Actors, Frames and Arguments: A Multi-Decade Computational Analysis of Climate Discourse in Financial News using Large Language Models

Ruiran Su[1] **Janet B. Pierrehumbert**[1] **Markus Leippold**[2]

[1]Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[2]Department of Finance, University of Zurich, Zurich, Switzerland

{ruiran.su, janet.pierrehumbert}@eng.ox.ac.uk, markus.leippold@df.uzh.ch

Abstract

Financial news media shapes trillion-dollar climate investment decisions, yet discourse in this elite domain remains underexplored. We analyze two decades of climate-related articles (2000–2023) from Dow Jones Newswire using an Actor–Frame–Argument (AFA) pipeline that extracts who speaks, how issues are framed, and which arguments are deployed. We validate extractions against 2,000 human-annotated articles using a Decompositional Verification Framework that evaluates completeness, faithfulness, coherence, and relevance. Our longitudinal analysis uncovers a structural transformation: pre-2015 coverage emphasized risk and regulatory burden; post-Paris Agreement, discourse shifted toward economic opportunity and innovation, with financial institutions becoming dominant voices. Methodologically, we provide a replicable paradigm for longitudinal media analysis with LLMs; substantively, we reveal how financial elites have internalized and reframed the climate crisis across two decades.

1 Introduction

Financial news media serves as the nervous system of the global economy, informing not only investors but also shaping corporate strategy and capital allocation decisions worth trillions of dollars (Shiller, 2017). How this elite domain portrays climate change is therefore consequential, influencing whether climate risks are framed as costs to be minimized or opportunities to be seized. While climate communication in general news has been extensively studied (Schmid-Petri et al., 2017a), the high-stakes arena of financial news remains underexplored, despite its significant role in shaping market expectations and influencing policy debates.

This paper addresses this gap through a longitudinal computational analysis of the *Dow Jones Climate News Corpus*, a collection of 980,061 climate-related articles (2000–2023). We focus our extraction on a stratified and uncertainty-enriched sample of 4,143 articles that preserves the temporal and thematic diversity of the full dataset. Moving beyond topic models and dictionary-based methods, we introduce an **Actor–Frame–Argument (AFA)** methodology that leverages Large Language Models (LLMs) to identify who speaks, how climate issues are framed, and which arguments are advanced. To ensure reliability across two decades of discourse, we validate our pipeline against a 2,000-article human-annotated gold standard and introduce a **Decompositional Verification Framework (DVF)** that evaluates extractions along completeness, faithfulness, coherence, and climate relevance with multi-judge validation.

We organize the study around three guiding research questions:

- **RQ1 (Actors):** Which types of financial and policy actors dominate climate discourse in elite financial news, and how has their prominence evolved over time?
- **RQ2 (Frames):** How have climate change frames evolved between 2000–2023, and what external events coincided with major shifts in framing?
- **RQ3 (Arguments):** What argumentative strategies and warrants are deployed to justify climate positions, and how do these differ across actor groups and eras?

Our contributions are threefold. First, we develop and validate an LLM-based AFA pipeline for longitudinal extraction of actors, frames, and arguments, anchored by DVF to ensure robustness and auditability. Second, we provide the most comprehensive map to date of climate framing in elite financial news, documenting systematic differences across time, actor groups, and argumentative repertoires. Third, we uncover a structural transformation in financial climate narratives: from a pre-2010

Findings of the Association for Computational Linguistics: EACL 2026, pages 1994–2014
March 24-29, 2026 ©2026 Association for Computational Linguistics

emphasis on *costs and compliance* to a post-Paris Agreement framing of *opportunity and innovation*, linked to external shocks such as the Paris Agreement and subsequent policy shifts.

By combining methodological innovation with substantive insight, this work establishes a replicable paradigm for large-scale discourse analysis.

2 Related Work

2.1 Argument Mining and Discourse Analysis

Argument mining seeks to recover argumentative structure, including claims, evidence, and warrants from text (Lawrence and Reed, 2019). Early work established end-to-end pipelines for component detection and relation parsing, often with global inference (Stab and Gurevych, 2017a). Subsequent neural approaches improved robustness by modeling non-tree structures and joint decoding of components and links (Liu et al., 2021). Parallel strands in discourse parsing provide complementary representations of coherence that have been leveraged for argument quality and relation modeling (Liu et al., 2020; Prasad et al., 2019).

Other research has expanded task coverage (stance, evidence retrieval, relation typing, and quality) and domains, utilizing larger and more diverse datasets (Hua et al., 2022; Chen et al., 2021; Gleize et al., 2020). Transformer baselines remain strong for span/link prediction (the task of identifying argument component boundaries and predicting the relationships between them), while graph-based architectures and discourse-informed models help capture long-range structure and argumentative coherence (Ruiz-Dolz et al., 2021; Toledo et al., 2021).

Most closely related to our aims are works that examine argument structure at scale and across time. However, existing studies typically operate on short-horizon corpora (essays, forums, debates) and evaluate with intrinsic metrics alone. Emerging evidence shows that properly prompted LLMs can match or outperform task-specific models for argument mining, but concerns remain about faithfulness, bias, and reproducibility (Gorur et al., 2025; Li et al., 2025).

2.2 Media Framing and Climate Communication

Framing theory emphasizes that how issues are presented strongly conditions public interpretation and policy response (Entman, 1993). Early studies relied on manual content analysis of small corpora, often in the U.S. or Europe, highlighting political polarization and media biases in the portrayal of climate change (Nisbet, 2009; Schmid-Petri et al., 2017b; Boykoff and Boykoff, 2004). More recent work has extended this line to multi-country comparisons, showing persistent divergences between U.S., European, and Chinese media in how climate debates are framed (Schmid-Petri and Arlt, 2016a; Duan and Miller, 2021).

Computational approaches have sought to scale framing analysis. Dictionary-based methods (Card et al., 2015) and topic models (Blei et al., 2003; Roberts et al., 2014) have been applied to climate corpora to identify high-level themes such as risk, adaptation, and responsibility (Schirmag et al., 2025).

However, such approaches often lack granularity. While topic modeling identifies *what* is being discussed, framing analysis reveals *how* it is discussed, highlighting specific aspects of reality to promote a particular interpretation. Recent advances in computational framing emphasize the need for models that capture not just what frames are invoked but also the argumentative strategies that support them (Demszky et al., 2019). Crucially, (Hofmann et al., 2022) demonstrate that LLMs can effectively capture dynamic shifts in framing and ideology over time, particularly regarding moral dimensions, which is a perspective that directly parallels our focus on the temporal evolution of responsibility and opportunity in financial discourse.

2.3 LLMs for Climate and Financial Discourse

Recent work has begun to apply large language models to climate and finance text, yielding methods and findings directly relevant to our Actor-Frame-Argument (AFA) pipeline. For example, (Leippold, 2023) explored structured interviews of GPT-3 on climate finance narratives, while (Jain et al., 2024) demonstrate LLM-based tools for climate-aware investment decision support. Domain-specialized models such as ClimateBERT (Webersinke et al., 2022) and SusGen-GPT (Wu et al., 2024) provide useful embeddings and generation priors for climate and sustainability text. Other studies used LLMs to estimate public opinion and targeting effects around climate topics (Lee et al., 2024; Islam and Goldwasser, 2024), and to analyze bias and representational differences in LLM outputs on sociopolitical issues, including climate

(Chen et al., 2023).

2.4 Our Contributions

Our work advances these three research strands in complementary ways. Relative to argument mining (§2.1), we target a multi-decade financial news corpus enabling longitudinal analysis, decompose extraction into an integrated Actor–Frame–Argument framework jointly analyzing who speaks, how issues are framed, and which arguments are deployed, and pair LLM extraction with a Decompositional Verification Framework auditing completeness, faithfulness, coherence, and relevance. Relative to framing research (§2.2), we focus on financial news, a high-stakes domain where framing influences capital allocation, and analyze actor-specific rhetorical strategies over time. While narrative framing models capture story-level conflicts and resolutions, our approach emphasizes the economic and regulatory dimensions distinctive to financial discourse. Relative to LLM applications (§2.3), we validate extraction through multi-judge verification against human annotations, emphasizing reproducibility and faithfulness over predictive performance alone.

3 The Dow Jones Climate News Corpus

3.1 Corpus Construction

We construct the Dow Jones Climate News Corpus in two steps. First, we filter articles in the Dow Jones Newswire (2000–2023) using Dow Jones Intelligent Identifiers (DJIDs), a proprietary subject taxonomy that consistently categorizes financial news (Dow Jones, 2024; ProQuest, 2024). We retain articles tagged with climate-related DJIDs spanning three domains: (1) *Core Climate Issues* (e.g.,N/CO2 [Carbon Dioxide]) (2) *Energy Transition* (e.g., N/BFL [Biofuels]), and (3) *Climate-Affected Sectors* (e.g., N/AGR [Agriculture]). A full list of DJIDs and their descriptions is provided in Appendix A.1.

Second, to quantify potential selection bias introduced by reliance on DJID codes, we conduct a validation study combining keyword-based re-screening and supervised classification. A summary of results indicates that DJID filtering achieves high precision (∼95%) and recall (∼92%), suggesting that our climate corpus is both accurate and broadly representative. Full methodological details are provided in Appendix A.2.

Preprocessing involved standard normalization and near-duplicate removal. Using locality-sensitive hashing, we removed ∼8.7% of articles with high overlap, yielding **980,061 unique climate-related articles** with full metadata and timestamps for diachronic analysis (Figure 4 in Appendix A.5 shows the temporal distribution of the Dow Jones Climate News Corpus, annotated with major climate policy events). Full preprocessing details are provided in Appendix A.4.

