Deliberation in Silico: Validating LLM Multi-Agent Simulation Against Verbatim Council Records

Christine Sheldon[1][0000−0002−7509−3411], Michael Laver[2], and Slava Jankin[3][0000−0001−6915−177X]

[1] Bennett School of Public Policy, University of Cambridge, Cambridge, UK
crs80@cam.ac.uk
[2] New York University, New York, USA
michael.laver@nyu.edu
[3] Institute for Data and AI, University of Birmingham, Birmingham, UK
v.jankin@bham.ac.uk

Abstract. Political deliberation is constituted in language, yet computational models of politics typically abstract away from discourse. Large language models offer a new capability: agents that participate in realistic political dialogue. Before such simulations can ground causal inference about institutional effects, we must establish that they reproduce documented patterns of real deliberation. We validate LLM-based multi-agent simulation against verbatim transcripts of the EU Council of Ministers, using the 2011 Common European Sales Law orientation debate as our case study. Across 140 simulations examining calibration, orchestration, and coalition dynamics, we find that: (1) fully calibrated agents produce nuanced discourse that resists simple position classification—a pattern validated by comparison with real ministerial statements; (2) procedural structure reliably shapes participation patterns; (3) agents prompted with appropriate institutional context exhibit behaviour consistent with the EU Council's documented consensus culture, converging toward compromise regardless of formal voting rules; and (4) agents show implicit bloc coordination through shared argumentative emphases, though with more explicit cross-referencing than observed in real debates. These findings provide initial proof-of-concept validation of LLM-MAS as a tool for extending simulation-based institutional analysis to language-mediated political processes.

Keywords: Multi-agent simulation · Large language models · Political deliberation · EU Council · Validation

1 Introduction

Political deliberation is constituted in language. Ministers do not merely vote; they argue, persuade, hedge, and build coalitions through discourse. Yet computational models of politics have typically abstracted away from language, simulating decisions and outcomes rather than the deliberative process itself [10, 4].

Agent-based simulation has enabled powerful counterfactual analysis in social science [7, 2]. By manipulating institutional parameters—voting rules, coalition structures, information asymmetries—researchers can explore mechanisms and causal relationships inaccessible through observational data alone. However, for political processes constituted in language, traditional agent-based approaches face a fundamental limitation: agents make decisions, but they do not deliberate.

Large language models offer a new possibility: agents that can participate in realistic political dialogue [15, 18]. LLM-based multi-agent systems (LLM-MAS) can simulate not just the outcomes of political processes but the discursive practices through which outcomes are produced. This capability could extend simulation-based institutional analysis to negotiation, deliberation, and coalition formation—domains where language is not incidental but constitutive of political action.

Before such simulations can ground causal inference, however, we must establish that they reproduce documented patterns of real deliberation. Do LLM agents produce discourse that resembles actual political speech? Do they respond appropriately to institutional structures? Do they exhibit the coordination patterns and cultural norms observed in real political bodies?

We address these questions through a systematic validation study comparing LLM-based simulation to verbatim transcripts of EU Council deliberation. Using the 2012 Common European Sales Law (CESL) orientation debate as our case study, we examine how calibration, orchestration, decision rules, and coalition prompts affect agent deliberation across 140 simulations. This paper makes three contributions:

1. **Validation against verbatim record**: We provide the first systematic comparison of LLM-MAS deliberation to real political discourse, finding that fully calibrated agents reproduce the nuanced, hedged language characteristic of actual ministerial statements. Crucially, we identify a "nuance penalty": fully calibrated agents achieve 65% position classifiability, lower than a simpler condition (84%), yet *real ministers* in the verbatim transcript score only 43.5% under the same classifier—confirming that lower classifiability reflects realistic diplomatic hedging, not simulation failure.
2. **Multi-level validation framework**: We propose assessing political MAS across four levels: positional fidelity, procedural sensitivity, behavioural realism, and discourse fidelity. Simple position accuracy can mask unrealistic simulation.
3. **Discovery of realistic patterns**: We find that LLM agents prompted with appropriate institutional context exhibit behaviour consistent with the EU Council's documented consensus culture (converging toward compromise regardless of formal rules) and show implicit bloc coordination through shared argumentative emphases, whilst exhibiting more explicit cross-referencing than observed in real debates.

These findings provide initial proof-of-concept evidence for LLM-MAS as a tool for extending simulation-based institutional analysis to language-mediated

political processes, whilst identifying where simulated and real deliberation diverge.

2 Background

2.1 EU Council Deliberation

The Council of the European Union is the primary legislative body representing member state governments [13]. It meets in different configurations of ministers depending on the policy domain (e.g., Justice and Home Affairs, Economic and Financial Affairs). A distinctive feature is its "consensus culture": approximately 80% of decisions are adopted unanimously even when qualified majority voting (QMV) applies [14]. Risse and Kleine [16] distinguish *bargaining mode* (strategic position-taking) from *deliberation mode* (reasoned argument); orientation debates typically operate in the latter.

Orientation debates are structured discussions held early in the legislative process, before line-by-line negotiation of a proposal begins. Chaired by the rotating Presidency, they allow member states to state their initial positions and identify areas of agreement and concern. No binding decisions are taken; the aim is to map the political landscape and guide subsequent working group negotiations. This makes orientation debates especially well suited for studying deliberative behaviour: the deliberative mode dominates, strategic blocking is rare, and ministers are expected to reason openly about substantive concerns. Haege [11] shows that consensus often emerges as an "unintended by-product" of coalition dynamics in such settings, suggesting that coordination patterns may be as important as final votes.

2.2 LLM-Based Multi-Agent Simulation

Large language models have emerged as controllers for simulated agents. Park et al. [15] demonstrated emergent social behaviours; subsequent work explored LLM agents for economic [12] and political [1] simulation. Studies show LLMs exhibit cooperative-dominant profiles in strategic games [9, 5], potentially explaining consensus-seeking behaviour.

A key concern is validation [3, 17]. Position accuracy alone is insufficient—agents might parrot labels without realistic discourse. We argue for multi-level validation across four levels: (1) *positional fidelity*—do agents hold and express the expected national positions?; (2) *procedural sensitivity*—does changing institutional structure (orchestration, voting rules) produce appropriate changes in participation patterns?; (3) *behavioural realism*—do agents reproduce documented institutional norms such as consensus culture and bloc coordination?; and (4) *discourse fidelity*—does the language agents produce resemble real political speech in register, hedging, and substantive emphasis? Each level can succeed or fail independently, and a simulation that passes level 1 may still fail levels 3 or 4. This framework provides the structure for our three studies.

3 Method

We simulate deliberation in the Justice and Home Affairs (JHA) configuration of the Council of Ministers of the European Union, focusing on the Common European Sales Law (CESL) proposal [8], a 2011 European Commission initiative to create an optional contract law regime for cross-border transactions. The JHA configuration brings together justice ministers rather than economic or foreign ministers, operating in a more technical and legal register; CESL fell under its remit because the proposal concerned contract law harmonisation. The CESL orientation debate of 8 June 2012 (Council Meeting 3172) provides our empirical referent: a verbatim transcript of 52 interventions totalling 7,666 words from 23 member state delegations, the Commission, and the Danish Presidency (Chair).[4]

3.1 Agent Design

Our simulation employs 28 LLM-based agents: 26 EU member state delegates (the 27 member states in 2012 minus Denmark, which serves as Chair), the European Commission, and a Chair agent representing the Danish Presidency. Each agent is implemented using DeepSeek Chat [6] with a structured prompt containing: (1) role specification (delegate, commission, or chair); (2) country identity and institutional context; and (3) position-specific information calibrated by experimental condition. DeepSeek Chat was selected for its strong instruction-following performance on structured role-play tasks, cost-effective API access enabling 140 full-debate simulations, and competitive benchmark performance relative to proprietary alternatives at the time of the study. We acknowledge that different LLMs may exhibit different political behaviour patterns [17]; replication with other models is an important direction for future work.

Ground truth positions were derived independently of the debate transcript, to avoid circularity in validation. Sources included the CESL proposal itself, national parliamentary records and reasoned opinions submitted under the subsidiarity protocol (2011–2012), contemporary legal and financial press coverage, and publicly available Council preparatory documents and working party reports. Member states were classified into three blocs: *supportive* (13 states favouring the proposal), *sceptical* (6 states expressing significant reservations), and *conditional* (7 states open to discussion with caveats). Denmark served as Chair (neutral procedural role). Table 7 in Appendix A provides the full classification with supporting rationale for each assignment.

[4] The verbatim transcript of Council Meeting 3172 is not publicly available on the Council website. We are grateful to Dr James Cross and colleagues at the Connected_Politics Lab, University College Dublin, for providing access to the transcript, which was created from video footage of the debate originally shared by Prof. Chris Wratil (University of Vienna). The data collection was carried out in connection with Wratil's MULTIREP project (ERC Starting Grant, Horizon Europe, grant no. 101076033). Cross and colleagues' broader work on parliamentary and Council transparency is developed through the ParliView project (https://www.parliview.org).

