

# ArgLLM-App: An Interactive System for Argumentative Reasoning with Large Language Models

Demonstration Track

Adam Dejl\*  
Imperial College London  
London, United Kingdom  
adam.dejl18@imperial.ac.uk

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

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

## ABSTRACT

Argumentative LLMs (ArgLLMs) are an existing approach leveraging Large Language Models (LLMs) and computational argumentation for decision-making, with the aim of making the resulting decisions faithfully explainable to and contestable by humans. Here we propose a web-based system implementing ArgLLM-empowered agents for binary tasks. ArgLLM-App supports visualisation of the produced explanations and interaction with human users, allowing them to identify and contest any mistakes in the system’s reasoning. It is highly modular and enables drawing information from trusted external sources. ArgLLM-App is publicly available at [argllm.app](http://argllm.app), with a video demonstration at [youtu.be/vzwlGOr0sPM](https://youtu.be/vzwlGOr0sPM).

## KEYWORDS

Argumentation; LLMs; RAG

### ACM Reference Format:

Adam Dejl\*, Deniz Gorur\*, and Francesca Toni. 2026. ArgLLM-App: An Interactive System for Argumentative Reasoning with Large Language Models: Demonstration Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. <https://doi.org/10.65109/NWEK7590>

## 1 INTRODUCTION

Large Language Models (LLMs) have recently emerged as a major class of AI models. Specifically, the diversity of knowledge encoded in LLMs and their ability to apply this knowledge zero-shot in a range of settings makes them promising candidates for use in decision-making. However, they are currently limited by their inability to reliably provide outputs that are faithfully explainable and contestable when mistaken. Freedman et al. [3] introduced *Argumentative LLMs* (ArgLLMs in short) to reconcile these strengths and weaknesses. ArgLLMs can utilise (arbitrary) LLMs to construct quantitative bipolar argumentation frameworks (QBAFs) [7], which then serve as the basis for formal reasoning in decision-making supported by gradual semantics for QBAFs [2]. The interpretable nature of these argumentation frameworks and formal reasoning means that any decision made by ArgLLMs may be faithfully explained to, and contested by, humans. To date, ArgLLMs have been

proven effective, as well as competitive with state-of-the-art techniques such as chain-of-thought [8], in the binary decision-making task of claim verification.

Our main contribution in this demo paper is a system implementing ArgLLM-empowered agents for any binary decision-making task (including claim verification as in [3]). We call the system *ArgLLM-App* to emphasise that it can be naturally used to build applications with ArgLLMs. ArgLLM-App is a web application supported by a server implementing the ArgLLM logic and mediating access to the base LLM. It is highly flexible with respect to the depth and breadth of generated QBAFs, and the gradual semantics for the evaluation of opinions in these QBAFs. The system also supports interaction with human users, enabling them to modify aspects of the generated QBAF, either by adjusting the base confidence in arguments (in QBAF terminology: the arguments’ base scores) or expanding the QBAF (in QBAF terminology: adding attackers/supporters of arguments). This can be done directly via a QBAF visualisation or indirectly through a chat interface. Such modifications lead ArgLLM-App to revise its assessment of the binary decision-making task at hand. Human users can also upload trusted sources, inputted as PDFs, to inform the generation of QBAFs, in the spirit of Retrieval-Augmented Generation (RAG) [5].

## 2 PRELIMINARIES

QBAFs [2] are quadruples  $(\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$  consisting of a finite set of *arguments*  $\mathcal{A}$ , disjoint binary relations of *attack*  $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$  and *support*  $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ , and a *base score function*  $\tau : \mathcal{A} \rightarrow [0, 1]$ .

