Final Term Project Report

Adrian Hoang

2025-05-04

1 Reflection on Model Behavior

The fine-tuned model achieved an accuracy of 4 out of 6 (66.7%) compared to the SMT verdicts on the evaluation set.

Table 1: Evaluation Summary: LLM vs. SMT Verdicts

Case ID	Fact Pattern Summary	Model Output	SMT Verdict	Status
Case 1	Marcus & Lily (23yo QC Student, >50% support, low income)	Yes	Yes	Correct
Case 2	Ellen & Zack (17yo Nephew, fails residency, fails self-support, high income)	Yes	No	Error
Case 3	Raj & Meena (68yo Mother, QR, low income, valid MSA assumed)	Yes	Yes	Correct
Case 4	Luis & Paco (35yo Cousin/HH Member, >50% support, fails QR income)	No	No	Correct
Case 5	Helen & Irene (90yo Mother, QR, low income, ambiguous/verbal MSA)	Yes	No	Ambiguous
Case 6	Brenda & Carl (85yo Father, QR, low income, ambiguous support level)	No	No	Ambiguous

Strengths

The model demonstrated competence in identifying several common dependency patterns. It correctly classified a standard Qualifying Child (QC) scenario involving the student age rule (Case 1) and a Qualifying Relative (QR) case failing the income test (Case 4). Encouragingly, it also correctly handled the complex Multiple Support Agreement scenario (Case 3, assuming the SMT reflects the MSA assertion) and the ambiguous support scenario (Case 6), aligning with the SMT results. This suggests the fine-tuning effectively captured core patterns related to parent/child relationships, student status, basic income thresholds, and potentially MSA contexts present in the training data.

Weaknesses and Errors

The model exhibited significant weaknesses when scenarios required strict application of multiple negative constraints or interpretation of nuanced legal requirements not easily represented by simple patterns.

- **Case 2 (Ellen & Zack):** The model incorrectly predicted "Yes". Zack failed multiple tests: QC residency ($<$ 6 months), QC/QR self-support ($>$50%), and QR income ($>$\$4700). The LLM likely overweighted positive signals (age 17, nephew relationship) and failed to apply the multiple disqualifying factors rigorously. This points to difficulty in handling complex conjunctions of negative constraints.
- **Case 5 (Helen & Irene):** The model predicted "Yes", while the SMT verdict was "No". This case involved a verbally agreed MSA, which lacks the required written declaration for legal validity. The SMT model, likely requiring an explicit assertion of a *valid* MSA (which couldn't be made), defaulted to "No" dependency via MSA. The LLM likely interpreted the description of contribution and agreement as sufficient, failing to capture the strict procedural requirement of a written declaration, potentially due to lack of such specific failure examples in training.

Handling of Ambiguity

The two cases designed with ambiguity (Case 5 - MSA validity, Case 6 - support level) were resolved to "No" by the SMT solver based on the provided Z3 output (likely because the unasserted conditions needed for a "Yes" verdict defaulted to false). The LLM correctly matched the SMT verdict for Case 6 (ambiguous support, predicting "No") but erred on Case 5 (ambiguous MSA, predicting "Yes"). This suggests the LLM might make optimistic or pattern-based guesses when faced with underspecified information (like assuming the verbal MSA was sufficient), whereas the SMT solver requires explicit assertion for positive conditions.

Overall Assessment

The fine-tuning process enabled the LLM to learn basic dependency patterns effectively. However, its performance degraded significantly on cases requiring strict adherence to multiple constraints, precise numerical boundaries, or nuanced legal formalities (like MSA requirements). The errors highlight the difference between pattern recognition learned during fine-tuning and the rigorous logical deduction performed by the SMT solver. This project underscores the limitations of current LLMs for high-stakes compliance tasks requiring guaranteed rule adherence and demonstrates the significant value of SMT as a tool for verification, evaluation, and potentially "guardrailing" LLM outputs in complex, rule-based domains like law. Future improvements would necessitate a much larger, more diverse dataset focusing explicitly on edge cases, numerical boundaries, and specific failure conditions for each rule.