Framework for Openly Augmented Mediation (FOAM): A Pluralistic Architecture for Explainable and Contestable AI

ANONYMOUS AUTHOR(S)

High-stakes AI systems increasingly mediate access to credit, healthcare, and public benefits, yet affected parties often cannot see why a decision was made or meaningfully contest it. Even post hoc review of chain-of-thought traces from individual models can be incomplete or strategically misleading, thereby limiting accountability. We propose FOAM (Framework for Openly Augmented Mediation), a pluralistic architecture that treats explanation as a *deliberative process* rather than post-hoc narration. FOAM instantiates differentiated agents with explicit value commitments, structures their interaction through cross-examination and rebuttal protocols, and outputs not just a recommendation but a *contestable record intended to support downstream review*: claims linked to sentence-level evidence provenance, surviving objections, and explicit points of disagreement. We evaluate FOAM in evidence-grounded policy debate generation, a domain where arguments must withstand adversarial scrutiny. In a double-blind tournament of 66 cases, FOAM outperforms human-expert and zero-shot baselines on overall quality (81.7 vs. 70.1 vs. 50.6) while achieving dramatically higher evidence verifiability (76.2% perfect validation vs. 8.7% and 0%). These results demonstrate that pluralistic deliberation can produce outputs that are simultaneously persuasive *and* auditable, a necessary condition for contestable AI by design.

Additional Key Words and Phrases: Algorithmic accountability; Contestable AI; Explainable AI (XAI); Multi-agent deliberation; Evidence provenance

ACM Reference Format:
Anonymous Author(s). 2026. Framework for Openly Augmented Mediation (FOAM): A Pluralistic Architecture for Explainable and Contestable AI. 1, 1 (January 2026), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Accountability gap in high-stakes AI

AI systems are now routinely embedded in high-stakes decision workflows—healthcare triage and documentation [26], hiring and workplace management [27], credit and insurance, public benefits, and criminal-legal risk assessments [2]. In these settings, "performance" cannot be reduced to predictive accuracy or user satisfaction: when a system's output influences outcomes that materially affect people's rights, opportunities, or safety, **accountability requires (i) intelligible reasons and (ii) effective avenues to challenge and revise those reasons**. Yet most deployed AI remains organized around a monolithic model that produces a single authoritative output, with limited transparency into *why* it said what it said and little procedural support for contesting it when it is wrong, biased, or normatively inappropriate.

This accountability gap has two tightly coupled dimensions. **Explainability** is often treated as a documentation problem—generate a rationale, a summary, or a list of features—rather than a *reason-giving* problem grounded in the kinds of explanations different stakeholders actually need (e.g., diagnostic vs. role-based explanations) [25, 40]. **Contestability**, meanwhile, is frequently bolted on as an afterthought (appeals processes, "report a problem" buttons, or generic feedback loops) rather than built into the architecture of reasoning itself. Meaningful contestability requires at least (a) visibility into decision logic,

2026. Manuscript submitted to ACM

(b) comprehensibility for affected parties, and (c) actionable mechanisms for challenge and revision [1]. A system that cannot surface its operative assumptions, show its evidentiary basis, and support structured disagreement cannot plausibly satisfy these conditions—especially in domains where reasonable stakeholders legitimately disagree about values, tradeoffs, and acceptable risk.

1.2 Why post-hoc "explanations" break: the faithfulness problem

A central reason current explainability tooling struggles is that it frequently relies on **post-hoc self-explanation from the same model that produced the decision**. For large language models in particular, chain-of-thought and rationale-style explanations can be fluent and persuasive while remaining weakly coupled to what actually drove the output. Chen et al. benchmark state-of-the-art reasoning models and report low overall faithfulness scores—e.g., **25% for Claude 3.7 Sonnet and 39% for DeepSeek R1** under their evaluation design—highlighting that models may omit or misrepresent key determinants of their answers even when explicitly prompted to "show their work" [6]. Related work similarly emphasizes that CoT can be misleading as an interpretability proxy, especially when users treat it as a reliable window into computation rather than a generated text artifact [37].

This "faithfulness gap" creates a direct accountability failure mode: if the explanation channel can drift from the decision channel, then transparency becomes performative—useful for persuasion, but unreliable for oversight, auditing, or recourse. In high-stakes contexts, that is not a subtle limitation; it is a design-level mismatch between what institutions need (verifiable reasons and traceable evidence) and what monolithic systems can robustly provide. The core implication is architectural: **if we want explanations that can support contestation, we need systems that can produce multiple, checkable reason-giving traces—not a single narrative generated by the same mechanism being explained.** This motivates pluralistic approaches that externalize disagreement, force explicit warrants, and attach provenance to claims so that challenges can target the actual moving parts of the reasoning.

1.3 What we propose (FOAM) and what is new

This paper develops and evaluates **pluralistic AI systems** that operationalize explainability and contestability through **structured multi-agent deliberation** rather than post-hoc narration. We introduce **FOAM (Framework for Openly Augmented Mediation)**, an architecture that treats accountable AI outputs as the product of a mediated process:

(1) **Differentiated agents** with distinct roles and epistemic commitments (e.g., advocate, skeptic, evidence-checker, values/impact assessor),
(2) **Deliberative protocols** that require agents to advance and respond to claims under explicit constraints (e.g., argument typing, cross-examination, and structured rebuttal), and
(3) **Sublation operators**—formal mechanisms for preserving what survives critique while revising what fails, so that the system's final output is not merely an average of perspectives but a documented transformation through contestation.