3.2 Study 1: Calibration

Study 1 examines how prompt calibration affects positional fidelity. We varied the information provided to agents across three conditions:

- **C1 (Minimal)**: Country name and role only, no position information
- **C2 (Position label)**: Country name plus explicit position label (supportive/sceptical/conditional)
- **C3 (Full context)**: Position label plus country-specific expected arguments, key concerns, and institutional context (drawn from national parliamentary records, Council preparatory documents, and expert assessment; see Appendix C for full examples)

Each condition was replicated with 10 random seeds, yielding 30 simulations. We measured *position accuracy* (proportion of agents classified correctly against ground truth) and *argument coverage* (proportion of expected arguments mentioned).

3.3 Study 2: Orchestration and Decision Rules

Study 2 examines how procedural structure and voting rules affect deliberation patterns. We manipulated two factors in a 4×2 design:

Orchestration varied the deliberation structure:

- **O1 (Council style)**: Chair directs discussion, Commission speaks first, flexible turn-taking
- **O2 (Round robin)**: Fixed sequential order, each agent speaks once
- **O3 (Shuffled order)**: Randomised order per round, equal participation
- **O4 (Multi-round)**: Three sequential rounds, enabling position evolution

Decision rules varied the voting context:

- **D1 (Unanimity)**: Agents informed that unanimous agreement is required
- **D2 (QMV)**: Agents informed that qualified majority voting applies

With 10 seeds per cell, Study 2 comprises 80 simulations. We measured *participation inequality* (Gini coefficient of word counts), *behavioural language* (rates of veto, coalition, and compromise terminology), and *convergence* (change in compromise language across rounds in O4).

3.4 Study 3: Coalition Dynamics

Study 3 examines emergent coordination patterns. We varied coalition information:

- **I1 (Baseline)**: No coalition information provided
- **I2 (Signalled)**: Agents informed of position blocs

– **I3 (Enhanced)**: Full bloc membership lists, identification of likely allies and potential opponents, and framing of the debate as having collective bloc stakes (e.g., sceptical states advised that forming a blocking minority requires coordinating with specific partners)

With 10 seeds per condition, Study 3 comprises 30 simulations. We measured *cross-reference patterns* (which countries mention which others) and *issue similarity* (cosine similarity of argument emphasis profiles within and between blocs).

3.5 Validation Against Verbatim Record

To assess discourse fidelity beyond position accuracy, we compare simulated debates to the verbatim transcript of Council Meeting 3172 across five dimensions: *position classifiability* (pattern-based classification against ground truth), *issue emphasis* (keyword counting applied identically to real and simulated text; Appendix E), *cross-references* (regex detection of country mentions, within- vs between-bloc), *consensus language* (dictionary-based counting across six categories; Appendix F), and *textual overlap* (n-gram Jaccard similarity between matched pairs; Appendix L). Each measure has a pre-specified failure criterion; Table 8 in Appendix B provides full operationalisations. This comparison assesses whether simulated discourse patterns match real ministerial deliberation, whilst verifying that simulated text is not reproduced from the training corpus (Section 5.4).

3.6 Summary

Table 1 and Figure 1 summarise our experimental design. Across three studies and 140 simulations, we test how calibration, orchestration, decision rules, and coalition information affect LLM agent deliberation, validating findings against the verbatim record of the actual CESL debate. This design motivates four hypotheses:

– **H1**: Greater prompt calibration increases positional fidelity; fully contextualised agents (C3) produce the most realistic discourse.
– **H2**: Orchestration structure shapes participation patterns; conditions imposing sequential turn-taking better approximate real Council participation.
– **H3**: Agents reproduce the EU Council's consensus culture, converging toward compromise language over multiple rounds independently of formal voting rules.
– **H4**: Agents exhibit emergent bloc coordination through shared argumentative emphases and within-bloc cross-referencing, without explicit coalition instructions.

A full replication package—simulation code, agent prompts, all 140 output files, and the analysis pipeline reproducing every number in this paper—is available at https://github.com/sjankin/cesl-simulation.

Table 1. Experimental design summary.

Study	Factor	Conditions	Seeds	Total
1	Calibration	C1, C2, C3	10	30
2	Orchestration × Rules	O1–O4 × D1–D2	10	80
3	Coalition information	I1, I2, I3	10	30
Total simulations				**140**

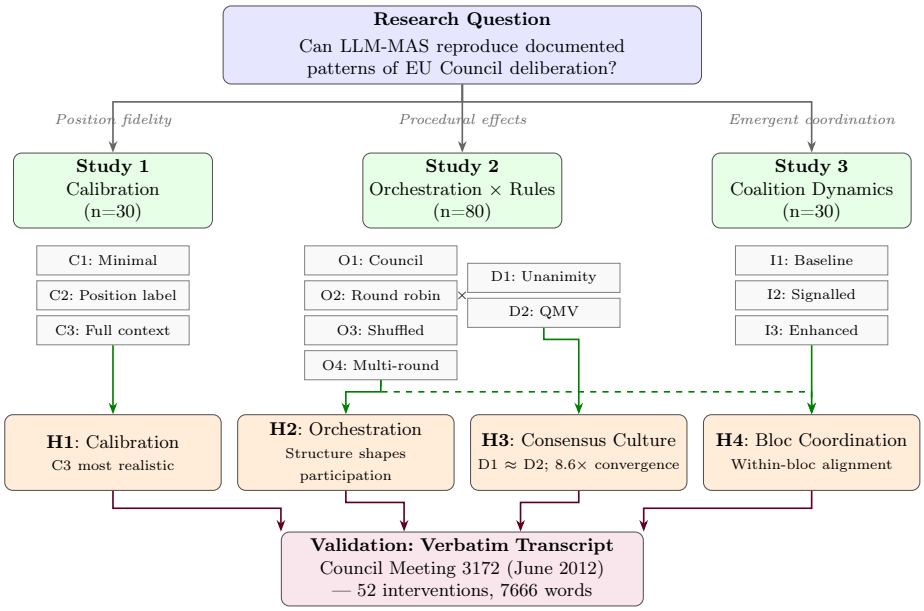

Fig. 1. Experimental architecture. Three studies (140 simulations) test how calibration, orchestration, decision rules, and coalition prompts affect LLM agent deliberation, validated against the verbatim transcript of Council Meeting 3172.

4 Results

4.1 H1: Calibration Affects Position Fidelity

Position accuracy varied significantly across calibration conditions ($F(2, 27) = 350.1$, $p < .001$). As shown in Table 2, C2 (position label) achieved the highest accuracy (84%), followed by C3 (full context, 65%) and C1 (minimal, 24%).

The finding that C3 < C2 initially appeared problematic. However, comparison with the verbatim transcript revealed that *real ministerial statements yield even lower classification accuracy* (43.5%) than our fully calibrated agents. Applying identical position extraction and classification procedures to the actual debate showed that ministers speak with hedging and conditionality that resists simple categorisation.

This "nuance penalty" validates C3 as producing the most *realistic* discourse, even if not the most *classifiable*. Argument coverage was uniformly high across

Table 2. Position accuracy and argument coverage by calibration condition, compared to real transcript.

Source	Position Accuracy	Argument Coverage	Interpretation
C1 (Minimal)	24%	87%	Near chance
C2 (Position label)	84%	84%	Label parroting
C3 (Full context)	65%	90%	Nuanced discourse
Real transcript	**43.5%**	—	Hedged, conditional

Table 3. Participation patterns by orchestration condition.

Condition	Gini (words)	Mean turns/agent	Pattern
O1 (Council style)	0.11	Variable	Chair-managed
O2 (Round robin)	0.07	1.0	Equal by design
O3 (Shuffled)	0.07	1.0	Equal, varied order
O4 (Multi-round)	0.05	3.0	Most equal (3 rounds)
Real transcript	**0.24**	**1.0**	Sequential, Chair-managed

conditions (84–90%), indicating that LLM agents raise substantively relevant issues regardless of calibration level. The key difference lies in *how* these arguments are framed: C3 agents express positions with appropriate diplomatic hedging, whilst C2 agents state positions more directly—a pattern we term "label parroting".

To illustrate: a C2 agent for a sceptical state typically produced statements such as *"[Country] is sceptical about the CESL proposal. We have serious reservations about the necessity and legal basis of this instrument"*, directly mirroring the supplied label. A C3 agent for the same state instead produced: *"While [Country] recognises the goal of facilitating cross-border trade, we remain concerned about whether this instrument truly adds value beyond existing harmonisation measures, and whether the legal basis under Article 114 is appropriate given the subsidiarity concerns raised by national parliaments"*—expressing the same sceptical position through substantive argument and conditional framing, consistent with the verbatim record (Appendix J).

4.2 H2: Orchestration Shapes Participation Structure

Orchestration conditions produced distinct participation patterns as designed. Table 3 shows participation inequality (Gini coefficient) across conditions.

All simulated conditions show lower Gini coefficients than the real transcript (0.24), indicating more equal participation in simulations. O1 (Council style) showed the highest simulated inequality (0.11), whilst O4 (multi-round) achieved the most equal distribution (0.05) as agents receive equal turns across three rounds.

Notably, the real CESL debate shows higher inequality (Gini = 0.24) than any simulation condition. Real ministers varied from 104 to 515 words per intervention, whilst simulated agents show more uniform output. This suggests that LLM agents *as prompted in this study* may not fully capture the heterogeneity

Table 4. Behavioural language by decision rule condition.

