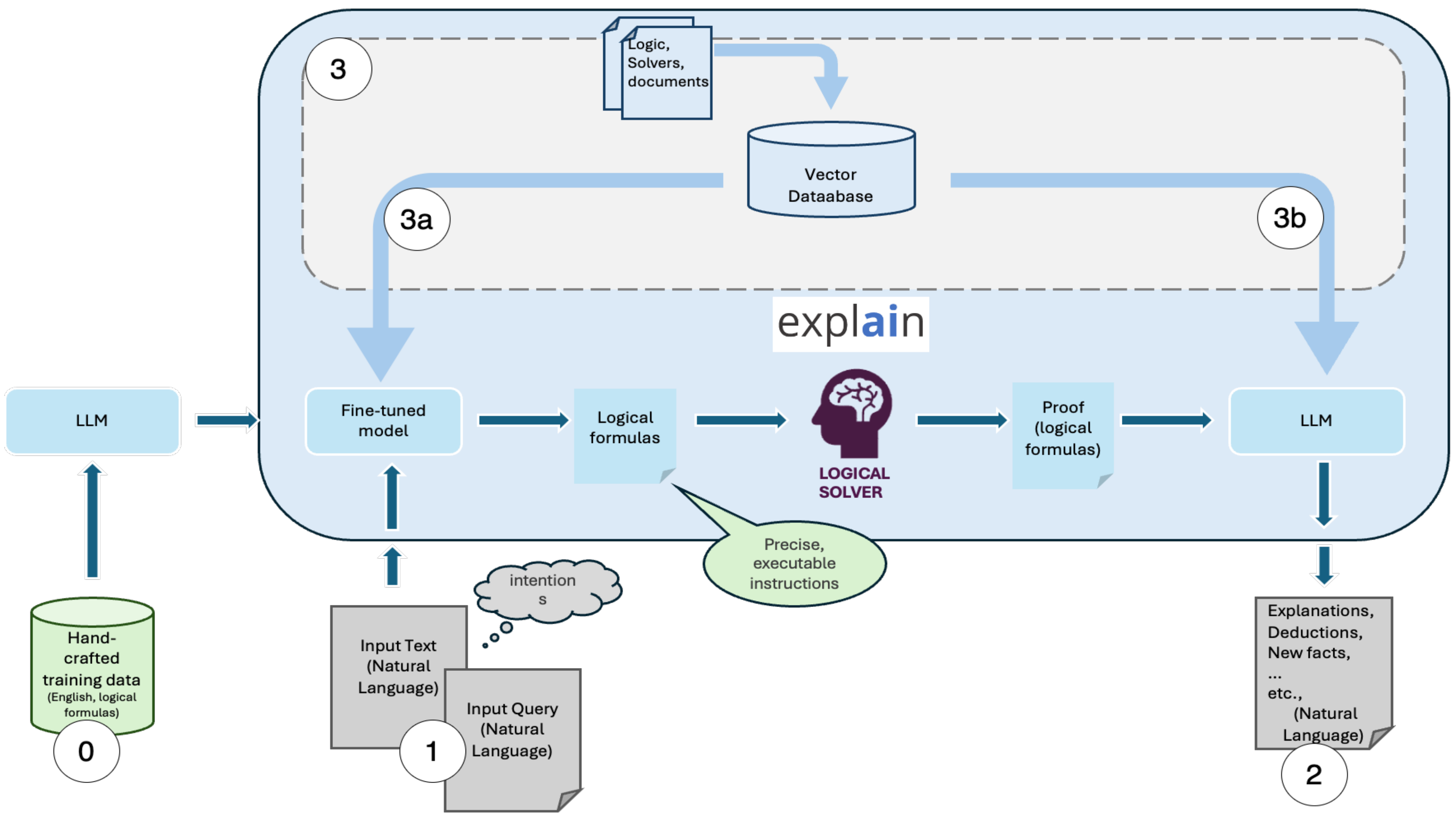
Generative AI with Logical Reasoning

Kalyan Krishnamani
NVIDIA


3
Logic, Solvers, documents
Vector Dataabase
3a
3b
explain
LLM
Fine-tuned model
Logical formulas
LOGICAL SOLVER
Proof (logical formulas)
LLM
Precise, executable instructions
intentions
Hand-crafted training data (English, logical formulas)
0
Input Text (Natural Language)
Input Query (Natural Language)
1
Explanations, Deductions, New facts, ... etc., (Natural Language)
2

