Retrieval- and Argumentation-Enhanced Multi-Agent LLMs for Judgmental Forecasting

Deniz Gorur
Imperial College London
United Kingdom
d.gorur22@imperial.ac.uk

Antonio Rago
King's College London
United Kingdom
antonio.rago@kcl.ac.uk

Francesca Toni
Imperial College London
United Kingdom
ft@imperial.ac.uk

ABSTRACT

Judgmental forecasting is the task of making predictions about future events based on human judgment. This task can be seen as a form of claim verification, where the claim corresponds to a future event and the task is to assess the plausibility of that event. In this paper, we propose a novel multi-agent framework for claim verification, whereby different agents may disagree on claim veracity and bring specific evidence for and against the claims, represented as quantitative bipolar argumentation frameworks (QBAFs). We then instantiate the framework with a variety of agents realised with Large Language Models (LLMs): (1) ArgLLM agents, an existing approach for claim verification that generates and evaluates QBAFs; (2) RbAM agents, whereby LLM-empowered Relation-based Argument Mining (RbAM) from external sources is used to generate QBAFs; (3) RAG-ArgLLM agents, extending ArgLLM agents with a form of Retrieval-Augmented Generation (RAG) of arguments from external sources. Finally, we conduct experiments with two standard judgmental forecasting datasets, with instances of our framework with two or three agents, empowered by six different base LLMs. We observe that combining evidence from agents can improve forecasting accuracy, especially in the case of three agents, while providing an explainable combination of evidence.

KEYWORDS

Argumentation, LLMs, Judgmental Forecasting, RAG

ACM Reference Format:
Deniz Gorur, Antonio Rago, and Francesca Toni. 2026. Retrieval- and Argumentation-Enhanced Multi-Agent LLMs for Judgmental Forecasting. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/SNBR1486

1 INTRODUCTION

Judgmental forecasting is the task of making predictions about future events by reasoning over incomplete, uncertain, and often conflicting information [26, 38] based on the judgment of agents. This task can be seen as a form of claim verification, where a future outcome is seen as a claim and the task is to assess the plausibility of that outcome. The claim verification task involves determining whether a given claim is supported or refuted, often requiring the assessment of conflicting evidence. This evidence for verifying forecasting claims could be, for example, generated by LLMs [13, 18, 41] or obtained from other repositories. However, despite LLMs' remarkable capabilities across a range of tasks, they fall short here due to the fact that they may hallucinate or provide logically inconsistent outputs, and they cannot faithfully explain or allow for the contestation of their own outputs [17]. For example, claims about future events may require up to date knowledge that a single LLM lacks, due to its incomplete training data, leading to unreliable or conflicting predictions [37]. One approach, which aims at targeting this issue, uses Argumentative LLMs (ArgLLMs) [17]. ArgLLMs leverage techniques from computational argumentation (see [2, 4] for overviews) to generate Quantitative Bipolar Argumentation Frameworks (QBAFs) [5] that provide structured debates on the claim to be verified. In doing so, ArgLLMs output a transparent decision for the claims along with supporting and attacking arguments acting as evidence. However, relying on a single ArgLLM is limiting, as its output is constrained by the knowledge and biases of its underlying LLM, potentially omitting crucial evidence.

To overcome the limitations of a single-agent approach, we propose a novel multi-agent framework for claim verification that combines argumentative reasoning with QBAFs from multiple, independent agents into a single, more robust QBAF. As illustrated in Figure 1, our *Multi-Agent QBAF Combinator* module aggregates the outputs from several agents by measuring the semantic similarity [9] between arguments, merging similar views to create a more robust framework. We chose QBAFs as they are core to ArgLLMs and have been deployed successfully in several applied settings, including judgmental forecasting [23] and decision-making [3, 15].

To instantiate our Multi-Agent QBAF Combinator module to support judgmental forecasting, we consider two novel kinds of agents realised with LLMs, in addition to ArgLLM agents. For these two novel kinds of agents, we leverage on the widely acknowledged fact that the integration of Retrieval-Augmented Generation (RAG) [40] enhances LLMs by incorporating external knowledge and solving issues like hallucination [18, 37]. We use two novel types of RAG-based agents: (1) using Relation-based Argument Mining (RbAM) [6, 20] to identify supporting, attacking, or neither relations between retrieved evidence from sources and claims; and (2) using the retrieved evidence from sources for generating supporting and attacking arguments, in the spirit of ArgLLMs.

Overall, our contributions, overviewed in Figure 1, are as follows:

(1) *Multi-Agent QBAF Combinator*: a novel method to combine independently generated QBAF outputs into a single QBAF.
(2) *RbAM Agents*: using RbAM to incorporate evidence retrieved from external sources, directly as arguments.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/SNBR1486

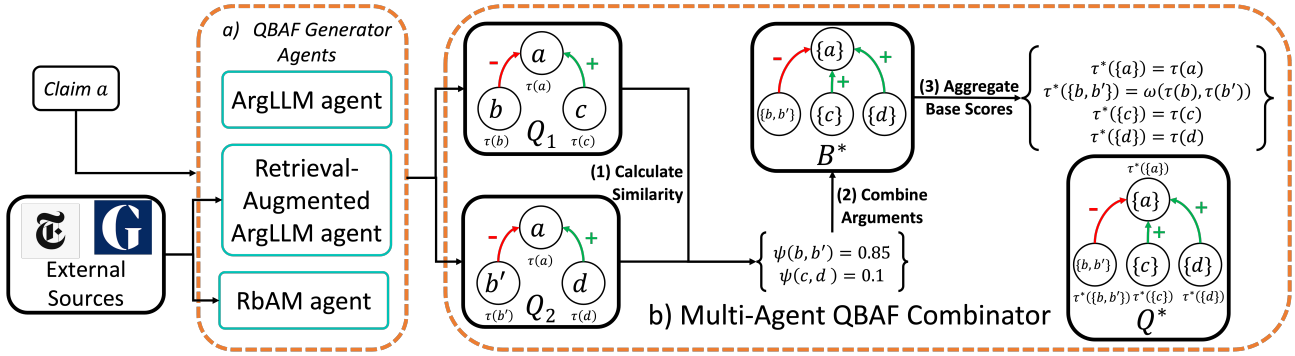

Figure 1: The overall pipeline. The 'QBAF Generator Agents' module can be instantiated with ArgLLM agents (baseline) or our two RAG-based methods (§5): Retrieval-Augmented ArgLLM agents and RbAM agents, both taking in input external sources. The 'Multi-Agent QBAF Combinator' (§4) module then takes the generated QBAFs (two in the figure, but our method applies to any number) and (1) calculates similarity between arguments in the QBAFs, (2) combines similar arguments to obtain a single BAF $\mathcal{B}^*$, and (3) aggregates the base scores of the combined arguments to obtain base scores τ^*, leading to a combined QBAF Q^*.

(3) *Retrieval-Augmented ArgLLM Agents*: using evidence retrieved from external sources, in prompts (in the spirit of [42]).
(4) Extensive experiments on two judgmental forecasting datasets, demonstrating that our multi-agent framework, particularly with three distinct agents, can improve forecasting accuracy and provide a transparent framework for claim verification in judgmental forecasting.

An extended version of this paper [21] includes additional information in the supplementary material.

2 RELATED WORK

Claim verification with LLMs and RAG. The field of claim verification has witnessed rapid advancements, particularly with LLMs, which can process extensive amounts of information and have remarkable generative capabilities. Despite their strengths, LLMs encounter limitations in this task as they may generate conflicting evidence due to their diverse and incomplete training data [13], which can lead to inaccurate predictions. To address this, Zhang and Gao [39] use a hierarchical step-by-step prompting method that breaks down the claim, which improves the accuracy of news claim verification. Another solution is CRAVE [41], which improves accuracy by first eliminating ambiguities and retrieving evidence from external sources, and then reasoning through two conflicting perspectives. Similarly, Ge et al. [18] propose using RAG combined with LLMs to resolve conflicting evidence and improve claim verification accuracy. Knowledge-based LLM approaches have been developed for structured claim verification, including generating argumentation frameworks to reason about claims (i.e. ArgLLMs [17]), extracting argumentation frameworks from RAG sources to reason about claims (i.e. ArgRAG [42]), generating first order logic representations to reason about claims [36], and using multi-agent LLM systems for claim verification [16]. We build on some of this earlier work on using LLMs to generate argumentation frameworks, but focus on combining the agents' generated frameworks.

