RUC-QA: Reasoning Until Convinced

Explainable Question Answering via Argumentation Graphs

Stefano Bistarelli (University of Perugia),
Marco Cuccarini (University of Naples & University of Perugia),
Nico Potyka (University of Cardiff)

February 5, 2026

Introduction

Motivation: Beyond the Black Box

- LLMs are powerful tools but produce **opaque** decisions and are prone to **hallucinations**.
- In critical domains (medical and finance), verifiable explanations are required.
- RUC-QA integrates **symbolic reasoning** (argumentation graph) and a **probabilistic approach** (LLM).

Objectives of RUC-QA

Task: Multiple Choice Questions
Mimic human decision-making by creating an argumentation graph that supports or attacks all possible hypotheses. Collect information until there is enough evidence to provide an answer.

- Transparency: every decision is mapped in a graph.
- Efficiency: greedy strategy.
- Independence: external knowledge sourced from Wikipedia.

Methodology

System Architecture

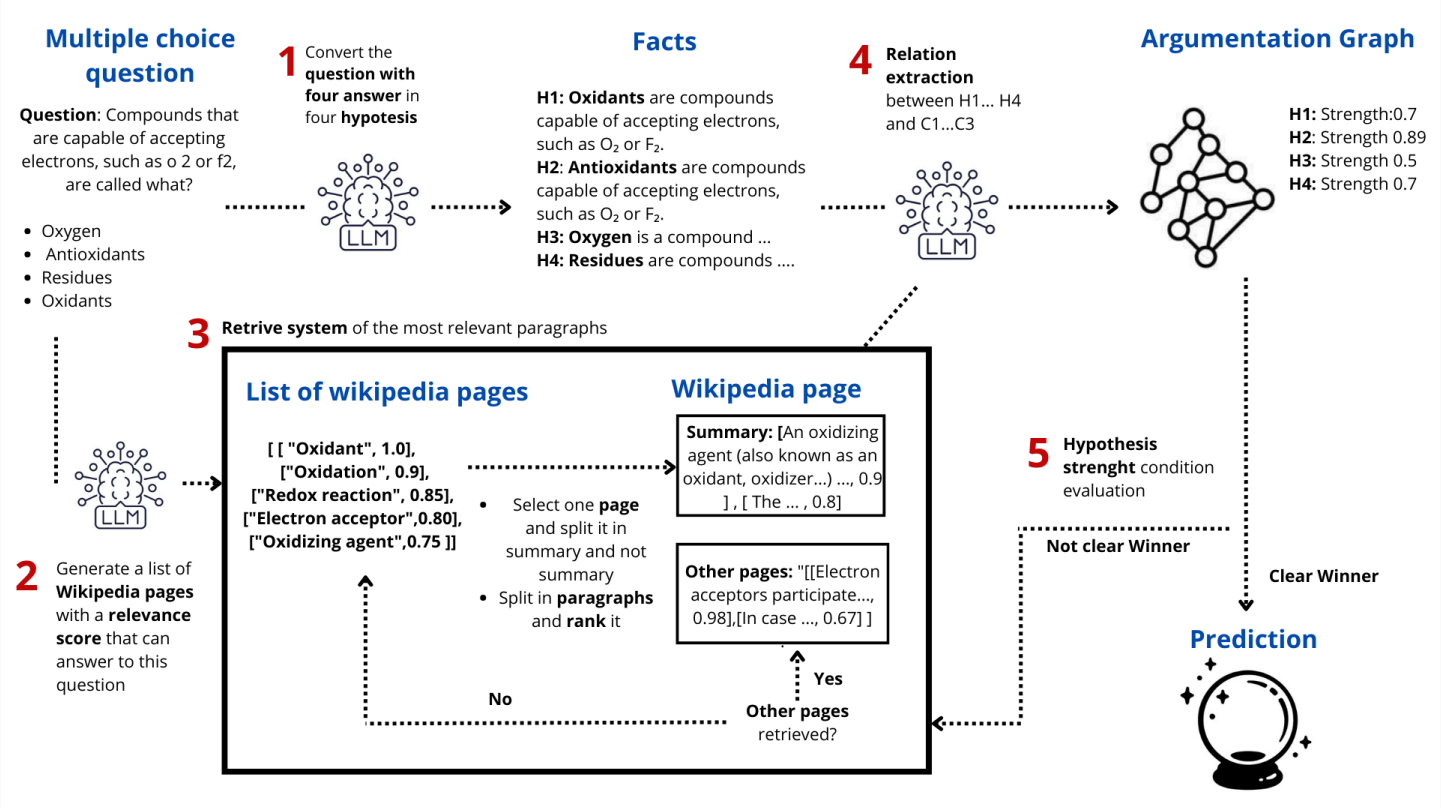


Figure 1: Pipeline from question to answer

(1) Conversion into Factual Hypotheses

Each option becomes an independent statement.

Example
Question: What is the star of the Solar System?

Options: Moon, Sun, Mars

Hypothesis 1: The Moon is the star of the Solar System.

Hypothesis 2: The Sun is the star of the Solar System.

