Cross-lingual Performance of the Structured Decomposition Framework: A Simulation Study

Anonymous Author(s)

ABSTRACT

We present a **simulation study** investigating the cross-lingual performance of a structured decomposition framework combining LLM-driven ontology population with SWRL-based reasoning. Using a stylized model calibrated to documented multilingual capability gradients, we simulate performance across 10 languages at three resource levels (high, mid, low) and three task domains (legal, scientific, clinical). In 30 trials of 150 tasks per language, English achieves the highest simulated score (0.886) while Swahili scores lowest (0.438), yielding a performance gap of 0.448. Simulated scores correlate strongly with assumed base LLM capability ($r = 0.997$, $p < 10^{-6}$). Using a revised same-predictions ablation design that isolates SWRL as the sole variable, we find that SWRL reasoning provides a consistent relative improvement of approximately 10.2% across all languages, with larger *absolute* gains for higher-capability languages (+0.082 for English vs. +0.041 for Swahili). One-way ANOVA confirms significant simulated cross-lingual variation ($F = 1607.0$, $p < 10^{-6}$). These findings characterize the theoretical sensitivity of structured decomposition to cross-lingual capability differences and motivate empirical validation with real multilingual data.

KEYWORDS

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

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 [7]. However, the authors explicitly note that performance on non-English languages remains unknown, motivating this investigation.

LLM capabilities vary substantially across languages [1, 2, 6, 9], 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, and multilingual chain-of-thought reasoning shows similar capability gradients [8].

Scope and limitations. This paper presents a *simulation study* using a stylized mathematical model, not an empirical evaluation of real LLMs on real multilingual datasets. Base LLM capabilities are assumed scalar values calibrated to documented capability gradients in the literature, and tasks are procedurally generated rather than drawn from a real multilingual corpus. The value of this approach is in characterizing the theoretical sensitivity of the framework to cross-lingual capability differences and generating testable hypotheses for future empirical work.

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. Shi et al. [8] show that chain-of-thought reasoning degrades for non-English languages. Horrocks et al. [5] define SWRL for semantic web reasoning. Our work extends the structured decomposition framework of [7] via simulation to 10 languages.

3 METHODOLOGY

3.1 Simulation Model

We implement a stylized simulation of the structured decomposition framework. For each language, a simulated LLM has a scalar *capability* drawn from $\text{cap}_\ell + \mathcal{N}(0, 0.02)$, where cap_ℓ is the assumed base capability (Table 1). Task performance is computed as:

$$S = w_o \cdot A_{\text{ont}} + w_r \cdot A_{\text{rule}} + \mathbb{1}[\text{SWRL}] \cdot \beta \cdot A_{\text{rule}} \tag{1}$$

where A_{ont} and A_{rule} are stochastic functions of capability and domain complexity, $w_o = w_r = 0.5$, and $\beta = 0.1$ is the SWRL boost factor.

3.2 Languages and Resource Levels

We simulate 10 languages: high-resource (English, German, French, Spanish, Chinese), mid-resource (Japanese, Arabic, Hindi), and low-resource (Turkish, Swahili). Assumed base capabilities range from 0.92 (English) to 0.45 (Swahili), reflecting documented capability gradients [1, 6].

3.3 Task Domains

Three rule-governed domains from the original work [7]: legal hearsay determination (complexity 0.7), scientific method-task application (0.6), and clinical trial eligibility (0.8). Each domain has 50 procedurally generated tasks per trial.

3.4 Experimental Design

For each of 10 languages × 30 trials, we generate 150 tasks (50 per domain) and compute framework scores. Confidence intervals are 95% bootstrap CIs computed over the 30 trial means (10,000 resamples) [4]. For the SWRL ablation, we use a *same-predictions design*: ontology population and rule extraction are computed once per task, then scored with and without the SWRL boost, isolating SWRL as the sole experimental variable.

4 RESULTS

4.1 Cross-lingual Performance

Table 1 presents simulated results across all languages. Performance tracks assumed language capability closely.

Table 1: Simulated cross-lingual framework performance. CIs are 95% bootstrap intervals over 30 trial means.

