A Study of Self-Verifying Multi-Agent AI for Reliable Decision Making

Dr. K. Andal[1], Abarna S[2], Anbarasi M[3], Angelin Jelis S[4], Deepalakshmi G[5]
[1]*Ph.D., Associate Professor, Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Puducherry, India*
[2,3,4,5]*Student, Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Puducherry, India*

***Abstract*—Large language models (LLMs) can produce fluent but inconsistent or weakly supported answers, creating a need for approaches that make disagreement and uncertainty easier to examine. This study considers how multi-agent generation, evidence assessment, and abstract argumentation may contribute to more reliable and explainable decision support. It synthesizes literature on requirements engineering, generative AI, argumentation, and quality and safety considerations, then develops a conceptual framework for comparing claims, evidence, conflicts, and uncertainty.**

The study distinguishes three questions that are often conflated: whether a claim is supported by evidence, whether an argument is accepted under a formal attack relation, and whether a response is preferable under stated decision criteria. Dung's grounded semantics can organize the second question, but cannot independently establish factual truth. As a conceptual study, the paper offers an analytical synthesis and identifies risks and questions for future research, without making empirical performance claims. Its central conclusion is that agent agreement and formal argument status should be treated as review signals, while evidence quality, unresolved disagreement, and human judgment remain explicit.

***Index Terms*—large language models, multi-agent systems, abstract argumentation, grounded semantics, explainable AI, decision support, claim provenance.**

I. INTRODUCTION

Large language models (LLMs) are used to answer questions, summarize information, draft explanations, and support decisions. Their fluency can make unsupported statements appear as persuasive as well-grounded ones. Errors may involve an individual fact, a relationship between facts, a missing qualification, or a recommendation that overlooks a constraint. In settings where a user must inspect or act on an answer, a plausible response alone is insufficient: the user also needs to understand what was compared, what evidence was considered, where disagreement remains, and when the system cannot make a reliable choice.

Prior research considers generative AI in requirements engineering, including elicitation, analysis, and multi-agent collaboration [10], [11], [20]– [22], [27]. Argumentation offers a complementary lens for representing competing claims and reasons [12], [18], [28]. These strands suggest that reliability depends not only on generating alternatives, but also on clarifying what counts as evidence, how conflicts are interpreted, and how uncertainty is communicated.

This paper is a conceptual study of self-verifying multi-agent AI for reliable decision making and does not claim experimental validation. The contribution is an analytical synthesis that separates claim-level evidence assessment from graph-based argument status and from candidate preference. The distinction matters because a formally accepted argument can still rely on incomplete or misinterpreted evidence, and a preferred answer can remain uncertain.

The study addresses four questions: (1) what roles can independent generation and critique play in examining LLM answers; (2) how can claims, evidence, and conflicts be represented without treating disagreement as proof; (3) what limitations arise when argumentation semantics is used to communicate answer status; and (4) which evaluation dimensions are needed to assess reliability and explainability? These questions guide a qualitative discussion and an evaluation agenda; the paper reports no experiment or measured outcome.

II. BACKGROUND AND RELATED WORK

A. Abstract Argumentation

Dung's Abstract Argumentation Framework represents a set of arguments and a binary attack relation between them [12]. Given a set of arguments A and an attack relation R, the framework is AF = (A, R). Grounded semantics produces a unique, conservative extension: accepted arguments are defended against all attackers, arguments attacked by accepted arguments are rejected, and remaining arguments are undecided. Recent work explores argumentation for conflict resolution in language models [18], while value-based and bipolar frameworks examine richer relationships among arguments [28], [29]. These formalisms help organize disagreement, but their conclusions remain conditional on how arguments and attacks are identified.

An argument's formal status does not verify a source's authenticity, relevance, freshness, or interpretation. Evidence assessment and argumentation therefore answer related but distinct questions: evidence review assesses what a source supports, while an argumentation framework assesses how represented positions interact. Maintaining this distinction is essential to avoid presenting a graph label as factual certification.

B. Multi-Agent Generation and Verification

Multi-agent approaches to requirements elicitation and analysis investigate how agents can contribute complementary interpretations, negotiate conflicts, and refine requirements [10], [11], [19], [27]. Broader reviews of generative AI in requirements engineering describe opportunities as well as limitations in using LLMs for elicitation and analysis [20]–[22]. This literature motivates structured comparison, but the presence of multiple agents does not guarantee independent evidence or correct conclusions. Agents may share model families, training data, prompts, and source material. Agreement should therefore be treated as an observation rather than proof, and disagreement should prompt examination of claims and evidence rather than a simple vote.

C. Quality and Safety Context

Requirements and quality research emphasizes that dependable outcomes depend on explicit goals, stakeholder needs, and trade-offs [1], [3], [4], [6], [17], [23]. Safety-oriented standards and regulation provide further context for the limits of automated decision support [13], [14]. For multi-agent AI, these perspectives imply that answer quality should not be reduced to fluency or consensus: relevance, completeness, evidence support, consistency, uncertainty, and consequences all matter. No single quality dimension or benchmark can establish reliability across unrelated tasks.