LLMs and Judgmental Forecasting. Recent research has also explored the capabilities of LLMs in the context of judgmental forecasting, with mixed findings. Several studies have found that using LLMs directly, without external grounding, often fails to improve forecasting accuracy compared to human judgment [1, 33]. In contrast, Halawi et al. [22] demonstrate that LLMs augmented with RAG can approach the performance of expert human forecasters. Moreover, they show that combining forecasts from humans and LLMs achieves the best forecasting accuracy. Other work suggests that human forecasters advised by LLM-generated inputs can further enhance performance [34]. Overall, while LLMs alone rarely outperform expert forecasters, RAG appears promising for achieving forecasting accuracy comparable to humans, which motivated us to incorporate it in our pipeline.

Combining Argumentation Frameworks. In computational argumentation, there has been some work on merging argumentation frameworks. Coste-Marquis et al. [10, 11] develop a merging pipeline that aligns agents' frameworks into a common domain, uses minimal edit distance operators to merge attack relations, and then votes on extensions to produce group level consensus arguments. Cayrol and Lagasquie-Schiex [8] generalise this by introducing weighted argumentation frameworks that embed relative strengths of disagreement, producing a single weighted argumentation framework. Leite et al. [27] merge argumentation frameworks using a semantic approach that selects arguments and attacks of an agent varying the least from other agents. Delobelle et al. [12] merge the sets of acceptable arguments rather than merging argumentation frameworks directly. Argumentative exchanges [30] allow combining knowledge from argumentation agents, but in a selected and distributed manner. In this paper, we focus on merging QBAFs, a topic that, to our knowledge, has not been studied before.

3 PRELIMINARIES

A *Bipolar Argumentation Framework (BAF)* [7] is a triple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, where: $\mathcal{X}$ is a set of *arguments*; $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ is a relation *of attack*; and $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$ is a relation of *support*, where $\mathcal{A}$ and $\mathcal{S}$ are disjoint ($\mathcal{A} \cap \mathcal{S} = \emptyset$). A *Quantitative Bipolar Argumentation Framework*

(QBAF) [5] is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ where: $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ is a BAF and $\tau : \mathcal{X} \to [0,1]$ is a total function, with $\tau(a)$ the *base score* of $a \in \mathcal{X}$.

In this paper, we adopt the notion of a *BAF/QBAF for an argument* $a \in \mathcal{X}$ from Rago et al. [30], as follows. For $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ a QBAF and $\mathcal{B} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ a BAF, for any $\alpha, \beta \in \mathcal{X}$, let a *path* from α to β be $p = \langle (\alpha_0, \alpha_1), \ldots, (\alpha_{n-1}, \alpha_n) \rangle$ for some $n > 0$ (referred to as the *length* of p, denoted $|p|$) where $\alpha_0 = \alpha$, $\alpha_n = \beta$ and, for any $1 \leq i \leq n$, $(\alpha_{i-1}, \alpha_i) \in \mathcal{A} \cup \mathcal{S}$. Let $\text{path}(\alpha, \beta)$ and $|\text{path}(\alpha, \beta)|$ indicate the set of all paths from α to β and the number of paths in $\text{path}(\alpha, \beta)$, respectively. Then, for $\alpha^* \in \mathcal{X}$, $\mathcal{Q}/\mathcal{B}$ is a *BAF/QBAF for* α^* iff (i) $\forall_{\alpha \in \mathcal{X}} \text{path}(\alpha^*, \alpha) = \emptyset$; (ii) $\forall_{\alpha \in \mathcal{X} \setminus \{\alpha^*\}} |\text{path}(\alpha, \alpha^*)| = 1$; (iii) $\forall_{\alpha \in \mathcal{X}} \text{path}(\alpha, \alpha) = \emptyset$. In essence, a BAF/QBAF for α^* is essentially a tree with α^* as the root, instead of a multi-tree considered in [30]. We also use *pro/con* arguments in QBAFs as in [30]. Let $\mathcal{Q}$ be a QBAF for α^*. Then, the *pro arguments* and *con arguments* for $\mathcal{Q}$ are, respectively: $\text{Pro}(\mathcal{Q}) = \{\alpha \in \mathcal{X} \mid \exists p \in \text{path}(\alpha, \alpha^*)$, where $|p \cap \mathcal{A}|$ is even$\}$; $\text{Con}(\mathcal{Q}) = \{\alpha \in \mathcal{X} \mid \exists p \in \text{path}(\alpha, \alpha^*)$, where $|p \cap \mathcal{A}|$ is odd$\}$.

In the remainder of the paper, we assume that the QBAFs are all QBAFs for some claim, i.e. trees (and thus acyclic graphs).

Arguments in QBAFs are evaluated using *gradual semantics*, i.e. total functions often in the form $\sigma : \mathcal{X} \to [0,1]$, assigning a *strength* to each argument. One such semantics is the *Discontinuity-Free Quantitative Argumentation Debate (DF-QuAD)* [31]. For a given QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, for any $x \in \mathcal{X}$ with $n \geq 0$ attackers with strengths $v_1, \ldots, v_n$, $m \geq 0$ supporters with strengths $v'_1, \ldots, v'_m$ and $\tau(x) = v_0$, DF-QuAD computes x's strength as

$\sigma(x) = C(v_0, \mathcal{F}(v_1, \ldots, v_n), \mathcal{F}(v'_1, \ldots, v'_m))$,

where, for any $w_1, \ldots, w_k$, $\mathcal{F}(w_1, \ldots, w_k)$ is 0 if $k = 0$ and $1 - \prod_{i=1}^{k}(|1 - w_i|)$ otherwise, while C is defined as follows: for $v_a = \mathcal{F}(v_1, \ldots, v_n)$ and $v_s = \mathcal{F}(v'_1, \ldots, v'_m)$, if $v_a = v_s$ then $C(v_0, v_a, v_s) = v_0$; else if $v_a > v_s$ then $C(v_0, v_a, v_s) = v_0 - (v_0 \cdot |v_s - v_a|)$; otherwise $C(v_0, v_a, v_s) = v_0 + ((1 - v_0) \cdot |v_s - v_a|)$.

4 MULTI-AGENT CLAIM VERIFICATION

To verify claims using multiple agents, we propose a method to combine the outputs of independently-generated QBAFs. Our method produces a combined QBAF by clustering the arguments across the independently-generated QBAFs using a similarity measure, followed by aggregating their base scores. This allows us to capture a diverse argumentative perspective. We assume that the independently-generated QBAFs are such that there is no (x, y) that is an attack in one of the QBAFs and a support in another, i.e. we assume a lingua franca for the relations as in [30].

To determine whether arguments should be combined, we first need a formal notion of their similarity. We thus define similarity functions for clustering arguments.

Definition 1. *Let $\mathcal{X}$ be a set of arguments. A* similarity function *$\Psi : \mathcal{X} \times \mathcal{X} \to [0,1]$ is such that for $x \in \mathcal{X}$, $\Psi(x, x) = 1$ and for $x, y \in \mathcal{X}$, $\Psi(x, y) = \Psi(y, x)$.*

This definition ensures that the similarity of an argument to itself is maximal and that the similarity function is independent of the order in which arguments are compared.

To aggregate multiple base scores assigned to similar arguments, we define base score aggregation functions. These functions allow for the aggregation of vectors of base scores into a single representative value.

