Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR meets Machine Learning and Explanation

# Argumentation for Explainable and Globally Contestable Decision Support with LLMs

**Adam Dejl**[1] , **Matthew Williams**[2] , **Francesca Toni**[1]
[1]Department of Computing, Imperial College London
[2]Department of Surgery & Cancer, Imperial College London
{adam.dejl18, matthew.williams, ft}@imperial.ac.uk

## Abstract

Large language models (LLMs) exhibit strong general capabilities, but their deployment in high-stakes domains is hindered by their opacity and unpredictability. Recent work has taken meaningful steps towards addressing these issues by augmenting LLMs with post-hoc reasoning based on computational argumentation, providing faithful explanations and enabling users to contest incorrect decisions. However, this paradigm is limited to pre-defined binary choices and only supports local contestation for specific instances, leaving the underlying decision logic unchanged and prone to repeated mistakes. In this paper, we introduce ArgEval, a framework that shifts from instance-specific reasoning to structured evaluation of general decision options. Rather than mining arguments solely for individual cases, ArgEval systematically maps task-specific decision spaces, builds corresponding option ontologies, and constructs general argumentation frameworks (AFs) for each option. These frameworks can then be instantiated to provide explainable recommendations for specific cases while still supporting global contestability through modification of the shared AFs. We investigate the effectiveness of ArgEval on treatment recommendation for glioblastoma, an aggressive brain tumour, and show that it can produce explainable guidance aligned with clinical practice.

## 1 Introduction

Trained on large corpora of general textual data as well as specific problem-solving tasks, large language models (LLMs) have demonstrated strong performance in a variety of settings (Bubeck et al. 2023; Van Veen et al. 2024; Luo et al. 2025; McDuff et al. 2025; Bermejo et al. 2025). However, due to their reliance on stochastic next-token prediction, LLMs can also be notoriously unreliable, as evidenced by issues such as hallucinations (Huang et al. 2025) and omissions of critical information (Busch et al. 2025). These issues pose a substantial obstacle to safely using LLMs in high-stakes domains. Importantly, the inherent opacity of LLMs makes it impossible to faithfully explain their outputs or to reliably redress issues. While methods

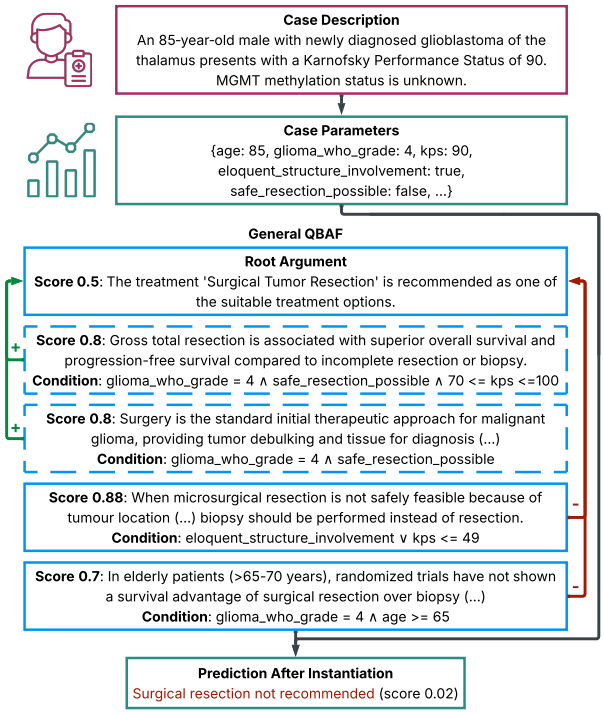


Figure 1: Illustration of ArgEval inference for one of the treatment options for glioblastoma[1]. **Case Parameters** extracted from the **Case Description** are used to instantiate the **General QBAF** associated with the given option (**Root Argument**), removing the dashed nodes whose conditions are not satisfied. The **Prediction After Instantiation** is obtained from the instantiated framework, which also serves as a faithful explanation.

such as chain-of-thought (Wei et al. 2022) or verbalised explanations can provide some degree of rationalization, they have been found to be unfaithful, i.e., inaccurate in describing the true internal reasoning of models (Chen et al. 2025).

In this work, we introduce *ArgEval*, a new framework for decision support with LLMs that leverages argumentative reasoning to produce explainable predictions that can be globally contested. Like existing approaches combining LLMs with argumentation (Freedman et al. 2025; Zhou et al. 2025; Zhu et al. 2025; Gorur, Rago, and Toni 2025), ArgEval uses an LLM for mining arguments for and

[1]Icons from Noun Project (Patient by Alzam and Data by Anna Riana, CC BY 3.0).

[2]Icons from Noun Project (Policy by Puspito, Decision by Kamin Ginkaew, report analysis by kartini 1, Patient by Alzam and Data by Anna Riana, CC BY 3.0).

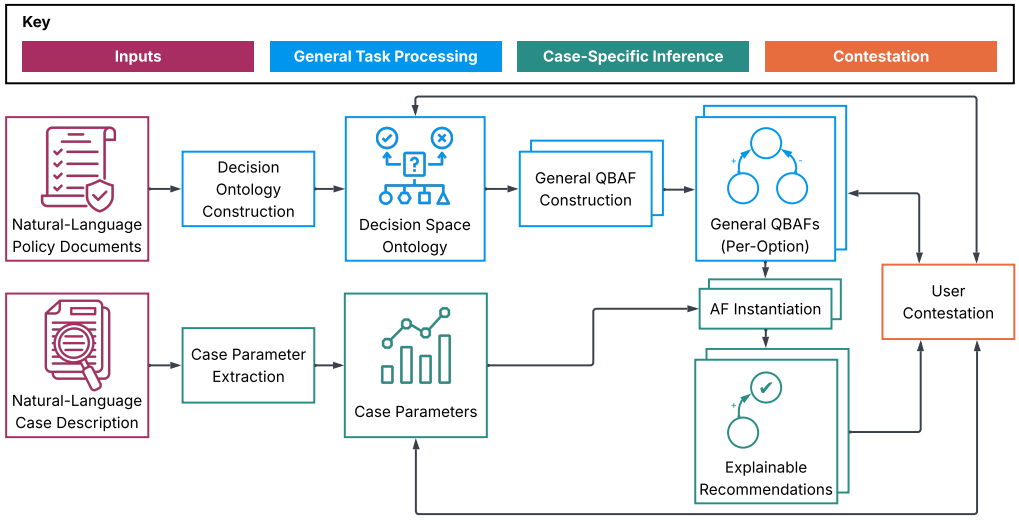


Figure 2: Overview of the ArgEval pipeline[2]. Top: given natural-language policy documents that specify general criteria for decision-making in a certain domain, ArgEval builds a decision-space ontology and constructs general QBAFs for each candidate decision in the ontology. Bottom: at inference time, the general QBAF is instantiated with the parameters of a specific case, providing faithfully explainable recommendations. Users can contest the decision-space ontology, the general QBAFs, the extracted case parameters and the general parameter schema specifying the properties to be extracted, in response to incorrect recommendations or explanations produced by the model.

against a particular decision as well as estimating numerical scores indicating their intrinsic strength (also called *base scores*). The arguments and scores are then used to form a *quantitative bipolar argumentation framework* (QBAF), enabling deterministic inference via gradual semantics (Baroni, Rago, and Toni 2019). Applying the semantics yields a set of final argument strengths, which indicate the degree of acceptability of each argument after considering its interactions with the other arguments in the QBAF. The strength of the argument representing the considered decision can then be used to make predictions, with the QBAF acting as an inherently faithful explanation for these predictions.