D. Requirements Engineering as a Decision Context

Requirements engineering offers a useful context for studying reliable multi-agent reasoning because it routinely involves incomplete information, competing stakeholder goals, and statements whose meaning depends on scope and context. Elicitation approaches gather and reconcile needs from different perspectives, while requirements analysis examines consistency, feasibility, and priority [1], [17], [23], [24]. These activities resemble the challenge of comparing candidate LLM answers: each response may capture a different aspect of the question, and apparent disagreement may reflect distinct assumptions rather than a direct contradiction.

Negotiation research emphasizes that conflicts are not always resolved by choosing the most frequently stated position. Stakeholders may assign different values to quality attributes, accept different risks, or use different contextual assumptions [26]. A multi-agent study can therefore treat disagreement as evidence of an unresolved decision structure. The analysis should identify which claims conflict, which interests or assumptions explain the conflict, and which evidence could help resolve it. This framing is more informative than a majority vote because it preserves the reasons behind competing conclusions.

The requirements perspective also clarifies why scope and provenance matter. A statement can be valid for one user group, operating environment, or time period and invalid for another. If such qualifiers disappear during summarization, later reasoning may produce a false contradiction or an unjustified consensus. An analysis that records claim context and its source is therefore better positioned to distinguish incompatible facts from differences in scope. These observations motivate evaluation at the level of claims and relations, rather than only at the level of a final answer.

E. Generative AI and Multi-Agent Analysis

Recent research examines the use of generative AI in requirements elicitation, prompt design, and analysis [8]–[11], [19]–[22]. This work spans both the assistance of individual analysts and collaboration among multiple agents. Multi-agent approaches are relevant when a task benefits from different roles, such as proposing interpretations, identifying missing information, or challenging assumptions [10], [11], [27]. The literature also motivates caution: an agent's contribution depends on its prompt, context, and information access, and a fluent response may still omit constraints or introduce unsupported details.

Prompting strategies can shape the structure of a task and encourage an LLM to ask questions, compare alternatives, or explain a rationale [8], [9]. Yet prompt structure is not equivalent to independent verification. If multiple agents receive the same incomplete information or share underlying model characteristics, their agreement may reproduce a common error. Conversely, disagreement can be useful even when no agent is correct, because it can reveal ambiguity or missing evidence. The analytical value of plurality therefore depends on whether outputs are compared against sources and explicit criteria rather than counted.

Multi-agent work in requirements analysis also suggests a connection between critique and negotiation. A critic can challenge an assumption, while another perspective can defend it or qualify its scope. Argumentation offers a way to represent these exchanges without treating every disagreement as equivalent. However, agent roles and interaction rules can influence which arguments are generated and which are overlooked. A study of the approach should consequently report the reasoning protocol and examine how outcomes change when agent roles or available evidence change.

F. Argumentation, Rationale, and Contestability

Argumentation and design-rationale research provides concepts for making a decision's grounds inspectable [15], [16], [25], [28]–[30]. A rationale may include a claim, supporting reasons, objections, and the context in which a choice is preferred. This is especially important when a decision is contestable: a reviewer should be able to identify not only the selected conclusion but also the strongest alternatives and the basis for rejecting them. Abstract argumentation formalizes one part of this task by representing arguments and attacks, while other frameworks consider support, values, or bipolar relations [12], [28], [29].

Contestability should not be confused with exhaustive explanation. A concise account can be useful if it faithfully identifies the decisive claims, relevant evidence, and unresolved questions. A longer rationale can still mislead if it omits important qualifications or presents a generated explanation as privileged access to internal model reasoning. Evaluation should therefore ask whether people can reconstruct the basis of a judgment and detect weaknesses, not merely whether they rate an explanation as clear or persuasive.

G. Synthesis of the Literature

Across these areas, four themes recur. First, requirements and decision contexts must be made explicit because relevance and priority depend on stakeholder goals [1], [3], [4]. Second, generative AI can support elicitation and comparison, but its outputs require scrutiny and the diversity of agents does not establish independence [8]–[11], [20]–[22]. Third, argumentation can make support and conflict structures visible, but formal labels are conditional on the graph and do not establish factual truth [12], [18], [28]–[30]. Fourth, quality and safety depend on transparent trade-offs, evidence, and human accountability [2], [6], [7], [13], [14].

These themes suggest that self-verification is better understood as a layered study problem than as a single correctness score. A robust analysis must consider the quality of candidate claims, the relevance of evidence, the validity of attack relations, the stability of preferences under reasonable changes in criteria, and the effect of explanations on human judgment. This synthesis informs the conceptual framework and future evaluation dimensions developed in the following sections; it does not assert that any particular configuration has already achieved these outcomes.

III. CONCEPTUAL SCOPE AND ANALYTICAL FRAMEWORK

A. Scope and Assumptions

This study concerns informational question answering and decision support, not autonomous clinical, legal,

financial, or other high-impact authority. It treats multi-agent verification as a conceptual process in which alternative responses are examined through claims, evidence, and explicit relations. The analysis does not assume that agent outputs are independent, that majority agreement is reliable, or that citations supplied in generated text are valid. Formal argument labels describe a position in a constructed graph, not factual truth. When evidence is absent, sources conflict, or alternatives cannot be distinguished, uncertainty should remain explicit.

B. Claims, Evidence, and Provenance

Let q denote a question and Cq a set of candidate's responses considered for it. A response can be decomposed into material claims, each linked to the response span from which it was derived. Evidence review may classify a claim as supported, refuted, unresolved, or not assess-able. These labels concern the relationship between the claim and the available sources; they are not argumentation labels. Provenance preserves which response and source passage motivated an interpretation, making omissions and scope shifts easier to inspect.

