Unpacking Unstructured Data: Large Language Models used to extract insights from Neuropathological Reports of Parkinson's Patients

Oleg Stroganov[1], Amber Schedlbauer[1], Emily Lorenzen[1] Alex Kadhim[1], Anna Lobanova[1], NIH NeuroBioBank Network[2]

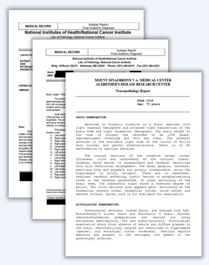
[1] Rancho BioSciences, LLC, 16955 Via Del Campo #220, San Diego, CA

[2] NIH NeuroBioBank Network

Email: Emily.Lorenzen@ranchobiosciences.com

Background and Objective

The NeuroBioBank (NBB) is a national resource for investigators conducting research on conditions of the nervous system using human post-mortem brain tissue and includes pathological reports. However, the **unstructured nature of neuropathological reports poses significant challenges** in extracting meaningful and consistent data. We investigated the **potential of Large Language Models (LLMs) in unpacking unstructured neuropathological reports**, with a focus on a subset of patients with Parkinson's Disease.



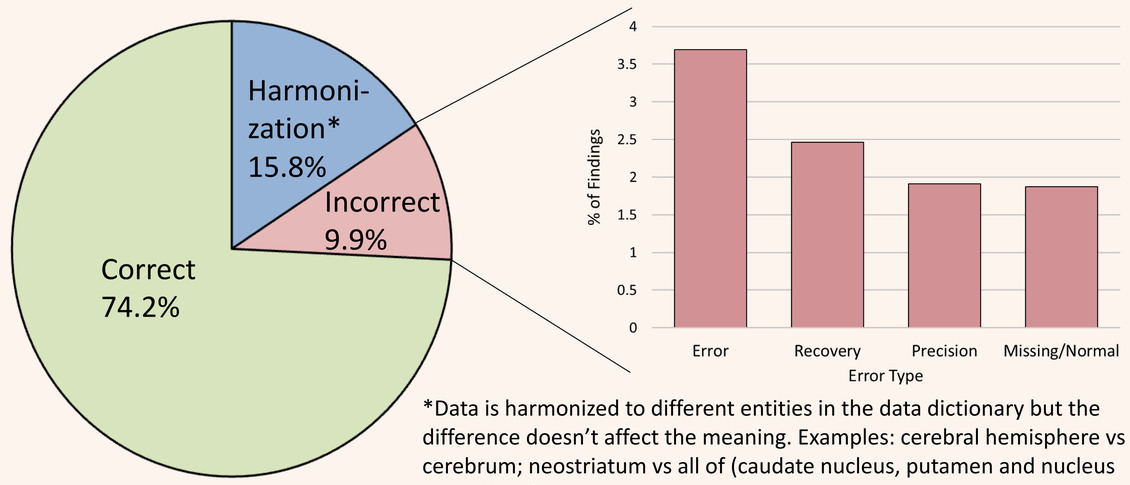
Unstructured data:
- 7 sites
- 15 report formats
- 816 PD reports for pilot

Stain/Bio	Location	Laterality	Sublocation	Cell Type	Pathology Finding	Descriptor
HE	Cerebellum			Granule cell	Thinning	Present
HE	Cerebellum		Dentate nucleus		Neuronal loss	Mild
HE	Cerebellum		Dentate nucleus		Non-classical Lewy bodies	Present
HE	Cerebellum		Dentate nucleus		Bright eosinophilic inclus	Present
HE	Vessels		Perivascular space		Expansion	Present
HE	Vessels		Perivascular space		Lymphocytic accumulatio	Moderate
HE	Neuropil				Rarefaction	Patchy
Biel	Hippocampus				Neurofibrillary tangles	Moderate
Biel	Hippocampus				Neuritic plaques	Moderate
Biel	Hippocampus				Pick's bodies	Not present
Biel	Entorhinal co				Neurofibrillary tangles	Moderate
Biel	Entorhinal co				Neuritic plaques	Moderate
Biel	Entorhinal co				Pick's bodies	Not present
Biel	Transentorhir				Neurofibrillary tangles	Moderate
Biel	Transentorhir				Neuritic plaques	Moderate
Biel	Transentorhir				Pick's bodies	Not present
AT8	Hippocampus				Neurofibrillary tangles	Moderate
AT8	Hippocampus				Pick's bodies	Not present
AT8	Hippocampus		Periventricle	Glial cell	Staining	Increase
AT8	Hippocampus		Subpial	Glial cell	Staining	Increase
AT8	Hippocampus		White matter	Astrocytes	Staining	Patchy
AT8	Hippocampus		Neuropil	Neurites	Staining	Diffuse
AT8	Entorhinal co				Neurofibrillary tangles	Moderate
AT8	Entorhinal co				Pick's bodies	Not present
AT8	Entorhinal co		Periventricle	Glial cell	Staining	Increase

