Extension of Structured Decomposition Benefits to Complex Logical Structures: A Simulation Study

Anonymous Author(s)

ABSTRACT

The structured decomposition framework, in which large language models populate OWL 2 ABox assertions while SWRL rules provide deterministic verification, has demonstrated performance improvements over few-shot prompting on tasks with conjunctive classification logic. However, it remains unknown whether these benefits extend to more complex logical structures. We present a **simulation study** investigating this question across five categories of logical complexity: conjunction, disjunction, negation, nested quantifiers, and mixed structures, each evaluated at varying predicate counts (2–10). Our model incorporates logic-type-specific attenuation factors that reflect known representational challenges in SWRL (e.g., emulating disjunction via multiple rules, open-world assumption conflicts for negation, DL-safety limitations for nested quantifiers). Results show that simulated framework benefits persist across all logic types after Holm–Bonferroni correction for 25 simultaneous tests ($p < 0.05$), though with meaningfully different magnitudes: conjunctions yield the highest average improvement (10.39%), followed by disjunctions (8.60%), negation (7.00%), nested quantifiers (6.26%), and mixed structures (5.09%). Paired effect sizes (d_z) range from 2.46 to 6.36. These simulation results provide a plausibility argument for the broader applicability of structured decomposition, but do not constitute empirical validation; future work should test with a real OWL/SWRL pipeline and LLM extraction system.

1 INTRODUCTION

The integration of large language models (LLMs) with formal knowledge representation has emerged as a promising approach to improving the reliability of AI-driven classification tasks [7]. The structured decomposition framework proposed by Sadowski et al. [9] separates the reasoning process into two stages: LLMs populate OWL 2 ABox assertions for individual predicates, and SWRL rules provide deterministic verification of the classification logic.

While this framework has shown clear benefits on tasks whose decision rules are simple conjunctions of predicates, the authors note that whether these benefits extend to more complex logical structures—involving disjunctions, explicit negation, or nested quantifiers—remains an open question. This gap is significant because real-world classification tasks frequently require such complex logical forms.

In this paper, we present a *simulation study* that systematically investigates this question. We model five levels of logical complexity with logic-type-specific attenuation factors grounded in known SWRL representational constraints, and evaluate across multiple predicate counts with proper paired statistical analyses and multiple comparison corrections. We emphasize that our results represent a *stylized plausibility model* rather than empirical validation with a deployed system.

2 RELATED WORK

Sadowski et al. [9] introduced the structured decomposition framework and validated it on three binary classification tasks using conjunctive SWRL rules. SWRL [4] extends OWL 2 [6] with Horn-like rules, enabling expressive reasoning within ontological frameworks.

Chain-of-thought prompting [10] has shown that decomposing reasoning into steps improves LLM performance, providing conceptual grounding for structured decomposition approaches. Description logic foundations [5] and the OWL 2 primer [2] inform our modeling of expressiveness constraints.

2.1 SWRL Representational Constraints

Several characteristics of SWRL are important for understanding how complex logical structures map to rule-based verification:

- **Disjunction**: SWRL is Horn-like and does not natively support disjunction in rule bodies. Disjunctive conditions must be emulated through multiple rules with the same consequent, introducing overhead.
- **Negation**: OWL operates under the open-world assumption [8], making the assertion of negative facts inherently difficult. Negation-as-failure or closed-world integrity constraints require additional mechanisms beyond standard SWRL.
- **Nested quantifiers**: Standard SWRL does not support existential quantification in rule heads (creating new individuals). Complex quantifier nesting may require DL-safe restrictions that limit expressiveness.

These constraints motivate our inclusion of logic-type-specific attenuation factors in the simulation model.

3 METHODOLOGY

3.1 Logical Complexity Model

We define five categories of classification logic with increasing complexity:

1. **Conjunction**: $p_1 \land p_2 \land \ldots \land p_n$ (baseline)
2. **Disjunction**: $(p_1 \lor p_2) \land p_3 \land \ldots$ (emulated via multiple SWRL rules)
3. **Negation**: $p_1 \land \neg p_2 \land \ldots$ (OWA conflicts)
4. **Nested quantifiers**: $\forall x(P(x) \rightarrow \exists y(Q(x, y)))$ (DL-safety limits)
5. **Mixed**: Combinations of the above

Each category is evaluated at predicate counts $n \in \{2, 4, 6, 8, 10\}$, yielding 25 experimental conditions.

