

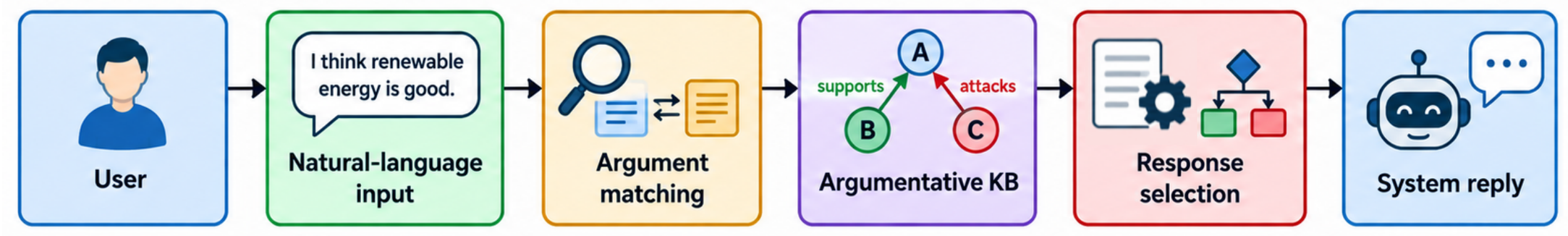
A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Extracting Supports from Abstract Argumentation Frameworks via Structural Patterns and LLM-based Filtering

