UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
THEORETISCHE INFORMATIK

Can semantic similarities of words enhance common sense reasoning? A case study with prover E and SUMO.

Können semantische Ähnlichkeiten von Wörtern die Schlussfolgerungen des gesunden Menschenverstands verbessern? Eine Fallstudie mit Prover E und SUMO.

Bachelorarbeit

verfasst am
Institut für Software Engineering und Programming Languages

im Rahmen des Studiengangs
Informatik
der Universität zu Lübeck

vorgelegt von
Julian Britz

ausgegeben und betreut von
Prof. Dr. Diedrich Wolter

mit Unterstützung von
Moritz Bayerkuhnlein

Lübeck, den 06. Juli 2025

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

_____

Julian Britz

Zusammenfassung

Zusammenfassung

Abstract

Abstract

Acknowledgements

Ackbowledgements

Contents

1 Introduction 1

1.1 Problem Statement 1
1.2 Entanglement Between Natural Language and Logical Grammar 1
1.3 Formal Hypothesis 1
1.4 Justification and Implications 2
1.5 Related Work 2
1.6 Structure of this Thesis 2

2 Preliminaries 3

2.1 Common Sense Reasoning 3
2.2 Word Embeddings 3
2.3 Grammars 3
2.4 Theorem Provers 3
2.5 Selection Strategies 4

3 Experiments 5

3.1 How WhiteboxTests are made 5
3.2 Reengineering of paper (**Schon2024**) 5
3.3 Analyze of reeingineering 7
3.4 Add core axioms 10
3.5 Comparing LLMs 11
3.6 Analyze different E settings 12
3.7 Different prover 13
3.8 Time study 14

4 Conclusion 16

5 Further work 17

1

Introduction

In this thesis, we investigate how adding frequently used axioms ("core axioms") to a subset of the Adimen SUMO grammar, selected based on a combination of syntactic and semantic criteria, enhances the success rate of automated theorem proving.

1.1 Problem Statement

- Given a formal logical grammar, such as Adimen SUMO, and a conjecture C, we aim to determine a subset of axioms that maximizes the probability of proving C.
- We use a combination of syntactic and semantic similarity measures to select an initial set of relevant axioms $\mathcal{A}_{\text{rel}}$.
- We hypothesize that augmenting this set with frequently used axioms $\mathcal{A}_{\text{core}}$ further improves proof success.

1.2 Entanglement Between Natural Language and Logical Grammar

- Adimen SUMO, partially derived from WordNet, reflects structured knowledge from natural language.
- Natural language inherently follows logical structures, even if it remains ambiguous and context-dependent.
- Language models like SBERT capture semantic relationships, supporting the idea that formal logical expressions share deep structural similarities with natural language.
- We propose an **entanglement hypothesis**, where natural language and logical structures influence each other:

$$\mathcal{L}_{\text{NL}} \bowtie \mathcal{L}_{\text{logic}} \tag{1.1}$$

1.3 Formal Hypothesis

- Define:

  - $\mathcal{A}$: The set of all axioms in Adimen SUMO.
  - C: A conjecture to be proven.
  - $d : \mathcal{A} \times C \to \mathbb{R}$: A similarity function measuring syntactic and semantic closeness.
  - $\mathcal{A}_{\text{rel}}$: The subset of axioms, made out of k nearest axioms to C:

$$\mathcal{A}_{\text{rel}} = \{A \in \mathcal{A} \mid A \text{ is among the } k \text{ closest axioms to } C \text{ w.r.t. } d\} \tag{1.2}$$

  - $\mathcal{A}_{\text{core}}$: A set of frequently used axioms in previous proofs.
  - $\mathcal{A}_{\text{enh}} = \mathcal{A}_{\text{rel}} \cup \mathcal{A}_{\text{core}}$: The enhanced axiom set.
  - $P(T, \mathcal{A}', C)$: A function that returns True if theorem prover T finds a proof of C using $\mathcal{A}'$, otherwise False.
- Hypothesis:

$$\forall C \in \mathcal{C}, \quad \Pr(P(T, \mathcal{A}_{\text{enh}}, C) = \text{True}) > \Pr(P(T, \mathcal{A}_{\text{rel}}, C) = \text{True}) \tag{1.3}$$

- This suggests that the success probability of proving C increases when $\mathcal{A}_{\text{core}}$ is included.

1.4 Justification and Implications

- Since Adimen SUMO's grammar is **entangled** with natural language structures, selecting axioms using semantic embeddings (e.g., SBERT) captures implicit logical dependencies.
- Frequently used axioms act as **structural anchors** within this entanglement, improving proof discovery.
- The theorem prover benefits from both the **nearest axioms** and the **core axioms**, leading to an overall higher proof success rate.

This thesis aims to empirically validate this hypothesis by evaluating proof success rates with different axiom selection strategies and demonstrating the effectiveness of incorporating core axioms.

1.5 Related Work

(**Schon2024**) (**Àlvez2014**) (**Hoder2011**) (**Roederer2009**) (**Sutcliffe2007**)

1.6 Structure of this Thesis

Structure of this Thesis.

