LLM-Based Simulation of Multilateral Climate Negotiations: A Multi-Agent Framework for Computational Climate Diplomacy

Ziqi Wei (ziqiwei@mit.edu)[1], Anna Papp[2]
[1]Department of Civil and Environmental Engineering, MIT, [2]Department of Economics, MIT

Massachusetts Institute of Technology

MIT Schwarzman College of Computing

GitHub Repo
github.com/alcohol-freeWhisky/climate_negotiation_simulation

Background & Motivation

Problem

Multilateral climate negotiations under the UNFCCC involve many country coalitions bargaining over specific options, but:

· Game-theoretic models ignore textual and procedural details;
· Qualitative case studies cannot be replicated, or scaled;
· No computational platform for counterfactual experiments.

Why LLM Agents

· Maintain stable, differentiated identities with system prompts;
· Respond to bridging text produced by a neutral Chair agent;
· Sustain negotiations without forgetting earlier rounds.

Research Questions

· Can LLM **agents sustain coherent, differentiated positions** reflecting real-world coalition dynamics across extended negotiations?
· Can the Chair agent **synthesize treaty-quality text** comparable to actual COP decisions?
· Does **agent ablation (removing one coalition) produce politically interpretable effects** aligning with expectations?

Model Architecture

Architecture

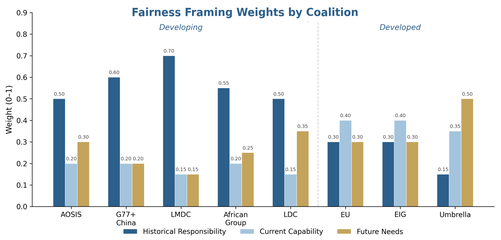
- **Scenario Configuration** — prompts, goals, constraints
- **Negotiation Engine** — state, analysis, counteroffers
- **Agents & Chair** — 8 agents + 1 chair
- **LLM Backend & Evaluator** — inference, scoring, analytics

Agent Settings

- Identity Schema
- Hard Facts
- Normative Frame
- Stance
- Behavioral Parameters
- Interaction Style
- LLM Overrides

Weights & Influence

- Historical Responsibility
- Historical Context
- Projected Future Needs
- Current Capability
- Capability Assessment
- Projected Future Needs
- Fairness Weights & Influence
- Agent Arguments



· Simulate UNFCCC Article 6.8 negotiations with 8 LLM-driven blocs, producing treaty text near-identical to real COP outcomes.
· Reproduce high-fidelity treaty language that fails politically with real deadlocks.
· Reveal through ablation that text quality and consensus are structurally opposed: vulnerable blocs enrich the treaty while G77's coalition pressure drives the collective veto.
· Extend to multi-article negotiations, add coalition side-channels, and test other LLM agents.

Experiment Workflow

Context & Key Issues

- UNFCCC COP30
- Paris Agreement Article 6.8
- North–South divide
- Climate finance
- Technology transfer
- Historical responsibility

Negotiating Blocs

Global South / Developing Blocs

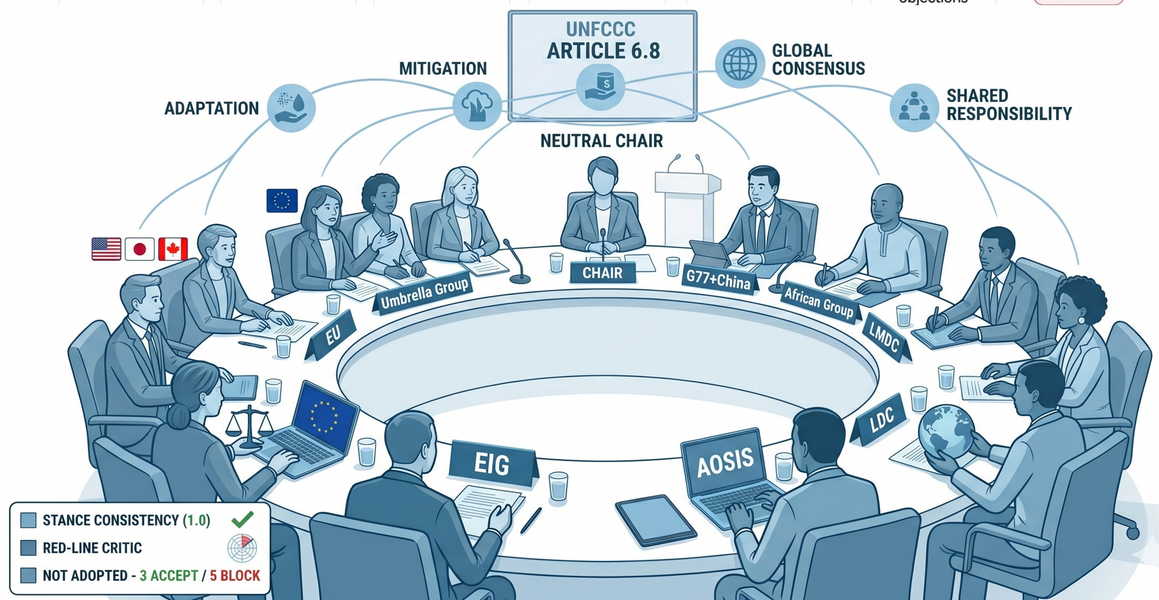
- **G77 + China** — Group of 77 and China
- **LMDC** — Like-Minded Developing Countries
- **African Group**
- **AOSIS** — Alliance of Small Island States
- **LDC** — Least Developed Countries