3.2 Sampling Methodology

To construct a tractable yet representative subset of our climate corpus, we implemented a four-stage hierarchical sampling procedure. This design ensures that the 4,143-article sample retains the temporal and thematic diversity of the full 980k-article corpus, while also enriching for complex argumentative texts.

3.2.1 Temporal Stratification

We stratified the corpus into four periods reflecting major phases in climate finance discourse: Pre-crisis (2000–2007), Financial Crisis (2008–2012), Post-crisis Recovery (2013–2018), and Climate Finance Surge (2019–2023). Within each stratum t, the sample size was determined by proportional allocation:

$$n_t = \left\lfloor n_{\text{total}} \times \frac{|C_t|}{|C|} \right\rfloor,$$

where $|C_t|$ is the size of stratum t and $|C|$ is the size of the full corpus.

3.2.2 Thematic Clustering

To preserve thematic coverage, each article was represented by a 2:1 weighted concatenation of its headline and first paragraph embeddings, computed using the SBERT `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). We applied agglomerative clustering with Ward linkage (Ward, 1963), selecting the optimal number of clusters k_t by maximizing the silhouette score (Rousseeuw, 1987) within the range $[5, \min(15, \lfloor n_t/20 \rfloor)]$. Pilot studies indicated that $k < 5$ conflated distinct themes, while $k > 15$ produced redundant micro-clusters.

3.2.3 Representative Article Selection

From each cluster, we applied Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to select representative articles that balance central-

Temporal Stratum	Original	Sample	%
Pre-crisis (2000–2007)	5,441	155	0.8
Financial crisis (2008–2012)	320,157	1,355	0.8
Post-crisis (2013–2018)	215,351	911	0.8
Climate surge (2019–2023)	439,112	1,722	0.8
Total	**980,061**	**4,143**	**0.8**

Table 1: Distribution of articles across temporal strata in original corpus and strategic sample.

ity and diversity:

$$\mathrm{MMR}(a_i) = \lambda \cdot \mathrm{sim}_c(a_i) - (1-\lambda) \cdot \max_{a_j \in S} \mathrm{sim}_s(a_i, a_j).$$

We set $\lambda = 0.7$, slightly favoring centrality. A sensitivity analysis across $\lambda \in [0.5, 0.9]$ showed stable downstream performance (Appendix B).

3.2.4 Active Learning Enrichment

To enrich the sample with harder argumentative cases, we incorporated an active learning loop. A preliminary RoBERTa-base argument detector (F1 = 79.5% on a seed set) was applied to the wider corpus (Appendix C). Articles with prediction entropy above the 90th percentile were prioritized. Iterative sampling converged after four rounds, measured by Jensen–Shannon divergence of entropy distributions (<0.05 between iterations).

3.2.5 Final Sample

The resulting 4,143-article sample achieves high fidelity to the original corpus: Jensen–Shannon divergence <0.1 across temporal and thematic distributions, cosine similarity >0.85 between sample and population centroids, and minimal degradation (-3.3 points) in downstream argument extraction performance compared to a 20k random sample.

4 Actor-Frame-Argument (AFA) Extraction Pipeline

We design a modular pipeline for extracting Actor–Frame–Argument (AFA) structures from financial news. The goal is to represent who speaks, how climate change is framed, and what argumentative strategies are deployed, in a form that is scalable and reproducible. The pipeline is sequential; each stage conditions on the previous one, ensuring coherence across actors, frames, and claims. Figure 1illustrates the pipeline architecture.

Actor–Stance Identification. The first stage identifies actors and their expressed stances on climate issues. Each extraction specifies the actor *name*, *type* (*company*, *financial institution*, *government*, *NGO/advocacy*, *individual*), and expressed *stance*. , and *stance* (*supportive*, *opposing*, *neutral*, or *mixed*). For attributional fidelity, every extraction is linked to a direct *evidence* quote from the article.

Frame Classification. The second stage assigns one *primary frame* and an optional *secondary frame* from our eight-frame typology.

The eight-frame typology adapts established frameworks from climate communication research (Nisbet, 2009; Schmid-Petri and Arlt, 2016b) to financial discourse. Key adaptations include: (1) splitting generic "Economic Development" into *Economic Risk* vs. *Economic Opportunity* to capture the critical distinction in financial contexts, (2) adding *Market Dynamics* to capture competitive positioning unique to financial news, and (3) refining "Morality/Ethics" to *Social Responsibility* reflecting corporate framing conventions. This typology emerged through pilot coding of 500 articles. Full mapping to source frameworks is in Appendix E.

Argument Extraction. Finally, the model decomposes the article's argumentation into its components: a central *claim*, supporting *evidence*, and a *warrant* connecting claim and evidence.

4.1 Models and Prompting

While domain-specific encoders such as ClimateBERT (Webersinke et al., 2022) or data-centric models like SusGen-GPT (Wu et al., 2024) offer attractive pretraining priors for climate text, our focus is on extraction robustness and longitudinal generalization across heterogeneous financial-news styles.

We therefore adopt a closed–open pairing strategy for extraction. The primary extractor is Gemini-2.5-flash (Comanici et al., 2025), which offers strong instruction-following reliability and efficiency. To ensure that findings are not from a single proprietary model, we benchmark the same prompts on a diverse open-weight model LLaMA-4 Maverick-17B (MetaAI, 2025). This pairing balances performance (closed-sourced model) with reproducibility and transparency (open-sourced model), following best practices in recent work on LLM-as-annotator pipelines (Tan et al., 2024; Törnberg, 2024; Pavlovic and Poesio, 2024).

While multiple open-weight models exist, we focus on one representative baseline in the main analysis for clarity. All prompts use a structured chain-

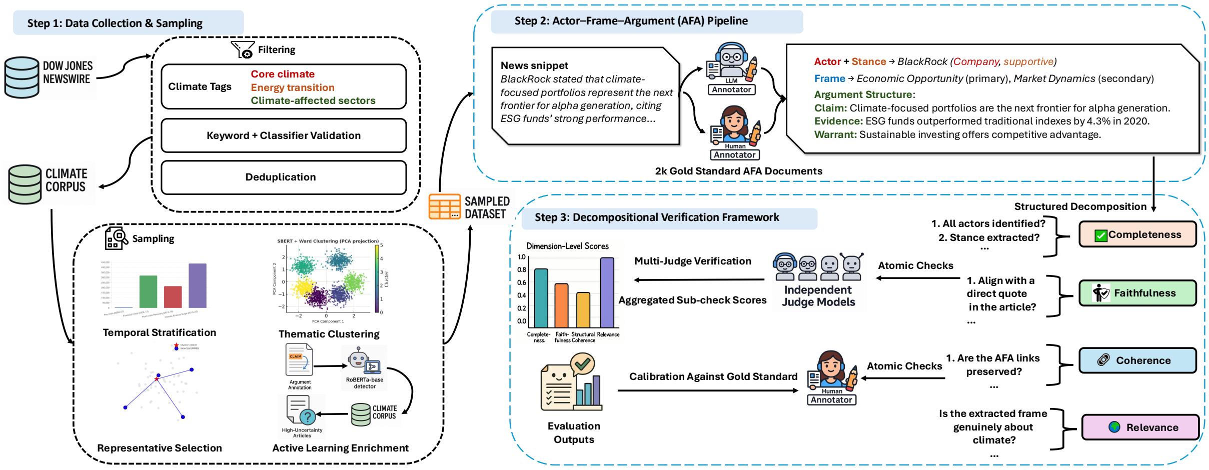

Step 1: Data Collection & Sampling
DOW JONES NEWSWIRE
CLIMATE CORPUS
Filtering
Climate Tags
Core climate
Energy transition
Climate-affected sectors
Keyword + Classifier Validation
Deduplication
SAMPLED DATASET
Sampling
Temporal Stratification
SBERT + Ward Clustering (PCA projection)
PCA Component 1
PCA Component 2
Cluster
Thematic Clustering
Cluster center
Selected (MMR)
Representative Selection
CLAIM
Argument Annotation
RoBERTa-base detector
CLIMATE CORPUS
High-Uncertainty Articles
Active Learning Enrichment
Step 2: Actor–Frame–Argument (AFA) Pipeline
News snippet
BlackRock stated that climate-focused portfolios represent the next frontier for alpha generation, citing ESG funds' strong performance...
LLM Annotator
Human Annotator
Actor + Stance → BlackRock (Company, supportive)
Frame → Economic Opportunity (primary), Market Dynamics (secondary)
Argument Structure:
Claim: Climate-focused portfolios are the next frontier for alpha generation.
Evidence: ESG funds outperformed traditional indexes by 4.3% in 2020.
Warrant: Sustainable investing offers competitive advantage.
2k Gold Standard AFA Documents
Step 3: Decompositional Verification Framework
Structured Decomposition
1. All actors identified?
2. Stance extracted?
...
Completeness
Dimension-Level Scores
1.0
0.8
0.6
0.4
0.2
0.0
Complete-ness.
Faith-fulness
Structural Coherence
Relevance
Multi-Judge Verification
Aggregated Sub-check Scores
Independent Judge Models
Atomic Checks
1. Align with a direct quote in the article?
...
Faithfulness
Evaluation Outputs
Calibration Against Gold Standard
Human Annotator
Atomic Checks
1. Are the AFA links preserved?
...
Coherence
Is the extracted frame genuinely about climate?
...
Relevance

Figure 1: End-to-end pipeline of our study. Step 1: Data collection and sampling from the Dow Jones Climate Corpus, including stratification, clustering, and active learning enrichment. Step 2: Actor–Frame–Argument (AFA) extraction pipeline, identifying actors, frames, and argument structures. Step 3: Decompositional verification framework, combining multi-judge evaluation with validation against a human-annotated gold standard.

of-thought design and are released Appendix F for reproducibility.