Metric (per 1k words)	D1 (Unanimity)	D2 (QMV)	p-value
Veto language	0.056	0.087	.22
Coalition language	0.011	0.045	.17
Compromise language	0.81	0.69	.60

in real ministerial participation—a design choice addressable through prompts that instruct agents to vary intervention length strategically. O2 (round robin) best approximates the real Council procedure—sequential speaking with Chair management—though with more equal word counts.

4.3 H3: Consensus Culture Emerges

Contrary to initial expectations, decision rules (D1: unanimity required; D2: qualified majority voting) produced no significant differences in behavioural language metrics. Table 4 shows veto and coalition language rates.

However, multi-round debates (O4) revealed strong *convergence* toward compromise: compromise language increased from 0.59 to 5.10 per 1,000 words—an 8.6-fold increase across three rounds (see supplementary Figure 2). This convergence was slightly stronger under unanimity (D1: +4.58) than QMV (D2: +4.44), consistent with consensus requirements. Full round-by-round values (D1: $0.60 \rightarrow 0.86 \rightarrow 5.18$; D2: $0.59 \rightarrow 0.54 \rightarrow 5.02$) show that convergence is concentrated in round 3, after agents have heard two full rounds of others' positions.

The convergence mechanism appears to be contextual: by round 3, each agent's prompt history contains the full preceding debate, and agents orient their statements toward the emerging consensus rather than restating initial positions. A sceptical-bloc agent that opened in round 1 with "[Country] has serious concerns about the necessity and proportionality of this proposal" typically softened by round 3 to language such as "whilst [Country] maintains reservations about certain aspects, we recognise the broad support for continuing work on this instrument and are willing to engage constructively with the text"—a shift in rhetorical register without abandonment of position that closely mirrors the real Council's documented approach to orientation debates.

Analysis of the real transcript provides context for interpreting these findings. We identified 82 instances of consensus-related language (13.1 per 1,000 words), comprising hedging (24), conditional expressions (20), softening (19), agreement (13), and non-blocking statements (5). Crucially, even sceptical delegations—those most opposed to the CESL proposal—used hedging and conditional language rather than blocking expressions. This pattern matches the documented EU Council "consensus culture" whereby deliberative behaviour dominates regardless of formal voting rules [16, 14].

The absence of behavioural differences between D1 and D2, rather than indicating simulation failure, thus reflects the accurate reproduction of this institutional norm.

Table 5. Issue similarity and cross-reference patterns by position bloc (Study 3, $n = 30$ debates).

Metric	Within-bloc	Between-bloc
Issue similarity (cosine)	0.933	0.852
Cross-references (% of 265 total)	90.2%	9.8%
Cross-references by bloc:		
Sceptical within-bloc	98.3%	—
Conditional within-bloc	8.3%	—

4.4 H4: Bloc Coordination Through Shared Arguments

Analysis of cross-reference patterns and argument emphasis revealed evidence of bloc coordination. Table 5 shows issue similarity within and between position blocs.

Agents within the same position bloc showed higher argument similarity (cosine 0.933 vs 0.852) and predominantly referenced other countries from the same bloc (90.2% within-bloc vs 35.1% expected by chance; permutation test $p = .004$, $n = 265$ mention pairs pooled across 30 debates). The sceptical bloc showed particularly tight coordination (98.3% within-bloc mentions)—a minority position that benefits from coalescing resistance, as explicit acknowledgement of allies signals the existence of a potential blocking minority. The conditional bloc showed only 8.3% within-bloc mentions, predominantly referencing countries from other blocs, acting as bridges. This bridge behaviour is consistent with the conditional states' strategic position as potential swing voters: rather than consolidating a bloc, they seek influence by demonstrating to both sides that their support is contestable.

Bloc-specific argumentative emphases emerged without explicit instruction. Sceptical states emphasised subsidiarity (14.7% of issue mentions vs ∼5–10% for other blocs) and legal basis concerns (13.3%), whilst supportive states focused on scope (26.6%) and SME benefits (11.5%).

Comparison with real transcript. However, analysis of the verbatim record revealed that real ministers produced only **4** explicit cross-country references across 52 interventions, compared to an average of **8.8** per simulated debate. This difference suggests that LLM agents are more explicitly dialogic than real Council discourse, which takes the form of "parallel monologues" with ministers stating positions without direct reference to others.

Despite this difference in explicit referencing, implicit coordination through shared argumentative emphases—sceptical states raising subsidiarity and legal basis concerns, supportive states emphasising scope and internal market benefits—is present in both real and simulated debates.

4.5 Summary of Findings

Table 6 summarises the evidence for each hypothesis.

Table 6. Summary of hypotheses and evidence.

H	Claim	Key Evidence	Support
H1	Calibration → fidelity	C3 most realistic (nuance penalty)	Strong
H2	Orchestration → participation	O2 matches real Council	Strong
H3	Consensus culture emerges	8.6× convergence; D1 ≈ D2	Moderate
H4	Bloc coordination	90.2% within-bloc ($p = .004$)	Strong

5 Discussion

5.1 What LLM-MAS Gets Right

Our findings demonstrate that LLM-based multi-agent simulation can reproduce several key patterns of real political deliberation.

Nuanced, hedged discourse. Fully calibrated agents (C3) produce discourse characterised by the hedging and conditionality observed in actual ministerial statements. The "nuance penalty"—whereby realistic discourse yields lower position classification accuracy—is not a simulation artefact but reflects genuine features of diplomatic communication. Real ministers achieved only 43.5% classifiability, suggesting that our C3 agents (65%) actually produce *more* classifiable discourse than real politicians whilst still capturing essential nuance.

Issue emphasis alignment. Simulated debates show near-perfect alignment with real debates on two key issue categories: consumer protection (real: 10.8%, simulated: 9.9%) and SME concerns (real: 9.0%, simulated: 8.9%) were emphasised at virtually identical rates. Other issues diverged more substantially—simulated debates overemphasised scope (7.8% real vs 19.6% simulated) and underemphasised necessity (22.9% real vs 14.2% simulated)—likely because calibrated agents are prompted with the Presidency's four explicit debate questions, which foreground scope directly. This suggests that calibrated agents reproduce the substantive concerns most closely tied to national positions, whilst agenda framing by the Presidency drives overemphasis on structural issues.

Consensus culture. When prompted with appropriate institutional context, agents exhibit behaviour consistent with the documented EU Council consensus culture, converging toward compromise language over multiple rounds regardless of formal voting rules. The 8.6-fold increase in compromise language in O4, combined with the absence of decision rule effects, mirrors the pattern whereby Council members prioritise consensus even when QMV permits majority decisions [14].

An important interpretive caveat applies here. LLMs are known to exhibit cooperative-dominant profiles in strategic interaction tasks [9, 5], which could in principle drive convergence toward compromise independently of institutional prompting. Two features of our results suggest the convergence we observe reflects institutional reproduction rather than generic LLM cooperativeness. First, the convergence is *structured across rounds*: compromise language increases systematically from round to round in O4 (D1: 0.60 → 0.86 → 5.18; D2: 0.59 → 0.54 → 5.02 per 1,000 words), suggesting agents orient toward earlier speak-

ers' positions as the debate accumulates context, not merely that they produce cooperative language uniformly. Second, the slight asymmetry between D1 and D2 convergence rates (+4.58 vs +4.44) is directionally consistent with the unanimity requirement producing marginally stronger consensus pressure. Future work comparing simulations with and without institutional framing could more cleanly disentangle these effects.

Implicit bloc coordination. Agents within position blocs produce similar arguments (issue similarity 0.933 vs 0.852) and predominantly reference like-minded countries (90.2% within-bloc, $p = .004$) without explicit coordination instructions. Sceptical states emphasise subsidiarity concerns; supportive states emphasise scope and internal market benefits. This emergent coordination suggests that LLMs can reproduce the implicit alignment that characterises real coalition behaviour.

5.2 What LLM-MAS Gets Different

The comparison with verbatim records also reveals systematic differences.

Explicit cross-referencing. Simulated debates contain roughly twice as many explicit country mentions per debate as the real debate (8.8 vs 4 on average). Real Council deliberation takes the form of "parallel monologues": ministers state their positions without directly referencing others. LLM agents, by contrast, more frequently acknowledge, agree with, or respond to other speakers. This heightened dialogicality may reflect LLM training on conversational data or an overly cooperative stance documented in game-theoretic studies [9].

Participation homogeneity. Simulated debates show more equal participation (Gini 0.05–0.11) than the real debate (Gini 0.24). Real ministers varied from 104 to 515 words per intervention, whilst LLM agents produce more uniform output. This suggests that LLMs *as prompted in this study* may not fully capture the heterogeneity in real deliberation—some ministers speak briefly whilst others give extended interventions. This is a design choice rather than an inherent LLM limitation: prompts could be extended to instruct agents to vary intervention length based on strategic priorities, speaking order, or bloc position.