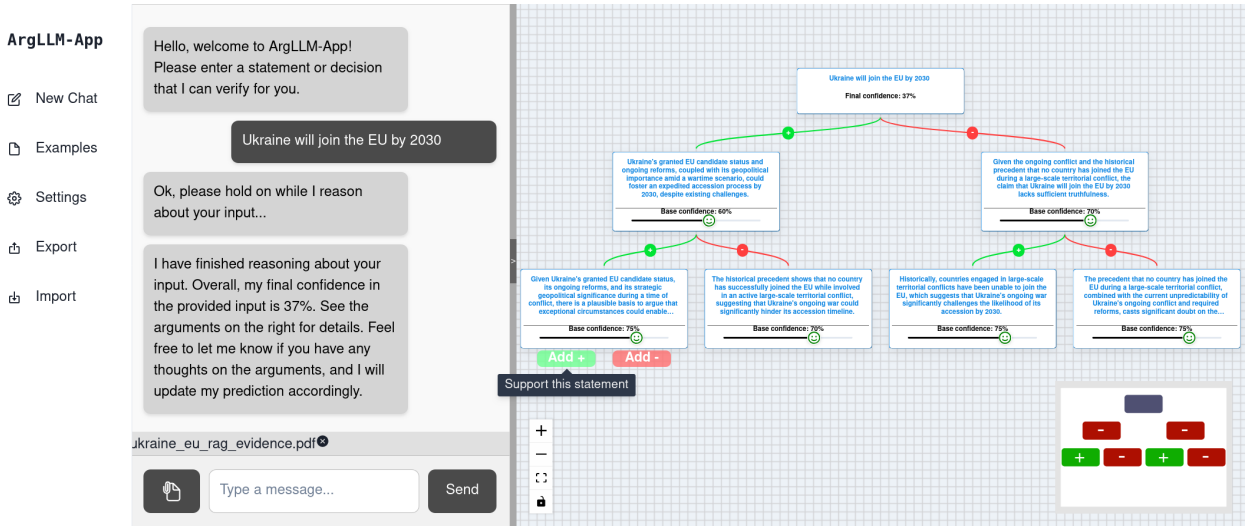
QBAFs can be seen as graphs with arguments as nodes and elements of the attack and support relations as edges. ArgLLMs adopt a graphical visualisation as illustrated in Figure 1, right. Here, the QBAF has seven arguments (given by boxes) and attack and support relations (given by red/- and green/+ edges, respectively) with three elements each. The arguments are a claim (“Ukraine will join the EU before 2030”), evidence for/against the claim, or evidence for/against other evidence. The base scores are indicated as the *base confidence*: their value is determined by the underpinning LLM by direct questioning [3]. In the figure, as in ArgLLMs, the *final confidence* of the claim argument is computed by a *gradual semantics*. For QBAFs, this is a function  $\sigma : \mathcal{A} \rightarrow [0, 1]$ , often involving an iterative procedure initialising strength values with the base scores and then repeatedly updating the strength of arguments based on the strength of their attackers and supporters. In the figure,  $\sigma$  is DF-QuAD [7], but ArgLLM-App also accommodates other semantics.

In ArgLLMs, QBAFs, seen graphically, are restricted to trees with claims as their root (as in Figure 1). When they are generated with the help of LLMs, they are either of *depth 1* (with a single layer of

\*These authors contributed equally. We thank Pranay Padavala for his early input.



This work is licensed under a Creative Commons Attribution International 4.0 License.



**Figure 1: Output of ArgLLM-App for claim “Ukraine will join the EU before 2030”, with settings as in Figure 2 and document-based QBAF generation (see the PDF document indication at the bottom of the chat). In the QBAF, the bottom left argument has been selected for addition of a supporter at depth 3. The abstract graph at the bottom right shows the QBAF structure.**

attackers and supporters for the claim) or of *depth 2* (with each attacker or supporter in turn having its own attacker and supporter, as in Figure 1). In both cases, in ArgLLMs, arguments have a single supporter and a single attacker (*breadth 1*). ArgLLM-App allows a choice between depth 1 and 2 options as well as a choice of breadth. Note that we limit the depth to 2 to avoid overloading the users.

### 3 ARGLLM-APP FEATURES

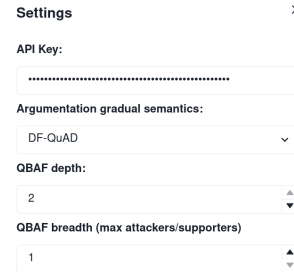
ArgLLM-App is highly customisable. It allows the choice of the following parameters:

- QBAF depth (1 or 2);
- QBAF breadth (up to 4, i.e. up to 4 attackers and 4 supporters per argument, the same number);
- gradual semantics (DF-QuAD [7], Euler-based [1] and Quadratic Energy [6]).