Stefano Bistarelli, Marco Cuccarini and Carlo Taticchi

AI³@AIxIA 2026

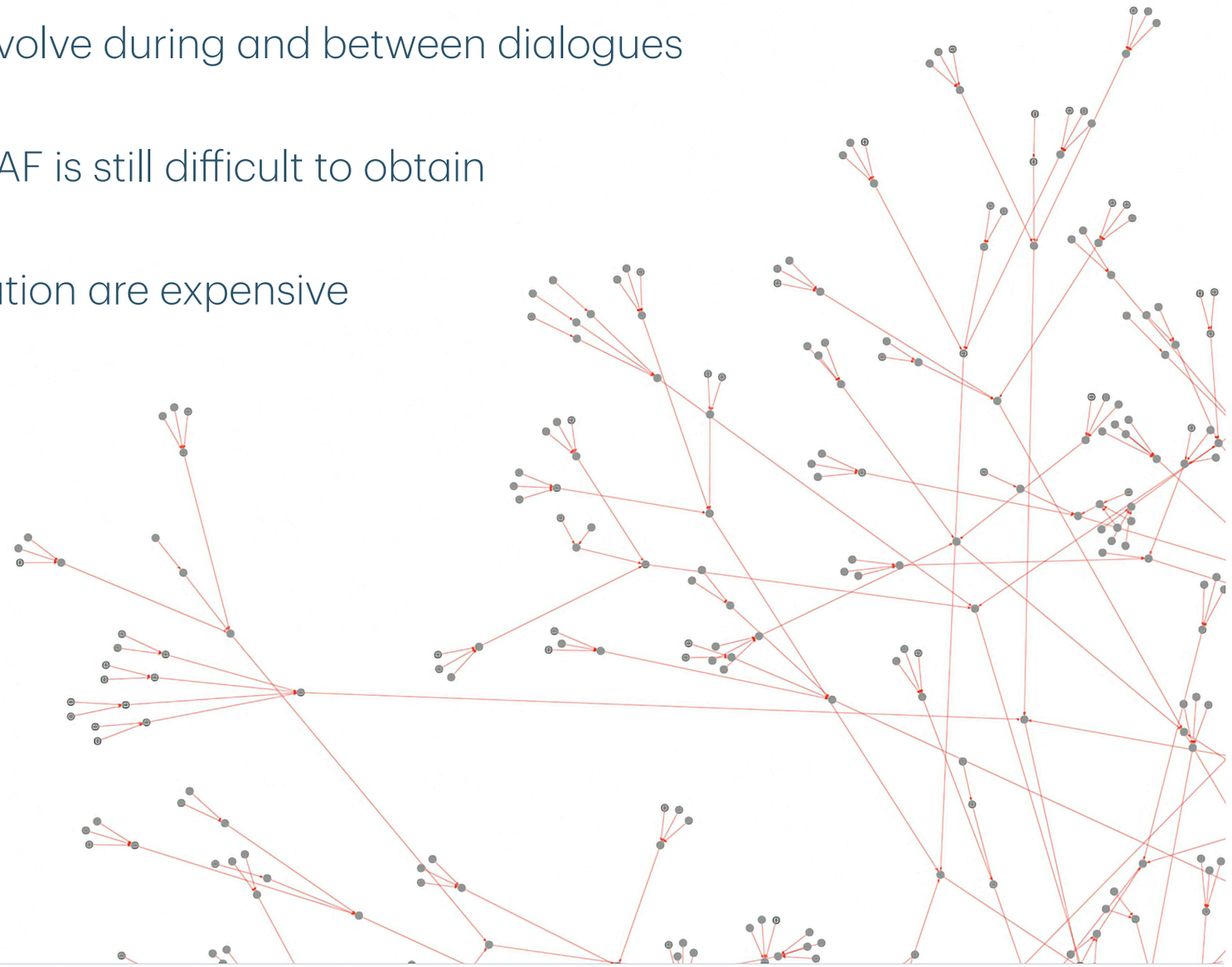
The broader context



- Combining natural-language interaction with computational argumentation
- Each topic uses a bipolar argumentation framework as its argumentative knowledge base
- Formal semantics guide the selection of acceptable dialogue responses
- Users may introduce new arguments and relations during the dialogue

A large knowledge-base…

- Where to find natural-language arguments and explicit attack and support relations?
- The topic knowledge base can evolve during and between dialogues
- A sufficiently large and reliable BAF is still difficult to obtain
- Manual construction and annotation are expensive



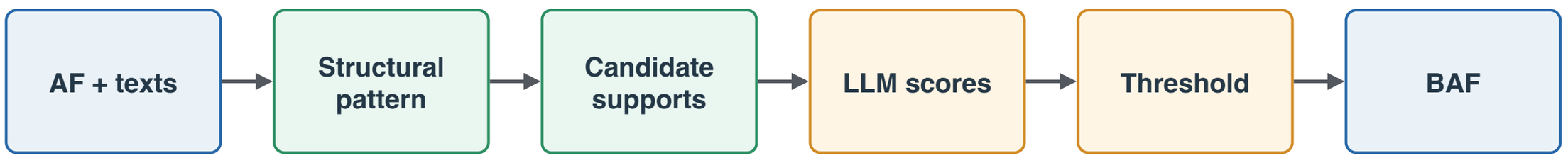
Enriching available AFs

- **Question: can an existing AF provide structural evidence for support?**
- Argument datasets and mining pipelines often provide arguments and attacks
- When argument texts are available, the existing attack structure can guide support extraction
- The aim is to avoid evaluating every ordered pair of arguments

The attack relation generates candidates. Text evaluates their semantic plausibility.

Approach overview

- The structural stage restricts the pairs submitted to the LLM
- The semantic stage assigns a directional confidence score to each candidate
- Thresholding produces an unweighted support relation

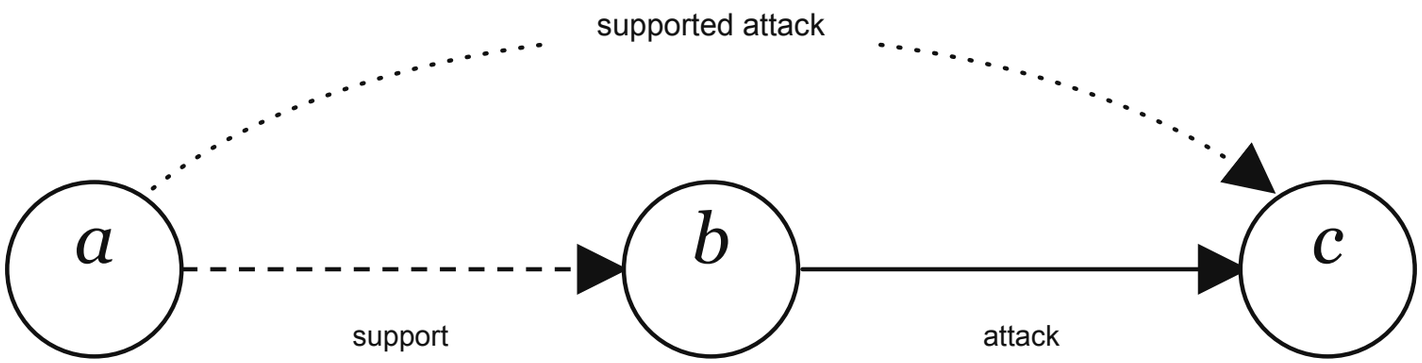


Supported attacks in BAFs

- $B = \langle \mathsf{Arg}, \mathsf{Att}, \mathsf{Supp} \rangle$
- A supported attack from a to c is a sequence $(a_1, a_2, \ldots, a_n), \quad a_1 = a, \quad a_n = c, \quad n \geq 3$

