MIT App Inventor Punya AI Reasoning Explainability

Brendan Capuzzo | Advisor: Dr. Oshani Seneviratne

Abstract

AI-powered mobile apps often fail to provide clear explanations for their decisions. MIT App Inventor Punya is an Android app development software that includes a rule-based reasoner, but it offers limited insight into its reasoning process. To foster user trust, interpretable explanations are essential for making complex AI/ML decision-making processes more transparent.

Introduction

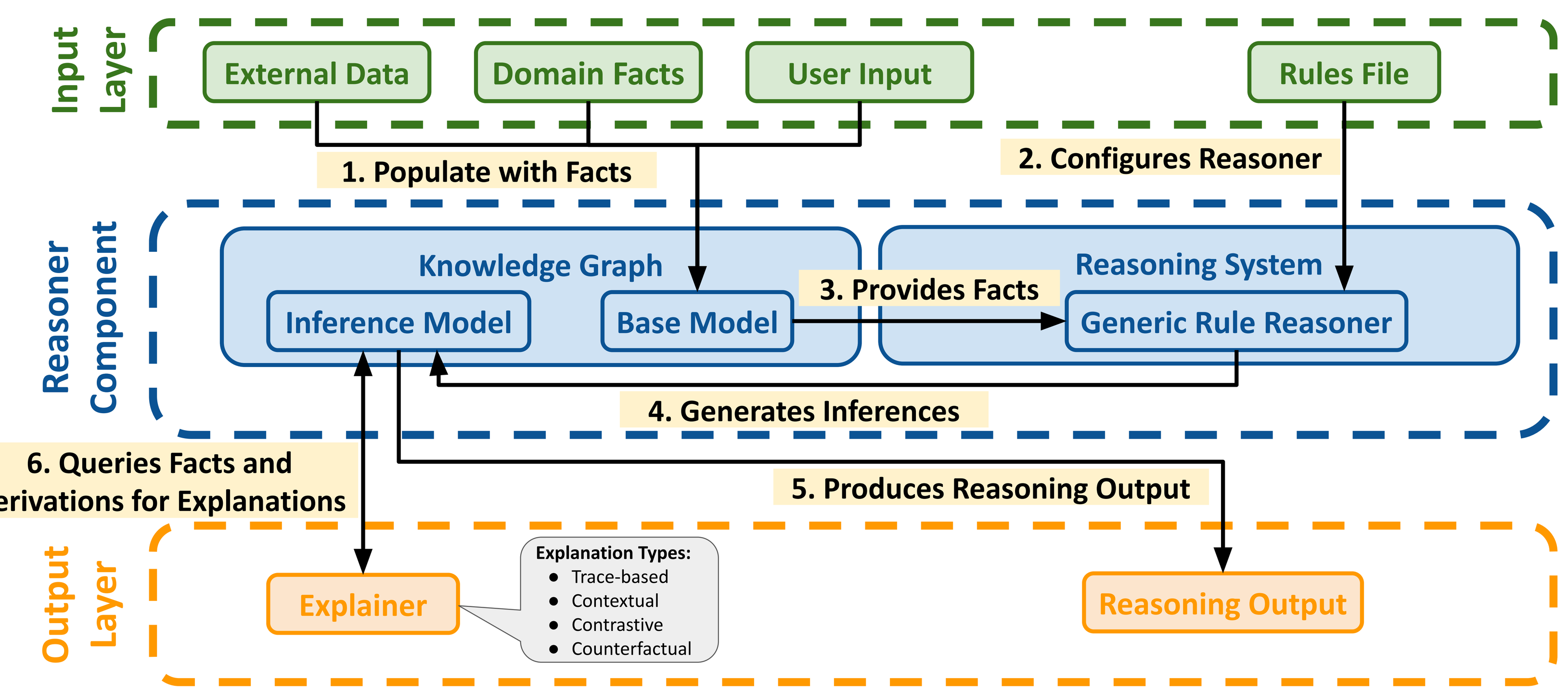
- AI systems often make decisions without revealing their reasoning, posing challenges in critical fields like healthcare, finance, and law
- Transparency is essential for users to trust and verify decisions that impact lives
- Decisions are represented using RDF (Resource Description Framework) triples:
  - Subject: The entity (e.g., loan applicant)
  - Predicate: The relationship/attribute (e.g., credit score)
  - Object: The value/outcome (e.g., "Eligible")