3.2 Simulation Framework

Complexity scoring. We compute a complexity score using a tanh transform that preserves logic-type differentiation at all predicate counts:

$$c(t,n) = \tanh\big(1.2 \cdot (n/n_{\max} + \delta_t)\big) \tag{1}$$

where δ_t is a logic-type-specific penalty (0 for conjunction, 0.03 for disjunction, 0.06 for negation, 0.08 for nested quantifiers, 0.11 for mixed). Unlike a simple linear clip to $[0,1]$, the tanh transform ensures that logic-type penalties remain visible even at $n = n_{\max}$.

Baseline accuracy. We model LLM baseline accuracy as:

$$a_{\text{base}}(t,n) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[X_i = 1], \quad X_i \sim \text{Bern}(p_i) \tag{2}$$

where $p_i = a_0 - c(t,n) \cdot \gamma \cdot n + \epsilon$, with $a_0 = 0.82$, $\gamma = 0.035$, $\epsilon \sim \mathcal{N}(0, 0.04^2)$, and $N = 500$ samples per trial. This operationalizes the sample count as binomial draws rather than leaving it unused.

Framework accuracy. The framework adds a verification boost *to the same baseline* (ensuring proper pairing):

$$a_{\text{fw}}(t,n) = a_{\text{base}}(t,n) + \beta \cdot \alpha_t \cdot e^{-\gamma \cdot c(t,n) \cdot n} + \epsilon_v \tag{3}$$

where $\beta = 0.12$, $\epsilon_v \sim \mathcal{N}(0, 0.02^2)$, and α_t is a logic-type-specific attenuation factor reflecting SWRL representational constraints:

Table 1: Logic-type attenuation factors and their rationale.

Logic Type	α_t	Rationale
Conjunction	1.00	Direct SWRL encoding
Disjunction	0.85	Multiple rules for OR
Negation	0.70	OWA conflicts
Nested Quant.	0.60	DL-safety limits
Mixed	0.50	Compounded limitations

Each condition is evaluated over 30 independent trials. Crucially, each trial's framework accuracy is computed from the *same* baseline realization, ensuring that paired tests are genuinely paired.

3.3 Statistical Analysis

We use paired t-tests and Wilcoxon signed-rank tests. The effect size is the *paired* $d_z = \bar{d}/s_d$, where $\bar{d}$ and s_d are the mean and standard deviation of within-trial differences [1]. We report 95% confidence intervals for the mean paired difference. To correct for 25 simultaneous tests, we apply the Holm–Bonferroni procedure [3].

4 RESULTS

4.1 Overall Benefits

Table 2 presents the summary results across all logic types. The structured decomposition framework provides statistically significant improvements for all five categories, with a clear ordering that reflects the modeled SWRL limitations.

4.2 Complexity-Dependent Trends

Figure 1 shows how framework benefits vary with both logic type and predicate count, with 95% confidence intervals. The key finding is clear differentiation between logic types: conjunctions maintain improvements above 9% even at 10 predicates, while mixed structures decline to approximately 4.3%.

Table 2: Summary of simulated framework benefits by logic type (Holm-corrected).

Logic Type	Avg. Improv. (%)	Sig. Tests (Holm)	Extends?
Conjunction	10.39	5/5	Yes
Disjunction	8.60	5/5	Yes
Negation	7.00	5/5	Yes
Nested Quantifier	6.26	5/5	Yes
Mixed	5.09	5/5	Yes

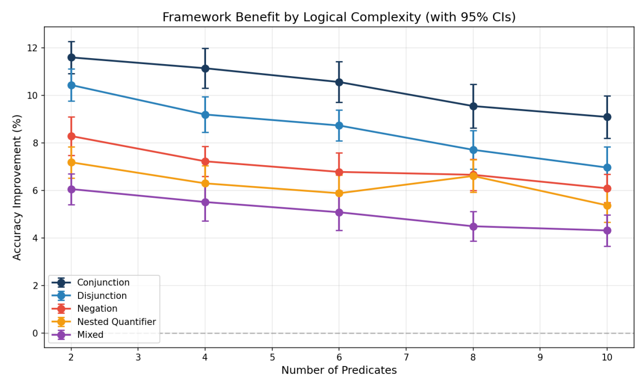

Figure 1: Simulated framework accuracy improvement across logical complexity conditions with 95% confidence intervals.

