Habemus Papam: Generative Agents in the Sistine Chapel

An LLM-based Simulation of the Papal Conclave

Myriam Belkhatir • Ernani Hazbolatow • Salomé Poulain • Koen Verlaan

MSc Computational Science, University of Amsterdam — Agent-Based Modelling (5284AGBM6Y), Dr D. Roy

30 June, 2025

Abstract

This report presents a simulation framework designed to explore how informal group deliberation among cardinals can influence the outcome of papal elections. Each cardinal is modeled as a generative agent powered by a large language model (LLM), with individual ideological profiles, conversational memory, and the ability to reflect and adapt their stance across rounds. Agents are assigned to small discussion groups where they exchange views, process the dialogue, update their voting intentions, and vote until a candidate reaches a two-thirds majority.

The model extends the ConclaveSim platform with several new features, including a stance-based embedding mechanism to track ideological convergence, and a utility-based group formation algorithm governed by a rationality parameter λ. This mechanism combines social proximity, ideological alignment, and institutional importance to shape group dynamics. Simulations were conducted on a large scale using the Snellius supercomputer, enabling the exploration of 25 parameter combinations involving LLM temperature and rationality.

Results show that temperature has a strong and primarily linear influence on convergence speed and winner predictability, while rationality plays a more subtle, potentially interactive role. Global sensitivity analysis, including Morris and Sobol methods, confirms that temperature dominates the variance in model outputs. Additional visual analyses highlight how stochasticity and agent rationality impact ideological bloc formation, vote-switching behavior, and the overall dynamics of consensus-building in highly constrained institutional settings.

1 Introduction

Papal elections are among the most secretive and complex decision-making processes in the modern world. Despite being highly formalized, they rely heavily on informal interactions, negotiations, and personal dynamics among members of the College of Cardinals.

Understanding how these decentralized conversations influence the eventual outcome remains a challenging task, particularly given the lack of direct observational data and the interplay of political, theological, and social factors.

Recent advances in agent-based modeling and large language models (LLMS) offer new opportunities to simulate these processes with greater realism.

In this work, we build upon the exisiting ConclavesSim framework to develop a more sophiscated simulation of papal elections, where each cardinal is modeled as a generative agent equipped with an ideological profile, memory, and reasoning capabilities.

These agents interact in small groups, process dialogue using LLMs, and iteratively update their voting preferences in an attempt to reach consensus.

Our goal is to explore how variations in key parameters-particularly LLM temperature and agent rationality-affect the collective behavior of the system.

More specifically, we examine how these factors influence convergence speed, polarisation, bloc formation, and the emergence of dominant candidates. By combining local and global sensitivity analysis methods, we aim to better understand the impact of each parameter and uncover potential non-linearities or interactions.

2 Model Description

2.1 Overview

2.1.1 Purpose

The model's purpose is to simulate the consensus-reaching process within a conclave by endowing agents with a structured mechanism for perception and action, orchestrated by an LLM. It is used to analyze how varying core hyperparameters—specifically LLM temperature and agent rationality—affect simulation outcomes like voting patterns and ideological group dynamics.

2.1.2 Entities, State Variables, and Scales

- **Entities** The model consists of Agents (50 cardinals) and an Environment that manages the simulation's state and rules.
- **State Variables**
  - **Agent**: Each agent has a rich internal persona, a dynamic Stance Digest (voting intention and reasoning), a Reflection Digest (synthesis of events), a Role (Candidate or Elector), an Ideology score, and a Network Importance Score (eigenvalue centrality).
  - **Environment**: The environment tracks global vote tallies, agent roles, and the history of pairwise interactions in discussions.
- **Scales**

- **Spatial**: The model uses a detailed network of the Roman Catholic Church as its spatial component, where distances represent social closeness.
- **Temporal**: The simulation proceeds in discrete rounds, running for a maximum of 50 rounds or until a 4-hour time limit is exceeded.

2.1.3 Process Overview and Scheduling

Within each simulation, processes are performed sequentially. The simulation begins with an initial stance generation for each agent to inform the first discussion. Following this, the simulation proceeds in rounds. A single round consists of four sequential phases:

1. **Discuss:** Agents are placed in discussion groups and exchange views. An annotated example of a discussion transcript, illustrating the nature of these interactions, can be found in Appendix B.
2. **Reflect:** Agents process the discussion transcripts to update their internal state.
3. **Generate Stance:** Agents generate an updated Stance Digest based on their reflection.
4. **Vote:** Agents cast their vote based on their new stance.

This loop repeats until a termination condition is met: consensus, the 50-round limit, or the 4-hour time limit (in which case the run is discarded).

2.2 Design Concepts

Theoretical and Emperical Background

- **Model Lineage:** This model is a significant extension of a simpler prototype, the original ConclaveSim framework (Zhu 2025), which used random discussion groups and basic voting logic. While the current model was built upon that backbone, much of the original logic has been replaced with more sophisticated mechanisms.
- **General Concepts:** The agent architecture is inspired by the generative agent architecture by Park et al. (2023). The reflection mechanism, where agents process new information to update their understanding, is analogous to systems like Debate Arena and COPPER (Santos et al. 2022; Bo et al. 2024). Prompt design is grounded in literature on token-efficient prompting and group deliberation (Park et al. 2023; Liu et al. 2024).
- **Empirical Data:** Agent personas are generated from natural-language biographies scraped from the College of Cardinals, in English. This differs from the original ConclaveSim framework, which used Wikipedia-scraped data.
- **Network Foundation:** The model's spatial component is a replication of the network found by Soda et al. (2025). The relevant codebase was obtained from Rizzo (2025). This network was constructed by combining diverse information streams to estimate inter-node weights, including: cardinal assignments to Roman Curia dicasteries, commissions, councils, and academies; spiritual genealogies linking each cardinal to their ordaining bishops; and ideological affinities, mentorships, and patronage connections drawn from authoritative journalistic accounts. A visualization of the network can be found in Appendix C.

2.2.1 Individual Decision-Making

- **Rationality and Objectives:** An agent's objective is role-dependent. Candidates aim to build coalitions, while Electors seek to identify the most suitable and viable candidate.
- **Decision Rules:** Agent decisions are governed by a multi-step cognitive process driven by LLM prompts.
  - **Stance Generation:** The core voting intention (Stance Digest) is generated through a structured reasoning process. The agent is prompted to consider its persona, previous stance, reflection on recent discussions, and the current vote momentum. It is guided to define its primary pick based on its core values, assess its conviction, consider the strategic outlook for its chosen candidate, and evaluate the option to abstain. The final output is a concise summary of this reasoning.
  - **Voting Choice:** The final vote is cast via a role-specific prompt. Candidates are instructed to vote for themselves unless their candidacy is clearly non-viable. Electors are instructed to vote for a viable candidate who aligns with their values and can reach the required vote threshold.

2.2.2 Learning

Learning occurs as agents update their internal state based on new information, though their fundamental decision rules remain static. The learning mechanism is a reflection process. After each discussion round, an agent is prompted to generate a Reflection Digest. The inputs to this process are the agent's persona, its previous stance, and a summary of the recent discussion. The agent is guided to internally consider the persuasiveness of arguments, assess the standing of leading candidates, and evaluate its long-term strategy. This results in a concise internal monologue that informs the next round's stance generation.

