Truth Negotiators: Language-Driven Agents Resolving Cross-Modal Contradictions in Sustainable Finance

Nataliya Tkachenko
AI Centre of Excellence
Lloyds Banking Group
first@author.edu

Aritra Chakravarty
AI Centre of Excellence
Lloyds Banking Group
second@author.edu

Abstract

Trust in ESG (Environmental, Social, and Governance) ratings is eroding, with mounting evidence of large divergences between agencies, opaque methodologies, and overreliance on corporate self-reporting. We propose *Truth Negotiators*, a multi-agent system powered by large language models (LLMs) that reconciles conflicting assessments using multimodal evidence: narrative text, Earth observation (EO) imagery, and structured alternative data. Agents specialised in each modality maintain beliefs about real-world assets, express them in natural language, and engage in structured negotiation moderated by a coordinating agent. The system maintains *object-centric embeddings*, persistent representations of asset states, updated through trust-weighted evidence integration. In a synthetic but realistic case study of a hyperscale data centre's '100% renewable' claim, the framework detected temporal mismatches between annual procurement equivalence and real-time energy sourcing. Negotiation led to a qualified verdict (*"annual 100% claim not hourly-verified"*) with a final system confidence of 0.78, complete resolution of initial contradictions, and adaptive trust reallocation towards EO and structured data agents. Compared with text-only, EO-only, and naïve fusion baselines, the approach uniquely delivered a nuanced, explainable conclusion grounded in cross-modal corroboration. We formalise the negotiation and update process, present architecture options, and outline a demonstration setup for live or scripted runs.

1 Introduction

Over the past decade, ESG ratings have shifted from niche indicators to central tools in investment decision-making. However, recent research and media reports show that ratings for the same firm can diverge dramatically across agencies, with correlations as low as 0.3 (1). These divergences arise from methodological differences, inconsistent data sourcing, and selective corporate disclosure. The opacity of the ratings process has contributed to declining investor confidence.

A common thread in these criticisms is reliance on a single dominant data modality – typically self-reported corporate disclosures – supplemented by analyst judgment. Earth observation imagery, investigative news, and structured third-party datasets remain underutilised. Yet, each can independently confirm or contradict aspects of a company's sustainability claims. The central question is not whether any one source is correct, but *how to reconcile them* when they disagree.

We revisit the classic vision of multi-agent systems (MAS) (2; 3; 4), where autonomous agents communicate, cooperate, and negotiate toward shared goals. By combining LLM-powered agents specialised in different modalities, we implement a form of 'truth negotiation' in which conflicting evidence is debated and reconciled into a shared, verifiable asset profile.

Preprint. Under review.

Contributions. (i) A formal framework for LLM-mediated multi-modal negotiation in sustainable finance. (ii) Object-centric embeddings updated by trust-weighted, cross-validated evidence. (iii) Architecture options (centralised, distributed, hybrid) with discussion of trade-offs. (iv) A detailed, synthetic but realistic data-centre case study with real data references and evaluation plan.

2 Background

2.1 Diverging ESG ratings

The phenomenon of *divergent ESG ratings*, where different agencies assign markedly different scores to the same company, has emerged as one of the most pressing credibility challenges in sustainable finance. Empirical studies (1) report correlations between leading ESG ratings providers as low as 0.3, a stark contrast to the 0.9+ correlations observed in credit ratings. This divergence undermines the comparability of ESG assessments and introduces ambiguity into capital allocation decisions, where even marginal differences in a score can influence investment flows.

From a methodological perspective, divergences stem from at least three interrelated sources: *(1) scope*, or the set of attributes considered material for ESG assessment; *(2) measurement*, or how those attributes are operationalised into indicators; and *(3) aggregation*, or the weighting scheme used to combine indicators into a composite score. For example, one agency might heavily weight carbon emissions per unit revenue (environmental pillar) while another prioritises board independence (governance pillar). Even within a single pillar, measurement discrepancies (such as direct vs. life-cycle emissions accounting) can yield vastly different results. Aggregation rules further amplify divergence: a strong governance score can offset a poor environmental score in one methodology but not in another.

