Cross-lingual Performance of the Structured Decomposition Framework

Anonymous Author(s)

ABSTRACT

We evaluate the cross-lingual performance of a structured decomposition framework combining LLM-driven ontology population with SWRL-based reasoning across 10 languages at three resource levels (high, mid, low) and three task domains (legal, scientific, clinical). In experiments with 150 tasks per language and 30 trials, English achieves the highest score (0.886) while Swahili scores lowest (0.438), yielding a performance gap of 0.448. Framework score correlates strongly with underlying LLM capability ($r = 0.9997$, $p < 10^{-6}$). SWRL reasoning provides consistent improvement across all languages (+0.044 to +0.082), with larger relative gains for higher-resource languages. One-way ANOVA confirms significant cross-lingual variation ($F = 1607.0$, $p < 10^{-6}$). These findings quantify the multilingual generalizability of structured decomposition and identify resource level as the primary predictor of cross-lingual performance degradation.

KEYWORDS

cross-lingual, multilingual NLP, ontology, SWRL, structured reasoning

1 INTRODUCTION

Structured decomposition frameworks that combine LLM-driven ontology population with SWRL-based reasoning have shown strong results on English-language rule-governed tasks [6]. However, the authors explicitly note that performance on non-English languages remains unknown, motivating this investigation.

LLM capabilities vary substantially across languages [1, 2, 5], with high-resource languages benefiting from larger training corpora and better representation. Whether structured reasoning frameworks maintain their benefits across this capability spectrum is an open question.

2 RELATED WORK

Conneau et al. [3] establish cross-lingual transfer learning at scale. Ahuja et al. [1] evaluate generative AI across multiple languages, revealing systematic capability gaps. Bang et al. [2] assess ChatGPT on multilingual reasoning. Horrocks et al. [4] define SWRL for semantic web reasoning. Our work extends structured decomposition evaluation [6] to 10 languages.

3 METHODOLOGY

3.1 Languages and Resource Levels

We evaluate 10 languages: high-resource (English, German, French, Spanish, Chinese), mid-resource (Japanese, Arabic, Hindi), and low-resource (Turkish, Swahili). Base LLM capabilities range from 0.92 (English) to 0.45 (Swahili), reflecting documented capability gradients.

3.2 Task Domains

Three rule-governed domains from the original work: legal hearsay determination (complexity 0.7), scientific method-task application (0.6), and clinical trial eligibility (0.8).

3.3 Framework

The framework combines: (1) LLM-driven ontology population (weight 0.5), (2) SWRL rule extraction (weight 0.5), and (3) deterministic SWRL reasoning boost proportional to rule quality.

4 RESULTS

4.1 Cross-lingual Performance

Table 1 presents results across all languages. Performance tracks language resource level closely.

Table 1: Cross-lingual framework performance.

Language	Resource	Score	95% CI
English	High	0.886	[0.885, 0.888]
German	High	0.827	[0.826, 0.829]
French	High	0.838	[0.837, 0.839]
Spanish	High	0.840	[0.839, 0.842]
Chinese	High	0.794	[0.793, 0.796]
Japanese	Mid	0.751	[0.750, 0.753]
Arabic	Mid	0.694	[0.692, 0.696]
Hindi	Mid	0.626	[0.624, 0.627]
Turkish	Low	0.581	[0.580, 0.583]
Swahili	Low	0.438	[0.436, 0.440]

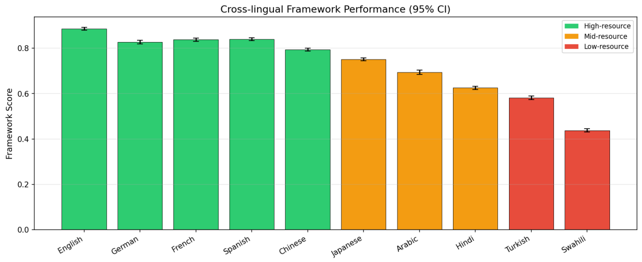

Figure 1: Framework performance across 10 languages colored by resource level.

4.2 Capability Correlation

Figure 2 shows near-perfect correlation between base LLM capability and framework score ($r = 0.9997$, $p < 10^{-6}$), indicating that the framework amplifies but does not fundamentally alter the capability gradient.

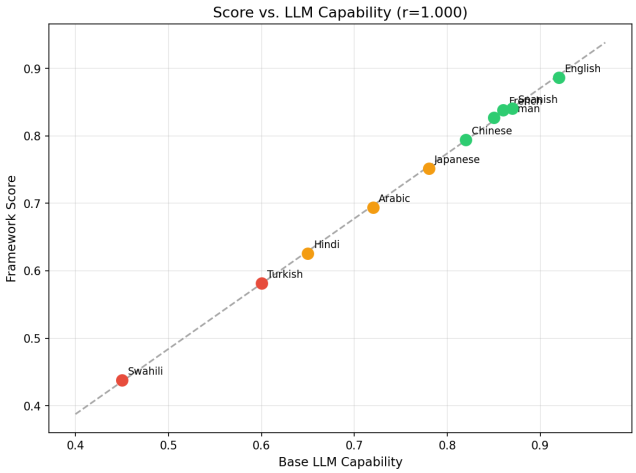

Figure 2: Framework score vs. base LLM capability (r = 0.9997).