Definition 2. *Let $K \in \mathbb{N}$. A* base score aggregation function w.r.t. *K, $\omega : \bigcup_{k=1}^{K} [0,1]^k \to [0,1]$ is such that:*

1. *for any $v \in [0,1]^k$, for any permutation v' of the elements of v, $\omega(v') = \omega(v)$ (*order-independence*);*
2. *for any $v \in [0,1]^k$, $\min(v) \leq \omega(v) \leq \max(v)$ (*boundedness*);*
3. *for any $v_i \in [0,1]$, $\omega((v_i, \ldots, v_i)) = v_i$ (*idempotence*);*
4. *for any $v, v' \in [0,1]^k$ with $v = (v_1, \ldots, v_k)$ and $v' = (v'_1, \ldots, v'_k)$, if $\forall_{i \in \{1,\ldots,k\}} (v_i \leq v'_i)$ then $\omega(v) \leq \omega(v')$ (*monotonicity*).*

This definition ensures that base score aggregation functions do not depend on the order in which the elements in their input are presented, always return a number which lies within the range of the provided base scores, if all elements in the input are the same then the output is the common element, and they are monotonic with respect to the inputs, respectively. We consider two instantiations of the notion of base score aggregation function: average aggregation, i.e. $\omega_{avg}((v_1, \ldots, v_k)) = \frac{1}{k} \sum_{i=1}^{k} v_i$; and maximum aggregation, i.e. $\omega_{max}((v_1, \ldots, v_k)) = \max_{i=1}^{k} v_i$. It is easy to see that both satisfy Definition 2 (in particular both functions are order-independent, bounded, idempotent, and monotonic – see [21]).

Next, we define what it means to aggregate a set of QBAFs into a combined QBAF, given a similarity function and a base score aggregation function. This combined QBAF captures argument clusters, relations between them, and their aggregated base scores.

Definition 3. *Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_n$ be $n > 1$ QBAFs, where, for $i \in \{1, \ldots, n\}$, $\mathcal{Q}_i = \langle \mathcal{X}_i, \mathcal{A}_i, \mathcal{S}_i, \tau_i \rangle$. Let $\mathcal{X} = \bigcup_{i=1}^{n} \mathcal{X}_i$, $\mathcal{A} = \bigcup_{i=1}^{n} \mathcal{A}_i$, and $\mathcal{S} = \bigcup_{i=1}^{n} \mathcal{S}_i$. Let $\Psi : \mathcal{X} \times \mathcal{X} \to [0,1]$ be a similarity function, and $\delta \in [0,1]$ be a similarity threshold. Let $\omega : \bigcup_{k=1}^{K} [0,1]^k \to [0,1]$ be a base score aggregation function w.r.t. $K = |\mathcal{X}|$. Then, the* combined QBAF *$\mathcal{Q}^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ is as follows:*

- *$\mathcal{X}^* \subseteq 2^{\mathcal{X}}$ satisfies the following properties:*
  1. *$\forall x \in \mathcal{X}, \exists_1 x^* \in \mathcal{X}^*$ such that $x \in x^*$;*
  2. *$\forall x, y \in \mathcal{X}, \exists x^* \in \mathcal{X}^*$ such that $x, y \in x^*$ iff (i) $\exists z^* \in \mathcal{X}^*$ and $\exists z, z' \in z^*$ with $(x, z), (y, z') \in \mathcal{A}$ or $(x, z), (y, z') \in \mathcal{S}$, and (ii) $\Psi(x, y) \geq \delta$;*
- *$\mathcal{A}^* = \{(x^*, y^*) \mid \exists_{x \in x^*} \exists_{y \in y^*}$ such that $(x, y) \in \mathcal{A}\}$;*
- *$\mathcal{S}^* = \{(x^*, y^*) \mid \exists_{x \in x^*} \exists_{y \in y^*}$ such that $(x, y) \in \mathcal{S}\}$;*
- *$\forall x^* = \{x_1, \ldots, x_k\} \in \mathcal{X}^*$, $\tau^*(x^*) = \omega(\tau(x_1), \ldots, \tau(x_k))$.*

The first property ensures assignment of each argument $x \in \mathcal{X}$ to a unique cluster, forming a singleton if no other properties apply. The second property ensures that any two arguments are in the same cluster if and only if they both have the same relation (attack or support) towards arguments in the same parent cluster and meet the threshold of the similarity function. Combined QBAFs preserve relations by adding relations between clusters and aggregating the base scores according to the clusters. Note that, not only each attack/support between clusters results from lifting an attack/support between arguments, but also each attack/support between arguments is captured by some attack/support between clusters. Formally:

Lemma 4.1. *Let $x^*, y^* \in \mathcal{X}^*$. Then, $(x^*, y^*) \in \mathcal{A}^*$ (or $(x^*, y^*) \in \mathcal{S}^*$) iff $\exists x \in x^*$ and $\exists y \in y^*$ such that $\exists i \in \{1, \ldots, n\}$ where $(x, y) \in \mathcal{A}_i$ (or $(x, y) \in \mathcal{S}_i$, respectively).*

Example 1. *Consider two QBAFs, $\mathcal{Q}_1 = \langle \mathcal{X}_1 = \{a, b, c\}, \mathcal{A}_1 = \{(b, a)\}, \mathcal{S}_1 = \{(c, a)\}, \tau_1(a) = 0.5, \tau_1(b) = 0.2, \tau_1(c) = 0.8 \rangle$ and $\mathcal{Q}_2 = \langle \mathcal{X}_1 =$*

$\{a, b', d, e, e'\}, \mathcal{A}_1 = \{(b', a)\}, \mathcal{S}_1 = \{(d, a), (e, b'), (e', b')\}, \tau_2(a) = 0.5, \tau_2(b') = 0.7, \tau_2(d) = 0.4, \tau_2(e) = 0.3, \tau(e') = 0.1\rangle$. *Consider a similarity function* Ψ *such that* $\Psi(b, b') = 0.9$, $\Psi(e, e') = 0.6$, *and all other possible argument pairs have similarity below the threshold* $\delta = 0.5$. *Let* ω *be the arithmetic mean* ω_{avg}. *We construct the combined QBAF* $Q^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ *as follows:* $\mathcal{X}^* = \{x_1^* = \{a\}$ *(singleton, no parent argument);* $x_2^* = \{b, b'\}$ *(since* $\Psi(b, b') = 0.9 > \delta$*);* $x_3^* = \{c\}$ *(singleton, no similar arguments);* $x_4^* = \{d\}$*(singleton, no similar arguments);* $x_5^* = \{e, e'\}$ *(since* $\Psi(e, e') = 0.6 > \delta$*)};* $\mathcal{A}^* = \{(x_2^*, x_1^*)$ *since* $(b, a) \in \mathcal{A}_1, (b', a) \in \mathcal{A}_2$; (x_5^*, x_2^*) *since* $(e, b') \in \mathcal{A}_2, (e', b') \in \mathcal{A}_2\}$; $\mathcal{S}^* = \{(x_3^*, x_1^*)$ *since* $(c, a) \in \mathcal{S}_1$; (x_4^*, x_1^*) *since* $(d, a) \in \mathcal{S}_2\}$; $\tau^*(x_1^*) = \omega(0.5, 0.5) = 0.5$; $\tau^*(x_2^*) = \omega(0.2, 0.7) = 0.45$; $\tau^*(x_3^*) = 0.8$; $\tau^*(x_4^*) = 0.4$; $\tau^*(x_5^*) = \omega(0.3, 0.1) = 0.2$.