C. Analytical Questions

The study uses the following questions to organize com-parison:

1. What material claims and qualifications does each response make?
2. Which claims have checkable support, and what uncertainty remains?
3. Which arguments support or attack other arguments, and why?
4. Which response best meets the stated criteria, or should the comparison remain inconclusive?

IV. CONCEPTUAL ANALYSIS OF CANDIDATE ANSWERS

A. Independent Candidate Generation

Multiple candidate answers can reveal different interpretations, assumptions, and factual claims. For a meaningful comparison, candidates should initially be produced independently; otherwise, later answers may merely repeat the first response. However, distinct agents or providers do not necessarily provide independent evidence. They may share training material, prompting conventions, retrieved sources, or model components. Agreement is therefore a signal for further examination, not a substitute for corroboration. Comparative analysis should consider the provenance and limitations of each response as well as its content.

Differences between candidates should be characterized rather than reduced to a vote. A disagreement may concern the central answer, a qualification, the time or population to which a claim applies, or the interpretation of a source. Some differences are stylistic and immaterial; others affect the decision. This motivates claim-level analysis that preserves the exact statement and its context. Failure to obtain a candidate, or an incomplete answer, should be understood as missing information rather than evidence for another answer.

B. Argumentation and Explainable Selection

An analytical graph can be represented as H = (A, S, R), where A is a set of arguments, S is a support relation, and R is a directed attack relation. A support edge records that an argument offers a reason or evidence for another. An attack edge (a_i, a_j) ∈ R records a justified challenge to a_j, such as an incompatible claim, an evidence conflict, a scope mismatch, or an invalid inference. Mere disagreement between candidates is not by itself sufficient to justify an attack edge.

The corresponding Dung framework is AF = (A, R) [12]. For any set of arguments X ⊆ A, its characteristic function is:

$$F(X) = \{ a \in A : \text{each attacker } b \text{ of } a \text{ is attacked by some } c \in X \} \quad (1)$$

The grounded extension is the least fixed point of F. An argument in this extension is labelled accepted; an argument attacked by an accepted argument is rejected; and any remaining argument is undecided. Support links can be shown in the explanatory graph but are not silently treated as attack edges in the Dung framework.

All three labels are relative to the graph and semantics and do not independently certify factual correctness. The grounded extension can be constructed iteratively from $G_0 = \emptyset$ and $G_{t+1} = F(G_t)$ until the sets stop changing. The finite argument set guarantees that this process terminates. The resulting extension is unique for a given attack graph, but changes to one claim or edge can change the labels of other nodes, so a

correction requires recomputation of the whole affected framework.

Table 1: Evidence judgments and argument labels are distinct

Label family	Example values	Question answered
Evidence reviews	Supported, refuted, unresolved, not assessable	What does the checked source indicate about the claim?
Grounded semantics	Accepted, rejected, undecided	What status follows from the constructed attack graph?

These labels answer different questions. An argument may be accepted in the graph while the evidence for its factual claim remains unresolved; that combination indicates uncertainty, not confirmation. Keeping evidence judgments distinct from formal labels reduces the risk of overstating what an argumentation analysis establishes.

Candidate comparison can consider evidence support, logical consistency, question coverage, unresolved conflicts, and unsupported material claims. The relative importance of these criteria depends on the decision context and should be stated openly. If evidence is insufficient or the criteria do not distinguish alternatives, an inconclusive finding is more defensible than an unjustified winner.

C. Claim Extraction and Evidence Provenance

Claim analysis identifies propositions that materially affect an answer, recommendation, or interpretation. Compound sentences may need to be separated when one part can be supported while another is contradicted. Exact text and context matter: a missing qualifier that changes a recommendation is more consequential than a stylistic variation. Analyses should therefore preserve enough provenance for a reviewer to compare an interpreted claim with the statement from which it was derived.

Evidence review asks whether a cited source can be located and whether its content is relevant to the claim's scope. A generated citation is a pointer to investigate, not evidence until it has been checked. Claims without checkable support remain unresolved or unsupported under the stated review criteria. When sources conflict, the disagreement should be acknowledged rather than resolved by selectively favoring one candidate. This layer supports factual review but remains separate from formal argument labels.

Relation identification requires explicit criteria. A contradiction should be recorded only when propositions are incompatible under their stated scope and time. A scope challenge may target an overgeneralized claim without refuting a narrower version; an undercut may challenge an inference while leaving its source intact. Stating why an edge is warranted makes the analysis reviewable and helps distinguish substantive conflict from superficial wording differences.

D. Comparative Criteria and Uncertainty

For a candidate c, let E(c) measure the proportion of material claims with adequate evidence, L(c) measure consistency among its material claims, and Q(c) measure coverage of the question's required aspects. Let U(c) measure unresolved material conflicts and N(c) measure material claims lacking adequate support. Larger values are desirable for E, L, and Q; smaller values are desirable for U and N. Exact rubrics and thresholds should be specified for the task under study.