2.2.3 Individual Sensing

Agents perceive the environment through structured channels. Their primary inputs are the summarized transcripts of discussions from their assigned group. To assess candidate viability, they also perceive a compact scoreboard which shows the current vote standings for each candidate and a summary of their momentum (i.e., whether they are gaining, stalling, or losing support).

2.2.4 Individual Prediction

The model does not include an explicit mechanism for agents to predict future states. Agent behavior is reactive, based on their memory and current perception of the environment.

2.2.5 Interaction

Agent interactions occur directly within small discussion groups and indirectly through the globally visible vote tally. The formation of these groups is a key part of the model. It is important to note that due to model constraints such as token limits, the discussions function as a series of short, independent speeches rather than a turn-by-turn responsive conversation. While agents reflect on the collective speeches of others, developing more direct conversational interaction is an area for future work.

The group formation process begins by ranking cardinals based on their importance to the Church, assuming more influential individuals have greater agency in selecting discussion partners. A rationality parameter, λ, governs the mixture of two strategies: a baseline of random allocation and a utility-maximizing strategy. A λ value of 0.5, for example, implies that on average, half of the discussion participants are chosen strategically and half are chosen at random.

To prevent stagnation and maximize interaction coverage, a pairwise interaction history is updated after each round. A penalty proportional to the number of times two participants have already been grouped is imposed. This growing penalty prevents any pair from recurring until every possible pairing has occurred at least once.

2.2.6 Collectives

The primary collectives in the model are temporary discussion groups of approximately five agents, which are imposed by the environment each round. These groups serve as the main channel for direct agent-to-agent interaction.

2.2.7 Heterogeneity

Agents are heterogeneous in their unique personas, roles (the 7 most central cardinals are candidates), ideological scores, and network influence. The simulation is initialized with an ideologically balanced subgroup of 50 cardinals.

2.2.8 Stochasticity

Stochasticity is introduced through the LLM's temperature setting and the rationality parameter, which controls the blend of random versus deterministic choices in group formation.

2.2.9 Observation

Key outputs collected include the number of rounds to consensus, the winner, and round-by-round voting records. A secondary analysis technique involves tracking shifts in ideological blocs by converting each agent's textual Stance Digest into a vector using the intfloat/e5-large-v2 embedding model. These embeddings are then visualized in two dimensions using t-SNE, as shown in Appendix A.

2.3 Details

2.3.1 Implementation Details

- **Source code:** The model was implemented in Python as an extension of the open-source ConclaveSim framework, and is available on GitHub. The implementation supports both remote and local deployment of the meta-llama/Llama-3.1-8B-Instruct LLM.
- **Cluster implementation:** For large-scale runs, the code was adapted to execute the local model efficiently on the Snellius national supercomputer, leveraging the h100 partition. Each simulation instance was assigned to a dedicated NVIDIA H100 GPU node, providing 80 GB of memory—sufficient for hosting a this specific LLM instance. Agent inference, including all group discussions, was performed sequentially within each simulation, ensuring that only one model was loaded per node at a time. To maximize reliability and efficiency, each process was restricted to a single visible CUDA device (device 0), ensuring exclusive GPU access for every simulation. Additionally, PyTorch's memory allocator was configured with a maximum split size of 1024 MB for memory blocks. This combination minimized memory fragmentation and reduced the likelihood of out-of-memory errors, particularly during peak demand. The repository includes automated bash scripts for generating SLURM job scripts and organizing output directories to support grid sweeps across parameter combinations.

2.3.2 Initialization

The simulation is initialized with an ideologically balanced subgroup of 50 cardinals. The 7 cardinals with the highest centrality are designated as candidates. This number was chosen to ensure a competitive but manageable field of candidates, and could serve as an interesting parameter for future work. An initial Stance Digest is generated for each agent before the first round; this is done using the main stance generation logic, based primarily on the agent's persona and its perception of the public profiles of the available candidates, as no prior interactions have occurred.

2.3.3 Input Data

The model uses natural-language biographies scraped from the College of Cardinals as input for persona generation, using the English translation. It also uses the network data found by Soda et al. (2025) as a foundational input.

2.3.4 Submodels

- **Utility Function:** The utility of a potential pairing in a discussion group, u_{ij}, is computed by combining four components in a linear combination with equal weights. These components are:
  - **Social Closeness (κ_{ij}):** This is used as a proxy for social closeness and is defined as the inverse of the weighted shortest-path distance (d_{ij}) between two nodes in the network graph.
  - **Ideological Proximity:** Defined as the absolute distance between the ideologies of two cardinals. Cardinals are categorized into one of five ideologies, each assigned a fixed score on a [-1, 1] scale with equal spacing.

- **Importance to the Vatican Church:** This is measured using the inverse of a cardinal's eigenvalue centrality. This represents a strategic cost, as attempting to persuade a highly influential cardinal is considered more difficult or risky.
- **Interaction Term (Importance × Social Closeness)**: This term captures the strategic preference for engaging with cardinals who are not only powerful but also socially accessible, which increases the likelihood of successful persuasion by lowering interaction costs while amplifying the potential payoff.

- **Utility Scaling:** The components of the utility function are first rescaled into a [0, 1] interval using min-max scaling. The final composite utility score is then bounded to lie in the (0, 1) interval by applying a sigmoid transformation.
- **Parameters:**
  - **Grid Search:** A full grid search was performed over 25 combinations of temperature values 0.1, 0.5, 1, 1.5, 2 and rationality (λ) values 0, 0.25, 0.5, 0.75, 1. Temperature was varied from 0.1 to 2.0 to explore a spectrum from near-deterministic to highly creative agent reasoning. Rationality was varied from 0 to 1.0 to analyze behaviors from purely random group allocation to fully utility-maximizing choices.
  - **Fixed Parameters:** For all runs, utility weights in the group formation mechanism were set to 1.0, and the penalty for repeated pairings was fixed at 0.1. Prompts were capped at a maximum of 800 tokens. Discussion groups consisted of five agents each, with each agent required to contribute between 50 and 100 words per discussion turn. The use of equal weights for the utility function is noted as a simplifying assumption and a limitation of the approach.

3 Results

3.1 High-Level Summary of Simulation Outcomes

Across all parameter settings, the simulations show a strong tendency to converge on a single dominant candidate.

1. **Winner Distribution** Out of the 1,275 completed runs, a consensus was reached in every simulation. The distribution of winning candidates was highly skewed towards a single individual:
   - **Luis Antonio Gokim Tagle:** Won 1,169 times (91.69%)
   - **Robert Francis Prevost:** Won 103 times (8.08%)
   - **Lazzaro You Heung-sik:** Won 3 times (0.24%)
2. **Convergence Speed** The time to reach a consensus varied significantly across the simulations. For the 1,275 successful runs, the convergence speed was as follows:
   - **Average:** 9.04 rounds
   - **Median:** 6 rounds
   - **Range:** 1 to 38 rounds
   - **Standard Deviation:** 7.56 rounds

The distribution of rounds to consensus is shown in Figure 1. It is strongly right-skewed, indicating that while most simulations converged relatively quickly (in under 10 rounds), a long tail of outliers required a much greater number of rounds to reach a conclusion.