  such that $(a_i, a_{i+1}) \in \mathsf{Supp}$ for $i = 1, \ldots, n-2$

  and $(a_{n-1}, a_n) \in \mathsf{Att}$



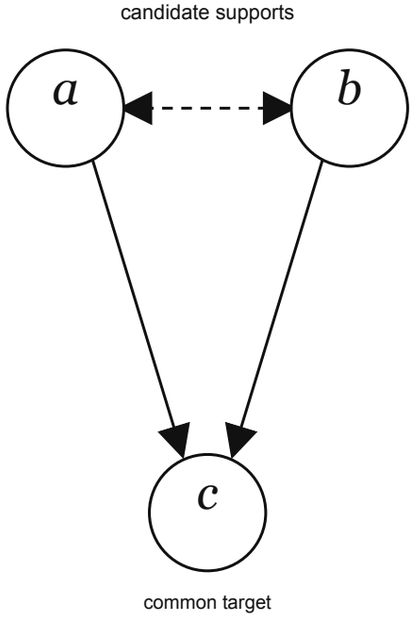
Proposal: co-attacking pattern

- In a supported attack, different arguments contribute to defeating the same target
- Two arguments attacking a common target may show potential supports for each other
- Co-attacking candidate support

$$\mathrm{CSupp} = \{(a,b) \in \mathrm{Arg} \times \mathrm{Arg} \mid \exists c : (a,c),(b,c) \in \mathrm{Att}\}$$

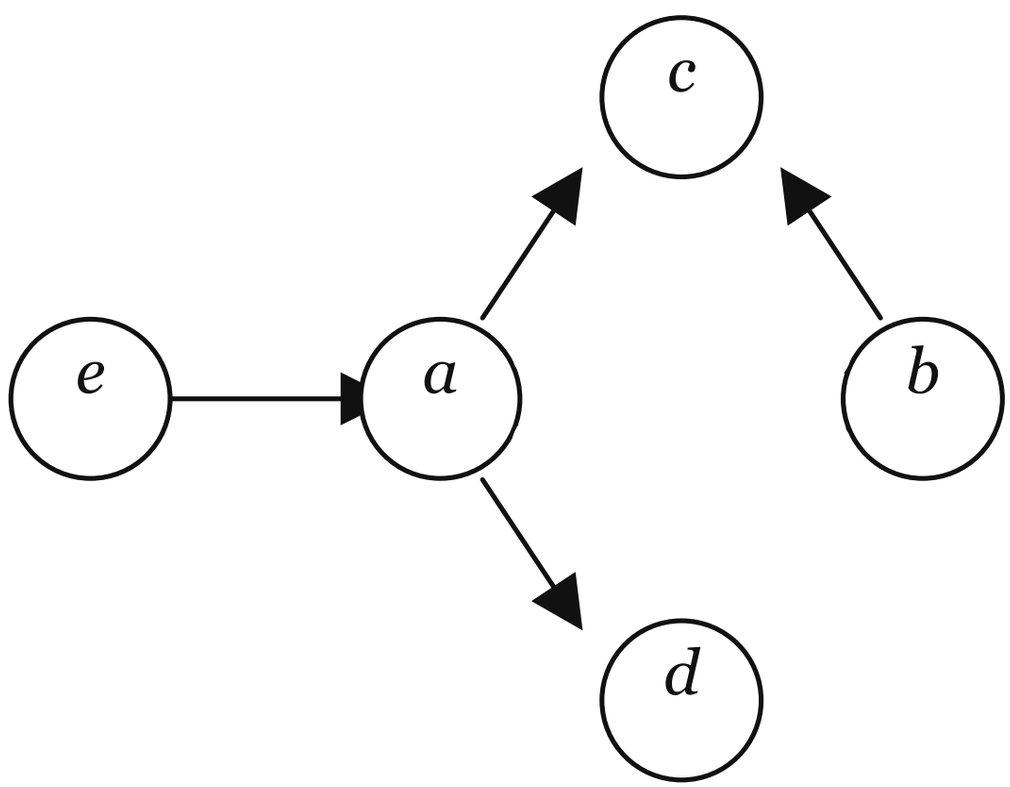
where (a), (b), and (c) are pairwise distinct

- **CSupp** is symmetric: both (a,b) and (b,a) are generated
- These relations are only candidates: their plausibility needs to be evaluated



