ReGMS: Retrieval-Grounded Multi-Agent Scenario Analysis for Climate Risk

Yun Wing Kiang
Department of Electrical and Computer Engineering
The University of Hong Kong
Hong Kong, Hong Kong
kiangyw@eee.hku.hk

King Hang Lam
Department of Electrical and Computer Engineering
The University of Hong Kong
Hong Kong, Hong Kong
khlam@eee.hku.hk

ABSTRACT

We study climate–related scenario analysis through the lens of multi–agent systems. Our goal is to support IFRS S2/TCFD reporting with scenarios that are grounded in public pathways and transparent in their assumptions. We present ReGMS, a retrieval–grounded agent architecture where specialized LLM agents (retrieval, scenario design, quantitative coupling, compliance) coordinate to build and verify transition and physical–risk scenarios. We cast the task as a constrained stochastic game in which retrieved evidence sets feasible actions and verification checks. We discuss convergence behavior of evidence–constrained best–response updates and empirically compare against a centralized planner baseline. For evaluation, we use the Network for Greening the Financial System (NGFS) Phase V transition pathways and NASA NEX–GDDP–CMIP6 downscaled daily projections (from which we derive simple heat and rainfall indices). We report diagnostics for internal consistency against NGFS reference trajectories, scenario diversity across canonical regulatory cases (Current Policies vs. Net Zero 2050), and standards alignment with citation provenance. Baselines include an IAM–style planner, independent agents, and a single–LLM pipeline. These experiments illustrate that ReGMS can produce traceable, standards–aligned scenarios under explicit constraints.

KEYWORDS

Multi–agent framework; Climate Scenario Analysis; Constrained Stochastic Games; Retrieval–Grounded Reasoning; ESG; TCFD/IFRS S2; GAAI

ACM Reference Format:
Yun Wing Kiang and King Hang Lam. 2026. ReGMS: Retrieval-Grounded Multi-Agent Scenario Analysis for Climate Risk. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/RAVT8220

1 INTRODUCTION

Climate scenario analysis has become a mandated practice for financial disclosures [7, 17], requiring organizations to assess their resilience under different climate futures. To be decision–useful, scenarios must be internally coherent (e.g., consistent assumptions across transition and physical impacts), grounded in credible data, and aligned with standards like TCFD and IFRS S2. IFRS S2 (ISSB climate disclosures) and the TCFD recommendations require scenario analysis spanning short–, medium–, and long–term horizons and covering both transition and physical risks. Disclosures should describe scenario sources and key assumptions so they can be audited. We operationalize these requirements by anchoring transition variables to the Network for Greening the Financial System (NGFS) pathways and producing citation provenance for key assumptions. Accordingly, we target four objectives: (i) transition fidelity to external pathways, (ii) hazard–adaptation alignment for physical risk, (iii) scenario–consistent mitigation dynamics, and (iv) standards–aligned compliance and traceability with auditable citations. Achieving these qualities is challenging: it involves integrating heterogeneous information (economic pathways, climate model outputs) and applying expert judgment to ensure plausibility and compliance. Large Language Models (LLMs) offer a promising tool for synthesizing quantitative and narrative information, but a naïve single–LLM approach can produce plausible–sounding yet factually ungrounded or inconsistent scenarios. LLMs are prone to hallucination and may violate hard constraints (e.g., deviating from reference data) if not carefully controlled [6, 8, 12, 14]. Recent advances suggest two key strategies to mitigate these issues: (1) *retrieval augmentation*, where external knowledge is fetched to improve factual accuracy and transparency [4, 9, 15], and (2) *multi–agent decomposition*, where a complex task is split among specialized agents that can reason and critique each other's outputs [10, 20, 24].

Inspired by these insights, we propose **ReGMS**, a retrieval–grounded multi–agent system for climate scenario analysis. ReGMS employs a team of cooperating LLM–based agents to construct climate risk scenarios that meet all the above requirements. The agents collaborate on distinct sub–tasks of scenario generation, such as selecting an appropriate reference pathway, quantifying transition trajectories, incorporating physical climate hazards, and verifying compliance with standards. By design, each agent's actions are constrained by retrieved evidence from authoritative sources (e.g., NGFS scenario data, climate model documentation, disclosure guidelines). This ensures that the scenario is built on a foundation of facts and external rules rather than the LLMs' internal guesswork. To our knowledge, our approach is the first to leverage multiple LLM agents for generating climate risk scenarios that are not only creative and flexible but also *verifiably grounded* in scientific and economic evidence.

Contributions. (1) We introduce a novel multi–agent LLM architecture for scenario planning, demonstrating that decomposing the task and injecting domain knowledge via retrieval yield more reliable outcomes than a monolithic approach. (2) We integrate

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/RAVT8220

climate–finance disclosure *standards* into the generation loop: the agents explicitly check and enforce scenario requirements (illustrated with IFRS S2/TCFD rules) so that the AI–generated scenarios automatically adhere to external frameworks. (3) We present a quantitative evaluation on real–world climate scenarios, showing that ReGMS produces internally consistent, standards–aligned scenarios with full traceability. We compare against strong baselines (a centralized planner, independent uncoupled agents, and a single–LLM pipeline) and define ablation variants to probe the role of each component (retrieval, multi–agent roles, standards feedback). Our results indicate that ReGMS achieves a superior balance of consistency, realism, and compliance, without sacrificing any key aspect.

2 RELATED WORK

2.1 Multi–Agent Frameworks for Planning

Complex planning and reasoning tasks can often be handled more effectively by a team of specialized agents rather than a single model. Recent work has explored orchestrating multiple LLMs in a coordinated framework. For example, *HuggingGPT* [24] uses one LLM as a controller that delegates subtasks to expert models, demonstrating that an LLM "manager" can break down a problem and integrate results from others. Li et al. [10] formalize an agent–oriented planning paradigm in which a meta–agent creates and allocates sub–tasks to specialist agents, using feedback loops to adjust the plan. Role–based multi–agent interaction has also been explored in conversational and simulated environments [11, 20]. These studies suggest that dividing cognitive labor among LLM–based agents can yield results more consistent and robust than one–shot prompting. ReGMS adopts this philosophy: instead of a single model trying to handle everything, we design an ensemble of agents (Design, Transition, Physical, Standards, Assurance, etc.), each with a clear role. They operate in a directed acyclic workflow and can critique or correct each other's outputs (e.g., the Standards and Assurance agents review others' outputs), akin to a multi–agent debate or verification process. This structure helps maintain consistency and incorporate domain–specific checks, as advocated by Parmar et al. [21], who introduce verification and constraint–checking agents to improve reliability in complex reasoning tasks. Our work extends these ideas to the climate scenario domain, illustrating the benefit of a "society of LLMs" that cooperate under defined rules.

