liberal partner (Open Vld) builds consensus, the centrist (CD&V) sits between – and N-VA's advantage concentrates in the *Direct* class the grounding analysis finds most likely to materialize, so the winner's influence is also the most durable.

4.3 Empirical validation via real-world grounding

The grounding distribution is stable across the three simulations. Averaged, points are present in 9.9 ± 2.0%, partial in 35.9 ± 0.9%, and absent in 54.2 ± 1.5%, with $G = 0.278 \pm 0.017$; the two passes per run differ by at most 0.3 points in present rate, so residual variation is evaluator noise.

Table 3. Grounding distribution per simulation (pooled).

Simulation	N	Present	Partial	Absent	G
Simulation 1	899	10.1%	35.0%	54.8%	0.276
Simulation 2	866	11.8%	35.7%	52.5%	0.296
Simulation 3	864	7.8%	36.9%	55.3%	0.262
Across runs	**2,629**	**9.9 ± 2.0**	**35.9 ± 0.9**	**54.2 ± 1.5**	**0.278 ± 0.017**

Splitting by MILT class (Table 4, pooled) confirms internal lineage predicts external materialization. The three provenance classes (71.1% of classifications) realize at $G = 0.32$–0.37, whereas the two error classes (28.9%) collapse to a combined $G = 0.164$: *Orphan*, the largest error class, is *Absent* in 75.4% of cases (438 of 581). Notably, *Synthesized* clauses realize marginally better than *Direct* ones ($G = 0.365$ vs. 0.333), so genuine cross-party compromise survives at least as reliably as unilaterally retained provisions.

Reproducibility is weaker for generation volume than grounding quality, and the gap is itself an artefact of hallucination: distinct proposals per dossier varied widely across runs (*Poverty Alleviation* 53, 28, 14; *Economy and Innovation* 33,

Table 4. Grounding outcomes by MILT topology class, pooled across runs.

MILT topology state	**Present**	**Partial**	**Absent**	G
Direct Lineage (τ_1)	13.9%	38.9%	47.2%	0.333
Diluted Lineage (τ_2)	4.5%	42.3%	53.2%	0.256
Synthesized Lineage (τ_3)	14.2%	44.7%	41.1%	0.365
Pipeline Artifact (τ_4)	5.6%	35.2%	59.2%	0.232
Orphan (Hallucination) (τ_5)	4.0%	20.7%	75.4%	0.143
Ideological Retention ($\tau_1 + \tau_2$)	11.7%	39.7%	48.6%	0.315
Systemic Error ($\tau_4 + \tau_5$)	4.3%	24.1%	71.6%	0.164

27, 58), and these swings concentrate in the ungrounded topologies, with *Orphan* and *Pipeline* together near 30% (760 of 2,629) and far less predictable than manifesto-anchored material. At non-zero temperature the model enumerates fluctuating unsupported elaborations around a stable ideological core, so we base all claims on proportions, not raw volumes.

On that basis, a substantial fraction of provisions reached a real-world counterpart despite only four rounds: 45.8 ± 1.5% were graded *Present* or *Partial* (pooled *Present* 9.9%, 95% CI [8.7, 11.0]; *Partial* 35.9%, [34.0, 37.7]; *Absent* 54.2%, [52.3, 56.1]). Realization is sharply uneven, from $G = 0.491$ (*Education*) to $G = 0.147$ (*Finance and Budget*). Per-dossier counts are too sparse for a single run, but run-invariant grounding (Table 3) licenses pooling the three runs; Table 5 reports the pooled estimates, ordered by G.

Two regularities organize this ranking, both following from the level of abstraction at which a proposal is pitched rather than its subject. First, full presence (*Present*) concentrates in programmatic dossiers – only *Education* (23.2%), *Foreign Policy and Tourism* (18.3%), and *Poverty Alleviation* (17.9%) approach or exceed 18% *Present* – and for the best-realized dossiers the *Partial* state is modal (*Heritage* 66.7%, *Education* 51.8%), indicating thematic correspondence over literal adoption. Second, the lowest-realized dossiers are saturated with quantified fiscal commitments or competences split with the federal level: *Brussels and Flemish Periphery* records no fully grounded proposal and *Finance and Budget* almost none (1.1%), the former partly because the theme occupies only a brief manifesto passage ($N = 52$). The dominant *Absent* category is thus chiefly fabricated operational detail – invented budgets, percentages, deadlines, institutions – not failed thematic reasoning. The *Partial* state localizes this: most partials identify the correct commitment then append a quantified or institutional specification absent from the agreement, while a smaller set encode genuine contradiction (e.g., a proposal to strengthen Unia where the agreement instead winds it down, or one making the social-housing means test voluntary where the agreement makes it compulsory), which our three-state scheme registers as *Absent*.