System Architecture

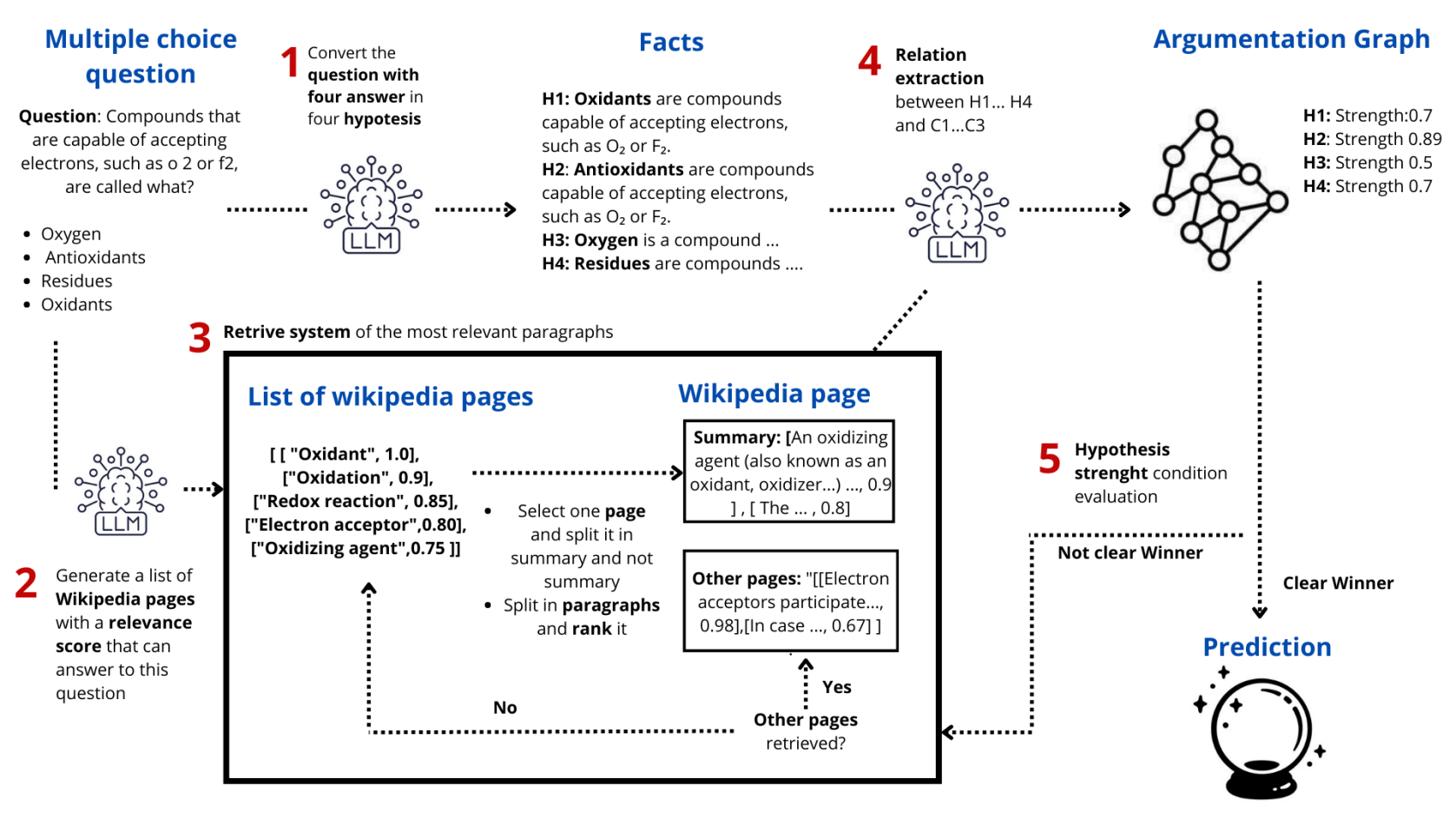


Figure 2: Pipeline from question to answer

(2-3) Semantic Search and Filtering

- **Source Identification:** The LLM acts as an intelligent search engine, suggesting the most promising Wikipedia page titles for the hypothesis.
- **Page Ranking:** Pages are ranked based on global semantic relevance to the original question.
- **Sparsity Filtering (S-BERT):**
  - We do not read the entire document; we split it into paragraphs.
  - Use **Sentence-BERT** to compute cosine similarity between the question and each paragraph.
  - Only the top-K paragraphs that pass a relevance threshold are selected.

Note: The abstract/summary is prioritized, diving deeper only if necessary.