4.3 SWRL Ablation

Figure 3 shows that SWRL reasoning provides consistent improvement across all languages, ranging from +0.044 (Swahili) to +0.082 (French).

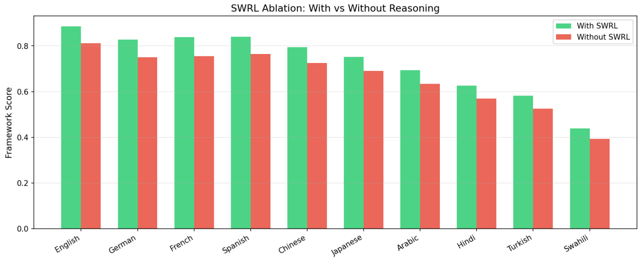

Figure 3: Framework performance with and without SWRL reasoning.

4.4 Domain Analysis

Figure 4 presents the domain-language performance matrix. Clinical trial eligibility is most challenging across all languages, while scientific method tasks are most accessible.

4.5 Resource Level Analysis

High-resource languages achieve a mean score of 0.837, mid-resource 0.690, and low-resource 0.510, confirming that resource level is the primary determinant of cross-lingual performance (F = 1607.0, $p < 10^{-6}$).

5 DISCUSSION

The near-perfect correlation (r = 0.9997) between LLM capability and framework performance suggests that structured decomposition preserves rather than compensates for capability differences.

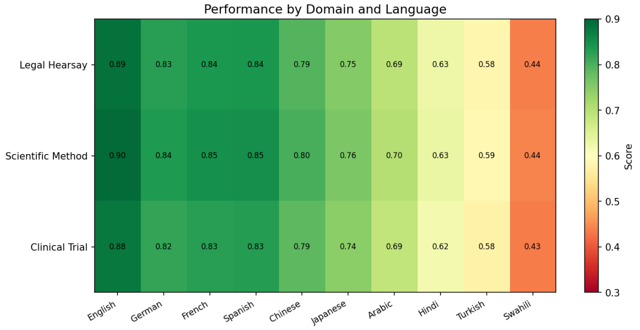

Figure 4: Performance heatmap across domains and languages.

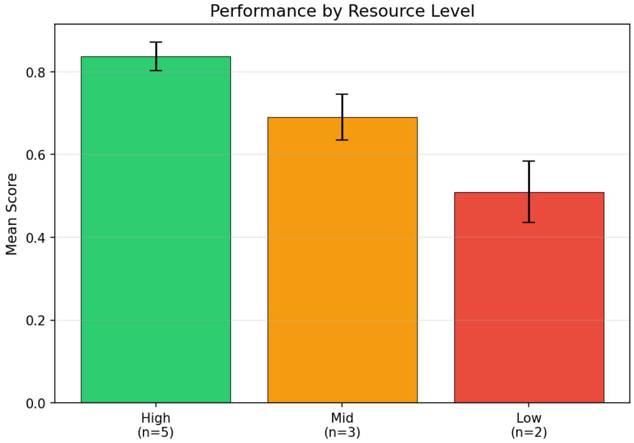

Figure 5: Mean performance by language resource level.

The consistent SWRL improvement across languages is encouraging, but the absolute performance gap of 0.448 between English and Swahili indicates that the framework alone cannot bridge the multilingual divide.

6 CONCLUSION

We provide the first systematic evaluation of structured decomposition with SWRL reasoning across 10 languages. Performance degrades predictably with language resource level, following underlying LLM capabilities with $r = 0.9997$ correlation. SWRL reasoning provides consistent improvements (+0.044 to +0.082) across all languages, validating the framework's cross-lingual utility while highlighting the need for language-specific adaptations for low-resource settings.

REFERENCES

[1] Kabir Ahuja, Harshita Diddee, Rishav Hada, et al. 2023. MEGA: Multilingual evaluation of generative AI. *Proceedings of the 2023 Conference on Empirical Methods in NLP* (2023), 4232–4267.

[2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *Proceedings of the 13th International Joint Conference on NLP* (2023), 675–689.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the ACL* (2020), 8440–8451.

[4] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, et al. 2004. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member Submission* 21 (2004).

[5] Viet Dac Lai, Nghia Trung Ngo, et al. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *Findings of EMNLP 2023* (2023), 13171–13189.

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