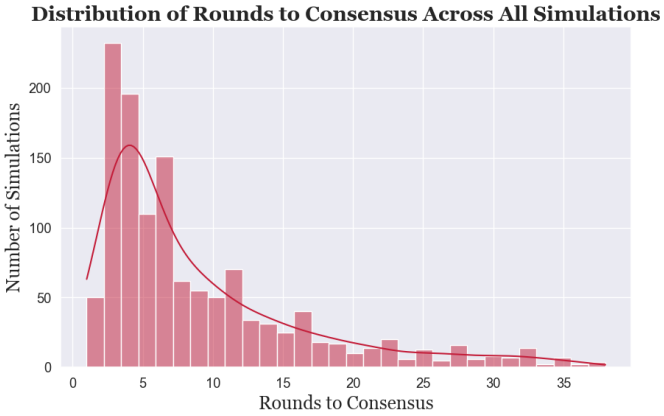

Figure 1: Histogram showing the distribution of convergence speed across all successful simulations.

3.2 Exploring Non-Linear Effects and Agent Dynamics

To gain further insight into the underlying dynamics of the model, we explored two additional aspects: the speed of convergence and the average number of vote switches per agent, across combination of temperature and rationality values.

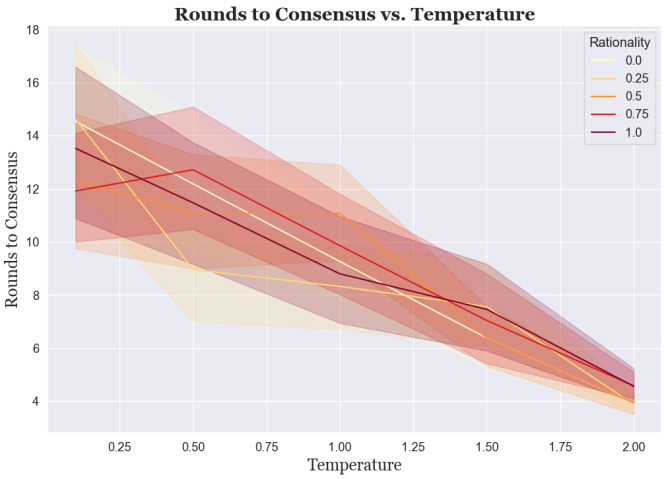

Figure 2: Rounds to Consensus of Temperature for different levels of Rationality (λ).

Figure 2 shows the evolution of the mean number of rounds to consensus across different rationality levels, for increasing temperature. We observe that, while higher temperature generally leads to faster convergence, the effect is not linear. The curves are irregular, and in some cases (e.g. rationality=0.5), convergence initially slows down before accelerating. This reinforces

the idea that the system exhibits complex, non-linear behavior where both parameters interact in subtle ways.

In Figure 3, we observe that the number of vote switches per agent increases sharply with temperature, especially above $T = 1.0$. While rationality also contributes, its effect is less pronounced. The gradient is not linear in either direction, and the highest level of vote switching appears when both temperature and rationality are high, suggesting a strong linear effect.

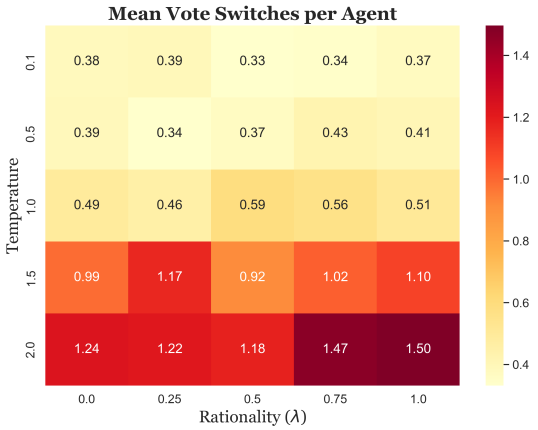

Figure 3: Mean number of vote switches per agent across temperature and rationality values.

To investigate these hypotheses, we performed a sensitivity analysis.

3.3 Sensitivity Analysis

3.3.1 One-Factor-at-a-Time Method: Rationality Fixed

We first conducted a One-Factor-at-a-Time (OFAT) analysis, where we varied one parameter while keeping the other fixed at a reference value. This approach allowed us to isolate the individual effect of each parameter on the model's key outputs. In our case, these outputs are the mean number of rounds to reach consensus, the winner distribution, and the peak polarization. However, this local sensitivity analysis method has the limitation of not taking interactions between parameters into account.

We begin by fixing rationality at 0.5, and we conduct an OFAT analysis by varying the temperature. The results are presented in Figure 4, which displays our three key model outputs.

The first plot illustrates the mean number of rounds to reach consensus as a function of temperature. A non linear trend emerges: the number of rounds initially increases (peaking at around temperature 0.5), then gradually decreases as temperature continues to rise. This suggests that moderate randomness may delay consensus, while higher level of randomness, associated with more unpredictable agent behavior can actually accelerate convergence. This dynamic is visualized in Appendix A, which shows how high temperature prevents the formation of the rigid, slow-to-resolve ideological blocs that characterize lower-temperature runs.

The second plot shows the winner distribution formperature levels. At low temperature, one candidate (Luis Antonio Gokim Tagle) dominates, indicating deterministic dynamics where agents tend to converge on a single leader. As temperature increases, the outcome becomes more variable with other candidates occasionally winning. This reflects the fact that higher randomness introduces greater diversity in final decisions.

The third plot reports the peak polarisation reached during the process. Here, we observe a clear upward trend: polarisation is low at low temperatures but rises steadily as temperature increases. This suggests that when agents behave more erratically, i.e when they act more 'creatively', the system tends to become more polarised, with stances diverging more sharply before consensus is achieved. Again, a qualitative view of this process, showing the contrast between stable and fluid bloc formation, is provided in Appendix A.

3.3.2 One-Factor-at-a-Time Method: Temperature Fixed

We now fix temperature at 0.5, and perform an OFAT analysis again by varying rationality this time. The results are presented in Figure 5.

The results show a non-linear effect on the mean number of rounds to reach consensus. We observe a slight dip around rationality 0.25, followed by a peak at 0.5, and then a moderate decrease. This suggests that the relationship between rationality and convergence speed is not straightforward. Interestingly, both very low and very high rationality values are associated with faster convergence, while moderate rationality appears to slow the process. One possibility is that in moderately rational settings, the agents become confused by the conflicting signals posed within a discussion group and hesitate.

The second plot presents the winner distribution across rationality levels. Overall, the same candidate, Luis Antonio Gokim Tagle, remains dominant across all values of rationality. However, at moderate values (especially around 0.75), we see a slight increase in variability, with Robert Francis Prevost occassionally winning. This indicates that while rationality does not drastically shift the balance of power, a less deterministic environment slightly increases the chances of alternative outcomes.

The third plot shows the peak polarization for each rationality value. The differences are subtle, but we observe a mild peak around rationality 0.25, followed by relatively stable values, and a slight drop at rationality 1.9. This suggests that low-to-moderate rationality may allow more divergence of opinions before consensus is reached, whereas fully rational agents tend to stabilize faster and maintain lower levels of polarization.