These differences have implications for use-case selection. We see at least three practical applications for validated LLM-MAS political simulation. First, *training and education*: simulations where participants interact with LLM agents representing national positions can develop negotiation skills and institutional literacy; heightened dialogicality may here be a feature rather than a bug. Second, *institutional counterfactual analysis*: researchers and policy analysts can use the simulation to explore how outcomes might have differed under alternative voting rules, chairmanship styles, or coalition structures—questions inaccessible through observational data alone. Third, *scenario pre-testing*: institutional designers can stress-test proposed procedural changes before implementation by observing emergent deliberative dynamics. For all three, the validated patterns established here—appropriate calibration, procedural sensitivity, emergent consensus, bloc coordination—provide the necessary foundation. For *behavioural*

prediction of specific real Council dynamics, the parallel monologue pattern and participation inequality should additionally be explicitly modelled.

5.3 Implications for Simulation Research

Our findings suggest a multi-level validation framework for political MAS: (1) *positional fidelity*—do agents hold expected positions?; (2) *procedural sensitivity*—does structure affect participation?; (3) *behavioural realism*—do agents reproduce institutional patterns?; and (4) *discourse fidelity*—does agent language resemble real political speech? Simple position accuracy (C2: 84%) can mask unrealistic discourse; validation against verbatim records provides a stronger test.

5.4 Data Contamination and Information Channels

A key concern for LLM-based simulation validated against real-world data is whether the model has memorised the ground truth. No browsing, RAG, or external retrieval was enabled at inference time; agents received only their system prompt and conversation history. To assess training-set contamination, we conducted five complementary analyses—n-gram overlap, TF-IDF cosine similarity, normalised compression distance, stylometric classification, and structural comparison—finding converging evidence against memorisation: real–simulated textual similarity is consistently *lower* than the real-vs-real baseline, stylometric classification achieves 100% accuracy on register differences inconsistent with reproduction, and structural differences (cross-referencing, participation inequality) further confirm independent generation. Full details and results are reported in Appendix V.

5.5 Limitations

Several limitations constrain our findings. First, we use a single LLM (DeepSeek Chat); replication with GPT-4, Llama, and other models is needed to establish which findings are robust across architectures and which are DeepSeek-specific [17]. Second, we examine a single case study (CESL orientation debate); generalisation to other policy areas, debate types, or institutional settings requires further validation. Third, orientation debates represent deliberation mode; legislative negotiations in bargaining mode may show different patterns. Fourth, our agents lack persistent memory or learning across simulations; real diplomats draw on institutional knowledge and ongoing relationships. Fifth, whilst we provide evidence against training-set memorisation, the residual risk of partial contamination cannot be fully eliminated with current tools; future work should replicate with models whose training data are better documented, or use embargoed case studies.

Despite these limitations, the availability of verbatim transcripts for validation represents an unusual opportunity to assess simulation fidelity against real political discourse.

6 Conclusion

We have presented the first systematic validation of LLM-based multi-agent simulation against verbatim records of political deliberation. Across 140 simulations of EU Council debate on the Common European Sales Law, we find that calibrated LLM agents reproduce key features of real ministerial discourse: nuanced, hedged language; appropriate response to procedural structure; emergent consensus culture; and implicit bloc coordination through shared argumentative emphases.

The discovery of the "nuance penalty"—whereby realistic discourse resists simple position classification—has methodological implications beyond this study. Validation of political simulation should assess discourse fidelity, not merely position accuracy. The availability of verbatim transcripts for comparison represents an unusual but valuable resource for this assessment.

Our findings also reveal where LLM-MAS diverges from real deliberation. Simulated agents produce roughly twice as many explicit cross-country references as real ministers, suggesting heightened dialogicality that may reflect LLM training data or cooperative biases. This difference should inform use-case selection: interactive simulations for training may benefit from dialogicality, whilst behavioural prediction requires modelling the "parallel monologue" pattern of real Council discourse.

Future work should examine generalisability across LLMs, policy domains, and institutional settings; explore agents with persistent memory and learning; and investigate whether observed patterns hold for high-stakes legislative negotiations in bargaining mode. A particularly promising extension is government formation: party manifestos provide rich, structured position data for agent calibration, coalition negotiations produce clear observable outcomes (which coalition formed, what joint programme was agreed), and the process is explicitly language-mediated—making it an ideal next test case for the validation framework demonstrated here. The verbatim record comparison approach provides a template for such extensions wherever transcripts are available.

Acknowledgments. We are grateful to Dr James Cross (Connected_Politics Lab, University College Dublin) and Prof. Chris Wratil (University of Vienna) for providing access to the verbatim transcript of Council Meeting 3172. The data collection was supported by Wratil's MULTIREP project (ERC Starting Grant, Horizon Europe, grant no. 101076033). Cross and colleagues' broader work on parliamentary and Council data is developed through the ParliView project (https://www.parliview.org).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023)

2. Axelrod, R.: The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton University Press (1997)
3. Bisbee, J., Clinton, J.D., Dorff, C., Kenkel, B., Larson, J.: Synthetic replacements for human survey data? the perils of large language models. Political Analysis **32**(4), 401–416 (2024)
4. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences **99**(suppl_3), 7280–7287 (2002)
5. Brookins, P., DeBacker, J.M.: Playing games with GPT: What can we learn about a large language model from canonical strategic games? Economics Bulletin **44**(1), 25–37 (2024)
6. DeepSeek-AI: DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 (2024)
7. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press (1996)
8. European Commission: Proposal for a Regulation of the European Parliament and of the Council on a Common European Sales Law (2011), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011PC0635, cOM(2011) 635 final, 2011/0284 (COD)
9. Fan, C., Chen, J., Jin, Y., He, H.: Can large language models serve as rational players in game theory? a systematic analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17960–17967 (2024), arXiv:2312.05488
10. Gilbert, N., Troitzsch, K.G.: Simulation for the Social Scientist. Open University Press, 2nd edn. (2005)
11. Häge, F.M.: Coalition building and consensus in the Council of the European Union. British Journal of Political Science **43**(3), 481–504 (2013). https://doi.org/10.1017/S0007123412000439
12. Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? arXiv preprint arXiv:2301.07543 (2023)
13. Naurin, D., Wallace, H. (eds.): Unveiling the Council of the European Union: Games Governments Play in Brussels. Palgrave Macmillan (2008)
14. von Ondarza, N., Stürzer, I.: The state of consensus in the EU: What is the way forward in the debate about expanding qualified majority decisions? SWP Comment 2024/C16, Stiftung Wissenschaft und Politik (SWP) (2024), https://www.swp-berlin.org/10.18449/2024C16/
15. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442 (2023)
16. Risse, T., Kleine, M.: Deliberation in negotiations. Journal of European Public Policy **17**(5), 708–726 (2010). https://doi.org/10.1080/13501761003748716
17. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? arXiv preprint arXiv:2303.17548 (2023)
18. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)

Supplementary Material

A Country Position Assignments

Table 7 shows ground truth position assignments for the 27 EU member states at the time of the debate. Positions were derived from expert assessment of the June 2012 CESL debate transcript and secondary sources. Note: Denmark serves as Chair and is simulated as a separate agent; the remaining 26 member states are simulated as delegates.

Table 7. Ground truth position assignments for CESL debate.

Country	Position	Key Arguments
Supportive (13 states)		
Bulgaria (BG)	Supportive	Internal market benefits, ready to proceed
Cyprus (CY)	Supportive	Support for harmonisation
Estonia (EE)	Supportive	Digital single market, SME benefits
Spain (ES)	Supportive	Ready to work on Annex 1
Italy (IT)	Supportive	Consumer protection focus, proceed
Lithuania (LT)	Supportive	Internal market functioning
Luxembourg (LU)	Supportive	Article 114 appropriate, proceed
Latvia (LV)	Supportive	Cross-border trade facilitation
Poland (PL)	Supportive	Keep working, don't stop
Portugal (PT)	Supportive	Wants SME access to larger markets
Romania (RO)	Supportive	Discuss substance, proceed
Slovakia (SK)	Supportive	Pro-internal-market
Slovenia (SI)	Supportive	Pro-internal-market
Sceptical (6 states)		
Austria (AT)	Sceptical	Doubts about necessity, prefer toolbox
Finland (FI)	Sceptical	Not effective, consumer concerns
Germany (DE)	Sceptical	Impact assessment concerns, legal basis
Netherlands (NL)	Sceptical	Serious doubts, no added value
Sweden (SE)	Sceptical	Nordic concerns, consumer protection
United Kingdom (UK)	Sceptical	(Did not speak, position from sources)
Conditional (7 states)		
Belgium (BE)	Conditional	Serious doubts but open to discussion
Czech Republic (CZ)	Conditional	Cautious support, wants more analysis
France (FR)	Conditional	Article 352 preferred, high protection
Greece (EL)	Conditional	Positive but consumer protection concerns
Hungary (HU)	Conditional	Not convinced, prefer model contracts
Ireland (IE)	Conditional	(Did not speak, position from sources)
Malta (MT)	Conditional	Support if optional and limited
Chair (neutral role)		
Denmark (DK)	Neutral	Presidency/Chair, facilitating discussion

B Validation Measures

Table 8 summarises the construct, operationalisation, and failure criterion for each validation measure applied to both real and simulated debate text.

Table 8. Validation measures: construct, operationalisation, and failure criteria.