4.2 Human-Annotated Gold Standard

To establish a reliable evaluation anchor, we constructed a 2,000-article gold-standard set drawn from our corpus. Each annotation unit is a complete article (mean=1,087 tokens, median=982 tokens). The "news snippet" in Figure 1 is illustrative; actual processing uses full articles. Articles were annotated on the Zooniverse citizen-science platform (see Appendix G), with each item receiving at least five independent annotations. For categorical tasks (stance, frame), final labels were obtained by majority vote. For span-based tasks (actors, arguments), we applied a two-step consensus procedure: token-level agreement scores were computed across annotators, tokens with majority support were retained, and contiguous majority-agreed tokens were merged into unified spans (see Appendix H.6). This aggregation preserves boundary precision for short mentions while preventing over-fragmentation in longer argument spans.

We computed inter-annotator agreement (IAA) using metrics appropriate to multi-annotator settings: pairwise F_1 for actor spans (Stab and Gurevych, 2017b), Krippendorff's α (Krippendorff, 2004) for stance, Krippendorff's α for frames, and span-level F_1 for argument claim extraction (Stab and Gurevych, 2014; Habernal et al., 2018). Table 2 summarizes observed IAA scores, all of which fall in the "substantial agreement" range ($>$ 0.7) (Landis and Koch, 1977), confirming the reliability of the citizen-science annotation protocol.

Component	Metric	IAA
Actor Identification	Pairwise F1	0.81
Stance Classification	Krippendorff's α	0.76
Frame Assignment	Krippendorff's α	0.72
Claim Extraction	Span-level F1 (macro)	0.74

Table 2: Inter-annotator agreement across AFA components under the Zooniverse citizen-science protocol.

4.3 Decompositional Verification Framework

To ensure reliable and auditable evaluation, we introduce a **Decompositional Verification Framework (DVF)** that pushes beyond single holistic scores. Instead of asking judges to rate an extraction globally, DVF decomposes the evaluation into fine-grained sub-checks, verifies them across multiple model families, and anchors them to human gold annotations.

Structured Decomposition. DVF evaluates extractions along four dimensions: (1) *Completeness* (all components captured), (2) *Faithfulness* (alignment with source text), (3) *Coherence* (schema and actor–frame–argument links), and (4) *Relevance* (domain specificity). Full sub-checks are provided in Appendix I.2.

Multi-Judge Verification. To mitigate self-grading bias, DVF employs a diverse set of judges: GPT-4o (OpenAI, 2024), Claude-Sonnet-4 (Anthropic, 2024), and two open-weight evaluators (Qwen3-30B A3B (Qwen, 2025), Mixtral-8×22B

(Mistral, 2024)). Judges provide sub-check ratings, which are aggregated into dimension-level scores. To anchor these automated scores, we created a human evaluation set of 500 randomly sampled LLM outputs. These outputs were annotated under the DVF rubric by coders distinct from the AFA gold-standard annotators. This separate evaluation set enables validation of automated DVF scores, ensuring that judge reliability is grounded in human assessments while avoiding circularity.

5 Results and Discussion

We present our findings in four parts. First, we validate the reliability of the AFA extraction pipeline (§ 5.1). Second, we analyze longitudinal trends in actor prominence using our five-category schema (§ 5.2, RQ1). Third, we document the structural transformation in framing strategies across two decades (§ 5.3, RQ2). Finally, we examine how actors combine frames with argumentative strategies, revealing distinct rhetorical profiles (§ 5.4, RQ3). A qualitative error analysis of 150 sampled mispredictions is provided in Appendix J.4.

5.1 Pipeline Validation

5.1.1 Justification for LLM-Based Extraction

A critical question is whether LLMs' computational cost is justified when fine-tuned Pre-trained Language Models (PLMs) could achieve similar performance on specific tasks. We test two hypotheses: (H1) *LLMs provide superior performance on complex, multi-class tasks* , and (H2) *sequential dependencies through in-context learning will improve the performance.*

Experimental Setup We partitioned our 2,000-article gold standard into 1,400 train / 300 validation / 300 test splits. We compared three RoBERTa-Large (Zhuang et al., 2021) baselines against our zero-shot LLM pipeline:

Independent RoBERTa: Each AFA component trained as an isolated model receiving only raw article text: (1) token-level NER for actors, (2) sequence classifier for stance, (3) multi-label classifier for frames, and (4) span extractor for arguments.

Sequential RoBERTa (Gold): Downstream components receive gold-standard upstream annotations as additional features (e.g., frame classifier receives gold actors concatenated with article text).

All RoBERTa models used identical hyperparameters. Training details are provided in Appendix K.

Model	Actor	Stance	Frame	Arg.
Independent RoBERTa	.762	.718	.504	.563
+ Gold upstream	.762	.742	.631	.595
Gemini-2.5	**.819**	**.791**	**.783**	**.767**
Δ (LLM – Best RoBERTa)	+5.7	+4.9	+15.2	+17.2

Table 3: Performance comparison (macro F1) on 300-article test set.

Results Table 3 shows LLMs outperform fine-tuned RoBERTa-Large across all components, with largest gaps on frame classification (+15.2 F1) and argument extraction (+17.2 F1), which supports H1.

For example, RoBERTa classifies "renewable energy fund outperformed traditional portfolios by 4.3%" as *Technological Solution* (triggered by "renewable energy"), while the correct frame is *Economic Opportunity* (centered on financial returns). LLMs correctly identify the argumentative focus through longer-range contextual reasoning.

RoBERTa also struggles with long-range dependencies in complex argumentative structures. We found out that it either truncates claims prematurely at clause boundaries or over-extends into supporting evidence, suggesting insufficient training data for learning nuanced argument boundaries.

Sequential Dependencies and Error Propagation Table 3 reveals the critical finding supporting H2: *Sequential conditioning is beneficial*: Gold upstream features improve RoBERTa performance substantially (+12.7 for frames, +3.2 for arguments), confirming that AFA components are not independent, frames depend on actors, and arguments depend on both. But independent RoBERTa classifiers can not naturally achieve that.

Implications To summarize, for our use case (multi-decade, multi-component extraction with limited human annotations), LLMs provide superior cost-performance trade-offs. They excel at: (1) complex multi-class tasks requiring semantic distinction and (2) sequential reasoning through in-context learning rather than brittle feature engineering.

5.1.2 LLM-based Performance

To establish the reliability of our LLM-based extraction pipeline, we evaluate model outputs

Component	Gemini-2.5	LLaMA-4
Actor–Stance Identification		
Actor Type	**0.847**	0.823
Stance	**0.791**	0.768
Overall F1	**0.819**	0.796
Frame Classification		
Primary Frame	**0.783**	0.761
Secondary Frame	**0.692**	0.671
Argument Extraction		
Claim	**0.806**	0.782
Evidence	**0.774**	0.753
Warrant	**0.721**	0.698
Overall F1	**0.767**	0.744
DVF Aggregate Scores		
Completeness	**0.831**	0.809
Faithfulness	**0.887**	0.863
Coherence	**0.792**	0.771
Relevance	**0.856**	0.834

Table 4: Pipeline performance against the 2,000-article gold standard (macro F1). Gemini-2.5-flash serves as the primary extractor; LLaMA-4-Maverick-17B provides open-weight baseline performance. Bolded values indicate the best performance per component.

against a 2,000-article gold standard created via Zooniverse annotations.

Table 4 reports system performance across all three AFA components for both Gemini-2.5-flash (primary extractor) and LLaMA-4-Maverick-17B (open-weight baseline). Gemini consistently outperforms LLaMA by roughly two to three points across components. Performance is highest for actor type identification ($F_1 = 0.847$) and lowest for second frame classification ($F_1 = 0.692$), reflecting the inherent ambiguity and sparsity of secondary frame labels. Unlike primary frames, which typically correspond to the dominant narrative focus of a passage, secondary frames often capture peripheral or overlapping interpretive cues (e.g., regulatory and technological frames co-occurring), making automatic prediction more difficult.

Beyond intrinsic F1, we validate extractions using the DVF aggregate scores, which confirm quality across four dimensions. Faithfulness in particular remains consistently high, indicating that extracted components are well-grounded in the source text. To ensure transparency, DVF scores are validated against a 500-sample human evaluation set. Detailed per-judge breakdowns and human validation statistics are reported in Appendix I.2, while we present only the aggregate results in the main text for clarity.

5.2 Actor Prominence Over Time

To address **RQ1 (Actors)**, we examine how different categories of actors participate in climate discourse in financial news. Using our five-category schema (*companies*, *financial institutions*, *governments*, *NGOs/advocacy*, *individuals*), we track longitudinal changes in actor prominence across four temporal strata (2000–2007, 2008–2012, 2013–2018, 2019–2023). Figure 2 displays actor distributions over time.

We observe a clear structural shift in discursive authority ($\chi^2 = 217.3$, $p < 0.001$). In the early 2000s, government ($\approx 31\%$) and companies ($\approx 21\%$) dominated climate-related financial reporting, consistent with a regulatory- and advocacy-driven framing of climate change. For instance, a 2006 article cites the World Wildlife Fund warning that "regulatory inaction risks locking in high-carbon infrastructure." During the financial crisis, individuals rose to $\approx 18\%$ of mentions, while governments began to decline in relative presence ($\approx 28\%$), reflecting a shift toward market-based voices.

In the post-crisis period, coinciding with the Paris Agreement, financial institutions surged from $\approx 15\%$ before the crisis to more than 25% after 2015, and grew to $\approx 34\%$ during recent years, becoming the single largest actor group. Asset managers and banks increasingly positioned themselves as climate leaders. For example, a 2021 Dow Jones piece quotes BlackRock: "Sustainable finance is now central to alpha generation."

In parallel, during Climate surge NGOs fell below 10%, while governments stabilized around 20%, often providing background regulatory context rather than leading the narrative.

