A Case Study of Cross-Lingual Zero-Shot Generalization for Classical Languages in LLMs

V.S.D.S. Mahesh Akavarapu[α], Hrishikesh Terdalkar[β], Pramit Bhattacharyya[γ], Shubhangi Agarwal[β], Vishakha Deulgaonkar[γ], Pralay Manna[γ], Chaitali Dangarikar[γ], Arnab Bhattacharya[γ]

[α]University of Tübingen, [β]University of Lyon 1, [γ]Indian Institute of Technology Kanpur

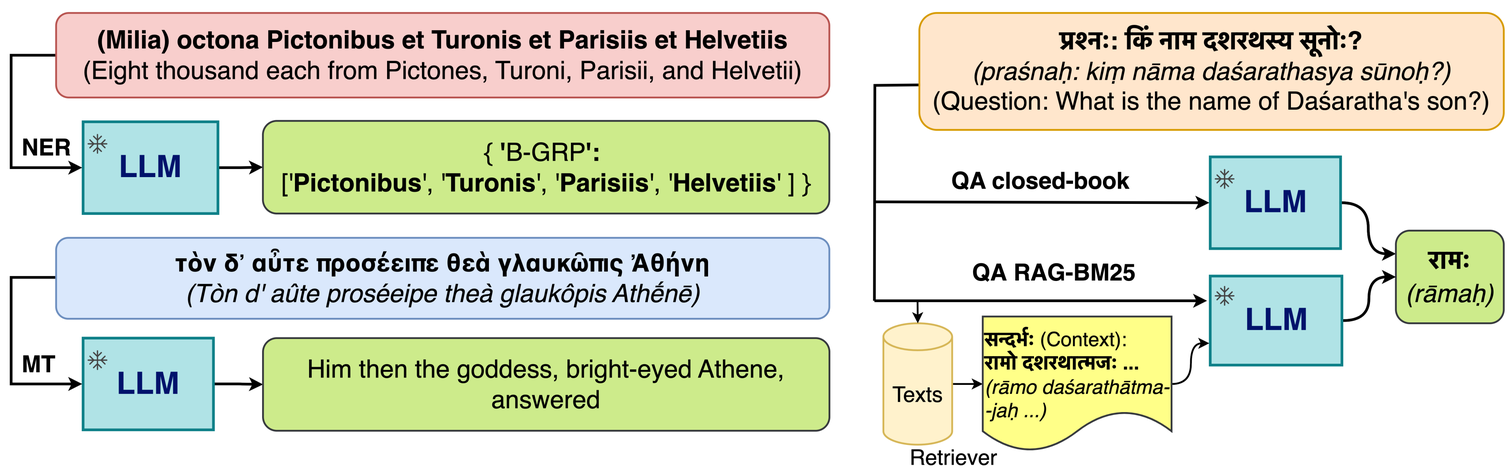
Motivation & Research Question

Classical Languages: Sanskrit, Ancient Greek, Latin

- A special case of low resourced languages
- Low-resource for NLU tasks
- Rich ancient literature available in digitized format
- High inflection present a challenge
- Influence high resourced languages — 28% of English vocabulary from Latin

Key Question: How well do LLMs generalize on Classical Languages, given that there is no evidence of instruction tuning on these languages?