Introduction

Today's Generative AI systems are limited in their reasoning capabilities. Even when they reason, it is achieved by training the LLM on a large reasoning dataset (*pattern matching based reasoning*). They hence appear to reason logically while their reasoning stems from the pattern trained on, or reward optimized for during their training. With Generative AI permeating more and more applications, *logical reasoning* is becoming imperative, especially in certain critical domains like finance, autonomous systems, healthcare, software engineering, etc. This work attempts to infuse logical reasoning into Generative AI.

The explain system

1. Transform the input natural language (context and queries) text into logical formulas.

2. Solve the logical formulas using a solver based on mathematical logic (e.g., a theorem prover).

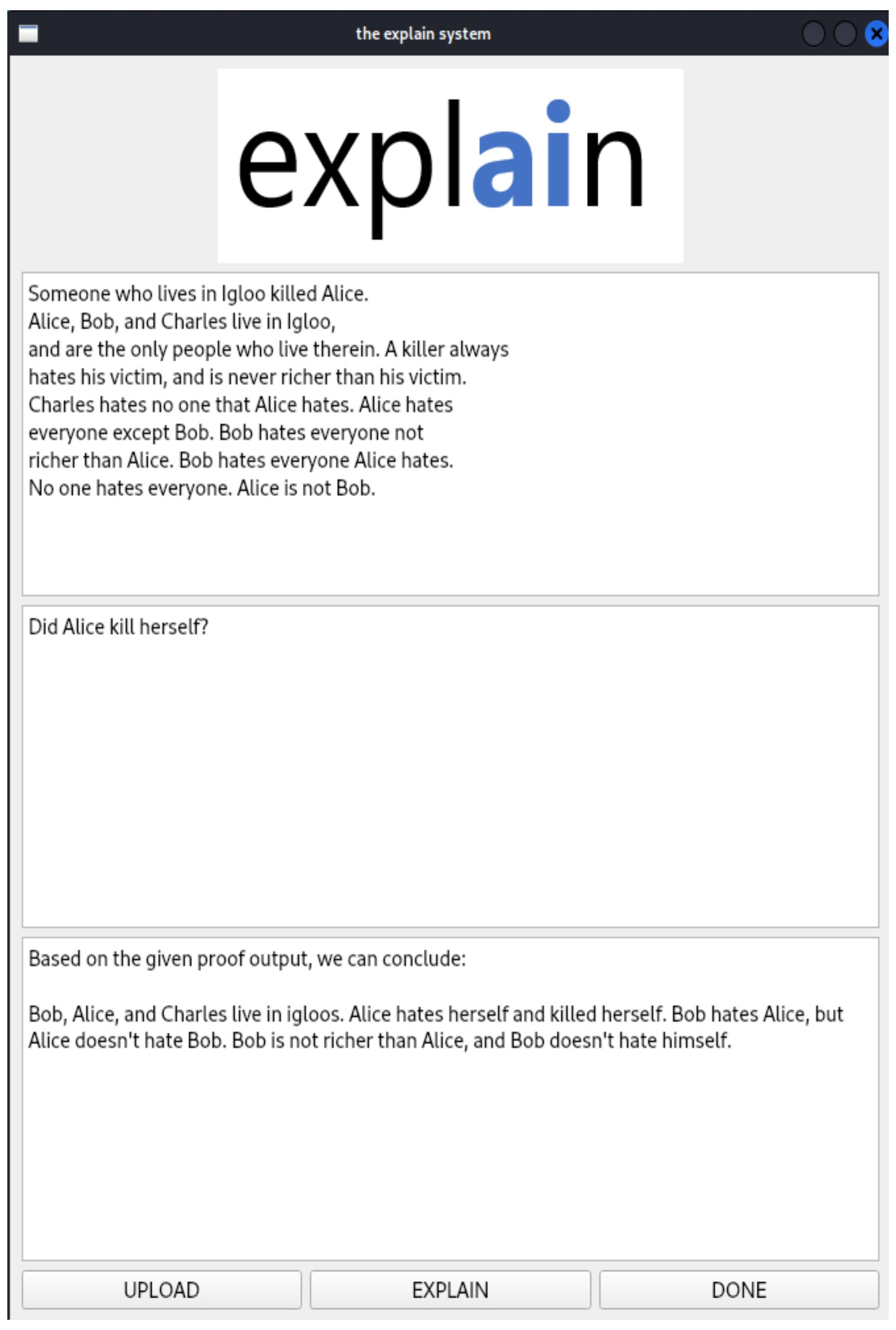
3. Transform the proof results from the solver back into natural language for the user.

