Argumentative Large Language Models for Explainable and Contestable Claim Verification (Extended Abstract)

Gabriel Freedman , **Adam Dejl** , **Deniz Gorur** , **Xiang Yin** , **Antonio Rago** and **Francesca Toni**
Department of Computing, Imperial College London
{gif22, adam.dejl18, d.gorur22, xy620, a.rago, ft}@imperial.ac.uk

Abstract

Large language models (LLMs) internalise large amounts of world knowledge yet lack mechanisms to reliably explain and justify their conclusions. We introduce *argumentative LLMs* (ArgLLMs) (Freedman et al. 2025), a model-agnostic wrapper that imbues any LLM with the formal guarantees of quantitative bipolar argumentation frameworks (QBAFs). For an input claim: (i) an LLM generates supporting and attacking arguments; (ii) the same LLM assigns each argument an intrinsic numerical strength; and (iii) a deterministic gradual semantics computes a final dialectical strength, leading to an assessment of the veracity of the claim. The resulting output is therefore *faithful*, *human-interpretable*, and provably *contestable*. Across three fact-verification benchmarks, ArgLLMs match chain-of-thought prompting in accuracy while additionally providing reliable rationales.

1 Motivation

LLMs exhibit strong zero-shot performance on diverse reasoning tasks, but their autoregressive generation offers no guarantee that the produced "explanations" correspond to the latent computation (Turpin et al. 2023). This misalignment jeopardises LLMs' explainability and eliminates any principled way to contest any faulty answers that they provide. Formal argumentation has been advocated within KR as a model of defeasible inference offering transparent rationales (Leofante et al. 2024). We therefore ask:

Can we enhance LLMs by adding a symbolic argumentative layer that preserves their native capabilities while enforcing faithful and contestable reasoning?

2 Background

Quantitative Bipolar Argumentation Framework A QBAF is a tuple $\langle A, R^-, R^+, \tau \rangle$ where A is a finite set of arguments, R^- (attack) and R^+ (support) are disjoint binary relations on A, and $\tau : A \to [0, 1]$ assigns each argument an intrinsic strength (Baroni, Rago, and Toni 2019).

The dialectical strength σ of every argument may then be computed by a gradual semantics. We employ DF-QuAD (Rago et al. 2016), which, for a given argument, aggregates the strengths of its attackers and supporters, along with its intrinsic strength, to compute σ.

Restricted Trees To ensure experimental consistency, we restrict generated QBAFs to depth ≤ 2 and breadth ≤ 2 (inclusive of supports and attacks). In practice this results in competitive results, however it is not a fundamental limitation of the framework, which can be adapted to accomodate an arbitrary number of nodes.

3 ArgLLM Pipeline

Figure 1 depicts the four stage process of the ArgLLM methodology, which is outlined below.

Stage 1: Argument Generation Given a claim x, we prompt an LLM to output supporting and attacking arguments. Each argument is generated in an independent context window. The prompt is recursively reused on the generated arguments until the pre-specified depth is reached.

Stage 2: Intrinsic Strength Attribution We then prompt an LLM to produce, for each argument α, a numerical argument strength. The returned integer is normalised to $[0, 1]$ and stored as $\tau(\alpha)$. We experiment with two options for the root x, using a fixed prior of $\tau(x) = 0.5$, and using an estimated prior as with the child arguments.

Stage 3: Dialectical Aggregation The populated QBAF is now a fixed symbolic object. Applying DF-QuAD yields σ for every node without additional calls to the LLM. As the semantics is monotonic in every support and attack argument, the reasoning is faithfully captured by the final graph.

Stage 4: Output Taking x as the root argument of our QBAF, in our experiments we consider the claim True whenever $\sigma(x) > 0.5$. However, it is very simple to adjust this threshold based on empirical validation, or corresponding to different risk tolerances.

4 Formal Guarantees

Let x be the root of a QBAF and let σ be DF-QuAD. We prove two monotonicity properties (details in the full paper).