While sharing the use of QBAFs with prior works, ArgEval differs in several ways. Unlike ArgLLMs (Freedman et al. 2025; Zhou et al. 2025) and like ArgRAG (Zhu et al. 2025), ArgEval uses external sources to guide the mining of QBAFs. However, these sources are used to compile a *decision-space ontology* and *general QBAFs*, respectively representing the available decision options and general knowledge concerning the decision-making problem of interest, differing from ArgRAG, which only uses them to make predictions relating to specific instances. Unlike both ArgLLMs and ArgRAG, ArgEval is not restricted to binary questions (claims) but is applicable to open-ended decisions, in the spirit of Evaluative AI (Miller 2023).

Uniquely, when making predictions for specific cases, ArgEval instantiates the general QBAFs associated with each decision option rather than constructing these QBAFs from scratch (see illustration in Figure 1). As with ArgLLMs and ArgRAG, the resulting instantiated QBAFs can be used as faithful explanations. Users can inspect these QBAFs and *contest* them to correct any mistakes, either by changing argument base scores or by adding additional arguments. However, ArgEval shifts contestation from case-specific, *local contestability*, as supported by ArgLLMs and ArgRAG, to *global contestability*, potentially affecting other future cases. Since the arguments in the instantiated QBAF directly correspond to arguments in the general QBAF, any changes to those arguments will directly affect predictions for any cases satisfying their applicability conditions. Apart from the QBAFs, users may also inspect and adjust the decision-space ontology, the extracted case parameters used for QBAF instantiation and the general parameter schema specifying the properties to be extracted for each case.

To evaluate the effectiveness of ArgEval, we apply it to the task of recommending suitable treatments for different cases of glioblastoma based on relevant clinical guidelines. We show that variants of ArgEval perform competitively against the reasoning LLM baseline and ArgLLMs-O, an adapted version of ArgLLMs using the extracted decision-space ontology for supporting open-ended decisions. Importantly, ArgEval achieves this performance using a fraction of the computational cost associated with the other methods while providing a significant qualitative advantage in the form of global contestability.

To summarise, our key contributions are as follows:

- We propose ArgEval (Figure 2), a new framework for decision support that provides faithful explainability and global contestability through argumentation.
- We apply ArgEval to treatment recommendation for glioblastoma brain tumours, showing that it exhibits competitive performance compared to other methods at a fraction of the inference cost.
- We demonstrate the benefits of global contestability in a case study where contestation on a single sample substantially improves the performance of the model.

| Property | Method | | | |
|---|---|---|---|---|
| | LLMs | ArgLLMs ArgRAG | ArgLLMs-O | ArgEval |
| Open-Ended Decisions | ✓ | ✗ | ✓ | ✓ |
| Faithful Explainability | ✗ | ✓ | ✓ | ✓ |
| Local Contestability | (✓) | ✓ | ✓ | ✓ |
| Global Contestability | ✗ | ✗ | ✗ | ✓ |
| Inference Speed | ↯↯ | ↯ | ↯ | ↯↯↯ |

Table 1: Qualitative comparison between ArgEval and other methods. (✓) in the third row signifies that local contestability for LLMs relies on their opaque mechanisms rather than being fully transparent. A higher number of ↯ symbols denotes faster inference. More details regarding the listed properties are provided in Section 4.3.

## 2 Related Work

**Argumentation and LLMs** The use of argumentation in combination with LLMs is advocated by several. Argumentative LLMs (ArgLLMs) (Freedman et al. 2025) have been recently proposed as a step towards making LLMs more reliable and responsive to user-driven error corrections (i.e., contestations), in the form of a paradigm combining general LLM knowledge and capabilities with formal argumentative reasoning. We adopt the same philosophy as ArgLLMs and use them as a baseline to compare against, but we are not restricted to binary decisions (such as determining whether a certain claim is truthful or not) or local contestation.

In a similar spirit, ArgRAG (Zhu et al. 2025) uses retrieval-augmented generation (RAG) with external sources for generating QBAFs, making binary judgements about claim validity. Like ArgRAG, we inform our QBAF generation with external sources, but these are compiled into a structured ontology instead of being retrieved ad hoc as in ArgRAG. Moreover, similarly to ArgLLMs and unlike ArgRAG, we restrict the QBAF extraction to acyclic graphs.

Both ArgLLMs and ArgRAG were originally proposed for the task of claim verification, and have since been adapted to other use-cases, such as judgemental forecasting (Gorur, Rago, and Toni 2025). In addition to approaches that extract full QBAFs, some works use LLMs for mining attack and support relations between pieces of text treated as arguments, as in (Gorur, Rago, and Toni 2025), or arguments, their components and/or relations between them, as in (Cabessa, Hernault, and Mushtaq 2025).

**Argumentation Schemes and Ontologies** Our general QBAFs, such as those used in the glioblastoma treatment recommendation experiment, can be informed by (existing or new) argumentation schemes with critical questions (Walton, Reed, and Macagno 2008), to be instantiated for specific cases of interest. This approach is inspired by (Ayoobi et al. 2026), though their method does not use LLMs or any ontology to instantiate the argument schemes. Our use of ontologies for QBAF instantiation is related to (Rago et al. 2025), which, however, uses them directly in combination with text mining rather than to form general QBAFs.

**Contestable AI** Contestability is increasingly advocated as important in human-in-the-loop AI systems (Lyons, Velloso, and Miller 2021), with argumentation indicated as particularly suitable to support AI contestability (Leofante et al. 2024; Dignum et al. 2025). However, existing approaches to contestability via argumentation, notably (Freedman et al. 2025; Yin et al. 2025), focus on local contestability, for single input-output pairs. (De Angelis, Proietti, and Toni 2025) proposes a form of global contestability, where learnt argumentation-based models are contested, but these are structured argumentation frameworks rather than QBAFs.

**Faithfulness of Explanations** While often proposed as a simple way of explaining LLM reasoning, chain-of-thought (CoT) has been found to be unfaithful to the true reasoning process of models, failing to mention factors that affect the outputs (Chen et al. 2025). Thus, CoT is insufficient for faithful interpretability (Barez et al. 2025). Some works, e.g. (Atanasova et al. 2023), study the faithfulness of various post-hoc explanation methods for LLMs' outputs, e.g. saliency maps and counterfactuals. Other works, e.g. (Siegel et al. 2024), propose faithfulness metrics for explanations generated as reasoning traces by LLMs. ArgEval, like ArgLLMs and ArgRAG, instead opts for drawing explanations from generated QBAFs, in the spirit of (Čyras et al. 2021), which are faithful by construction.

## 3 Preliminaries

A *QBAF* is a quadruple $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ where:

- $\mathcal{A}$ is a finite set of *arguments*;
- $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$ is a binary *attack* relation;
- $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ is a binary *support* relation;
- $\mathcal{R}^- \cap \mathcal{R}^+ = \emptyset$;
- $\tau : \mathcal{A} \to [0, 1]$ is a *base score function*.

Arguments in QBAFs may be evaluated by *gradual semantics* (Baroni, Rago, and Toni 2019), a total function $\sigma : \mathcal{A} \to [0, 1]$ which, for any $\alpha \in \mathcal{A}$, assigns a *strength* $\sigma(\alpha)$ to $\alpha$. While $\tau(\alpha)$ can be seen as the *intrinsic strength* for $\alpha$, the strength $\sigma(\alpha)$ can be seen as "dialectical", following the debate captured by $\mathcal{R}^-$ and $\mathcal{R}^+$. For a specific QBAF $\mathcal{Q}$, we let $\sigma_{\mathcal{Q}}(\alpha)$ denote the strength of $\alpha \in \mathcal{A}$ in $\mathcal{Q}$.

In this paper we adopt the following restrictions on QBAFs. Each QBAF is a tree, with the root being an argument representing a decision option (similarly to ArgLLMs, where the root is a claim). Each such tree has either *depth 1* or *depth 2*. At depth 1, the root is attacked and/or supported by a finite number of arguments; at depth 2, each of those arguments can be attacked and/or supported by any number of further arguments. Unlike ArgLLMs, which allow at most one attacker and one supporter for each argument, our QBAFs can have arbitrary breadth.