Steps 1. and 3. are achieved using an LLM. Step 2. is achieved using a logical solver, in this case a theorem prover. Optionally, the system also has a RAG component **(3)** that can help with steps 1. and/or 3.

Who killed Alice?

"Someone who lives in Igloo killed Alice. Alice, Bob, and Charles live in Igloo, and are the only people who live therein. A killer always hates his victim and is never richer than his victim. Charles hates no one that Alice hates. Alice hates everyone except Bob. Bob hates everyone not richer than Alice. Bob hates everyone Alice hates. No one hates everyone. Alice is not Bob."

Did Alice kill herself?



Someone who lives in Igloo killed Alice.
Alice, Bob, and Charles live in Igloo,
and are the only people who live therein. A killer always
hates his victim, and is never richer than his victim.
Charles hates no one that Alice hates. Alice hates
everyone except Bob. Bob hates everyone not
richer than Alice. Bob hates everyone Alice hates.
No one hates everyone. Alice is not Bob.

Did Alice kill herself?

Based on the given proof output, we can conclude:

Bob, Alice, and Charles live in igloos. Alice hates herself and killed herself. Bob hates Alice, but Alice doesn't hate Bob. Bob is not richer than Alice, and Bob doesn't hate himself.

Natural Language to Logical Language

An accessible LLM is fine-tuned using a dataset of natural language statements and logical formula pairs **(0)**. This fine-tuned LLM is used to generate axioms, hypotheses and conjectures (logical formulas) from user provided natural language text - context text and an optional user query **(1)**. The logical formulas generated by the fine-tuned LLM would be:

- A set of axioms
- A set of hypotheses
- A conjecture (user provided query)

Hypotheses (assumptions) are generated for aspects that are not explicitly stated but if true could affect the result of the system. An example encoding of an axiom and a conjecture is provided below (in TPTP syntax):

```
fof(killer_hates_victim, axiom,
    ![X, Y] : (killed(X, Y) => hates(X, Y)).

fof(alice_kills_alice, conjecture,
    killed(alice, alice)).
```

Challenges

- Frame problem in Logic[1]:
  If $\neg DoorOpen(0) \wedge \neg LightOn(0) \wedge DoorOpen(1)$,
  Does $\neg LightOn(1)$ hold?

- Encoding implicit semantic information:
  Kal lives in NYC vs. *Kal spends time in Big Apple*

Note: The goal of this work is *not to* develop an LLM that can reason, but *to infuse logical reasoning into pattern-matching based reasoning* (present day Generative AI). The explain system will improve and scale as the underlying LLM improves and scales.

Logical Reasoning

The logical solver:

- Checks consistency of axioms.
- Tries to prove the conjecture from the axioms and hypotheses
- Identifies hypotheses that prove/disprove the conjecture.

The logical formulas themselves can involve several theories (integers, reals, etc.) and quantifiers catering to different application domains (e.g., software engineering). The logical solver output is the form of deductions, resolution rules, etc. An LLM is used to extract the main aspects of the proof, with respect to the user query, and present it in natural language **(2)**. This includes deduction of new information that is not explicitly stated, logical derivation of the conjecture or its negation, etc.