A study comparing candidates may use a declared scalar summary such as:

$$J(c) = w_E * E(c) + w_L * L(c) + w_Q * Q(c) - w_U * U(c) - w_N * N(c) \quad (2)$$

where nonnegative weights encode task priorities. This expression is a conceptual example, not an empirically validated scoring rule. A vector or lexicographic comparison may be preferable when criteria are not commensurable. Critical-claim constraints can take precedence over aggregate quality so that numerous minor supported details do not mask one consequential unsupported assertion. If no candidate meets the stated conditions, or leading candidates remain indistinguishable, the conclusion should remain inconclusive.

E. Explanation and Contestability

An explanation should identify the claims that influenced a comparative judgment, the evidence considered, the decisive support or attack relations, and the remaining uncertainties. It should distinguish observations from interpretations and avoid implying

access to a model's private internal reasoning. A reviewer should be able to challenge a claim, source, or relation; disagreements and corrections are part of the analysis rather than nuisances to conceal. This emphasis on rationale and negotiation is consistent with research on design rationale and requirements negotiation [15]–[17], [26].

V. ANALYTICAL SYNTHESIS

A. Comparison Sequence

A conceptual comparison can be organized into the following analytical steps:

1. Clarify the question, its scope, and the criteria relevant to answering it.
2. Compare independently generated responses and identify their material claims and qualifications.
3. Examine claims against relevant sources, noting support, contradiction, and unresolved evidence.
4. Represent justified support and attack relations, giving a rationale for each relation.
5. Apply the selected argumentation semantics and interpret its labels relative to the graph.
6. Compare the alternatives under explicit criteria and report a preference or an inconclusive result.

This sequence is a study framework, not a claim that each step can be automated reliably. Missing or weak evidence limits the comparison but does not itself determine an argument's formal label. Similarly, a graph label cannot repair a claim-extraction error or establish that a cited source is authoritative.

B. Ambiguity and Revision

A comparative judgment should identify which criteria influenced it and which issues remain unresolved. A single composite score is not always appropriate: criteria may conflict, evidence may be incomplete, or alternatives may be indistinguishable. In such cases an inconclusive conclusion is preferable to false precision. Any revision to a claim, source interpretation, or attack relation can change the resulting labels and should therefore be treated as a substantive analytical change, not a cosmetic update.

C. Illustrative Reasoning Example

Consider a question asking which of two software releases introduced a documented feature, with release notes available. Two candidate answers state that the feature first appeared in release 2 and release 3. A source passage may support the first claim while a scope qualification challenges the second. The comparison should distinguish the claims' exact scope and the source's relevance before assigning an attack relation. If the release notes do not resolve the conflict, the conclusion should remain uncertain. This example illustrates the analytical framework; it is not an empirical finding.

D. Illustrative Grounded Labelling

Let a_1 state that a feature first appeared in release 2, a_2 state that it first appeared in release 3, and a_3 represent a passage that supports the release-2 claim and challenges the scope of a_2. For this schematic graph, the attack relation is $R = \{(a_1, a_2), (a_2, a_1), (a_3, a_2)\}$. The passage's authenticity and interpretation still require review; the graph illustrates only how statuses follow from represented attacks.

The grounded construction begins with unattacked arguments. Here a_3 is unattacked and enters the grounded extension. Since it attacks a_2, that argument is rejected. The attack from a_2 against a_1 is counterattacked by accepted a_3, so a_1 is defended and is accepted in the next iteration. These labels are consequences of the graph, not an independent judgment that the release note is true or authoritative. If its interpretation or the scope relation were mistaken, the graph and its labels would require reassessment.

VI. LIMITATIONS AND ANALYTICAL RISKS

Reliable interpretation requires that confidence in one analytical layer not conceal errors in another. Table 2 summarizes principal risks and conservative responses.

Table 2: Failure modes and conservative responses

Analytical risk	Possible indication	Appropriate response
Claim omission or merger	A material statement is absent, or distinct claims are combined.	Preserve context, acknowledge extraction limits, and seek independent review.
Unverified citation	A cited source or passage cannot be located or checked.	Treat support as unresolved; citation appearance alone is insufficient.

Conflicting evidence	Relevant sources support incompatible scopes or conclusions.	Explain the disagreement and retain an inconclusive judgment where needed.
Missed or spurious attack	A relevant conflict is absent, or a relation lacks justification.	Reassess the relation and the resulting argument labels.
Correlated agents	Candidates repeat shared assumptions or sources.	Avoid treating agreement as independent corroboration.
Privacy or misuse risk	Sensitive inputs or high-impact claims lack appropriate review.	Minimize exposure and require qualified human judgment.

Analysis should distinguish an incorrect conclusion from a correct conclusion reached through an invalid argument. A claim may happen to be true even if its apparent evidence is fabricated or irrelevant. Evaluation should therefore consider answer quality and intermediate judgments, including extraction coverage, evidence relevance, relation quality, formal labelling, and the stated rationale. Agreement among agents should not be counted as correctness because they can share sources and failure modes.

Particularly informative cases include conflicting evidence, claims whose truth depends on a qualification, cyclic attack graphs, and source material that contains instructions unrelated to the factual question.

Such cases expose how scope, ambiguity, and untrusted content affect interpretation. Grounded semantics may leave arguments undecided in cycles; this is a meaningful result to report, not necessarily an error to suppress.