Structural evidence is insufficient

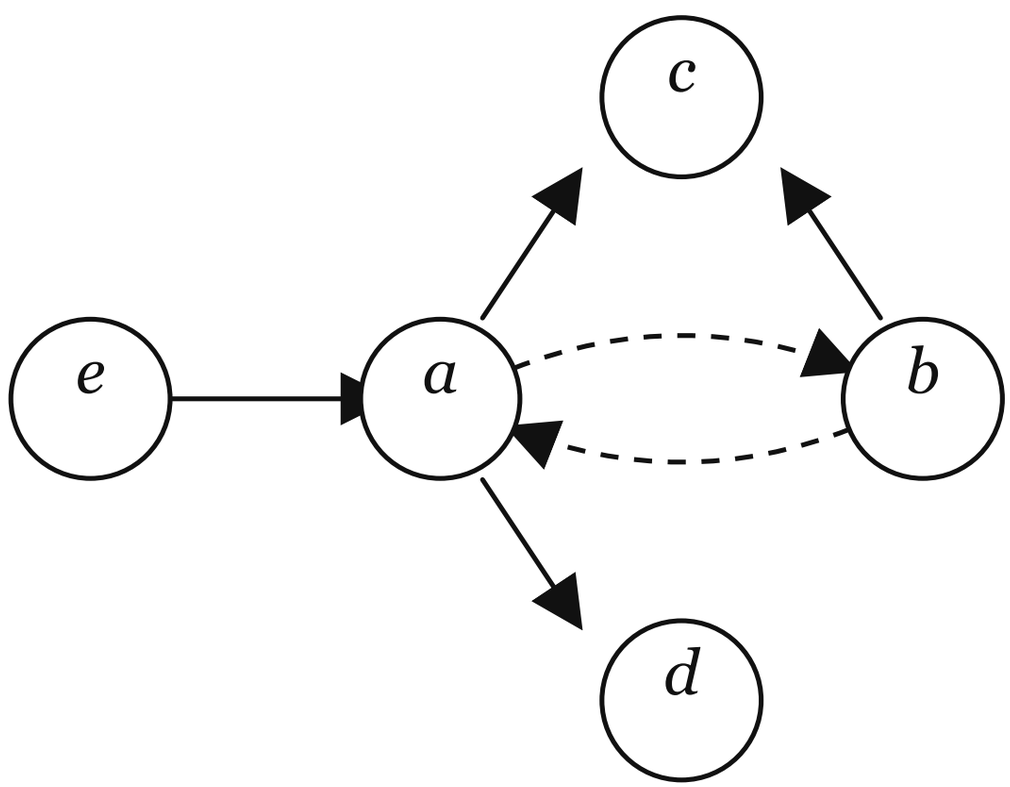
- The induced attacks can change defence and semantic outcomes
- Example: the set $\{e\}$ defends d



Structural evidence is insufficient

- The induced attacks can change defence and semantic outcomes
- Example: the set $\{e\}$ defends d

  a and b both attacking c introduce the co-attacking supports (a, b) and (b, a)



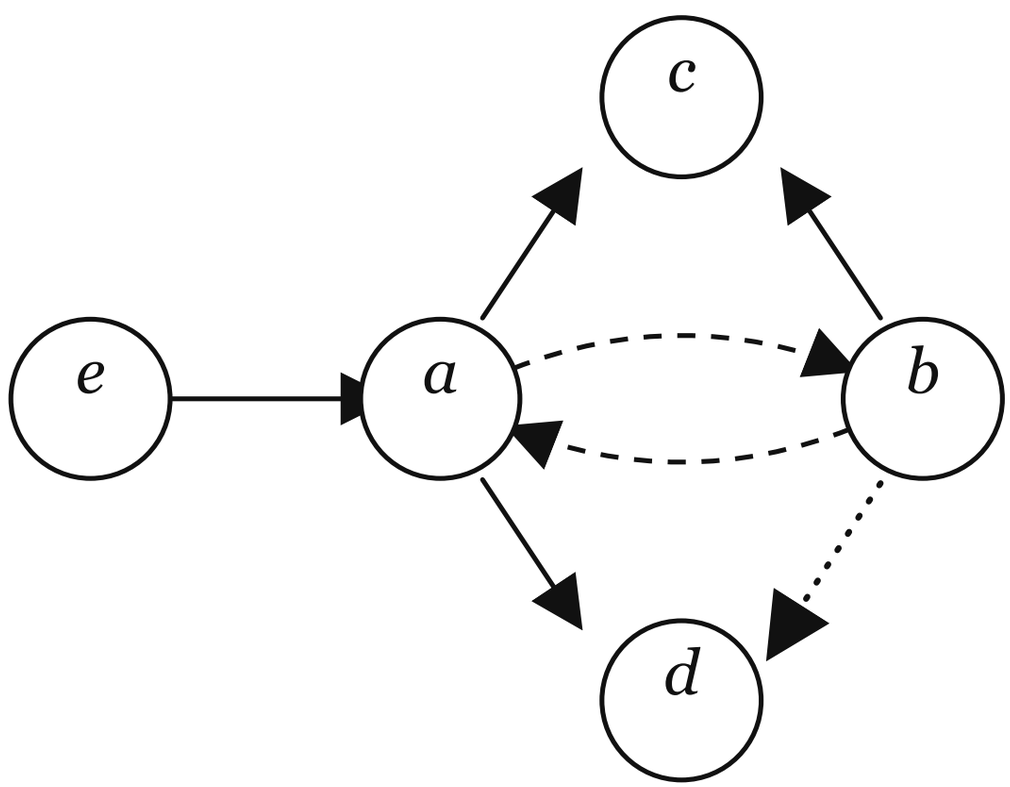
Structural evidence is insufficient

- The induced attacks can change defence and semantic outcomes
- Example: the set $\{e\}$ defends d

  a and b both attacking c introduce the co-attacking supports (a, b) and (b, a)

