

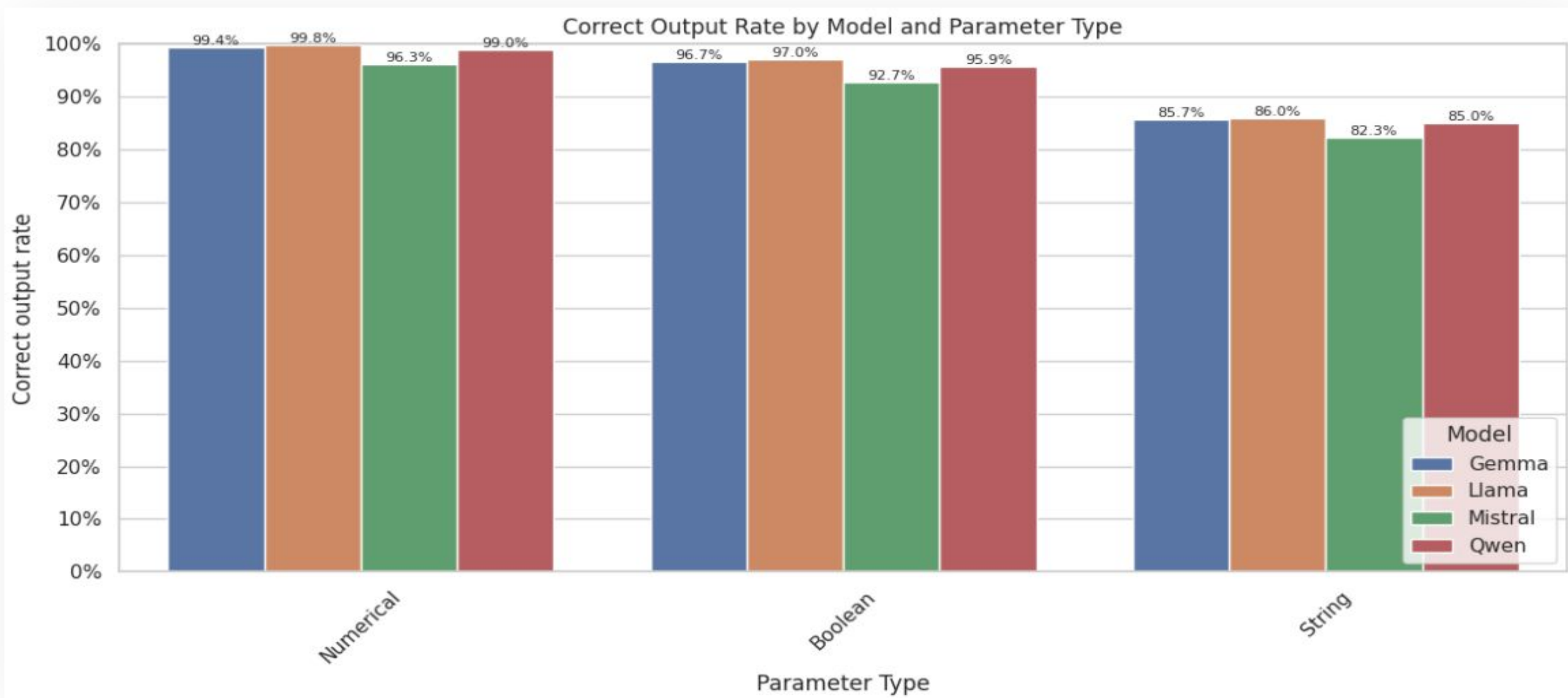
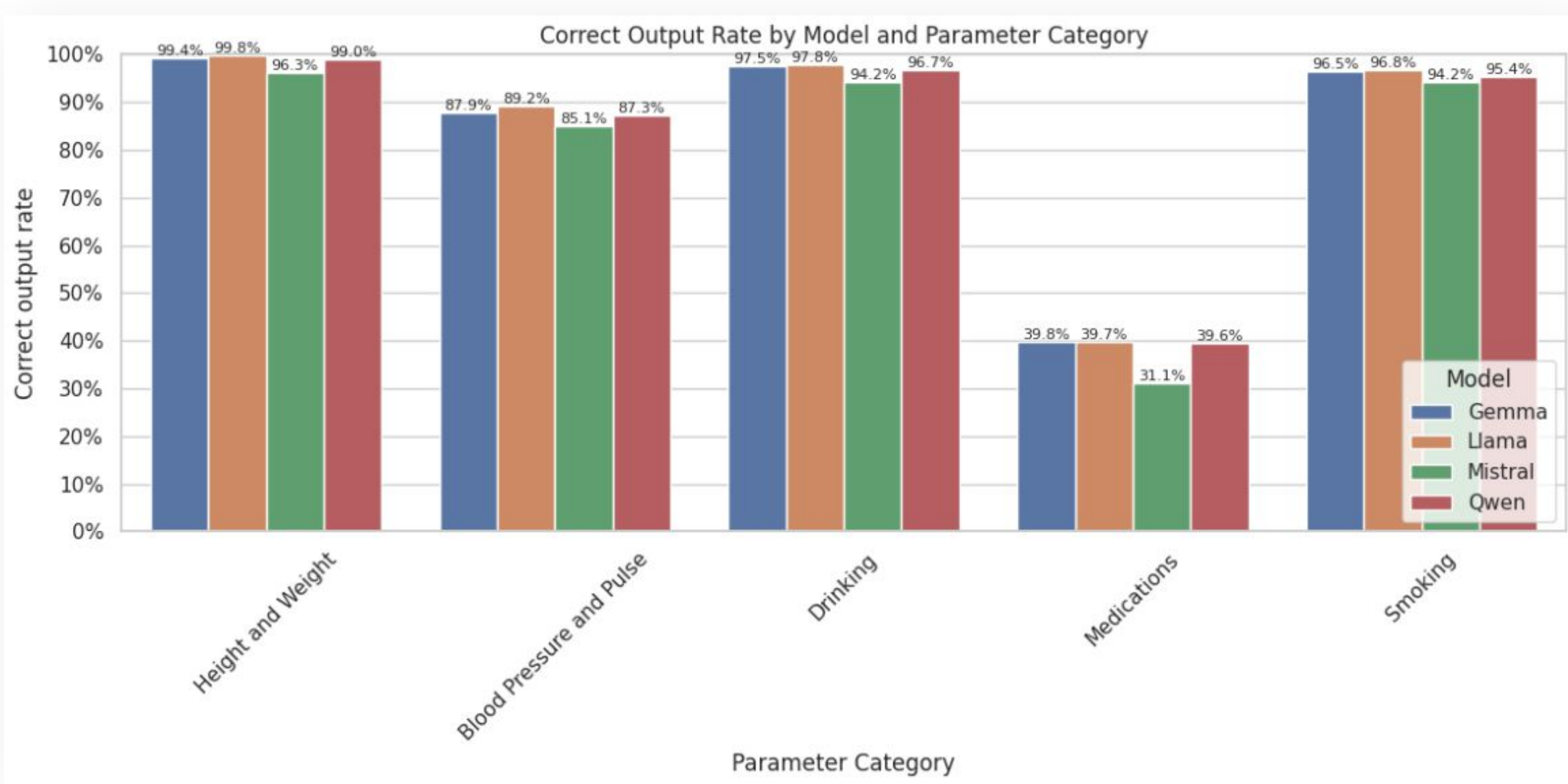
High-quality targeted parameter extraction from free text documents using LLMs is feasible due to very high overall accuracy. After post-processing and mapping to the OMOP standard, the acquired data can be integrated into our data warehouse.