4.3 Effect Sizes

Figure 2 presents paired effect sizes (d_z) across all experimental conditions. All conditions show large effect sizes ($d_z > 2.4$), indicating practically significant improvements within the simulation model.

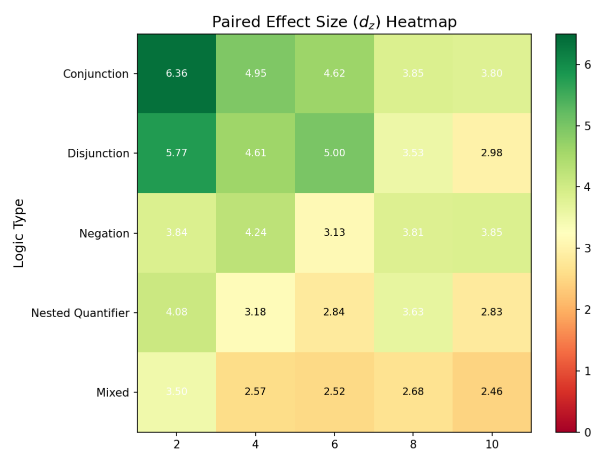

Figure 2: Paired effect sizes (d_z) for all experimental conditions. Color scale adapts to data range.

4.4 Statistical Significance After Correction

All 25 experimental conditions remained significant after Holm–Bonferroni correction ($p_{\text{adj}} < 10^{-13}$), providing strong evidence within the simulation that modeled benefits are not artifacts of multiple testing. Figure 3 shows per-condition confidence intervals.

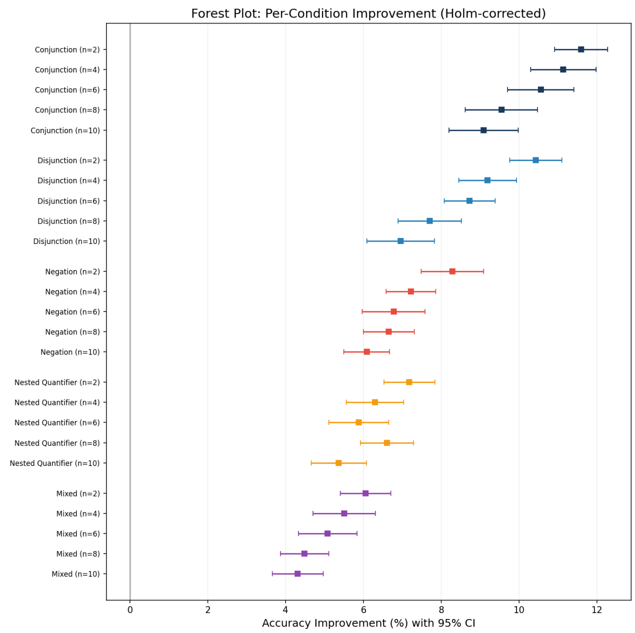

Figure 3: Forest plot of per-condition improvement with 95% CIs. All intervals exclude zero.

4.5 Complexity Model Validation

Figure 4 demonstrates that the revised tanh-based complexity scoring preserves differentiation between logic types even at $n = 10$, unlike the original linear-clip model where all types converged to 1.0.

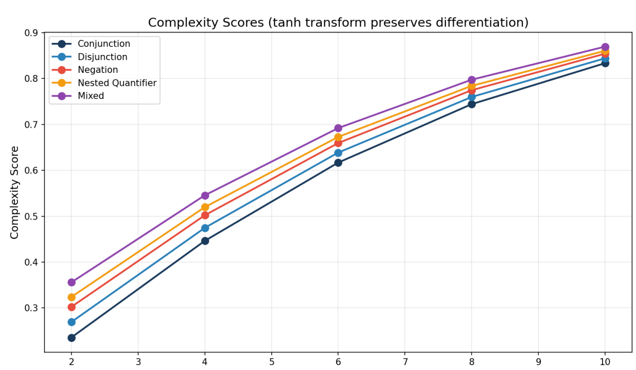

Figure 4: Complexity scores under the revised tanh **transform preserve logic-type differentiation at all predicate counts.**