System Architecture

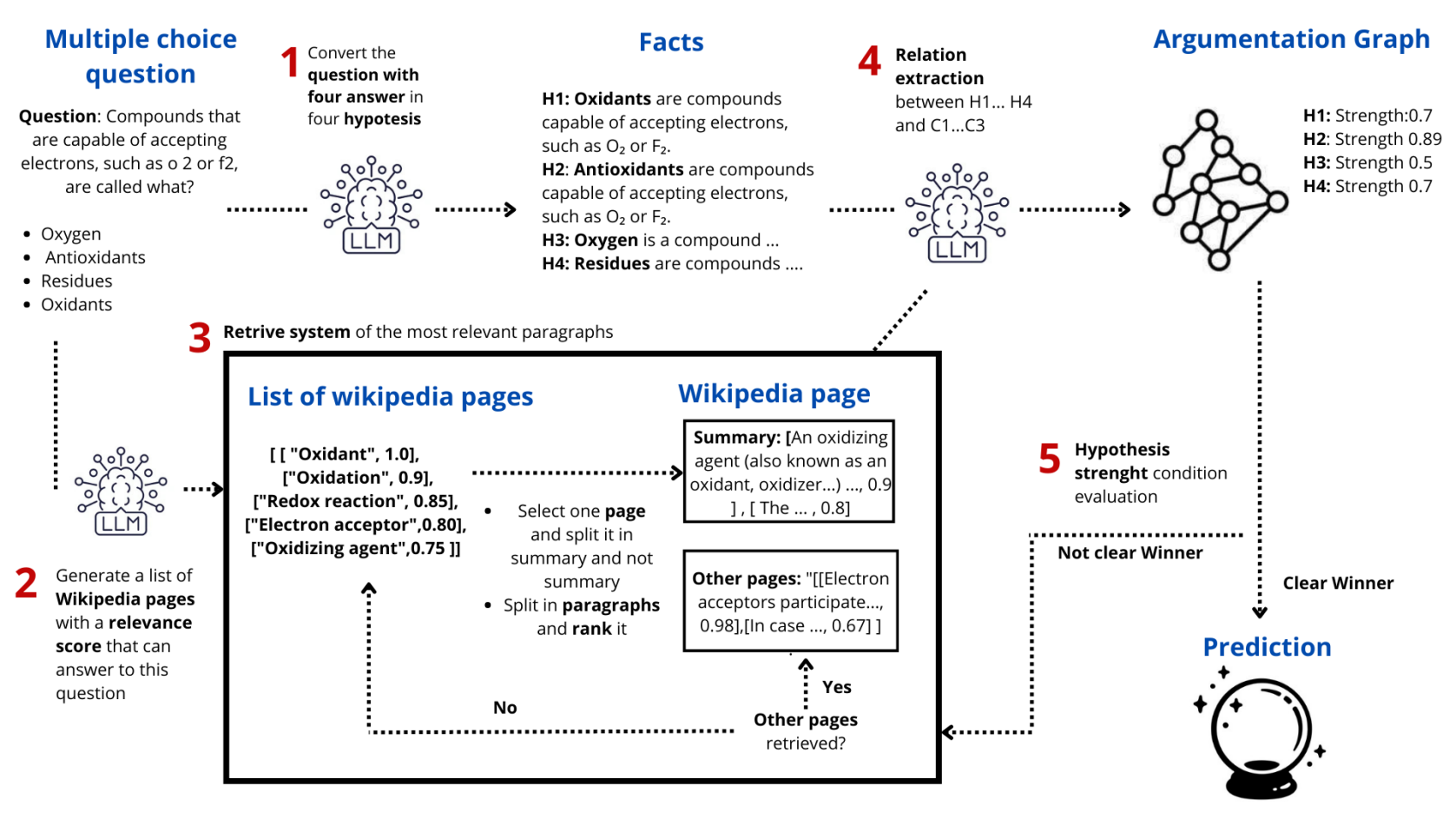


Figure 3: Pipeline from question to answer

(4) Argument and Relation Extraction

- **Claims (Atomic Propositions):** The LLM extracts only essential and independent information from Wikipedia text and converts it into readable claims.
- **Bipolar Relations:** For each claim, the system identifies logical connections to hypotheses:
  - **Support (+):** Evidence that confirms a specific hypothesis.
  - **Attack (−):** Evidence that contradicts or refutes a hypothesis.
- **Noise Control:** A maximum number t of arguments per iteration is defined to keep the graph manageable and focused.

