

Structured Clinical Data Extraction with Large Language Models (LLMs)

From Benchmarking to Knowledge Graphs

Karthik Prathaban^{1,2}, Douwe J. Spaanderman¹, Farhan Akram², Stefan Klein¹, Martijn P.A. Starmans^{1,2}

¹Department of Radiology and Nuclear Medicine, Erasmus MC

²Department of Pathology and Clinical Bioinformatics, Erasmus MC

Correspondance: k.prathaban@erasmusmc.nl

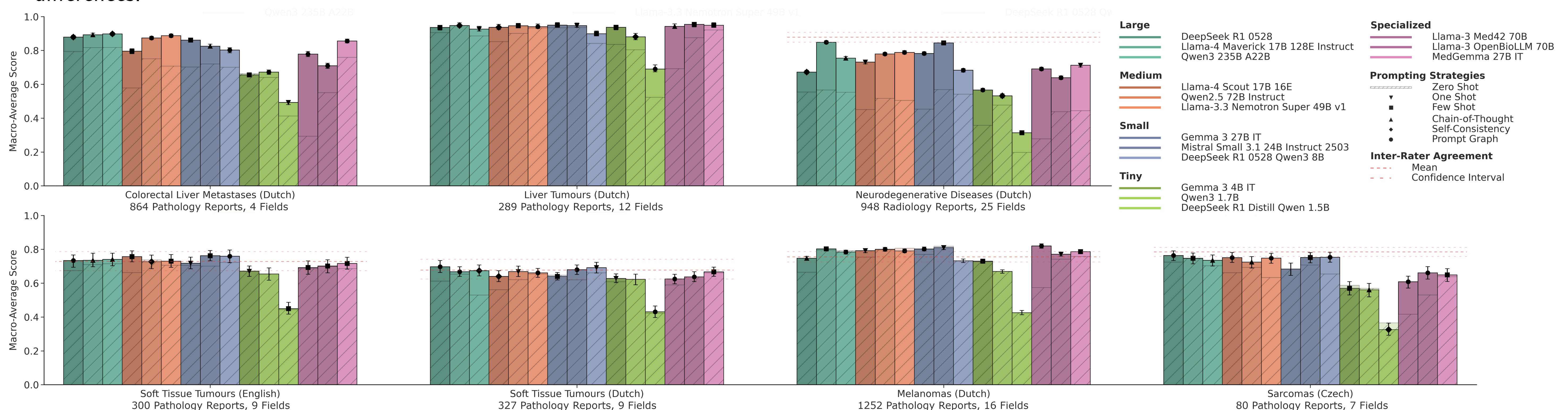
Motivation - Clinical pathology and radiology reports contain rich diagnostic information, but are typically stored as free text. Manual structuring is time-consuming and inconsistent across institutions and languages. Scalable and reproducible extraction methods are needed.

Methods - We benchmarked open-weight LLMs for structured clinical data extraction across diseases, languages, and report types using diverse prompting strategies. We additionally evaluated schema-based (JSON) and knowledge-graph extraction pipelines for accuracy, structure, and downstream usability.

Conclusion - Optimized prompting enables small-to-medium open-weight LLMs to achieve near-expert performance for structured clinical data extraction. Knowledge-graph outputs preserve relational structure with comparable accuracy, supporting scalable and flexible clinical data reuse.

Results: Systematic Evaluation of Open-Weight LLMs for Structured Clinical Data Extraction.

- ▶ Twelve general-purpose and three medical-specialized models were benchmarked across seven pathology and radiology use cases, three languages, and six prompting strategies, with performance compared to inter-rater agreement. Performance was summarized per use case using a macro-average of scores across extracted variables. Macro-average performance is shown, comparing zero-shot prompting with the best-performing strategy per model.
- ▶ Results show that optimized prompting allows small-to-medium models to achieve near-expert performance, with task complexity dominating performance differences.



From Free-Text to Structured Text to Knowledge Graphs

- ▶ In schema-based outputs such as JSON, entities may be encoded inconsistently as strings, lists, or nested dictionaries, complicating normalization and reuse.
- ▶ Knowledge graph triples provide a uniform representation of entities and relations, preserving structure with comparable accuracy and enabling flexible downstream clinical AI applications.

Report (Translated Excerpt):

MACROSCOPY:
Received in consultation: 1 block and 17 slides, numbered. External report received and scanned.

MICROSCOPY:
Lesional needle biopsy consisting of mature fat cells interspersed... without significant mitotic activity, and incipient necrosis.