For the experiments, we use the *discontinuity-free quantitative argumentation debate* (DF-QuAD) gradual semantics (Rago et al. 2016), which is defined as follows. For any $\alpha \in \mathcal{A}$ with $n \geq 0$ attackers with strengths $v_1, \ldots, v_n$, $m \geq 0$ supporters with strengths $v'_1, \ldots, v'_m$ and $\tau(\alpha) = v_0$, $\sigma(\alpha) = \mathcal{C}(v_0, \mathcal{F}(v_1, \ldots, v_n), \mathcal{F}(v'_1, \ldots, v'_m))$, where

**Algorithm 1** Decision-Space Ontology Construction

**Input**: A set of documents $\mathcal{D}$ and a set of text chunks $\mathcal{T}'_d$ for each document $d \in \mathcal{D}$
**Output**: Constructed ontology $\mathcal{O}$

Initialise $\mathcal{E} \leftarrow \emptyset$, $\mathcal{T} \leftarrow \emptyset$, $\mathcal{H} \leftarrow \emptyset$, $\mathcal{S} \leftarrow \emptyset$
**for** $d \in \mathcal{D}$ **do**
  **for** $t \in \mathcal{T}'_d$ **do**
    ▷ *Extract chunk entities and hierarchy relations*
    $\langle \mathcal{E}_t, \mathcal{H}_t \rangle \leftarrow \texttt{mine\_ontology}(\mathcal{E}, \mathcal{H}, t)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_t$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$
    $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_t$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(e, t) \mid e \in \mathcal{E}_t\}$
  **end for**
**end for**
$\mathcal{O} \leftarrow \langle \mathcal{E}, \mathcal{T}, \mathcal{H}, \mathcal{S} \rangle$
**return** $\mathcal{O}$

- 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 $\mathcal{C}(v_0, v_a, v_s) = v_0$; else if $v_a > v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0 - (v_0 \cdot |v_s - v_a|)$; otherwise $\mathcal{C}(v_0, v_a, v_s) = v_0 + ((1 - v_0) \cdot |v_s - v_a|)$; and
- given $n$ arguments with strengths $v_1, \ldots, v_n$, if $n = 0$ then $\mathcal{F}(v_1, \ldots, v_n) = 0$, otherwise $\mathcal{F}(v_1, \ldots, v_n) = 1 - \prod_{i=1}^{n}(|1 - v_i|)$.

*Argument schemes* (Walton, Reed, and Macagno 2008) are templates for debate composed of major/minor premises, conclusions and critical questions about the minor premises' validity. The schemes guide how arguments are built and the critical questions point to ways to attack them.

## 4 ArgEval

The ArgEval pipeline (illustrated in Figure 2) consists of two main stages: general task processing, which is responsible for mapping the task domain and identifying general rules for making decisions in that domain, and case-specific inference, which provides decision recommendations for specific instances. We describe the specifics of each stage below.

### 4.1 General Task Processing

**Decision-Space Ontology Construction** The initial step of the general task processing automatically maps the decision space associated with the given domain, constructing a structured ontology of possible decision options. As source material for constructing the ontology, the pipeline relies on a corpus of natural-language policy documents $\mathcal{D}$ describing the general procedures for making decisions on the given task. For example, when aiming to provide decision support for choosing between different treatments for a certain medical condition, $\mathcal{D}$ may consist of the relevant clinical guidelines. Each document $d \in \mathcal{D}$ is decomposed into a set of text chunks $\mathcal{T}_d$, split semantically according to the document sections. Optionally, these chunks can be filtered according to their relevance to the considered task using an LLM, yielding a filtered set $\mathcal{T}'_d$ (or $\mathcal{T}'_d = \mathcal{T}_d$ when filtering is not applied).

**Algorithm 2** General QBAF Construction

**Input**: Ontology $\mathcal{O} = \langle \mathcal{E}, \mathcal{T}, \mathcal{H}, \mathcal{S} \rangle$, mining depth $d$, argument scheme $s_{\text{arg}}$ (optional), boolean flag `score_root`
**Output**: A set of general QBAFs $\mathbf{G} = \{\mathcal{G}_e\}_{e \in \mathcal{E}}$ and global parameter schema $\Pi$

Initialise $\mathbf{G} \leftarrow \emptyset$, $\Pi \leftarrow \emptyset$
**for** $e \in \mathcal{E}$ **do**
  Initialise $\tau_e \leftarrow \emptyset$, $\chi_e \leftarrow \emptyset$
  $\mathcal{T}_e \leftarrow \{t \in \mathcal{T} \mid (e, t) \in \mathcal{S}\}$
  ▷ *1. Mine a bipolar argumentation framework*
  ▷ *with a natural-language condition function* $\chi_e^N$
  $\langle \mathcal{A}_e, \mathcal{R}_e^-, \mathcal{R}_e^+, \chi_e^N \rangle \leftarrow \texttt{mine\_baf}(e, \mathcal{T}_e, d, s_{\text{arg}})$
  **for** $a \in \mathcal{A}_e$ **do**
    ▷ *2. Estimate base scores*
    **if** $\texttt{is\_root}(a)$ **and not** `score_root` **then**
      $\tau_e(a) \leftarrow 0.5$
      $\chi_e(a) \leftarrow \top$ ▷ *Use tautology for root*
    **else if** $\texttt{is\_root}(a)$ **then**
      $\tau_e(a) \leftarrow \texttt{score}(a, \mathcal{T}_e)$
      $\chi_e(a) \leftarrow \top$ ▷ *Use tautology for root*
    **else**
      $p \leftarrow \texttt{get\_parent}(a, \mathcal{R}_e^-, \mathcal{R}_e^+)$
      $\texttt{is\_sup} \leftarrow (a, p) \in \mathcal{R}_e^+$
      $\tau_e(a) \leftarrow \texttt{score}(a, \mathcal{T}_e, p, \texttt{is\_sup}, \chi_e^N(a))$
      ▷ *3. Formalise conditions, updating schema*
      $\langle \chi_e(a), \Pi \rangle \leftarrow \texttt{form\_cond}(a, \chi_e^N(a), \Pi)$
    **end if**
  **end for**
  $\mathcal{G}_e \leftarrow \langle \mathcal{A}_e, \mathcal{R}_e^-, \mathcal{R}_e^+, \tau_e, \chi_e \rangle$
  $\mathbf{G} \leftarrow \mathbf{G} \cup \{\mathcal{G}_e\}$
**end for**
**return** $\mathbf{G}, \Pi$

The obtained chunks are then used to construct an ontology $\mathcal{O} = \langle \mathcal{E}, \mathcal{T}, \mathcal{H}, \mathcal{S} \rangle$ where $\mathcal{E}$ is a set of decision entities in the ontology, $\mathcal{T}$ is the associated set of source text chunks, $\mathcal{H} \subseteq \mathcal{E} \times \mathcal{E}$ is a hierarchy relation linking specific variants of decision options to their general categories and $\mathcal{S} \subseteq \mathcal{E} \times \mathcal{T}$ is a provenance relation linking each decision entity to all the relevant chunks in which it is mentioned.

The ontology is constructed in an iterative way, starting from a tuple of empty sets and being progressively expanded with newly appearing decision entities as well as links recording all entity mentions from the individual text chunks, as shown in Algorithm 1. The `mine_ontology` function is implemented by an LLM provided with the current hierarchy of entities and the text of the current chunk in each iteration, and outputting data about new entities and entity mentions as part of a structured JSON. An example decision option ontology for the glioblastoma treatment recommendation task is shown in Figure 3.