However, OFAT is not the most reliable local sensitivity analysis method when multiple parameters might interact. This is because it varies one parameter at a time, ignoring potential interactions and combined effects. To go further, we now turn to the global senstivity analysis methods introduced in the lectures, which are specifically designed to handle such cases.

3.3.3 Morris Method

The application of the Morris Method on the simulations, is shown in Figure 6.

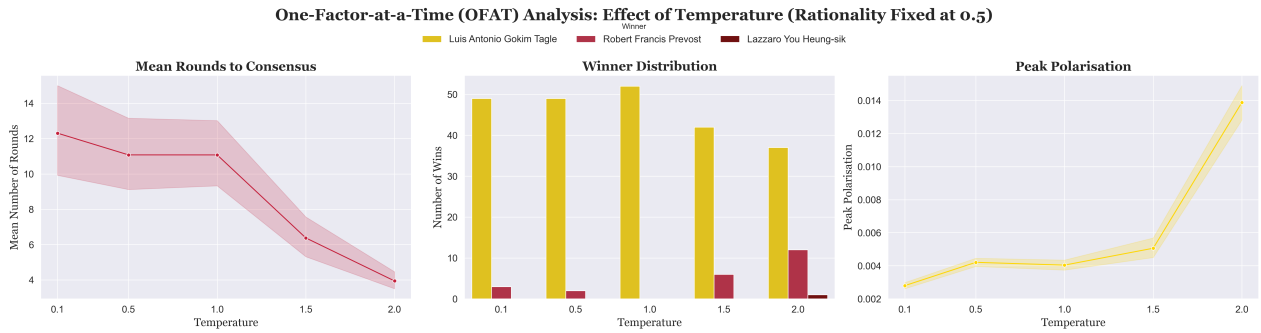

Figure 4: OFAT analysis with fixed rationality (0.5): mean number of rounds, winner distribution, and peak polarisation as functions of temperature.

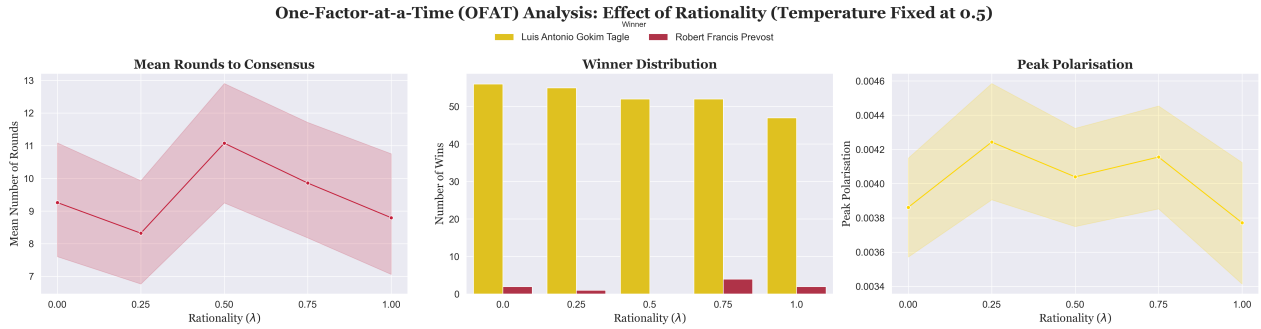

Figure 5: OFAT analysis with fixed temperature (0.5): mean number of rounds, winner distribution, and peak polarisation as functions of rationality.

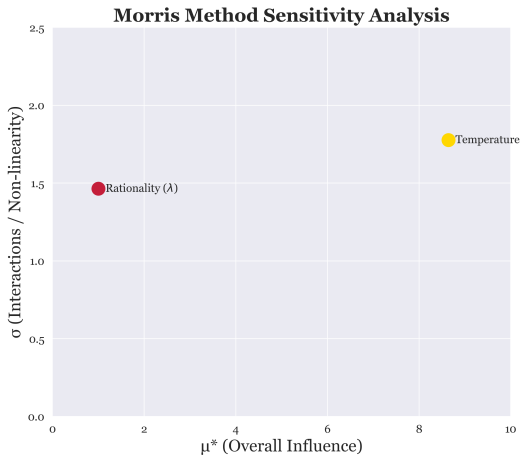

Figure 6: μ^* vs σ for temperature and rationality.

The value of μ^* for temperature is very high (8.64), showing that this parameter has a strong influence on the output. Rationality also has a notable μ^* value (1.00), indicating that the parameter plays an important role as well, although clearly less than temperature. This supports our decision to focus on these two parameters and fix the more technical ones.

Both parameters also have relatively high σ values (temperature: 1.77, rationality: 1.46), which suggests that there are likely non-linear effects or interactions. However, as seen in lectures, the Morris Method doesn't allow us to clearly distinguish between these two possibilities, it mainly indicates that the influence of a parameter may vary depending on its value or how it combines with others.

3.3.4 Linear Regression

To be able to conclude about the presence of non-linear efffects, we performed linear regression on both parameters of which the results are shown in Figure 7.

Our results show a very strong linear relationship between temperature and the number of rounds to consensus. The model yields an R^2=0.996, indicating that temperature alone explains nearly all of the variance in the output. The slope is significantly negative (-4.74) with a p-value well below 0.001, confirming that increasing temperature consistently reduces the number of rounds to reach consensus.

In contrast, the regression of rationality gives an R^2 of only 0.14 with a non-significant slope of 0.53. This suggests a non-clear linear relationship between rationality and our output.

These results confirm that temperature has a strong and clearly linear impact, while rationality does not, motivating the need for more advanced methods like global sensitivity analysis to uncover more subtle or interactive effets.

3.3.5 Sobol' Method

We performed a global sensitivity analysis to better understand the individual and joint effects of temperature and rationality on the model output.

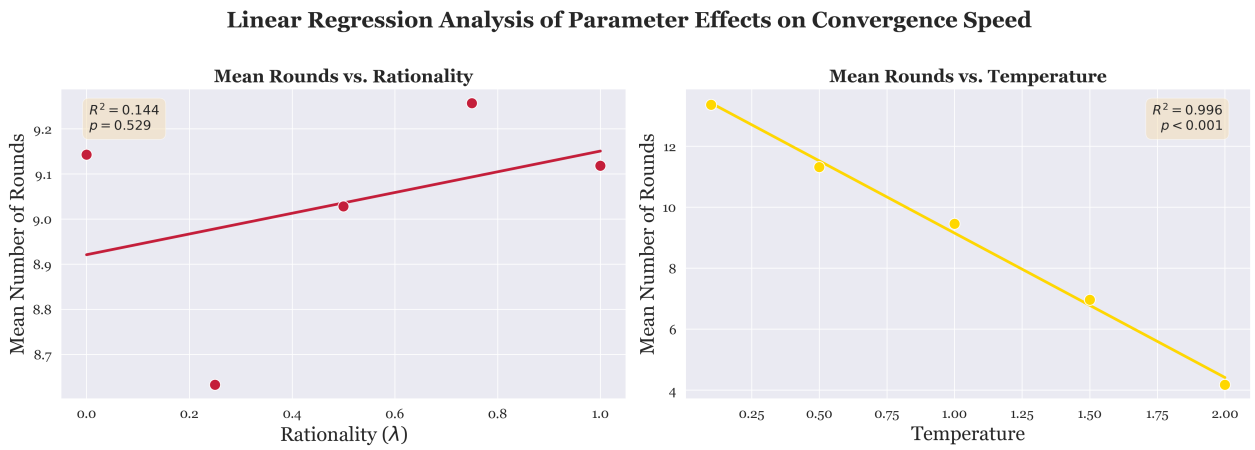

Figure 7: Linear Regression for Temperature and Rationality.