Explanation Types

Explanation	Definition
Trace-based	▪ Shows step-by-step reasoning chain ▪ Explains "how" the system reached its conclusion ▪ Maps reasoning rules to input facts
Contextual	▪ Considers surrounding circumstances ▪ Includes user situation and environment ▪ Explains relevance of external factors
Contrastive	▪ Compares different outcomes ▪ Highlights key differences between scenarios ▪ Explains why a result occurred instead of another
Counterfactual	▪ Explores "what-if" scenarios ▪ Shows how changing inputs affects outcomes ▪ Identifies minimal changes needed for different results

[1]

Reasoning Architecture



Example Explanations

Base Knowledge

Model evaluates loan applications using RDF triples that represent:
- Applicant attributes (credit score, monthly income, monthly debt)
- Calculated metrics (DTI ratio = monthly debt/monthly income)
- Decision outcomes (Eligible or Not Eligible)

Input Triple: (applicant1, loanEligibility, Not Eligible)

Rules

1. DTI Rule: Calculates debt-to-income ratio from monthly debt and income
2. Eligibility Rules:
   - If DTI ratio > 0.35: Not Eligible
   - If credit score < 620: Not Eligible
   - Otherwise: Eligible

Applicant1 Facts

- Monthly Income: $5000
- Monthly Debt: $2000
- DTI Ratio: 0.4
- Credit Score: 680
- Eligibility: Not Eligible

Trace-Based

Conclusion: applicant1 has Loan Eligibility: Not Eligible used the following matches:
Match: applicant1 has Type: Person
Match: applicant1 has DTI Ratio: 0.4
Conclusion: applicant1 has DTI Ratio: 0.4 used the following matches:
Match: applicant1 has Type: Person
Match: applicant1 has Monthly Debt: 2000.0
Match: applicant1 has Monthly Income: 5000.0
And paired them with the following rule:
[[DTIRule: (?applicant type Person) (?applicant monthlyDebt ?debt) (?applicant monthlyIncome ?income) quotient(?debt ?income ?dti) -> (?applicant dtiRatio ?dti)]]
to reach this conclusion.

And paired them with the following rule:
[[NotEligibleDTIRule: (?applicant type Person) (?applicant dtiRatio ?dti) greaterThan(?dti '0.349999') -> (?applicant loanEligibility 'Not Eligible')]]
to reach this conclusion.

Contextual

Shallow Explanation:
Conclusion: applicant1 has Loan Eligibility: Not Eligible
Based on rule: [NotEligibleDTIRule: (?applicant type Person) (?applicant dtiRatio ?dti) greaterThan(?dti '0.349999' -> (?applicant loanEligibility 'Not Eligible')]
Using the following facts:
• applicant1 has Type: Person
• applicant1 has DTI Ratio: 0.4

Simple Explanation:
applicant1 has Loan Eligibility: Not Eligible because applicant1 has Type: Person and applicant1 has DTI Ratio: 0.4.