Two implications follow. Lineage topology is an actionable, ex-ante reliability signal: anchored provisions realize far more often than *Orphans* and *Pipeline Artifacts*, so topology can triage output before any real-world referent exists, and

Table 5. Per-dossier grounding distribution, pooled across runs and ordered by G.

Policy domain	N	Present	Partial	Absent	G
Education	112	23.2%	51.8%	25.0%	0.491
Poverty Alleviation	95	17.9%	47.4%	34.7%	0.416
Foreign Policy and Tourism	82	18.3%	43.9%	37.8%	0.402
Heritage	48	6.3%	66.7%	27.1%	0.396
Justice	108	12.0%	41.7%	46.3%	0.329
Well-being	131	14.5%	35.1%	50.4%	0.321
Cohesion, Onboarding and Integration	140	14.3%	35.7%	50.0%	0.321
Economy and Innovation	118	11.9%	38.1%	50.0%	0.309
Energy and Climate	176	12.5%	32.4%	55.1%	0.287
Housing	152	7.9%	40.8%	51.3%	0.283
Environment	117	6.0%	39.3%	54.7%	0.256
Media and Public Broadcaster (VRT)	131	6.1%	38.9%	55.0%	0.256
Animal Welfare	121	4.1%	42.1%	53.7%	0.252
Mobility and Public Works	159	11.9%	25.8%	62.3%	0.248
Government Operations and Efficiency	159	5.0%	37.7%	57.2%	0.239
Work and Social Economy	141	9.2%	27.7%	63.1%	0.230
Agriculture and Marine Fishing	116	6.9%	31.0%	62.1%	0.224
Local and Urban Governance	191	11.5%	20.4%	68.1%	0.217
Equal Opportunities	82	4.9%	31.7%	63.4%	0.207
Culture, Youth, and Sport	106	3.8%	32.1%	64.2%	0.198
Brussels and Flemish Periphery	52	0.0%	36.5%	63.5%	0.183
Finance and Budget	92	1.1%	27.2%	71.7%	0.147

the agreement between internal lineage and external grounding makes the two evaluations mutually reinforcing. And because hallucinated operational specificity is both the dominant and a localized failure mode, it is the most addressable – amenable to constrained decoding or post-hoc verification without altering the thematic reasoning the simulation already performs well.

5 Conclusion

We set out to determine not merely which coalition a multi-agent simulation predicts, but how it reaches agreement and which party shapes it. We built a coalition simulator coupling ASH chunking of manifestos, a two-stage protocol that adapts one base model by SFT then steers it into distinct partisan personas via DPO, and a manifesto-grounded retrieval layer keeping each agent bounded to its platform throughout a hub-and-spoke negotiation arbitrated by a tuned formateur. Unlike debate frameworks that assume cooperative agents, our agents are neither neutrally helpful nor arbitrarily opinionated but faithfully partisan – the property that makes the emergent dynamics worth measuring.

On this simulator we presented the first end-to-end empirical evaluation of *Multi-Layered Information Lineage Topology* (MILT). Over 22 dossiers across three simulations, each evaluated twice (N = 2,629 points), we showed that (i) the grounding distribution is highly reproducible – proportions barely move

and $G = 0.278 \pm 0.017$ even as individual edge classifications drift; (ii) 57.4% of provisions are anchored in manifestos (*Direct* or *Diluted*), a further 13.7% arise as genuine synthesis, against a 28.9% systemic-error rate dominated by debate-phase hallucination rather than initialization bias; and (iii) the topology decomposes into per-dossier regimes tracking each area's statutory and substantive structure. Validating these labels against the historical 2019 Flemish agreement, internal provenance and real-world materialization reinforce one another: provisions with verifiable lineage realize far more often than orphaned or pipeline content, so topology is an actionable, ex-ante reliability signal even before ground truth exists. Averaged across runs, $45.8 \pm 1.5\%$ of provisions reached a real-world counterpart after only four rounds – notable external fidelity – though realization is uneven across dossiers, peaking in *Education*, with the dominant absent category reflecting fabricated operational detail rather than failed thematic reasoning. Building on these classifications, the *Coalition Influence Score* (CIS), a principled $+3/+2/+1$ mapping excluding hallucinated content by construction, identifies N-VA as the clear winner ($40.0 \pm 3.3\%$ of attributable weight, ahead of CD&V and Open Vld) – robust across all runs and passes, consistent with N-VA being both largest party and formateur, and concentrated in the *Direct* class most likely to materialize. MILT thus delivers what an end-to-end classifier cannot: a temporally-decomposed, reproducibility-quantified, externally-grounded account of *how* a simulated coalition reaches agreement, and *who* wins it.