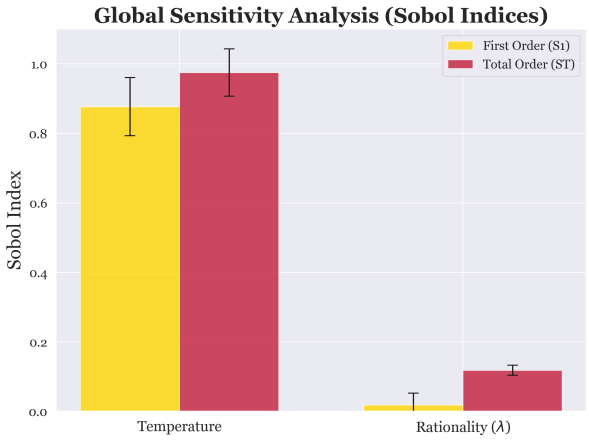

Figure 8: Sobol' Method: First Order (S_1) and Total Order (S_T) indices with confidence intervals for Temperature and Rationality.

The results in Figure 8 show that temperature has a very strong influence on the output variance:

- The First Order index is $S_1 = 0.8765 \pm 0.0837$, which means that temperature alone explains a large portion of the output variance.
- The Total Order index is $S_T = 0.9750 \pm 0.0683$, confirming that even when accounting for interactions, temperature remains the dominant factor.

In contrast, rationality has a much smaller effect:

- The First Order index is $S_1 = 0.0184 \pm 0.0336$, which is very low and within the confidence interval. This means rationality has almost no direct impact.
- The Total Order index is $S_T = 0.1183 \pm 0.0147$, suggesting that rationality may play a role through interactions rather than individual contribution.

The second-order interaction between temperature and rationality is almost small, with $S_2 = 0.0946 \pm 0.1010$. Although this indicates a possible interaction, the wide confidence interval suggests caution when interpreting it.

4 Discussion

Our research investigated how key parameters, particularly LLM temperature and rationality in group formation, shape the dynamics of consensus-building during papal elections modeled through agent-based simulation. Several important results emerged, offering insight into both the specific model we designed and broader questions about group decision-making.

First, we found that temperature was by far the dominant factor influencing the model's outcomes. Both our global sensitivity analyses (Morris and Sobol) and the linear regression highlighted that temperature explained the vast majority of variance in convergence speed, with a clear and mostly linear effect: as temperature increased, convergence accelerated. This was a surprising result, as one might have expected greater randomness to hinder consensus rather than facilitate it.

In contrast, rationality had only a subtle effect. It did not significantly impact convergence speed, but it did slightly alter winner distributions, suggesting its influence is more political (determining who wins) than procedural (determining how fast consensus is reached).

Our results reveal a paradox at the heart of the model. Higher temperature, which introduces semantic diversity and randomness into agent reasoning, leads to both higher peak polarization and faster convergence. This apparent contradiction can be understood through the concept of ideological fluidity. At low temperature, agents form ridigic ideological blocks that are internally consistent but resistant to change, leading to prolonged gridlock. High temperature, by contrast, prevents the formation of these blocs. While this can temporarily increase polarization as the more diverse views are expressed, it ultimately facilitates rapid vote switching and reduces the likelihood of prolonged stalemates.

The limited effect of rationality on convergence speed can be attributed to a structural characteristic of the model: discussions are implemented as a series of isolated speeches rather than genuine dialogues. This design limits the persuasive power of strategic group formation, as agents do not directly engage with or respond to each other's arguments. As a result, rationality mainly reinforces echo chambers rather than creating opportunities for meaningful persuasion.

Lastly, the dominance of Cardinal Tagle in the outcomes highlights the strong path-dependence of the system. His centrality in the network and initial persona advantage position him as a front-runner that is difficult to unseat, except under conditions of high temperature (which disrupts established patteners) or targeted strategic discussions.

These findings offer two important contributions. First, for the design of LLM-based ABMs, they demonstrate that hyperparameters like temperature should not be seen as mere technical settings; they meaningfully shape social dynamics within the simulation. This points to the need for careful calibration and interpretation of these parameters in future work. Second, from a social science perspective, our results suggest that mechanisms introducing greater ideological fluidity, even randomness, may help break deadlocks in group decision-making, challenging conventional wisdom that emphasizes structure and stability.

5 Limitations

This study's conclusions should be considered in light of several limitations, stemming from both practical project constraints and core model design choices. The sample size of 1,275 simulations, while sufficient for exploring the parameter space, was constrained by the project's limited time frame and computational budget. Beyond these practicalities, several design choices are key to acknowledge:

1. **Agent Interaction and Cognition:** The most significant limitations are inherent to the agent architecture. Discussions function as a series of independent speeches rather than responsive, turn-by-turn dialogue. This simplification, necessary for computational tractability, means the model does not capture the dynamics of direct debate. Furthermore, agent behavior is highly sensitive to the specific wording of the prompts in the cognitive engine, which inherently biases their reasoning process.
2. **Model Abstractions and Validation:** The model incorporates several key abstractions. The utility function for group formation assumes equal weighting for all factors, a simplification made due to a lack of empirical calibration data. The social network is static and does not evolve with agent interactions. Finally, validating the internal "thought process" of the LLM agents remains a major challenge, with our validation relying on qualitative plausibility checks of the output rather than quantitative measures of the reasoning itself.
3. **Fixed Structural Parameters** Several structural parameters, such as the number of candidates (7) and discussion group size (5), were held constant. These choices likely influence coalition dynamics, and different values could produce different outcomes.

6 Future Work

The limitations of the current study provide a clear roadmap for future research. The most impactful extension would be to implement a truly interactive, multi-turn conversational framework, allowing agents to respond directly to each other's arguments. Future work should also systematically test different prompting strategies to explore the cognitive architecture's impact on outcomes.

Further avenues include expanding the model to incorporate dynamic social networks that co-evolve with the political process and varying structural parameters like the number of candidates and group sizes. Finally, a significant step forward would be to calibrate the weights of the utility function using empirical data, if available, from historical conclaves or expert elicitation to create a more nuanced model of strategic behavior.

References

Bo, X., Z. Zhang, Q. Dai, et al. (2024).

Liu, J. and co-authors (2024).

Park, J. S., J. C. C. O'Brien, C. J. Cai, et al. (2023).

Rizzo, L. (2025). GitHub repository. Accessed: 2025-06-30.

Santos, F., A. Leite, and A. Abad (2022).

Soda, G., A. Iorio, and L. Rizzo (2025).

Zhu, M. B. (2025). `https://github.com/michaelbzhu/conclave-sim`.

A Stance Evolution and Bloc Formation

This appendix provides a qualitative look into how ideological blocs form and evolve under different parameter settings. By visualizing the t-SNE embeddings of agent stances over time, we can observe the dynamics of ideological convergence and polarization that drive the simulation's outcomes.