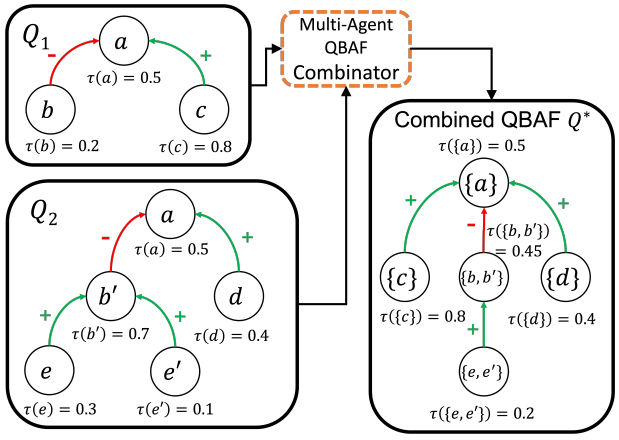

Figure 2: The Multi-Agent QBAF Combinator takes two initial QBAFs, Q_1 (top-left) and Q_2 (bottom-left), as input and outputs a single, merged QBAF Q^* (right). (See Example 1)

Proposition 1. *Let* $Q_1, \ldots, Q_n$ *be QBAFs, for* $n > 1$, *where, for* $i \in \{1, \ldots, n\}$, $Q_i = \langle \mathcal{X}_i, \mathcal{A}_i, \mathcal{S}_i, \tau_i \rangle$. *Then, the combined QBAF* $Q^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ *is a QBAF.*

Proposition 2. *Let* $Q_1, \ldots, Q_n$ *be QBAFs for the same argument* f, *for* $n > 1$, *where, for* $i \in \{1, \ldots, n\}$, $Q_i = \langle \mathcal{X}_i, \mathcal{A}_i, \mathcal{S}_i, \tau_i \rangle$. *Then, the combined QBAF* $Q^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ *is a QBAF for* $\{f\}$.

Proposition 3. *Let* $Q_1, \ldots, Q_n$ *be QBAFs for the same argument* f, *for* $n > 1$, *where, for* $i \in \{1, \ldots, n\}$, $Q_i = \langle \mathcal{X}_i, \mathcal{A}_i, \mathcal{S}_i, \tau_i \rangle$. *Then, the combined QBAF* $Q^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ *preserves pro/con arguments in the original QBAFs, i.e., for* $\mathrm{Pro}(Q_i)$ *and* $\mathrm{Con}(Q_i)$ *the set of pro-arguments and con-arguments, respectively, for* Q^**:*

$\forall_{x^* \in \mathcal{X}^*} \Big(x^* \in \mathrm{Pro}(Q^*) \Leftrightarrow (\exists_{x \in x^*, i \in \{1,\ldots,n\}} x \in \mathrm{Pro}(Q_i)) \Big)$, *and*

$\forall_{x^* \in \mathcal{X}^*} \Big(x^* \in \mathrm{Con}(Q^*) \Leftrightarrow (\exists_{x \in x^*, i \in \{1,\ldots,n\}} x \in \mathrm{Con}(Q_i)) \Big)$.

Note that arguments without any similarities with other arguments (e.g. d in Example 1) form singleton clusters (e.g. $\{e\}$ in the same example). Note also that we do not cluster arguments with different parents or with the same parent but in different relations. Thus, in Example 1, we do not cluster c and e (as they do not share a parent) or b and d (as they have opposite stances for the same parent) even if they are similar. Clustered arguments are made to share the same incoming relations in the combined QBAF. Thus, in Example 1, b and b' in the cluster $\{b, b'\}$ share supporter $\{e, e'\}$.

To construct a combined QBAF satisfying Definition 3, we define a bottom-up algorithm that clusters arguments layer by layer, based on the similarity function. The Multi-Agent QBAF Combinator algorithm starts from a given claim argument a and then clusters its supporters and attackers, recursively clustering their children, until the maximum depth of the input QBAFs is reached. For each cluster, the algorithm aggregates their base scores.

Algorithm 1 Multi-Agent QBAF Combinator

Require: $F = \{Q_1, \ldots, Q_n\}$: QBAFs for claim a, $n > 1$
Require: Ψ: similarity function
Require: δ: similarity threshold
Require: ω: base score aggregation function
1: Initialize empty QBAF $Q^* \leftarrow \langle \{\{x\} \mid x \in \mathcal{X}\}, \emptyset, \emptyset, \tau^*(a) = \omega(\tau_1(a), \ldots, \tau_n(a)) \rangle$
2: previous layer $\leftarrow \{a\}$
3: **for** $d \leftarrow 1$ to max depth level$(\{Q_1, \ldots, Q_n\})$ **do**
4: merged $\leftarrow \{\}$
5: **for** $z^* \leftarrow$ previous layer **do**
6: **for all** pairs (x, y) where $x, y \in \mathcal{X}$ such that $(x, z), (y, z') \in \mathcal{A}$ or $(x, z), (y, z') \in \mathcal{S}$ where $z, z' \in z^*$ **do**
7: **if** $\Psi(x, y) \geq \delta$ **then**
8: merge x^* and y^*, where $x \in x^*$ and $y \in y^*$
9: add x, y to same cluster in merged
10: **else**
11: add x to cluster in merged
12: add y to another cluster in merged
13: **for** $z^* \leftarrow$ previous layer and $x^* \leftarrow$ merged **do**
14: set $\tau^*(x^*) = \omega(\tau(x) \mid x \in x^*)$
15: **if** $\exists_{x \in x^*}(x, z) \in \mathcal{A}$ such that $z \in z^*$ **then**
16: add (x^*, z^*) to $\mathcal{A}^*$
17: **if** $\exists_{x \in x^*}(x, z) \in \mathcal{S}$ such that $z \in z^*$ **then**
18: add (x^*, z^*) to $\mathcal{S}^*$
19: previous layer $\leftarrow$ merged
20: **return** Q^*: combined QBAF

Lines 1-2 initialise the combined QBAF Q^* with singleton clusters for every argument in $\mathcal{X}$ in the union QBAF. We set the 'previous layer' to $\{a\}$, the root. The algorithm then executes recursively layer by layer, Lines 4-12 merge clusters in $\mathcal{X}^*$ and build a set of clusters 'merged' at the current depth. Arguments are in the same cluster if they support or attack the same parent argument and are similar. Otherwise, they are placed in separate clusters. Lines 13-18 add relations between 'merged' clusters to the combined QBAF, and assign to clusters their aggregated base score. Line 19 updates 'previous layer' to contain all merged arguments in the previous layer, enabling the algorithm to iterate to the next layer.

Example 2. *We apply Algorithm 1 to Example 1's* Q_1 *and* Q_2.

In line 1, the combined QBAF Q^* *is initialized with singleton clusters of all arguments in the union QBAF:* $\{\{a\}, \{b\}, \{c\}, \{b'\}, \{d\}, \{e\}, \{e'\}\} = \mathcal{X}^*$, *and the base score of* a *is set to* $\tau^*(a) = \omega(\tau_1(a), \ldots, \tau_n(a))$. *In line 2, the 'previous layer' is set to* $\{a\}$.

At depth 1, 'merged' is set to {} in line 4. Lines 5-12 process all arguments in 'previous layer' (a in depth 1). Lines 6-12 goes through $(b, b'), (c, d)$ *as* $(b, a), (b', a) \in \mathcal{A}$ *and* $(c, a), (d, a) \in \mathcal{S}$. $\Psi(b, b') > 0.5$ *so 'merged'=*$\{\{b, b'\}\}$. $\Psi(c, d) < 0.5$ *so 'merged'=* $\{\{b, b'\}, \{c\}, \{d\}\}$. *Line 8 merges clusters in the combined QBAF,* $\{\{b, b'\}, \{c\}, \{d\}\} \subset \mathcal{X}^*$.

Lines 13-18 goes through all merged clusters $\{b, b'\}, \{c\}, \{d\}$ *and all arguments in 'previous layer'. Then, the base scores are added in line 14,* $\tau^*(\{b, b'\}) = \omega(\tau(b), \tau(b')) = \omega(0.5, 0.5) = 0.5$, $\tau^*(\{c\}) = 0.8$, $\tau^*(\{d\}) = 0.4$. *Lines 15-18 adds support (attack) relations if there is an argument in the cluster that is supporting (attacking, respectively) an argument in the 'previous layer'. So,* $\{(\{b, b'\}, a)\} \subset \mathcal{A}^*$ *as* $(b, a) \in \mathcal{A}$, *and* $\{(\{c\}, a), (\{d\}, a)\} \subset \mathcal{S}^*$ *as* $(c, a), (d, a) \in \mathcal{S}$.

Before starting depth 2, line 19 assigns the union of all arguments that were in 'merged' to 'previous layer'. 'Previous layer'= $\{b, b', c, d\}$.