**General QBAF Construction** After extracting the decision-space ontology $\mathcal{O} = \langle \mathcal{E}, \mathcal{T}, \mathcal{H}, \mathcal{S} \rangle$ for a specific task, ArgEval uses this ontology to mine the general QBAFs $\mathbf{G} = \{\mathcal{G}_e\}_{e \in \mathcal{E}}$ for each identified option $e \in \mathcal{E}$. Each general QBAF is a tuple $\mathcal{G}_e = \langle \mathcal{A}_e, \mathcal{R}_e^-, \mathcal{R}_e^+, \tau_e, \chi_e \rangle$ where $\mathcal{A}_e, \mathcal{R}_e^-, \mathcal{R}_e^+, \tau_e$ denote arguments, a binary attack relation,

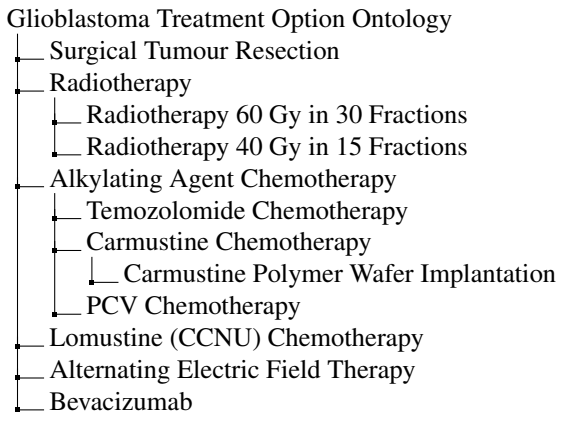


Figure 3: Subset of a glioblastoma treatment option ontology automatically constructed from the relevant clinical guidelines. Only the entities and the hierarchical relations are visualised without the corresponding text chunks and provenance relations. Note that the used LLM has incorrectly categorised Lomustine as a separate treatment rather than a variant of Alkylating Agent Chemotherapy, although this has no effect on the rest of the ArgEval pipeline. The nine leaves of the ontology are used in our main experiments.

a binary support relation and a base score function as for standard QBAFs (see Section 3), while $\chi_e : \mathcal{A}_e \to \Phi$ is a condition function mapping arguments to formal conditions specifying the cases in which the given argument applies. $\Phi$ denotes the domain of conditions expressible in some suitable formal language.

Apart from the general QBAFs themselves, the procedure also produces a global parameter schema $\Pi$, containing the definitions of all parameters appearing in the extracted argument conditions. In our instantiation of the method, both the global parameter schema and the formalised conditions are expressed using JSON schemas,[3] which are relatively easy for LLMs to generate while establishing structure and value constraints that can be straightforwardly verified in an automated fashion. An example JSON schema formalising the condition for one of the arguments from Figure 1 is given in Listing 1. Here, `eloquent_structure_involvement` and `kps` are parameters defined in the global parameter schema $\Pi$.

The general QBAF construction procedure itself iterates over all the identified options $e \in \mathcal{E}$ in the ontology, proceeding in three main steps: bipolar argumentation framework mining, argument base score estimation and argument condition formalisation (see Algorithm 2).

*1) Bipolar Argumentation Framework Mining*. In the first step, the procedure employs an LLM to recursively mine arguments supporting and attacking the given decision option up to a specified depth $d$, forming a tree bipolar argumentation framework structure. This process is similar to the one used for ArgLLMs (Freedman et al. 2025) with some important deviations: the LLM mining the arguments is also provided with all text chunks referring to the given decision, may be optionally supplied with an argument scheme $s_{\text{arg}}$

[3]https://json-schema.org/

**Listing 1** JSON schema formalising the condition of the first attacking argument from Figure 1.

```
{
 "$schema": "https://json-schema.org/
     draft/2020-12/schema",
 "type": "object",
 "anyOf": [
  {
   "properties": {
    "eloquent_structure_involvement": {
     "type": "boolean",
     "const": true
    }
   }
  },
  {
   "properties": {
    "kps": {
     "type": "integer",
     "maximum": 49
    }
   }
  }
 ]
}
```

to guide the argument generation using domain-specific criteria (see an example in Table 2), can generate frameworks of an arbitrary breadth, with potentially different numbers of attackers and supporters, and produces natural-language conditions $\chi_e^N$ describing the circumstances under which each argument applies.

*2) Argument Base Score Estimation*. Once the arguments have been generated, the second step of the construction process estimates their base scores. Depending on a boolean hyperparameter `score_root`, this can either be done only for non-root arguments with the root receiving a midpoint score of $0.5$ or for all arguments in the framework including the root. The former option makes ArgEval predictions more sensitive to the influences of the root argument's attackers and supporters. Apart from the argument text itself, the scoring function is also provided with all text chunks associated with the option under consideration, and, for non-root arguments, the text of the parent argument, the relation of the evaluated argument to this parent and the natural-language condition associated with the argument. Similarly to ArgLLMs, our scoring function prompts an LLM to directly output the estimated base scores, as this simple strategy has been empirically found to outperform other, more complex approaches (Zhou et al. 2025). This process results in base scores such as those shown in the example in Figure 1.

*3) Argument Condition Formalisation*. The final stage of the QBAF construction process converts the natural-language conditions $\chi_e^N$ associated with the arguments to the corresponding formal representations $\chi_e$. This procedure also iteratively updates the global parameter schema $\Pi$ with any new parameters referenced in the formalised conditions. Similarly to the other stages of the construction pipeline, this

| | |
|---|---|
| **Major premise** | Generally, if a clinical intervention provides a net medical benefit and is comparatively suitable among its mutually exclusive alternatives for a given patient, then it is recommended. |
| **Minor premise** | The clinical intervention provides a positive benefit-risk balance for the considered patient. |
| **Minor premise** | The clinical intervention is superior or equivalent to its incompatible alternatives for the considered patient. |
| **CQ1** | Are the clinical features of the patient free of any contraindications against the considered intervention? |
| **CQ2** | Does the available evidence from clinical guidelines establish that the intervention provides a net benefit to the given patient? |
| **CQ3** | Is there no alternative intervention that offers a better benefit-risk profile for the given patient and is incompatible with the currently considered intervention? |

Table 2: Premises and critical questions of the argument scheme $s_{\text{arg}}$ used for the glioblastoma treatment recommendation task. The scheme was developed specifically for this task in collaboration with a domain expert.

step leverages an LLM to generate each condition as well as the updated parameter schema.

After the construction process is finished, the obtained general QBAFs $\mathbf{G} = \{\mathcal{G}_e\}_{e\in\mathcal{E}}$ and the parameter schema $\Pi$ can be used for case-specific inference, as described in the following section.

## 4.2 Case-Specific Inference

The case-specific inference procedure of ArgEval first uses an LLM to extract a set of structured case parameters $\mathcal{P}$ from a natural-language description of the considered case $c$, guided by the definitions in the global parameter schema $\Pi$. For each general QBAF $\mathcal{G}_e \in \mathbf{G}$, these parameters can then be checked against the conditions associated with its arguments, instantiating these frameworks by removing any irrelevant arguments whose conditions are not satisfied, their descendants and any associated relations. This procedure yields regular, instantiated QBAFs $\mathbf{Q} = \{\mathcal{Q}_e\}_{e\in\mathcal{E}}$, which can be evaluated by a gradual semantics $\sigma$. The final strengths of the root arguments produced by the semantics can then be used as recommendation scores $\mathbf{S} = \{s_e\}_{e\in\mathcal{E}}$ quantifying the relative and absolute suitability of the associated decision options. An illustration of the inference process for a single decision option is shown in Figure 1, with the full procedure detailed in Algorithm 3.

## 4.3 ArgEval Properties

ArgEval exhibits several useful qualitative properties summarised in Table 1. Firstly, as we outlined above, the automatic decision-space ontology construction process used by ArgEval enables its application in *open-ended domains* where the different decision options are not specified prior to applying the method. Additionally, since the predictions at inference time are derived by the deterministic argumentation semantics based on the instantiated QBAFs, these QBAFs can act as inherently *faithful explanations* of the used reasoning process. This explainability and reliance on formal argumentative reasoning also facilitate *local contestability*, enabling users to correct mistakes associated with inference for specific samples.