The values of the parameters can be configured before interacting with the system (see Figure 2). For binary decisions given in the chat, ArgLLM-App generates outputs as illustrated in Figure 1.

ArgLLM-App is highly interactive:

- Users can modify the generated outputs by using the base confidence slider to express what they deem to be the correct confidence in evidence for/against claims or (in the case of depth 2) other evidence.
- Users can add additional supporters or attackers by using the Add buttons associated with arguments (Figure 1 illustrates how to initiate the addition of a supporter).
- Users can augment the internal LLM knowledge through document-based QBAF generation with documents specified in a PDF format. These PDFs are parsed to Markdown format and incorporated into the LLM prompts. Alternatively, users can provide additional information in the chat, which ArgLLM-App can autonomously turn into attackers or supporters of the relevant arguments.



**Figure 2: Settings: in addition to the API key for access to the base LLM, the user can configure the gradual semantics, depth and breadth to be used by ArgLLM-App.**

### 4 FUTURE WORK

ArgLLM-App is a prototype that can be extended in several directions. The system could be adjusted to support QBAF depths higher than 2 while avoiding cognitive overload of users. It could also be enhanced to enable the upload of documents in formats other than PDF. Other ways to compute base confidence may also be useful [9]. The current realisation allows the use of only base LLMs from OpenAI, and only single LLMs. We envisage the support of LLMs from other providers, and multi-agent variants where different agents rely upon different LLMs, in the spirit of [4]. We are currently allowing debate of a single binary decision: multiple decisions/question answering would also be useful. Further, it would be interesting to fully integrate RAG [5] into ArgLLM-App, allowing our agents to find relevant sources autonomously and extract relevant arguments therefrom, as in [4]. Finally, we accommodate interactions with single users: multiple users interacting concurrently with the system and each other may bring additional value.

## ACKNOWLEDGEMENTS

This research was partially supported by ERC under the EU’s Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX). We thank all members of the Computational Logic and Argumentation group at Imperial College London, and in particular Antonio Rago and Gabriel Freedman, for their suggestions and feedback on earlier versions of this work. We also thank Pranay Padavala for his early contributions.

## REFERENCES

- [1] Leila Amgoud and Jonathan Ben-Naim. 2017. Evaluation of Arguments in Weighted Bipolar Graphs. In *ECSQARU 2017*. Springer. [https://doi.org/10.1007/978-3-319-61581-3\\_3](https://doi.org/10.1007/978-3-319-61581-3_3)
- [2] Pietro Baroni, Antonio Rago, and Francesca Toni. 2018. How Many Properties Do We Need for Gradual Argumentation?. In *AAAI 2018*. AAAI Press. <https://doi.org/10.1609/aaai.v32i1.11544>
- [3] 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 2025*. AAAI Press. <https://doi.org/10.1609/AAAI.V39I14.33637>
- [4] Deniz Gorur, Antonio Rago, and Francesca Toni. 2026. Retrieval- and Argumentation-Enhanced Multi-Agent LLMs for Judgmental Forecasting. In *AAMAS 2026*. International Foundation for Autonomous Agents and Multiagent Systems / ACM. <https://doi.org/10.65109/SNBR1486>
- [5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS 2020*. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [6] Nico Potyka. 2018. Continuous Dynamical Systems for Weighted Bipolar Argumentation. In *KR 2018*. AAAI Press. <https://aaai.org/ocs/index.php/KR/KR18/paper/view/17985>
- [7] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *KR 2016*. AAAI Press. <http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874>
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS 2022*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html)
- [9] Kevin Zhou, Adam Dejl, Gabriel Freedman, Lihu Chen, Antonio Rago, and Francesca Toni. 2025. Evaluating Uncertainty Quantification Methods in Argumentative Large Language Models. In *EMNLP 2025*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.1184>