2

Preliminaries

2.1 Common Sense Reasoning

- **Explanation:** Common sense reasoning refers to the ability of a system to make presumptions about the everyday world, similar to human reasoning.
- **What is it?** It involves reasoning about typical situations, handling incomplete information, and making reasonable assumptions.
- **Applications and Use Cases:**
  - Natural language understanding.
  - Automated reasoning.
  - Robotics and AI decision-making.
- **Current State:** Modern AI systems, including deep learning models, struggle with common sense reasoning, but efforts such as knowledge graphs and large-scale pre-trained models aim to improve it.
- **Challenges:**
  - Ambiguity in natural language.
  - Lack of structured common-sense knowledge.
  - Computational limitations in reasoning.

2.2 Word Embeddings

- **Definition:** Word embeddings are dense vector representations of words that capture semantic meanings.
- **Creation:** Generated using neural network-based models like Word2Vec, GloVe, and FastText.
- **Properties:**
  - Captures semantic similarity.
  - Enables algebraic operations like $king - man + woman = queen$.
- **Applications:**
  - Natural language processing (NLP).

  - Machine translation.
  - Information retrieval.
- **Limitations and Challenges:**
  - Contextual ambiguity.
  - Bias in training data.
  - Scalability for large vocabularies.

2.3 Grammars

- **Explanation:** Grammars define the structure and rules governing a language, whether natural or formal.
- **Types of Grammars:** Context-free grammars, dependency grammars, formal grammars.
- **Structure and Organization:** Grammars consist of syntax rules, lexicons, and transformations.
- **Why Are They Needed?** They provide the foundation for parsing, generating, and interpreting language in NLP and logic systems.
- **SUMO:** Suggested Upper Merged Ontology, a formal ontology integrating linguistic and logical structures.

2.4 Theorem Provers

- **Explanation:** Theorem provers are automated reasoning tools designed to validate logical statements based on formal rules.
- **Types of Theorem Provers:** Resolution-based, tableau-based, and interactive provers.
- **Functionality:** They work by refutation—attempting to derive contradictions from the negation of a theorem.
- **Why Are They Needed?** Essential for verifying mathematical proofs, formal verification, and logic-based AI.
- **Limits and Challenges:**
  - Computational complexity.
  - Handling undecidable problems.
  - Scalability for large proof spaces.
- **Prover E:**
  - **How Does E Work?** A saturation-based theorem prover using superposition calculus.
  - **Satauto Mode:** An automatic mode that selects strategies dynamically for proof search.

2.5 Selection Strategies

- **Syntactic:** Selects axioms based on structural similarity and pattern matching.
- **Semantic:** Uses embeddings and meaning-based similarity to find relevant axioms.
- **Combination - Union:** Integrates syntactic and semantic methods to enhance selection efficiency.

3

Experiments

3.1 How WhiteboxTests are made

- Paper of Álvez about building WhiteBox Tests (**Álvez2017**)
  1. In the context of Adimen SUMO, **falsity tests** are formulated by taking specific statements or conditions and applying **negation** to them.
  2. These negated statements serve as a basis for falsity tests, which are essential in logical proof systems.
  3. A **theorem prover** such as Prover E then performs a process known as **refutation**.
  4. Through refutation, the theorem prover attempts to demonstrate the **inconsistency** or **falsehood** of the original statements.
  5. This is achieved by attempting to derive a **contradiction** from the negated statements, indicating that the original (non-negated) statements cannot all be true simultaneously.

3.2 Reengineering of paper (**Schon2024**)

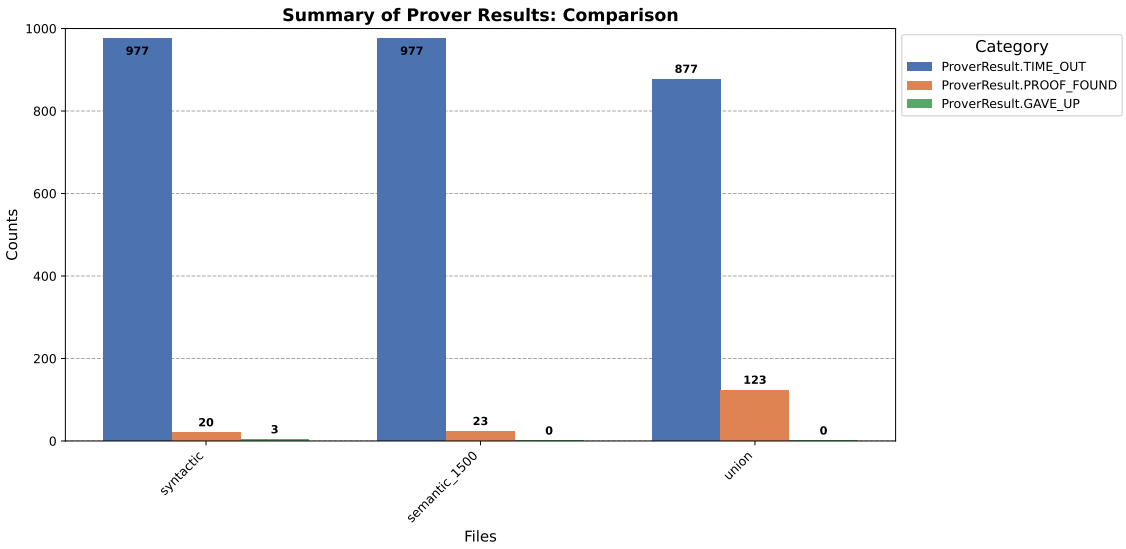

Figure 3.1: Reengineering of results of paper (**Schon2024**)