Structured data:
- Neuropathological findings, diagnosis and staging
- FAIR* data

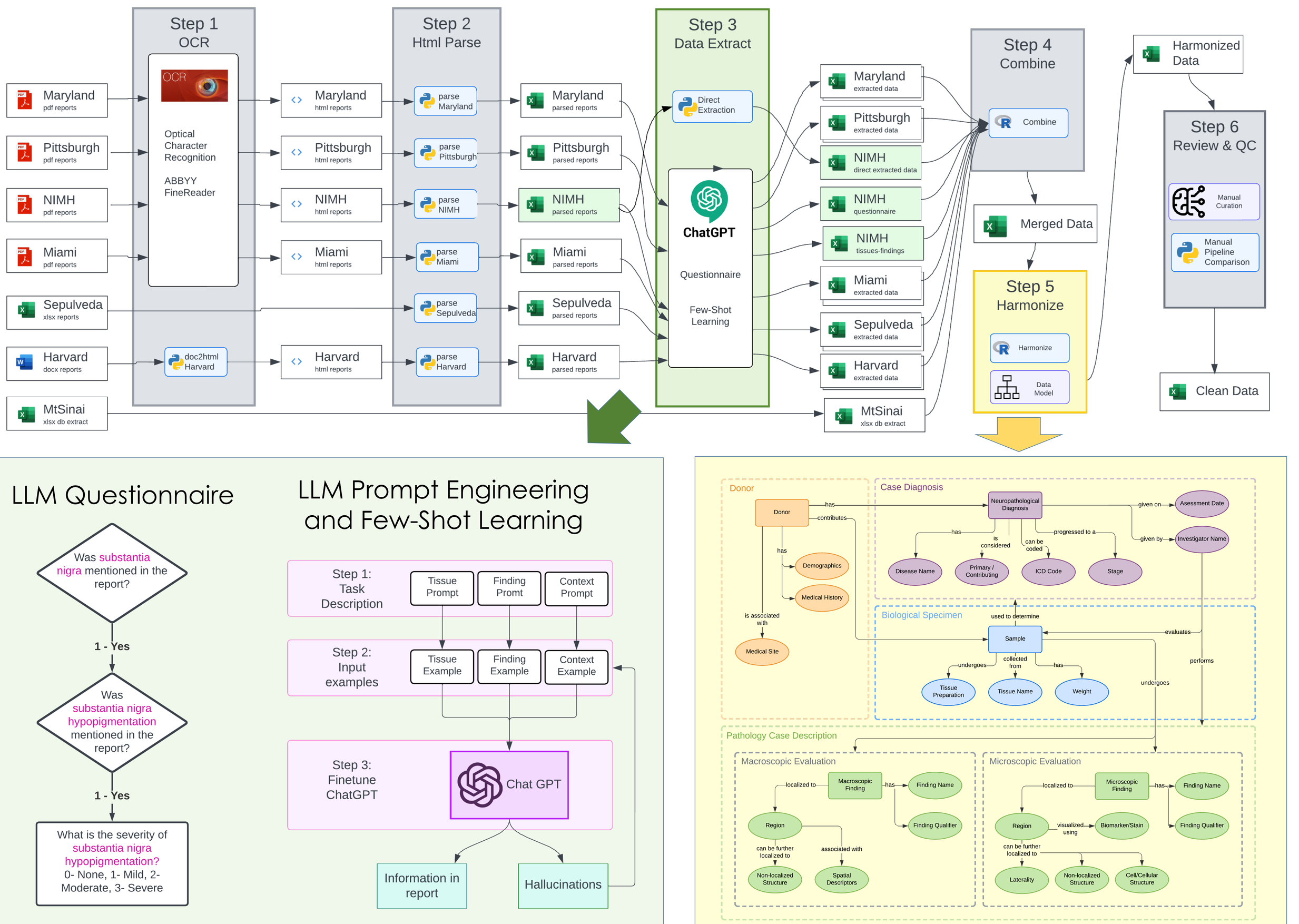
Results and Conclusions

The pilot study demonstrated the potential of LLMs in structuring unstructured neuropathological reports of Parkinson's Disease patients available in the NBB. The developed data model and extraction pipeline allowed for the accurate extraction of 74.2% findings from microscopic evaluations. These findings were then mapped and harmonized with a data model specifically developed for neuropathological conditions. This approach could be applied to other conditions and facilitate research efforts.



*Data is harmonized to different entities in the data dictionary but the difference doesn't affect the meaning. Examples: cerebral hemisphere vs cerebrum; neostriatum vs all of (caudate nucleus, putamen and nucleus accumbens)

Data Preprocessing and Parsing | Information Extraction | Harmonization | QC and Review



LLM Questionnaire

LLM Prompt Engineering and Few-Shot Learning

* data which meet principles of findability, accessibility, interoperability, and reusability.