University of Applied Sciences and Arts Northwestern Switzerland
School of Life Sciences

Exploring the Role of LLMs in LOINC Code Mapping

Alfredo CREGO
Master-Thesis

Customer: Dr. Nicole Göbel (USB), Dr. Bram Stieltjes (USB)
Expert: Dr. Peter Janes (Abdagon AG), Alexander Leichtle (Insel Spital)
Supervisor: Dr. Emanuele Laurenzi (FHNW), Pr. Charuta Pande (FHNW)

Abstract

PROBLEM: To share clinical data for research purposes on the Swiss Personalized Health Network (SPHN), data provider institutions, such as university hospitals, need to map their internal laboratory tests to a standard terminology like Logical Observation Identifiers Names and Codes (LOINC).

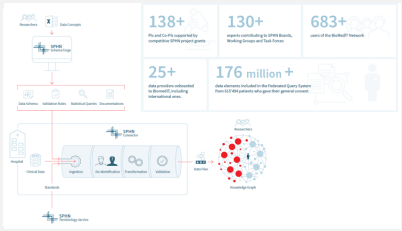
This is a non-trivial task requiring medical knowledge and laboratory expertise and is time consuming for human coders. But it is a necessary step to comply with the SPHN requirements and provide researchers with semantically rich and interoperable information.

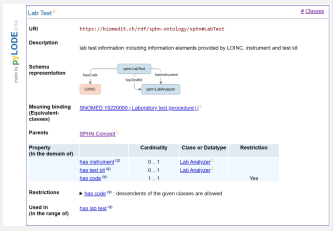
New pre-trained large language models (LLMs) exhibit emerging abilities that could be leveraged to ease this process supporting laboratory experts, but they also present some limitations that need to be addressed (such as hallucinations, context window size, domain specific knowledge, etc.).

GOALS & METHODS: This master thesis research aims to evaluate the feasibility of applying *foundational* LLMs to automate the mapping of local lab tests to standard LOINC codes. Medprompt (Nori et al., 2023) makes the point that *foundational* models can improve their performance through multi-step prompt strategies avoiding the costly fine-tuning. Adopting a similar approach, experimental research has been conducted applying several prompt strategies, from zero-shot, embedding vectors, nearest neighbors, retrieval augmentation, question answering, chain-of-thoughts reasoning, few-shots learning, web search tools and agents.

RESULTS: The results evaluate the precision, completeness and validity of the outputs, in a mixed approach of quantitative and qualitative metrics to compare with a provided ground truth and evaluate, with the help of a human expert, the soundness of the reasonings produced with the models.

SPHN





LOINC & UCUM

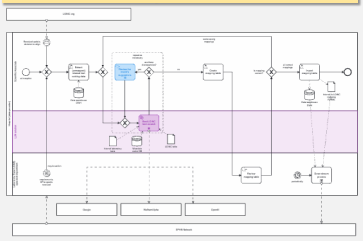
Each LOINC code corresponds to a single laboratory test or panel and is uniquely identifiable by combining six axes (Srinivasan et al., 2006):

1) **Component** or analyte, (e.g., potassium);
2) **Property** measured (e.g., a mass concentration);
3) **Timing** (point in time, or integrated over duration of time in longitudinal studies);
4) Type of **sample** (system) (e.g., blood or organs);
5) Type of **scale** (numerical, quantitative, qualitative, categorical, hierarchical, ordinal, nominal or narrative); and
6) **Method** used to produce the result (where relevant).

Valid UCUM Code	Descriptive Name	Common Synonyms (non-UCUM)
%	Percent	%
/uL	PerMicroliter	/uL
{IU}/L	InternationalUnitsPerLiter	IU/L
10*3/uL	ThousandsPerMicroliter	K/uL, x10^3/mm^3 or
10*6/uL	MillionsPerMicroliter	M/uL, x10^6/mm^3
fL	Femtoliter	fL
g/dL	GramsPerDeciliter	g/dL
g/L	GramsPerLiter	g/L
g/mL	GramsPerMilliliter	g/mL