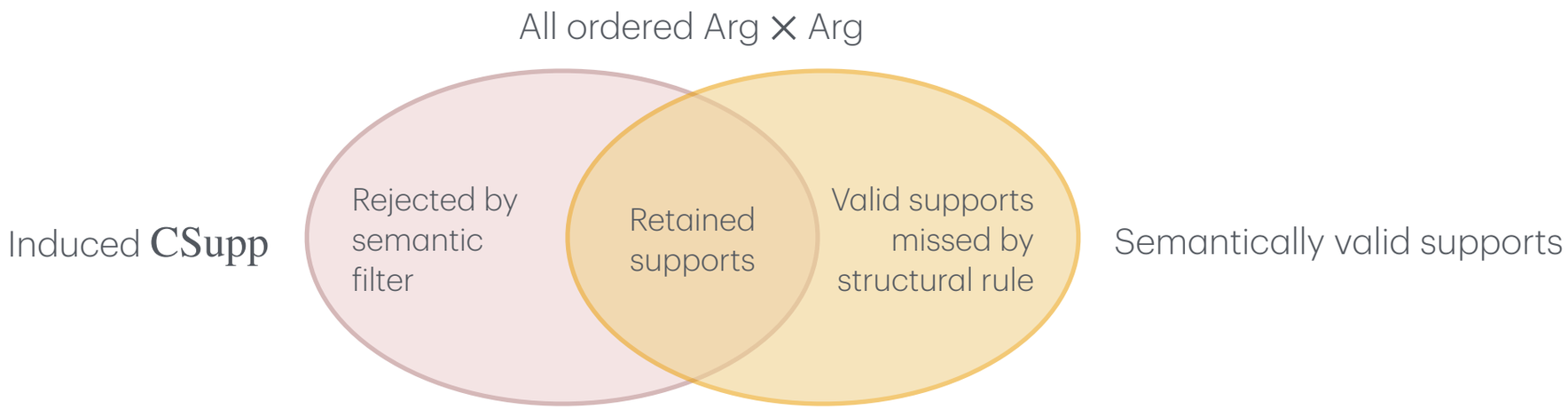
  (b, a, d) is a supported attack

  the set $\{e\}$ no longer defends d



Candidate-selection trade-off

- The structural rule reduces the number of LLM evaluations
- Only arguments sharing an explicit target enter **CSupp**
- A semantic filter can reject unsuitable candidates
- It cannot recover valid supports excluded by the structural stage



LLM-based semantic filtering

- Each argument a has a sentence m(a)
- $\Theta(a, b) = f(m(a), m(b)) \in [0,1]$
- The model evaluates an ordered candidate support
- A zero-shot prompt asks whether a supports b
- The output is a single numerical confidence score

ZERO-SHOT PROMPT

Argument a: [sentence]
Argument b: [sentence]

"How confident are you that a supports b?
Return one value between 0 and 1."

Support value: 0.84

From confidence scores to BAFs

- Supp $= \{(a, b) \in \text{CSupp} \mid \Theta(a, b) \geq \theta\}$
- Semantic confidence is directional, so $\Theta(a, b)$ may differ from $\Theta(b, a)$
- The filtered support relation need not be symmetric
- The score measures confidence in the existence of a support relation
- After filtering, the BAF remains unweighted

a → b 0.87

b → a 0.42

With θ = 0.80, only (a,b) is retained

Threshold-selection objectives

- **Data-tailored** (no predefined target)
  Select the knee of the curve $s(\theta)$, adapting θ to the observed distribution of LLM scores

- **G_1** (Target number of supports)
  Select θ so that $|\mathsf{Supp}_\theta|$ is as close as possible to a desired number n