2.2 Retrieval–Augmented Generation for LLMs

LLMs augmented with retrieval mechanisms have demonstrated improved factual correctness and transparency in knowledge–intensive tasks. In *WebGPT*, for example, a GPT–3 model issues web searches and cites webpages, resulting in more accurate answers with verifiable sources [15]. Lewis et al. [9] showed that feeding retrieved documents into the prompt helps the model generate content grounded in those documents, reducing hallucinations. Surveys of retrieval–augmented generation [4] highlight that this approach not only boosts factual accuracy but also increases user trust, as outputs can be accompanied by citations. ReGMS leverages retrieval throughout its multi–agent pipeline. Each agent has access to a domain–specific knowledge corpus—including NGFS scenario documentation, climate datasets and reports, and disclosure guidelines—and can query this corpus for relevant snippets. Retrieved evidence is used to justify decisions and may directly constrain the agent's output. For instance, when the Transition agent needs the carbon price trajectory for a given scenario, it retrieves the official NGFS data rather than relying on memory; the Standards agent retrieves IFRS S2 text to verify required disclosure elements. By making retrieval an integral part of generation, we ensure that every aspect of the scenario can be traced back to an external source, addressing traceability and factual accuracy by construction. This approach builds on prior work in tool–augmented LLMs and self–consistency checks (e.g., LLMs that use calculators, databases, or code execution to validate outputs).

2.3 Climate Scenario Analysis under Standards

In climate finance, scenario analysis has been used to stress–test portfolios and assess risk under hypothetical futures. Early work by Battiston et al. [1] quantified how climate scenarios could impact the financial system, underscoring the importance of consistent assumptions (small changes in inputs led to nonlinear changes in outcomes). Recent studies have examined the value of climate disclosures: Ding et al. [2] find that investors reward firms that provide detailed scenario analyses, indicating a demand for credible and comprehensive scenarios. Regulators and standard–setters have responded by formalizing expectations: the Task Force on Climate–related Financial Disclosures (TCFD) set forth recommendations for scenario disclosures [17], and the ISSB's IFRS S2 (2023) now mandates reporting of climate risks using scenario analysis, with specific guidance on considering multiple time horizons and both transition *and* physical risk [7]. Our work addresses these practical needs by ensuring that generated scenarios align with such standards by design. We use reference pathways from the Network for Greening the Financial System (NGFS) [3] as fixed baselines (e.g., "Current Policies" vs. "Net Zero 2050"). We also incorporate physical climate data (downscaled CMIP6 projections) to add a hazard dimension to scenarios [25], as is increasingly expected in disclosures. By automating the assembly of scenarios that adhere to external frameworks and data sources, ReGMS bridges advanced AI planning techniques with the domain of climate risk management, providing a tool to generate scenarios that are not only plausible but also defensible to regulators and stakeholders. This aligns with recent calls to apply multi–agent methods to climate challenges [28].

3 METHODOLOGY

Our methodology constructs climate–change risk scenarios through a retrieval–grounded multi–agent system designed to satisfy the four objectives above. We leverage authoritative public data for both transition pathways and physical impacts, and enforce reporting constraints during generation.

3.1 Data and Preprocessing

Transition pathways (NGFS). We use climate transition scenarios from the Network for Greening the Financial System (NGFS) [3] as the backbone for transition dynamics. In particular, we focus on two archetypal NGFS scenarios: *Current Policies* and *Net Zero 2050*. From the NGFS scenario data (Phase V), we extract key variables

on a global (World) level at 5–year intervals from 2020 to 2050. We specifically take the carbon price (shadow price of CO_2) trajectory, since carbon pricing is a central transition indicator, along with other variables (like primary energy mix and emissions) used qualitatively in prompts. We interpolate the carbon price to an annual time series for 2020–2050 (using a monotonic spline to avoid overshooting). This yields a reference path π_t^{ref} for carbon price that *must be adhered to* by our scenario (we treat it as an immutable external constraint). We preserve the scenario's internal units and scale—e.g. the absolute price levels differ between Current Policies and Net Zero—so that ReGMS directly works with the official data without recalibration. Using standard scenarios as fixed references aligns with regulatory guidance that organizations should use externally published pathways in disclosures [7, 17].

Physical hazard data. To incorporate physical climate risk, we include a simplified hazard time series derived from downscaled climate projections. We use a dataset of climate hazard indices built from CMIP6 models (e.g. NASA NEX–GDDP downscaled data) for a chosen region. In our case study, we focus on the Asia–Pacific (APAC) region and use a composite hazard index h_t that combines extreme heat and extreme precipitation indicators (z–scored). We consider two forcing scenarios to pair with the NGFS transitions: for a high–emissions transition (Current Policies), we use a corresponding high–emissions physical pathway (e.g. SSP5–8.5), whereas for Net Zero, we use a lower–emissions physical scenario (e.g. SSP2–4.5 or similar moderate pathway) to reflect mitigation. The hazard data is not available for every year; instead, we have historical baseline values for 2015–2019 and projections for 2046–2050 (a common situation where physical impact models provide mid–century snapshots). Years 2020–2045 contain no hazard data (treated as missing). Our system is designed to handle this sparsity. Notably, the lack of near– and mid–term hazard values means the scenario must assume minimal physical impacts in those periods. The Standards agent (described later) ensures the final scenario still addresses short–, medium–, and long–term horizons by focusing adaptation measures in the late period where hazards manifest. We treat h_t as an exogenous input series (for 2046–2050) and ignore it in years where it's not available (no spurious interpolation of hazards). This setup ensures scenarios cover both transition and physical risks over multiple time horizons, in line with disclosure expectations [7, 17].

