

Specification	Implication
<i>Accurate, stable grounding</i>	<i>Return OMOP-standardised terms without hallucination or unpredictable model output</i>
<i>Respect OMOP-style constraints</i>	<i>Generalisable methods to enforce CDM conventions (stronger than just enum lists)</i> <i>Specify task-dependent preferences for term sets and concept hierarchies</i>
<i>Zero-shot, configurable pipeline</i>	<i>New targets and vocabularies specified declaratively without retraining</i>
<i>Reusable & portable</i>	<i>Support sharing of validated configurations - no re-development required</i>
<i>Operate on standard professional grade machines</i>	<i>Must produce workable results with models that can run without dedicated GPU i.e. typically no larger than ~3-7B parameters</i> <i>High-level abstraction to allow more powerful or bespoke models where resourcing and throughput/reasoning demands allow</i>
<i>Run under heavily restricted environments</i>	<i>Support linking to locally hosted models, vocabularies & configuration resources</i> <i>Integrate with securable model hosts</i>

Specification	Implication
<i>Accurate, stable grounding</i>	<i>Return OMOP-standardised terms without hallucination or unpredictable model output</i>
<i>Respect OMOP-style constraints</i>	<i>Generalisable methods to enforce CDM conventions (stronger than just enum lists)</i> <i>Specify task-dependent preferences for term sets and concept hierarchies</i>
<i>Zero-shot, configurable pipeline</i>	<i>New targets and vocabularies specified declaratively without retraining</i>
<i>Reusable & portable</i>	<i>Support sharing of validated configurations - no re-development required</i>
<i>Operate on standard professional grade machines</i>	<i>Must produce workable results with models that can run without dedicated GPU i.e. typically no larger than ~3-7B parameters</i> <i>High-level abstraction to allow more powerful or bespoke models where resourcing and throughput/reasoning demands allow</i>
<i>Run under heavily restricted environments</i>	<i>Support linking to locally hosted models, vocabularies & configuration resources</i> <i>Integrate with securable model hosts</i>

Semantically Grounded

Specification	Implication
<i>Accurate, stable grounding</i>	<i>Return OMOP-standardised terms without hallucination or unpredictable model output</i>
<i>Respect OMOP-style constraints</i>	<i>Generalisable methods to enforce CDM conventions (stronger than just enum lists)</i> <i>Specify task-dependent preferences for term sets and concept hierarchies</i>
<i>Zero-shot, configurable pipeline</i>	<i>New targets and vocabularies specified declaratively without retraining</i>
<i>Reusable & portable</i>	<i>Support sharing of validated configurations - no re-development required</i>
<i>Operate on standard professional grade machines</i>	<i>Must produce workable results with models that can run without dedicated GPU i.e. typically no larger than ~3-7B parameters</i> <i>High-level abstraction to allow more powerful or bespoke models where resourcing and throughput/reasoning demands allow</i>
<i>Run under heavily restricted environments</i>	<i>Support linking to locally hosted models, vocabularies & configuration resources</i> <i>Integrate with securable model hosts</i>

Semantically Grounded

↔


Configurable & Sharable

Specification	Implication
<i>Accurate, stable grounding</i>	<i>Return OMOP-standardised terms without hallucination or unpredictable model output</i>
<i>Respect OMOP-style constraints</i>	<i>Generalisable methods to enforce CDM conventions (stronger than just enum lists)</i> <i>Specify task-dependent preferences for term sets and concept hierarchies</i>
<i>Zero-shot, configurable pipeline</i>	<i>New targets and vocabularies specified declaratively without retraining</i>
<i>Reusable & portable</i>	<i>Support sharing of validated configurations - no re-development required</i>
<i>Operate on standard professional grade machines</i>	<i>Must produce workable results with models that can run without dedicated GPU i.e. typically no larger than ~3-7B parameters</i> <i>High-level abstraction to allow more powerful or bespoke models where resourcing and throughput/reasoning demands allow</i>
<i>Run under heavily restricted environments</i>	<i>Support linking to locally hosted models, vocabularies & configuration resources</i> <i>Integrate with securable model hosts</i>








JOURNAL ARTICLE

Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning

J Harry Caufield , Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglou, HyeonSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel ... [Show more](#)

Bioinformatics, Volume 40, Issue 3, March 2024, btae104, <https://doi.org/10.1093/bioinformatics/btae104>
Published: 21 February 2024 **Article history** ▼

 PDF  Views ▼  Cite  Permissions  Share ▼




Abstract

Motivation

Creating knowledge bases and ontologies is a time consuming task that relies on manual curation. AI/NLP approaches can assist expert curators in populating these knowledge bases, but current approaches rely on extensive training data, and are not able to populate arbitrarily complex nested knowledge schemas.