While the faithful explainability and local contestability properties have previously been considered and found to hold for other argumentative approaches (Freedman et al. 2025; Zhu et al. 2025), ArgEval is unique in also supporting *global contestability*. We consider a method to be globally contestable if it provides a faithfully interpretable mechanism $M$ such that contesting some aspect of $M$ for an instance $x$ can result in an updated mechanism $M'$, for which the resulting method performs better than the original on other instances $x' \neq x$ drawn from the same distribution, as measured by some pre-specified evaluation metric. That is, for a globally contestable method, user contestations on a specific input can persist and improve future predictions on different instances. ArgEval facilitates this through its reuse of general reasoning patterns captured in the decision-

**Algorithm 3** Case-Specific Inference

**Input**: A set of general QBAFs $\mathbf{G} = \{\mathcal{G}_e\}_{e\in\mathcal{E}}$, global parameter schema $\Pi$, natural-language case description $c$, argumentation semantics $\sigma$
**Output**: Sets of instantiated QBAFs $\mathbf{Q} = \{\mathcal{Q}_e\}_{e\in\mathcal{E}}$ and decision recommendation scores $\mathbf{S} = \{s_e\}_{e\in\mathcal{E}}$

Initialise $\mathbf{Q} \leftarrow \emptyset, \mathbf{S} \leftarrow \emptyset$
▷ *1. Extract parameters from case description*
$\mathcal{P} \leftarrow \texttt{extract\_params}(c, \Pi)$
**for** $\mathcal{G}_e \in \mathbf{G}$ **do**
  $\langle \mathcal{A}_e, \mathcal{R}_e^-, \mathcal{R}_e^+, \tau_e, \chi_e \rangle \leftarrow \mathcal{G}_e$
  ▷ *2. Instantiate general QBAF into regular QBAF*
  Initialise $\mathcal{A}'_e \leftarrow \mathcal{A}_e, \mathcal{R}'^-_e \leftarrow \mathcal{R}_e^-, \mathcal{R}'^+_e \leftarrow \mathcal{R}_e^+$
  **for** $a \in \mathcal{A}_e$ **do**
    ▷ *Remove arguments with unsatisfied conditions*
    ▷ *and the associated relations*
    **if** $a \in \mathcal{A}'_e$ **and not** $\texttt{eval\_cond}(\chi_e(a), \mathcal{P})$ **then**
      $\mathcal{A}_{\text{rm}} \leftarrow \{a\} \cup \texttt{descendants}(a, \mathcal{R}'^-_e, \mathcal{R}'^+_e)$
      $\mathcal{A}'_e \leftarrow \mathcal{A}'_e \setminus \mathcal{A}_{\text{rm}}$
      $\mathcal{R}'^-_e, \mathcal{R}'^+_e \leftarrow \texttt{rm\_rels}(\mathcal{A}_{\text{rm}}, \mathcal{R}'^-_e, \mathcal{R}'^+_e)$
    **end if**
  **end for**
  $\tau'_e \leftarrow \tau_e$ restricted to $\mathcal{A}'_e$
  $\mathcal{Q}_e \leftarrow \langle \mathcal{A}'_e, \mathcal{R}'^-_e, \mathcal{R}'^+_e, \tau'_e \rangle$
  ▷ *3. Calculate the recommendation score*
  $r_e \leftarrow \texttt{get\_root}(\mathcal{Q}_e)$
  $s_e \leftarrow \sigma_{\mathcal{Q}_e}(r_e)$
  $\mathbf{Q} \leftarrow \mathbf{Q} \cup \{\mathcal{Q}_e\}$
  $\mathbf{S} \leftarrow \mathbf{S} \cup \{s_e\}$
**end for**
**return** $\mathbf{Q}, \mathbf{S}$

space ontology, general QBAFs and the general parameter schema, which can all be updated during contestation. These pre-compiled components also improve *inference speed*, as case-specific inference only uses an LLM to extract the relevant case parameters, substantially reducing token usage.

# 5 Experiments

Here, we describe the experiments applying ArgEval and additional baseline methods to the task of recommending suitable treatments for glioblastoma, an aggressive brain tumour with poor prognosis. Apart from reporting the initial performance on this task, we also conduct a case study illustrating the impact of contestation on the key evaluation metrics[4].

## 5.1 Experimental Setup

**Task** In our experiments, we consider the task of providing patient-specific recommendations regarding suitable interventions for glioblastoma. The intention is to partially simulate the application of ArgEval and other baselines as a clinical decision support system advising clinicians on therapies that should be considered as part of the overall treatment plan. We believe this task poses a realistic and relevant evaluation setting, as the deployment of explainable AI systems would likely provide most substantial benefits in high-stakes domains such as healthcare.

**Policy Documents** We use four established clinical guidelines on treating high-grade glioma or glioblastoma as our policy documents: the European Society for Medical Oncology (ESMO) clinical practice guidelines for high-grade glioma (Stupp et al. 2014), the National Comprehensive Cancer Network® (NCCN®) guidelines on central nervous system cancers[5] (NCCN 2025), the guideline on primary brain tumours and brain metastases in over 16s from the National Institute for Health and Care Excellence (NICE 2018), and the Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on glioblastoma management in adults (Wen et al. 2025). Due to the length of the selected guidelines, spanning more than 200 pages in one case, we manually extracted pages relevant to glioblastoma treatment.

To parse and preprocess the unstructured guideline PDFs, we used the Docling library (Auer et al. 2024) along with its heron document layout analysis model (Livathinos et al. 2025) and a hybrid chunker splitting the documents into shorter chunks. These chunks were postprocessed by gpt-oss-20b (OpenAI 2025), merging semantically related chunks that were separated by page breaks, figures or other PDF elements and filtering out chunks unrelated to glioblastoma treatment. This process resulted in a total of 67 chunks.

Given the refined chunks, we used gpt-oss-20b to construct a decision-space ontology using the procedure described in Section 4, identifying a total of 179 treatment entities. To constrain our experiments to the most common atomic treatment options, we only retained leaf entities associated with mentions in at least three out of the four considered guideline documents that were not combination treatments (i.e., they had a single root ancestor in the ontology hierarchy). This yielded an ontology with 9 different treatment options captured in Figure 3. A review by an expert oncologist determined that the ontology captures the most commonly considered non-combination therapies for glioblastoma, along with some less relevant options, providing a useful basis for the subsequent experiments.

**Case Descriptions** To source the patient case descriptions, we again liaised with an expert clinical oncologist to obtain a set of four key parameters that substantially affect treatment decisions for glioblastoma patients along with suitable value ranges: age (ranging over the values 50, 60, 75 and 85), MGMT methylation status (ranging over the values "methylated", "unknown" and "unmethylated"), Karnofsky Performance Status (KPS, ranging over the values 10, 30, 50, 70 and 90) and tumour location (ranging over the values "non-dominant frontal lobe", "thalamus" and "brainstem"). In addition to these properties, we also add a spurious parameter indicating the patient sex ("male" or "female"), which is typically seen as having no impact on the choice of appropriate therapy for glioblastoma. Considering all possible combinations of parameter values provides us with $4 \times 3 \times 5 \times 3 \times 2 = 360$ unique patient parameter sets, which we then convert into brief patient vignettes using an LLM (see Figure 1 for an example). We performed manual checks to ensure that the vignettes accurately represent the input patient parameters with no omissions or extraneous information.

**Labels** The ground-truth treatment recommendation labels were determined according to an algorithm with rules informed and reviewed by the clinical oncologist, aiming to capture the general clinical practice when deciding on treatments recommended to different patient subgroups. The labels are assigned to each patient vignette/treatment combination, resulting in $360 \times 9 = 3240$ labels in total. Each label indicates whether a specific treatment is "recommended", "maybe recommended" or "not recommended" for a patient with the given characteristics, with the "maybe recommended" label indicating potential suitability with relatively weak or conflicting evidence.