VII. STUDY METHOD AND REPRODUCIBILITY

This paper uses a conceptual, literature-informed analysis rather than an empirical system evaluation. It brings together research on requirements elicitation and negotiation, generative AI, multi-agent collaboration, argumentation, and quality or safety concerns [1]–[30]. The purpose is to identify concepts and evaluation questions, not to claim that a particular model configuration or decision procedure has been tested.

A future empirical study should define its task domain, question set, source materials, and annotation criteria in advance. It should record the model families and versions, prompts, generation settings, candidate-selection criteria, and relevant dates so that changes in model behavior are not mistaken for effects of the analysis method. Evidence criteria should distinguish source availability from source relevance, and annotation should preserve the exact claim and context. Sensitive data and provider terms may limit what can be published; any such restriction should be disclosed.

Reproducibility also requires transparent treatment of interpretive choices. Researchers should report how material claims are segmented, how support and attack relations are assigned, which argumentation semantics is used, and how uncertainty affects comparative judgments. When annotation is subjective, independent reviewers and adjudication can characterize disagreement without erasing genuine ambiguity. No numerical findings are asserted in this conceptual study.

VIII. FUTURE EVALUATION FRAMEWORK

A. Study Design

An empirical study can compare single-response generation, multi-agent critique, and argumentation-informed comparison on paired questions. Study conditions, question scope, source materials, claim-extraction rules, relation criteria, and selection preferences should be specified before evaluation. Results should be stratified by task type and evidence conditions rather than combined into a universal score. Repeated sampling can characterize variability where model access and research ethics permit it.

Human annotation can assess claim materiality, evidence support, contradiction, and completeness. Independent annotators and adjudication can characterize disagreement, but adjudication does not turn an interpretive source dispute into certainty. The question, rather than each claim, should be the primary sampling unit because claims within a response are not independent observations. Annotation guidance should define materiality, adequate support, contradiction, scope mismatch, and unresolved evidence. Reviewers should examine the source

passage and exact claim context; agreement should be reported separately for each label type. Where feasible, reviewers can be blinded to the analysis condition.

B. Baselines and Ablations

Useful comparisons include a single response, same-agent revision, independent multi-agent generation, and multi-agent generation followed by explicit claim-and-argument analysis. Component analyses can examine the contribution of evidence review, attack relations, and grounded semantics separately. Conditions should use the same questions, context, and permitted evidence. The number of generation and critique opportunities should be comparable, so any observed difference is not simply caused by giving one condition more chances to answer. Model versions, prompts, and study conditions should be reported to support interpretation. These are proposed comparisons, not results of the present study.

C. Measures

The following outcomes separate answer quality from the behavior of intermediate components:

- Claim correctness and unsupported-claim rate: the proportion of material claims supported by reference evidence, and the proportion lacking adequate support.
- Extraction and relation quality: precision, recall, and F1 for claims, evidence links, and annotated support or attack relations.
- Grounded labelling: agreement between computed accepted, rejected, and undecided labels and labels derived from an adjudicated graph.
- Selection and ambiguity: quality of selected candidates, tie frequency, abstention or ambiguity rate, and coverage.
- Explanation utility: whether reviewers can identify decisive claims, evidence, relations, and remaining uncertainty.
- Decision calibration: whether confidence and abstention correspond to the strength and completeness of available evidence.

Argument acceptance and external evidence judgments must be reported separately. Explanation quality is not correctness: an accurate-looking graph may encode a faulty interpretation, and a correct response may have an incomplete explanation.

Table 3: Evaluation outcomes and their role

Measure	Operational definition	Purpose
Claim support	Material claims supported by reference evidence, separate from graph status.	Answer reliability
Unsupported rate	Unsupported material claims divided by all material claims.	Grounding risk
Extraction quality	Precision, recall, and F1 for claims, spans, evidence links, and relations.	Component quality
Grounded labels	Agreement with labels computed from an adjudicated attack graph.	Semantic correctness
Selection coverage	Fraction of questions receiving a comparative judgment.	Abstention behavior
Explanation utility	Reviewer accuracy reconstructing claims, evidence, and rationale.	Clarity
Decision calibration	Alignment between confidence, evidence strength, and abstention.	Appropriate reliance

For N study questions, let S be the subset for which a comparative judgment is made and let E_i indicate whether that judgment meets an external factuality rubric. Coverage is |S|/N. Error among judged cases is 1 - |S|^-1 * sum_{i ∈ S} 1[E_i], with zero coverage reported separately rather than treated as zero error. Abstention and the fraction of undecided graph arguments are distinct quantities.

In the absence of reported measurements, no numerical performance or superiority claim is made here.

D. Hypotheses and Statistical Analysis

An empirical evaluation could test four hypotheses: graph-based comparison reduces unsupported material claims relative to a single-response baseline; structured claims and relations improve reviewers' ability to reconstruct a judgment; extraction and relation errors predict changes in grounded labels and comparative outcomes; and any quality gains justify the additional analytical effort. These hypotheses are not assumed to hold. A null or negative result would

indicate that the added reasoning structure is not justified under the tested conditions.

Analysis should account for claims clustered within questions and for repeated runs. A question-level paired bootstrap or a preselected mixed-effects model can be used for paired conditions; the method and primary outcome should be fixed before test scoring. Report effect sizes and uncertainty intervals, and declare any correction for multiple ablation comparisons. Subgroup results by dataset, evidence conflict, answer length, and provider agreement should be identified as exploratory unless powered in advance. Sample size should follow a power analysis based on pilot annotation, rather than be chosen after observing test outcomes.