3.3 Analyze of reeingineering

– Mean variable count

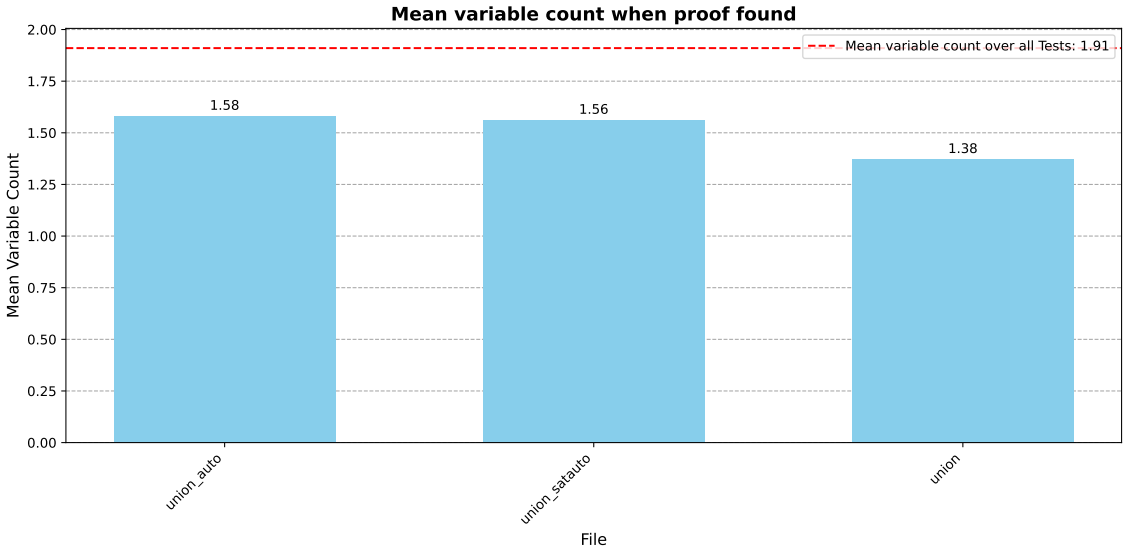

Figure 3.2: Variable Count

– Count signs

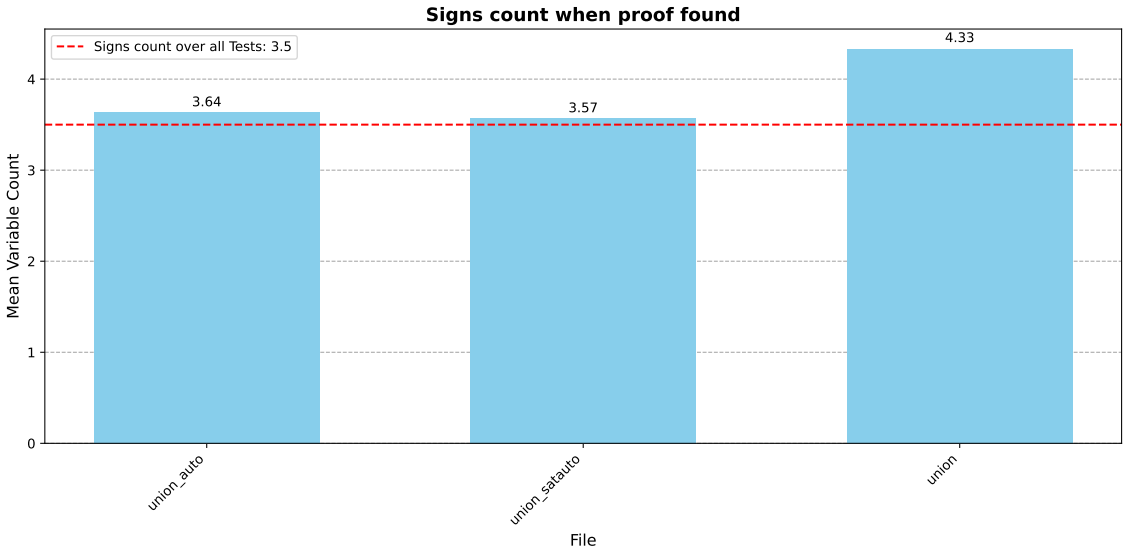

Figure 3.3: Count signs

– Character Count

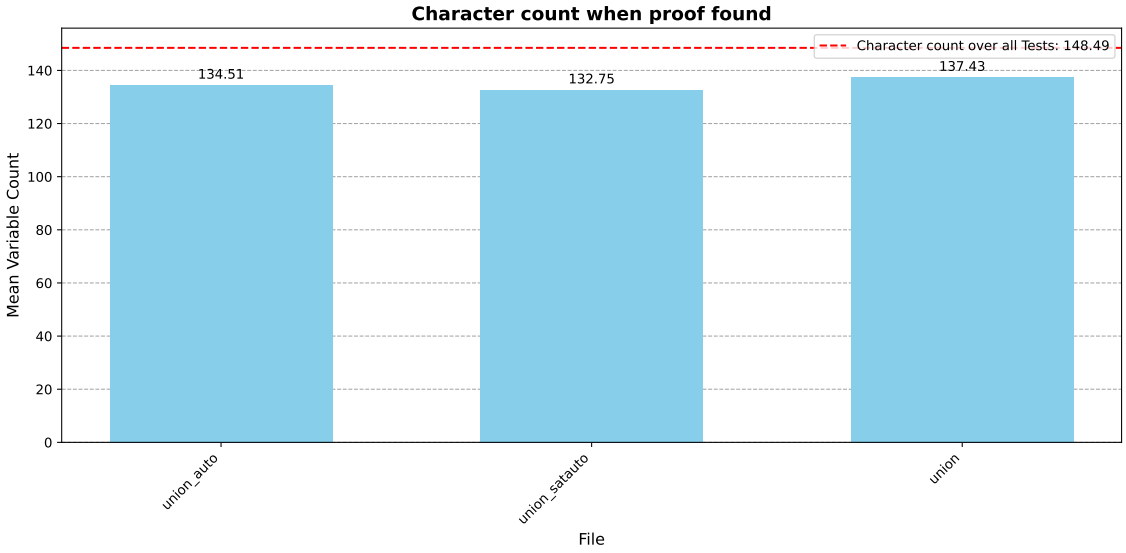

Figure 3.4: Character count