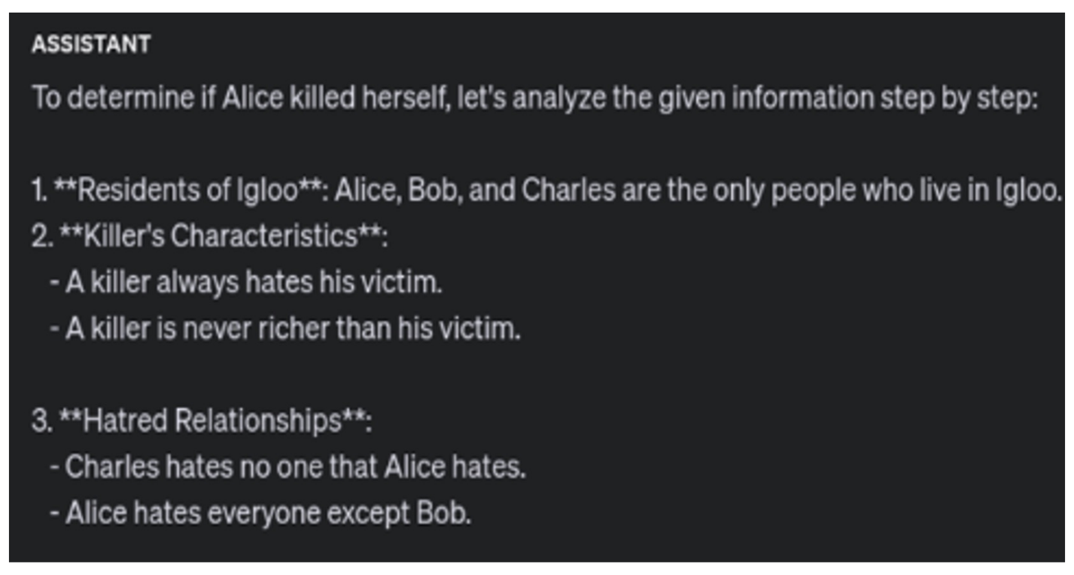
Advantages

- Logical reasoning
- Deriving new facts (logically)
- Provable correctness

Future Work

- Identifying and generating frame axioms and relevant hypotheses.
- Employing knowledge graphs[2] and/or Ontology matching to augment the formula generation.

GPT-4o



ASSISTANT

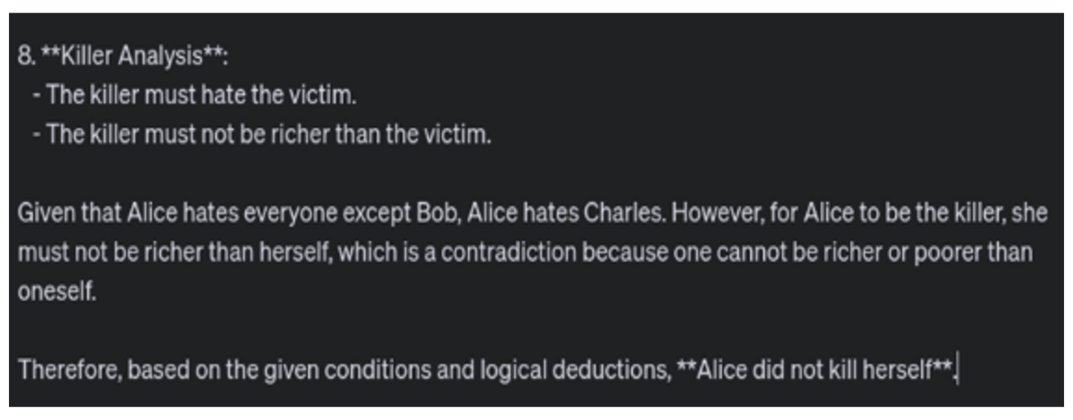
To determine if Alice killed herself, let's analyze the given information step by step:

1. **Residents of Igloo**: Alice, Bob, and Charles are the only people who live in Igloo.
2. **Killer's Characteristics**:
 - A killer always hates his victim.
 - A killer is never richer than his victim.

3. **Hatred Relationships**:
 - Charles hates no one that Alice hates.
 - Alice hates everyone except Bob.

...

...