'merged' is set to {} in line 4. Lines 5-12 goes through all arguments in 'previous layer' $(b, b', c, d$ *in depth 1). Lines 6-12 goes through* (e, e') *as* $(e, b'), (e', b') \in \mathcal{S}$. $\Psi(e, e') > 0.5$ *so 'merged' =* $\{\{e, e'\}\}$. *Line 8 merges the clusters in the combined QBAF,* $\{\{e, e'\}\} \subset \mathcal{X}^*$.

Lines 13-18 goes through all merged clusters $\{e, e'\}$ *and all arguments in 'previous layer'. Then, the base scores are added in line 14,* $\tau^*(\{e, e'\}) = \omega(\tau(e), \tau(e)) = \omega(0.3, 0.1) = 0.2$. *Lines 15-18* $\{(\{e, e'\}, \{b, b'\})\} \subset \mathcal{A}^*$ *as* $(e, b) \in \mathcal{S}$.

Finally, the algorithm returns $\mathcal{Q}^*$ *as the combined QBAF.*

Proposition 4. *Algorithm 1 terminates and returns a combined QBAF in polynomial time.*

Note that we have defined the combination process algorithmically, however the Multi-Agent QBAF Combinator could be seen as a *judge agent*. In this view, the generating agents act as independent experts providing potentially conflicting evidence $(\mathcal{Q}_1, \ldots, \mathcal{Q}_n)$ and the judge reconciles these opinions.

5 LLM-AGENTS FOR QBAF GENERATION

In this section, we describe the three LLM-agent variants we consider: Argumentative LLM (ArgLLM) agents, Retrieval-Augmented ArgLLM (RAG-ArgLLM) agents (extending ArgLLMs with RAG, to incorporate external textual evidence), and Relation-based Argument Mining (RbAM) agents (constructing QBAFs by directly using the external sources as arguments and classifying these arguments' stances towards the claims). Each of these agents independently generates QBAFs. RAG-ArgLLM agents are a variant of ArgRAG [42]: whereas ArgRAG could generate any QBAFs, including with cycles, the QBAFs generated by RAG-ArgLLM agents are QBAFs for claims (and thus trees, in the spirit of ArgLLMs [17]).

5.1 ArgLLM Agents

ArgLLM agents extract QBAFs from LLMs and formally reason over them using gradual semantics, as in [17]. ArgLLMs agents have three components: (1) *Argument Generation*, producing a BAF $\mathcal{B}$ for a given claim a, given a generative model (LLM) G, and parameters for argument generation θ, such as the prompt used, and the number of arguments to be generated in depth and breadth, denoted formally as $\Gamma(x, G, \theta) \to \mathcal{B}$; (2) *Intrinsic Argument Strength Attribution*, assigning base scores to the arguments in $\mathcal{B}$ using an evaluative function E to obtain a QBAF $\mathcal{Q}$, defined as $\mathcal{E}(\mathcal{B}, E) \to \mathcal{Q}$; (3) *Argument Strength Calculation*, applying a gradual semantics σ to $\mathcal{Q}$ to obtain an assessment for a, denoted as $\Sigma(a, \mathcal{Q}, \sigma) \to \sigma(a)$.

5.2 RAG-ArgLLM Agents

To strengthen the factual grounding of the generated arguments, we extend ArgLLM agents with RAG, by prompting with textual evidence from external sources so as to generate more complete and informed debates.

Let $T = \{t_1, ..., t_k\}$ be a set of textual evidence relevant to a claim a, retrieved from a collection of sources, and let G be a generative model (LLM). We define RAG-ArgLLM agents as ArgLLM agents where θ in the argument generation function is a modified prompt (*RAG prompt for ArgLLMs*) which incorporates the retrieved evidence T. Figure 3 illustrates how this prompt can improve results.

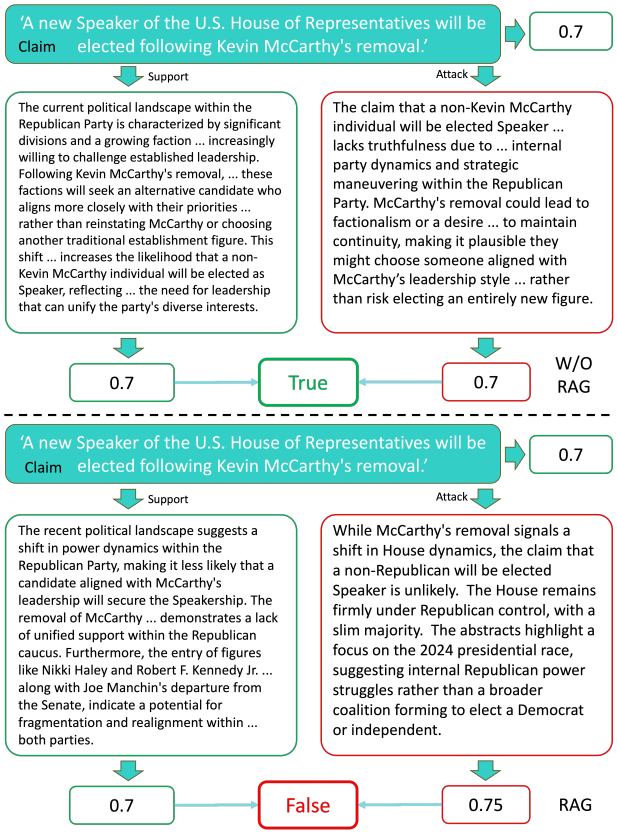

Figure 3: An example of how an RAG-ArgLLM agent (bottom) can improve the results compared to an ArgLLM agent (top). The example is taken from the Metaculus dataset and its claim is false. The ArgLLM agent incorrectly predicts it as false whereas the RAG-ArgLLM agent correctly predicts it.

5.3 RbAM Agents

In this setting, textual evidence is seen directly as arguments. The RbAM component in RAG-ArgLLM agents classifies the relation between each textual evidence and the claim as support, attack, or neither, yielding a QBAF in which the retrieved evidence may

support or attack the claim, if applicable. In other words, RbAM Agents treat each retrieved evidence as an argument, and the RbAM component determines whether it is related to the claim and how.

6 EXPERIMENTAL SET-UP

First, we evaluate our chosen individual agents (from §5): (1) ArgLLM agents (as a baseline), (2) RAG-ArgLLM agents, and (3) RbAM agents. As in [17], we considered two variants of ArgLLM agents: depth 1 (one attacker and one supporter for the claim, and no other arguments) and depth 2 (the attacker and supporter at depth 1 have one attacker and one supporter each). For RAG-ArgLLM agents, we use two different external sources (The NYTimes and The Guardian), chosen as they are openly accessible via an API. As for ArgLLM agents, we consider both depth 1 and depth 2 variants of RAG-ArgLLM agents. For RbAM agents, we adopt the best performing RbAM method from [20]: a few-shot prompt-based classification approach using Mixtral-8x7B [25], which Gorur et al. [20] show significantly outperforms prior baselines on the RbAM task. RbAM agents only generate depth-1 QBAFs for claims.

Then, we evaluate our multi-agent framework by combining the outputs of these agents in various configurations: (1) pairs of ArgLLM agents, (2) pairs of RAG-ArgLLM agents, (3) pairs consisting of different types of agents: one ArgLLM agent and one RAG-ArgLLM agent or two RAG-ArgLLM agents but using different external sources (The NYTimes and The Guardian), (4) three-agent combinations: combining individual ArgLLM agents with the two RAG-ArgLLM agents, each using a different source.

In all experiments, we use the Jina-V3 embeddings [35] with cosine similarity for the similarity function Ψ, selected due to its strong performance on semantic similarity tasks and compact size.[1] Also, we use the similarity threshold $\delta = 0.5$.[2]

6.1 External Sources

To support RAG, we retrieve evidence from two news article sites.

NYTimes. We collected all NYTimes article abstracts published between January 2023 and December 2024 using the NYTimes API.

Guardian. For each forecasting argument, we used GPT-4o-mini to generate five targeted search queries (see [21] for details). To get article abstracts these queries were submitted to the Guardian API.