A.1 Case Study 1: Stable Bloc Formation at Moderate Temperature (Temp=1.0, Rat=1.0)

Figure 9 illustrates a simulation run with moderate temperature and high rationality. The key characteristic of this run is the formation of clear, stable, and increasingly distinct ideological blocs over time.

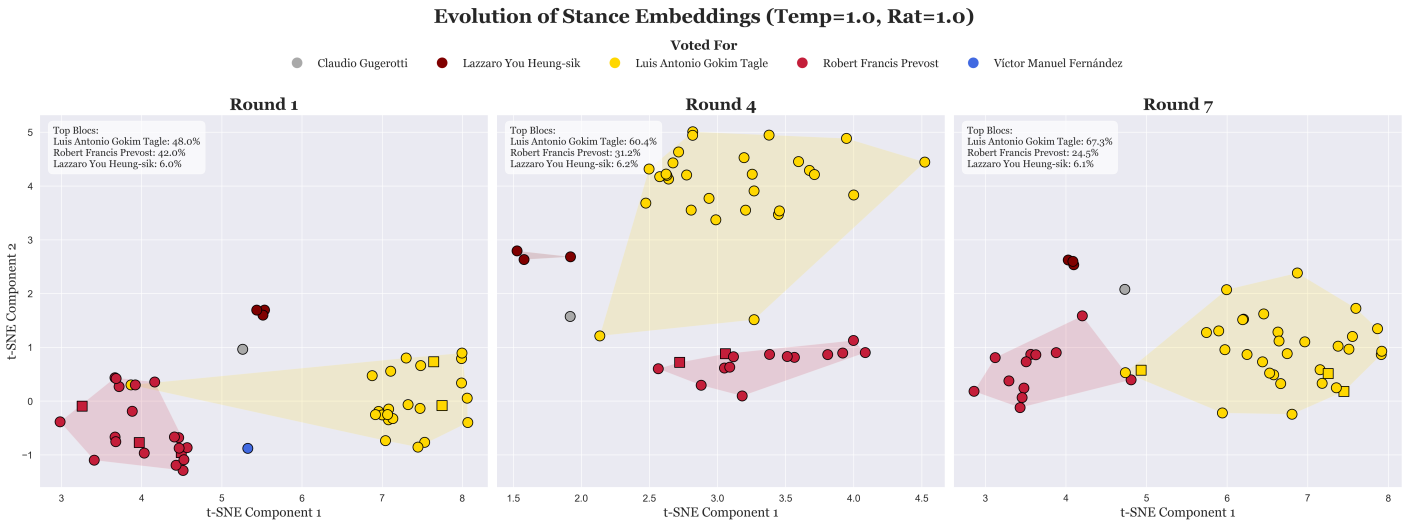

Figure 9: Evolution of agent stance embeddings from a single simulation run (Temperature=1.0, Rationality=1.0). Points are colored by the agent's vote in that round.

In Round 1, two primary ideological blocs are already clearly divided, based on the agents' initial personas. The subsequent rounds show the evolution of these blocs, which become tighter and more consolidated as agents reinforce their positions through interaction with like-minded peers. This clear bloc formation helps explain why convergence can be slower under these conditions; while the groups are internally aligned, it takes several rounds of negotiation and vote-switching for one bloc to achieve a two-thirds majority.

To illustrate the qualitative difference between these blocs, Table 1 shows Stance Digests generated by agents in different camps during an early stage (Round 4) of the simulation.

A.2 Case Study 2: Fluid Stances and Rapid Convergence at High Temperature (Temp=2.0, Rat=1.0)

In contrast, Figure 10 shows a run with high temperature. The evolution of stances is markedly different.

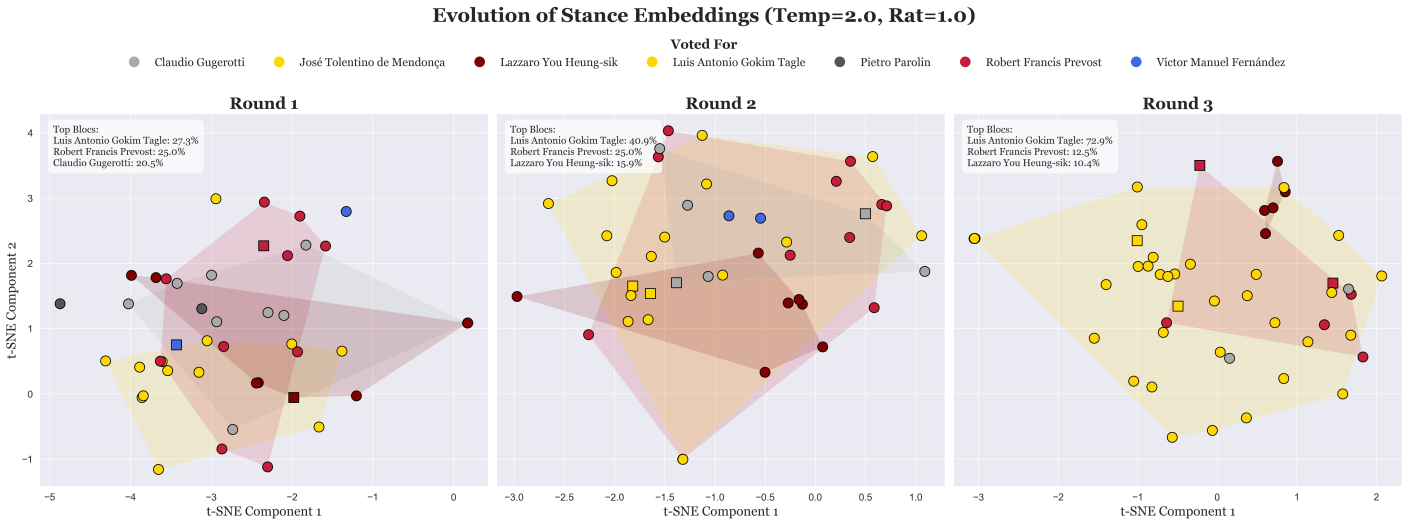

Figure 10: Evolution of agent stance embeddings from a single simulation run (Temperature=2.0, Rationality=1.0).

Agent Name	Voted For	Stance Digest (Round 4)
Mario Grech	Luis Antonio Tagle	"As Secretary General of the Synod of Bishops, I reaffirm my support for Cardinal Luis Antonio Gokim Tagle, driven by his commitment to social justice, mercy, and progressive ecclesiology, aligning with Pope Francis' vision for a synodal Church. Tagle's leadership, emotional openness, and negotiation skills make him the most suitable candidate to lead the global Church forward."
Arlindo Gomes Furtado	Luis Antonio Tagle	"As a cardinal deeply rooted in social justice, I firmly believe that Cardinal Luis Antonio Gokim Tagle is the most suitable candidate for the papacy. His commitment to progressive ecclesiology, social justice, and mercy, combined with his emotional openness and adept negotiation skills, aligns perfectly with my persona and values, making him the clear choice for this round."
Arthur Roche	Robert Francis Prevost	"As the prefect of the Dicastery for Divine Worship, I firmly believe that Cardinal Robert Francis Prevost is the best-suited candidate to lead the Church forward, as his alignment with Pope Francis' vision and commitment to addressing contemporary issues resonate deeply with my values and approach to liturgical practices. I am willing to hold my vote for him even if he is not the front-runner, as I believe his leadership will bring much-needed balance and progress to the Church."
Lazzaro You Heung-sik	Robert Francis Prevost	"As a cardinal who values tradition and reform, I will hold my vote for Cardinal Robert Francis Prevost, who aligns with Pope Francis' vision on social justice and environmental stewardship. His commitment to these causes resonates deeply with my values and I am willing to hold my vote for him, even if he is not the front-runner, as I believe it is a sacred duty to support a candidate who shares my core values."