These findings suggest a rebalancing of discursive power: from a regime led by governments to a financial-market regime where financial institutions and companies are the dominant voices in elite climate discourse. Exact proportions are reported in Appendix J.1.

5.3 Narrative Transformation in Frames

To address **RQ2 (Frames)**, we analyze temporal shifts in how climate issues are framed in elite financial news. We track the eight-frame typology introduced earlier (economic risk, regulatory compliance, economic opportunity, technological solution, market dynamics, environmental urgency, social responsibility, uncertainty/skepticism) across

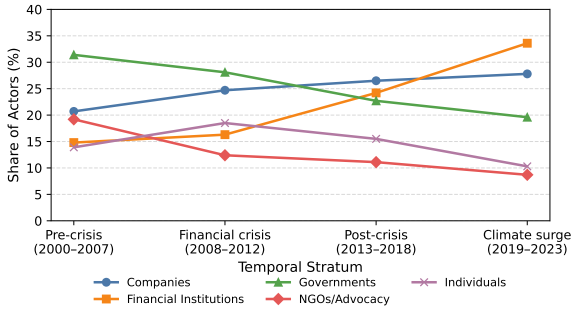


Figure 2: Actor prominence over time, showing proportional distribution across five actor categories in four temporal strata.

four temporal strata, and test for significant distributional changes over time.

From risk/compliance to opportunity/innovation. A visual overview in Figure 3 shows a clear trend: risk and compliance-oriented coverage recedes while opportunity and innovation-oriented coverage grows. A changepoint analysis (PELT) identifies a statistically significant structural break in **2015Q4–2016Q2**, aligning with the Paris Agreement and related policy announcements, after which *economic opportunity* and *technological solution* become increasingly salient.

Statistical evidence. We observe a reversal in the relative weights of *economic risk* and *economic opportunity* frames before vs. after the 2015–2016 changepoint identified by our PELT analysis (see §5.3). Opportunity rises substantially in the late period, while risk declines; *technological solution* also increases and frequently co-occurs with opportunity (positive association), consistent with a forward-looking investment narrative. Exact per-period percentages and significance tests are reported in Appendix J.2.

Illustrative contrast. Early-period articles often emphasize compliance costs or exposure (e.g., mandated adjustments, liabilities), whereas post-2015 coverage increasingly highlights growth and competitive positioning (e.g., green finance pipelines, innovation leadership). Short examples are shown below (paraphrased for brevity):

- **Risk (pre-2015):** “New carbon rules will raise operating costs and pressure margins.”
- **Opportunity (post-2015):** “Climate-focused portfolios and clean-tech investments open new alpha and market-leadership avenues.”

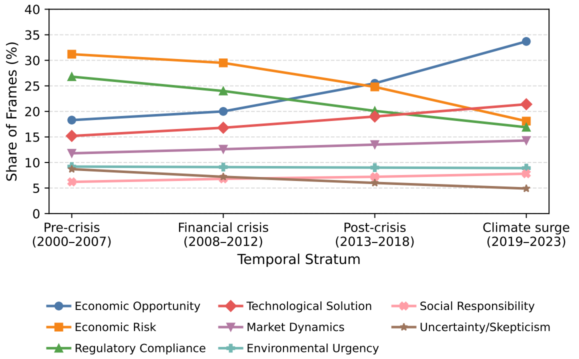


Figure 3: Temporal dynamics of climate frames across four strata.

There is a clear, statistically grounded pivot in financial-news framing of climate: the narrative transitions from *risk and regulatory burden* to *opportunity and technological innovation*, with the turning point concentrated around 2015–2016. This reframing is consistent with the sector’s shift toward green finance and investment-led rationales.

5.4 Actor–Frame–Argument Alignment (RQ3)

To answer **RQ3 (Arguments)**, we analyze how actor groups combine frames with arguments to construct climate narratives. Beyond overall shifts, this analysis reveals distinct rhetorical repertoires that different actors use in elite financial news.

Actor–Frame Associations. Table 17 reports standardized residuals from a χ^2 test of independence, capturing which frames are over- or under-represented by each actor category. Companies and financial institutions show a pronounced preference for *economic opportunity* frames, while NGOs disproportionately emphasize *environmental urgency*. Governments and regulators display more balanced use of *risk* and *regulatory compliance* frames, reflecting their institutional roles.

Argument Strategies by Actor. Warrant analysis highlights divergent argumentative logics:

- **Companies** often invoke *competitive advantage* and *innovation opportunity* warrants, presenting climate action as a business growth strategy.
- **Financial Institutions** emphasize *market leadership* and *risk-adjusted returns*, linking climate investments to fiduciary responsibility.

- **Governments/Regulators** rely heavily on *regulatory necessity* and *compliance* warrants.
- **NGOs/Advocacy groups** foreground *environmental urgency* and *ethical responsibility*.
- **Individuals (researchers/experts)** emphasize *scientific evidence* and risk communication, often reinforcing NGO frames but with technical grounding.

Argument Complexity. We measure elaboration as the median number of premises per claim. Opportunity-oriented arguments show greater complexity (median 2.8–3.2 premises), while risk arguments are terser (median 1.4–1.7 premises), which suggests that forward-looking frames demand more justificatory investment.

Illustrative Examples.

- **NGO (2010)**: "Governments must act now; unchecked emissions will accelerate ecological collapse." (Frame: Environmental Urgency).
- **Company (2018)**: "Investing in clean tech secures long-term competitiveness in global markets." (Frame: Economic Opportunity).
- **Financial Institution (2021)**: "Climate-focused portfolios represent the next frontier for alpha generation." (Frame: Economic Opportunity).

Actor groups not only favor different frames but also use distinctive argumentative strategies. This alignment of actors, frames, and arguments demonstrates how the narrative has shifted: financial institutions and companies increasingly position themselves as market leaders in the climate transition, while NGOs emphasize urgency and governments focus on compliance. Together, these patterns reveal a fragmented but structured rhetorical field in financial climate discourse.

These financial-news findings align with LLM-based studies of climate discourse outside the financial domain. For example, Lee et al. (Lee et al., 2024) show that model-derived framing signals correlate with public opinion trends on global warming, suggesting that automated frame extraction can reflect wider societal narratives. This cross-domain concordance lends external validity to our inferences about the reframing of climate in financial media.

6 Conclusion

We presented an Actor–Frame–Argument (AFA) pipeline for longitudinal discourse analysis that combines LLM-based extraction with a decompositional verification framework. Applied to a corpus of climate-related financial news spanning two decades, the approach yields both methodological robustness and substantive insight.

We document a structural transformation in elite financial coverage: from early emphasis on economic risk and regulatory compliance toward a post-2015 discourse centering on opportunity and technological solutions. We further show that actors deploy distinct rhetorical strategies, companies and financial institutions emphasize competitive advantage and risk-adjusted returns, governments stress compliance and policy instruments, and NGOs foreground environmental urgency, revealing a patterned but heterogeneous field of climate finance narratives.

Methodologically, our pipeline offers a replicable pipeline for integrating actor identification, framing, and argument mining with auditable evaluation. The results underscore how financial media can shape expectations about the pace and direction of the climate transition.

By bridging advances in NLP with questions central to climate communication and finance, this study shows how computational social science can illuminate the stories that steer capital and policy.

Limitations

Our study is based on a single, English-language news source (Dow Jones Newswire). Future work should diversify sources to assess the generality of our findings. While we validated our LLM-based annotation, LLMs have inherent biases. Our validation mitigates this, but cannot eliminate it entirely. Recent work highlights that LLMs can reproduce or amplify biases in climate and sociopolitical communication (Chen et al., 2023; Islam and Goldwasser, 2024). Our Decompositional Verification Framework reduces some risks by multi-judge validation and faithfulness checks, but a comprehensive fairness audit and subgroup robustness tests remain important future directions, especially for applications that could influence investment or policy decisions. Finally, our study maps discourse but does not directly measure its real-world impact on capital flows, a critical avenue for future research. Future work should extend coverage beyond a single newswire and language, link discursive shifts to measurable market outcomes and policy cycles and explore real-time monitoring tools for emerging narrative turns.

Ethics Statement

Our analysis does not involve private or personally identifiable information. We acknowledge the environmental costs of large model training and inference. Recent analyses indicate that improvements in hardware efficiency and software optimizations have begun to reduce marginal carbon intensity for many workloads (Patterson et al., 2022), but these costs remain nontrivial. To reduce our footprint, we used stratified sampling, efficient prompting, and multi-stage extraction rather than retraining large models end-to-end. We report these operational choices to enable reproducibility and to encourage energy-aware replication. Future work should additionally explore lightweight domain adapters and retrieval-augmented approaches to further minimize compute and emissions. We also acknowledge that LLMs may reflect biases from their training data; our validation against human annotators is a crucial check. The methods presented could be used to monitor "greenwashing," but could also be used to craft more persuasive disinformation. We release our findings in the spirit of open academic inquiry, believing the benefit of transparently understanding media narratives outweighs the risk of misuse.

Acknowledgments

Ruiran Su is supported by a Jardine Foundation Scholarship. We gratefully acknowledge the contributions of the volunteers on the Zooniverse platform, whose dedication made the construction of the gold-standard dataset possible. We also thank Dow Jones for providing access to the news corpus used in this study. We also acknowledge the use of the University of Oxford's Advanced Research Computing (ARC) facility for the computational resources used in this work. Finally, we thank the anonymous reviewers for their constructive feedback that helped improve this paper.

References

Anthropic. 2024. Claude sonnet 4. Model documentation / system card. Claude Sonnet 4, proprietary model.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Maxwell T. Boykoff and Jules M. Boykoff. 2004. Balance as bias: Global warming and the u.s. prestige press. *Global Environmental Change*, 14(2):125–136.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) & 7th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 438–444. Association for Computational Linguistics.