Retrieval Pipeline. For both sources, to optimise evidence retrieval, we embedded each abstract using the Jina-V3 embedding model [35]. The resulting embeddings were stored in a vector database, ChromaDB, to support real-time retrieval. We then retrieved the top five relevant articles for each claim, restricted to abstracts dated before the closing date of the claim.

6.2 Base LLMs

We evaluate all configurations using the following six LLMs (including both closed- and open-weight models): Mixtral (Mixtral-8x7B-Instruct-v0.1) [25], Mistral (Mistral-7B-Instruct-v0.3) [24], Gemma (Gemma-7b-it) [28], Gemma-2 (Gemma-2-9b-it) [32], Llama-3 (Meta-Llama-3-8B-Instruct) [14], and GPT-4o (GPT-4o-mini) [29] models. We chose these models as their training data cut-off dates preceded the forecasting events, thereby avoiding data contamination.

6.3 Datasets

All configurations are executed on two popular judgmental forecasting datasets.

GJOpen. [3] A forecasting arguments dataset [19], which contains 2923 rephrased question-answer pairs from Good Judgment Open. The dataset covers both binary and multiple-choice forecasting questions, and is publicly available.

Metaculus. [4] A subset of the Forecasting dataset from [22], which originally includes 8881 binary forecasting questions. We picked only the questions that have been resolved and were open between the dates of September 2023 and September 2024, which yielded 388 forecasting questions.

Obtaining Claims. For our experiments, for each dataset, we converted each question-answer pair into a natural-language forecasting argument (claim) using the Mistral-7B-Instruct-v0.3 LLM [24], followed by manual review, similarly to [19].

7 RESULTS

7.1 Individual Agents

ArgLLM and RAG-ArgLLM Agents. We first evaluate the ArgLLM and RAG-ArgLLM agents (using the NYTimes and the Guardian) with their four variations depth 1 (with 0.5 base score and estimated base score assigned to the claim) and depth 2 (with 0.5 base score and estimated base score assigned to the claim) as in [17].

Table 1 (minus the last column) reports results on GJOpen and Metaculus. We note that retrieval improves forecasting accuracy on Metaculus in nearly all variations. The highest accuracy increase comes when the external sources are used with Gemma-2 (with depth 1, estimated base score), where accuracy improves from 68% to 81% (for both NYTimes and Guardian). The RAG-ArgLLM agents that did not improve accuracy were already performing well and the external sources did not seem to help. On GJOpen, the improvement is less consistent. However, using the Guardian seems to improve forecasting accuracy more compared to using the NYTimes. The highest accuracy increase occurs when using the NYTimes with Llama-3 (with depth 2, 0.5 base score), where the accuracy improves from 63% to 75%.

We additionally analysed the overlap between supporting and attacking arguments generated by RAG-ArgLLM agents when different external sources are used. Overall, we observe that overlap is generally high (see [21] for details), indicating that different sources often lead to substantially similar argumentative structures.

RbAM Agents. The last column in Table 1 shows the results for RbAM agents. Overall, RbAM agents do not perform well for GJOpen. On inspection of the generated QBAFs, this may be because the base scores generated are very small due to the abstracts not having a proper argumentative structure. One possible solution, left to future work, may be to prompt a separate model to extract or generate structured arguments from each abstract. The Guardian

[1] At the time of experimentation, Jina-V3 ranked fourth on the MTEB leaderboard: https://huggingface.co/spaces/mteb/leaderboard

[2] In our initial, exploratory experiments this value yielded the best results.

[3] https://www.gjopen.com/

[4] https://www.metaculus.com/

as an external source helped more than the NYTimes on GJOpen, possibly because of its more relevant context. On the other hand, RbAM agents perform better for Metaculus, although it still does not surpass the best ArgLLM or RAG-ArgLLM agents. Therefore, we decided not to combine QBAFs obtained from RbAM agents as the accuracy was very low.

Dataset	Depth	Source	Mixtral	Mistral	Gemma	Gemma-2	Llama-3	GPT-4o	RbAM
GJOpen	1	W/O	**67**/70	65/71	**73**/57	**79**/**77**	78/75	**74**/**69**	-
		NYTimes	61/71	65/73	71/**66**	76/75	78/44	72/66	43/29
		Guardian	61/**72**	**66**/**74**	72/64	78/76	**79**/**76**	71/67	**49**/**36**
	2	W/O	**56**/**69**	**56**/67	**70**/56	**79**/**77**	63/**75**	69/**68**	-
		NYTimes	49/67	52/**68**	64/**63**	75/75	**75**/50	71/66	-
		Guardian	49/68	51/**68**	64/**63**	77/76	74/52	**72**/67	-
Metaculus	1	W/O	66/74	61/70	76/73	68/66	**81**/65	**81**/74	-
		NYTimes	**71**/**78**	**79**/**80**	**81**/73	**81**/**78**	**81**/**78**	78/73	**69**/**79**
		Guardian	69/73	**79**/**80**	80/**76**	**81**/**78**	79/69	79/**76**	63/78
	2	W/O	**61**/**73**	62/67	71/73	67/66	**78**/65	**82**/**76**	-
		NYTimes	52/72	61/**78**	**72**/**74**	76/74	**78**/66	79/73	-
		Guardian	55/**73**	**66**/77	68/73	**80**/**78**	73/**68**	80/**76**	-

Table 1: Performance of individual agents on the GJOpen and Metaculus datasets, with depth 1 and depth 2. The table compares the baseline ArgLLM agents (W/O source) and the RAG-ArgLLM agents (NYTimes and Guardian source) with the base LLM models as per columns, as well as, in the last column, the RbAM agents (using the NYTimes and the Guardian). For each agent configuration, we report accuracy of 0.5 /estimated base scores. Bold indicates best results for each agent configuration, with/out the use of sources.

7.2 Multi-Agent Claim Verification

ArgLLM Agent Pairs. We first evaluate our multi-agent framework with pairs of ArgLLM agents, where each agent in the pair independently generates QBAFs, and then our multi-agent framework combines them. We experiment across six pairs of ArgLLM agents with their four variations depth 1 (with 0.5 base score and estimated base score assigned to the claim) and depth 2 (with 0.5 base score and estimated base score assigned to the claim). For each variation, we apply two aggregation functions, average and maximum, to combine argument base scores.

Table 2 (top) reports some results (full results are available in [21]). In summary, the results show variation across individual ArgLLM agents. This variation reflects differences in the agents' argumentative reasoning capabilities. Across all variations, the combination of argumentative reasoning produced by the ArgLLM agents frequently outperforms at least one of the individual agents, and in several cases outperforms both. We hypothesise that our framework improves performance when the paired agents bring complementary argumentative reasoning. For example, Table 2 (top) shows that combining an ArgLLM agent with Llama-3 as base model and another ArgLLM agent tends to improve on accuracy: this could mean that Llama-3 generates argumentative reasoning that complements that of other LLMs. The instances where the framework does not perform well are when an ArgLLM agent already performs well individually. This could suggest that the second agent contributes less useful or redundant arguments. For instance, Gemma-2 already performs well, so combining it with other agents does not improve results.

Overall, the multi-agent framework can improve forecasting performance. Average aggregation generally outperforms max aggregation, suggesting that a moderate base score is more effective than selecting the strongest base score.