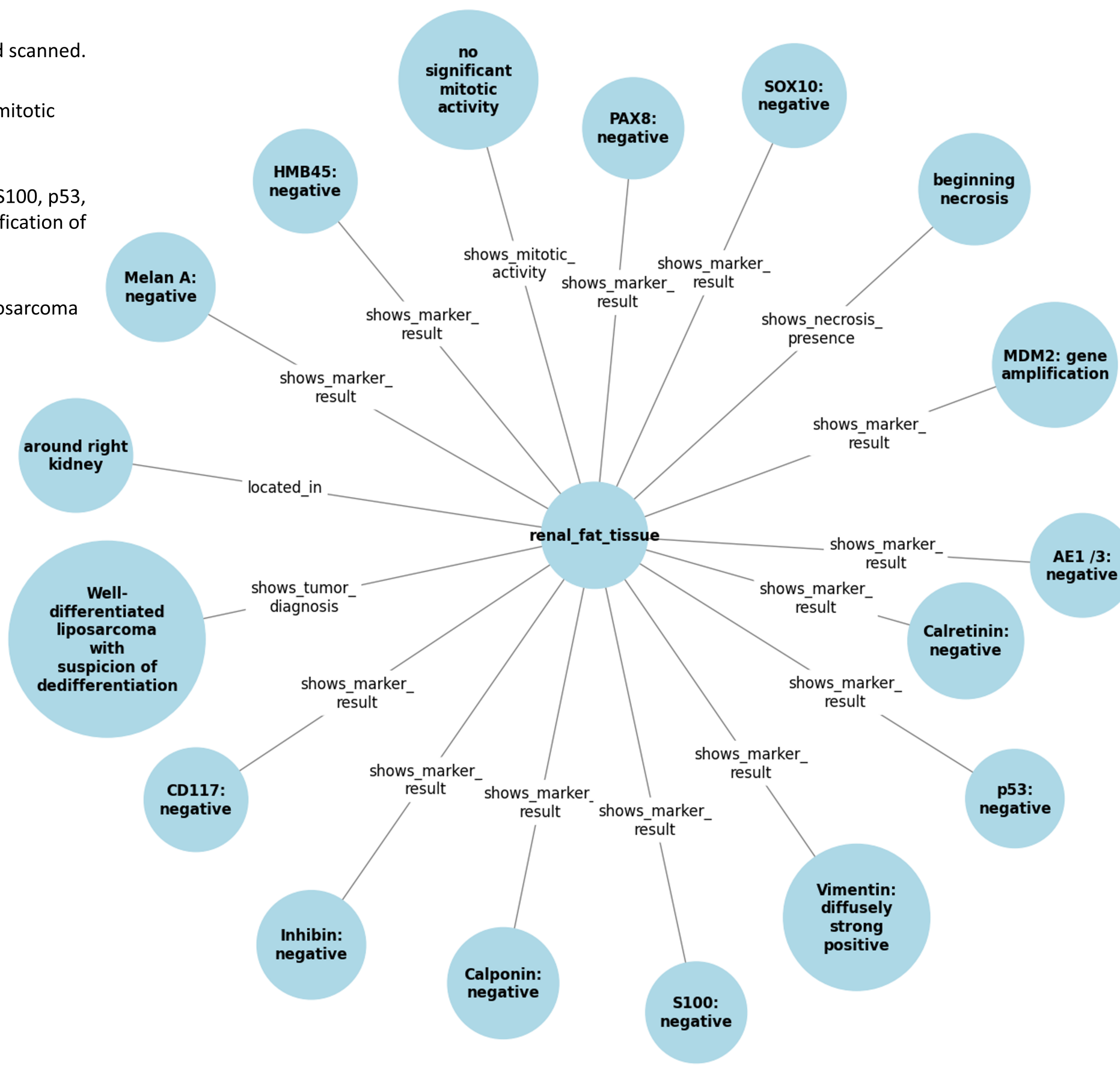
IMMUNOHISTOCHEMISTRY:
Vimentin: diffusely strong positive. AE1/3, SOX10, HMB45, Inhibin, Calponin, Melan A, S100, p53, PAX8, Calretinin and CD117: negative. According to an external report, ISH shows amplification of the MDM2 gene...

CONCLUSION:
... (soft tissue biopsy around the right kidney dated...) ... Cellular well-differentiated liposarcoma with suspicion of dedifferentiation...