Expanding the Semmelweis Clinical Database with parameters extracted from free text documents using large language models

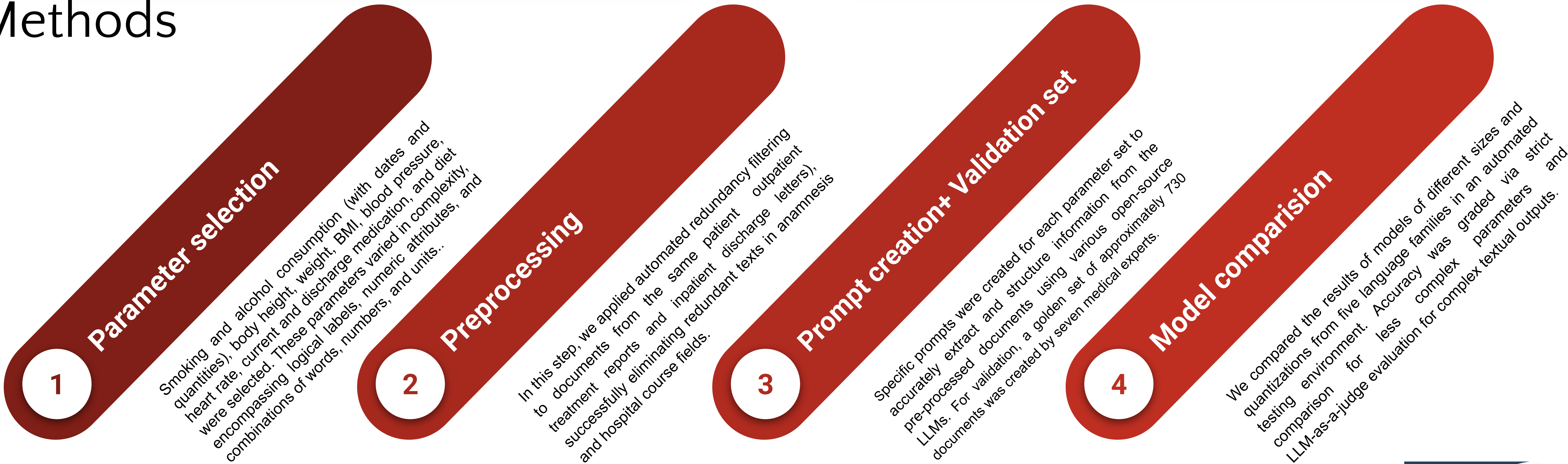
Background: The Semmelweis Clinical Database stores OMOP-standardized clinical data from approximately 2 million patient visits since 2011, including diagnoses, procedures, drugs, and laboratory observations. However, significant information, such as family history, health behaviors, physical status, medication, and diet, remains in unstructured free-text format, hindering research. Applying Large Language Models (LLMs) for medical data extraction lacks out-of-the-box solutions. Our aim was to identify techniques and test the usability of different LLMs for extracting and structuring selected parameters.

Result 1: To efficient data extraction different prompting techniques were needed to be utilized. Finding these category specific prompts was a crucial step during the research, and seemingly successful, given the overall good results.

Result 2: While the numerical and boolean output types grasped easier by the models, the string variables might pose greater complexity. This is due to the higher variance of the potential outputs and the greater importance of understanding complex context.



Methods



LLMs successfully identified parameter instances not found through manual labeling. This highlights the superior efficiency and comprehensiveness of our LLM-based method compared to traditional manual extraction, underscoring its power in uncovering valuable insights that might otherwise be overlooked.