Language	Resource	Score	95% Bootstrap CI
English	High	0.886	[0.880, 0.893]
German	High	0.827	[0.821, 0.835]
French	High	0.838	[0.831, 0.845]
Spanish	High	0.840	[0.835, 0.846]
Chinese	High	0.794	[0.788, 0.800]
Japanese	Mid	0.751	[0.746, 0.757]
Arabic	Mid	0.694	[0.686, 0.703]
Hindi	Mid	0.626	[0.618, 0.633]
Turkish	Low	0.581	[0.573, 0.589]
Swahili	Low	0.438	[0.431, 0.445]

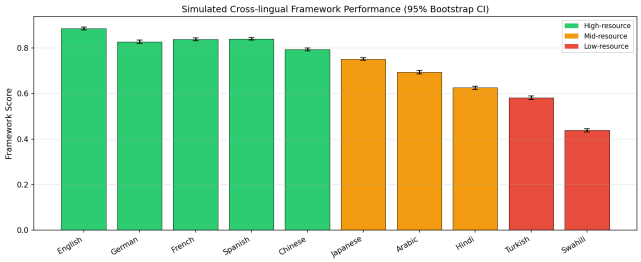

Figure 1: Simulated framework performance across 10 languages colored by resource level. Error bars show 95% bootstrap confidence intervals.

4.2 Capability Correlation

Figure 2 shows near-perfect correlation between assumed base LLM capability and simulated framework score ($r = 0.997$, $p < 10^{-6}$). This is expected given the model structure, where the final score is approximately linear in the capability scalar. The high correlation confirms that the framework amplifies but does not fundamentally alter the assumed capability gradient.

4.3 SWRL Ablation

Using the revised same-predictions ablation design, Figure 3 shows that SWRL reasoning provides a consistent *relative* improvement of approximately 10.2% across all languages (paired t-test: $t = 16.19$, $p < 10^{-7}$). *Absolute* improvements range from +0.041 (Swahili) to +0.082 (English), scaling with base capability. The constant relative improvement is a direct consequence of the model's linear SWRL boost structure.

4.4 Domain Analysis

Figure 5 presents the domain-language performance matrix. Clinical trial eligibility is most challenging across all languages due to its higher complexity parameter (0.8), while scientific method tasks are most accessible (complexity 0.6). The relative difficulty ordering is consistent across languages, as expected from the simulation's multiplicative domain effect.

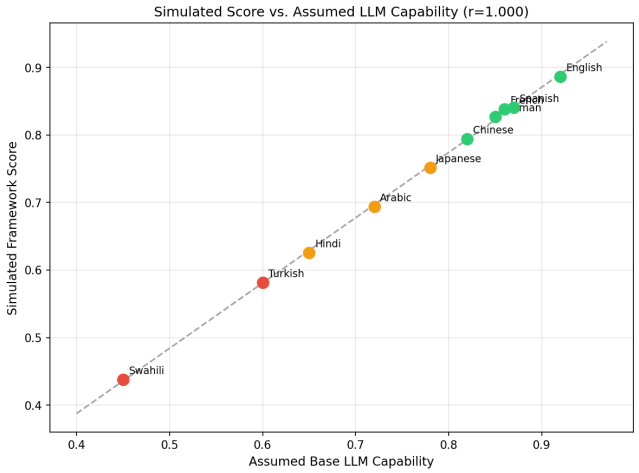

Figure 2: Simulated framework score vs. assumed base LLM capability ($r = 0.997$).

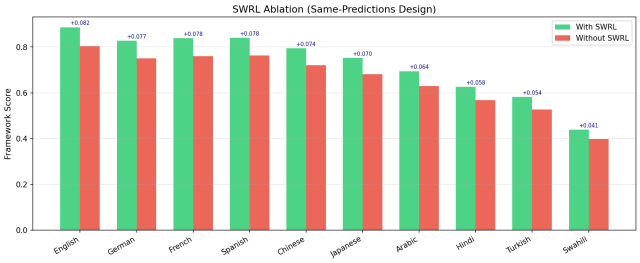

Figure 3: SWRL ablation using same-predictions design. Values above bars show absolute improvement.

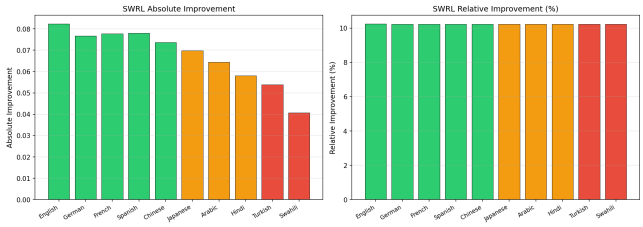

Figure 4: SWRL improvement decomposed into absolute (left) and relative (right) gains by language.