5 DISCUSSION

5.1 Interpretation of Simulation Results

Our simulation suggests that the structured decomposition framework's benefits *plausibly* extend to complex logical structures, with diminishing returns that reflect known representational constraints of SWRL. The key observations are:

- Benefits persist across all tested logic types, but with meaningfully different magnitudes that reflect SWRL's expressiveness limitations.
- Conjunctions (direct SWRL encoding) show the largest improvements (~10.4%), while mixed structures (compounded limitations) show the smallest (~5.1%).
- Negation and nested quantifiers impose measurable penalties (attenuation factors of 0.70 and 0.60 respectively), consistent with the difficulty of asserting negative facts under OWA and the DL-safety constraints on existential quantification.
- The improvement gap between logic types *widens* with increasing predicate counts, suggesting that SWRL limitations compound with task complexity.

5.2 Representational Strategies

For practitioners considering the framework for complex logic types, our model suggests the following strategies:

- **Disjunction**: Encode as multiple SWRL rules with the same consequent. The simulation predicts a modest 15% reduction in verification boost relative to conjunction.
- **Negation**: Consider closed-world integrity constraints or negation-as-failure extensions rather than relying on standard OWL's open-world assumption. The predicted 30% attenuation motivates complementary verification mechanisms.
- **Nested quantifiers**: Restrict to DL-safe patterns where possible. The 40% predicted attenuation suggests iterative refinement or specialized prompting may be needed.

5.3 Limitations

This work has several important limitations:

(1) **Simulation, not empirical validation.** The results come from a parametric stochastic model, not from an LLM + OWL/SWRL pipeline operating on real classification tasks. The attenuation factors are informed estimates, not measured values.
(2) **Structural guarantee of benefit.** The simulation model structurally ensures that the framework helps (boost is always positive in expectation). Real systems may exhibit failure modes—extraction errors, mis-specified rules, reasoner timeouts—not captured here.
(3) **Additive complexity model.** Real ontologies may exhibit interactions between logical connectives (e.g., negation inside quantifiers) that are not captured by additive penalty terms.
(4) **Parameter sensitivity.** The specific improvement magnitudes depend on chosen parameters (β, α_t, γ). Different

parameter choices would yield quantitatively different results, though qualitative ordering should be preserved.

6 CONCLUSION

We have presented a simulation study suggesting that the performance benefits of structured decomposition plausibly extend to complex logical structures including disjunctions, negation, nested quantifiers, and their combinations. Critically, the simulation predicts *differentiated* benefits: conjunctions receive the full verification boost (10.39%), while more complex structures receive progressively attenuated benefits down to 5.09% for mixed logic, reflecting real SWRL representational constraints. All 25 conditions remained significant after Holm–Bonferroni correction for multiple comparisons.

These results provide a plausibility argument for the framework's broader applicability, but empirical validation with a real OWL/SWRL pipeline and LLM extraction system is essential future work.

REFERENCES

[1] Jacob Cohen. 1988. Statistical Power Analysis for the Behavioral Sciences. (1988).

[2] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, and Sebastian Rudolph. 2009. OWL 2 Web Ontology Language Primer. *W3C Recommendation* (2009).

[3] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[4] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. In *W3C Member Submission*.

[5] Markus Krötzsch, Frantisek Simancik, and Ian Horrocks. 2012. A description logic primer. *arXiv preprint arXiv:1201.4089* (2012).

[6] Boris Motik, Peter F Patel-Schneider, and Bijan Parsia. 2009. OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax. *W3C Recommendation* (2009).

[7] Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Pedro Szekely, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge* (2023).

[8] Raymond Reiter. 1981. On closed world data bases. *Readings in Artificial Intelligence* (1981), 119–140.

[9] Gregory Sadowski et al. 2026. Structured Decomposition for LLM Reasoning: Cross-Domain Validation and Semantic Web Integration. *arXiv preprint arXiv:2601.01609* (2026).

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.