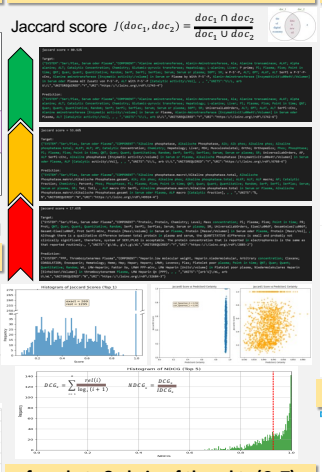
Experimental Research and Results

hybrid-AI

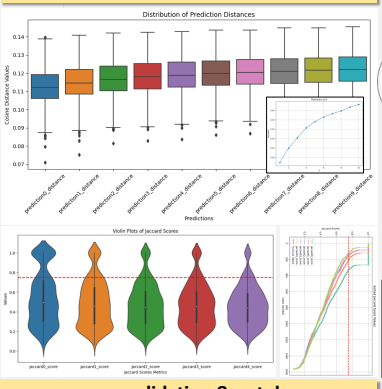


info retrieval metrics

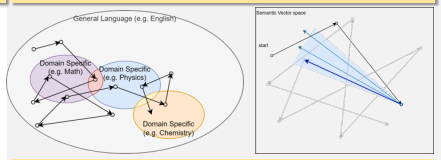
Jaccard score $J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2}$



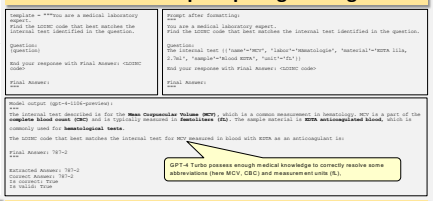
kNN retriever



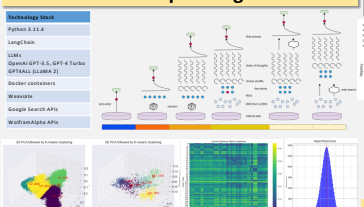
LLMs



zero-shots & prompt engineering



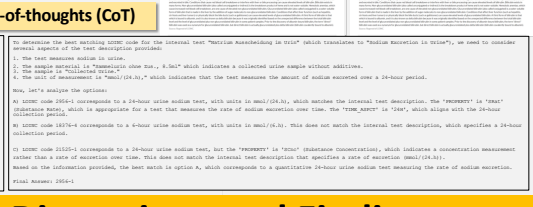
multi-step strategies



parser, validation & ontology

web search, tools & agents

few-shots & chain-of-thoughts (CoT)



Discussions and Findings

LangChain framework was used to prompt OpenAI GPT-4 Turbo, GPT-3.5 and GPT4ALL. Largest model performed best, smaller model was faster, local model was not up to standard. Zero-shots prompt caused LLMs to hallucinate, requiring a retriever to produce valid LOINC codes. Web search retrieved relatively good results. CoT provides expert with explanations to evaluate the outputs. For information expansion, the model needs to elicit tacit knowledge to fill in the missing information from a sparse input.

Countering hallucinations with reliable retrievers is challenging. Retrievers need to integrate reliable data sources (databases and other web services) and enable the models to search for the information they are missing. Analysis of issues limiting the discriminatory power of the embedding vectors with nearest neighbor search show how a relevance metric is needed.

Conclusions

"*If it was a human, I would hire him/her*" said N.G. during the evaluation. Although it still makes mistakes, it gives sound and well-structured responses with reasonings that are easy to read, understand and evaluate for an expert. Even when it is wrong or does not have the response, it triggers further thoughts and lines of investigation that are valuable.

Several prompt strategies were applied to instruct LLMs for a specific task that requires domain knowledge. Foundational models, without fine-tuning, are flexible enough to produce satisfactory results and be a valuable help to support humans in their work. Still, development effort needs to be spent in combining, testing and evaluating different strategies, including formatting instructions, parsing and validation (e.g. ontology) of the outputs to make it usable and reliable in a productive environment.

References
Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine (arXiv:2311.16452). arXiv. https://doi.org/10.48550/arXiv.2311.16452
Srinivasan, A., Kuruppaddiy, N., Mithal, P., & Cascolio, S. W. (2006). Semantic Web Representation of LOINC: An Ontological Perspective. AMIA Annual Symposium Proceedings, 2006, 1107.
SPHN Fact Sheet. (2023). https://sphn.ch/wp-content/uploads/2023/10/SPHN-Fact-Sheet-2023.pdf
LOINC https://loinc.org/

Abdagon

Universitätsspital Basel