– Cosine similarity Why Would Provable Conjectures Have Lower Axiom Similarity?

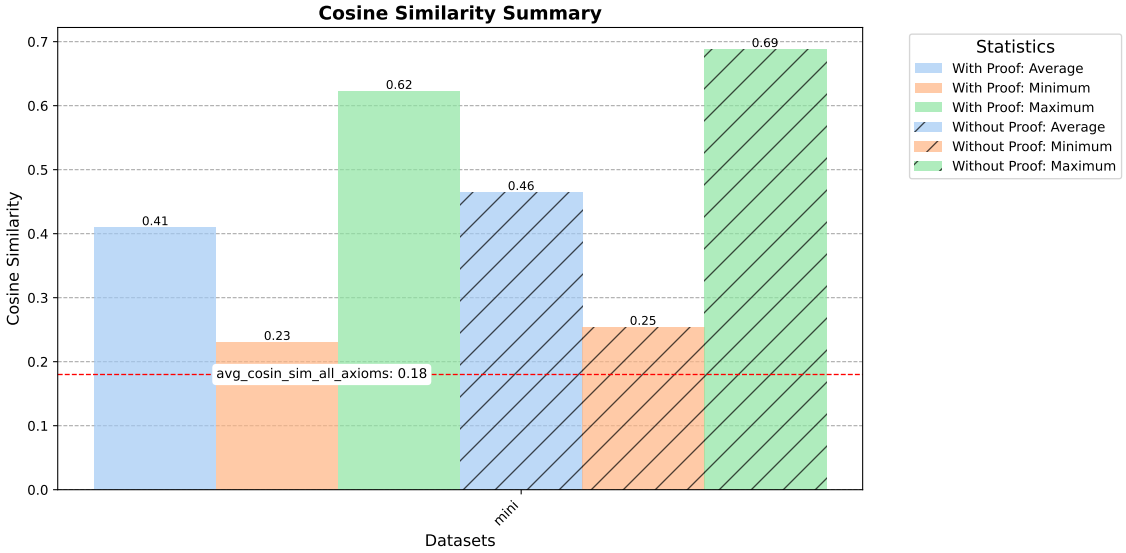

Figure 3.5: Cosine similarity

This suggests that proving the conjecture involves bridging different concepts, rather than just refining a single closely related idea.