Retrieval corpus. A distinctive aspect of ReGMS is that all agents have access to a shared knowledge corpus and can perform retrievals. We compile a domain–specific text database consisting of: NGFS scenario documentation (descriptions of scenario assumptions and data dictionaries), climate model and dataset documentation (explaining variables like the hazard index and data availability), and excerpts from climate disclosure standards (IFRS S2 and TCFD technical guidance). This corpus is indexed for semantic search (e.g. using an embedding–based retriever). When an agent is invoked, it can query this corpus with a relevant question (via a toolkit integration, similar to [27]). Retrieved passages are then provided to the LLM agent as additional context or "tools," and key facts from them are expected to be cited or used explicitly in the agent's output. For example, the Design agent might retrieve text describing what the "Net Zero 2050" scenario entails (to correctly set the stage), or the Physical agent might retrieve documentation about the hazard data timeframe to know that 2046–2050 are the only future years with data. By grounding decisions in retrieved evidence, we reduce the risk of the LLM inventing unsupported claims and ensure traceability—each part of the scenario can be linked back to a source. Retrieval is performed over a *closed, curated* corpus (not the open web) containing NGFS documentation/data dictionaries, CMIP6/NEX–GDDP dataset notes, and IFRS S2/TCFD guidance. Each chunk stores source metadata (document and section/page), which the Assurance agent propagates into a citation trace. If a required fact is not retrieved, agents mark it unavailable rather than inventing values.

3.2 Retrieval–Grounded Multi–Agent Architecture

At the core of ReGMS is a collection of specialized LLM–based agents that collaborate to build the scenario sequentially. Each agent is assigned a distinct role and set of responsibilities, corresponding to different aspects of scenario design or verification. The agents communicate through a shared state (which contains the evolving scenario data and accumulated evidence) and are orchestrated in a fixed order (with optional feedback loops). Minor iterations are allowed if the Standards or Assurance agent signals that a revision is needed (in practice we found one pass was usually sufficient, as the agents are constrained by strict evidence–based rules). Figure 1 illustrates the system architecture of our proposed ReGMS framework. This modular pipeline design follows known patterns in multi–agent LLM systems (a central controller invoking specialized sub–agents) [5, 26]. Below, we summarize each agent's role:

- **Design Agent**: Formulates a high–level scenario blueprint. Given a prompt specifying the scenario (e.g. "Current Policies in APAC" or "Net Zero 2050 global"), it selects key parameters: which NGFS scenario (and underlying IAM model) to follow for transition data, the region of interest and appropriate climate model/SSP for physical hazards, the time horizon (e.g. 2020–2050), and any necessary "guardrails." It uses retrieved facts to ensure choices are compatible (e.g. confirming that hazard data is available for the chosen region and scenario). The output of the Design agent is a structured brief that fixes these exogenous choices for downstream agents, for example: *"Use NGFS Net Zero 2050 (REMIND model) for transition; focus on APAC region hazards with GFDL–CM4 under SSP2–4.5; consider years 2020–2050; include a short–term (2020–2025) and late–term (2046–2050) window as required."*
- **Transition Agent**: Proposes the trajectory for transition–related variables under the scenario. This includes: the annual carbon price π_t (which in fact is not freely chosen—π_t must follow the NGFS reference π_t^{ref} provided by the design), the green technology share g_t (fraction of energy from low–carbon sources), and an initial adaptation level a_t (representing baseline adaptive capacity or policy in early years). Essentially, the Transition agent has two main degrees of freedom: (i) it sets a path for g_t (e.g. starting value and growth rate) consistent with the scenario narrative (e.g. modest growth under Current Policies vs. rapid under Net Zero), and (ii) it may set an initial a_t or adaptation policy stance (though

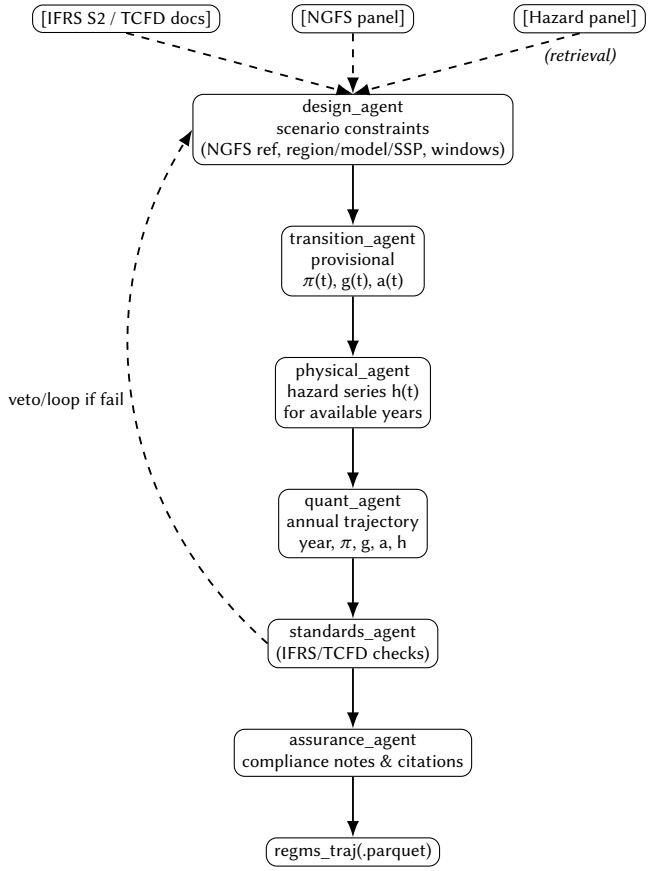

Figure 1: ReGMS system architecture. Inputs (NGFS, hazard, IFRS/TCFD) are available to all agents via retrieval; for readability, we show retrieval into `design_agent` which initializes constraints. Agents then flow down to `assurance_agent`, with a feedback loop from `standards_agent`.

later the Physical agent will adjust adaptation in response to hazards). Crucially, the carbon price path is locked to the provided reference data — the agent is instructed to *use the exact reference* π_t^{ref} for each year's carbon price. This guarantees perfect fidelity to the chosen scenario's transition pathway, enhancing credibility and comparability (mirroring how analysts use official scenarios directly). The Transition agent retrieves relevant scenario descriptions or data (e.g. emissions trends in the scenario) to guide g_t and a_t choices. Its output is a preliminary sequence for $\{\pi_t, g_t, a_t\}$ over the analysis period.