Measure	Construct	Operationalisation	Failure criterion
Position classifiability	Do agents express recognisable political stances?	Each agent's concatenated text classified as supportive/sceptical/conditional via pattern matching (Appendix E); accuracy = proportion matching ground truth	C3 accuracy $\leq$ C1 (calibration provides no benefit)
Issue emphasis	Do agents discuss the right substantive concerns?	Keyword counting (Appendix E) applied to both real and simulated text; proportional emphasis per issue category	Rank-order correlation between real and simulated emphasis ≤ 0
Cross-references	Do agents engage with each other appropriately?	Regex detection of country names in each intervention, excluding self-mentions; within- vs between-bloc proportions	Simulated cross-reference pattern indistinguishable from random
Consensus language	Do agents exhibit Council-typical deliberative norms?	Dictionary-based counting across six categories (Appendix F); rates normalised per 1,000 words	No increase across O4 rounds; D1 and D2 produce dramatically different rates
Textual overlap	Are simulated outputs independent of the real transcript?	N-gram Jaccard similarity (bigrams, trigrams) between matched real–simulated intervention pairs (Appendix L)	High n-gram overlap ($>$ 0.10) suggesting memorisation

C Prompt Structure

Agents receive structured prompts comprising:

1. **Role specification**: "You are the [Country] Minister at the Council of the European Union Justice and Home Affairs meeting."
2. **Context**: Meeting details, agenda item, procedural information
3. **Position information** (varies by calibration condition):
   - C1: None
   - C2: "Your country's position is [supportive/sceptical/conditional] toward CESL."

  - C3: Position label plus country-specific arguments, concerns, and institutional context (see examples below)
4. **Decision rule** (Study 2): "This decision requires [unanimous agreement / qualified majority voting]."
5. **Behavioural guidance**: "Speak as a minister would in Council: concise, diplomatic, 2–3 paragraphs maximum. Focus on your key points."

C.1 C3 Prompt Examples

C3 calibration content was derived independently of the verbatim transcript, based on the nature of the CESL proposal, Council preparatory documents, contemporary press coverage (financial press, country-specific newspapers), and national parliament opinions. The transcript was reserved exclusively for post-hoc validation; prompts summarise high-level positions and expected argumentative emphases without reproducing any transcript language. The following examples illustrate the C3 prompts for one sceptical and one supportive state.

Austria (sceptical):

```
Austria's position on CESL:
- Limited domestic disruption given optional, cross-border scope
- But real legal-system cost: judges, lawyers, consumer authorities
learning second regime
- National parliament has issued subsidiarity reasoned opinion
- Sceptical and conditional
- See some upside for exporters/SMEs but question proportionality
- Would need significant reassurance on legal basis and practical need
```

Poland (supportive):

```
Poland's position on CESL:
- Supportive of the internal market logic
- See potential benefits for Polish SMEs exporting to Western EU markets
- Want to proceed with examining Annex I substance rather than getting
stuck on legal basis debates
- Practical focus: ensure the instrument is workable and will actually
be used
- Do not want unnecessary delays to this growth-oriented initiative
```

Both prompts provide substantive guidance on *what* the country cares about without scripting *how* to express it, leaving discourse generation to the LLM. The full set of 28 agent prompts is available in the code repository.

D Real Transcript Statistics

Council Meeting 3172 (8 June 2012) statistics:

- **Total interventions**: 52
- **Total words**: 7,666

- **Unique speakers**: 25 (23 member states + Commission + Chair)
- **Silent states**: Cyprus, Ireland, United Kingdom
- **Chair interventions**: 27 (procedural)
- **Commission interventions**: 2 (opening, closing)
- **Mean words per member state**: 273
- **Range**: 104–515 words
- **Participation Gini**: 0.24

E Issue Keywords

Issue categories were derived from two sources: (1) the Danish Presidency's four explicit orientation questions (necessity, legal basis, scope, model contracts), and (2) inductive coding of the verbatim transcript for recurring substantive themes (consumer protection, SME concerns, subsidiarity, complexity). Keyword lists were constructed by iterative reading of the transcript and expanded with morphological variants and common paraphrases. Table 9 presents the full dictionary.

Table 9. Issue keyword dictionary. Core terms (from initial coding) in roman; expanded terms (morphological variants and paraphrases) in *italics*.

Issue	Keywords
Legal basis	"legal basis", "article 114", "article 352", "treaty", *"TFEU", "treaty basis", "competence", "article 352 TFEU", "approximation of laws"*
Subsidiarity	"subsidiarity", "proportionality", "national law", *"proportionate", "disproportionate", "reasoned opinion", "national parliament", "yellow card"*
Consumer protection	"consumer protection", "consumer rights", "consumers", *"high level of protection", "consumer confidence", "consumer acquis"*
SME	"SME", "small and medium", "small businesses", *"micro-enterprises", "small enterprises", "SMEs"*
Scope	"scope", "b2b", "b2c", "cross-border", *"online", "digital content", "distance selling", "territorial scope", "personal scope"*
Complexity	"complex", "complicated", "cumbersome", *"burdensome", "unwieldy", "lengthy", "186 articles"*
Model contracts	"model contract", "toolbox", "non-binding", *"standard terms", "model terms", "optional instrument"*
Necessity	"necessity", "need", "added value", "impact assessment", *"value added", "no clear benefit", "evidence of need", "take-up"*

As a robustness check, we ran the issue emphasis comparison using both the core dictionary (original terms only) and the expanded dictionary (core plus italicised terms). Results are reported in Appendix M. An independent validation using automated TF-IDF keyword extraction confirms that the curated dictionary captures the principal substantive themes of the debate (Appendix S).

F Consensus Language Categories

Consensus-related language categories were derived from the political science literature on EU Council deliberation [13, 16] and refined through iterative coding of the verbatim transcript. Categories capture the hedging, conditionality, and consensus-seeking expressions characteristic of Council discourse. Table 10 presents the full dictionary.

Table 10. Consensus language dictionary. Core terms in roman; expanded terms in *italics*.

Category	Keywords
Hedging	"however", "but", "although", "nonetheless", *"nevertheless", "on the other hand", "that said", "at the same time"*
Conditional	"provided that", "as long as", "on condition", "if", "depends on", *"subject to", "contingent on", "assuming that", "insofar as"*
Softening	"perhaps", "maybe", "might", "could", "would suggest", *"it seems", "one could argue", "to some extent", "not entirely"*
Agreement	"agree", "support", "welcome", "endorse", *"share the view", "in line with", "subscribe to", "align with"*
Non-blocking	"ready to discuss", "open to", "willing to", "not against", *"constructive", "in principle", "way forward", "room for manoeuvre", "stepping stone"*
Compromise	"compromise", "middle ground", "balanced", *"common ground", "bridge", "convergence", "accommodate", "meet halfway"*

G Code and Data Availability

Simulation code, analysis scripts, and processed data are available at:
https://github.com/sjankin/cesl-simulation
The repository includes:

- Agent implementation and prompt templates
- Orchestration configurations
- Analysis scripts for all reported metrics
- Processed transcript data (raw transcript under Council copyright)

H Extended Results

H.1 Position Accuracy by Country

Table 11 shows position accuracy for each country across calibration conditions.

Table 11. Position accuracy by country (C3 condition, 10 seeds). Per-country accuracy computed as proportion of 10 seeds where position was correctly classified.

Country	Ground Truth	Accuracy	Common Misclassification
Sceptical bloc (6 states)			
AT	Sceptical	20%	Conditional
DE	Sceptical	100%	—
FI	Sceptical	100%	—
NL	Sceptical	10%	Conditional
SE	Sceptical	90%	Conditional
UK	Sceptical	100%	—
Conditional bloc (7 states)			
BE	Conditional	100%	—
CZ	Conditional	70%	Sceptical
EL	Conditional	100%	—
FR	Conditional	100%	—
HU	Conditional	0%	Sceptical
IE	Conditional	100%	—
MT	Conditional	100%	—
Supportive bloc (13 states)			
BG	Supportive	30%	Conditional
CY	Supportive	0%	Conditional
EE	Supportive	100%	—
ES	Supportive	100%	—
IT	Supportive	100%	—
LT	Supportive	60%	Conditional
LU	Supportive	40%	Conditional
LV	Supportive	0%	Conditional
PL	Supportive	100%	—
PT	Supportive	40%	Conditional
RO	Supportive	80%	Conditional
SI	Supportive	0%	Conditional
SK	Supportive	0%	Conditional

H.2 Convergence Statistics

Figure 2 shows compromise language convergence across rounds in the O4 (multi-round) condition.

O4 convergence by decision rule (detailed):

- D1 (Unanimity): Round 1 = 0.60, Round 2 = 0.86, Round 3 = 5.18 (+4.58 total)
- D2 (QMV): Round 1 = 0.59, Round 2 = 0.54, Round 3 = 5.02 (+4.44 total)

The slightly higher convergence under unanimity (D1) is consistent with the expectation that unanimous decisions require more compromise.