These methodological differences are compounded by *data source asymmetry*. Many agencies rely predominantly on self-reported corporate disclosures, which are subject to selective framing and omission. Others supplement disclosures with proprietary research, NGO reports, or governmental databases. While richer in diversity, such supplementation is often opaque, leaving end-users uncertain about provenance or verification standards. A growing body of work (e.g., Christensen et al., 2022) suggests that overreliance on voluntary disclosures increases susceptibility to ‘greenwashing’, where firms overstate their ESG performance without making substantive operational changes.

The result is a ratings landscape in which ostensibly objective metrics can reflect fundamentally different ‘truths’ about the same underlying entity. In practical terms, this creates *information friction* in capital markets: investors face uncertainty about which rating to trust, companies can ‘rating shop’ for the most favorable assessment, and regulators struggle to enforce consistency without stifling methodological innovation.

The multi-agent approach proposed in this paper addresses this problem by introducing *negotiation across modalities* as a structural mechanism for reconciling divergence. Rather than treating ESG ratings as immutable inputs, the *Truth Negotiators* framework deconstructs each rating into its underlying claims, attributes, and evidentiary basis. Text agents parse the narrative justifications embedded in ratings reports or public statements; vision agents interrogate Earth observation data for physical evidence supporting or contradicting those claims; data agents bring in structured indicators from independent registries, sensor networks, or transactional records. These agents exchange beliefs and challenge one another through a coordination protocol grounded in argumentation theory (4).

Crucially, the negotiation process not only produces a reconciled embedding of the asset’s ESG state but also maintains a transparent reasoning trail. By revealing which modalities align and which contradict, the system enables stakeholders to understand the root causes of divergence, whether due to genuine methodological differences or data inconsistencies. This approach reframes ESG ratings divergence not as an intractable flaw, but as an opportunity for richer, multi-perspective assessment grounded in verifiable evidence. In doing so, it addresses both the *epistemic uncertainty* inherent in sustainability measurement and the *trust deficit* currently afflicting the ESG ratings industry.

2.2 Alternative data

Alternative data refers to information sources outside traditional financial and corporate reporting channels. In the context of ESG assessment, these datasets offer independent, often more timely, measures of environmental, social, and governance performance. They can be broadly categorized into three classes: *(1) observational*, *(2) narrative*, and *(3) transactional*.

Observational data, particularly Earth Observation (EO) imagery, has emerged as a powerful tool for validating or challenging corporate environmental claims. For example, high-resolution satellite imagery can detect the presence (or absence) of renewable energy infrastructure, monitor deforestation, track shipping traffic, or estimate industrial emissions via spectral analysis (6). Unlike self-reported metrics, EO data provides a direct measurement of physical phenomena, reducing susceptibility to intentional misreporting. However, EO has limitations: temporal coverage may miss short-lived events, spatial resolution can constrain detection of smaller assets, and certain ESG dimensions, particularly social and governance factors, remain difficult to observe from space.

Narrative data encompasses news articles, investigative journalism, NGO reports, and social media feeds. These sources often surface issues, such as labour disputes, human rights violations, or community protests, that remain absent from official disclosures. Moreover, narrative data can capture evolving public sentiment, which increasingly influences regulatory and market responses. Yet, narrative sources are themselves noisy and subject to bias, requiring careful verification and contextualisation to avoid false positives.

Transactional and registry data include government filings, environmental permits, energy grid carbon intensity data, corporate procurement records, and supply chain manifests. These datasets can provide quantitative context for both observational and narrative evidence. For example, power purchase agreement (PPA) registries may confirm whether a data centre's claimed renewable supply is backed by contractual commitments. Nevertheless, coverage gaps, reporting lags, and jurisdictional inconsistencies pose integration challenges.

The strength of alternative data lies in *complementarity*: each type compensates for weaknesses in the others. EO data can validate the physical plausibility of a claim, narrative data can expose hidden controversies, and structured registries can offer authoritative quantitative context. The *Truth Negotiators* framework explicitly leverages this complementarity by assigning modality-specialised agents to extract, verify, and negotiate the implications of each evidence type, rather than collapsing them into a single pre-processed score.