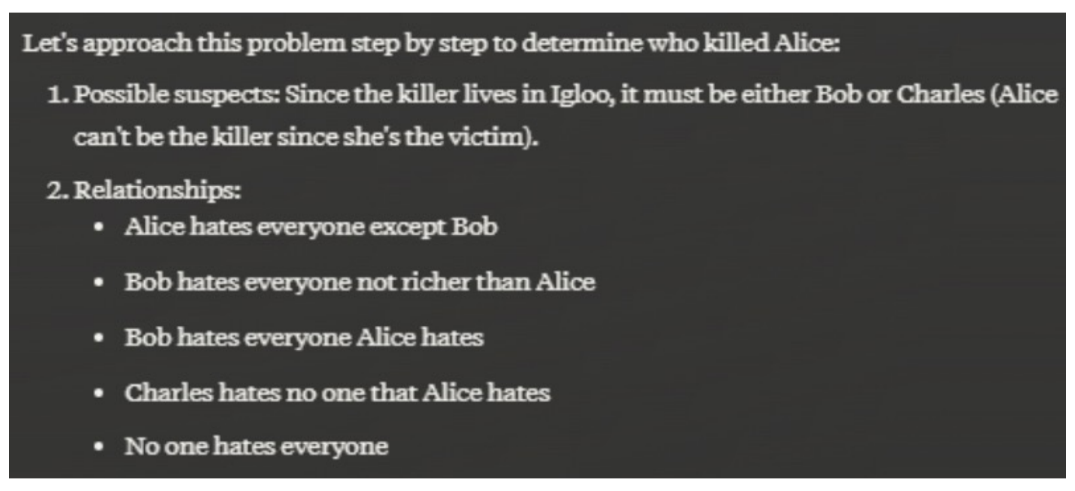


8. **Killer Analysis**:
 - The killer must hate the victim.
 - The killer must not be richer than the victim.

Given that Alice hates everyone except Bob, Alice hates Charles. However, for Alice to be the killer, she must not be richer than herself, which is a contradiction because one cannot be richer or poorer than oneself.

Therefore, based on the given conditions and logical deductions, **Alice did not kill herself**.

Claude-3.5-Sonnet

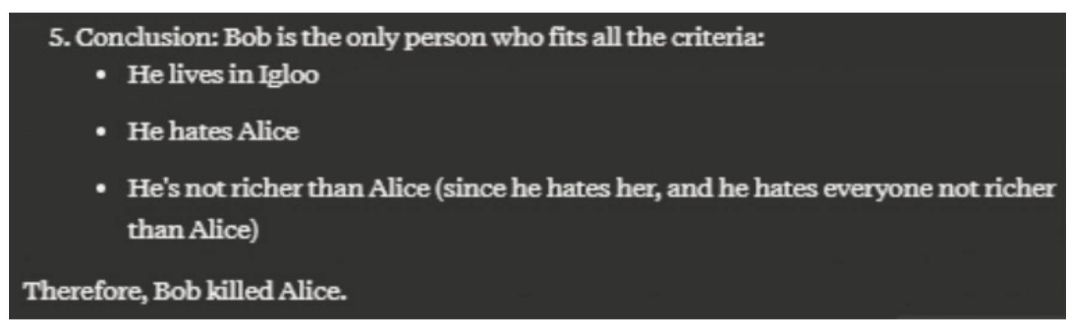


Let's approach this problem step by step to determine who killed Alice:

1. Possible suspects: Since the killer lives in Igloo, it must be either Bob or Charles (Alice can't be the killer since she's the victim).
2. Relationships:
   - Alice hates everyone except Bob
   - Bob hates everyone not richer than Alice
   - Bob hates everyone Alice hates
   - Charles hates no one that Alice hates
   - No one hates everyone

...

...



5. Conclusion: Bob is the only person who fits all the criteria:
   - He lives in Igloo
   - He hates Alice
   - He's not richer than Alice (since he hates her, and he hates everyone not richer than Alice)

Therefore, Bob killed Alice.

References

1. The Frame Problem *https://plato.stanford.edu/entries/frame-problem/*
2. Yasunaga, M., Bosselut, A., Ren, H., Zhang, X. Manning, C.D., Liang, P. & Leskovec, J. *Deep Bidirectional Language-Knowledge Graph Pretraining.*, NeurIPS 2022.