3.4 Add core axioms

– Results of prover

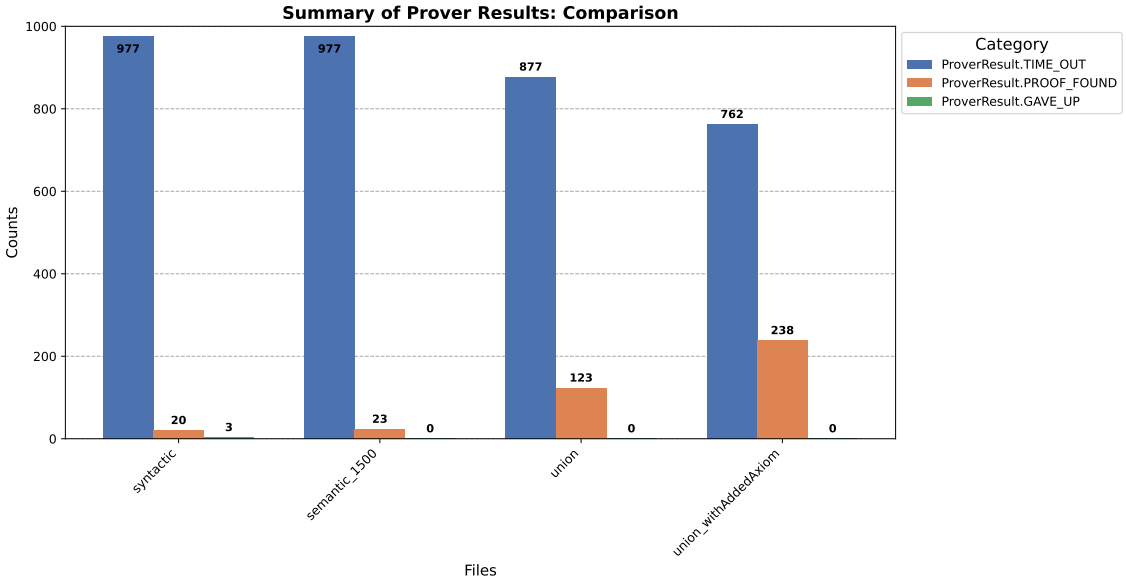

Figure 3.6: Summary of Prover Results with core axioms

– Cosine similarity

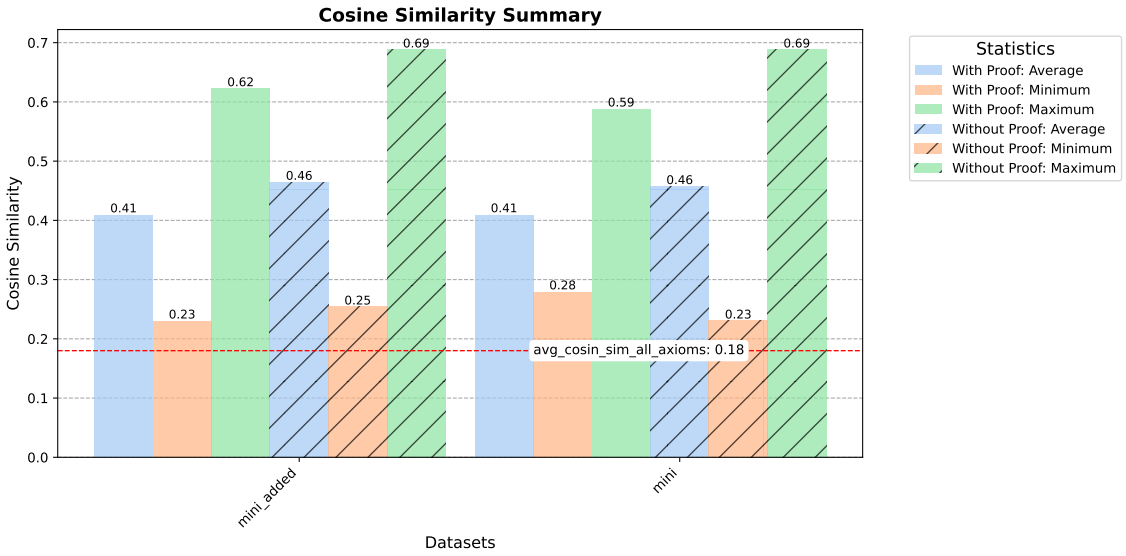

Figure 3.7: Summary of Prover Results: Standard Mode

3.5 Comparing LLMs

– Results of prover

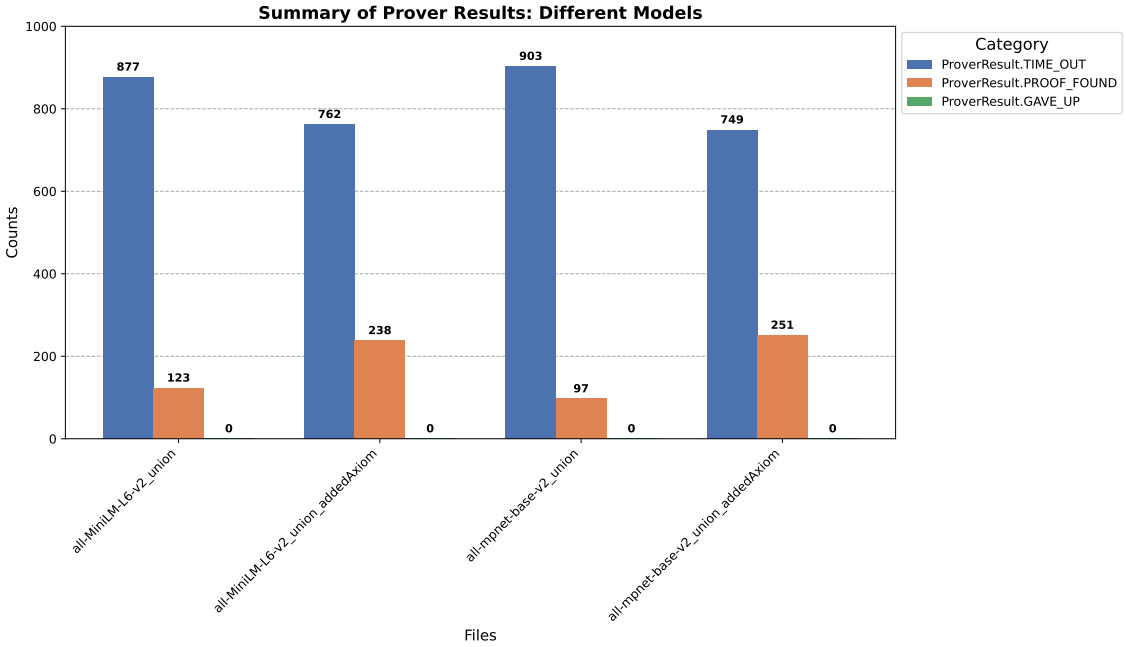