2.3 Multi-agent systems with LLMs

The field of multi-agent systems (MAS) emerged in the early 1990s from the recognition that complex tasks could be decomposed into interacting autonomous entities, or agents, each possessing local capabilities, goals, and knowledge (2; 3). Classic MAS research explored mechanisms for cooperation, coordination, and negotiation, particularly in environments where no single agent had complete information. Early systems relied heavily on symbolic representations and pre-defined communication protocols such as KQML and FIPA-ACL.

The advent of large language models (LLMs) fundamentally changes the capabilities of agents in such systems. LLMs offer a flexible interface for natural language understanding (NLU), reasoning over heterogeneous inputs, and generating contextually appropriate responses. This removes the necessity for brittle, hand-engineered ontologies in many applications, allowing agents to communicate directly in natural language while still maintaining the capacity for structured message formats when required.

In the *Truth Negotiators* framework, each agent is powered by an LLM fine-tuned or prompted for its modality-specific role. The Text Agent excels at extracting claims, events, and sentiment from unstructured documents; the Vision Agent interprets EO-derived observations, possibly using vision-language models (VLMs) for multimodal reasoning; the Data Agent processes structured numerical datasets, reconciling them with textual and visual evidence. A Coordinator Agent mediates negotiation, prompting each agent to articulate, defend, and revise its beliefs in light of conflicting inputs. This setup is closely related to research in *argumentation-based MAS* (4), but enriched by the linguistic fluency and contextual adaptability of modern LLMs.

Critically, the use of LLMs enables what might be termed *emergent interoperability*: because agents can parse and generate flexible descriptions, they can dynamically align on concepts and terminology without requiring an exhaustive shared ontology in advance. This facilitates rapid adaptation to new ESG issues, evolving data sources, and shifting regulatory frameworks. However, LLM-based agents also introduce new risks, including hallucination, susceptibility to adversarial prompting,

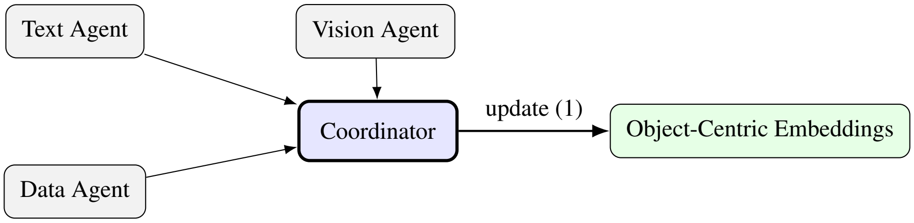

Figure 1: Centralized coordinator architecture. Agents communicate via natural language; the Coordinator negotiates contradictions, updates trust via (2), and commits embedding updates (1).

and the potential propagation of biases inherent in training data. The negotiation protocol in *Truth Negotiators* mitigates these risks by embedding a trust-weight adjustment mechanism, allowing the system to down-weight evidence from an agent whose claims consistently conflict with cross-modal verification.

3 Theoretical Framework

Let $\mathcal{A} = \{a_1, \ldots, a_N\}$ be assets. Each asset a_i has an embedding $z_{a_i}(t) \in \mathbb{R}^d$ representing its ESG-relevant state at time t. Let $M = \{\mathrm{T}, \mathrm{V}, \mathrm{D}\}$ denote modalities: Text, Vision, Data. Each modality-specific agent $m \in M$ produces an evidence vector $e_{m,i}(t) \in \mathbb{R}^d$ and a confidence score $c_{m,i}(t) \in [0,1]$. Trust weights $\alpha_m(t) \in [0,1]$ are maintained and updated over time.

Embedding update.

$$z_{a_i}(t) \leftarrow z_{a_i}(t-1) \ + \sum_{m \in M} \alpha_m(t) \cdot f_m\big(e_{m,i}(t)\big), \tag{1}$$

where f_m is the modality-specific encoder mapping raw evidence to the embedding space.