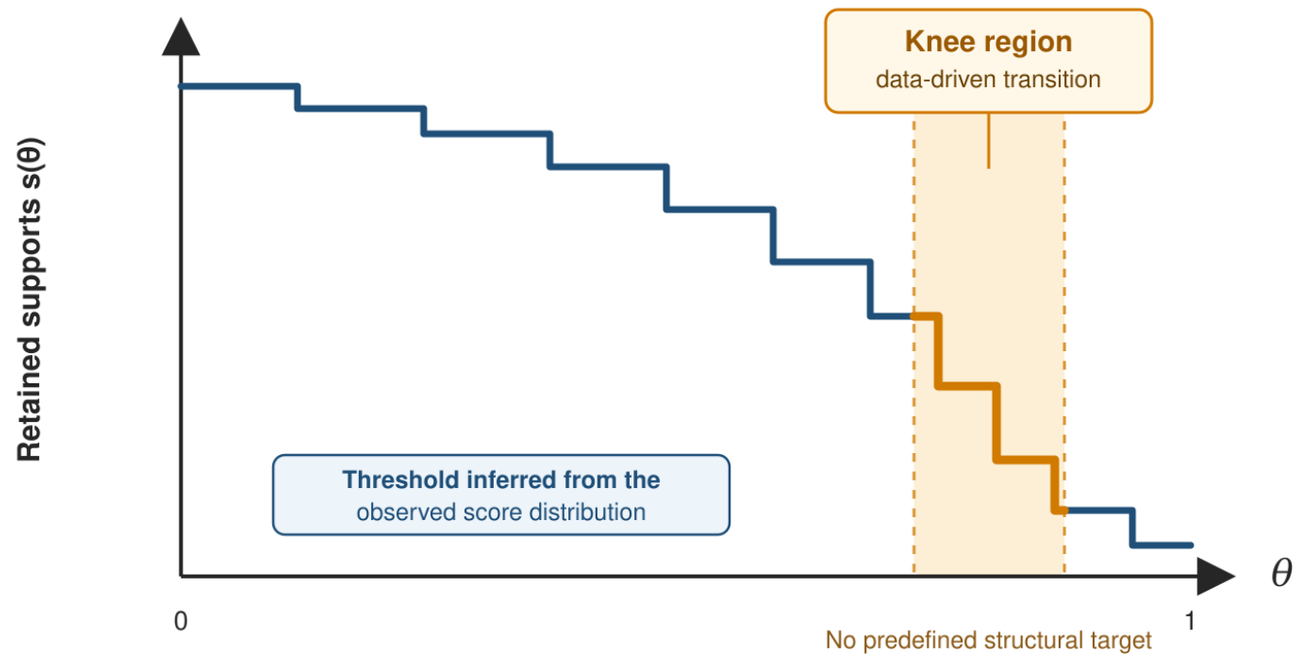
- **G_2** (Target support density)
  Select θ to approximate a desired density ρ relative to the number of possible support relations

- **G_3** (Target support-to-attack ratio)
  Select θ so that $|\mathsf{Supp}_\theta| / |\mathrm{Att}|$ is as close as possible to a desired ratio r

Both **G_2** and **G_3** reduce to **G_1**

Data-tailored threshold

- Data-driven selection, without a predetermined structural property
- Non-increasing step function $s(\theta) = |\text{Supp}_\theta|$
- The knee marks a region where the effect of increasing θ changes noticeably
- The selected value depends on the observed confidence scores