Figure 3.8: Summary of Prover Results: Different models

– Cosine similarity

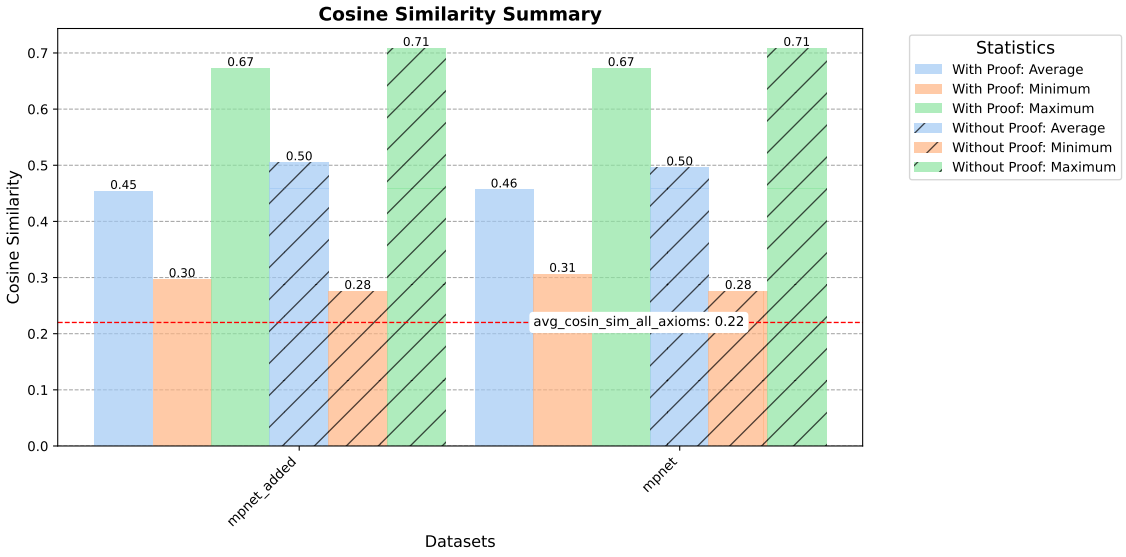

Figure 3.9: Cosine Similarity Mpnet

3.6 Analyze different E settings

– Satauto Using search heuristic

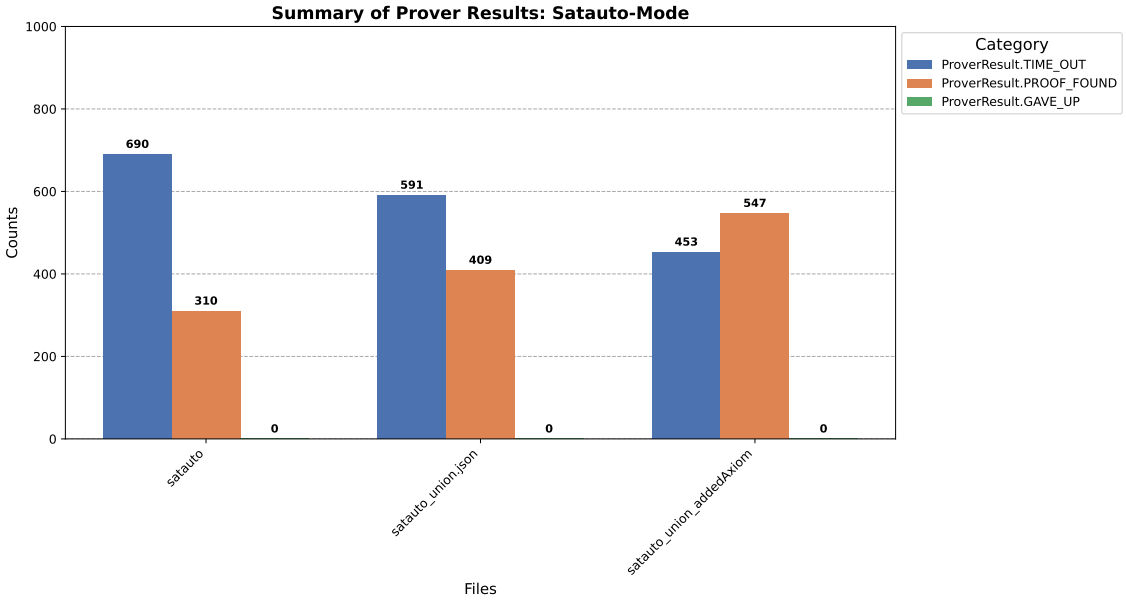

Figure 3.10: Summary of Prover Results: Satauto Mode