The intended artifact is not just a recommendation, but a contestable record: claims, counterclaims, evidentiary supports, explicit points of disagreement, and the rationale for any resolution.

We make three contributions:

(1) **Framework:** we provide a unified account of explainability *and* contestability as a single design target, arguing that they should be treated jointly and realized through pluralistic mediation rather than monolithic self-report.

(2) **Architecture and mechanisms:** we formalize FOAM as an implementable blueprint—agents, protocols, and revision operators—paired with provenance-oriented design choices that make challenges actionable (e.g., grounding claims in checkable evidence rather than free-form summarization).

(3) **Empirical validation:** we report results from an evaluation of pluralistic debate generation in a double-blind tournament of **66 policy debate cases**, where our structured multi-agent system achieved an overall score of **81.7** compared to **70.1** for human experts and **50.6** for zero-shot AI, while also achieving **76.2%** perfect evidence validation compared to **8.7%** for human experts and **0%** for unstructured AI—demonstrating that pluralistic architectures can produce outputs that are simultaneously more persuasive *and* more verifiable in an adversarial, evidence-sensitive setting.

We close by discussing implications for AI governance and by outlining a research agenda for **contestable AI by design**.

2 ACCOUNTABILITY REQUIREMENTS AND RELATED WORK

2.1 Explainability requirements beyond transparency

Contemporary calls for "explainable AI" often conflate **transparency** (exposing internal mechanisms) with **explanation** (providing reasons meaningful for a particular audience). Lipton argues that interpretability is not a single property and that many "explanations" function as *post-hoc rationalizations* whose relationship to actual model behavior is ambiguous [24]. Doshi-Velez & Kim emphasize that interpretability claims must be made relative to **use context**—including the user's expertise and stakes—because what counts as satisfactory differs across settings [7], yet few XAI papers explicitly address end-user perspectives [11]. In high-stakes domains, this motivates either inherently interpretable models or explanation mechanisms that achieve *reliability and auditability* rather than superficial plausibility [32].

For accountability, explanations must be **diagnostically useful** and **robust to strategic manipulation**. The NLP interpretability literature distinguishes *plausibility* (does an explanation look reasonable?) from *faithfulness* (does it track the true basis of the output?), arguing that faithful explanations require designs that go beyond "nice-sounding" rationales [14]. Explainability requirements should thus be stated in terms of **checkability**: tracing claims to concrete support and isolating points of disagreement [14, 25].

2.2 Contestability as a system property

Explainability alone does not guarantee meaningful challenge; contestability is best treated as a **system-level governance property**. Alfrink et al. frame "contestable AI by design" as building systems to *support* contestation—through traceability, structured justification, and pathways for challenge—rather than treating contestation as an external process [1]. Legal scholarship similarly emphasizes that decision-subjects need procedures to *question, rebut, and obtain redress* [20]. This matters because the scope of a "right to explanation" under GDPR is contested [39].

Operationally, contestability implies three requirements: (1) **visibility** that an AI-assisted decision occurred; (2) **comprehensibility** of stated grounds; and (3) **actionability**—a pathway to present counterevidence and obtain revision [1, 20]. The EU's Trustworthy AI guidance treats accountability as including mechanisms for redress and capacity to challenge outcomes [9]. These sources motivate a design target: **contestability must be an end-to-end workflow** linking reasons to evidence, rather than a static artifact [29].

2.3 Pluralistic and deliberative approaches to accountability

In high-stakes settings, disagreement is often normative ("which values should dominate?") not merely empirical. Feminist epistemology argues that knowledge claims are situated and that "view from nowhere" objectivity can mask whose assumptions are operationalized [12]. For AI accountability, this motivates an architectural stance: systems should make **value trade-offs explicit** and preserve dissenting considerations in contestable form [25].

Recent work emphasizes that "alignment" is underdetermined when stakeholders disagree about objectives and risks. Kasirzadeh distinguishes alignment approaches that presume a single value target from those treating plural values as first-class constraints [18]. "Society-in-the-loop" framings argue that algorithmic systems require institutionalized interfaces for dispute and revision [28]. These perspectives justify **pluralistic explanation** as a governance mechanism helping stakeholders identify where reasoning depends on contestable assumptions.

2.4 Multi-agent deliberation and debate in AI

A technical pathway to operationalizing pluralism is **structured multi-agent deliberation**. Constitutional AI introduced principle-guided self-critique [4], and Plurals demonstrates that diverse-persona LLM deliberation produces preferred outputs [3]. In AI safety, "debate" was proposed as a scalable oversight mechanism where adversarial argumentation surfaces flaws a single system might hide [13]. Multi-agent debate among LLMs has been reported to improve factuality [8], and recent work demonstrates that debate with more persuasive models helps non-expert judges achieve higher accuracy on difficult questions [19]. However, most results are evaluated in terms of accuracy; they do not guarantee that justifications are **auditable** or that third parties can contest specific premises [14, 32].

Computational argumentation provides complementary foundations via explicit representations of **claims, warrants, attacks, and normative priorities**. Toulmin's model analyzes argument structure in terms of claims supported by warrants and backing [36]. Surveys connecting argumentation and XAI argue these representations support explanation as a structured object of inquiry—stakeholders can contest particular premises and observe how conclusions change [38]. This motivates the claim that a *contestable* AI system should produce a **dispute-ready argumentative record**: reasons decomposed into contestable units, linked to supporting materials, and amenable to revision [20, 23, 38].