Kaiping Chen, Ruitong Duan, Ying Peng, and Jingwen Zhang. 2023. How gpt-3 responds to different publics on climate change and black lives matter: A critical appraisal of equity in conversational ai. *arXiv preprint arXiv:2209.13627*. Preprint, last revised March 2023.

Zheng Chen, Yulan He, and Yu Zhang. 2021. Hitting your marq: Multimodal argument quality assessment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6958–6970. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, and Marcel Blistein et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google / DeepMind. Preprint / technical report, includes the Gemini 2.5 Flash model.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Dow Jones. 2024. Dow Jones Intelligent Identifiers (DJID) Taxonomy API. Proprietary subject classification system for financial news content.

Ran Duan and Serena Miller. 2021. Climate change in china: A study of news diversity in party-sponsored and market-oriented newspapers. *Journalism*, 22(10):2493–2510.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Martin Gleize, Eyal Shnarch, Ran Levy, Ben Bogin, Ranit Aharonov, and Noam Slonim. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5772–5781. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can large language models perform relation-based argument mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of NAACL-HLT 2018*, pages 1930–1940.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550, Seattle, United States. Association for Computational Linguistics.

Xinyu Hua, Zichao Yang, and Yiming Yang. 2022. Iam: A comprehensive and large-scale dataset for integrated argument mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2052–2064. Association for Computational Linguistics.

Tunazzina Islam and Dan Goldwasser. 2024. Post-hoc study of climate microtargeting on social media ads with llms: Thematic insights and fairness evaluation. *arXiv preprint arXiv:2410.05401*. Preprint.

Ayush Jain, Manikandan Padmanaban, Jagabondhu Hazra, Shantanu Godbole, and Hendrik Hamann. 2024. Empowering sustainable finance: Leveraging large language models for climate-aware investments. In *ICLR 2024 Workshop on Tackling Climate Change with Machine Learning*. Climate Change AI.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

Markus Leippold. 2023. Thus spoke gpt-3: Interviewing a large-language model on climate finance. *Finance Research Letters*, 53:103617.

Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadić. 2025. Large language models in argument mining: A survey.

Yang Liu, Christian Stab, and Iryna Gurevych. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2135–2149. Association for Computational Linguistics.

Yang Liu, Sheng Zhang, Jingbo Shang, and Jiawei Han. 2020. A top-down neural architecture towards text-level parsing of discourse relations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6386–6395. Association for Computational Linguistics.

MetaAI. 2025. LLaMA 4: Maverick (17b) — a multimodal mixture-of-experts model. Model card / documentation for LLaMA-4 Maverick-17B (128 experts), accessed via ModelScope / model repositories.

Mistral. 2024. Mixtral 8×22b. https://huggingface.co/mistralai/Mixtral-8x22B-v0.1. Open-weight mixture-of-experts language model.

Matthew C. Nisbet. 2009. Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development*, 51(2):12–23.

OpenAI. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276. Multimodal "omni" model by OpenAI.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. In *NLPerspectives 2024*. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn discourse treebank 3.0. Linguistic Data Consortium.

ProQuest. 2024. Home - Dow Jones Factiva - LibGuides. `https://proquest.libguides.com/factiva`. Describes DJID as containing approximately 350,000 classification codes.

Qwen. 2025. Qwen3 30b a3b. https://huggingface.co/unsloth/Qwen3-30B-A3B-GGUF. Open-weight evaluator model.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Sharon Gadarian, Bethany Albertson, and David Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Maria Ruiz-Dolz, Santiago Cortes, Javier Garcia, and Ana Garcia-Serrano. 2021. Transformer-based models for automatic identification of argument relations. *IEEE Intelligent Systems*, 36(4):54–62.

Tatjana Schirmag, Jakob H. Wedemeyer, Annika Stechemesser, and Leonie Wenz. 2025. Neural topic modeling reveals german television's climate change coverage. *Communications Earth & Environment*, 6(1):441.

Hannah Schmid-Petri, Silke Adam, Ivo Schmucki, and Thomas Häussler. 2017a. A changing climate of skepticism: The factors shaping climate change coverage in the us press. *Public Understanding of Science*, 26(4):498–513.

Hannah Schmid-Petri, Silke Adam, Ivo Schmucki, and Thomas Häussler. 2017b. A changing climate of skepticism: The factors shaping climate change coverage in the u.s. press. *Public Understanding of Science*, 26(4):498–513.

Hannah Schmid-Petri and Dorothee Arlt. 2016a. Constructing an illusion of scientific uncertainty? framing climate change in german and british print media. *Communications*, 41(3):265–289.

Hannah Schmid-Petri and Dorothee Arlt. 2016b. Constructing an illusion of scientific uncertainty? framing climate change in german and british print media. *Communications*, 41(3):265–289.

Robert J. Shiller. 2017. Narrative economics. *American Economic Review*, 107(4):967–1004.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. In *Proceedings of the 2017 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 985–995, Los Angeles, California. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017b. Parsing argumentation structures in persuasive essays. In *Computational Linguistics*, volume 43, pages 619–659.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Assaf Toledo, Yoav Kantor, Gabriel Stanovsky, and Ido Dagan. 2021. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1427–1437.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. In *Proceedings of the AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. ArXiv:2110.12010.

Qilong Wu, Thibault Masson, Aritra Mukherjee, Yihao Zhao, Chaima Driouich, Xingjian Xing, Patrick Paroubek, and Mickael Coustaty. 2024. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation. *arXiv preprint arXiv:2412.10906*. Preprint.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Corpus Curation and Validation

A.1 DJID Code Reference

Table 5 lists all Dow Jones Intelligent Identifier (DJID) codes used in our corpus construction, grouped into the three domains introduced in Section 3.1: Core Climate Issues, Energy Transition, and Climate-Affected Sectors.

Domain	Code	Description
Core Climate Issues	N/ENV	Environment
	N/CO2	Carbon Dioxide / Emissions
	N/RNW	Renewables
Energy Transition	N/BFL	Biofuels
	N/COA	Coal
	N/NUK	Nuclear Energy
	N/NGS	Natural Gas
Climate-Affected Sectors	N/AGR	Agriculture
	N/FST	Forestry

Table 5: Dow Jones Intelligent Identifier (DJID) codes used in constructing the climate news corpus.

A.2 Validation of DJID Filtering

To evaluate the reliability of DJID-based filtering, we conducted a two-part validation study.

False-Negative Analysis. We sampled 1,000 articles not tagged with climate-related DJIDs and re-screened them using:

- **Keyword Lexicon:** A curated set of 152 climate-related terms (see Appendix A.3) matched against article text.
- **Supervised Classifier:** A RoBERTa-base model fine-tuned on 5,000 climate vs. non-climate news articles (macro F1 = 0.91). Articles flagged as "climate-related" by either method were manually verified.

Out of 1,000 articles, 73 were flagged; manual inspection confirmed 60 as genuinely climate-related. This implies a false-negative rate of ∼8%.

False-Positive Analysis. We randomly sampled 500 articles tagged with climate DJIDs. Two annotators independently verified whether each article substantively discussed climate issues (Krippendorff's $\alpha = 0.79$). Only 21 (4.2%) were deemed false positives, yielding estimated precision >95%.

Summary. Overall, DJID filtering achieves recall of ∼92% and precision >95%. While highly accurate, complementary methods (lexicon or classifier-based augmentation) can further enhance coverage in future work.

A.3 Keyword Lexicon Construction

The climate keyword lexicon was designed to capture articles that may not be tagged with relevant DJIDs. It was constructed from prior climate communication research, IPCC glossaries, and domain-specific terminology in financial reporting. The final lexicon contains 152 terms across five thematic categories. Representative examples are provided in Table 6; the full lexicon is released with our supplementary materials for reproducibility.

Category	Example Keywords
General Climate	climate change; global warming; greenhouse effect
Carbon	carbon; CO2; carbon tax; carbon capture
Energy	renewables; solar; wind; fossil fuels; biofuels
Finance	ESG; green bonds; carbon markets
Policy	Paris Agreement; Kyoto Protocol; net zero

Table 6: Representative subsets of the climate keyword lexicon. The full lexicon (152 terms) is provided in the supplementary release.

A.4 Preprocessing Details

Normalization. All articles were standardized through boilerplate removal (e.g., repeated headers, copyright notices), whitespace normalization, and Unicode normalization. Token-level cleaning was avoided to preserve domain-specific terminology.

Near-Duplicate Detection. Because wire services frequently release slightly modified repeats of the same story, we applied MinHash-based locality-sensitive hashing (LSH) to identify duplicates. Pairs with Jaccard similarity ≥ 0.9 were marked as near-duplicates. We retained one canonical version per cluster, removing ∼8.7% of articles. This prevents redundancy and avoids overweighting syndicated stories.

Final Corpus. After preprocessing, the corpus contains 980,061 unique climate-related articles with preserved metadata (timestamps, DJIDs, article source identifiers, etc.), ensuring integrity for diachronic analysis.

A.5 Corpus Temporal Distribution

B Sampling Hyperparameter Analysis

B.1 Sensitivity Analysis of MMR λ

The choice of the Maximal Marginal Relevance (MMR) parameter λ governs the trade-off between selecting articles central to a theme (relevance) and those that are novel (diversity). We performed a sensitivity analysis by generating five distinct subsamples using different λ values and evaluating the

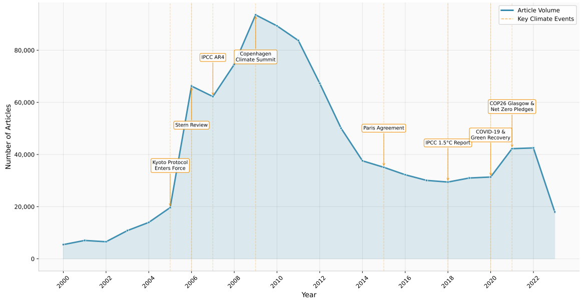


Figure 4: Temporal distribution of Dow Jones climate-related articles (2000–2023), annotated with key climate events.