Several caveats and extensions remain. The CIS weights clauses equally regardless of budgetary and political stakes (e.g., *Education* outweighing *Animal Welfare*). The simulation is closed, omitting socio-economic, federal, media-salience, and electorate-feedback channels, while realization is conditioned on the single 2019 referent. Finally, the three-state grounding scheme records contradiction merely as absence, and the automated NLI labels remain transparent only through their audit trails. Future work could enrich the simulation: a two-level-game formulation [29] coupling regional and federal bargaining with a media-salience process, architectural ablations (peer-to-peer, partisan, or multi-formateur committees, and alternative formateur instantiations beyond the largest-party convention adopted here), component ablations quantifying the marginal contribution of SFT, DPO, and RAG against prompt-only or non-tuned baselines, and perturbed inputs such as alternative manifestos (e.g., 2024) or counterfactual coalitions. Cross-system replication would then test whether the narrow attrition band, frame-anchoring dominance, and lineage typology are system-level properties of proportional-representation coalition politics or artefacts of the Flemish segmented party system [9,10]. Finally, the evaluation could be hardened by validating a sample of the automated NLI labels against human expert annotators, and by moving the CIS beyond uniform clause weighting, including a sensitivity analysis of its 3/2/1 gradient under alternative weight vectors.

References

1. Agiza, A., Mostagir, M., Reda, S.: PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in Large Language Models.

In: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. p. 2–12. AIES '24, AAAI Press (2025)
2. Amgoud, L., Belabbes, S., Prade, H.: Towards a formal framework for the search of a consensus between autonomous agents. In: Proceedings of the 4th international joint conference on Autonomous agents and multiagent systems. pp. 537–543 (2005)
3. Bandara, E., Hewa, T., Gore, R., et al.: Towards Responsible and Explainable AI Agents with Consensus-Driven Reasoning (2025)
4. Becker, J., Kaesberg, L.B., Bauer, N., et al.: MALLM: Multi-Agent Large Language Models Framework. In: Proceedings of the 2025 Conference on EMNLP: System Demonstrations. pp. 418–439. ACL, Suzhou, China (2025)
5. Briman, E., Shapiro, E., Talmon, N.: AI-Generated Compromises for Coalition Formation (2026)
6. Choi, H.K., Zhu, J., Li, S.: Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models? In: Advances in Neural Information Processing Systems. vol. 38, pp. 101732–101764. Curran Associates, Inc. (2025)
7. Chroma-core: Chroma: The AI-native open-source embedding database (2023)
8. Cui, Y., Fu, H., Zhang, H., Wang, L., Zuo, C.: Free-MAD: Consensus-Free Multi-Agent Debate (2025)
9. De Winter, L., Swyngedouw, M., Dumont, P.: Party system(s) and electoral behaviour in Belgium: From stability to balkanisation. West European Politics **29**(5), 933–956 (2006)
10. Deschouwer, K., Reuchamps, M.: The Belgian Federation at a Crossroad. Regional & Federal Studies **23**(3), 261–270 (2013)
11. Du, Y., Li, S., Torralba, A., et al.: Improving factuality and reasoning in language models through multiagent debate. In: ICML'24: Proceedings of the 41st ICML. JMLR.org (2024)
12. Gemma Team, Kamath, A., Ferret, J., et al.: Gemma 3 Technical Report (2025)
13. Gritta, M., Paul, D., Li, X., et al.: Process Evaluation for Agentic Systems. In: Findings of the ACL: EACL 2026. pp. 2678–2692. ACL, Rabat, Morocco (2026)
14. Guo, T., Chen, X., Wang, Y., et al.: Large language model based multi-agents: a survey of progress and challenges. In: Proceedings of the 33rd IJCAI (2024)
15. Hannun, A., Digani, J., Katharopoulos, A., Collobert, R.: MLX: Efficient and flexible machine learning on Apple silicon (2023)
16. Jiang, H., Zhang, X., Cao, X., et al.: PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In: Findings of the ACL: NAACL 2024. pp. 3605–3627. ACL, Mexico City, Mexico (2024)
17. Kaesberg, L.B., Becker, J., Wahle, J.P., et al.: Voting or Consensus? Decision-Making in Multi-Agent Debate. In: Findings of the ACL: ACL 2025. pp. 11640–11671. ACL, Vienna, Austria (2025)
18. Kim, J., Yang, N., Jung, K.: Persona is a Double-edged Sword: Mitigating the Negative Impact of Role-playing Prompts in Zero-shot Reasoning Tasks (2024)
19. Kröhling, D.E., Chiotti, O.J., Martínez, E.C.: Artificial Theory of Mind in contextual automated negotiations within peer-to-peer markets. Engineering Applications of Artificial Intelligence **120**, 105887 (2023)
20. Li, H., Gong, R., Jiang, H.: Political Actor Agent: Simulating Legislative System for Roll Call Votes Prediction with Large Language Models (2024)
21. Li, X.: A Review of Prominent Paradigms for LLM-Based Agents: Tool Use, Planning (Including RAG), and Feedback Learning. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 9760–9779. ACL, Abu Dhabi, UAE (2025)