– Auto Using SInE + search heuristic

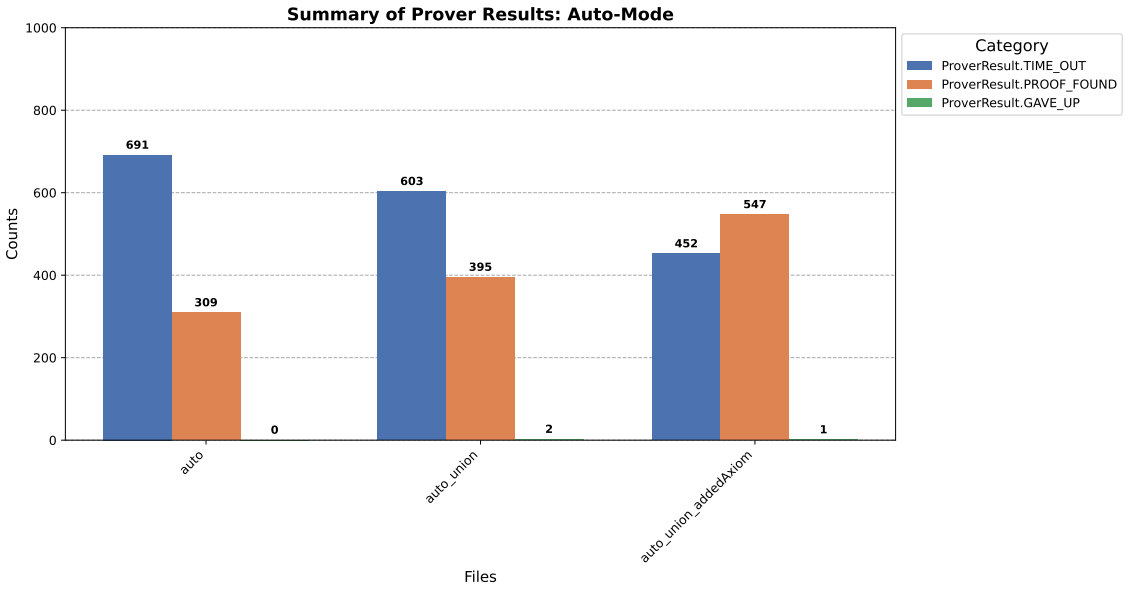

Figure 3.11: Summary of Prover Results: Auto Mode

3.7 Different prover

– Vampire

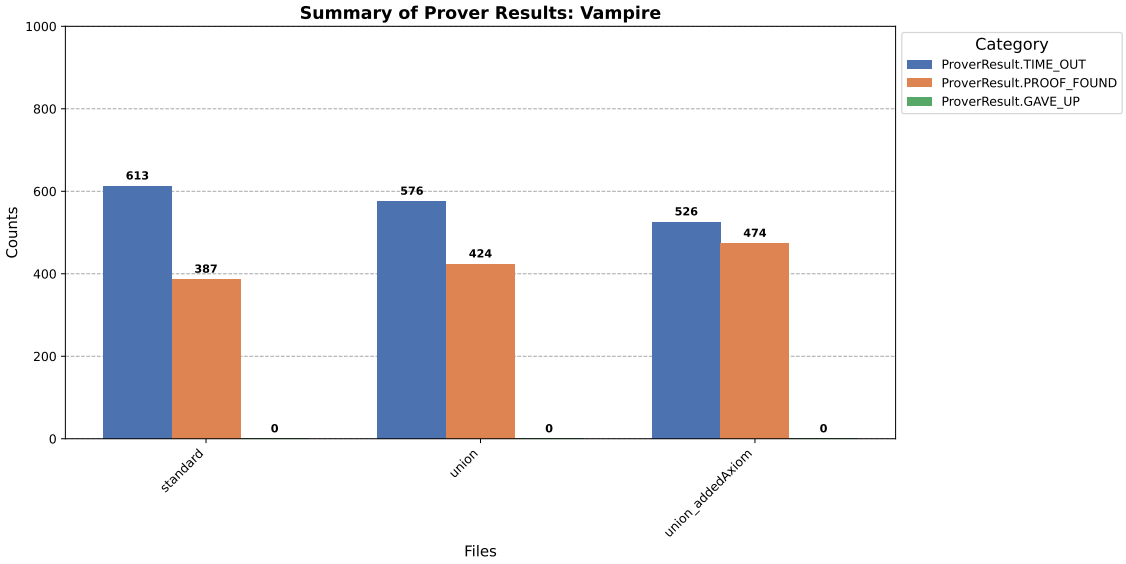

Figure 3.12: Summary of Prover Results: Vampire Mode

3.8 Time study

– Proofs found in first named

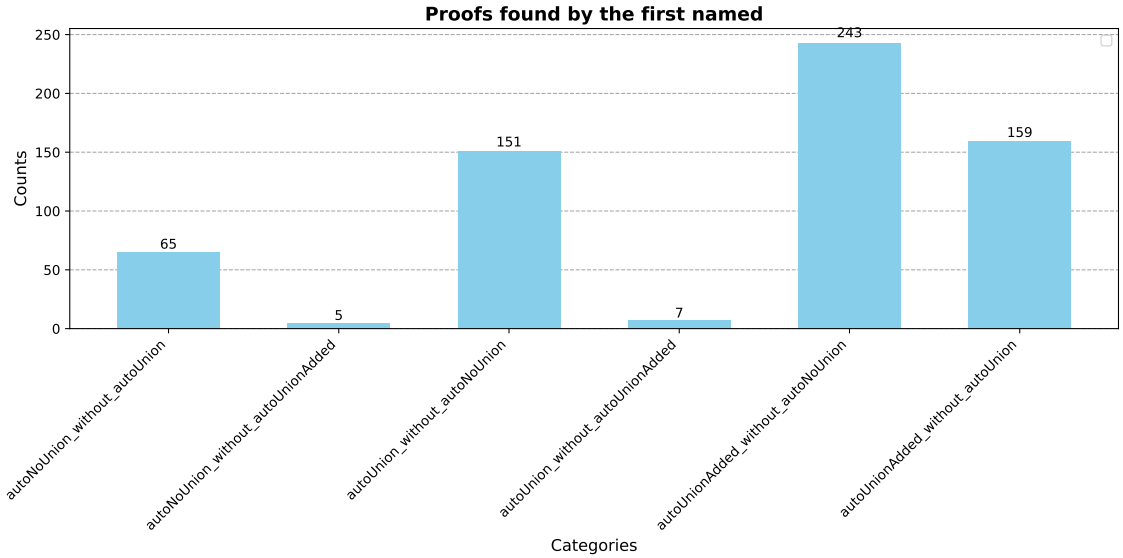