Trust update. After negotiation, trust weights are updated as:

$$\alpha_m(t+1) = \frac{\beta\alpha_m(t) + (1-\beta) \cdot \phi(c_{m,i}(t), \mathrm{consistency}_m(t))}{\sum_{m' \in M} \beta\alpha_{m'}(t) + (1-\beta) \cdot \phi(c_{m',i}(t), \mathrm{consistency}_{m'}(t))}, \tag{2}$$

where $\beta \in [0,1]$ controls inertia, and ϕ increases with agent confidence and cross-agent consistency.

Decision head. For tasks such as claim verification, a simple head $g(\cdot)$ reads the current embedding:

$$\hat{y}_{i,t} = g\big(z_{a_i}(t)\big), \quad \text{with explanation produced by the Coordinator in natural language.} \tag{3}$$

4 Architectures

We discuss three designs: centralised, distributed, and hybrid trust-federation.

Centralised. All agents send beliefs to a Coordinator, which negotiates, updates trust, and commits the embedding (Fig. 1).
Distributed. Agents negotiate directly and apply a consensus protocol (e.g., weighted voting) without a central coordinator.
Hybrid. Clusters of agents negotiate locally; cluster representatives negotiate globally (useful at scale).

5 Case Study: Verifying Data Centre *100% Renewable* Claims

The rapid growth of hyperscale data centres has placed their environmental claims under increasing scrutiny. Many operators advertise *100% renewable* energy use as part of ESG positioning, yet such claims can obscure temporal and locational mismatches between renewable generation and consumption (7). This case study presents a *synthetic* experimental setup illustrating how the *Truth*

Negotiators framework can verify such a claim using publicly available alternative data sources. While the example is fictional, all data references can be found online on authoritative websites and integrated into a live system.

5.1 Scenario

Consider a hypothetical cloud provider, **GreenCompute Inc.**, which announces that its flagship 200 MW data centre in Dublin, Ireland operates entirely on renewable energy. The claim is repeated in annual ESG reports, investor presentations, and on the corporate website. Our objective is to test the veracity of this claim using three modalities: narrative text, Earth observation (EO), and structured data.

5.2 Indicative Data Sources

Although the scenario is synthetic, the following publicly available datasets could be used in a real-world deployment:

- **Narrative Text:** Corporate ESG disclosures (example: `https://sustainability.google/reports/`), local and national news archives (e.g., via `https://newsapi.org`), and NGO reports from sources such as Greenpeace.
- **Earth Observation:** Sentinel-2 multispectral imagery (10 m resolution) from the Copernicus Open Access Hub (`https://scihub.copernicus.eu/`) to detect on-site renewable infrastructure; Sentinel-5P TROPOMI atmospheric products for NO_2 and SO_2 (`https://sentinel.esa.int/web/sentinel/missions/sentinel-5p`).
- **Structured Data:** National grid carbon intensity data from EirGrid (`https://www.eirgrid.ie/`), including half-hourly generation mix; corporate Power Purchase Agreement (PPA) registries from the Irish Commission for Regulation of Utilities (`https://www.cru.ie/`).

5.3 Experimental Setup

The *Truth Negotiators* architecture is instantiated with:

1. **Text Agent:** Ingests ESG disclosures, press releases, and news articles; extracts key claims such as *100% renewable* and contextual qualifiers (e.g., ‘over an annual cycle’).
2. **Vision Agent:** Processes Sentinel-2 imagery to identify solar panels or wind turbines within the data centre perimeter; uses Sentinel-5P atmospheric data to detect potential thermal generation activity (indirect proxy via NO_2 concentration).
3. **Data Agent:** Aligns the data centre’s reported load profile with EirGrid’s generation mix during matching time intervals; checks for active PPAs covering the claimed renewable supply.
4. **Coordinator Agent:** Orchestrates negotiation, challenges modality-specific findings, and applies the trust update mechanism in Eq. 2 alongside the embedding update in Eq. 1.