3 FOAM APPROACH: PLURALISTIC ARCHITECTURE FOR EXPLAINABILITY AND CONTESTABILITY

3.1 Design goals and accountability threat model

Building on Section 2, we treat *explainability* and *contestability* as properties of an **epistemic process**, not a post-hoc narrative. We introduce **FOAM (Framework for Openly Augmented Mediation)**: a pluralistic, multi-agent architecture producing an answer *plus* a structured record of how it was stress-tested and synthesized. FOAM is organized around three primitives: (i) *differentiated agents* parameterized by explicit stance data structures, (ii) *deliberative protocols* forcing critique and revision, and (iii) *sublation* operators that synthesize without erasing disagreement. Figure 1 provides a system overview.

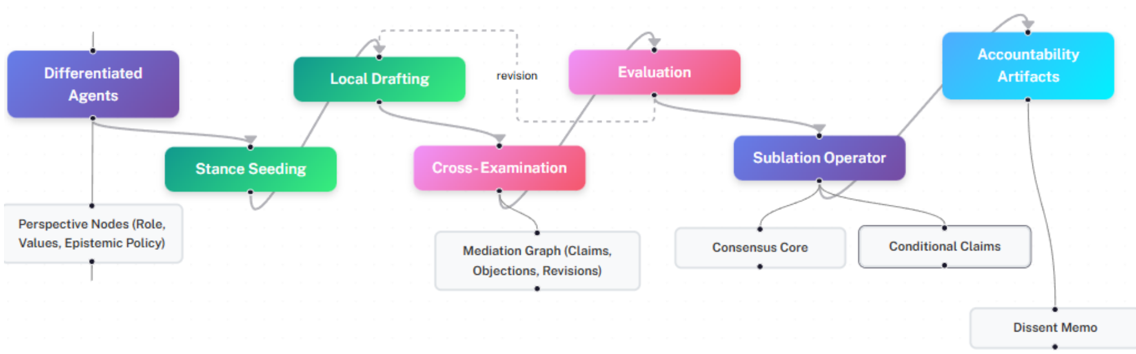

Fig. 1. FOAM system architecture. Differentiated agents with explicit perspective nodes engage in deliberative protocols producing accountability artifacts including a consensus core, conditional claims, and dissent memo.

Our threat model assumes base generative models can (a) produce fluent but false claims ("hallucination") [15], (b) rationalize decisions after the fact [37], (c) collapse multiple perspectives into a dominant frame, and (d) bury value tradeoffs inside unstructured prose. FOAM's core design makes *points of potential failure* explicitly addressable: disagreements are surfaced, objections are first-class objects, and synthesis preserves traceability from contested premises to recommendations.

3.2 Differentiated agents via explicit perspective representation

FOAM instantiates agents each assigned an explicit *Perspective Node* encoding *who the agent is epistemically*—domain role, value priorities, and reasoning schema. This implements "situated" explanation in an auditable way: the system discloses positions and enables critique of *perspective selection* itself [12]. Perspective nodes are operational constraints shaping what evidence is legitimate, which impacts are foregrounded, and which argument schemes are preferred.

A perspective node has three components: (1) **role** (e.g., regulator, clinician, community advocate), (2) **normative weighting** (e.g., safety vs autonomy vs equity), and (3) **epistemic policy** (e.g., acceptable support standards). During deliberation, FOAM enforces *stance coherence*: if generated warrants contradict the declared stance, the system flags the inconsistency.

Perspective nodes enable **second-order contestation**: stakeholders can dispute not only conclusions, but the *legitimacy of the perspective configuration* (e.g., “Why is utilitarian cost-effectiveness in scope here?”). FOAM makes the stance set an explicit input and target for governance [17]. This means FOAM can be rerun with added perspectives, reweighted priorities, or altered evidentiary rules, producing *comparative, contestable* outcomes.

3.3 Deliberative protocol: dialectical refinement and mediation trace

FOAM’s deliberation is a **mediation loop**: (1) *seeding* (instantiate agents + perspectives), (2) *local drafting* (independent proposals), (3) *cross-examination* (structured objections), (4) *evaluation* (scoring draft–objection pairs), and (5) *revision + synthesis*. The accountability point: **deliberation guarantees structured opportunities to find and localize error**, and records what happened when error was raised.

Cross-examination produces a **mediation graph**: a trace linking *which agent* made *which claim*, what objections were raised, how claims were revised, and which survived. This is the audit primitive: stakeholders can point to *the specific node* where they disagree. The trace can be expressed using standard provenance representations (e.g., PROV-O) [22].

3.4 Sublation: synthesis without erasure

After critique, FOAM applies a **sublation operator**: synthesis preserving what is valuable in competing positions while retaining unresolved tensions. Synthesis is disallowed from silently discarding material objections or collapsing incompatible frames into unmarked compromise. Sublation emits three artifacts: **a consensus core** (claims surviving cross-stance critique), **conditional claims** (branching on unresolved priorities), and a **dissent memo** (recording conflicts and contested premises).

3.5 Inspectable argument structure: Toulmin decomposition and typed syllogisms

To make contestation actionable, FOAM constrains outputs into **inspectable argument structure**. We adopt Toulmin-style decomposition—claim, grounds, warrant, backing, qualifier, rebuttal—because it maps to “what can be challenged”: stakeholders can contest evidence, the inferential link, scope conditions, or missing counterevidence [36, 38].

FOAM employs **typed syllogisms**—argument templates enforcing completeness (e.g., Advantage = Uniqueness + Link + Impact). These function as contestability scaffolds: if a stakeholder disputes the conclusion, the system points to the *specific weak component*, and the mediation graph shows whether it was raised in critique [34].