	Dataset	Model	Llama-3 (D=1)			Llama-3 (D=2)		
				Avg	Max		Avg	Max
ArgLLM	GJOpen	Single	78/75			Single	63/75	
		Mixtral 67/70	68/75	77/75		56/69	43/74	49/75
		Mistral 65/71	69/75	75/75		56/67	49/75	49/75
		Gemma 73/57	73/**76**	78/**78**		50/70	70/56	50/70
		Gemma-2 79/77	78/**78**	78/**78**		79/77	50/70	50/70
	Metaculus	Single	81/65			Single	81/65	
		Mixtral 66/74	66/74	66/74		61/73	66/**76**	66/**76**
		Mistral 61/70	61/70	61/70		62/67	61/**74**	61/**74**
		Gemma 76/73	68/66	76/73		76/73	71/73	68/66
		Gemma-2 68/66	68/66	68/66		67/66	68/66	68/66
RAG-ArgLLM NY-Times	GJOpen	Single	78/44			Single	75/50	
		Mixtral 61/71	61/**71**	61/71		49/67	37/49	48/51
		Mistral 59/68	65/**73**	65/**73**		52/68	41/49	48/50
		Gemma 71/66	76/**75**	76/**75**		64/63	53/54	37/57
		Gemma-2 76/75	76/75	37/57		75/75	66/71	76/75
	Metaculus	Single	79/68			Single	78/66	
		Mixtral 71/78	71/78	71/78		52/72	41/63	52/66
		Mistral 79/80	79/80	79/80		61/78	53/67	56/65
		Gemma 81/73	81/78	81/73		72/74	64/70	44/47
		Gemma-2 81/78	81/78	81/78		76/74	71/70	44/47
RAG-ArgLLM Guardian	GJOpen	Single	78/76			Single	74/52	
		Mixtral 61/72	61/72	61/72		46/68	36/50	51/52
		Mistral 66/74	66/74	66/74		51/68	39/51	48/51
		Gemma 72/64	78/76	78/76		64/63	52/56	38/59
		Gemma-2 78/76	78/76	78/76		77/76	71/73	38/59
	Metaculus	Single	79/70			Single	73/68	
		Mixtral 69/73	69/73	69/73		55/73	45/64	55/66
		Mistral 79/80	79/80	79/80		66/77	56/66	61/66
		Gemma 80/76	**81**/**78**	80/76		68/73	62/69	45/48
		Gemma-2 81/78	81/78	81/78		80/78	76/75	45/48

Table 2: Accuracy results of our multi-agent combination framework applied to pairs of agents of the same type: (Llama-3 and one of Mixtral, Mistral, Gemma and Gemma-2-based) (i) ArgLLM agents; (ii) RAG-ArgLLM agents using NYTimes; (iii) RAG-ArgLLM agents using the Guardian, for Depth 1 (left) and Depth 2 (right). We show accuracy results using 0.5/estimated base scores and ω_{avg} (Avg) and ω_{max} (Max) base score aggregations. Bold indicates improvement over both (single) agents; underline indicates improvement over one and parity with the other.

RAG-ArgLLM Agent Pairs. Next, we evaluate our multi-agent framework across six pairs of RAG-ArgLLM agents with their four variations depth 1 (with 0.5 base score and estimated base score assigned to the claim) and depth 2 (with 0.5 base score and estimated base score assigned to the claim) and the NYTimes and the Guardian as external sources. For each pair, each RAG-ArgLLM agent is using the same source.

Full results are provided in [21]. Table 2 (middle and bottom) gives partial results. Results show that our multi-agent framework with the RAG-ArgLLM agents often outperforms at least one of the individual RAG-ArgLLM agents, and in some cases exceed both. This similarly confirms that combining QBAFs can improve performance when the paired agents bring complementary argumentative

reasoning, even when arguments are retrieved from the external sources. Notably, looking at Table 2, combinations involving Llama-3 often improves accuracy despite Llama-3 not performing well individually with the RAG-ArgLLM agents, suggesting their arguments may remain complementary when combined with other models. We also observe that average aggregation typically outperforms max, aligning with earlier findings with pairs of ArgLLM agents. Performance gains are more consistent on Metaculus, where the paired agents often match or exceed both individual agents. On the other hand, improvements on GJOpen are more dependent on the choice of the external source: the Guardian tends to be more effective than the NYTimes, consistently with the findings for the individual RAG-ArgLLM and RbAM agents.

Overall, our experiments show that combining RAG-ArgLLM agents can also be beneficial, particularly, as with pairs of ArgLLM agents, when the combination is done with average aggregation.

ArgLLM and RAG-ArgLLM Agent Pairs. We then explore the performance of combining two different agents, one standard ArgLLM agent with one RAG-ArgLLM agent (using either the NYTimes or the Guardian) or two different RAG-ArgLLM agent, each using a different source.

Full results for combining two RAG-ArgLLM agents, where one uses the NYTimes and the other uses the Guardian as external sources, are provided in [21]. In summary, the combination of the RAG-ArgLLM agents with different sources often yields significant improvements, particularly in depth 1. RAG-ArgLLM agents with Mistral as base LLM and The Guardian as the source also proves to be a consistently strong complementary agent in depth 1, improving or maintaining accuracy in most pairings across both datasets (five out of six in both). However, performance in depth 2 is less consistent and often leads to a degradation in accuracy. This is likely due to the generation of irrelevant arguments in the second layer. Nonetheless, even in depth 2 some improvement is still observable over individual agents.

Results for pairings of one ArgLLM agent and one RAG-ArgLLM agent using NYTimes or Guardian as external sources are provided in [21]. In summary, the results indicate that this hybrid approach is highly effective. The RAG-ArgLLM agents introduce externally grounded arguments that can complement the internal reasoning of the ArgLLM agents. The benefit is particularly clear when using the NYTimes. For example, Mistral (NYTimes) often improves or maintains the accuracy of its ArgLLM partner on the GJOpen dataset (five out of six instances) for average aggregation. The Guardian source also shows strong performance in specific pairings. However, the RAG-ArgLLM agents using the Guardian do not seem to improve accuracy as much when they are combined with the ArgLLM agents.

Overall, our experiments demonstrate that combining different external sources can be beneficial. The most significant gains are achieved when agents leverage different external sources.

Three-Agent Combination. We extend the analysis of our multi-agent framework to three agents, where a standard ArgLLM agent is combined with two RAG-ArgLLM agents, one using the NYTimes and the other the Guardian. The results, provided in [21], demonstrate the potential to increase accuracy by increasing the diversity of argumentative reasoning. The addition of a third agent often leads to more robust and accurate forecasts, particularly on the GJOpen dataset. The most consistent improvements are observed when the base ArgLLM agent is a strong performer itself, such as with Gemma-2 or GPT-4o. This suggests that when an agent's internal reasoning is strong, it provides a solid foundation that can be effectively enhanced by complementary information from external sources, thereby improving the forecasting accuracy. While the three-agent combination does not always outperform the best individual agent in the trio, it consistently shows a clear improvement over the other two individual agents. On the Metaculus dataset, the benefits are also apparent, although the strong individual agent performance sometimes sets a high bar for improvement. Both aggregation methods show instances of improvement. However, average aggregation performs better than max aggregation in most instances, supporting the evidence from previous experiments.

Overall, the three-agent framework, by increasing the diversity of complementary reasoning, can improve forecasting accuracy whether from different models or different external data sources.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a novel multi-agent framework which integrates multiple perspectives by combining QBAFs generated by diverse LLM-based agents. We designed and evaluated three distinct agent types: baseline ArgLLM agents (which rely on internal LLM knowledge), RAG-ArgLLM agents and RbAM agents (both grounding their reasoning in external sources).

Our empirical evaluation on the GJOpen and Metaculus datasets yielded several key insights. First, the RAG-ArgLLM agents consistently outperformed the ArgLLM agents, especially on the Metaculus dataset. This confirms that grounding argumentative reasoning with external evidence can improve forecasting accuracy. While the RbAM agents showed promise on the Metaculus dataset, they did not perform well on the GJOpen dataset. Second, our multi-agent framework applied to pairs of the same agent types, in some instances improved forecasting accuracy. Performance gains were most significant when the agents generated complementary argumentative perspectives. Pairing an ArgLLM agent with a RAG-ArgLLM agent, or combining two RAG-ArgLLM agents using different external sources (the NYTimes and the Guardian), proved highly effective. This suggests integrating both internal LLM reasoning and external evidence can be beneficial. Finally, instantiating our framework with three agents, an ArgLLM agent and two RAG-ArgLLM agents with different sources, yielded further improvements. By increasing the diversity of reasoning and external sources, the three-agent framework produced more robust reasoning. This suggests that combining complementary perspectives yields better results.