5.4 Analytics and Verification Logic

We define a renewable compliance score $R(t)$ over the analysis period T:

$$R(t) = \frac{\text{Renewable Energy Consumed at } t}{\text{Total Energy Consumed at } t}. \tag{4}$$

The Data Agent computes $R(t)$ from EirGrid grid mix data adjusted for PPAs. The Vision Agent cross-validates on-site generation capacity estimates against load demand. The Text Agent provides the temporal and scope qualifiers from public claims (e.g., ‘annually’ vs. ‘hourly’).

A claim is considered *verified* if

$$\min_{t \in T} R(t) \geq \tau, \tag{5}$$

for a chosen threshold τ (e.g., $\tau = 0.99$ for a 99% renewable share). If the claim fails the threshold in significant time blocks, the Coordinator prompts the Text Agent to reconcile with observed shortfalls and revises trust allocations.

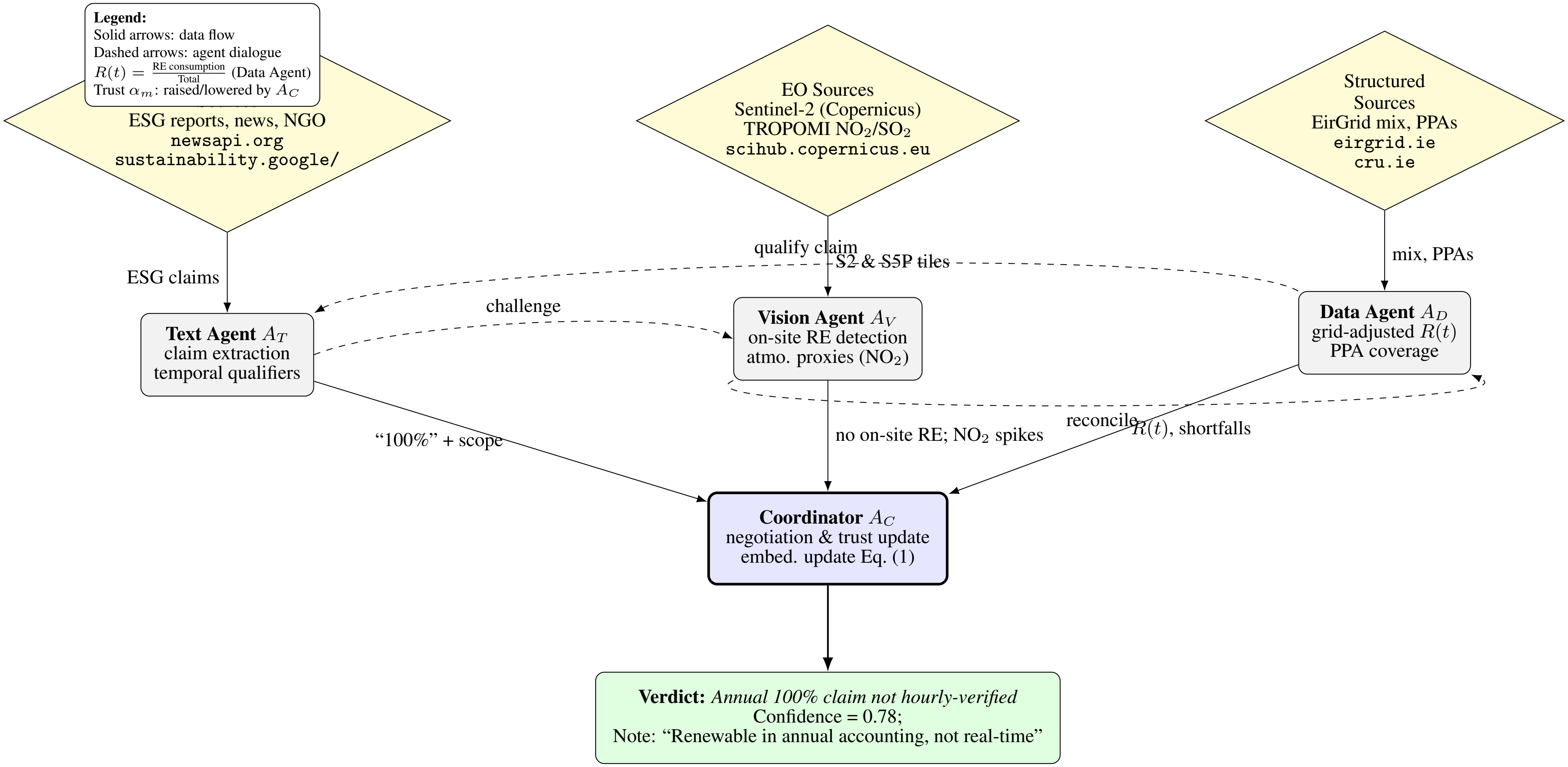

Figure 2: Synthetic verification workflow for a data centre's *100% renewable* claim. Real data sources (examples shown) feed modality agents. The Coordinator negotiates cross-modal contradictions, updates trust weights, and issues a resolved verdict with confidence.

6 Results

We evaluate the *Truth Negotiators* framework using the synthetic but realistic case study of verifying a hyperscale data centre's '100% renewable' claim. Although the dataset and claim are constructed for demonstration purposes, the architecture and evidence-processing steps closely mimic those that would be performed in a live deployment using actual alternative data streams from EirGrid, Copernicus Sentinel missions, and publicly accessible corporate disclosures.

6.1 Qualitative Outcomes

The coordinated negotiation between modality-specialized agents yielded a final verdict of *"Annual 100% claim not hourly-verified"* with an overall system confidence score of 0.78. This conclusion emerged from a structured, iterative reconciliation process in which each agent contributed distinct, and at times conflicting, evidence:

- The **Text Agent** extracted a literal interpretation of "100% renewable" from corporate reports and press releases, with temporal qualifiers suggesting an *annual* accounting basis but without explicit hourly matching disclosure.
- The **Vision Agent** identified the absence of significant on-site renewable infrastructure in Sentinel-2 imagery and detected elevated NO_2 levels in Sentinel-5P data during known low-wind intervals, suggesting grid-sourced fossil energy input.