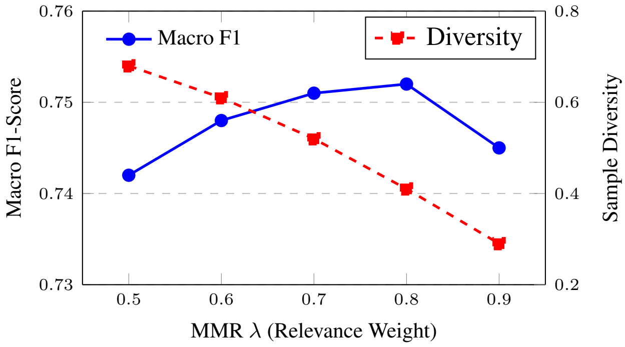


Figure 5: Trade-off between model performance (Macro F1) and sample diversity across MMR λ values.

performance of our downstream argument extraction model on each. As shown in Table 7 and Figure 5, a value of $\lambda = 0.7$ provides a near-optimal balance, achieving high F1-score while retaining substantial sample diversity, measured as the average pairwise cosine distance between selected articles within a cluster. While $\lambda = 0.8$ yields a marginal F1-score improvement, this comes at the cost of a significant drop in diversity, making it less suitable for our goal of capturing both mainstream and outlier arguments.

Table 7: Performance of the downstream argument extraction model and sample diversity as a function of the MMR parameter λ. The highest F1-score is in bold.

Metric	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
Macro F1	0.742	0.748	0.751	**0.752**	0.745
Precision	0.739	0.745	0.750	0.758	0.761
Recall	0.745	0.751	0.752	0.746	0.730
Diversity	0.68	0.61	0.52	0.41	0.29

Table 8: Examples of emergent cluster themes at different values of k for the 2019–2023 stratum.

Value of k	**Illustrative Cluster Themes**
$k = 4$	- Renewable Energy & Policy (Conflated) - Corporate ESG & Finance (Conflated)
$k = 8$	- Renewable Energy Technology - Climate Finance Policy - Corporate Sustainability Reports - Carbon Markets
$k = 16$	- Utility-Scale Solar Projects (Redundant) - Onshore Wind Projects (Redundant) - Voluntary Carbon Credits (Redundant)

B.2 Justification for Thematic Cluster Count (k)

The range for the number of thematic clusters, k_t, was constrained to $[5, 15]$ based on qualitative analysis. Table 8 provides illustrative examples from the 2019-2023 stratum, demonstrating that $k < 5$ conflates distinct themes, while $k > 15$ creates spurious, overly granular micro-clusters.

C Active Learning Details

C.1 Preliminary Model Architecture

The active learning loop was driven by a preliminary argument component detector based on *RoBERTa-base*. Key training hyperparameters are detailed in Table 9. The model achieved a macro F1-score of 79.5% on a held-out portion of the 1,000-article seed set.

Table 9: Hyperparameters for the preliminary argument component detector used in the active learning loop.

Hyperparameter	**Value**
Base Model	‘RoBERTa-base‘
Learning Rate	2e-5
Optimizer	AdamW
Batch Size	16
Max Sequence Length	256
Epochs	4
Warmup Steps	100

To clarify the relationship between our various annotated sets:

1. **Seed set for active learning** (1,000 articles): Sampled from full 980k corpus and annotated in-house before final sampling to bootstrap the preliminary argument detector.

2. **Final sample** (4,143 articles): Constructed using active learning with the seed-trained detector.

3. **Gold standard** (2,000 articles): Subset of the 4,143 sample, annotated via Zooniverse for pipeline evaluation.

4. **20k validation corpus** (Appendix D): Separate stratified random sample from the 980k corpus, annotated in-house, used *only* to validate that our 4,143 sample preserves downstream task performance. Not used for training or primary analysis.

C.2 Active Learning Convergence

The active learning process converged after 4 iterations. We defined convergence as the point where the Jensen-Shannon divergence (JSD) between the entropy distributions of articles selected in consecutive iterations fell below a threshold of 0.05. Figure 6 plots this trend, demonstrating that the set of "hardest" articles identified by the model stabilized quickly.

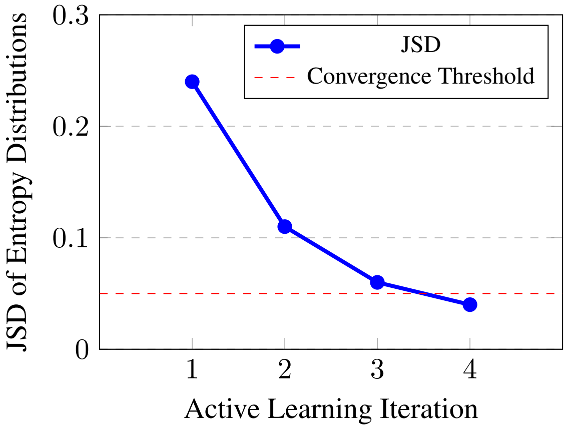


Figure 6: Convergence of the active learning loop. The JSD between selected article sets drops below the 0.05 threshold at iteration 4.

D Extended Validation Results

To verify the integrity of our 4,143-article sample, we conducted a rigorous validation protocol focused on two key criteria: (1) the preservation of downstream task performance and (2) the fidelity of metadata distributions compared to the parent corpus.

D.1 Downstream Task Performance Preservation

The most critical test of a sub-corpus is its utility for the intended downstream task. We evaluated this by training an argument extraction model on our sample and comparing its performance against the same model architecture trained on a much larger 20,000-article stratified random sample.

Model Architecture. The argument extraction model is a token classifier built upon `RoBERTa-large`. We added a linear layer on top of the final hidden states of the RoBERTa model to classify each token into IOB format (Inside, Outside, Beginning) for our target components ('Claim', 'Premise'). The model was fine-tuned end-to-end. Key hyperparameters, chosen via a limited search on a development set, are detailed in Table 10.

Table 10: Hyperparameters for the `RoBERTa-large` argument extraction model used in our validation experiment.

Hyperparameter	**Value**
Base Model	`roberta-large`
Learning Rate	1e-5
Optimizer	AdamW with linear decay
Weight Decay	0.01
Batch Size	8
Max Sequence Length	512 tokens
Training Epochs	5

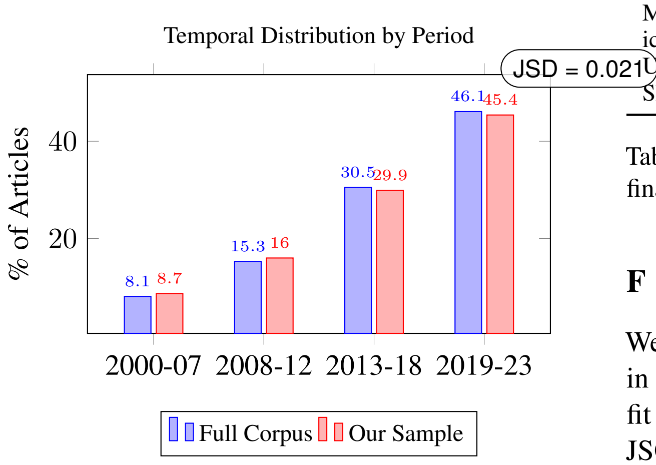
Results. As shown in Table 11, the model trained on our 4,143-article sample achieved a macro F1-score of 75.1%. This represents a performance degradation of only 3.3 percentage points compared to the 78.4% F1-score from the model trained on the 20k-article baseline. This result is highly encouraging, demonstrating that our sampling strategy preserves over 95% of the performance while using less than 21% of the training data, thus confirming its high data efficiency.

Table 11: Detailed performance comparison for the argument extraction task. Both models were evaluated on the same held-out test set of 1,500 articles. The best scores for each metric are in bold.

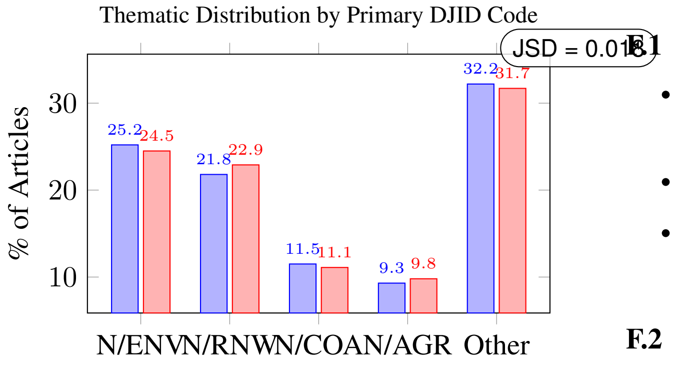
Training Sample	**Claim**			**Premise**			**Macro**
	P	R	F1	P	R	F1	F1
20k Stratified Random	**0.801**	**0.753**	**0.776**	**0.769**	**0.824**	**0.795**	**0.784**
Our Sample (4k)	0.782	0.725	0.752	0.723	0.781	0.751	0.751

D.2 Distributional Similarity Fidelity

We further validated that our sample's metadata distribution faithfully mirrors that of the parent 980k-article corpus. We measured the Jensen-Shannon divergence (JSD) for key distributions, where a lower JSD score indicates higher similarity. Figure 7 provides a visual comparison for temporal and thematic distributions, confirming a very high degree of fidelity (JSD < 0.03 for both).



(a) Comparison of temporal distributions.