(4) Argument and Relation Extraction

Logical Analysis Example **Extracted text:** "The Sun is classified as a yellow dwarf star."

System action:

- **Relation:** Supports H_{Sun} (The Sun is the star of the Solar System).
- **Relation:** Attacks H_{Moon} (The Moon is the star of the Solar System).

Note: This step transforms unstructured text into a logical network where each decision is traceable to the Wikipedia source.

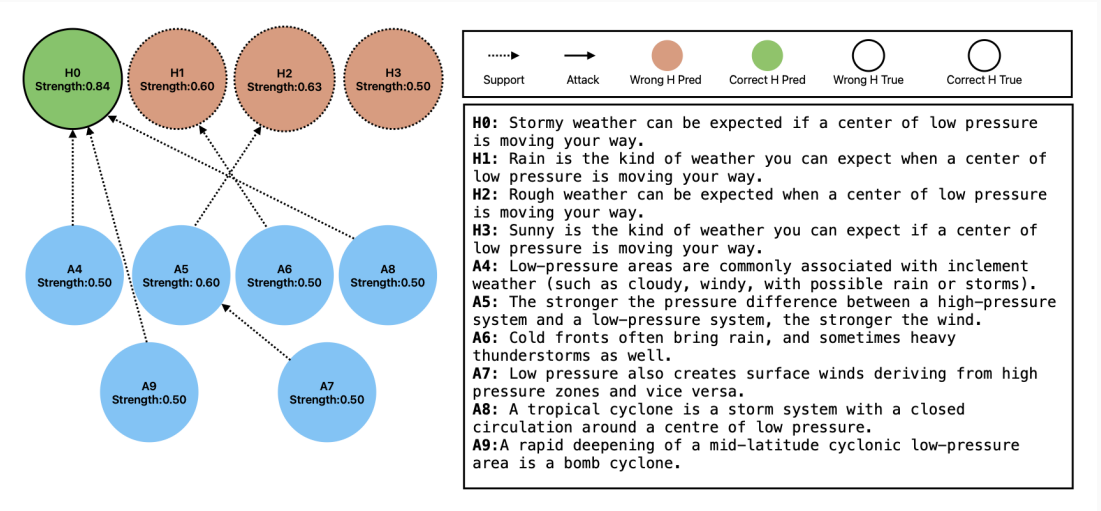
(4) Building the Argumentation Graph

Logical Connections

The graph maps the reasoning structure:

- **Inter-argumentative**
- **Hypothesis-Argument**

This structure makes the decision-making process **auditable** and transparent.



figureExample of Generated Bipolar Graph

System Architecture

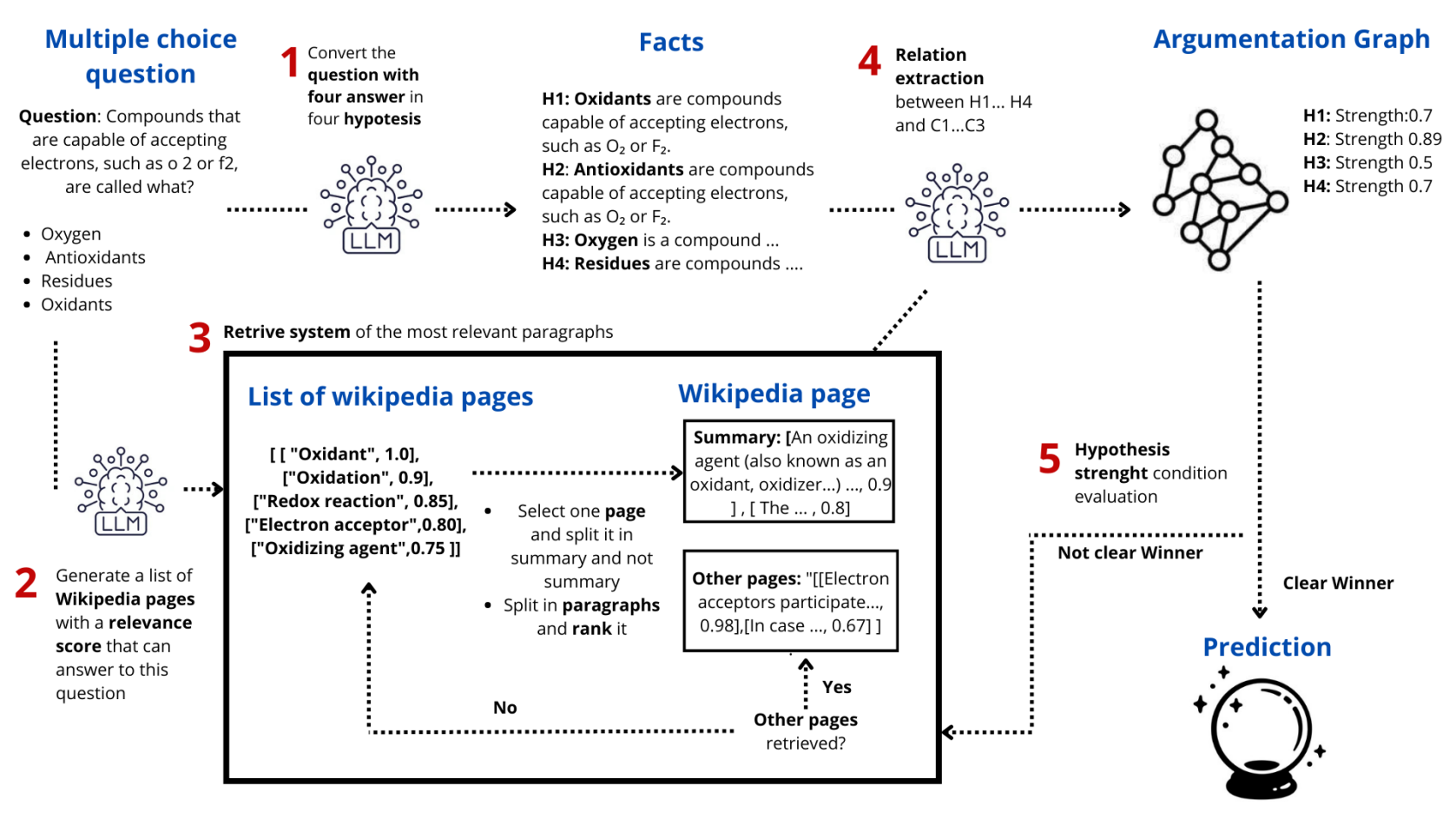


Figure 4: Pipeline from question to answer

(5) Reasoning Until Convinced: Stopping Strategy

- **Dynamic Evaluation:** The strength of each hypothesis is recalculated whenever new arguments are added to the graph.
- **Greedy Heuristic:** The system analyzes the most relevant information first (e.g., Wikipedia summary) and proceeds only if necessary.
- **"Clear Winner" Conditions:** The process stops as soon as a hypothesis dominates others according to two key parameters:

(5) Reasoning Until Convinced: Stopping Strategy

Stopping Criteria

1. **Strength Threshold (h):** The best hypothesis must exceed a minimum confidence value (e.g., 0.6 or 0.7).
2. **Safety Margin (m):** The difference between the first and second hypothesis must be sufficiently large (e.g., 0.15 or 0.2).

- **Efficiency:** If conditions are met, unnecessary paragraphs are skipped, saving time and tokens.
- **Fallback:** If all sources are exhausted without a clear winner, the hypothesis with the highest strength is chosen.

Experiments

Experimental Setup

The system was tested using open-source models to ensure reproducibility.

Selected Models

- **GPT-oss 20B**
- **Qwen 14B**
- **Sentence-BERT**: (*all-mpnet-base-v2*) for ranking.

Benchmark

- **SciQ**: Science questions.