Contrastive

Similarities:
• applicant1 has Monthly Income: 5000.0

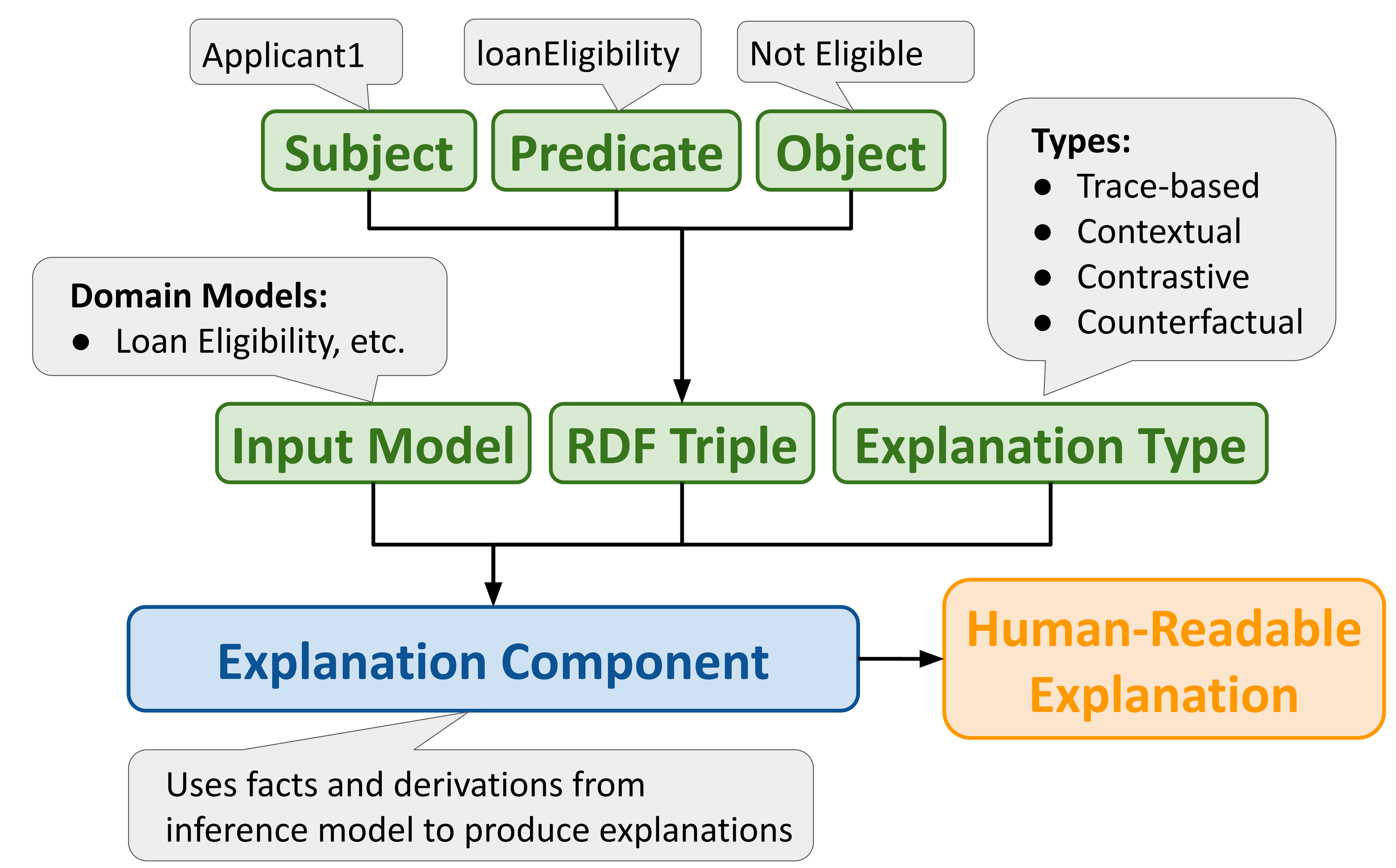
Differences:
• For Monthly Debt: this model has 2000.00 while alternate model has 1000.00
• For Loan Eligibility: this model has Not Eligible while alternate model has Eligible
• For Credit Score: this model has 680 while alternate model has 700
• For DTI Ratio: this model has 0.40 while alternate model has 0.20

Counterfactual

To change the outcome for applicant1 has Loan Eligibility: Not Eligible, you could look at these examples:

applicant3 has Loan Eligibility: Eligible because:
- Their applicant3 has DTI Ratio: 0.2 while yours is applicant1 has DTI Ratio: 0.4
- Their applicant3 has Monthly Debt: 1000.0 while yours is applicant1 has Monthly Debt: 2000.0
- Their applicant3 has Credit Score: 700 while yours is applicant1 has Credit Score: 680

Explanation Component



Conclusion

- Successfully implemented multiple explanation types for MIT App Inventor Punya reasoning component
- Created a framework for future expansion to more explanation types
- Demonstrated feasibility for explainer component on mobile devices

Future Work

- Integration with more complex AI models (Neural networks)
- Expand offerings for explanation types
- Optimize explanation outputs with NLP and accuracy scores
- Perform user studies on explanation effectiveness
- Deploy explanation component into MIT App Inventor Punya subproject

References

[1] S. Chari et al., "Explanation Ontology: A general-purpose, semantic representation for supporting user-centered explanations," Semantic web, pp. 1–31, May 2023, doi: https://doi.org/10.3233/sw-233282.