E. Human Review and Interpretation

Human review is especially important for unresolved source conflicts, critical claims, and high-impact decisions; it should not make an automated judgment appear certain. Reviewers should examine the original claim context, relevant passages, and objections, and be able to challenge an extraction, source interpretation, or relation with a stated reason. A user study can assess whether structured argumentation helps reviewers reconstruct decisive evidence and detect unsupported statements, while also measuring whether graph displays increase overreliance.

The grounded framework evaluates attacks, whereas evidence review assesses factual support. Relevant challenges may include direct contradiction, scope rebuttal, source-quality concerns, and invalid inference. Evaluation should compare computed labels with those derived from an adjudicated graph and measure relation precision and recall. Neither agreement with a graph nor a stable graph establishes truth.

Any assessment of practical value should consider the cognitive and resource costs of multi-agent comparison alongside answer quality. Comparisons should distinguish improvements attributable to argumentation from those due simply to additional opportunities for generation or critique.

Human review can be evaluated by presenting reviewers with a question, candidate answers, relevant evidence, and either a structured argumentation summary or a baseline explanation. Questions should assess which claims are decisive, what evidence challenges them, and why an argument received its label. Reconstruction accuracy, review time, and reviewer confidence can be measured separately. A preferred diagram is not necessarily a more useful explanation; the relevant criterion is whether reviewers identify decisive evidence and unsupported statements more accurately.

IX. THREATS TO VALIDITY AND RESPONSIBLE USE

Claim interpretation depends on extraction and annotation. Compound statements, qualifiers, and temporal scope can be lost during segmentation; evaluation labels may not represent every nuanced answer. Assessing extraction separately can make these errors visible but does not eliminate them. Grounded labels also depend on the graph: missing or spurious attack relations may change which arguments are accepted, rejected, or undecided.

Agent outputs may be correlated through shared training data, prompting conventions, source material, or model families. Agent diversity is therefore not equivalent to independent evidence. Model versions and non-deterministic generation may change over time, so study conditions and repeated observations should be documented. Results from a particular task do not establish reliability in unrelated domains or high-impact settings. Dataset contamination and narrow label conventions may inflate apparent performance, while expert-curated evaluation introduces sampling and annotation costs. Findings should not be pooled into a universal reliability claim.

Questions and source material may contain confidential information, fabricated references, or malicious instructions. Sensitive inputs should not be shared externally without authorization. Retrieved material should be treated as untrusted evidence rather than authoritative instruction. High-impact decisions require qualified human review; multi-agent analysis is a decision-support concept, not an autonomous authority.

Confidence thresholds require calibration against held-out human judgments, and an LLM's self-reported confidence is not automatically a probability. Reports should explain unresolved issues rather than imply certainty through an unvalidated confidence label.

X. DISCUSSION

The study separates three questions that are often conflated: which response is preferable under stated criteria, which arguments are accepted, rejected, or undecided in a constructed graph, and which claims are supported by available evidence. Independent generation can expose differences; argumentation can structure conflicts; evidence assessment can examine factual support. None of these alone proves an answer true.

The separation also helps locate sources of error. An unreliable judgment may arise from claim extraction, evidence assessment, relation construction, or the choice of evaluation criteria. Additional agents may surface useful alternatives but can also repeat shared assumptions. Thus, the value of multi-agent analysis must be assessed against its cognitive and resource costs in the intended context. An explicit inconclusive result may be more responsible than selecting a candidate when evidence does not distinguish the alternatives.

A self-verifying method cannot completely verify itself. Agents may share blind spots, evidence can be incomplete or incorrect, and argument graphs can omit relevant challenges. Human review, source inspection, and explicit uncertainty remain essential. The potential contribution is a more inspectable and contestable reasoning process, not an appearance of certainty.

A. Interpreting the Analytical Layers

The proposed analysis can be viewed as a chain of dependent judgments. First, a response is interpreted and divided into material claims. Next, claims are compared with evidence and with one another. Relations among the resulting arguments then determine formal labels, after which task criteria inform a comparative judgment. Each layer inherits uncertainty from the earlier layers. A missed qualifier can create a false contradiction; a mistaken source interpretation can change an attack relation; and a correctly computed grounded extension can still produce a misleading impression if the graph itself is incomplete. An evaluation that measures only the final answer cannot identify these sources of error.

This dependency also clarifies why argumentation and factual verification should not be collapsed. A graph can establish that one represented argument attacks another and that a third argument defends it under a selected semantics. It cannot establish that a source is genuine, that its contents are current, or that an inference is relevant to the user's context. Those judgments require evidence criteria and, in many domains, expert knowledge. Dung's framework is valuable precisely when interpreted narrowly: it formalizes acceptability relations over a stated graph rather than certifying the underlying world.

The distinction between typed relations and binary attack is another important consideration. Direct contradiction, scope rebuttal, and a challenge to source quality can have different implications even if all are encoded as attacks for a particular analysis. Preserving the reason for each relation helps reviewers understand whether an argument is denied, qualified, or merely weakened. Research on bipolar and value-based argumentation provides possible conceptual resources for representing more than a single conflict type [28], [29]. The choice of formalism should follow the question being studied rather than the convenience of a single graph representation.