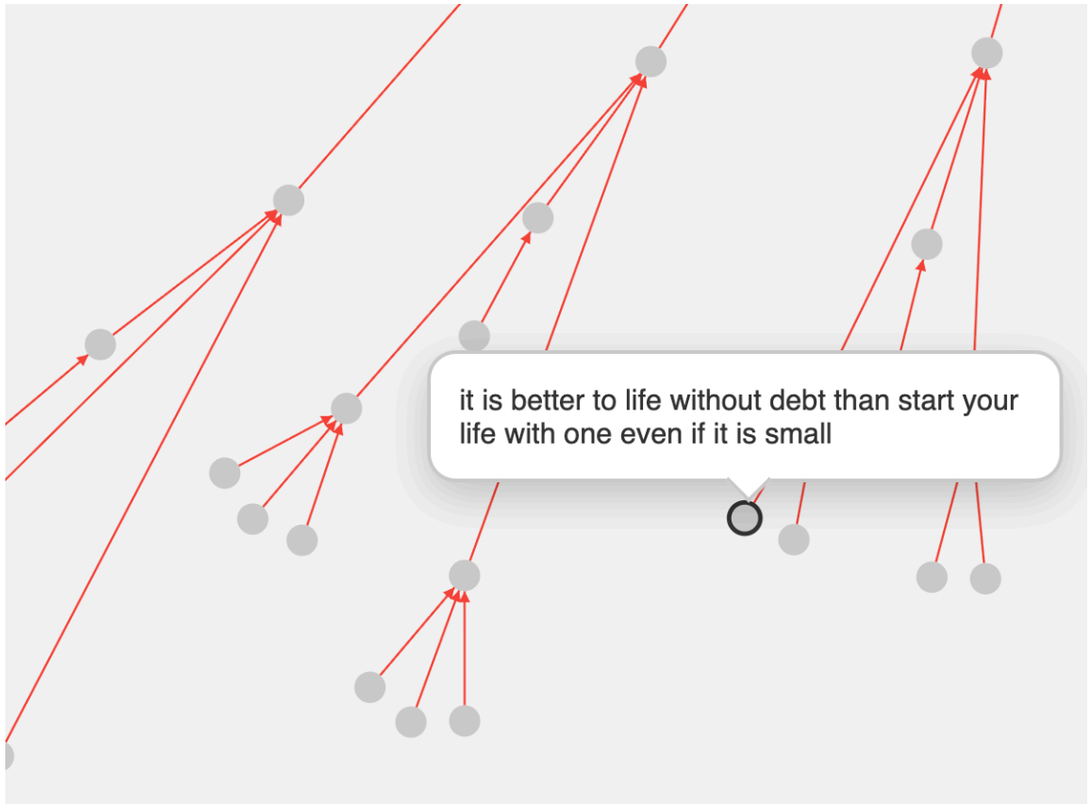
Experimental setting

1,285
natural-language arguments

1,270
attacks

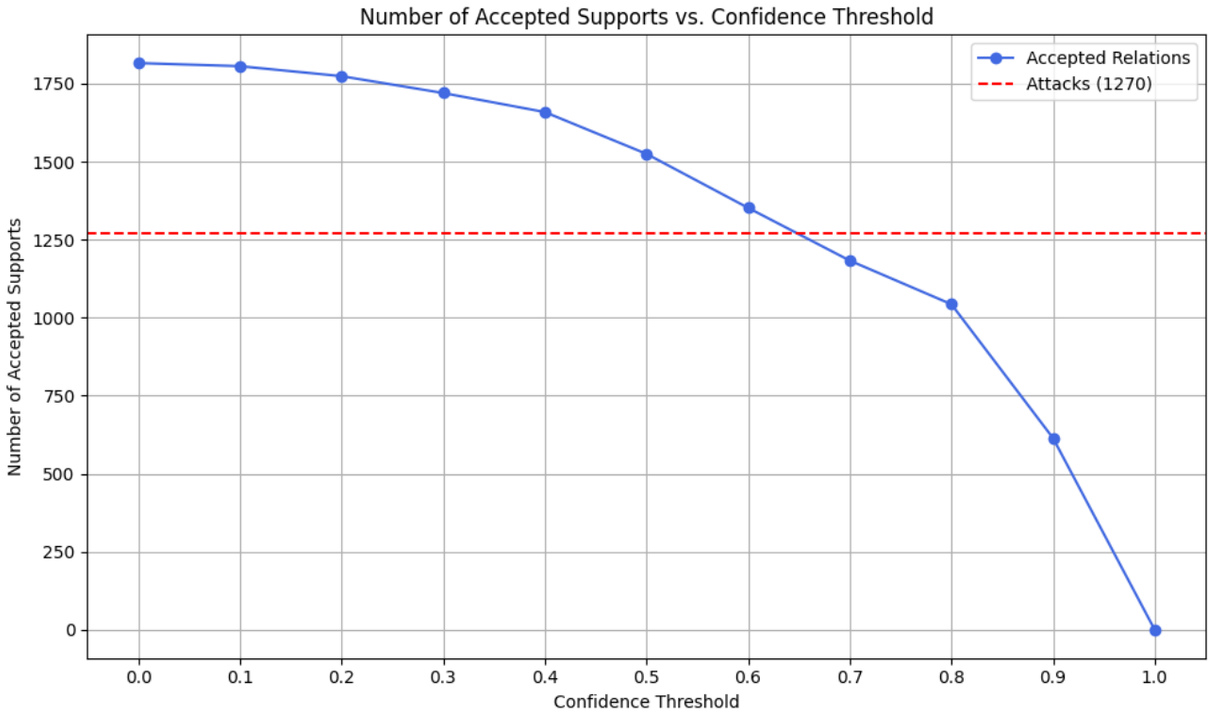
15
subgraphs

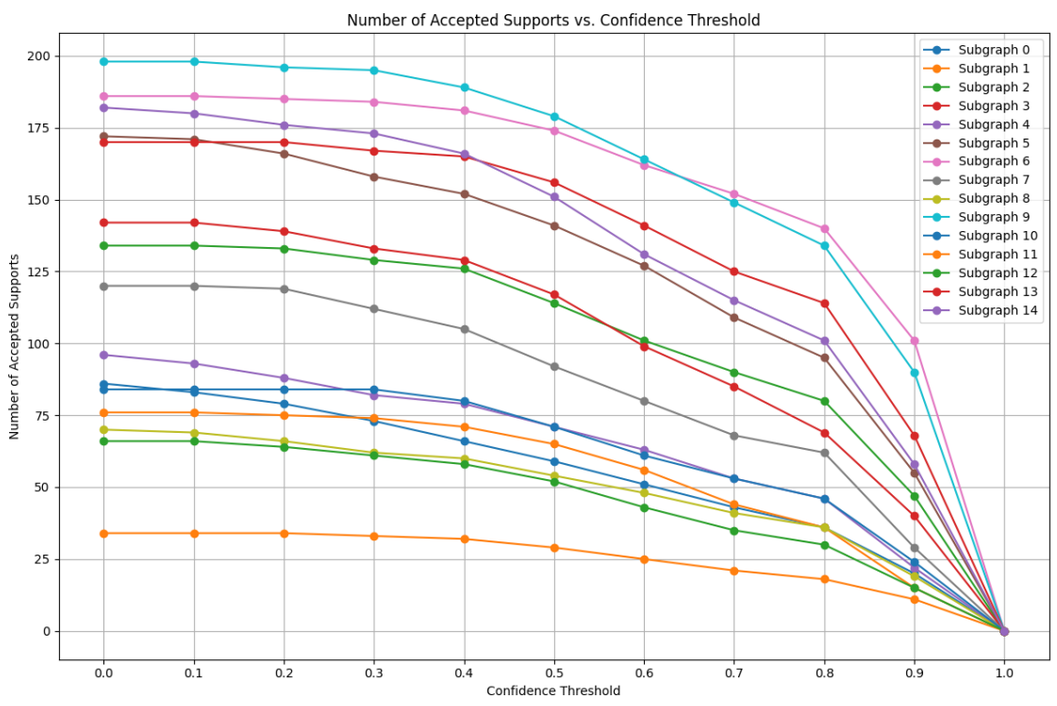
- Dataset (*Chalaguine and Hunter, 2019*): arguments about whether students should pay university fees
- Model: LLaMA 3.1 with 8 billion parameters
- Prompting: zero-shot, with temperature set to 0
- Each candidate support was evaluated once



Threshold analysis

- Possible knee region around $\theta = 0.8$, used for semantic evaluation
- Similar behaviour appears across the subgraphs
- Supports and attacks are balanced between $\theta = 0.6$ and 0.7





Human evaluation of the semantic filter

- LLM scores for 78 directed candidate support pairs were converted into binary support/non-support decisions (threshold $\theta = 0.8$)
- Agreement was ~70% between the LLM and each annotator
- The LLM therefore captures a substantial proportion of human support judgements, but important differences remain
- The experiment evaluates classification within the candidate set, not the recall of the complete extraction pipeline