Template tree traversal operationalizes structural contestability. At each branch point, the system records which template was selected (e.g., “traditional 1AC” vs. “kritik”), what resource allocation was applied, and whether novel templates were generated. Stakeholders can dispute not only *what* claims were made, but *why the structure took this form*. Unlike chain-of-thought where reasoning and response are interwoven, template traversal is a discrete prior step serving as foundational infrastructure to drafting.

4 CASE STUDY SYSTEM: EVIDENCE-GROUNDED POLICY DEBATE GENERATION

4.1 Why policy debate is an accountability crucible

We instantiate FOAM in a domain where *contestability is native to the task*: American competitive policy debate. Policy debate is a two-team adversarial format in which teams argue for and against a policy proposal under strict procedural constraints. In this ecosystem, argument quality is not evaluated purely as rhetorical fluency; instead, the activity is structured around *traceable evidentiary support* and explicit clash, so claims can be challenged in real time and revisited across subsequent speeches. Critically, policy debate operationalizes "grounding" through an established evidence artifact: the *debate card*. A card typically includes (i) a short biased summary intended to support a specific argumentative function, (ii) a full citation, and (iii) verbatim quoted source text, often with token-level highlighting that marks precisely what will be read into the round. Competitive success is strongly coupled to evidence quality and its deployment, creating an evaluation environment where provenance and verifiability are not optional.

4.2 Pipeline overview

Figure 2 summarizes our **five-phase pipeline** for generating an evidence-grounded constructive speech (the 1AC, in our evaluation setting). Phases 1–3 produce an inspectable argumentative plan in typed components (perspective assignment → strategic plan → template traversal), Phase 4 binds each argumentative component to *verbatim evidence at sentence granularity* (sentence-level provenance), and Phase 5 compiles and verifies the result (structural conformance, evidence/claim alignment, and perspective consistency). The key design principle is to keep the model in a role where it can be audited: rather than "write a persuasive case and cite sources," the system decomposes "case construction" into a sequence of constrained decisions that leave a machine-checkable trail.

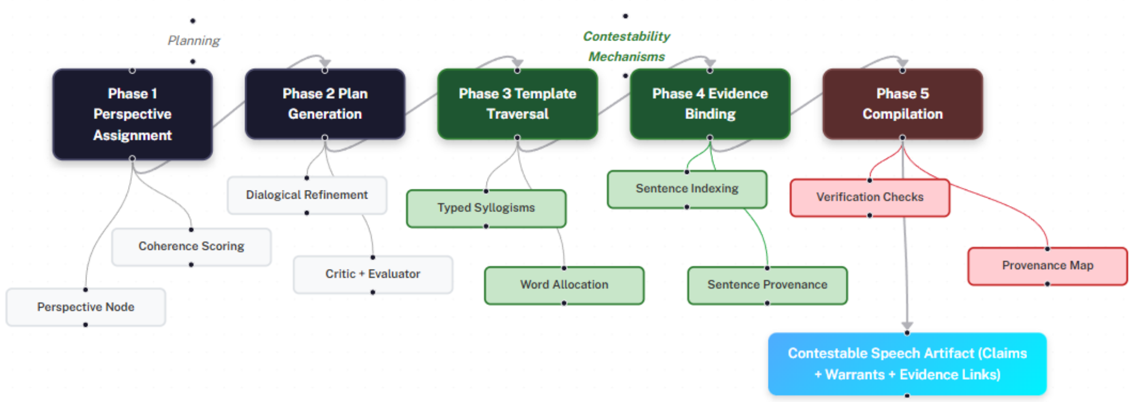

Fig. 2. Five-phase pipeline with accountability mechanisms. Phases 1–3 (Perspective Assignment, Plan Generation, Template Traversal) handle argumentative planning. Phase 4 (Evidence Binding) creates sentence-level provenance by selecting specific sentence IDs rather than paraphrasing. Phase 5 (Compilation) enforces verification checks. The output is a contestable speech artifact with claims, warrants, and traceable evidence links.

4.3 Phases 1–3: perspective assignment, planning, and template traversal

Phases 1–3 produce an inspectable argumentative plan through three contestability-relevant operations. In **Phase 1**, the system assigns an explicit perspective node (Section 3.2), making the evaluative frame a first-class auditable choice. In **Phase 2**, a dialectical refinement loop stress-tests the strategic plan: a Critic agent issues typed objections (logical gap, missing evidence, value conflict, scope overreach), an Evaluator scores each objection's materiality, and the Proposer revises or rebuts. This cycle iterates at least three times, and *all objections—including dismissed ones—remain in the mediation graph*, enabling downstream reviewers to inspect whether a weakness was raised and why the response was deemed adequate.

In **Phase 3**, template tree traversal expands the plan into a typed syllogism scaffold (e.g., Advantage = Uniqueness + Link + Impact). At each branch point, the system records which template was selected, what word allocation was applied (e.g., 30% impact, 40% link), and whether novel templates were generated. This trace enables a distinct class of challenges: stakeholders can dispute not only *what* claims were made, but *why the argumentative structure took this form rather than another*—for instance, contesting that a utilitarian impact calculus was chosen when the underlying values favor a rights-based framing.

4.4 Phase 4: sentence-level provenance

Motivation. Retrieval-augmented generation can reduce hallucinations, but it does not eliminate a central accountability failure mode: models may still produce claims that are *unsupported by*, *in conflict with*, or *misattributed to* retrieved text. Recent benchmarks explicitly document that, even under RAG setups, LLM outputs can contain unsupported or contradictory content relative to the retrieved passages [10]. Phase 4 therefore implements a stronger constraint than "retrieve then paraphrase": it forces the model to operate over *sentence identifiers* rather than free-form rewriting of source material.