- **ARC-Challenge**: Hard reasoning questions.
- **PlausibleQA**: Anti-memorization test.

Methodology: Extraction and Reasoning

- **Claim Extraction**: Transform Wikipedia text into atomic propositions.
- **Bipolar Relations**: Identify Supports ($+$) and Attacks ($-$) towards hypotheses.
- **Greedy Strategy**: The system iteratively analyzes data until sufficient confidence is reached.

Results: SciQ Benchmark

Model / Condition	SciQ (Acc.)
Qwen 3:14B (Ours)	
Complete	85.00
Uncertainty removed	**88.69**
GPT-oss:20B (Ours)	
Complete	88.10
Uncertainty removed	**89.07**
Open-Source (Baselines)	
Gemma 2 9B	91.00
LLaMA 3.1 70B	95.00

SciQ Insight Removing uncertainty allows 14B/20B models to approach the performance of much larger models (70B).

Results: ARC Easy Benchmark

Model / Condition	ARC Easy
Qwen 3:14B (Ours)	
Uncertainty removed	**89.44**
GPT-oss:20B (Ours)	
Uncertainty removed	**88.89**
Commercial Models	
GPT-3.5 Turbo	80.30
GPT-4	89.70
Open-Source	
LLaMA 3.1 8B	92.07

ARC Analysis The RUC-QA framework significantly outperforms GPT-3.5 Turbo and nearly matches GPT-4.

Results: PlausibleQA Hard

Model / Condition	PlausibleQA Hard
Qwen 3:14B (Ours - Uncertainty removed)	**65.81**
GPT-oss:20B (Ours - Uncertainty removed)	**65.08**
Open-Source (Large Models)	
Qwen 2.5 72B	60.00
LLaMA 3.1 70B	60.10
Baseline GPT-oss:20B	
No context	57.84
Top 10 Retrieval	65.85

Contamination Note On PlausibleQA Hard, RUC-QA outperforms 70B+ models, showing that graph reasoning prevents reliance on model memory alone.