- **Physical Agent**: Next, the Physical agent brings in the physical risk dimension. Based on the design brief's region and scenario, it loads the corresponding hazard data series (the exogenous h_t for available years). It then examines the relationship between hazards and adaptation: since hazard data is only present in 2046–2050, the agent notes that adaptation measures would mainly be needed in that late period. It may suggest how a_t should respond to h_t — for example, keeping adaptation low in years with no or low hazards (early/mid–term) and increasing it when hazards peak. It can also impose constraints like: adaptation level should not jump unrealistically or exceed certain bounds (e.g. you cannot go from no adaptation to full adaptation overnight). The Physical agent uses scientific context (possibly retrieving info on plausible adaptation responses or the severity of projected hazards) to ensure it interprets h_t correctly. Its output is essentially an enriched view of the hazard timeline plus any recommendations for adjusting the adaptation trajectory to align with expected impacts.
- **Quantitative Coupling Agent**: The Quantitative (Quant) agent takes the proposals from Transition and Physical agents and synthesizes them into a single coherent scenario timeline. It produces the final annual time series for all variables: carbon price π_t, green share g_t, adaptation level a_t, and hazard h_t. This involves interpolation and smoothing as needed and enforcing any hard constraints or logical consistency. The Quant agent merges proposals, enforces monotonicity/bounds, and outputs annual (π_t, g_t, a_t, h_t) consistent with constraints.
- **Standards Agent**: Once a full draft scenario timeline is assembled, the Standards agent checks it against disclosure standards and other external criteria. It verifies that the scenario includes all required components: e.g. does it cover multiple time horizons (short–, medium–, long–term as defined by IFRS S2)? Does it address both transition and physical risks clearly? It also checks if key assumptions are documented with sources (looking at the evidence collected so far). If any requirement is unmet, the Standards agent will flag an issue and possibly trigger a revision. For example, if the scenario somehow neglected a long–term horizon (post–2040) or if no physical risk was included, this agent would prompt the Design or Transition agent to incorporate that missing piece and run again. In our implementation, the Standards agent uses a set of hard rules derived from IFRS S2 and TCFD guidance (e.g. *"Scenario must include at least one time frame beyond 20 years"*, *"Both a transition scenario and a physical scenario aspect must be present"* [7, 17]). It ensures the final scenario output is something an auditor or regulator would accept as properly constructed.
- **Assurance Agent**: Finally, the Assurance agent conducts an internal consistency audit and prepares a traceability report. It computes diagnostic metrics on the finalized trajectory, such as the RMSE between $\{\pi_t\}$ and the NGFS reference (transition consistency), the correlation between h_t and a_t (physical risk alignment), etc. If it finds any major issues (e.g. adaptation does not respond at all to hazard changes), it could signal for a scenario revision (in our experiments, this was rarely needed since earlier agents already enforce these links). The Assurance agent also compiles all evidence references to justify the scenario. For instance, it will annotate that the carbon prices came from [3], hazard figures from a specific climate model source, and note compliance with IFRS S2 expectations by citing the standard. The end product is not just the data trajectory but also a short explanatory document with footnotes. This provides end–users with an

explanation of *why* each part of the scenario was chosen. The Assurance agent ensures the scenario is not only correct and standard–aligned, but also *explainable* and defensible — a crucial aspect for real–world adoption where stakeholders require justification for the scenario assumptions [17].

Throughout this process, a simple **Coupling mechanism** monitors for any conflicts between agents' outputs and resolves them according to predefined priorities. For example, if (hypothetically) the Transition agent attempted to set a carbon price that deviated from the NGFS reference, the system would override that to enforce $\pi_t = \pi_t^{\text{ref}}$. If the Transition and Physical agents gave conflicting suggestions for a_t in overlapping years, the coupling logic might dictate that the Physical agent's input (hazard–driven) takes precedence in years where hazard data is available, whereas Transition's policy holds in years without hazard data. These rules ensure the agents ultimately converge to a single feasible scenario without contradiction. We enforce four non–negotiable constraints: (i) carbon price is pinned to the NGFS reference π_t^{ref} for 2020–2050; (ii) hazards are used only in supported years (2046–2050) and are not interpolated; (iii) $g_t \in [0, 1]$ and is non–decreasing; (iv) a_t is bounded and may change only where hazard data exist. Violations trigger correction in Quant and can cause a Standards/Assurance veto. In our implementation using LangGraph, the orchestration order is `Design` → `Transition` → `Physical` → `Quant` → `Standards` → `Assurance`, with potential feedback loops from Standards (or Assurance) back to earlier stages if needed. In practice, one pass sufficed given the strong constraints we imposed. All agents (except the trivial coupling logic) are realized via calls to an LLM (GPT–5 [18, 19] in our case) using carefully crafted prompts that encode their role and the current state. We run the LLM deterministically (temperature 0) for consistency. The retrieval is handled via a toolkit (using a LangChain interface) that injects retrieved snippets into the prompts as needed. Our implementation logs every agent's decisions and sources, producing a rich trace that can be audited.

3.3 Game–Theoretic Formulation

To formalize the coordination among agents, we frame the scenario generation as a *constrained stochastic game* [13, 23] played over discrete time steps $t = 2020, \ldots, 2050$. Each agent is modeled as a player in the game with certain actions and preferences, subject to shared evidence–based constraints. This provides a theoretical lens to analyze incentive alignment and convergence.