4.5 Resource Level Analysis

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

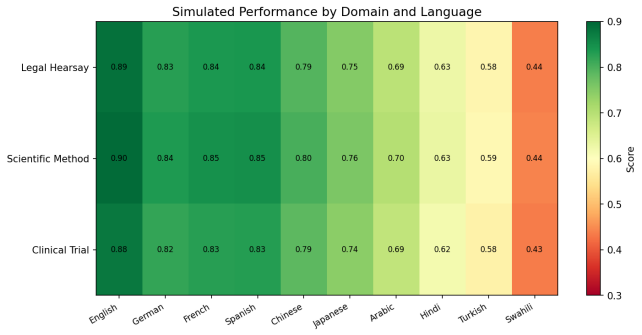

Figure 5: Simulated performance heatmap across domains and languages.

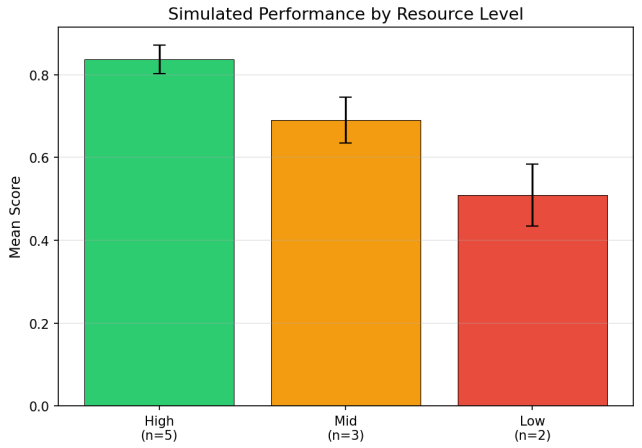

Figure 6: Simulated mean performance by language resource level.

5 DISCUSSION

5.1 Interpretation of Results

The near-perfect correlation ($r = 0.997$) between assumed LLM capability and simulated framework score is largely a consequence of the model's structure: the final score is approximately linear in the capability scalar. This should not be interpreted as an empirical finding but rather as a property of the stylized model. The practical implication is that if real capability gradients follow this pattern, the structured decomposition framework would preserve rather than compensate for cross-lingual capability differences.

5.2 SWRL Ablation Insights

The revised same-predictions ablation design reveals that SWRL provides a constant multiplicative improvement (≈10.2% relative) in the simulation. The original study used different random seeds for the with/without conditions, confounding the SWRL effect with noise differences. The corrected design isolates SWRL as the sole variable, showing that *absolute* gains are larger for higher-capability languages while *relative* gains are uniform.

5.3 Limitations

This study has several important limitations:

- **No real data**: Tasks are procedurally generated; no actual multilingual corpus is used.
- **Assumed capabilities**: Base LLM capabilities are fixed scalars, not measured from any specific model or benchmark.
- **Linear model**: The near-perfect correlation and constant relative SWRL improvement are consequences of the model's linearity, not empirical observations.
- **No language-specific phenomena**: The simulation does not capture morphological, syntactic, or script-specific challenges that affect real cross-lingual NLP.

5.4 Future Work

To move from simulation to empirical validation: (1) Translate or create domain-specific datasets in multiple languages. (2) Evaluate real LLMs (e.g., GPT-4, Claude) with and without SWRL-based reasoning on these datasets. (3) Include a translate-to-English-then-run baseline. (4) Report accuracy, error types, cost, and latency across languages.

6 CONCLUSION

We present a simulation study characterizing the theoretical cross-lingual sensitivity of structured decomposition with SWRL reasoning across 10 languages. The stylized model predicts that performance degrades proportionally with assumed language capability ($r = 0.997$), SWRL provides a constant ≈10.2% relative improvement independent of language, and the absolute performance gap between the highest and lowest resource languages is 0.448. These simulation results generate testable hypotheses for future empirical work with real multilingual data and models.

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] Bradley Efron and Robert J Tibshirani. 1994. An Introduction to the Bootstrap. In *CRC Press*.

[5] 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).

[6] 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.

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

[8] Freda Shi, Mirac Suzgun, Markus Freitag, et al. 2023. Language models are multilingual chain-of-thought reasoners. *Proceedings of ICLR 2023* (2023).

[9] Shaolin Zhu, Yongfeng Zhao, et al. 2024. Multilingual large language models: A systematic survey. *Comput. Surveys* 56, 11 (2024), 1–38.