Tasks



Experiments and Findings

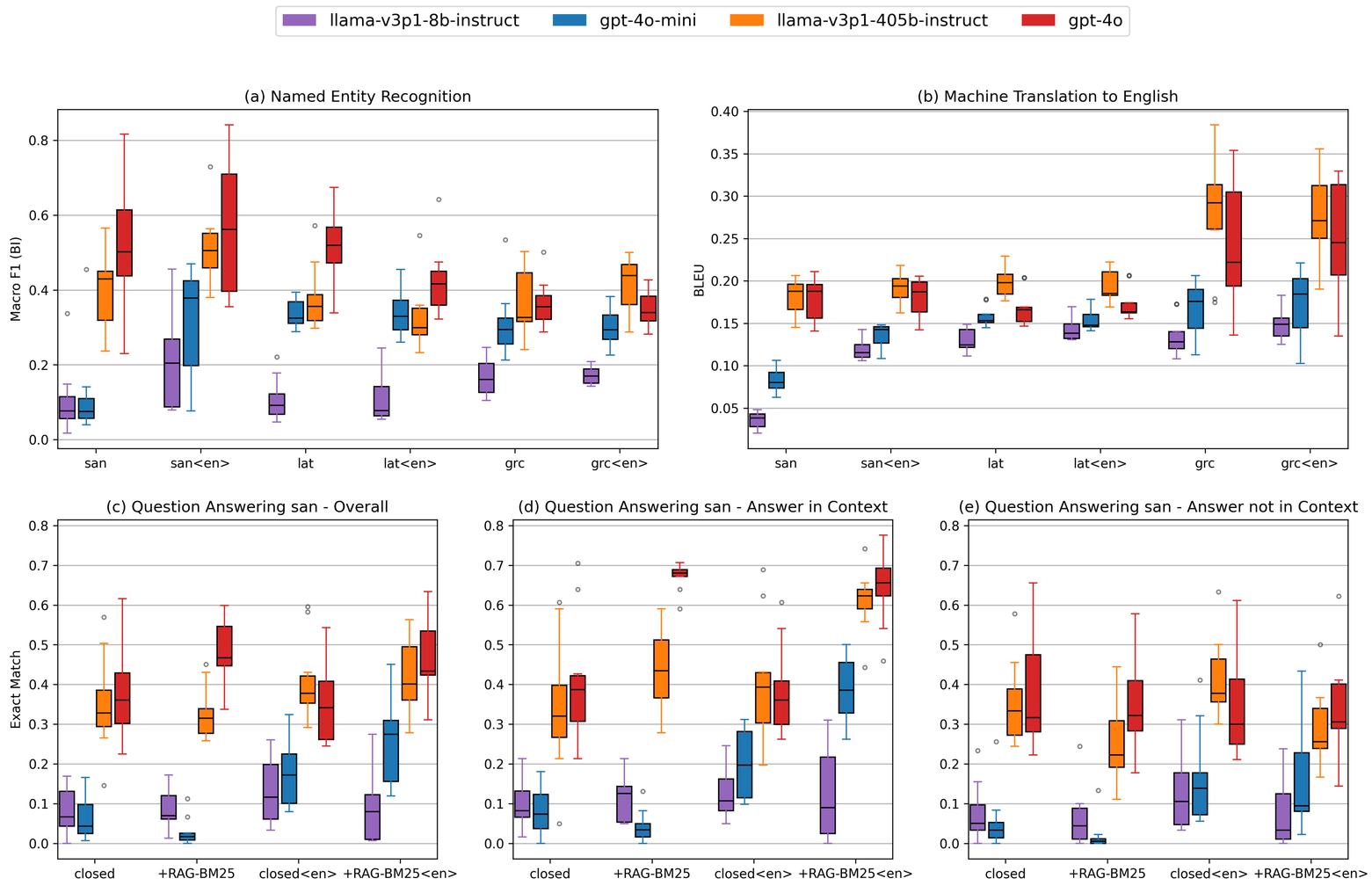
Datasets:

Task	Language	Test Size	Source
NER	Sanskrit	139	Terdalkar (2023)
	Latin	3,410	Erdmann et al. (2019)
	Ancient Greek	4,957	Myerston (2025)
MT	Sanskrit	6,464	Maheshwari et al. (2024)
	Latin	1,014	Rosenthal (2023)
	Ancient Greek	274	Palladino et al. (2023)
QA	Sanskrit	**1,501**	**This work**

Key Findings:

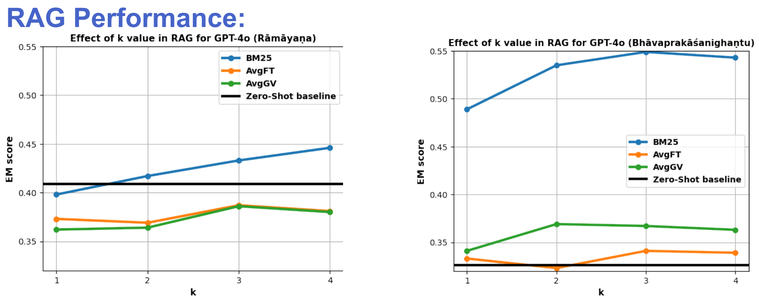
- **Larger models match/exceed fine-tuned baselines**
- **Significant performance gap** between large and small models
- **RAG significantly improves** QA performance
  - Smaller models fail to leverage context effectively
- **English prompts outperform** native language prompts
  - Especially true for smaller models
  - Evidence that models not instruction-tuned on classical languages
  - **Implication**: Performance due to cross-lingual generalization, not direct training.

Results



Sanskrit QA Insights

RAG Performance:



- BM25 retriever optimal with k=4
- Outperforms embedding-based methods — FastText and GloVe