Building on these findings, we plan to explore several promising directions: (1) investigating in more depth which models to pick while combining their argumentative reasoning; (2) studying whether an interactive debate could improve forecasting accuracy (e.g. using a similar framework to Rago et al. [30]); (3) evaluating the framework's robustness against real-world data irregularities such as noisy/misleading data, biased sources, and a wider variety of question types; and (4) investigating the practical deployment of the system, including optimising the computational cost and addressing the integration challenges into existing workflows.

ACKNOWLEDGMENTS

This research was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (grant agreement no. 101020934, ADIX), by J.P. Morgan and by the Royal Academy of Engineering, UK, under the Research Chairs and Senior Research Fellowships scheme (grant agreement no. RCSRF2021\11\45).

REFERENCES

[1] Mahdi Abolghasemi, Odkhishig Ganbold, and Kristian Rotaru. 2023. Humans vs Large Language Models: Judgmental Forecasting in an Era of Advanced AI. *CoRR* (2023). https://doi.org/10.48550/ARXIV.2312.06941

[2] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. 2017. Towards Artificial Argumentation. *AI Magazine* 38, 3 (2017), 25–36. https://doi.org/10.1609/aimag.v38i3.2704

[3] Marco Aurisicchio, Pietro Baroni, Dario Pellegrini, and Francesca Toni. 2015. Comparing and Integrating Argumentation-Based with Matrix-Based Decision Support in Arg&Dec. In *Theory and Applications of Formal Argumentation - Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers*. Springer. https://doi.org/10.1007/978-3-319-28460-6_1

[4] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre (Eds.). 2018. *Handbook of Formal Argumentation*. College Publications.

[5] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From Fine-Grained Properties to Broad Principles for Gradual Argumentation: A Principled Spectrum. *International Journal of Approximate Reasoning* 105 (Feb. 2019), 252–286. https://doi.org/10.1016/j.ijar.2018.11.019

[6] Lucas Carstens and Francesca Toni. 2015. Towards Relation Based Argumentation Mining. In *Proc. 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*. ACL. https://doi.org/10.3115/V1/W15-0504

[7] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *ECSQARU (Lecture Notes in Computer Science)*. Springer. https://doi.org/10.1007/11518655_33

[8] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2011. Weighted Argumentation Systems: A Tool for Merging Argumentation Systems. In *IEEE 23rd Int'l Conf. on Tools with Artificial Intelligence, ICTAI 2011, Boca Raton, FL, USA, November 7-9, 2011*. IEEE Computer Society. https://doi.org/10.1109/ICTAI.2011.99

[9] Dhivya Chandrasekaran and Vijay Mago. 2022. Evolution of Semantic Similarity - A Survey. *ACM Comput. Surv.* 54, 2 (2022), 41:1–41:37. https://doi.org/10.1145/3440755

[10] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. 2005. Merging Argumentation Systems. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*. AAAI Press / The MIT Press. http://www.aaai.org/Library/AAAI/2005/aaai05-096.php

[11] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. 2007. On the Merging of Dung's Argumentation Systems. *Artif. Intell.* 171, 10-15 (2007), 730–753. https://doi.org/10.1016/J.ARTINT.2007.04.012

[12] Jérôme Delobelle, Adrian Haret, Sébastien Konieczny, Jean-Guy Mailly, Julien Rossit, and Stefan Woltran. 2016. Merging of Abstract Argumentation Frameworks. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*. AAAI Press. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12872

[13] Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim Verification in the Age of Large Language Models: A Survey. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2408.14317

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2407.21783

[15] Valentinos Evripidou and Francesca Toni. 2014. Quaestio-it.com: a social intelligent debating platform. *J. Decis. Syst.* 23, 3 (2014), 333–349. https://doi.org/10.1080/12460125.2014.886496

[16] Giuseppe Fenza, Domenico Furno, Vincenzo Loia, and Pio Pasquale Trotta. 2025. Multi-LLM Agents Architecture for Claim Verification. In *Proc. Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), Bologna, Italy, February 03-08, 2025 (CEUR Workshop Proceedings)*. CEUR-WS.org. https://ceur-ws.org/Vol-3962/paper20.pdf

[17] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. In *AAAI-25, February 25 - March 4, 2025, Philadelphia, PA, USA*. AAAI Press. https://doi.org/10.1609/AAAI.V39I14.33637

[18] Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. 2025. Resolving Conflicting Evidence in Automated Fact-Checking: A Study on Retrieval-Augmented LLMs. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2505.17762

[19] Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Argumentatively Coherent Judgmental Forecasting. In *ECAI 2025 - 28th European Conference on Artificial Intelligence, October 25 to 30, 2025, Bologna, Italy*.

[20] Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can Large Language Models perform Relation-based Argument Mining?. In *Proc. 31st Int'l Conf. on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*. ACL. https://aclanthology.org/2025.coling-main.569/

[21] Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Retrieval- and Argumentation-Enhanced Multi-Agent LLMs for Judgmental Forecasting (Extended Version with Supplementary Material). *CoRR* (2025). arXiv:2510.24303 URL https://doi.org/10.48550/arXiv.2510.24303. Full version of this paper.

[22] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching Human-Level Forecasting with Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. http://papers.nips.cc/paper_files/paper/2024/hash/5a5acfd0876c940d81619c1dc60e7748-Abstract-Conference.html

[23] Benjamin Irwin, Antonio Rago, and Francesca Toni. 2022. Forecasting Argumentation Frameworks. In *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel, July 31 - August 5, 2022*. https://proceedings.kr.org/2022/55/

[24] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *CoRR* (2023). https://doi.org/10.48550/ARXIV.2310.06825

[25] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of Experts. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2401.04088

[26] Michael Lawrence, Paul Goodwin, Marcus O'Connor, and Dilek Önkal. 2006. Judgmental Forecasting: A Review of Progress Over the Last 25 Years. *International Journal of Forecasting* 22, 3 (2006), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

[27] Lucas Leite, Thiago Alves Rocha, and João F. L. Alcântara. 2015. Merging Argumentation Systems. In *2015 Brazilian Conference on Intelligent Systems, BRACIS 2015, Natal, Brazil, November 4-7, 2015*. IEEE Computer Society. https://doi.org/10.1109/BRACIS.2015.45

[28] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2403.08295

[29] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[30] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proc. 20th Int'l Conf. on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*. https://doi.org/10.24963/kr.2023/57

[31] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*. AAAI Press. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874

[32] Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2408.00118

[33] Philipp Schoenegger and Peter S. Park. 2023. Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament. *CoRR* (2023). https://doi.org/10.48550/ARXIV.2310.13014

[34] Philipp Schoenegger, Peter S. Park, Ezra Karger, and Philip E. Tetlock. 2024. AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2402.07862

[35] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2025. Jina Embeddings V3: Multilingual Text Encoder with Low-Rank Adaptations. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V (Lecture Notes in Computer Science)*. Springer. https://doi.org/10.1007/978-3-031-88720-8_21

[36] Haoran Wang and Kai Shu. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *Findings of ACL: EMNLP 2023, Singapore, December 6-10, 2023*. ACL. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.416

[37] Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments. In *Proc. 62nd Annual Meeting of ACL (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. ACL. https://doi.org/10.18653/V1/2024.ACL-LONG.556

[38] Maximilian Zellner, Ali E. Abbas, David V. Budescu, and Aram Galstyan. 2021. A Survey of Human Judgement and Quantitative Forecasting Methods. *Royal Society Open Science* 8, 2 (2021), 201187. https://doi.org/10.1098/rsos.201187

[39] Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In *Proc. 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of ACL, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*. ACL. https://doi.org/10.18653/V1/2023.IJCNLP-MAIN.64

[40] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2402.19473

[41] Yingming Zheng, Xiaoliang Liu, Peng Wu, and Li Pan. 2025. CRAVE: A Conflicting Reasoning Approach for Explainable Claim Verification Using LLMs. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2504.14905

[42] Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. ArgRAG: Explainable Retrieval Augmented Generation using Quantitative Bipolar Argumentation. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2508.20131