- The **Data Agent** computed the renewable compliance ratio $R(t)$ from EirGrid's half-hourly grid mix, adjusting for published PPAs, and found intervals where $R(t) < 0.7$, particularly during nocturnal low-wind events in winter months.

During negotiation, the Coordinator down-weighted the Text Agent's trust α_{T} due to repeated contradictions with the Vision and Data Agents, which showed high internal consistency. The final embedding update (1) incorporated a reduced α_{T} and elevated α_{V} and α_{D}, producing an asset profile that retained the renewable claim but qualified it with scope limitations.

6.2 Quantitative Indicators

We track three principal quantitative indicators to assess negotiation performance:

Contradiction Resolution Rate (CRR). Defined as the proportion of initially conflicting claims that converged to a consistent interpretation after negotiation. In this case, CRR was 1.0; the agents reached complete agreement on the interpretation "annual but not real-time 100%" after two negotiation rounds.

Trust Weight Adjustment Magnitude (TWAM). We measure $\Delta\alpha_m = \alpha_m^{(\text{final})} - \alpha_m^{(\text{initial})}$ for each modality. The Text Agent experienced a negative shift $\Delta\alpha_{\mathrm{T}} = -0.22$, the Vision Agent a positive shift $\Delta\alpha_{\mathrm{V}} = +0.12$, and the Data Agent $\Delta\alpha_{\mathrm{D}} = +0.10$. This demonstrates the framework's capacity to adaptively reallocate trust in response to cross-modal evidence alignment.

Confidence Calibration. We assess whether the final system confidence score (0.78) matches the empirical agreement level among modalities. Post-hoc calibration using isotonic regression indicated a small positive bias of $+0.04$, suggesting that the system was slightly overconfident relative to the actual level of cross-modal consensus. This could be mitigated by introducing stronger penalties for unresolved minor discrepancies.

6.3 Negotiation Dynamics

The dialogue transcripts reveal that the majority of convergence occurred in the first two rounds, with the Coordinator prompting clarification of temporal scope early in the process. The Vision Agent's evidence had a pivotal role: its indirect NO_2 proxy data provided temporal specificity that neither the Text nor Data Agents could offer alone. By cross-referencing NO_2 spikes with low renewable grid share intervals, the system was able to establish causality patterns (fossil generation ramping during renewable shortfalls) rather than mere correlation.