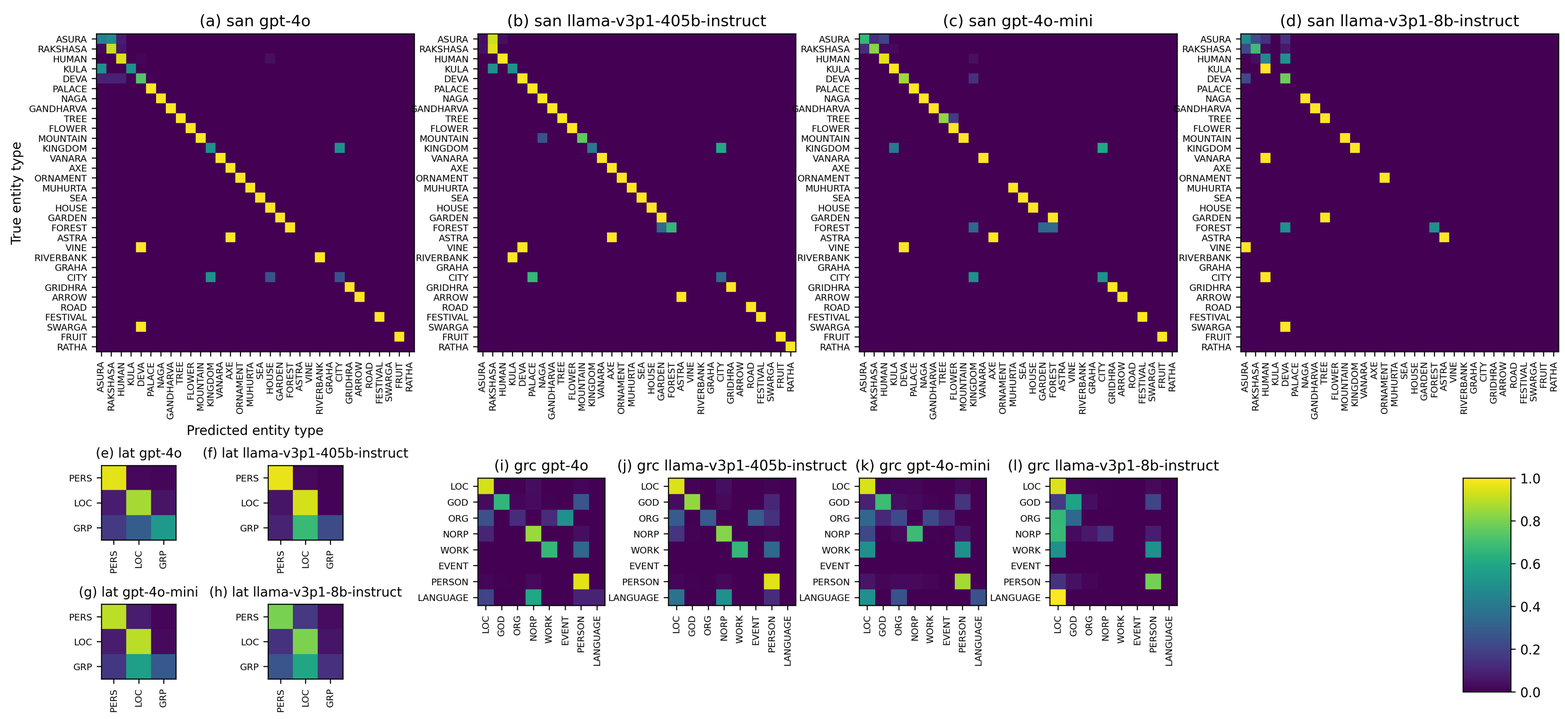
Inflection Handling:

- Models handle Sanskrit inflection well
- Minimal EM difference when lemmatized

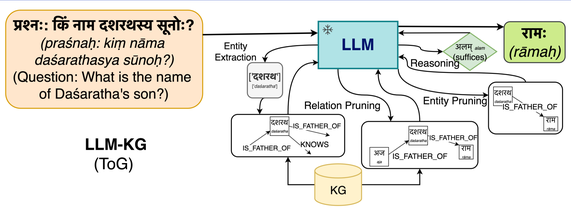
Orthographic Transfer:

- Slighlty better performance with Devanagari than Roman-based IAST
- Evidence of transfer from Hindi/Marathi

Entity Confusion in NER



LLM-KG Integration



Key Takeaways

- **Model scale crucial** for classical languages
- **Zero-shot competitive** with fine-tuned models
- **Retrieval helps** but needs capacity
- **Orthographic transfer** important
- **New Sanskrit QA dataset** (1,501 questions)

References

Terdalkar. 2023. *Sanskrit Knowledge-based Systems: Annotation and Computational Tools.*. PhD Thesis. IIT Kanpur

Erdmann et al. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. *NAACL*

Myerston. 2025. *NEReus: A named entity corpus of ancient greek*. `github.com/jmyerston/NEReus`.

Maheshwari et al. 2024. Samayik: A benchmark and dataset for English-Sanskrit translation. *LREC-COLING*

Rosenthal. 2023. *Machina cognoscens: Neural machine translation for latin, a case-marked free-order language*. Master's Thesis. Uni. of Chicago.

Palladino et al. 2023. Translation alignment for ancient greek: Annotation guidelines and gold standards. *Journal of Open Humanities Data*.

ACL 2025 VIENNA JULY 27 - AUGUST 1

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Lyon 1

mahesh-ak/SktQA