(b) Comparison of thematic distributions.

Figure 7: Comparison of key metadata distributions between the full 980k-article corpus and our 4k-article sample, confirming high distributional fidelity.

E Frame Typology

Our analysis employs a predefined eight-frame typology for climate discourse in financial news. This typology synthesizes categories from communication studies (Nisbet, 2009; Schmid-Petri and Arlt, 2016b) with finance-specific frames (e.g., economic risk, market dynamics) observed in preliminary corpus exploration. Table 12 provides definitions of each frame.

Frame	Definition
Economic Opportunity	Frames climate change as growth, innovation, and investment potential.
Economic Risk	Highlights financial losses, stranded assets, and risks to firms or markets.
Regulatory Compliance	Focuses on laws, policies, and regulatory burdens or incentives.
Technological Solution	Emphasizes innovation, R&D, and technical fixes to climate challenges.
Environmental Urgency	Stresses ecological severity and the need for rapid action.
Social Responsibility	Invokes ethics, corporate responsibility, and societal expectations.
Market Dynamics	Frames climate in terms of competition, supply-demand, and positioning.
Uncertainty Skepticism	Expresses doubt or skepticism about climate science, policies, or impacts.

Table 12: Eight-frame typology for climate discourse in financial news.

F Prompt Templates and Schemas

We release all prompt templates and schemas used in the Actor–Frame–Argument (AFA) pipeline. To fit the two-column format, we present compact JSON schemas and keep explanatory text brief. Full JSON schema files and scripts are included in our supplementary materials.

F.1 General Protocol

- Output must be a valid JSON object (no prose).
- Missing fields: return null or empty list.
- Frames must match exactly with the predefined typologies.

F.2 Stage 1: Actor–Stance

Task: Extract actors, type, stance, supporting quote. **Schema:**

```
{
 "actors": [{
   "name": "string",
  "actor_type": "company|gov|indiv|org",
   "stance": "supportive|opposing|,
           neutral|mixed",
   "quote_text": "string",
   "climate_relevance": "string"
 }]
}
```

F.3 Stage 2: Frame Classification

Task: Assign one primary and optional secondary frame. **Schema:**

```
{
 "primary_frame": "frame_enum",
 "secondary_frame": "frame_enum_or_null",
 "justification": "string",
 "climate_connection": "string"
}
```

F.4 Stage 3: Argument Extraction

Task: Extract claim, evidence, warrant, impact, and optional supporting arguments. **Schema:**

```
{
 "claim": "string",
 "evidence": ["string", "..."],
 "warrant": "string",
 "impact": "string",
 "supporting_arguments": [{
   "claim": "string",
   "evidence": ["string", "..."],
   "warrant": "string"
 }]
}
```

F.5 DVF Judge Prompts

Dimensions: Completeness, Faithfulness, Coherence, Climate Relevance. Each atomic check is scored on a continuous scale in $[0, 1]$, where 0 indicates complete failure, 1 indicates full success, and intermediate values capture partial correctness.

Schema (example):

```
{
 "completeness": {"actors":0.0-1.0,
                  "stance":0.0-1.0,
                  "frames":0.0-1.0,
                  "arguments":0.0-1.0},
 "faithfulness": {"quote_alignment":0.0-1.0,
            "para_equivalence":0.0-1.0},
 "coherence": {"links_preserved":0.0-1.0,
           "schema_wellformed":0.0-1.0},
 "climate_relevance": {"on_topic":0.0-1.0,
               "peripheral_excluded":0.0-1.0}
}
```

F.6 Inference Settings

Temperature = 0.2, Top-p = 0.9, max tokens = 512 (stages 1–2), 768 (stage 3), 256 (DVF). Stop sequences: `\{"\{\n  ","“‘\}`. Outputs validated against schemas; retries on failure.

G Human-Annotated Gold Standard

To establish a reliable evaluation anchor for our AFA pipeline, we constructed a gold-standard dataset of 2,000 articles sampled from the corpus. This section provides full details of the annotation protocol, guidelines, adjudication process, and agreement metrics.

G.1 Annotator Training and Background

Annotations were collected via *Zooniverse*, a widely used citizen science platform for distributed human annotation. Instead of a small team of in-house coders, we leveraged a large pool of volunteers. To ensure annotation quality, we implemented the following measures:

- **Onboarding Tutorial:** All contributors completed an interactive tutorial with examples of actor, frame, and argument annotations, as well as practice tasks with feedback.
- **Redundancy:** Each article was annotated independently by at least $k = 5$ volunteers to mitigate individual errors.
- **Aggregation:** Final labels were assigned via majority vote (for categorical tasks such as stance and frames) or consensus heuristics (for argument spans, using token-level agreement).

While annotators did not have formal training in linguistics or climate communication, the redundancy and aggregation protocol yielded high inter-annotator agreement (see Table 2). This approach demonstrates the feasibility of scalable citizen science annotation for complex discourse tasks.

G.2 Annotation Guidelines

The schema covered all components of the Actor–Frame–Argument (AFA) pipeline:

- **Actor–Stance**: identify all actors (company, government, NGO, individual) and classify stance as *supportive*, *opposing*, *neutral*, or *mixed*. Each actor annotation required a supporting quote.
- **Frame Classification**: assign one *primary frame* and (if present) one *secondary frame* from the eight-frame typology (see Appendix E). Annotators provided a brief justification.
- **Argument Extraction**: decompose argument structure into (i) central claim, (ii) evidence spans, (iii) warrant linking claim and evidence, and (iv) impact (stated or implied). Additional supporting arguments were annotated if present.

Examples of annotated documents and excerpts from the guidelines are released with our supplementary materials.

G.3 Annotation Procedure

Each of the 2,000 articles was annotated on Zooniverse by at least five independent contributors. For categorical tasks (stance, frame), final labels were determined by majority vote. For span-based tasks (actors, evidence, warrants), we aggregated annotations using token-level agreement and retained spans marked by at least 60% of contributors. Disagreements were resolved through consensus aggregation rather than individual adjudication. This redundancy protocol provides a robust approximation of expert annotation quality while leveraging the scale of citizen science.

H Annotation Guidelines (Excerpt)

This section provides an excerpt from the full annotation guidelines used for training human coders. These examples illustrate how annotators distinguished labels and handled common edge cases in the Actor–Frame–Argument (AFA) schema. The complete guidelines and codebook are released with our supplementary materials.

H.1 Actor–Stance Annotation

Task: Identify actors (organization, company, government, NGO, or individual) and assign stance.

Rules:

- An actor must be explicitly mentioned in the text. Generic terms such as "analysts" or "critics" are excluded unless tied to a named entity.
- If multiple subsidiaries or affiliates are mentioned, treat them as separate actors only if they present distinct stances.
- Stance definitions:
  - **Supportive:** Explicitly endorses climate action, investment, or regulatory compliance.
  - **Opposing:** Rejects or resists climate measures, or highlights negative impacts.
  - **Neutral:** Mentions climate issues without evaluative judgment.
  - **Mixed:** Expresses both supportive and opposing positions within the same context.
- Each stance annotation must be supported by a verbatim quote.

Example: *"ExxonMobil said new carbon rules will increase costs for energy producers."* → Actor: ExxonMobil; Stance: Opposing; Quote: verbatim.

H.2 Frame Classification

Task: Assign one primary frame and, if present, one secondary frame from the eight-frame typology.

Rules:

- Primary frame = the dominant way the issue is presented.
- Secondary frame = optional, used only if the text clearly invokes a secondary dimension.
- If multiple frames appear, select the one most central to the claim or argument as primary.

Example: *"Investors see renewable energy as the next big growth opportunity."* → Primary frame: Economic Opportunity; Secondary frame: Technological Solution.

H.3 Argument Extraction

Task: Decompose into claim, evidence, warrant, and impact.

Rules:

- **Claim:** Main proposition advanced by the actor.
- **Evidence:** Factual or statistical support cited in the text. Use verbatim snippets.
- **Warrant:** The reasoning linking claim and evidence. This can be implicit but should be paraphrased.
- **Impact:** Stated or implied consequence if the claim is accepted.

Example: Sentence: "BlackRock argues that green portfolios outperform traditional funds, with ESG funds gaining 4.3% in 2020."

- Claim: "Green portfolios outperform traditional funds."
- Evidence: "ESG funds gaining 4.3% in 2020."
- Warrant: "Financial returns justify sustainable investment."

H.4 Ambiguity Resolution

- If a passage fits multiple plausible labels, select the one best supported by evidence.
- Use "Mixed" stance only if a single actor explicitly expresses both support and opposition.
- Annotators flagged unclear cases for adjudication rather than guessing.

H.5 Inter-Annotator Agreement

Agreement was computed separately for each component type, using measures appropriate to multi-annotator settings:

- **Actor Identification**: token-level precision, recall, and F1 computed pairwise across annotators, averaged with macro-F1.
- **Stance Classification**: Fleiss' κ and Krippendorff's α (nominal scale) over four stance categories.
- **Frame Assignment**: Krippendorff's α on the eight-frame set for primary frames; Jaccard similarity averaged across annotator pairs for secondary frames.
- **Argument Components**: span-level F1 (claims, evidence, warrants, impacts), aggregated by majority vote. Macro-F1 reported across components.

H.6 Consensus Derivation for Span-Based Tasks

For span-based tasks (actors, arguments), we derive a single gold-standard annotation from multiple human labels using a two-step token-aggregation and span-reconstruction procedure. This approach consolidates fine-grained token agreement while avoiding over-fragmentation of long argument components.