Mechanism. Phase 4 is a two-step procedure:

Step (a): sentence indexing and retrieval. The system queries (i) a debate-evidence store (implemented in our current system as a vector database over a large set of debate "cards") and (ii) any other preprocessed sources permitted by the pipeline. Retrieved documents are segmented into sentences, each assigned a stable index, and returned to the deliberation workspace as a set of candidates with identifiers of the form `(document_id, sentence_id)` plus immutable citation metadata.

Step (b): evidence selection and tagging. The LLM is then prompted to (1) select which sentence IDs support each argument slot created in Phase 3 and (2) generate only a short "tag" that states what the selected evidence is being used to establish. Importantly, the model is not asked to restate the evidence; the evidence content in the final speech is assembled from the retrieved sentences themselves. This design eliminates an entire class of failure (fabricated quotations and invented citations) by construction: the model can be wrong about *which* sentences to use, but it cannot invent sentences that are not in the retrieved set.

Accountability and contestability properties. Sentence-level provenance changes the contestation workflow from "argue about what the model meant" to "inspect exactly what the model relied on." A stakeholder can challenge (i) *relevance* ("this sentence does not establish the warrant you claim"), (ii) *adequacy* ("the evidence is too weak/out of context"), or (iii) *selection bias* ("you ignored stronger counterevidence available in the same corpus")—and each challenge targets a concrete object (a sentence ID and its parent source). This is especially aligned with policy debate's evidence norms, which already treat quoted and highlighted text as the unit of disputation under cross-examination.

4.5 Phase 5: compilation and verification checks

Phase 5 compiles the typed argument scaffold (Phase 3) and the evidence bindings (Phase 4) into a final speech artifact suitable for evaluation. Compilation preserves the provenance map: each substantive claim in the rendered speech remains traceable to one or more sentence IDs plus citation metadata. The system then runs verification checks that are directly tied to the accountability requirements:

(1) **Structural completeness** (template validators—e.g., required components are present),
(2) **Evidence/claim alignment** (each slot has at least one bound sentence; missing bindings fail closed), and
(3) **Perspective consistency** (warrants and impacts do not contradict the declared perspective node from Phase 1).

Figure 2 highlights where provenance is created (Phase 4) and where it is enforced (Phase 5).

5 EMPIRICAL EVALUATION

5.1 Research questions

We evaluate FOAM's accountable-generation claims using an *audit-style* design: we define explicit research questions, compare against salient baselines, and report both performance outcomes and traceability outcomes as first-class metrics. This approach aligns with established work on internal algorithmic auditing and emerging "assurance audit" perspectives, which emphasize that accountability requires not only outcome quality, but also artifacts and procedures that make decisions inspectable and challengeable [21, 29].

We ask whether FOAM improves:

- **RQ1:** Quality/persuasiveness
- **RQ2:** Evidence verifiability
- **RQ3:** Whether gains are attributable to the accountability mechanisms rather than model strength

5.2 Experimental design and baselines

Task selection. We evaluate in evidence-grounded policy debate generation because it combines (i) long-horizon argumentative planning, (ii) adversarial robustness expectations (arguments must survive challenge), and (iii) strict evidentiary norms (claims are conventionally supported with citations). In computational argumentation, even highly resourced systems have historically relied on constrained debate settings and bespoke pipelines; the Project Debater line of work illustrates both the ambition of debate as a benchmark and the practical need to structure and constrain the task for reliable evaluation [33].

Debate artifact. We focus on the **first affirmative constructive (1AC)** as the most demanding generative unit in competitive policy debate: it must introduce a full strategic position (advantages/disadvantages/solvency framing), anticipate common lines of negative attack, and do so under tight length constraints while maintaining evidentiary support. This makes the 1AC a strong proxy for high-stakes accountable generation: arguments must be *comprehensible*, *internally coherent*, and *traceable to evidence* to be meaningfully contestable.

Corpus and baselines. We ran a **double-blind tournament of 66 cases** drawn from three sources:

(1) **FOAM-based structured system** ($n = 22$), generated via differentiated perspectives, iterative dialectical refinement, typed syllogisms, and sentence-level provenance;
(2) **Human expert baseline** ($n = 23$), sampled from expert-authored training materials from highly competitive policy debate programs; and
(3) **Zero-shot AI baseline** ($n = 21$), produced by frontier models (Gemini/Claude/ChatGPT/Grok) using prompt engineering and web-research access but without debate-specific pluralistic architecture.

Baseline controls (zero-shot AI). To reduce confounding from artifact format and resource constraints, we generated the zero-shot baseline using Claude 4.5 in research mode, GPT-5 in deep research mode, SuperGrok Heavy, and Gemini 2.5 in research mode. We used a single standardized "mega-prompt" that enforced the same 1AC conventions and constraints used by elite debate program materials and by our FOAM case-building pipeline: **8 minutes of read-time target (1300–1700 words)**; debate formatting (ALL-CAPS tags, short analytic warrants above evidence); a fixed advantage/solvency structure; explicit impact calculus; and comparable evidence-density targets (**3–7 cards per advantage; 2–5 in solvency**). The prompt also enforced a strict **no-fabrication policy**: when reliable bibliographic details and quotations could not be produced, models were required to generate high-precision search strings and to mark uncertainty as [**EVIDENCE NEEDED**]. When the interface supported browsing, web access was enabled to reduce evidence-access confounds. Unlike FOAM, these baselines did not use multi-agent deliberation, typed syllogisms enforcement, or sentence-level provenance binding; thus, baseline citations remained unconstrained natural-language references and were evaluated under the same automated validation pipeline. We generated **one** case per topic per condition and used outputs **as-is** (no manual editing beyond uniform formatting normalization).