Players and State. Let the players

$$\mathcal{P} = \{\text{Transition, Physical, Quant, Standards, Assurance}\}$$

(we omit Design for simplicity since it just initializes context). The state at year t is defined as

$$s_t = (\pi_t,\ g_t,\ a_t,\ h_t,\ \mathcal{E}),$$

comprising the carbon price, green share, adaptation level, hazard level, and the evidence set $\mathcal{E}$ accumulated so far. The evidence $\mathcal{E}$ includes retrieved facts such as the NGFS reference values for π_t, the years where hazard data exists, definitions from standards (e.g. what counts as "short–term"), etc. Evidence effectively becomes common knowledge that constrains what actions are allowed. The state evolves as agents sequentially choose actions that update parts of (π, g, a) or add to $\mathcal{E}$ (through retrieval).

Actions and Constraints. Each agent i has an action space A_i, but at any state s_t the *feasible* actions $A_i(s_t) \subseteq A_i$ are restricted by the evidence and prior agent decisions. For example, the Transition agent's action of setting $\{\pi_t\}$ must satisfy $\pi_t = \pi_t^{\text{ref}}$ for all t (if the NGFS data is in $\mathcal{E}$, any deviation is invalid). Similarly, evidence that hazard data exists only in 2046–2050 means the Physical agent cannot base adaptation on hazard for $t < 2046$ (it might still suggest low adaptation in those years). The Standards agent's actions are essentially to "accept" or "reject and demand fixes," which depend on whether all requirements in $\mathcal{E}$ (e.g. IFRS checks) are met. Formally, evidence imposes constraints like: $|\Delta g_t| \le \delta_g$ for Transition (limit on change in green share per year based on realistic policy speeds, as derived from data or reports); physical must choose one of the available climate models/SSPs for hazard (cannot invent data); standards must veto if required horizon windows are missing, etc. These constraints keep the game within the bounds of plausible and compliant scenarios.

Utility (Payoffs). We define utility functions for each agent corresponding to the scenario objectives (these are conceptual, since our agents are not learning via reward but are prompted to aim for these goals). Key utilities include:

$$u_{\text{trans}} = -\text{RMSE}(\{\pi_t\}, \{\pi_t^{\text{ref}}\}),$$

the negative root–mean–square error of the carbon price path vs. the NGFS reference, for the Transition agent (maximized when π_t exactly matches reference, yielding zero error).

$$u_{\text{phys}} = \rho(\{h_t\}, \{a_t\}),$$

the Spearman rank correlation between hazard and adaptation series (late–period) for the Physical agent, rewarding higher correlation (adaptation increases when hazard increases).

$$u_{\text{quant}} = -\text{Gap}(\{h_t\}, \{a_t\}),$$

for the Quant agent, where Gap is the mean absolute deviation between a normalized hazard and adaptation level (years 2046–2050). This gap measures magnitude alignment (it's 0 if adaptation levels are proportional to hazard each year; larger if adaptation is too low or high relative to hazard). We compute Gap as the mean absolute deviation between a min–max scaled hazard series and the (unclipped) adaptation series over 2046–2050; thus large over/under–responses can yield Gap > 1. Smaller gap (closer to 0) is better. We also define binary utilities for the checking agents:

$$u_{\text{std}} = \mathbf{1}\{\text{IFRS/TCFD checks pass}\},$$

for the Standards agent (1 if the scenario satisfies all required disclosure criteria, 0 if not), and

$$u_{\text{ass}} = \mathbf{1}\{\text{traceability/consistency checks pass}\},$$

for the Assurance agent (1 if the final scenario is internally consistent and all statements are backed by citations, etc.). Essentially, Standards and Assurance agents seek to enforce hard constraints rather than optimize a continuous metric.

In an ideal scenario generation outcome, all agents achieve their objectives: Transition attains zero RMSE (perfect fidelity), Physical/Quant reach strong hazard–adaptation alignment (high ρ and near–zero gap), and Standards/Assurance have all checks passing. These correspond to objectives (i), (ii), and (iv), while objective (iii)

is reported via the g_t trend. The multi–agent process is designed to reach or approach this optimum.

Dynamics and Equilibrium. The game is played sequentially according to the pipeline order. In our fixed–order deterministic implementation, "convergence" means reaching a constraint–satisfying terminal state in finitely many orchestration steps (typically one pass, with optional veto–triggered re–runs), rather than convergence of a learned policy. At each stage (node in the orchestration graph), agent i chooses an action that moves the state $s_t \to s'_t$. The Standards and Assurance agents have the ability to veto or request changes, which introduces a backtracking mechanism if something unacceptable is found. The process terminates when a *terminal state* is reached where no agent wishes to unilaterally deviate (i.e. the scenario cannot be improved without breaking a rule) and all constraints are satisfied. We can view this terminal state as a form of equilibrium: all agents' local incentives (as embodied by u_i) are satisfied given the evidence constraints, so the scenario is stable. In practice, because our agents are implemented via deterministic rules/prompting rather than learning, we ensure convergence by construction using the coupling logic. We can note that the centralized Planner baseline (described next) is akin to a single–player version of this game, where one agent optimizes a weighted combination of all u_i (essentially an oracle that tries to achieve all criteria at once, with full access to evidence). This game–theoretic view helps reason about how dividing the task among agents (with their specialized objectives) compares to a monolithic approach. When the evidence constraints are satisfiable and not directly conflicting, iterative best–response updates can reach a stable constraint–satisfying terminal state; we use the centralized Planner baseline as a point of comparison.

3.4 Baselines and Ablations

We evaluate ReGMS against several baseline approaches: **(i) Planner** – an IAM–style planner that follows hard–coded rules (essentially using the NGFS data directly for transition and setting adaptation proportional to hazard), **(ii) Independent agents** – a variant where each subtask is handled by a separate LLM agent *without* any coordination or shared memory (so they do not influence each other), and **(iii) Single–LLM** – a pipeline where a single large model is prompted to produce the entire scenario (with retrieval access, but no task decomposition or inter–agent critique). These represent, respectively, an ideal consistent policy, a completely uncoordinated policy, and an unconstrained end–to–end policy. We also define ablation variants to probe each key component of ReGMS: for example, removing retrieval, collapsing all roles into one agent, and removing standards feedback, but omit detailed ablation results due to space constraints. Additionally, we experimented with a **multi–agent reinforcement learning** approach, where the agents' policies were learned via MARL instead of prompted LLMs. This baseline (denoted *ReGMS–MARL*) allows agents to learn coordination through trial–and–error [22], but it does not guarantee strict adherence to constraints unless those are directly encoded in the reward.