22. Liang, T., He, Z., Jiao, W., et al.: Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In: Proceedings of the 2024 Conference on EMNLP. pp. 17889–17904. ACL, Miami, Florida, USA (nov 2024)
23. Liang, Y., Yang, L., Wang, C., et al.: Benchmarking LLMs for Political Science: A United Nations Perspective. Proceedings of the AAAI Conference on Artificial Intelligence **40**(1), 738–745 (2026)
24. Liu, X., Gu, S., Song, D.: AgenticPay: A Multi-Agent LLM Negotiation System for Buyer-Seller Transactions (2026)
25. McKie, J., Liu, R.: PyMuPDF Documentation (2024)
26. Moghimifar, F., Li, Y.F., Thomson, R., Haffari, G.: Modelling Political Coalition Negotiations Using LLM-based Agents (2024)
27. Ollama Team: Ollama Documentation (2023)
28. Penrose, L.S.: The Elementary Statistics of Majority Voting. Journal of the Royal Statistical Society **109**(1), 53–57 (1946)
29. Putnam, R.D.: Diplomacy and Domestic Politics: The Logic of Two-Level Games. International Organization **42**(3), 427–460 (1988)
30. Qdrant Team: FastEmbed: Fast, light, accurate library built for retrieval embedding generation (2023)
31. Qwen Team: Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model (2026)
32. Rafailov, R., Sharma, A., Mitchell, E., et al.: Direct preference optimization: Your language model is secretly a reward model. Adv. Neural Inf. Process. Syst. **36**, 53728–53741 (2024)
33. Smit, A., Grinsztajn, N., Duckworth, P., et al.: Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
34. Sumers, T.R., Yao, S., Narasimhan, K., Griffiths, T.L.: Cognitive Architectures for Language Agents (2024)
35. Vaccaro, M., Caosun, M., Ju, H., Aral, S., Curhan, J.R.: Advancing AI Negotiations: A Large-Scale Autonomous Negotiation Competition (2025)
36. Vlaamse Regering: Regeerakkoord van de Vlaamse Regering 2019–2024 [Coalition Agreement of the Flemish Government 2019–2024] (2019)
37. Wang, L., Ma, C., Feng, X., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)
38. Xi, Z., Chen, W., Guo, X., et al.: The rise and potential of large language model based agents: a survey. Science China Information Sciences **68**(2), 121101 (2025)
39. Yang, L., Li, S., Deng, A.: Dynamic Consensus Communication Mechanism for Large Language Model-Based Multi-Agent Systems. Journal of Signal Processing Systems **98**(1), 10 (2026)
40. Yehudai, A., Eden, L., Li, A., et al.: Survey on Evaluation of LLM-based Agents (2025)
41. Yin, Z., Sun, Q., Chang, C., et al.: Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication. In: Proceedings of the 2023 Conference on EMNLP. pp. 15135–15153. ACL, Singapore (2023)
42. Zhang, Z., Wang, X., Yi, M., et al.: PoliCon: Evaluating LLMs on Achieving Diverse Political Consensus Objectives (2026)
43. Zhao, R., Zhang, W., Chia, Y.K., et al.: Auto-Arena: Automating LLM Evaluations with Agent Peer Battles and Committee Discussions. In: Proceedings of the 63rd Annual Meeting of the ACL (Volume 1: Long Papers). pp. 4440–4463. ACL, Vienna, Austria (2025)