Evidence corpus for provenance. FOAM's evidence retrieval and validation leverage a structured debate-evidence corpus derived from OpenDebateEvidence, which (as released) contains **3.5M+** competitive debate documents with metadata useful for downstream argument mining and citation [30]. Operationally, our system queries a vector database of ∼85,000 curated "cards" plus any newly processed sources, and the generation pipeline preserves *sentence-level identifiers* so that downstream reviewers can trace claims to exact supporting spans.

5.3 Judging rubric and scoring

Tournament format and blinding. All submissions were anonymized and assigned unique IDs (e.g., `Case_001`), and judging proceeded purely on content without revealing origin. Cases advanced through a modified Swiss-style bracket with double elimination, and pairings were balanced by strategic approach (e.g., traditional policy vs. kritik) to reduce "judge adaptation" artifacts. Ties within a narrow score band triggered evidence validation as a tiebreaker, keeping accountability-relevant verifiability salient in advancement decisions. **All 66 cases were scored once under the rubric; Tables 1–2 report aggregate statistics over the full set and do not depend on bracket advancement.**

Rubric and judge. Following established LLM-as-judge methodology [41], a Claude Opus 4 judge evaluated each case on five weighted dimensions:

- **Argumentation Strength** (25%)

- **Evidence Quality** (25%)
- **Strategic Coherence** (20%)
- **Innovation** (15%)
- **Competitive Viability** (15%)

The rubric was designed to reward both argumentative competence and evidence-groundedness, while preserving enough structure for reproducibility.

5.4 Evidence validation methodology

Why evidence validation is an accountability metric (not just "anti-hallucination"). In contestable systems, stakeholders must be able to *locate* and *evaluate* the grounds of a claim—especially where persuasive language can obscure weak or missing support. Audit frameworks similarly emphasize that assurance depends on traceable evidence artifacts rather than outcome plausibility alone [21, 29]. We therefore operationalize verifiability as a measurable property of each case's citations.

Automated citation checks and categories. Each citation was automatically checked against the referenced source (via URL or resolvable reference), and classified into one of four buckets: **exact match**, **partial match**, **paraphrase**, or **fabricated**. We summarize results primarily via **Perfect Validation**, a stringent metric that counts only **exact matches**—i.e., the cited claim can be located verbatim in the referenced source span. This is intentionally conservative: Perfect Validation corresponds to the strongest form of contestability, where an affected party can directly inspect the cited text without interpretive debate about semantic similarity.

How FOAM changes the validation problem. FOAM's sentence-level provenance changes citation validation from a semantic retrieval problem into a *pointer integrity* problem: the model is never asked to reproduce source text, but instead selects sentence indices from retrieved documents and attaches them to specific argument components. This design greatly reduces degrees of freedom for fabrication and enables deterministic re-checking of a case's evidentiary backbone.

5.5 Results

Main tournament outcomes. Table 1 reports aggregate performance by source. The FOAM-based system achieved the highest overall score (**81.7**) relative to human experts (**70.1**) and zero-shot AI (**50.6**). The largest gap appears in **Evidence Quality** (**86.7** vs. **56.9** vs. **27.1**), consistent with the claim that provenance-constrained generation shifts the system from persuasive-but-unreliable outputs toward persuasive-and-grounded outputs.

Table 1. Tournament Results by Source

Metric	FOAM	Human Expert	Zero-shot AI
Overall Score	81.7	70.1	50.6
Evidence Quality	86.7	56.9	27.1

Evidence validation and verifiability. Table 2 reports Perfect Validation rates. FOAM achieved **76.2%** Perfect Validation, compared to **8.7%** for the human expert baseline and **0%** for zero-shot AI. This is the central accountability result: the FOAM pipeline does not merely produce arguments that a judge model rates as "good," but produces arguments whose evidentiary support can be mechanically verified at scale.

Table 2. Perfect Validation Rates

Source	Perfect Validation (%)
FOAM System	76.2
Human Expert	8.7
Zero-shot AI	0.0

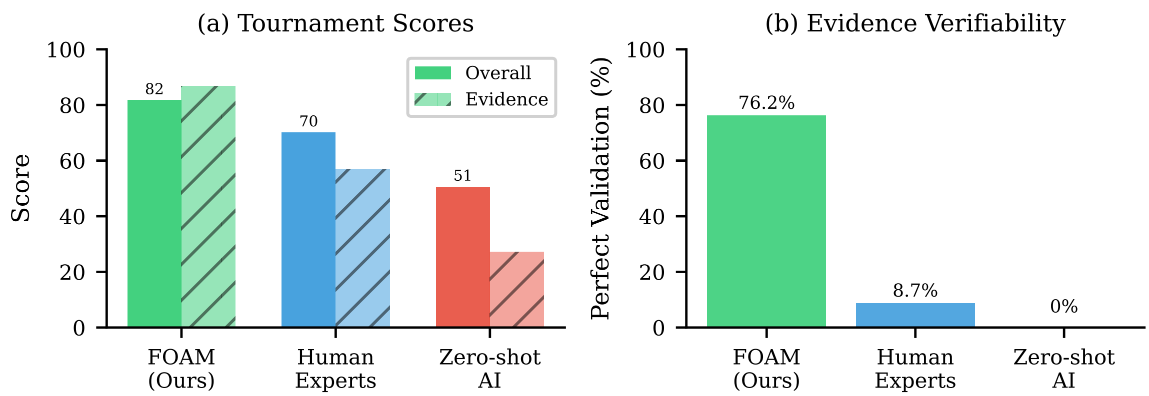

Fig. 3. Tournament results comparing FOAM, human expert baselines, and zero-shot AI. (a) Overall and Evidence Quality scores. (b) Perfect Validation rates—the percentage of citations that exactly match source text. FOAM achieves 76.2% perfect validation vs. 8.7% for human experts and 0% for zero-shot AI.