Global North / Mixed Blocs

- **EU** — European Union
- **Umbrella Group**
- **EIG** — Environmental Integrity Group

1. **Initialization** — Load draft text; initialize agents
2. **Opening Statements** — Agents present baseline positions
3. **Coalition Caucus** — Align within blocs; memory injection
4. **First Reading** — Review text; propose edits
5. **Informal Consultations** — Chair-led negotiation loop
6. **Final Plenary** — Red-line review; adopt or log objections

→ **Adopted** / **Not Adopted**

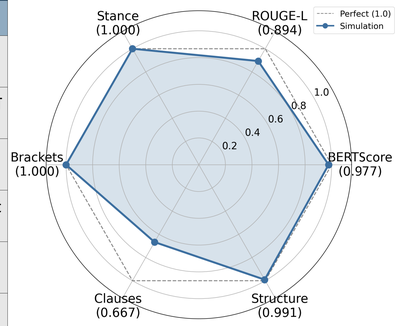


Results

Baseline Profile

Outcome Only **3 developed-world-aligned blocs accepted** the final text while **5 developing-country blocs blocked**, due to the absence of CBDR language and inadequate means-of-implementation provisions, reproducing **the North-South veto dynamic** in real negotiations.

Metric Definitions	
Metric	**What it Measures**
ROUGE-1/2/L	Lexical n-gram overlap with reference
BERTScore	Semantic similarity via contextual BERT embeddings
Key Clause Coverage	Essential policy elements present in agreed text
Structural Similarity	Paragraph/item/word-count alignment with reference
Bracket Resolution	Fraction of paragraphs fully de-bracketed
Stance Consistency	Red-line maintenance across agent statements

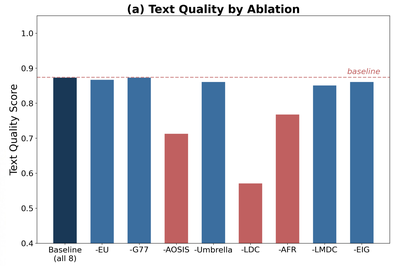


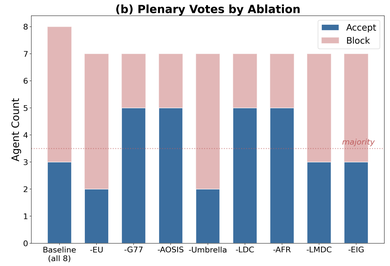
Evaluation The **simulation achieves near-perfect scores on textual dimensions but falls short on Key Clauses** (shown in above metrics), revealing that LLMs replicate the form and style of treaty language far more easily than the substantive policy content that defines real agreements.

Ablation Findings

Text quality vs. consensus

· In Fig (a) and (b), removing LDC or AOSIS improves acceptance but collapses text quality, which means the actors who enrich the treaty are not the actors who block it.
· G77+China is the only bloc whose removal achieves majority acceptance without degrading text quality; developed-country ablations have negligible impact on either metric.





Structural voting dynamics

· In Fig (c), EU and EIG always accept; G77 and LMDC always block; removing G77 flips other developing blocs to accept, exposing **coalition pressure as the driver of collective opposition**.
· In Fig (d), **developing-country ablations achieve majority at the cost of substantive depth**, which means the baseline produces quality text that fails politically, mirroring real equity deadlocks.