Theorem 1 (Base-score contestability)**.** *If a supporting leaf increases its intrinsic strength, $\sigma(x)$ cannot decrease. Likewise, strengthening an attacking leaf cannot increase $\sigma(x)$.*

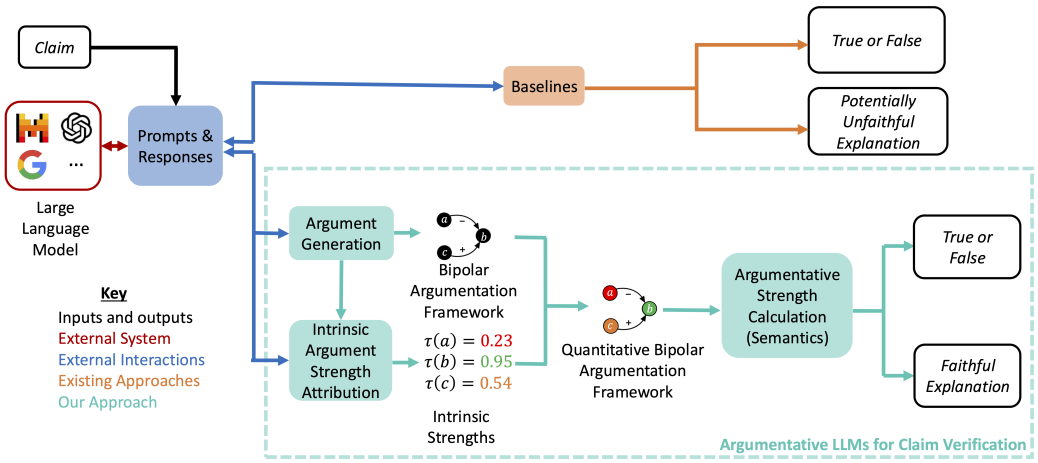

Figure 1: ArgLLM workflow - Stages 1 and 2 (Argument Generation and Intrinsic Argument Strength Attribution) query the LLM, Stage 3 (Argumentative Strength Calculation) is deterministic post-processing.

Theorem 2 (Structural contestability)**.** *Adding (removing, respectively) a new supporting branch never decreases (increases) $\sigma(x)$, and vice-versa for attacking branches.*

These properties guarantee that human edits to the argument graph have a predictable effect on the outcome, enabling systematic human interaction.

5 Example

An example of a basic instantiation of the framework:

Claim x: 'Taking ibuprofen can help you recover from a cold more quickly.'

1. Generate arguments

Support a_1: 'Ibuprofen reduces inflammation and fever, helping the body fight the virus.'

Attack a_2: 'Ibuprofen only masks symptoms and does not affect viral replication, so recovery time is unchanged.'

2. Score arguments

$\tau(x) = 0.5$ (neutral prior), $\tau(a_1) = 0.7$, $\tau(a_2) = 0.6$

3. Apply semantics

With one supporter and one attacker the semantics yields $\sigma(x) = 0.55$.

4. Verdict

Since $\sigma(x) > 0.5$, ArgLLM returns the label *True*.

6 Relevance to Knowledge Representation

1. **Argumentation and Non-Monotonic Reasoning** ArgLLMs blend symbolic argumentation frameworks with neurally generated components, extending comutational argumentation to a neuro-symbolic setting.
2. **Explainability and Contestability** The generated QBAF is a faithful representation of the decision-making process: every prediction is reproducible and locally contestable via Theorems 1 and 2.
3. **Bridging Symbolic and Sub-Symbolic AI** The pipeline showcases how lightweight KR formalisms can supervise, audit, and improve black-box LLMs without fine-tuning.

7 Conclusion and Outlook

ArgLLMs demonstrate that attaching a formal argumentation layer to off-the-shelf LLMs yields decisions that are simultaneously *accurate*, *transparent*, and *contestable*. These properties are particularly valuable in open-ended settings, where the correctness of the final answer is difficult or impossible to verify (e.g. forecasting). Future work includes generating richer argument graph topologies and using ensembles of LLMs to populate the frameworks.

References

Baroni, P.; Rago, A.; and Toni, F. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.* 105:252–286.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2025. Argumentative large language models for explainable and contestable claim verification. *Proceedings of the AAAI Conference on Artificial Intelligence* 39(14):14930–14939.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; Yin, X.; Zhang, D.; and Toni, F. 2024. Contestable AI Needs Computational Argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, 888–896.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *KR*, 63–73.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*.