Interpreting what is doing the work. Two mechanisms plausibly drive the observed gap: (i) **pluralistic deliberation** (multi-perspective critique and refinement) improves strategic coherence and argument coverage, while (ii) **sentence-level provenance** directly improves evidence integrity and sharply limits fabrication opportunities. Several high-scoring FOAM cases achieved perfect validation (fidelity = 1.0), indicating that high persuasive quality and high verifiability can co-occur under the FOAM constraint regime.

6 IMPLICATIONS FOR ACCOUNTABLE AI SYSTEMS

FOAM reframes explanation as a contestable record rather than a post-hoc narrative. Instead of producing a single rationale, the system outputs (i) an auditable argument structure (claims, warrants, rebuttals), (ii) explicit perspective configurations, and (iii) sentence-level provenance linking each substantive claim to a checkable source span. This shifts accountability from "did the explanation sound plausible?" to "which premises and evidence does the output depend on, and where can a challenge be lodged?"

Operationally, FOAM supports contestation at three levels [1]: (1) **evidence disputes** (a cited sentence does not support the tagged claim; missing counterevidence), (2) **inferential disputes** (the warrant connecting evidence to conclusion is invalid or incomplete), and (3) **normative disputes** (the perspective/value configuration is illegitimate or incomplete for the context). Because these objects are explicit, a reviewer can localize disagreement to specific nodes and request revision without reopening the entire output as free-form prose.

Institutionally, the resulting artifact functions as an auditable dossier that can support downstream review within existing governance workflows (internal review, incident response, assurance audits); dispute resolution itself requires institutional process beyond what the technical system provides. The technical contribution is not replacing due process, but supplying the structured, traceable materials that make procedural review feasible at scale.

7 LIMITATIONS AND FUTURE WORK

7.1 Methodological limitations and validity threats

First, our primary outcome measure relies on an automated judge (Claude Opus 4) to score debate artifacts under a fixed rubric. While LLM-as-judge evaluation is increasingly standard at scale, it is known to exhibit systematic biases (e.g., position effects, verbosity/style sensitivity, and self-enhancement tendencies) and may be vulnerable to prompt- or framing-based perturbations that shift preferences without corresponding semantic differences [5, 35, 41]. We reduce—but do not eliminate—these threats via double-blinding, standardized prompts, and by pairing judge scores with an independent evidence-validation audit. Nevertheless, the reported tournament results should be interpreted as descriptive for this evaluation setup, and future replications should triangulate across multiple judge models and human adjudication.

Second, our system's accountability guarantees are conditioned on the properties of the underlying evidence substrate. Sentence-level provenance constrains the model to point to specific source sentences rather than inventing citations, but it does not ensure that the retrieved evidence is complete, representative, or up to date. Coverage gaps, topical skew, and retrieval errors can shape which arguments are discoverable, and can yield outputs that are "well-cited" yet misleading due to selection effects, over-aggregation, or missing context [31]. These concerns are not unique to debate generation: any contestability mechanism built on curated corpora inherits the corpus' blind spots. Accordingly, FOAM should be viewed as an approach to making claims auditable and challengeable—not as a guarantee that the selected evidence is normatively "best" or epistemically sufficient.

Third, our evaluation scope is intentionally narrow and therefore limits external validity. We benchmark a specialized argumentative domain (policy debate) and a bounded artifact type (constructive case generation), and we do not yet measure downstream stakeholder contestation behaviors (e.g., whether affected parties can efficiently detect, understand, and successfully challenge specific warrants or citations). Additionally, our "perfect validation" metric is strict by design: it favors verbatim traceability and can under-credit faithful paraphrase or correct claims supported by multiple dispersed sentences. Conversely, the metric may fail to detect other fidelity failures (e.g., cherry-picked quoting or context stripping) that require richer contextual checks. These are appropriate trade-offs for an audit-style evaluation, but they motivate follow-on studies with complementary human-centered and context-sensitive validation protocols.

7.2 Safety and misuse considerations

Systems optimized for persuasive argumentation can be dual-use; we address misuse risks, affected groups, and mitigations in the Adverse Impacts statement (Endmatter).

7.3 Future work

A first priority is human-subject evaluation of contestability as an interaction property rather than a static artifact property. We plan controlled studies in which participants (including domain experts and affected stakeholders) attempt to (i) locate supporting evidence for a contested sentence, (ii) challenge a warrant or inference step, and (iii) request or compare alternative perspective nodes. Primary outcomes should include time-to-challenge, challenge success rates, perceived procedural fairness, and the degree to which the system supports actionable revision pathways (e.g., retracting a claim, swapping evidence, or surfacing counter-arguments) rather than merely producing longer explanations.

A second priority is extending FOAM with optimization and training methods while preserving contestability constraints. Preliminary results in iterative preference learning suggest that tactic selection and evidence integration can be improved, but also reveal failure modes that matter for accountable deliberation. Future work should explore training objectives that explicitly reward faithful warrant-evidence alignment (not only persuasiveness) and contestation-aware curricula.