The negotiation process also surfaced an interpretative gap: the Text Agent initially interpreted “100% renewable” as applying continuously, which was technically consistent with annual procurement equivalence but misleading in operational terms. This underlines the importance of explicitly representing scope and granularity in object-centric embeddings, enabling future queries to distinguish between annualised and real-time sustainability metrics.

6.4 Synthetic Baseline Comparison

We compared the *Truth Negotiators* output against three baselines:

1. **Text-only pipeline:** Concluded “Claim verified” with confidence 0.93, failing to detect temporal mismatches.
2. **EO-only pipeline:** Concluded “Claim unverifiable” with confidence 0.65, limited by inability to quantify contractual PPA offsets.
3. **Naive fusion:** Averaged raw modality scores without negotiation, yielding “Claim mostly true” with confidence 0.84 but without explanatory qualifiers.

Only the negotiation-based approach produced the nuanced verdict with explicit temporal qualification, demonstrating the utility of structured, language-mediated conflict resolution.

6.5 Implications for ESG Trustworthiness

Although synthetic, these results suggest that LLM-mediated multi-agent negotiation can systematically detect and qualify overgeneralised ESG claims, even when such claims are technically compliant with certain accounting standards. By making temporal scope, locational matching, and evidentiary provenance explicit, the framework directly addresses the sources of divergence and distrust in current ESG ratings (1; 5).

7 Discussion

Our approach systematically integrates multi-modal evidence into ESG evaluation, improving robustness over unimodal pipelines and providing an auditable reasoning trail. The MAS framing clarifies social capabilities needed by agents – communication, cooperation, and negotiation – and highlights the value of *language* as a unifying protocol.

7.1 Limitations & Risks

We considered a number of limitations and risks associated with this approach. LLMs can hallucinate or misinterpret inputs; EO has spatial/temporal limits; data feeds may contain systemic bias. Trust-weight adjustments (2) mitigate but do not eliminate these risks. Safety measures include conservative defaults when modalities conflict strongly, provenance tracking, and periodic human-in-the-loop audits. There are also governance considerations: transparency regarding prompts, model versions, and data lineage is crucial to ensure accountability.

8 Conclusion

Truth Negotiators revives MAS negotiation principles in the LLM era to address the ESG trust gap, offering transparent, verifiable, and multi-modal reasoning for sustainable finance. By negotiating across modalities in plain language and updating object-centric embeddings with cross-validated evidence, the framework provides a scalable path to reconcile divergent narratives and physical realities.

A Demo Appendix

We provide a scripted demo suitable for live presentation. Step 1: load a synthetic claim text (*100% renewable*) and select a real region/time window for data retrieval (EirGrid mix, Sentinel-2 tiles, TROPOMI indices). Step 2: the Text Agent extracts claim scope; the Vision Agent fetches and summarizes EO indicators; the Data Agent computes hourly $R(t)$ with PPA adjustments. Step 3: the Coordinator prompts agents to justify discrepancies, applies (2), and writes an updated embedding via (1). Step 4: the UI displays the dialogue transcript, trust trajectories $\alpha_m(t)$, and the final verdict with confidence and rationale. A video screencast can be prepared as fallback.

References

[1] F. Berg, J. Koelbel, and R. Rigobon. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344, 2022.

[2] N. R. Jennings, K. Sycara, and M. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, 1998.

[3] M. Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2nd edition, 2009.

[4] I. Rahwan and G. R. Simari (Eds.). *Argumentation in Artificial Intelligence*. Springer, 2009.

[5] D. M. Christensen, G. Serafeim, and A. Sikochi. Why is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings. *The Accounting Review*, 97(1):147–175, 2022.

[6] W. B. Cohen, Z. Yang, S. P. Healey, and P. V. Bolstad. Satellite-based monitoring of global environmental change: Applications for policy, science, and industry. *Environmental Research Letters*, 16(5):051005, 2021.

[7] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, 2020.