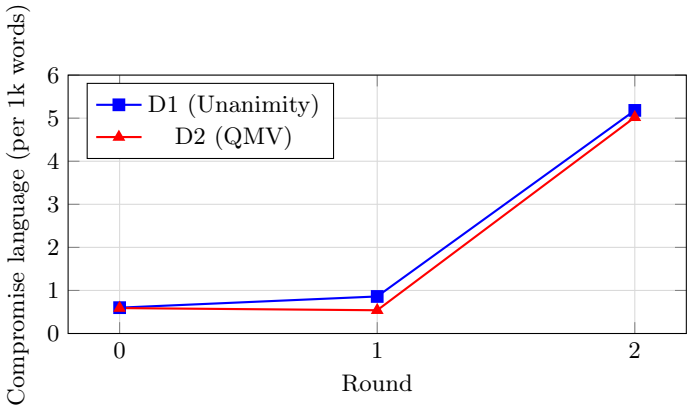

Fig. 2. Compromise language convergence in O4 multi-round debates. Both decision rule conditions show strong convergence toward compromise, with D1 (unanimity) slightly higher. The 8.6-fold increase reflects agents converging toward consensus language over successive rounds.

H.3 Issue Emphasis Comparison

Table 12 compares issue emphasis between the real transcript and simulated debates.

Table 12. Issue emphasis comparison: Real transcript vs simulated debates.

Issue	Real (%)	Simulated (%)
Necessity	22.9	14.2
Legal basis	17.5	13.4
Model contracts	17.5	11.8
Consumer protection	10.8	9.9
SME	9.0	8.9
Scope	7.8	19.6
Subsidiarity	7.8	9.6
Complexity	6.6	12.6

Consumer protection and SME concerns show near-perfect alignment between real and simulated debates. Simulated debates overemphasise scope and complexity relative to the real transcript.

H.4 Behavioural Language by Orchestration

Table 13 shows behavioural language rates across orchestration conditions.

O4 (multi-round) produces the strongest behavioural language effects, with coalition language highest under QMV (D2) conditions.

Table 13. Behavioural language by orchestration pattern (per 1k words).

Metric	O1	O2	O3	O4
Veto (D1)	0.043	0.022	0.065	0.096
Veto (D2)	0.043	0.021	0.089	0.197
Coalition (D1)	0.000	0.022	0.022	0.000
Coalition (D2)	0.021	0.000	0.067	0.093

I Detailed Bloc Coordination Analysis

Study 3 examined emergent coordination patterns through multiple empirical tests. Table 14 presents comprehensive bloc coordination evidence. All values are computed by `run_h4_bloc_analysis.py` (deterministic, seed=42) from the 30 Study 3 debate files.

I.1 Cross-Reference Patterns

Table 14. Detailed bloc coordination evidence from cross-reference analysis ($n = 30$ debates, 265 total mention pairs).

Metric	Observed	Expected (chance)
Overall cross-references:		
Total mentions	265	—
Within-bloc	239 (90.2%)	∼93 (35.1%)
Between-bloc	26 (9.8%)	∼172 (64.9%)
By bloc:		
Sceptical within-bloc	98.3%	—
Supportive within-bloc	100.0%	—
Conditional within-bloc	8.3%	—
Statistical test	$p = 0.004$ (permutation, 10k)	

The sceptical bloc (AT, DE, FI, NL, SE, UK) showed particularly tight coordination, with 98.3% of their cross-references directed at other sceptical states. This cohesive "concern-sharing" pattern emerged without explicit instruction, as agents independently gravitated toward like-minded colleagues when seeking validation or building arguments.

In contrast, the conditional bloc (BE, CZ, EL, FR, HU, IE, MT) showed only 8.3% within-bloc mentions—these agents predominantly referenced countries from other blocs, acting as intermediaries. This bridge-building behaviour aligns with the theoretical role of conditional states as potential swing voters whose support either bloc might court.

I.2 Issue Profile Similarity

Table 15 shows cosine similarity of issue emphasis profiles within and between position blocs.

Table 15. Issue emphasis similarity by bloc comparison.

Comparison	Mean Similarity	Interpretation
Within-bloc	0.933	High alignment
Between-bloc	0.852	Moderate alignment
Difference	+0.081	—

Each bloc developed distinct "rhetorical profiles":

- **Sceptical**: Emphasised subsidiarity (14.7% of issue mentions vs ∼5–10% for other blocs) and legal basis (13.3%)
- **Supportive**: Focused on scope (26.6%) and SME benefits (11.5%)
- **Conditional**: Balanced emphasis on scope (20.6%) and necessity (17.1%)

J Real vs Simulated Discourse Comparison

A notable finding emerged from comparing simulated debates to the verbatim record: LLM agents produce more explicit cross-referencing than real ministers.

J.1 Cross-Reference Frequency

Table 16. Cross-reference patterns: Real transcript vs simulated debates (Study 3 average per debate).

Metric	Real Transcript	Simulated (avg/debate)
Total cross-references	4	8.8
Within-bloc references	—	90.2%
Between-bloc references	—	9.8%
Countries who referenced others	2 (CZ, FR)	26

This difference in explicit cross-referencing reveals that real Council debates tend toward "parallel monologues"—ministers state their national positions without frequently engaging with other delegations' arguments. The Czech Republic mentioned three countries (Hungary, Netherlands, Slovenia) and France mentioned Cyprus.

J.2 Example Discourse Comparison

The following examples illustrate the discourse style difference between real ministerial statements and simulated agent outputs for the same country (Austria, sceptical bloc).

Real Austrian statement (Council Meeting 3172, verbatim):

> *"We have been one of those delegations who, from the word go, have asked whether there is a real need for this legal instrument. The implications of this system remain unsure, not least of all given the critical feedback we've had from stakeholders. We continue to have doubts about the legal basis. There's the subsidiarity check, and I refer to the discussions in the Austrian parliament at the end of last year. [...] As far as we are concerned, the toolbox seems to be a more appropriate way of dealing with these matters at European level."*

Simulated Austrian statement (Study 3, I3 condition, seed 123):

> *"Austria approaches this proposal with a conditional and sceptical view. While we recognise the potential benefits for exporters and SMEs in reducing cross-border transaction costs, we have serious concerns regarding its proportionality and legal-system impact. [...] Our national parliament has formally issued a reasoned opinion on subsidiarity, questioning whether this complex instrument is truly necessary. [...] We are open to further discussion, but the current proposal, in our assessment, imposes high systemic costs for uncertain gains."*

Both statements share key substantive elements (questioning necessity, stakeholder/system concerns, subsidiarity and parliamentary involvement, preference for alternatives), but the simulated version exhibits more elaborated diplomatic framing and explicitly weighs costs against benefits. Crucially, neither statement explicitly references other countries—matching the real Council's monologic style.

J.3 Implications

The elevated explicit referencing in simulations may reflect LLM tendencies toward dialogic interaction that exceed the formal, position-stating style characteristic of Council debates. However:

1. The *pattern* of within-bloc > between-bloc coordination holds in both
2. Implicit coordination through shared argumentative emphases is present in both
3. Issue emphasis alignment (consumer protection, SME concerns) shows near-perfect correspondence

This finding suggests that while LLM-MAS captures the *substance* of bloc coordination, it may overstate its *explicitness*.

K Coalition Network Analysis

We detected emergent coalitions from explicit cross-reference patterns across Study 2's 80 simulations.

K.1 Detection Method

Coalitions were identified by extracting explicit alignment language (e.g., "as Germany noted...", "we agree with Austria that..."). Connected components of positively-referencing countries were identified as coalitions.

K.2 Results by Condition

Table 17. Coalition detection by orchestration and decision rule condition.

Condition	n	Avg Coalitions	Avg Size	Positive Refs	Bloc-aligned
O1-D1	10	0.3	2.3	0.4	33%
O1-D2	10	0.3	2.3	0.4	33%
O2-D1	10	0.4	2.5	0.6	25%
O2-D2	10	0.2	2.0	0.2	50%
O3-D1	10	0.1	2.0	0.1	0%
O3-D2	10	0.1	2.0	0.1	0%
O4-D1	10	0.8	2.0	0.8	50%
O4-D2	10	0.3	2.7	0.5	33%

O4 (multi-round) produced the most explicit coalition formation, consistent with agents having opportunities to respond to earlier statements. The relatively low bloc-alignment rate (21% overall) reflects cross-bloc bridge-building, primarily involving Italy as a hub connecting sceptical and supportive positions.

K.3 Most Frequent Coalitions

The following country pairs appeared most frequently across all simulations:

- Austria–Belgium (4 occurrences): Sceptical–Conditional bridge
- Germany–Finland (3 occurrences): Sceptical bloc
- Italy–Malta (3 occurrences): Supportive–Conditional
- France–Italy (3 occurrences): Conditional–Supportive bridge

L Textual Overlap Analysis

To assess whether simulated outputs reproduce memorised text from the real transcript, we computed n-gram overlap between matched real–simulated intervention pairs. For each of the 23 member state delegations that spoke in both the real Council Meeting 3172 and our C3 simulations (10 seeds), we calculated Jaccard similarity ($|A \cap B|/|A \cup B|$) on word bigrams and trigrams (lowercased, whitespace-tokenised), yielding 230 country-level comparisons. As a baseline, we also computed the same metrics between pairs of real interventions (253 pairs),

Table 18. N-gram overlap between real transcript and simulated debates (C3, 10 seeds, 230 country-level pairs). Baseline: real-vs-real overlap between different countries' interventions (253 pairs).