**LLM Backbones** We consider two reasoning LLMs from two different providers, gpt-oss-20b (OpenAI 2025) and Qwen3-30B-A3B-FP8 (Yang et al. 2025). Both models rely on the Mixture-of-Experts (MoE) architecture for more efficient inference, with 3.6B and 3.3B active parameters, respectively. We selected these models due to their high scores on the Artificial Analysis Intelligence Index[6] among major open-weight models that fit within our resource constraints. In our experiments, both models were run in reasoning mode, setting the reasoning effort level to "medium" for gpt-oss-20b. Inference was performed using the vLLM engine (Kwon et al. 2023) with the sampling temperature

[4]Our code is available at https://github.com/adamdejl/argeval.

[5]Disclaimer: NCCN makes no warranties of any kind whatsoever regarding their content, use or application and disclaims any responsibility for their application or use in any way.

[6]https://artificialanalysis.ai/#intelligence

| LLM Backbone | Method | $d$ | Est. Root | Arg. Scheme | LMR$^{\uparrow}$ | NDCG$^{\uparrow}$ | I/O Tokens (M)$^{\downarrow}$ |
|---|---|---|---|---|---|---|---|
| **gpt-oss-20b (medium)** | | | | | | | |
| gpt-oss-20b | Base LLM | – | – | – | 0.8775 | 0.9578 | 16.95 / 6.59 |
| gpt-oss-20b | ArgLLMs-O | 1 | ✗ | ✗ | 0.8614 | 0.9739 | 80.51 / 20.12 |
| gpt-oss-20b | ArgLLMs-O | 1 | ✗ | ✓ | 0.8676 | 0.9629 | 80.41 / 19.89 |
| gpt-oss-20b | ArgLLMs-O | 1 | ✓ | ✗ | 0.8429 | 0.9718 | 95.91 / 22.69 |
| gpt-oss-20b | ArgLLMs-O | 1 | ✓ | ✓ | 0.8444 | 0.9707 | 95.98 / 22.72 |
| gpt-oss-20b | ArgEval | 1 | ✗ | ✗ | 0.6910 | 0.9330 | 1.18 / 0.79 |
| gpt-oss-20b | ArgEval | 1 | ✗ | ✓ | 0.8009 | 0.9654 | 1.20 / 0.62 |
| gpt-oss-20b | ArgEval | 1 | ✓ | ✗ | 0.4432 | 0.9293 | 1.18 / 0.66 |
| gpt-oss-20b | ArgEval | 1 | ✓ | ✓ | 0.6923 | 0.8983 | 1.15 / 0.68 |
| gpt-oss-20b | ArgEval | 2 | ✗ | ✗ | 0.8485 | 0.9422 | 3.96 / 1.93 |
| gpt-oss-20b | ArgEval | 2 | ✗ | ✓ | 0.7006 | 0.8514 | 4.09 / 2.02 |
| gpt-oss-20b | ArgEval | 2 | ✓ | ✗ | 0.5244 | 0.9480 | 4.09 / 1.92 |
| gpt-oss-20b | ArgEval | 2 | ✓ | ✓ | 0.6914 | 0.9241 | 4.51 / 2.29 |
| **Qwen3-30B-A3B-FP8** | | | | | | | |
| Qwen3-30B | Base LLM | – | – | – | 0.8630 | 0.9527 | 17.83 / 5.44 |
| Qwen3-30B | ArgLLMs-O | 1 | ✗ | ✗ | 0.8373 | 0.9752 | 78.06 / 22.11 |
| Qwen3-30B | ArgLLMs-O | 1 | ✗ | ✓ | 0.8617 | 0.9735 | 78.95 / 21.73 |
| Qwen3-30B | ArgLLMs-O | 1 | ✓ | ✗ | 0.8216 | 0.9815 | 93.93 / 25.80 |
| Qwen3-30B | ArgLLMs-O | 1 | ✓ | ✓ | 0.8417 | 0.9804 | 96.06 / 25.75 |
| Qwen3-30B | ArgEval | 1 | ✗ | ✗ | 0.8812 | 0.9750 | 0.90 / 0.67 |
| Qwen3-30B | ArgEval | 1 | ✗ | ✓ | 0.8623 | 0.9389 | 0.89 / 0.67 |
| Qwen3-30B | ArgEval | 1 | ✓ | ✗ | 0.7417 | 0.9138 | 0.88 / 0.78 |
| Qwen3-30B | ArgEval | 1 | ✓ | ✓ | 0.7497 | 0.9118 | 1.01 / 0.55 |
| Qwen3-30B | ArgEval | 2 | ✗ | ✗ | 0.7886 | 0.9782 | 2.31 / 1.38 |
| Qwen3-30B | ArgEval | 2 | ✗ | ✓ | 0.8818 | 0.9771 | 2.65 / 1.38 |
| Qwen3-30B | ArgEval | 2 | ✓ | ✗ | 0.6284 | 0.8957 | 2.36 / 1.36 |
| Qwen3-30B | ArgEval | 2 | ✓ | ✓ | 0.7694 | 0.9123 | 3.01 / 1.73 |

Table 3: Performance on glioblastoma (GBM) treatment prediction across models and inference configurations. LMR denotes the label match rate, indicating the share of the absolute treatment recommendation scores matching the labels, while NDCG quantifies the level of agreement between the ground-truth treatment ranking and the predicted ranking. The I/O Tokens column reports prompt/completion token usage in millions (M) without the initial decision-space ontology construction, which is shared for all methods.

set to 1.0 and other generation parameters taking their default values. We opted to use this temperature value since it is the recommended sampling temperature for gpt-oss-20b and since high temperatures are often needed to avoid degenerate behaviours in current reasoning LLMs, particularly looping (Pipis et al. 2025). While the main experiments use both considered models, the initial document processing and ontology construction were only performed using gpt-oss-20b, ensuring that all the experimental results are based on a consistent ontology.

**Baselines** We compare our method to two baselines: base LLMs and ArgLLMs-O. Base LLMs are directly instructed to provide a numerical score ranging from 0 to 1 indicating the degree to which a treatment is recommended for a given patient, with this process being repeated for each of the 9 treatment options. The LLM is provided with all the text chunks associated with the currently evaluated treatment option as well as the patient case description, and can use the reasoning chain for intermediate work before giving the final answer. The ArgLLMs-O baseline is an adapted version of the original ArgLLMs (Freedman et al. 2025) using our treatment option ontology and generating a separate QBAF for each possible treatment option. During both the argument generation and base score estimation stages, the LLM performing these operations is provided with all the relevant guideline chunks as well as the patient case description. Thus, both baselines can use the patient information during their entire reasoning process, while ArgEval only gains access to it during the case-specific inference stage.

**Method Variants** We consider eight different variants of ArgEval arising from the choices of three hyperparameters: the general QBAF depth (either 1 or 2), root base score estimation (either enabled or disabled) and argument scheme inclusion (either enabled or disabled). The maximum QBAF depth is limited to two, as considering larger depths would likely provide diminishing returns while considerably increasing computational costs and cognitive load on the users. See Section 4, Algorithm 2 and Table 2 for more information on these parameters. For ArgLLMs-O, we similarly consider variations with or without root base score