Comparison	Agreement
Human 1 vs Human 2	**85.90%**
Human 1 vs LLM	**69.23%**
Human 2 vs LLM	**70.51%**

Conclusions

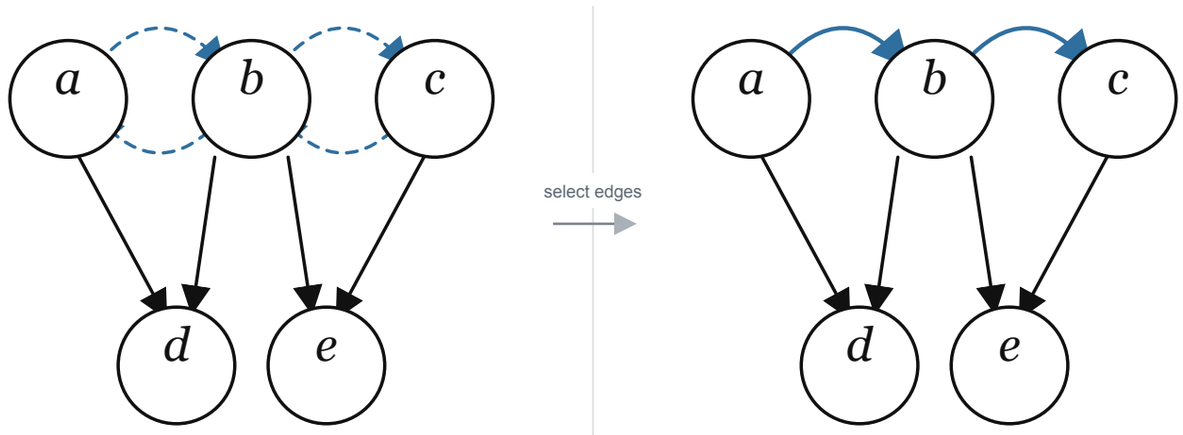
- The structural rule reduces the search space before the LLM evaluation
- Structural patterns generate candidate supports from an existing AF
- LLM scores assess their semantic plausibility
- Threshold selection can be data-driven or goal-oriented

Current limits

- Supports outside candidates cannot be recovered
- Limited experiments settings and evaluation
- Some arguments depend on their original target and lose meaning when evaluated alone

Open directions

- Broader candidate generation through additional patterns or semantic retrieval
- Common-target information in the semantic evaluation
- Quantitative or incomplete BAFs for uncertain supports



A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Thank you! Questions?

Extracting Supports from Abstract Argumentation Frameworks via Structural Patterns and LLM-based Filtering

Stefano Bistarelli, Marco Cuccarini and Carlo Taticchi

AI³@AIxIA 2026