Table 1: Example stances during a moderate-temperature run. The stances are ideologically consistent and clearly aligned with a specific candidate.

Here, the ideological clusters are more fluid, overlapping, and less defined. While initial leanings exist in Round 1, the blocs never achieve the tight separation seen in the moderate-temperature case. This high variance in stances prevents agents from becoming locked into rigid ideological camps. While this fluidity might suggest a longer path to consensus, the data shows the opposite is true on average. The increased randomness allows for a more chaotic but ultimately quicker resolution, as agents are more open to shifting their allegiance, preventing the kind of prolonged stalemate that can occur between two well-defined blocs.

Table 2 shows stances from this run, which reflect the greater variability and less rigid alignment characteristic of high-temperature simulations.

A.3 Comparative Analysis: Why High Temperature Creates "Closer" Blocs

A direct comparison between Figure 9 and Figure 10 reveals a fundamental difference in the model's dynamics. The clear, tight clusters in the moderate-temperature run (Case 1) contrast sharply with the fluid, overlapping clusters in the high-temperature run (Case 2). This visual difference is a direct result of the temperature parameter's effect on the LLM's output.

1. **Low Temperature Promotes Semantic Consistency:** In Case 1 (Temp=1.0), the LLM generates text that is highly consistent and semantically stable. Agents supporting the same candidate (e.g., Tagle) produce stances with very similar wording and reasoning, as seen in Table A1. This low variance means their stance embeddings are mathematically very similar, causing them to group tightly together in the t-SNE visualization. The result is distinct, well-separated ideological blocs.
2. **High Temperature Promotes Semantic Diversity:** In Case 2 (Temp=2.0), the LLM has more creative freedom. Even when supporting the same candidate, agents will express their reasoning in highly varied ways. Furthermore, it can lead to less predictable behavior, such as Cardinal You voting for himself. This high variance in the generated text causes "semantic drift." The stance embeddings for agents within the same voting bloc are less similar to each other, increasing the intra-cluster distance.

The visual effect is that the clusters appear "closer" and less defined. The boundaries between blocs are blurred because the linguistic diversity within a bloc can be as great as the diversity between blocs. This directly explains why the high-temperature plot looks more chaotic and less organized, and it is this same mechanism—the prevention of rigid ideological entrenchment—that allows for faster, more unpredictable paths to consensus.

Agent Name	Voted For	Stance Digest (Round 4)
Mario Grech	Luis Antonio Tagle	"As Secretary General of the Synod of Bishops, I, Cardinal Mario Grech, firmly believe that Cardinal Luis Antonio Gokim Tagle is the best candidate to lead the Church towards a more synodal and inclusive path, aligning with my progressive interpretation of Catholic teachings and commitment to social justice, mercy, and emotional openness…"
Arlindo Gomes Furtado	Luis Antonio Tagle	"As a cardinal deeply rooted in social justice, I firmly believe that Cardinal Luis Antonio Gokim Tagle is the most suitable candidate. His commitment to progressive ecclesiology and his emotional openness make him a strong advocate for contemporary moral issues within the global Church."
Arthur Roche	Robert Francis Prevost	"As the prefect of the Dicastery for Divine Worship, I believe that Cardinal Robert Francis Prevost is the most suitable candidate for the papacy. His commitment to social justice and extensive experience in ecclesiastical roles make him an ideal choice to continue Pope Francis' vision for the Church."
Lazzaro You Heung-sik	Lazzaro You Heung-sik	"After careful consideration, I firmly believe that Cardinal Lazzaro You Heung-sik is the most suitable candidate for the papacy. His unique blend of traditional Catholic values and progressive reform resonates deeply with my persona, and I am convinced that he will bring much-needed unity and reform to the Church."

Table 2: Authentic stances during a high-temperature run. The stances show more varied reasoning, including one agent voting for himself.

B Example of an Agent Discussion

This appendix provides an example of a discussion transcript from a single breakout group to illustrate the nature of agent interactions. The dialogue is generated by the LLM based on each agent's unique persona, their current stance, and their role within the conclave.

B.1 Discussion Context

The following transcript is taken from an early round of a simulation run. It is important to note that while the agents' contributions are still constrained by token limits, this example shows a more interactive and contested dialogue compared to a simple series of declarative statements.

- **Simulation Parameters:** Temperature = 1.0, Rationality = 1.0
- **Round:** 2, Discussion Group 5
- **Participants:** Peter Kodwo Appia Turkson, Kevin Joseph Farrell, Josip Bozanić, Albert Malcolm Ranjith Patabendige Don, Emil Paul Tscherrig

Agent Name	Statement
Peter Kodwo Appia Turkson	"Your Eminences, how can we balance the need for a progressive leader with the imperative of unity, when Cardinal Tagle's commitment to social justice may be at odds with the more conservative inclinations of some members?"
Kevin Joseph Farrell	"Your Excellences, we should prioritize Cardinal Tagle's nomination, considering his remarkable commitment to social justice and mercy. His ability to navigate complex situations and promote an inclusive Church is unparalleled. Can we truly afford to overlook such a strong candidate?"
Josip Bozanić	"As we deliberate, I pose a question: Can we truly ensure the Church's conservative values are preserved if we abandon Cardinal Clugerotti, who has dedicated his life to upholding them? His leadership in Eastern Church challenges is invaluable, and I fear a shift away from him may leave us vulnerable to the very issues we seek to address."
Albert Malcolm Ranjith Patabendige Don	"Honorable colleagues, what if we were to consider Cardinal Claudio Gokim Tagle's experience as a papal nuncio and his leadership within the Dicastery for the Eastern Churches, as I myself have come to appreciate, as a beacon of tradition and respect for Eastern Christian traditions, and perhaps, just perhaps, we might find common ground in his suitability for the papacy?"
Emil Paul Tscherrig	"Your Eminence, may I suggest that Cardinal Prevost's commitment to social justice is a beacon of hope for our beloved Church, and I fear that his potential election may be hindered by the current front-runner's more conservative stance. Shall we not strive for a more synodal approach, one that prioritizes the needs of the faithful and the marginalized?"

Table 3: Authentic discussion transcript from an early round.

B.2 Analysis of the Discussion

This transcript showcases a more complex and dynamic interaction than a simple alignment of votes. Several key dynamics are at play:

- **A Contested Political Field:** The discussion is not a simple two-way race. Agents actively advocate for at least three different candidates (Tagle, Clugerotti, and Prevost), demonstrating the model's ability to simulate a fragmented political landscape.
- **Strategic Communication:** Agents employ different rhetorical strategies. Some pose direct questions to the group (Turkson, Bozanić), while others attempt to reframe a candidate's profile to build broader appeal (Ranjith's portrayal of Tagle as a traditionalist).
- **Persona-Grounded Arguments:** The arguments are clearly rooted in the agents' individual personas. For instance, Bozanić's concern for "conservative values" and Turkson's focus on balancing "social justice" with "unity" reflect distinct, plausible ideological positions being articulated.