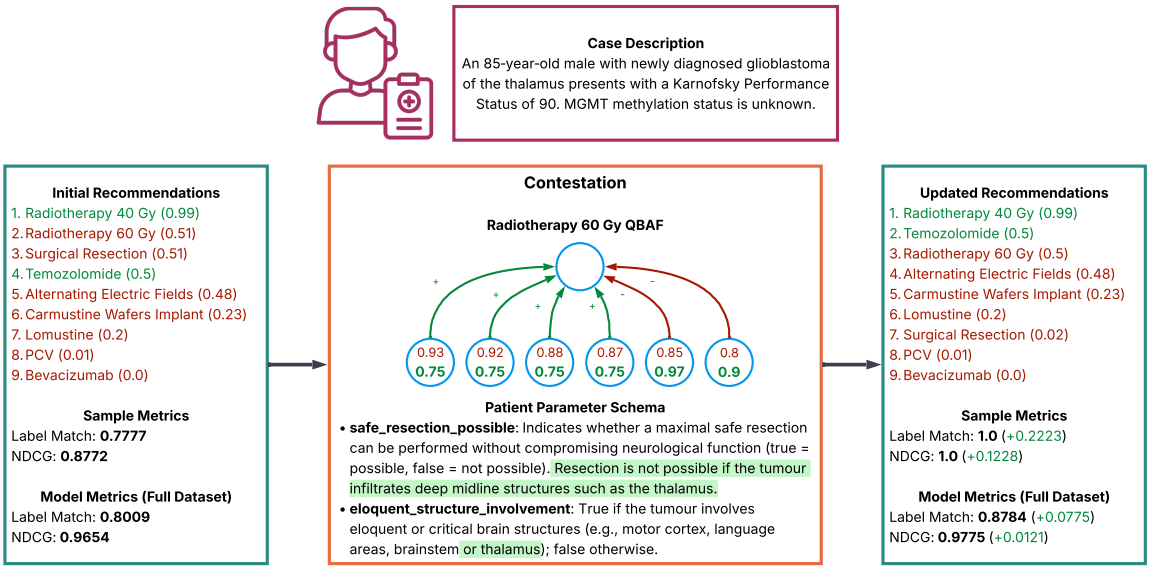


Figure 4: Illustration of the ArgEval contestability experiment[7]. To correct the initially suboptimal recommendations, we make small adjustments to the base scores in the general argumentation framework for the radiotherapy 60 Gy treatment option (with initial scores shown in smaller red font and updated scores in larger green font) and clarify the descriptions of two parameters associated with surgical resection in the parameter schema. These modifications are sufficient to achieve a perfect score on this instance while also substantially improving the overall performance. Treatments recommended by the ground-truth labels are shown in green, with those not recommended in red.

estimation and argument scheme inclusion. Due to the very high computational requirements of ArgLLMs-O, we only consider depth 1 versions for this baseline. All argumentative models use the DF-QuAD gradual semantics (Rago et al. 2016), which is also the default semantics for ArgLLMs.

**Metrics** We use three key metrics to compare the performance of the different methods. The label match rate (LMR) indicates the share of the absolute treatment recommendation scores matching their labels, with the "recommended" label considered to be matching for scores between $0.5$ and $1.0$, the "maybe recommended" label matching scores between $0.25$ and $0.75$ and the "not recommended" label matching scores between $0.0$ and $0.5$. Note that the used score intervals intentionally overlap in order to reflect inherent uncertainty associated with decision-making in medicine. This design choice is also intended to reduce the sensitivity of the metrics to the precise score scales used by different methods. Since the recommendation scores are intended to be helpful in ranking potential treatments from the most to least suitable, we also report the normalised discounted cumulative gain (NDCG) metric (Järvelin and Kekäläinen 2002). NDCG indicates the level of agreement between the ground-truth ranking (starting from the "recommended" treatments and ending with the "not recommended" ones) and the ranking defined by the predicted treatment recommendation scores. Finally, we report the total prompt and completion tokens used during the LLM inference for each method. These measures include the general QBAF construction for ArgEval but exclude the initial ontology mining, which is shared for all methods.

## 5.2 Performance Results

The results of our experiments are summarised in Table 3. There are several trends that can be observed.

Generally, the best-performing versions of ArgEval achieve competitive results compared to the baselines. For example, the variant using the Qwen3-30B model with depth 2, no root estimation and argument scheme inclusion achieves the best overall LMR of $0.8818$ with a comparatively high NDCG of $0.9771$, which is only surpassed by 3/25 other method variants (one of which is another instance of ArgEval). However, ArgEval seems to be more sensitive to the chosen hyperparameters with substantial variations between the different versions. This is likely due to the fact that any deficiencies during general QBAF construction affect all subsequent inferences. ArgEval variants using root base score estimation perform particularly poorly, probably because treatment options assigned high base scores may still be inappropriate for specific patients.

Nevertheless, the high performance of certain ArgEval variants is notable given their low computational requirements. In particular, thanks to its global reasoning, ArgEval requires substantially fewer inference tokens compared to the other baselines, with even the most expensive depth 2 variant requiring $\sim 2.4\times$ and $\sim 8.7\times$ fewer completion tokens compared to the cheapest base LLM version and the cheapest ArgLLMs-O version, respectively.

Overall, we argue that ArgEval provides a good trade-off between recommendation performance and computational costs, with a substantial qualitative benefit associated with global contestability. In the next section, we demonstrate

[7]Icon from Noun Project (Patient by Alzam, CC BY 3.0).

that this benefit can translate into direct performance gains.

### 5.3 Contestability Case Study

To demonstrate the explainability and global contestability capabilities of ArgEval, we conduct an additional experiment showing the performance effects of contesting the recommendations for a single output. We intentionally choose an ArgEval variant with moderate performance, specifically, the version using gpt-oss-20b with depth 1, no root score estimation and argument scheme inclusion.

As illustrated in Figure 4, the model correctly recommends Radiotherapy 40 Gy, but incorrectly ranks the unsuitable options Radiotherapy 60 Gy and Surgical Resection above Temozolomide, another possible therapy. Inspecting the instantiated QBAFs for Radiotherapy 60 Gy, we observe that there are four supporting and two attacking arguments. While the attacking arguments correctly identify Radiotherapy 40 Gy as typically being the more appropriate regimen for elderly patients, they are outweighed by the supporters arguing that Radiotherapy 60 Gy is generally suitable for patients with good performance status. Aiming to perform a minimal contestation correcting the recommendations, we slightly increase the base scores of the attackers and decrease the base scores of the supporters in the general QBAF, reducing the Radiotherapy 60 Gy score to slightly below 0.5 (not visible in Figure 4 due to rounding).

For Surgical Resection, our inspection of the instantiated QBAFs reveals that ArgEval has correctly captured conditions associated with contraindications to surgery, but is unable to infer that these contraindications apply to patients with glioblastoma of the thalamus during parameter extraction. To address this issue, we refine the corresponding parameter descriptions in the schema as illustrated. This results in correct extraction of the corresponding parameters and a substantial reduction in the Surgical Resection score, with the inference process after these changes illustrated in Figure 1.

Notably, performing the two contestations based on a single sample results in a substantial performance improvement on the full dataset, reaching an LMR of $0.8784$ and NDCG of $0.9775$, which surpasses the performance of all other methods using gpt-oss-20b.

## 6 Conclusion

In this paper, we introduced ArgEval, a novel framework automatically mapping task decision spaces, mining the corresponding option ontologies and constructing general QBAFs capturing the benefits and drawbacks of each decision option depending on the characteristics of a specific case. Uniquely, this shift to general argumentative reasoning enables global contestability while also providing substantial computational cost savings. We evaluated the performance of ArgEval on the task of treatment decision recommendation for glioblastoma, finding it to be highly competitive, with contestability providing opportunities for further performance improvements. Future work could further explore the utility of ArgEval's explainability and contestability to end users through user studies or consider its application in domains beyond healthcare.

## Acknowledgments

Dejl and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX). Toni was also funded by EPSRC (grant UKRI3928, NeSyDebates). Williams was partially funded by NIHR Imperial BRC (grant no. RDB01 79560) and by the Brain Tumour Research Centre of Excellence at Imperial.

## AI Declaration

We employed AI tools for coding and writing assistance as well as for providing general feedback on the paper. All code produced by AI tools has been manually reviewed and often substantially edited before use. Writing assistance consisted solely of polishing manually drafted text, with the outputs undergoing further review and editing.

## References

Atanasova, P.; Camburu, O.-M.; Lioma, C.; Lukasiewicz, T.; Simonsen, J. G.; and Augenstein, I. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 283–294. Toronto, Canada: Association for Computational Linguistics.