Step 1: Token-Level Agreement. Let each article be tokenized as a sequence $T = [t_1, t_2, \dots, t_m]$. Each annotator a_i provides a binary label sequence $y_i(t_j) \in \{0, 1\}$, where 1 indicates that token t_j belongs to a span of interest (e.g., actor mention, claim, evidence). For each token, we compute an agreement score:

$$A(t_j) = \frac{1}{N} \sum_{i=1}^{N} y_i(t_j),$$

where N is the number of annotators (typically $N = 5$). Tokens satisfying $A(t_j) \geq \tau$ (with $\tau = 0.5$) are retained, producing a consensus mask $M = [m_1, m_2, \dots, m_m]$ where $m_j = 1$ iff $A(t_j) \geq \tau$.

Step 2: Span Reconstruction. Contiguous retained tokens are merged into unified spans: $S_k = \{t_p, \dots, t_q\}$ such that $m_p = \cdots = m_q = 1$, $(m_{p-1}, m_{q+1}) = 0$. If multiple annotator spans overlap semantically, we compute pairwise span-level F1:

$$\mathrm{F1}(H_i, S_k) = \frac{2 \times |\mathrm{Tokens}(H_i) \cap \mathrm{Tokens}(S_k)|}{|\mathrm{Tokens}(H_i)| + |\mathrm{Tokens}(S_k)|},$$

where H_i is a human span. Spans with F1 > 0.7 are merged by maximal coverage (minimal start, maximal end).

Handling Non-Span Tokens. Tokens not marked by any annotator (e.g., function words) are assigned the outside label O:

$$y_i(t_j) = 0 \quad \forall i, \text{ if } t_j \text{ unmarked.}$$

They are excluded from aggregation but retained for evaluation. Token-level F1 is computed as

$$\mathrm{F1} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}},$$

where consistently labeled O tokens contribute neither to the numerator nor denominator.

Illustrative Example. Sentence: *"BlackRock stated that climate-focused portfolios will drive growth."*
Three annotators mark:

- A$_1$: [BlackRock]; [climate-focused portfolios will drive growth]
- A$_2$: [BlackRock Inc.]; [climate-focused portfolios will drive growth]
- A$_3$: [BlackRock]; [portfolios will drive growth]

After token alignment and majority voting, only *BlackRock* and *climate-focused portfolios will drive growth* reach consensus. All other tokens remain labeled O. This yields clean, non-overlapping gold spans.

H.7 Use in Evaluation

The adjudicated consensus labels form the gold standard for all intrinsic evaluations in Section 5. Model outputs are evaluated against this set using precision, recall, and F1. In addition, the gold standard provides the validation baseline for the Decompositional Verification Framework (DVF), ensuring that automatic verification scores are anchored to reliable human annotations.

I Decompositional Verification Framework

I.1 Decompositional Verification Sub-Checks

DVF decomposes evaluation into atomic sub-checks that are easier for models to verify and humans to audit.

- **Completeness:** (1) Are all actors identified? (2) Is stance extracted? (3) Are frames assigned? (4) Are argument structures fully captured?
- **Faithfulness:** (1) Does each extracted component align with a direct quote? (2) Is paraphrase semantically equivalent?
- **Structural Coherence:** (1) Are actor–frame–argument links preserved? (2) Is the schema well-formed?
- **Climate Relevance:** (1) Is the extracted frame genuinely about climate? (2) Are peripheral issues (e.g., generic market news) excluded?

I.2 Decompositional Verification Framework Results

This appendix provides the full breakdown of DVF evaluations, complementing the aggregate results reported in Section 5.1. Recall that DVF evaluates extractions across four dimensions: *completeness*, *faithfulness*, *structural coherence*, and *climate relevance*. Scores are averaged over a 2,000-article gold-standard set, with validation against an independent 500-sample human evaluation.

I.3 Per-Judge Performance

DVF employs four distinct judge models to mitigate self-grading bias: GPT-4o (primary), Claude-Sonnet-4, Qwen3-30B A3B, and Mixtral-8×22B. Table 13 reports per-judge scores before aggregation. GPT-4o and Claude achieve the highest consistency, while open-weight models provide competitive but slightly noisier estimates, ensuring robustness via cross-family triangulation.

Dim.	**GPT-4o**	**Claude**	**Qwen3**	**Mixtral**
Completeness	0.842	0.837	0.814	0.808
Faithfulness	0.902	0.895	0.871	0.868
Coherence	0.801	0.794	0.772	0.769
Relevance	0.869	0.861	0.842	0.836

Table 13: DVF dimension-level scores by individual judge, averaged over the 2k-article gold standard.

I.4 Human Validation Study

To anchor automated DVF scores, we constructed a separate human evaluation set of 500 randomly sampled model outputs. Each output was annotated by three trained coders using the DVF rubric, independent from the gold-standard annotators. Agreement across coders was high: $\kappa = 0.78$ (faithfulness), $\alpha = 0.74$ (completeness), and $\alpha = 0.71$ (coherence), indicating substantial reliability.

Table 14 compares aggregated model-judge scores against human annotations. Automated DVF evaluations track human ratings closely, with deviations under 0.03 across dimensions. This validation confirms that DVF provides a faithful proxy for expert assessment while scaling efficiently.

Dimension	**Human Score**	**DVF (avg.)**
Completeness	0.842	0.821
Faithfulness	0.895	0.857
Coherence	0.781	0.792
Relevance	0.864	0.816

Table 14: Comparison of human-coded DVF scores and aggregated automated DVF scores over the 500-item validation set.

These results demonstrate that while judge-specific variation exists, the aggregate DVF scores remain well-validated against human judgment, justifying their use as the main evaluation metric in the body of the paper.

J Results

J.1 Actor Category Shares

Table 15 reports the exact proportions of each actor category across the four temporal strata, complementing the trend visualization in Figure 2.

J.2 Frame Distributions Over Time

Table 16 reports exact frame proportions (%) in early (2000–2014) vs. recent (2019–2023) periods,

Period	Comp.	Fin. Inst.	Gov.	NGO	Indiv.
2000–2007	20.7	14.8	31.4	19.2	13.9
2008–2012	24.7	16.3	28.1	12.4	18.5
2013–2018	26.5	24.2	22.7	11.1	15.5
2019–2023	27.8	33.6	19.6	8.7	10.3

Table 15: Actor category shares (%) across temporal strata.

with significance tests for temporal independence. These values complement Figure 3 and the change-point analysis described in Section 5.3.

Frame	2000–2007	2019–2023	Δ
Economic Opportunity	18.3	33.7	+15.4
Economic Risk	31.2	18.1	−13.1
Regulatory Compliance	26.8	16.9	−9.9
Technological Solution	15.2	21.4	+6.2
Market Dynamics	11.8	14.3	+2.5
Environmental Urgency	9.2	8.9	−0.3
Social Responsibility	6.2	7.8	+1.6
Uncertainty/Skepticism	8.7	4.9	−3.8

Table 16: Frame distribution shift between early (2000–2007) and recent (2019–2023) periods (percentages). Chi-square tests indicate significant temporal change for most frames (*** p<.001, ** p<.01, * p<.05; n.s. not significant), matching the main-text trend narrative.

J.3 Actor–Frame Association Matrix

Table 17 reports standardized residuals from a χ^2 test of independence between actor groups and frame usage. Positive values indicate over-representation; negative values indicate under-representation. Significance levels are adjusted with Bonferroni correction.

J.4 Qualitative Error Analysis

To better understand the model's limitations, we manually examined 150 randomly sampled mispredictions across all Actor–Frame–Argument (AFA) components. Each case was annotated by two authors following the same schema used in the main evaluation. Three dominant error categories emerged, together explaining over 80% of observed failures.

Actor Ambiguity (32%). Ambiguities often arise in passages quoting multiple entities or spokespersons. For instance:

> "BlackRock and several market analysts said green bonds will outperform traditional fixed income."

Actor Type	Econ. Opp.	Econ. Risk	Tech. Sol.	Env. Urg.
Companies	$+4.2^{***}$	-2.8^{**}	$+2.1^{*}$	-3.9^{***}
Financial Inst.	$+5.7^{***}$	-1.9^{*}	$+1.6$	-4.2^{***}
Govt./Regulators	$+0.8$	$+1.2$	-0.6	$+1.8^{*}$
NGOs/Advocacy	-5.1^{***}	$+0.9$	-1.3	$+6.8^{***}$
Researchers/Experts	-1.4	$+2.3^{*}$	$+3.4^{***}$	$+2.7^{**}$

Table 17: Actor–frame association matrix (standardized residuals from χ^2 test). Positive values indicate over-representation. Significance: *** p<.001, ** p<.01, * p<.05.

The model frequently merges "BlackRock" and "analysts" into a single actor span or assigns a collective "financial institutions" type. Disentangling speaker roles may require discourse-level coreference or dependency cues.

Frame Overlap (27%). When both opportunity and technological themes are expressed, the system struggles to assign a dominant frame. Example:

> "Investments in renewable innovation secure competitiveness for firms adapting to net-zero policies."

Here, human annotators favored *Economic Opportunity* as primary and *Technological Solution* as secondary, whereas the model reversed them. Introducing hierarchical or multi-label frame modeling could mitigate this confusion.

Argument Boundary Drift (25%). In complex sentences with subordinate clauses, claim boundaries tend to over-extend. Example:

> "Experts argue that stricter disclosure rules, which may initially burden firms, will ultimately enhance transparency and investor confidence."

The model extracted the entire sentence as a claim, omitting separation between claim and evidence. Span-level attention or discourse segmentation could help constrain extraction.

Less Frequent Errors. Other minor issues include stance misclassification for sarcastic tone (8%) and actor-type misidentification for supranational organizations such as the IMF or World Bank (6%).

Implications. Overall, qualitative inspection indicates that most errors stem from rhetorical or syntactic complexity rather than lexical gaps. These findings suggest that future work should combine LLM annotation with discourse-aware parsers and fine-grained validation to better handle ambiguity in financial commentary.