Finally, a transparent comparison should make room for incomplete evidence. When sources disagree or do not cover the question, an unresolved status can be more informative than a forced ranking. The user's decision context determines whether such an outcome calls for additional evidence, expert review, or no action. This is especially important where errors have unequal consequences: a modest gain in answer coverage may not justify a meaningful increase in unsupported critical claims. The framework therefore treats abstention and uncertainty as substantive findings to analyze, not defects to hide.

B. Reliability, Coverage, and Decision Context

Reliability is multidimensional. A response may be factually accurate but incomplete, or comprehensive but include an unsupported detail that changes its practical meaning. Claim-level measures can reveal these differences, yet a simple fraction of supported statements may give equal weight to a minor detail and a consequential recommendation. Study criteria should therefore identify material claims and explain how criticality affects interpretation. This is consistent with quality-oriented approaches that treat product quality as a set of distinct attributes rather than a single label [3], [7].

Coverage and abstention also create a decision trade-off. A method that answers only when evidence is

strong may lower error among selected cases while leaving more questions unresolved. A method that always selects an answer may have higher coverage but greater risk. Both outcomes should be reported together; otherwise, a cautious method may appear ineffective solely because it abstains, or an overconfident method may appear useful because it always provides an answer. The appropriate balance depends on the consequence of error and the availability of human review.

Decision context also affects how conflicting criteria should be handled. For an exploratory question, broad coverage may be valuable; for safety-sensitive reasoning, one unsupported critical claim may outweigh many supported peripheral details. A weighted score can make preferences explicit, but its numerical appearance should not imply objective precision when its weights reflect value judgments. Sensitivity analysis is useful because it reveals whether conclusions depend on small changes in those choices. If the preferred candidate changes under plausible settings, the report should describe the result as unstable rather than conceal the dependence.

Finally, explanations should be evaluated by their effect on judgment. A graph can improve visibility of disagreement, but it can also create an impression of rigor that exceeds the quality of its underlying claims or evidence.

Reviewers should be asked to identify unsupported assertions, scope mismatches, and unresolved conflicts, rather than merely rate the graph's clarity. This shifts explainability from presentation alone toward whether a person can question and appropriately use the analysis.

C. Research Priorities

Future studies should validate claim extraction, evidence relevance, and relation identification separately before attributing answer quality to the complete analytical approach. Comparisons should include simpler baselines and matched generation opportunities. Evaluation should report unsupported claims, relation errors, reviewer understanding, abstention, and uncertainty, while distinguishing observed results from proposed expectations. Formal semantic consistency alone does not establish factual reliability.

D. Sensitivity of the Selection Policy

Evaluation criteria are policy choices. Prioritizing coverage may favor a detailed response, while prioritizing evidence support may favor a narrower claim or abstention. These priorities should be stated for the task and not adjusted after observing test outcomes. Sensitivity analysis can vary declared weights and thresholds to determine whether comparative judgments are stable. If small changes reverse the ranking, the result is fragile and should be reported as such.

Human disagreement with a machine-generated judgment is not an error to conceal. Reviewers may have context unavailable to the analysis, and their reasons can inform future evaluation. Human and automated judgments should remain distinct so researchers can assess when review changes an outcome without treating reviewer agreement as automatic proof.

XI. CONCLUSION

This study examined how multi-agent generation and abstract argumentation may support more reliable and explainable decision making. It presented a conceptual comparison framework that separates claim-level evidence, graph-based argument status, and preference under explicit criteria. Grounded semantics offers a formal account of acceptance, rejection, and undecided arguments, but its labels remain conditional on the claims and relations represented.

No empirical performance result is presented. Reliability remains dependent on claim interpretation, source quality, relation identification, and the degree of independence among agents. Future studies should evaluate these components separately, compare them with simpler alternatives, and report uncertainty and human-review outcomes. The appropriate role of the approach is to support critical examination, not replace source expertise or present automated agreement as proof.

REFERENCES

[1] K. Pohl, "Requirements engineering: An overview," in *Encyclopedia of Computer Science and Technology*, vol. 36. New York, NY, USA: Marcel Dekker, 1996.

[2] S. U. Rehman and V. Gruhn, "An effective security requirements engineering framework for cyber-physical systems," *Technologies*, vol. 6, no. 3, Art. no. 65, Jul. 2018, doi: 10.3390/technologies6030065.

[3] Systems *and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—System and Software Quality Models*, Standard ISO/IEC 25010, 2011.

[4] A. van Lamsweerde, "Requirements engineering in the year 00: A research perspective," in *Proc. 22nd Int. Conf. Software Engineering (ICSE)*, 2000, pp. 5–19, doi: 10.1145/337180.337184.

[5] Standish Group, "CHAOS Report 2020," 2020. [Online]. Available: https://www.standishgroup.com/

[6] D. M. Fernández *et al*., "Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice," *Empirical Software Engineering*, vol. 22, no. 5, pp. 2298–2338, Oct. 2017, doi: 10.1007/s10664-016-9451-7.

[7] J. D. Musa, Software Reliability Engineering: More Reliable Software Faster and Cheaper, 2nd ed. Bloomington, IN, USA: AuthorHouse, 2004.

[8] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design," arXiv:2303.07839, 2023.