Auer, C.; Lysak, M.; Nassar, A. S.; Dolfi, M.; Livathinos, N.; Vagenas, P.; Ramis, C. B.; Omenetti, M.; Lindlbauer, F.; Dinkla, K.; Mishra, L.; Kim, Y.; Gupta, S.; de Lima, R. T.; Weber, V.; Morin, L.; Meijer, I.; Kuropiatnyk, V.; and Staar, P. W. J. 2024. Docling technical report. *CoRR* abs/2408.09869.

Ayoobi, H.; Potyka, N.; Rapberger, A.; and Toni, F. 2026. Argumentative debates for transparent bias detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 40(23):18944–18952.

Barez, F.; Wu, T.-Y.; Arcuschin, I.; Lan, M.; Wang, V.; Siegel, N.; Collignon, N.; Neo, C.; Lee, I.; Paren, A.; et al. 2025. Chain-of-thought is not explainability. *Preprint*.

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.

Bermejo, V. J.; Gago, A.; Gálvez, R. H.; and Harari, N. 2025. LLMs outperform outsourced human coders on complex textual analysis. *Scientific Reports* 15(1):40122.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S. M.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR* abs/2303.12712.

Busch, F.; Hoffmann, L.; Rueger, C.; van Dijk, E. H.; Kader, R.; Ortiz-Prado, E.; Makowski, M. R.; Saba, L.; Hadamitzky, M.; Kather, J. N.; Truhn, D.; Cuocolo, R.; Adams, L. C.; and Bressem, K. K. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine* 5(1):26.

Cabessa, J.; Hernault, H.; and Mushtaq, U. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, 6624–6635. Abu Dhabi, UAE: Association for Computational Linguistics.

Chen, Y.; Benton, J.; Radhakrishnan, A.; Uesato, J.; Denison, C.; Schulman, J.; Somani, A.; Hase, P.; Wagner, M.; Roger, F.; Mikulik, V.; Bowman, S. R.; Leike, J.; Kaplan, J.; and Perez, E. 2025. Reasoning models don't always say what they think. *CoRR* abs/2505.05410.

De Angelis, E.; Proietti, M.; and Toni, F. 2025. Learning to contest argumentative claims. In *RuleML+RR*, 237–255.

Dignum, V.; Michael, L.; Nieves, J. C.; Slavkovik, M.; Suarez, J.; and Theodorou, A. 2025. Contesting black-box AI decisions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, 2854–2858. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

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.

Gorur, D.; Rago, A.; and Toni, F. 2025. Retrieval and argumentation enhanced multi-agent LLMs for judgmental forecasting. *CoRR* abs/2510.24303.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* 43(2).

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20(4):422–446.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, 611–626. New York, NY, USA: Association for Computing Machinery.

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.

Livathinos, N.; Auer, C.; Nassar, A. S.; de Lima, R. T.; Lysak, M.; Ebouky, B.; Berrospi, C.; Dolfi, M.; Vagenas, P.; Omenetti, M.; Dinkla, K.; Kim, Y.; Weber, V.; Morin, L.; Meijer, I.; Kuropiatnyk, V.; Strohmeyer, T.; Gurbuz, A. S.; and Staar, P. W. J. 2025. Advanced layout analysis models for Docling. *CoRR* abs/2509.11720.

Luo, X.; Rechardt, A.; Sun, G.; Nejad, K. K.; Yáñez, F.; Yilmaz, B.; Lee, K.; Cohen, A. O.; Borghesani, V.; Pashkov, A.; Marinazzo, D.; Nicholas, J.; Salatiello, A.; Sucholutsky, I.; Minervini, P.; Razavi, S.; Rocca, R.; Yusifov, E.; Okalova, T.; Gu, N.; Ferianc, M.; Khona, M.; Patil, K. R.; Lee, P.-S.; Mata, R.; Myers, N. E.; Bizley, J. K.; Musslick, S.; Bilgin, I. P.; Niso, G.; Ales, J. M.; Gaebler, M.; Ratan Murty, N. A.; Loued-Khenissi, L.; Behler, A.; Hall, C. M.; Dafflon, J.; Bao, S. D.; and Love, B. C. 2025. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour* 9(2):305–315.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum. Comput. Interact.* 5(CSCW1):106:1–106:25.

McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; Hou, L.; Cheng, Y.; Liu, Y.; Mahdavi, S. S.; Prakash, S.; Pathak, A.; Semturs, C.; Patel, S.; Webster, D. R.; Dominowska, E.; Gottweis, J.; Barral, J.; Chou, K.; Corrado, G. S.; Matias, Y.; Sunshine, J.; Karthikesalingam, A.; and Natarajan, V. 2025. Towards accurate differential diagnosis with large language models. *Nature* 642(8067):451–457.

Miller, T. 2023. Explainable AI is dead, long live explainable AI!: Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, 333–342. ACM.

NCCN. 2025. Central nervous system cancers, version 3.2025.

NICE. 2018. Brain tumours (primary) and brain metastases in over 16s. NICE.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *CoRR* abs/2508.10925.

Pipis, C.; Garg, S.; Kontonis, V.; Shrivastava, V.; Krishnamurthy, A.; and Papailiopoulos, D. 2025. Wait, wait, wait... Why do reasoning models loop? *CoRR* abs/2512.12895.

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

Rago, A.; Cocarascu, O.; Oksanen, J.; and Toni, F. 2025. Argumentative review aggregation and dialogical explanations. *Artificial Intelligence* 340:104291.

Siegel, N.; Camburu, O.-M.; Heess, N.; and Perez-Ortiz, M. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 530–546. Bangkok, Thailand: Association for Computational Linguistics.

Stupp, R.; Brada, M.; van den Bent, M.; Tonn, J.-C.; and Pentheroudakis, G. 2014. High-grade glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* 25:iii93–iii101. ESMO Updated Clinical Practice Guidelines.

Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.-B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E. P.; Seehofnerová, A.; Rohatgi, N.; Hosamani, P.; Collins, W.; Ahuja, N.; Langlotz, C. P.; Hom, J.; Gatidis, S.; Pauly,

J.; and Chaudhari, A. S. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 30(4):1134–1142.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wen, P. Y.; Weller, M.; Lee, E. Q.; Touat, M.; Khasraw, M.; Rahman, R.; Platten, M.; Lim, M.; Winkler, F.; Horbinski, C.; Verhaak, R. G. W.; Huang, R. Y.; Ahluwalia, M. S.; Albert, N. L.; Tonn, J.-C.; Schiff, D.; Barnholtz-Sloan, J. S.; Ostrom, Q.; Aldape, K. D.; Batchelor, T. T.; Bindra, R. S.; Chiocca, E. A.; Cloughesy, T. F.; DeGroot, J. F.; French, P.; Galanis, E.; Galldiks, N.; Gilbert, M. R.; Hegi, M. E.; Lassman, A. B.; Le Rhun, E.; Mehta, M. P.; Mellinghoff, I. K.; Minniti, G.; Roth, P.; Sanson, M.; Taphoorn, M. J. B.; von Deimling, A.; Weiss, T.; Wick, W.; Zadeh, G.; Reardon, D. A.; Chang, S. M.; Short, S. C.; van den Bent, M. J.; and Preusser, M. 2025. Glioblastoma in adults: A Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro-Oncology* 27(11):2751–2788.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; et al. 2025. Qwen3 technical report. *CoRR* abs/2505.09388.

Yin, X.; Potyka, N.; Rago, A.; Kampik, T.; and Toni, F. 2025. Contestability in quantitative argumentation. *CoRR* abs/2507.11323.

Zhou, K.; Dejl, A.; Freedman, G.; Chen, L.; Rago, A.; and Toni, F. 2025. Evaluating uncertainty quantification methods in argumentative large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 21700–21711. Suzhou, China: Association for Computational Linguistics.

Zhu, Y.; Potyka, N.; Hernández, D.; He, Y.; Ding, Z.; Xiong, B.; Zhou, D.; Kharlamov, E.; and Staab, S. 2025. ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation. *CoRR* abs/2508.20131.

Čyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4392–4399. International Joint Conferences on Artificial Intelligence Organization. Survey Track.