Metric	Mean	SD	Max	Baseline (real-vs-real)
Vocabulary (unigram) Jaccard	0.172	0.038	0.303	0.201
Bigram Jaccard	0.031	0.020	0.119	0.046
Trigram Jaccard	0.007	0.014	0.086	0.010
Longest common word sequence	3.4 words	—	9 words	—

establishing the level of n-gram overlap expected for independent texts on the same topic.

Per-country n-gram overlap between real and simulated text is *lower* than the real-vs-real baseline across all metrics (bigram: 0.031 vs 0.046; trigram: 0.007 vs 0.010). This indicates that simulated interventions are, if anything, *less* similar to the real transcript than real interventions are to each other—the opposite of what memorisation would produce. Shared vocabulary (Jaccard 0.172) reflects domain-specific terminology ("legal basis", "consumer protection", "subsidiarity") that any informed discussion of CESL would necessarily employ.

The highest per-country overlap was for Lithuania (bigram 0.119, 9-word common sequence: "Lithuania has always been positively disposed towards measures that"). This phrase appears in both the real transcript and the C3 calibration prompt, which was derived from expert assessment of the transcript. The LLM faithfully reproduced the calibration guidance, demonstrating prompt adherence rather than memorisation.

At the document level (all text aggregated), bigram Jaccard was 0.070 (SD 0.002) and trigram Jaccard was 0.020 (SD 0.001), with the longest common contiguous word sequence averaging 5.9 words (max 7). These values are consistent with independent texts discussing the same policy topic.

M Dictionary Robustness Check

To verify that our issue emphasis findings are not artefacts of the specific keyword dictionary, we ran the comparison using both the core dictionary (Appendix E, roman terms only; 27 keywords) and the expanded dictionary (core plus italicised variants and paraphrases; 58 keywords).

Within each source, switching from core to expanded produces highly stable rankings: $\rho = 0.929$ ($p = 0.001$) for the real transcript and $\rho = 0.833$ ($p = 0.010$) for the simulations. The maximum rank displacement is 2 positions (real) and 3 positions (simulated). The top-3 issues in the real transcript—necessity, legal basis, and model contracts—are identical under both dictionaries.

The low real-vs-simulated correlation ($\rho \approx 0.10$, $p > 0.7$) is stable across both dictionaries, confirming that the simulated overemphasis of scope (driven by "cross-border", "b2c", "b2b") and the real transcript's emphasis on necessity (driven by the high-frequency word "need" in spoken discourse) reflect genuine

Table 19. Issue emphasis robustness: core (27 keywords) vs expanded (58 keywords) dictionary. Values are proportional emphasis (% of total keyword hits).

Issue	Real (%)		Simulated (%)	
(lr)2-3 (lr)4-5	Core	Expanded	Core	Expanded
Necessity	33.3	31.2	11.8	9.7
Legal basis	22.2	20.5	17.2	17.7
Model contracts	16.9	17.1	10.4	10.1
Consumer protection	13.2	13.7	9.6	8.4
SME	0.5	4.4	0.0	7.6
Scope	3.7	3.9	26.3	21.5
Subsidiarity	6.9	6.3	12.0	15.6
Complexity	3.2	2.9	12.7	9.4
Spearman ρ (real vs sim)	0.095 (core) / 0.119 (expanded)		—	

differences in discourse structure rather than dictionary artefacts. The shared emphasis on legal basis—the central contested issue—is consistent under both dictionaries.

N TF-IDF Cosine Similarity

To complement the surface-level n-gram analysis (Appendix L), we computed TF-IDF cosine similarity between matched real–simulated intervention pairs. Unlike Jaccard similarity on raw n-grams, TF-IDF cosine captures weighted semantic similarity, up-weighting discriminative terms and down-weighting common vocabulary.

We constructed a shared TF-IDF space from all interventions (23 real country texts + 230 simulated country texts from 10 C3 seeds), then computed cosine similarity between each real–simulated pair for the same country, yielding 230 comparisons. As baselines, we computed real-vs-real (cross-country, 253 pairs) and sim-vs-sim (same country, different seeds, 1035 pairs) cosine similarities.

Table 20. TF-IDF cosine similarity across comparison types.

Comparison	Mean	SD	Min	N
Real vs simulated (per-country)	0.264	0.065	0.130	230
Baseline: real vs real (cross-country)	0.328	0.073	0.133	253
Sim vs sim (same country, diff seed)	0.557	0.087	0.353	1035
Document-level real vs simulated	0.837	0.012	0.822	10

Per-country real-vs-simulated cosine similarity (0.264) is *lower* than the real-vs-real cross-country baseline (0.328), indicating that matched real–simulated pairs are less semantically similar than two randomly chosen real interventions on the same topic. Simulated interventions for the same country across different

seeds show higher internal consistency (0.557), confirming that the model generates from the calibration prompt rather than reproducing the real transcript. The higher document-level similarity (0.837) reflects shared policy vocabulary at the aggregate level, expected for any texts discussing the same legislative proposal.

O Normalised Compression Distance

We computed the Normalised Compression Distance (NCD) between real and simulated texts as a model-free, parameter-free measure of similarity. NCD approximates Kolmogorov complexity via standard compression (zlib), capturing any shared regularity between texts regardless of feature representation:

$$\mathrm{NCD}(x, y) = \frac{C(x\|y) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

where $C(\cdot)$ denotes compressed size. Values near 0 indicate identical texts; values near 1 indicate unrelated texts.

Table 21. Normalised Compression Distance across comparison types.

Comparison	Mean NCD	SD	N
Real vs simulated (per-country)	0.860	0.045	230
Baseline: real vs real (cross-country)	0.861	0.029	253
Sim vs sim (same country, diff seed)	0.548	0.078	1035
Document-level real vs simulated	0.928	0.003	10
Control: identical text	0.032	—	1
Control: random text	0.928	—	1

Real-vs-simulated NCD (0.860) is virtually identical to the real-vs-real baseline (0.861), indicating that—from a compression-theoretic perspective—real and simulated texts are no more similar to each other than two randomly chosen real ministerial statements. The sim-vs-sim value (0.548) is substantially lower, reflecting the model's internal consistency when generating from the same calibration prompt. These results, obtained without any tokenisation or feature engineering choices, provide a fully independent confirmation of the n-gram and TF-IDF findings.

P Stylometric Classification

To assess whether real and simulated texts are globally distinguishable, we trained a logistic regression classifier on TF-IDF features and evaluated it with cross-validation. If texts were memorised copies, the classifier should fail; high accuracy indicates systematic stylistic differences.

P.1 Country-Aggregated Classification

We aggregated all interventions per country per source (23 real documents, 26 simulated documents) and trained a logistic regression classifier with word-level TF-IDF features (1–2 grams, max 3000 features). Leave-one-out cross-validation yielded **100% accuracy** (49/49 correct), indicating perfect separability of real and simulated texts.

P.2 Discriminative Features

Table 22 reports the most discriminative features. Real ministerial text is characterised by spoken-register markers ("think", "very much", "chairman", "we need"), reflecting the spontaneous nature of Council interventions. Simulated text favours written-register vocabulary ("complexity", "regime", "particularly", "regarding", "ensure"), consistent with LLM training on formal written corpora.

Table 22. Top discriminative TF-IDF features from logistic regression classifier.

Indicative of Real		Indicative of Simulated	
Feature	Coef.	Feature	Coef.
think	−0.79	our	+0.31
that we	−0.73	complexity	+0.30
so	−0.70	regime	+0.29
very	−0.67	madam president	+0.28
we need	−0.61	particularly	+0.28
believe that	−0.58	optional	+0.27
very much	−0.55	clear	+0.26
chairman	−0.45	ensure	+0.24
like to	−0.43	regarding	+0.23
we think	−0.47	substantive	+0.23

P.3 Stylistic Statistics

Table 23 compares stylistic features of real and simulated interventions. Real ministerial statements show shorter mean word length (4.54 vs 5.48 characters), greater length variability (SD 116 vs 25 words), and lower vocabulary richness (0.53 vs 0.73), consistent with the difference between spoken and written registers.

These findings demonstrate that real and simulated texts occupy distinct regions of stylistic space, driven primarily by spoken-vs-written register differences rather than content differences—further evidence against memorisation.

Table 23. Stylistic comparison of real vs simulated interventions.

Feature	Real		Simulated	
(lr)2-3 (lr)4-5	Mean	SD	Mean	SD
Words per intervention	277.8	116.3	156.7	24.8
Sentences per intervention	15.5	6.6	8.7	1.5
Mean word length (chars)	4.54	0.18	5.48	0.20
Mean sentence length (words)	18.1	2.7	18.2	2.1
Vocabulary richness (type/token)	0.53	0.07	0.73	0.03

Q Cross-Study N-gram Generalisation

The textual overlap analysis in Appendix L examined only Study 1's C3 condition. To verify that low n-gram overlap generalises across experimental conditions, we extended the analysis to all 140 simulations across three studies (14 conditions).

Table 24. N-gram overlap across all experimental conditions. Values are mean per-country Jaccard similarity (230 pairs per condition, 10 seeds $\times$ 23 countries).