Prediction Error

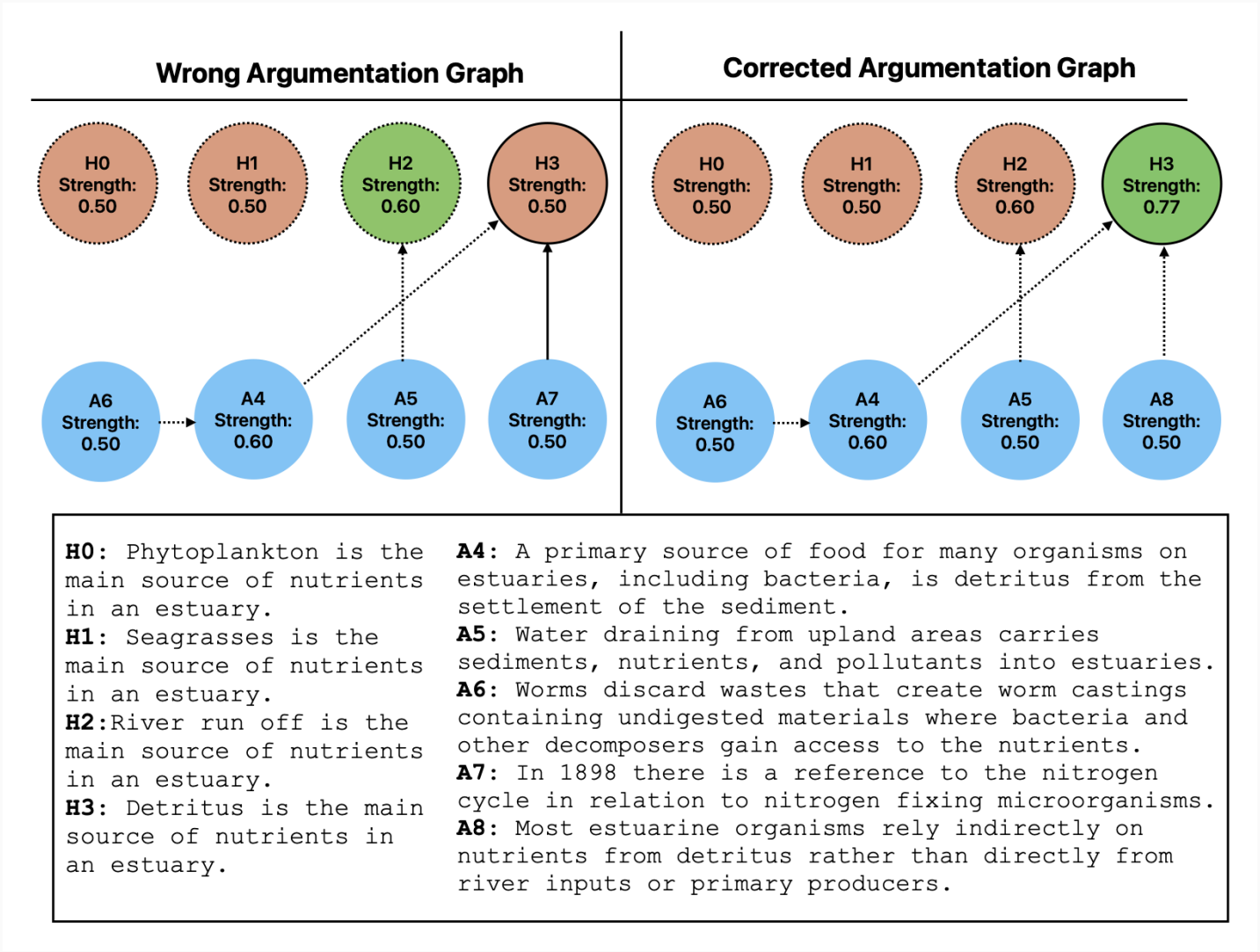


Figure 5: Pipeline from question to answer

Efficiency and Validation

- **Impact of "Uncertainty Removed"**: Improvements up to 22% on PlausibleQA, indicating correct identification of ambiguous cases.
- **Optimal Parameters**: Threshold $h = 0.7$ and Margin $m = 0.15$ balance rigor and speed.
- **Early Stopping**: Reduces computational load by 25-30% compared to standard RAG.

Thank you for your attention!