4 FINDINGS

We evaluated ReGMS and the baselines on the two NGFS scenarios: Current Policies and Net Zero 2050. Each method generated a scenario trajectory from 2020 to 2050. We then computed quantitative diagnostic metrics for (i) transition fidelity, (ii) physical risk alignment, and (iii) mitigation ambition; we additionally report (iv) compliance & traceability via qualitative checks of time–horizon coverage and citation provenance. Table 1 summarizes the quantitative metrics for all methods.

Overall, **ReGMS achieved the best balance across all criteria**: it strictly satisfied the transition pathway constraints and significantly outperformed the Single–LLM approach in hazard alignment, while also providing traceable, standards–compliant outputs. These findings reinforce the observation that structured multi–agent frameworks can outperform a single LLM on complex tasks by maintaining consistency across specialized subtasks [10, 21].

4.1 Transition Consistency

A fundamental requirement is that the scenario's key transition trajectory (especially carbon price or emissions) remains consistent with the chosen reference scenario. By design, ReGMS exactly follows the reference in all years for both scenarios (RMSE = 0 in Table 1). The Planner baseline and Independent baseline also achieve perfect tracking, since they were configured to do so. In contrast, the Single–LLM baseline showed notable deviations: under Current Policies it started on track but then drifted upward, overshooting the reference by about $10 by 2050. This resulted in a substantial RMSE of 2.5 (see Table 1) and a final–year error of +$10. In the Net Zero case, the Single LLM matched the 2050 target price, but it still had minor inconsistencies along the way (perhaps because the prompt explicitly mentioned a net–zero outcome), but it still had minor inconsistencies along the way (RMSE 3.1 over the years). In contrast, ReGMS's Transition agent was explicitly constrained to use the reference data, eliminating any chance of drift. As a result, ReGMS (as well as Planner and Independent) achieved *zero* RMSE and zero final–year drift in both scenarios, whereas Single–LLM did not (especially under Current Policies). ReGMS guarantees such fidelity by construction, demonstrating the benefit of retrieval grounding for quantitative consistency.

Observation. In Net Zero 2050 the trend is steep and obvious; Single–LLM still shows minor deviations, while ReGMS, Planner, and Independent exactly match the NGFS reference (RMSE = 0).

4.2 Physical Risk Alignment

A novel aspect of our evaluation is measuring how well each method aligns *adaptation* actions with *physical hazard* impacts. We use two metrics: the Spearman rank correlation between hazard h_t and adaptation a_t in the late period (2046–2050 when hazard data is available), and an "adaptation response gap" defined as the mean absolute deviation between the min–max scaled hazard series and the (unclipped) adaptation series over 2046–2050 (so extreme over/under–responses can yield gaps > 1). Higher correlation and lower gap indicate better alignment, as adaptation levels should rise and fall proportionately with hazard levels.

In *Current Policies*, ReGMS achieved a Spearman $\rho \approx 0.87$, indicating a very strong positive association: years with higher hazard saw higher adaptation levels. The Planner baseline unsurprisingly had $\rho \approx 0.95$ (essentially a perfect correlation, since it explicitly sets $a_t \propto h_t$ by design). The Independent baseline had $\rho = 0$ because a_t was static (no response to hazard, so the points form a flat horizontal line). The Single–LLM managed $\rho \approx 0.45$, a moderate correlation—its adaptation tended to increase in some high–hazard years, but inconsistently (the scatter is noisy). The adaptation response gap metrics tell a similar story: in Current Policies, ReGMS had a gap of ≈ 0.21, whereas Single–LLM's gap was ≈ 0.32. Independent's gap was 0.5, reflecting a largely flat adaptation profile despite varying late–period hazards. The Planner's gap was 0 (perfect alignment). These numbers confirm that ReGMS tightly coupled adaptation to hazard, reducing misalignment by roughly one–third relative to Single–LLM. In practical terms, under ReGMS the severe late–century climate impacts prompted a commensurate adaptation response (e.g., increasing adaptation level from 0 to 2 as the hazard index rose from moderate to high), whereas Single–LLM might under–respond or respond inconsistently.

In *Net Zero 2050*, differences were even more pronounced. Because aggressive mitigation in Net Zero limits physical climate change, the hazard index by 2046–2050 is relatively low. ReGMS again achieved $\rho \approx 0.87$ (high correlation) with an extremely small adaptation gap (≈ 0.01, essentially near–perfect proportionality). The Planner remained high as well ($\rho \approx 0.89$, gap 0). Strikingly, the Single–LLM produced $\rho \approx -0.63$ – a *negative* correlation – and a huge adaptation gap > 1.1. In other words, the Single–LLM's adaptation strategy in Net Zero was actually counterproductive: it allocated more adaptation in some years with lower hazard and less when hazard was higher, essentially getting the relationship backwards. This suggests the single model misunderstood the scenario's needs, perhaps assuming that if warming is limited then adaptation isn't urgent. However, even a low–warming scenario still involves some physical risk that must be addressed [16]. ReGMS, by design, has separate agents to ensure that even a low–hazard future still prompts appropriate adaptation: the Physical agent sees that some hazard exists and recommends proportional adaptation, and the Standards agent enforces that both risk types be covered. Thus ReGMS's Net Zero scenario remains coherent ("even with lower physical risk, we still adapt to whatever risk remains"), whereas the Single–LLM's output would fail a basic consistency check. This underscores the value of our multi–agent coordination: by assigning an agent specifically to hazard–adaptation alignment, ReGMS avoided the oversight made by the single holistic model.

Across both scenarios, ReGMS exhibited strong positive hazard–adaptation coupling, second only to the idealized Planner. The Independent approach had no coupling, and the Single–LLM was inconsistent (and in one case completely wrong–way). These findings show that ReGMS achieves objective (ii) of internal consistency between the transition and physical domains. In practice, this means an analyst using ReGMS can confidently say, "Because our scenario entails high physical risks in late years, our adaptation investments increase accordingly," and point to the data as evidence. This is consistent with NGFS Phase V, which highlights (chronic) physical risk as a major driver of projected losses and notes impacts are already non–negligible by 2030 across scenarios [3].