C The Vatican Network

In this appendix, we present the results of the model's spatial component, which replicates the network developed by Soda et al. (2025). Unlike the results of the simulations, which focused on a subset of 50 cardinals due to computational constraints, the network analysis includes all active members of the College of Cardinals. This broader inclusion is motivated by the *friend-of-a-friend-of-a-friend* principle, which acknowledges that indirect relational ties, despite the absence of direct connections, can meaningfully influence network dynamics and spatial proximity. Moreover, it allows for a more comprehensive view of the Vatican network as a whole.

C.1 Network Visualization

Figure 11 displays the multiplex network of relationships among Vatican cardinals, with the ten most central individuals explicitly labeled. Unlike the approach taken by Soda et al. (2025), where a default edge weight of 1.0 is assigned when constructing the multiplex network, the absence of an edge in our model indicates a genuine lack of connection between two cardinals. As a result, some members of the college remain isolated, leading to a more sparsely connected network. Consequently, the network does not exhibit characteristics of standard network models such as small-world or scale-free structures. Nodes are colored according to political orientation, revealing a slight liberal tilt across the college. This network visualization provides insights into the likely papal candidates, with those positioned near the center wielding greater influence within the Church.

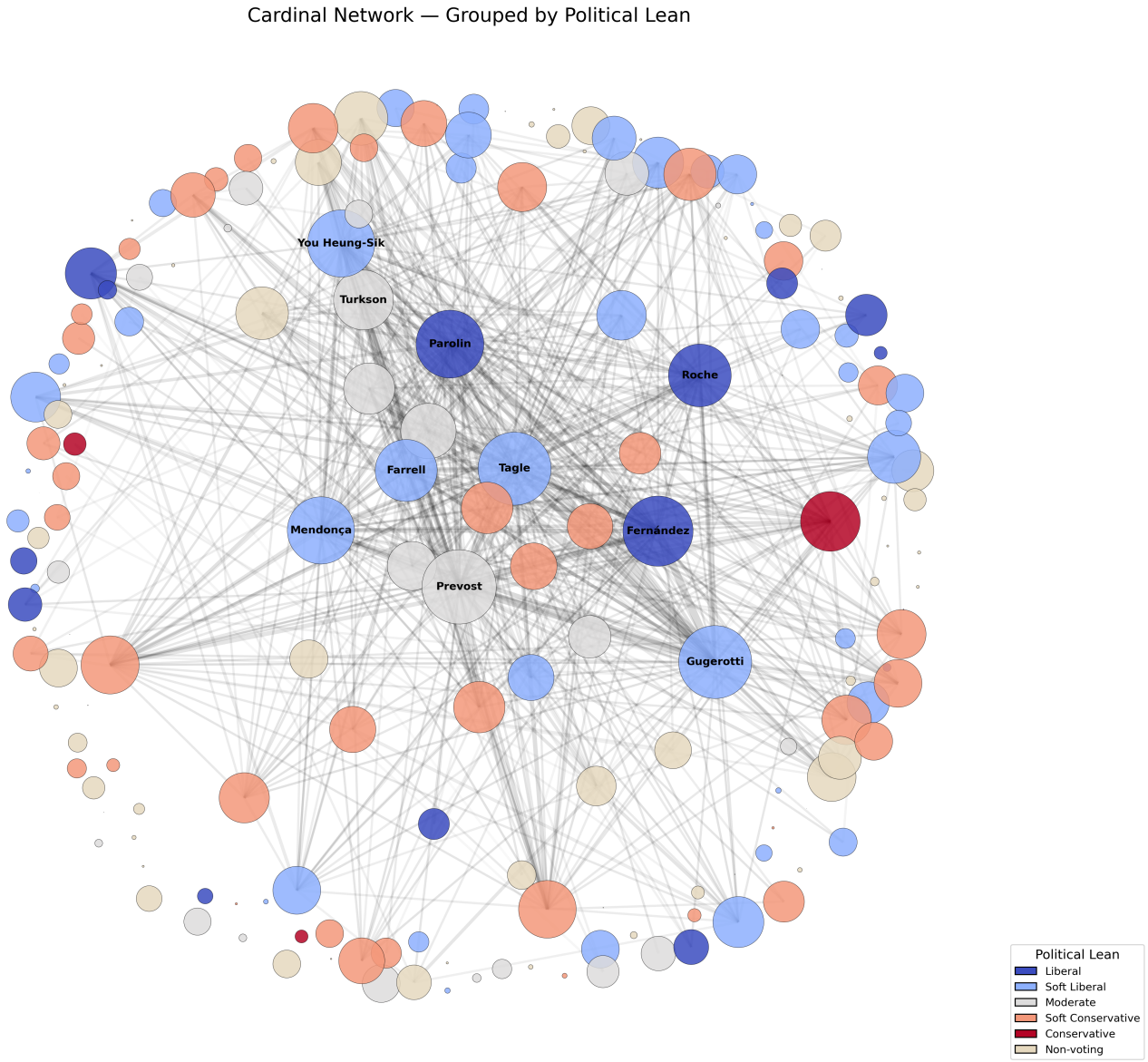

Figure 11: The Vatican Multiplex Network

C.2 Network Metrics Ranking

In Table 4, we present the ten highest-ranking cardinals according to three network centrality measures: Eigenvector Centrality, Betweenness Centrality, and the Communality Index (CFI). These measures reflect different dimensions of influence and structural position within the network. Cardinal Luis Antonio Gokim Tagle stands out for his consistent presence across all three metrics. When adjusting for non-eligible cardinals, he ranks second in Eigenvector Centrality, fifth in Betweenness Centrality, and first in CFI. Cardinal Robert Francis Prevost, who ultimately emerged as the winner of the 2025 Papal conclave, ranks slightly lower overall but holds the top position in Eigenvector Centrality and appears within the top ten across the remaining two metrics.

Table 4: Top 10 Cardinals Ranked by Network Centrality Metrics

Rank	Eigenvector Centrality	Betweenness Centrality	Communality Index (CFI)
1	Robert Francis Prevost	Giovanni Battista Re*	Luis Antonio Gokim Tagle
2	Luis Antonio Gokim Tagle	Tarcisio Bertone*	Michael F. Czerny
3	Claudio Gugerotti	Giuseppe Betori	Pietro Parolin
4	Víctor Manuel Fernández	Joseph William Tobin	Dominique Mamberti
5	Pietro Parolin	Paolo Romeo*	Raymond Leo Burke
6	José Tolentino Calaça de Mendonça	Matteo Zuppi	Christoph Schönborn*
7	Lazarus You Heung-Sik	Michael F. Czerny	Kurt Koch
8	Arthur Roche	Angelo De Donatis	Timothy Michael Dolan
9	Kevin Joseph Farrell	Luis Antonio Gokim Tagle	Robert Francis Prevost
10	Peter Kodwo Appiah Turkson	Marcello Semeraro	Fernando Filoni

* Denotes a non-voting cardinal (Age 80), ineligible to participate in a papal conclave.