[9] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, and C. Berger, "Requirements engineering using generative AI: Prompts and prompting patterns," in Generative AI for Effective Software Development, A. Nguyen-Duc, P. Abrahamsson, and F. Khomh, Eds. Cham, Switzerland: Springer, 2024, pp. 109–127, doi: 10.1007/978-3-031-55642-5_5.

[10] D. Jin, Z. Jin, X. Chen, and C. Wang, "MARE: Multi-agents collaboration framework for requirements engineering," arXiv:2405.03256, 2024.

[11] D. Jin, W. Sun, J. Huang, P. Liang, J. Xuan, Y. Liu, and Z. Jin, "IReDev: A knowledge-driven multi-agent framework for intelligent requirements development," arXiv:2507.13081, 2025.

[12] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," Artificial Intelligence, vol. 77, no. 2, pp. 321–357, Sep. 1995, doi: 10.1016/0004-3702(94)00041-X.

[13] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," Official Journal of the European Union, 2024.

[14] Road Vehicles—Functional Safety, Standard ISO 26262:2018, 2018.

[15] J. Lee, "Extending the Potts and Bruns model for recording design rationale," in Proc. 13th Int. Conf. Software Engineering, 1991, pp. 114–125, doi: 10.1109/ICSE.1991.130629.

[16] A. Tang, M. A. Babar, I. Gorton, and J. Han, "A survey of architecture design rationale," Journal of Systems and Software, vol. 79, no. 12, pp. 1792–1804, Dec. 2006, doi: 10.1016/j.jss.2006.04.029.

[17] I. Jureta, J. Mylopoulos, and S. Faulkner, "Revisiting the core ontology and problem in requirements engineering," in Proc. 16th IEEE Int. Requirements Engineering Conf., May 2008, pp. 71–80, doi: 10.1109/RE.2008.13.

[18] Z. Li, X. Fang, C. Chen, M. Li, and B. Liao, "Enhancing conflict resolution in language models via abstract argumentation," Neurocomputing, vol. 664, Art. no. 132093, Feb. 2026, doi: 10.1016/j.neucom.2025.132093.

[19] M. Ataei, H. Cheong, D. Grandi, Y. Wang, N. Morris, and A. Tessier, "Elicitron: A large language model agent-based simulation framework for design requirements elicitation," Journal of Computing and Information Science in Engineering, vol. 25, no. 2, Art. no. 021012, Feb. 2025, doi: 10.1115/1.4067388.

[20] C. Arora, J. Grundy, and M. Abdelrazek, "Advancing requirements engineering through generative AI: Assessing the role of LLMs," in Generative AI for Effective Software Development. Cham, Switzerland: Springer, 2024, pp. 129–148.

[21] K. Ronanki, C. Berger, and J. Horkoff, "Investigating ChatGPT's potential to assist in requirements elicitation processes," in Proc. 49th Euromicro Conf. Software Engineering and Advanced Applications (SEAA), Sep. 2023, pp. 354–361, doi: 10.1109/SEAA60479.2023.00061.

[22] H. Cheng, J. H. Husen, Y. Lu, T. Racharak, N. Yoshioka, N. Ubayashi, and H. Washizaki, "Generative AI for requirements engineering: A systematic literature review," Software: Practice and Experience, vol. 56, no. 2, pp. 141–170, 2026, doi: 10.1002/spe.70029.

[23] B. Nuseibeh and S. Easterbrook, "Requirements engineering: A roadmap," in Proc. Conf. Future of Software Engineering, May 2000, pp. 35–46, doi: 10.1145/336512.336523.

[24] D. Zowghi and C. Coulin, "Requirements elicitation: A survey of techniques, approaches, and tools," in Engineering and Managing Software Requirements, 2005, pp. 19–46.

[25] P. K. Murukannaiah and M. P. Singh, "Xipho: Extending Tropos to engineer context-aware personal agents," in Proc. Int. Joint Conf. Autonomous Agents and Multiagent Systems, May 2014, pp. 309–316.

[26] P. Grünbacher and N. Seyff, "Requirements negotiation," in Engineering and Managing Software Requirements. Berlin, Germany: Springer, 2005, pp. 143–162, doi: 10.1007/3-540-28244-0_7.

[27] H. Cheng, M. Kim, F. Khomh, T. Racharak, N. Yoshioka, N. Ubayashi, and H. Washizaki, "QUARE: Quality-aware requirements analysis through multi-agent dialectic negotiation," in Proc. 34th IEEE Int. Requirements Engineering Conf. (RE), Aug. 2026, pp. 1–11.

[28] T. J. M. Bench-Capon, "Persuasion in practical argument using value-based argumentation frameworks," Journal of Logic and Computation, vol. 13, no. 3, pp. 429–448, Jun. 2003, doi: 10.1093/logcom/13.3.429.

[29] C. Cayrol and M.-C. Lagasquie-Schiex, "On the acceptability of arguments in bipolar argumentation frameworks," in *Proc. 8th European Conf. Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Mar. 2005, pp. 378–389, doi: 10.1007/11518655_33.

[30] I. Mirbel and S. Villata, "Enhancing goal-based requirements consistency: An argumentation-based approach," in *Proc. 13th Int. Conf. Computational Logic in Multi-Agent Systems (CLIMA)*, vol. 7486, 2012, pp. 110–127.