4.3 Mitigation Pathways and Trends

Another perspective is how each method handles the *mitigation* trajectory—here represented by the green technology share g_t over time. The NGFS narratives qualitatively describe expectations (e.g., under Net Zero, rapid decarbonization; under Current Policies, slower progress), but they do not provide an explicit g_t timeseries, leaving this as a degree of freedom for our agents.

For *Current Policies*, ReGMS set a steadily rising g_t from a low initial value (reflecting today's energy mix) to a higher value by 2050. The fitted slope is about +0.0125 (i.e., +1.25 percentage points per year). The Planner was fixed at +0.0100 (+1.0%/yr), so ReGMS is slightly more ambitious on mitigation than the Planner in this scenario. The Independent baseline kept g_t essentially flat (slope 0, no growth in renewables—an implausibly pessimistic assumption, but it serves as a contrast). Interestingly, the Single–LLM in Current Policies also produced an increasing g_t trajectory with slope ≈ 0.0123 (+1.23%/yr), nearly matching ReGMS. This suggests the LLM did infer that even under current policies some technological progress would occur (perhaps because real–world trends or minor policies imply some growth in renewables). In this case, Single–LLM wasn't far off: ReGMS and Single–LLM both envision modest green transitions in a no–new–policy world, whereas Independent unrealistically assumes none. The slight edge is that ReGMS's Transition agent had access to evidence or context about expected renewable growth even without new policies, prompting it to introduce a bit more growth than the Planner. Overall, in Current Policies, all methods except Independent agree on a low but non–zero mitigation trajectory.

In *Net Zero 2050*, a much steeper climb in g_t is expected (since reaching net–zero CO_2 by 2050 implies a near–complete shift to clean energy). Indeed, ReGMS's g_t skyrockets — ending with a slope of ≈ 0.0128 (+1.28%/yr) and achieving a high final share (though not quite 100%). The Single–LLM ramped up even faster: slope ≈ 0.0142 (+1.42%/yr), apparently pushing towards an even more aggressive end state (perhaps assuming nearly full decarbonization by 2050). The Planner remained at its fixed +0.0100 (since it was not scenario–aware in our simple implementation, it undershot what Net Zero demands). The Independent again flat–lined (0 slope, clearly implausible for a Net Zero future). Thus, in Net Zero, both ReGMS and Single–LLM recognized the need for a transformative increase in green share; Single–LLM went a bit further (overshooting what policy might feasibly achieve by 2050), while ReGMS provided a slightly more moderated (yet still very ambitious) path. The Independent was 0 in both, and the Planner was consistently +1.0 in both (hence in Net Zero it undershot the needed transition). In both scenarios ReGMS increases g_t appropriately; Planner is conservative; Independent flat; Single-LLM steeper in NZ.

5 DISCUSSION

Combining the results, we find that ReGMS consistently delivers high–quality scenarios meeting multiple criteria simultaneously, whereas each baseline falls short on at least one dimension:

- **Planner:** The centralized Planner excels in numeric consistency and alignment (by construction, it nails carbon prices and hazard coupling), but it is a fixed policy rather than a flexible AI approach. It cannot explain why it made choices

Table 1: Quantitative comparison of scenario outcomes for ReGMS and baselines under two scenarios: *Current Policies (CP)* and *Net Zero 2050 (NZ)*. Transition fidelity is measured by carbon price RMSE (lower is better). Hazard–adaptation alignment is measured by Spearman ρ (higher is better) and adaptation gap (lower is better). Mitigation trend is the annual increase in green tech share Δg. ReGMS achieves the best overall balance: perfect transition fidelity and strong cross–domain consistency.

2*Method	Current Policies (CP)				Net Zero 2050 (NZ)			
	Price RMSE	Haz–Adapt ρ	Adapt gap	Δg (%/yr)	Price RMSE	Haz–Adapt ρ	Adapt gap	Δg (%/yr)
ReGMS	0.0	0.87	0.21	1.25	0.0	0.87	0.01	1.28
Planner	0.0	0.95	0.00	1.00	0.0	0.89	0.00	1.00
Independent	0.0	0.00	0.50	0.00	0.0	0.00	0.50	0.00
Single–LLM	2.5	0.45	0.32	1.23	3.1	−0.63	1.11	1.42
ReGMS–MARL	0.9	0.89	0.00	1.00	1.3	0.95	0.00	1.00

(no citations or narrative) and cannot easily adapt to scenarios beyond its hard–coded rules (e.g., it did not adjust g_t for Net Zero vs. Current Policies). ReGMS matches the Planner on transition fidelity (zero carbon price error) and attains strong hazard–adaptation alignment while being far more extensible: it can ingest new evidence, handle different scenario narratives, and produce explanatory text.

- **Independent:** The uncoupled baseline fails on integrated thinking. By neglecting mitigation progress and hazard–triggered adaptation, it produces a scenario that would be deemed implausible or incomplete in a disclosure context (imagine claiming "despite rising hazards, our adaptation response is nil"—clearly unacceptable). ReGMS strongly outperforms it, showing the value of coordinating transition and physical planning (objectives (ii) and (iii)). For example, in both scenarios the Independent approach had an adaptation gap of 0.5 (indicating negligible adaptation response), whereas ReGMS reduces it to 0.21 (CP) and 0.01 (NZ).
- **Single–LLM:** The one–shot LLM baseline shows that even a powerful model with data access struggles to satisfy all constraints simultaneously without structured guidance. It performed reasonably on the mitigation trend (aligning with obvious scenario cues), but had trouble with exact quantitative fidelity and especially with linking physical and transition aspects. The Net Zero hazard–adaptation failure is a cautionary example: an end–user of that scenario might be misled to think "perhaps hazards will decrease, so we can afford less adaptation," which is a confused takeaway. In contrast, ReGMS provides a coherent story: even in a low–warming scenario, whatever physical risks remain are addressed proportionately. This aligns with recent findings that orchestrating multiple LLMs with tools and retrieval can yield more accurate, validated results in decision–support contexts [21, 24]. Furthermore, Single–LLM lacks any verification stage: it had no mechanism to catch its own errors (no "assurance" step), so mistakes like misaligning adaptation went uncorrected. ReGMS's Assurance agent, by contrast, would have flagged such inconsistencies (e.g., a low Spearman ρ) and prompted a revision if they occurred.
- **Multi–agent RL:** We also tested a multi–agent reinforcement learning baseline (ReGMS–MARL). It achieved near–perfect hazard–adaptation alignment (comparable to Planner) but did not strictly enforce the transition path (incurring a small price RMSE; see Table 1). This suggests that an end–to–end MARL approach can learn coordinated policies [22], but our LLM–based approach ensures critical constraints by design and offers greater transparency in reasoning.