8 CONCLUSION

High-stakes deployments of LLM-based systems demand more than *transparent-seeming* narratives; they require explanations that can be *challenged, audited, and revised*. Recent evidence suggests that post-hoc "reasoning traces" are often not a reliable proxy for what drives model behavior: when a prompt-injected hint changes a model's answer, state-of-the-art reasoning models reveal that hint in their chain-of-thought only about **25–39%** of the time, indicating substantial unfaithfulness of verbalized rationales to causal drivers of outputs [6]. This paper contributes (1) **FOAM**, a pluralistic deliberation architecture for explainability-and-contestability-by-design; (2) an **inspectable provenance mechanism** that makes sentence-level claims traceable to source spans and contestable at the level stakeholders actually dispute; and (3) an **audit-style empirical evaluation** in evidence-grounded policy debate generation. In a double-blind tournament of 66 cases, the FOAM-based system achieves higher overall scores than expert-human and zero-shot baselines (Table 1) and dramatically higher perfect evidence validation rates (Table 2), demonstrating that accountable generation can be simultaneously *high-quality* and *verifiable*.

For the FAccT community, the central implication is a practical shift from explanation-as-disclosure to **contestable explanations**: outputs whose *claims, warrants, and evidence links* are explicit, inspectable, and designed to invite targeted challenge (e.g., disputing a cited sentence, contesting a warrant, or requesting an alternative perspective node). This orientation is consistent with due-process motivations for a meaningful right to contest consequential automated decisions [16]. Where governance requires reason-giving that can withstand scrutiny, pluralistic deliberation plus verifiable provenance offers a concrete design pattern for building AI systems whose decisions can be examined, contested, and improved without relying on "black-box" rationalizations.

ENDMATTER

Generative AI Usage Statement

This research investigates the use of large language models (LLMs) within a structured multi-agent deliberation framework. The FOAM system described in this paper uses LLMs as components within the deliberation pipeline. The paper text itself was drafted by human authors with AI assistance limited to copy-editing and formatting suggestions. All substantive claims, experimental design, and analysis reflect human judgment and interpretation.

Ethical Considerations

This work develops AI systems with persuasive capabilities, which raises dual-use concerns. We address these in Section 7 and Section 6, discussing safeguards including transparency requirements, evidence provenance constraints, and the deliberate choice to evaluate in a domain (competitive debate) with established norms for scrutinizing persuasive claims. The evaluation involved no human subjects; all baselines were drawn from publicly available debate materials or generated outputs.

Adverse Impacts Statement

Systems that generate persuasive, evidence-grounded arguments could be misused for misinformation, manipulation, or to overwhelm human review capacity. Affected groups include decision-subjects in high-stakes domains and information consumers generally. We mitigate these risks through: (1) provenance requirements that make claims auditable; (2) evaluation in a domain with adversarial scrutiny norms; (3) architectural transparency (the deliberation trace is inspectable). Deployment in sensitive domains should include access controls, logging, human oversight, and institutional review processes.

REFERENCES

[1] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica.

[3] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2024. Plurals: A system for guiding LLMs via simulated social ensembles. In *CHI*.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073* (2022).

[5] Guiming Chen et al. 2024. Humans or LLMs as the judge? A study on judgement bias. *arXiv preprint* (2024).

[6] Yanda Chen et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint* (2025).

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. In *ICML Workshop on Human Interpretability in Machine Learning*.

[8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *arXiv preprint arXiv:2305.14325*.

[9] European Commission High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI.

[10] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. RARR: Researching and revising what language models say, using language models. In *ACL*.

[11] Abdul KM Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable AI from the user perspective: A systematic literature review. In *Pacific Asia Conference on Information Systems*.

[12] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.
[13] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. In *arXiv preprint arXiv:1805.00899*.
[14] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Proceedings of ACL*. 4198–4205.
[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
[16] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.
[17] Atoosa Kasirzadeh. 2024. Plurality of value pluralism and AI value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*. https://openreview.net/forum?id=AOokh1UYLH
[18] Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology* 36, 2 (2023), 1–23.
[19] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kris Sachan, Ansh Raber, Arthur Gretton, et al. 2024. Debating with more persuasive LLMs leads to more truthful answers. In *ICML*. Best Paper Award.
[20] Joshua A Kroll, Joanna Huey, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165 (2017), 633.
[21] Khoa Lam et al. 2024. Assurance audits for AI systems. In *Proceedings of FAccT*.
[22] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. PROV-O: The PROV ontology. W3C Recommendation.
[23] Francesco Leofante and Francesca Toni. 2024. Contestable AI needs computational argumentation. In *Proceedings of KR*.
[24] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.
[25] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
[26] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342
[27] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of FAccT*. 469–481. https://doi.org/10.1145/3351095.3372828
[28] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[29] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of FAccT*. 33–44.
[30] Allen Roush et al. 2024. OpenDebateEvidence: A massive-scale dataset for argument mining and summarization. *arXiv preprint* (2024).
[31] Allen Roush et al. 2025. A superpersuasive autonomous policy debating system. *arXiv preprint* (2025).
[32] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
[33] Noam Slonim et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384.
[34] Alfred C Snider. 2008. *Code of the debater: Introduction to policy debating*. IDEA Press Books.
[35] Aman Singh Thakur et al. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv preprint* (2024).
[36] Stephen E Toulmin. 1958. *The uses of argument*. Cambridge University Press.
[37] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*.
[38] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: A survey. *Knowledge Engineering Review* 36 (2021).
[39] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.
[40] Yuan Yao. 2024. Explanatory pluralism in explainable AI. *AI Magazine* 45, 1 (2024), 82–93.
[41] Lianmin Zheng et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *NeurIPS*.