Figure 3.13: Proof found on other modes

– Time to find proof

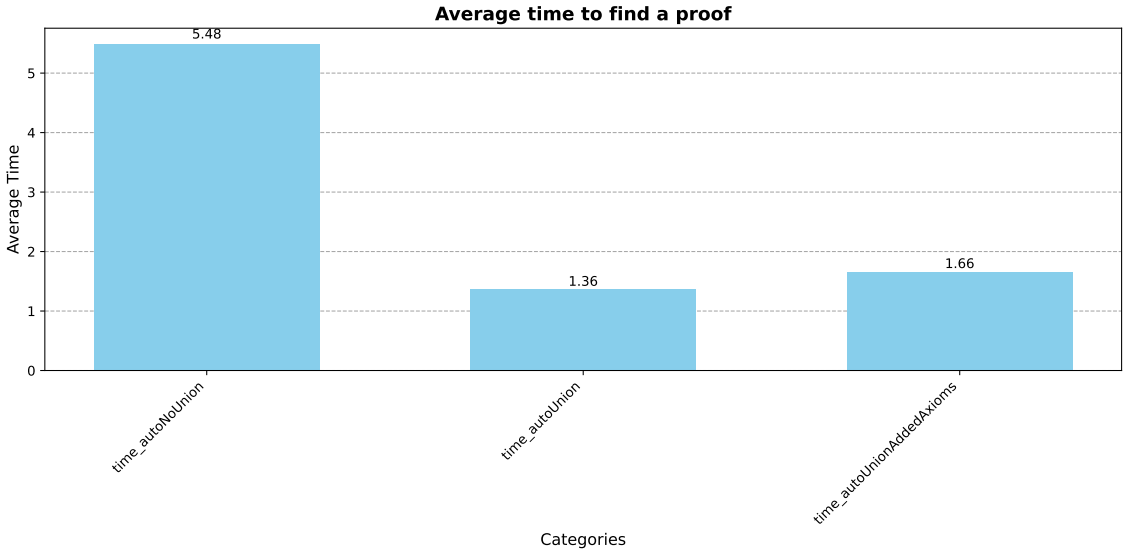

Figure 3.14: Proofs found time

– Time to find common proof

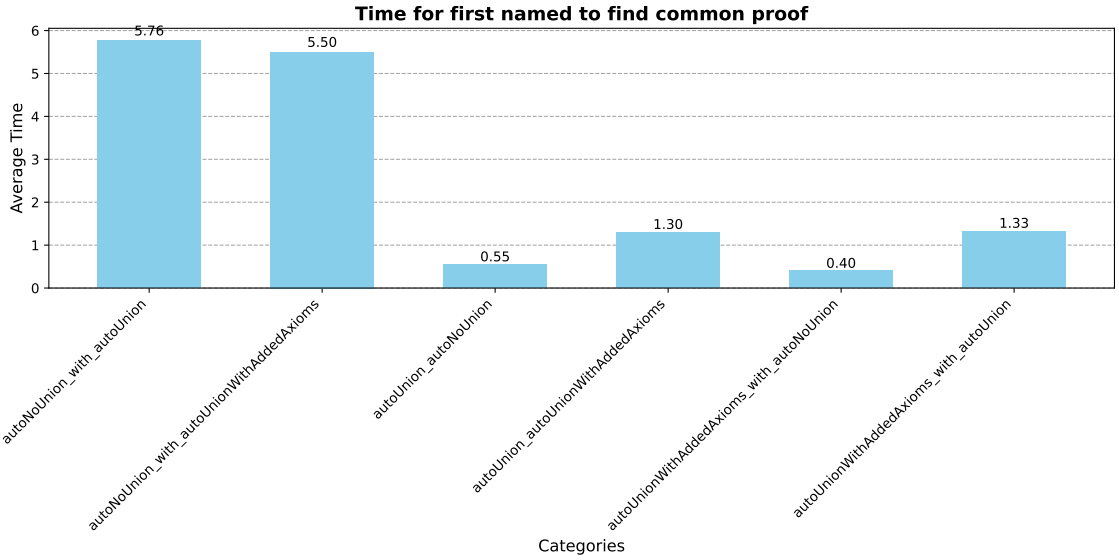

Figure 3.15: Common proofs found time

4

Conclusion

Conclusion.

5

Further work

Further work.

$$