In terms of compliance and traceability (objective (iv)), ReGMS has clear advantages. Every output from ReGMS is backed by source citations (e.g., each carbon price point is footnoted to "NGFS 2024 data," hazard figures to climate model data, and statements about time horizons to IFRS S2 [7]). This provides an audit trail that regulators explicitly value (IFRS S2 expects companies to disclose the sources of scenarios and assumptions used [7]). The Single–LLM, while it might produce a fluent narrative, cannot provide reliable citations for the numbers it generated (unless one augmented it with retrieval, in which case it's no longer a single pure LLM approach). We did not assign a numeric score for "standards compliance" because ReGMS was constructed to always pass those checks (the Standards agent ensures it) and the baselines didn't produce full narrative reports to assess. Qualitatively, however, ReGMS outputs would meet IFRS S2 expectations (multiple time horizons, clear articulation of scenario methodology, etc.), Single–LLM might miss some elements (e.g., explicitly mentioning the timeframes or source data), and the Independent baseline would fail to demonstrate resilience and completeness. The Planner, if augmented with narrative, could be made to report required information, but it would still lack explanation beyond "we forced these values."

In summary, our findings confirm that the retrieval–grounded multi–agent strategy (ReGMS) succeeds in generating climate risk scenarios that are consistent with established pathways and internally coherent across domains. It uses external knowledge to lock down critical constraints (yielding perfect transition adherence), and it coordinates agents so that physical impacts and adaptation responses are logically linked. The result is a scenario that is plausible, defensible, and rich in information—qualities essential for real–world climate risk analysis and decision support. Overall, ReGMS provides the strongest trade–off among learning–based baselines, while the rule–based Planner remains an upper bound on alignment metrics but lacks flexible narration and citation provenance.

REFERENCES

[1] Stefano Battiston, Antoine Mandel, Irene Monasterolo, Franziska Schütze, and Gabriele Visentin. 2017. A climate stress-test of the financial system. *Nature Climate Change* 7 (2017), 283–288.

[2] Tongqing Ding, Jonathan Jona, Brad Potter, and Naomi Soderstrom. 2025. Are climate scenario analysis disclosures valued by investors? *Journal of Accounting and Public Policy* 51 (2025), 107313.

[3] Network for Greening the Financial System (NGFS). 2024. *NGFS Long-term Scenarios for Central Banks and Supervisors.* Technical Report. Network for Greening the Financial System.

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2024).

[5] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. In *Proceedings of the 12th International Conference on Learning Representations (ICLR).*

[6] Leo Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43 (2025), 1–55.

[7] International Sustainability Standards Board (ISSB). 2023. *IFRS S2 Climate-Related Disclosures.* Technical Report. IFRS Foundation.

[8] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55 (2023), 1–38.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS).*

[10] Ao Li, Yuexiang Xie, Songze Li, Fugee Tsung, Bolin Ding, and Yaliang Li. 2025. Agent–Oriented Planning in Multi–Agent Systems. In *Proceedings of the 13th International Conference on Learning Representations (ICLR).*

[11] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems (NeurIPS).*

[12] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).* 3214–3252.

[13] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML).*

[14] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).* 1906–1919.

[15] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser–assisted question–answering with human feedback. *arXiv preprint arXiv:2112.09332* (2022).

[16] Intergovernmental Panel on Climate Change (IPCC). 2022. *Climate Change 2022: Impacts, Adaptation and Vulnerability.* Cambridge University Press.

[17] Task Force on Climate-related Financial Disclosures. 2017. *Final Report: Recommendations of the Task Force on Climate-related Financial Disclosures.* Technical Report.

[18] OpenAI. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf.

[19] OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/.

[20] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th ACM Symposium on User Interface Software and Technology (UIST).*

[21] Mihir Parmar, Xin Liu, Palash Goyal, Yanfei Chen, Long Le, Swaroop Mishra, Hossein Mobahi, Jindong Gu, Zifeng Wang, Hootan Nakhost, Chitta Baral, Chen-Yu Lee, Thomas Pfister, and Hamid Palangi. 2025. PlanGEN: A Multi-Agent Framework for Generating Planning and Reasoning Trajectories for Complex Problem Solving. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[22] James Rudd-Jones, Mirco Musolesi, and María Pérez-Ortiz. 2025. Multi-Agent Reinforcement Learning Simulation for Environmental Policy Synthesis. In *Proceedings of 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* 2890–2895.

[23] Lloyd S. Shapley. 1953. Stochastic Games. In *Proceedings of the National Academy of Sciences*, Vol. 39. 1095–1100. https://doi.org/10.1073/pnas.39.10.1095

[24] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. In *Advances in Neural Information Processing Systems (NeurIPS).*

[25] Bridget Thrasher, Weile Wang, Andrew Michaelis, Forrest Melton, Tsengdar Lee, and Ramakrishna Nemani. 2022. NASA Global Daily Downscaled Projections, CMIP6. *Scientific Data* 9 (2022), 262.

[26] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *Proceedings of the First Conference on Language Modeling (COLM).*

[27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR).*

[28] Vahid Yazdanpanah, Sara Mehryar, Nicholas R. Jennings, Swenja Surminski, Martin J. Siegert, and Jos van Hillegersberg. 2020. Multiagent Climate Change Research. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* 1726–1731.