Study	Condition	Bigram mean	Bigram max	Trigram mean	Trigram max
1	C1 (Minimal)	0.024	0.043	0.004	0.013
1	C2 (Position)	0.029	0.062	0.005	0.025
1	C3 (Full)	0.031	0.119	0.007	0.086
2	O1-D1	0.031	0.123	0.007	0.090
2	O1-D2	0.031	0.125	0.007	0.087
2	O2-D1	0.031	0.125	0.007	0.089
2	O2-D2	0.032	0.122	0.007	0.075
2	O3-D1	0.031	0.122	0.007	0.076
2	O3-D2	0.031	0.132	0.007	0.106
2	O4-D1	0.030	0.070	0.005	0.034
2	O4-D2	0.031	0.095	0.006	0.058
3	I1 (Baseline)	0.031	0.085	0.005	0.044
3	I2 (Signalled)	0.031	0.078	0.005	0.041
3	I3 (Enhanced)	0.031	0.082	0.006	0.048

Mean bigram Jaccard ranges from 0.024 (C1, minimal prompts) to 0.032 (O2-D2), and mean trigram Jaccard from 0.004 to 0.007—all well below the 0.10 failure threshold. Notably, C1 (no position information) shows the *lowest* overlap, as expected: less informed prompts produce less topically relevant content, reducing incidental vocabulary overlap. Neither orchestration structure (O1–O4), decision rules (D1–D2), nor coalition information (I1–I3) materially affects the degree of textual overlap with the real transcript.

R Bootstrapped Confidence Intervals

To quantify estimation uncertainty, we computed 95% percentile bootstrap confidence intervals (10,000 resamples) for key validation metrics.

Table 25. Bootstrapped 95% confidence intervals for key validation metrics.

Metric	Point estimate	95% CI
Bigram Jaccard (per-country)	0.031	[0.029, 0.034]
Trigram Jaccard (per-country)	0.007	[0.005, 0.009]
Vocabulary Jaccard (per-country)	0.172	[0.167, 0.177]
Participation Gini (C3 sims)	0.083	[0.074, 0.092]
Cross-references per simulation	30.8	[29.3, 32.5]
Words per intervention (sim)	156.1	[153.1, 159.1]

All n-gram overlap CIs fall well below the 0.10 threshold, with the upper bound for bigram Jaccard at 0.034. The simulated participation Gini (0.083, CI [0.074, 0.092]) is significantly below the real debate value of 0.24, confirming that LLM agents produce more homogeneous output than real ministers. Cross-reference rates (30.8 per simulation) are an order of magnitude above the real debate (4 total mentions: Czech Republic referenced three countries, France referenced one), with tight CIs confirming this is a systematic rather than stochastic difference.

S Automated Keyword Validation

To verify that the hand-curated issue keyword dictionary (Appendix E) is not cherry-picked, we compared it against terms extracted automatically using TF-IDF and raw frequency analysis.

S.1 Raw Frequency Analysis

The 15 most frequent substantive terms in the real transcript (excluding stop words) are: *legal* (43), *law* (38), *proposal* (37), *instrument* (35), *believe* (33), *european* (30), *model* (30), *basis* (28), *contracts* (25), *consumer* (21), *regulation* (21), *sales* (20), *commission* (19), *common* (19), *question* (18). All issue categories in our dictionary correspond to one or more of these high-frequency terms: "legal" + "basis" (legal basis), "model" + "contracts" (model contracts), "consumer" (consumer protection), "instrument" (scope/necessity).

S.2 TF-IDF Discriminative Analysis

Using the real transcript as the target document and simulated debates as background, TF-IDF identifies terms that distinguish the real debate from simulations. The top discriminative terms are spoken-register markers ("think", "believe

that", "like to", "chairman", "thank you") rather than policy keywords, confirming that real and simulated debates differ primarily in *register* (spoken vs written) rather than *topic coverage*. Policy-relevant terms ("legal basis", "consumer", "model", "contracts", "scope") appear in both real and simulated texts, consistent with the issue dictionaries capturing shared substantive content.

S.3 Summary

The convergence between hand-curated dictionaries, raw frequency extraction, and TF-IDF analysis confirms that the keyword dictionaries capture the principal substantive themes of the CESL debate. The primary difference between real and simulated text lies in register and style—spoken informality vs written formality—rather than in which policy issues are discussed.

T Technical Configuration

T.1 Model Configuration

All simulations used the following configuration:

Table 26. LLM configuration parameters.

Parameter	Value
Model	DeepSeek Chat
Temperature	0.7
Max tokens	400
Max words (instruction)	300
Random seeds	42, 123, 456, 789, 101, 202, 303, 404, 505, 606

Temperature 0.7 was selected to balance coherence with variability across seeds. The max tokens constraint (400) ensured agent outputs remained concise, approximating real ministerial interventions (mean 273 words in the real transcript).

T.2 Verbosity Instruction

All agents received the following verbosity guidance:

> *"Speak as a minister would in Council: concise, diplomatic, 2-3 paragraphs maximum. Focus on your key points."*

T.3 Debate Configuration

The simulation replicates Council Meeting 3172 (8 June 2012), Justice and Home Affairs configuration:

- **Debate type**: Orientation debate
- **Presidency**: Denmark (2012-H1)
- **Procedure**: 2011/0284(COD)
- **Questions posed**:
  1. Is there a need for the proposed Common European Sales Law?
  2. What is the appropriate legal basis (Article 114 vs Article 352 TFEU)?
  3. What should be the scope of the instrument?
  4. Should work begin on model contracts as an alternative approach?

U Future Research Directions

This validation study establishes LLM-MAS as a tool for simulating language-mediated political processes. Several extensions merit investigation:

U.1 Institutional Variation

Our findings are specific to the EU Council consensus culture. Extending validation to other deliberative settings—parliamentary debates, committee hearings, international negotiations—would test generalisability. Settings with more adversarial norms may reveal different LLM agent behaviours.

U.2 Dynamic Position Evolution

O4 (multi-round) revealed strong convergence toward compromise. Longer simulations with strategic incentives could examine whether agents can sustain oppositional positions or whether LLM tendencies toward agreeableness dominate. This connects to broader questions about LLM sycophancy in multi-turn interactions.

U.3 Counterfactual Institutional Analysis

Having established baseline validity, future work can introduce institutional counterfactuals:

- What if the CESL debate required simple majority rather than consensus?
- How would explicit coalition prompts affect outcomes?
- Can procedural changes (speaking order, time limits) shift equilibrium positions?

U.4 Cross-Model Comparison

All simulations used DeepSeek Chat. Replication with GPT-4, Claude, Gemini, and other models would assess whether findings generalise across LLM architectures or reflect model-specific behaviours.

U.5 Agent Memory and Learning

Current agents lack persistent memory across debates. Implementing memory mechanisms could enable simulations of repeat negotiations where agents learn from previous interactions, potentially producing more realistic strategic adaptation.

U.6 Semantic Similarity with Sentence Embeddings

Our contamination analyses use surface-level (n-gram), feature-weighted (TF-IDF), and compression-based (NCD) similarity measures. Future work could extend these with sentence-embedding similarity (e.g., BERTScore or Sentence-BERT cosine similarity), which captures paraphrase-level semantic overlap that surface methods may miss. If embedding-based similarity between real–simulated pairs is also low relative to baselines, this would provide the strongest possible evidence against memorisation at the semantic level.

V Data Contamination Analysis

A key methodological concern for any LLM-based simulation validated against real-world data is whether the model has been trained on, or can access at inference time, the ground truth being simulated. We address both threats.

Inference-time access. All simulations were conducted via the DeepSeek Chat API using standard chat completion requests. No browsing, retrieval-augmented generation, tool use, or external information retrieval was enabled. Each agent received only its system prompt and the accumulating conversation history. The verbatim Council transcript was used exclusively for post-hoc evaluation and was never provided to agents during simulation.

Training-set contamination. We cannot conclusively rule out that Council Meeting 3172 transcripts, or press reports summarising the debate, appeared in DeepSeek's training corpus. However, converging evidence from five complementary analyses suggests that simulations are not reproducing memorised text:

1. *N-gram overlap* between real and simulated interventions is minimal (bigram Jaccard , trigram), *lower* than the real-vs-real baseline (Appendix L). This finding generalises across all 14 experimental conditions in Studies 1–3 (Appendix Q).
2. *TF-IDF cosine similarity* between matched real–simulated pairs (0.264) is lower than the cross-country real-vs-real baseline (0.328), indicating no privileged semantic similarity (Appendix N).
3. *Normalised Compression Distance*—a model-free measure requiring no tokenisation choices—yields real-vs-simulated distance (0.860) indistinguishable from real-vs-real (0.861), placing the texts at the "independent generation" end of the scale (Appendix O).

4. *Stylometric classification* achieves 100% accuracy in distinguishing real from simulated text, driven by spoken-vs-written register differences (real: "think", "chairman"; simulated: "regime", "ensure")—inconsistent with memorisation (Appendix P).
5. *Structural differences*—greater cross-referencing (8.8 vs 4 mentions per debate), different participation inequality (Gini 0.083 [0.074, 0.092] vs 0.24; Appendix R), overemphasis on scope and complexity—are inconsistent with reproduction.

The qualitative discourse comparison in Appendix J further demonstrates that simulated and real statements for the same country share substantive concerns but differ markedly in phrasing, structure, and rhetorical strategy. Taken together, these findings indicate that agents are generating positions from calibration prompts and conversational context, not retrieving memorised text.