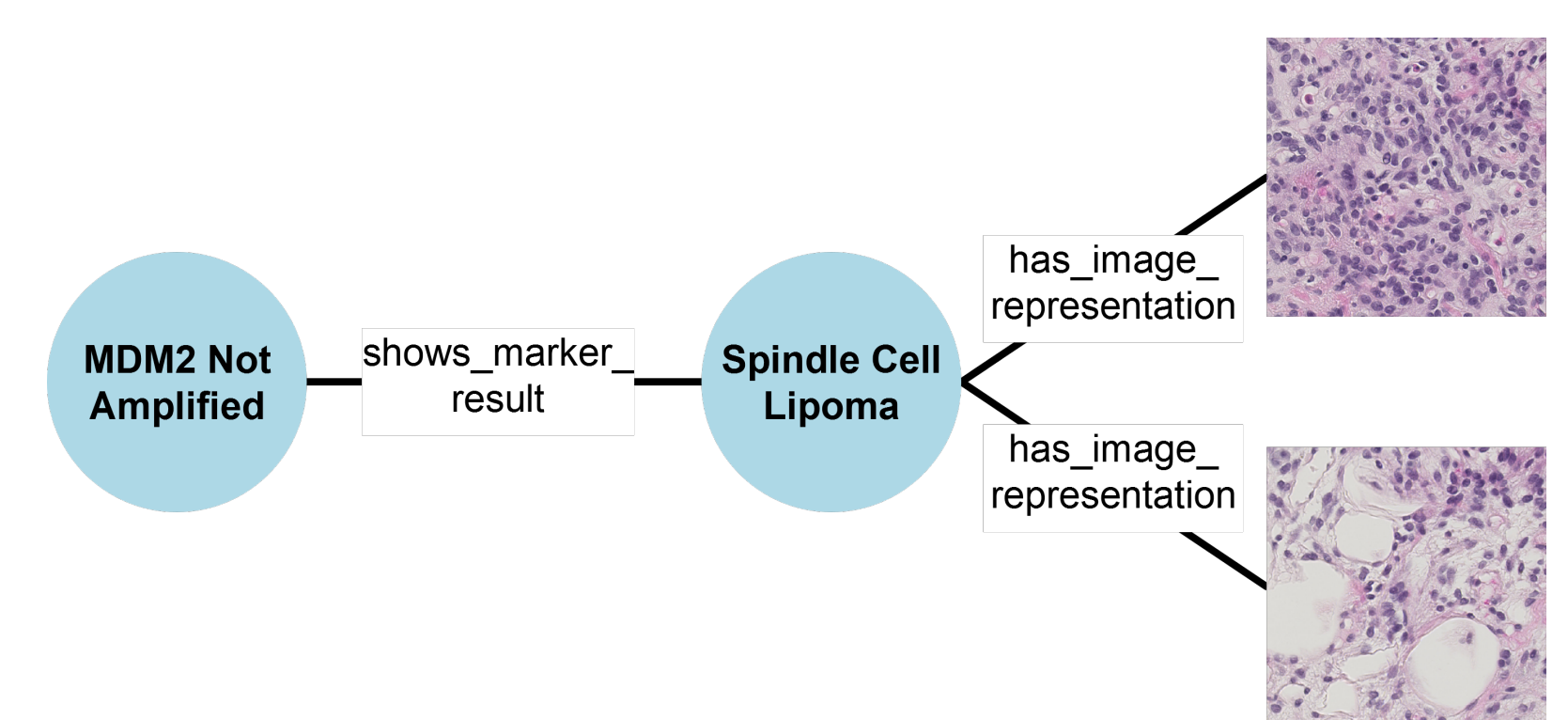
Extracted JSON:

```
{
  "organ_or_site": "around right kidney",
  "tumor_diagnosis": "Well-differentiated liposarcoma with suspicion of dedifferentiation",
  "mitotic_activity": "no significant mitotic activity",
  "necrosis_presence": "beginning necrosis",
  "marker_results": {
    "Vimentin": "diffusely strongly positive",
    "AE1/3": "negative",
    "SOX10": "negative",
    "HMB45": "negative",
    "Inhibin": "negative",
    "Calponin": "negative",
    "Melan A": "negative",
    "S100": "negative",
    "p53": "negative",
    "PAX8": "negative",
    "Calretinin": "negative",
    "CD117": "negative",
    "MDM2": "amplification"
  }
}
```

Extracted Graph:



Future Work



We aim to extend knowledge graphs derived from pathology reports with features extracted directly from whole-slide images. This will enable multimodal graph representations that jointly encode textual, spatial, and visual information for downstream diagnostic and prognostic tasks.

References

1. Spaanderman DJ, Prathaban K, et al. *Evaluating open-weight large language models for structured data extraction from narrative medical reports*. arXiv, 2025.
2. Prathaban K et al. *From reports to relations: large language models for knowledge graph extraction in digital pathology*. Proc. Int. Workshop on Emerging LLM/LMM Applications in Medical Imaging, 2025.