- *Works OK, with some limitations*
 - *Brittle parsing of outputs*
 - *No OMOP-specific grounding files*
 - *Issues working in restricted / airgapped environments*
 - *Strict cardinality requirements don't reflect reality of clinical notes*

(N.B. for the most part, these limitations are specific to the implementation, not conceptual. i.e. solvable)


OntoGPT



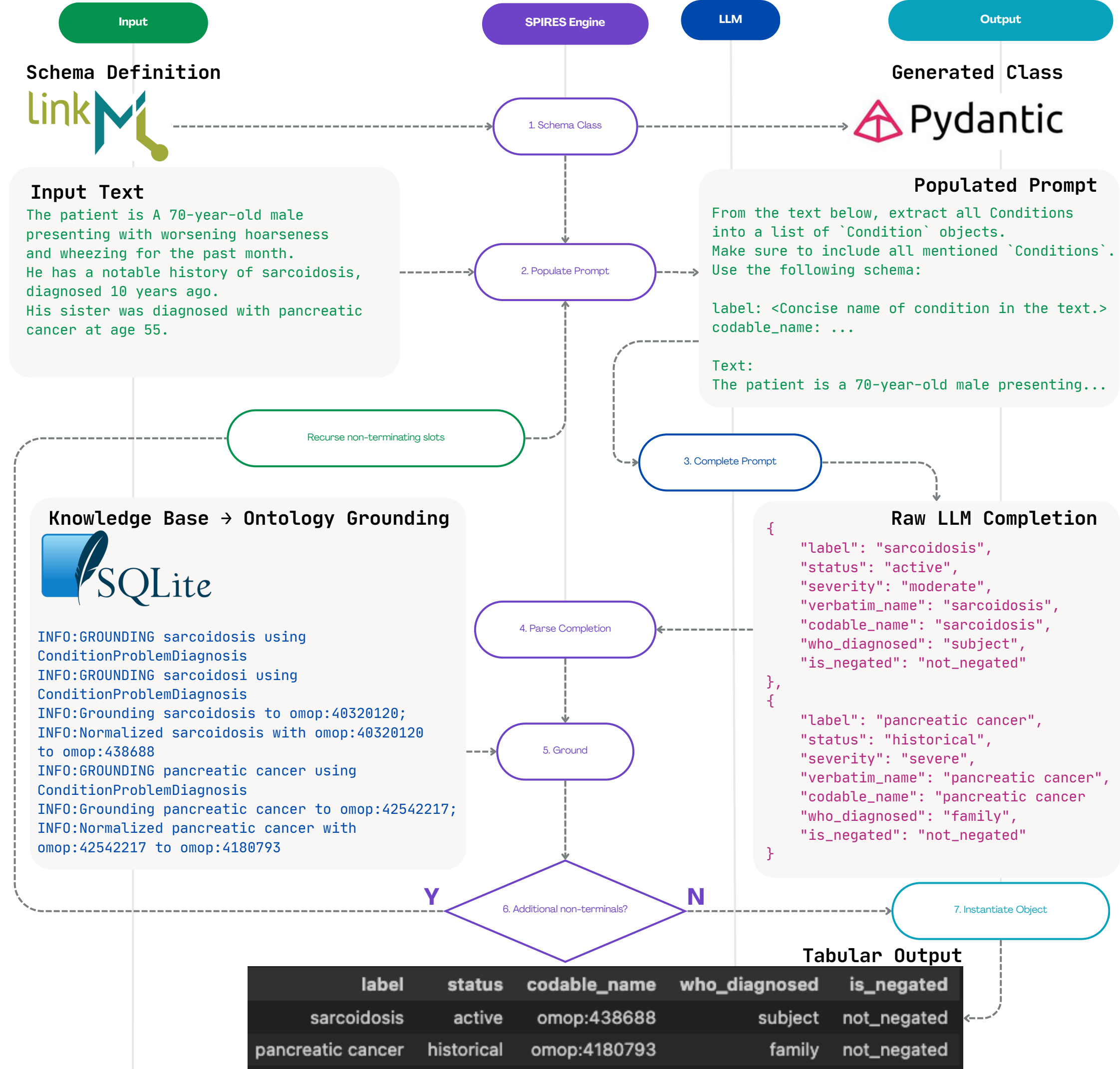
Introduction

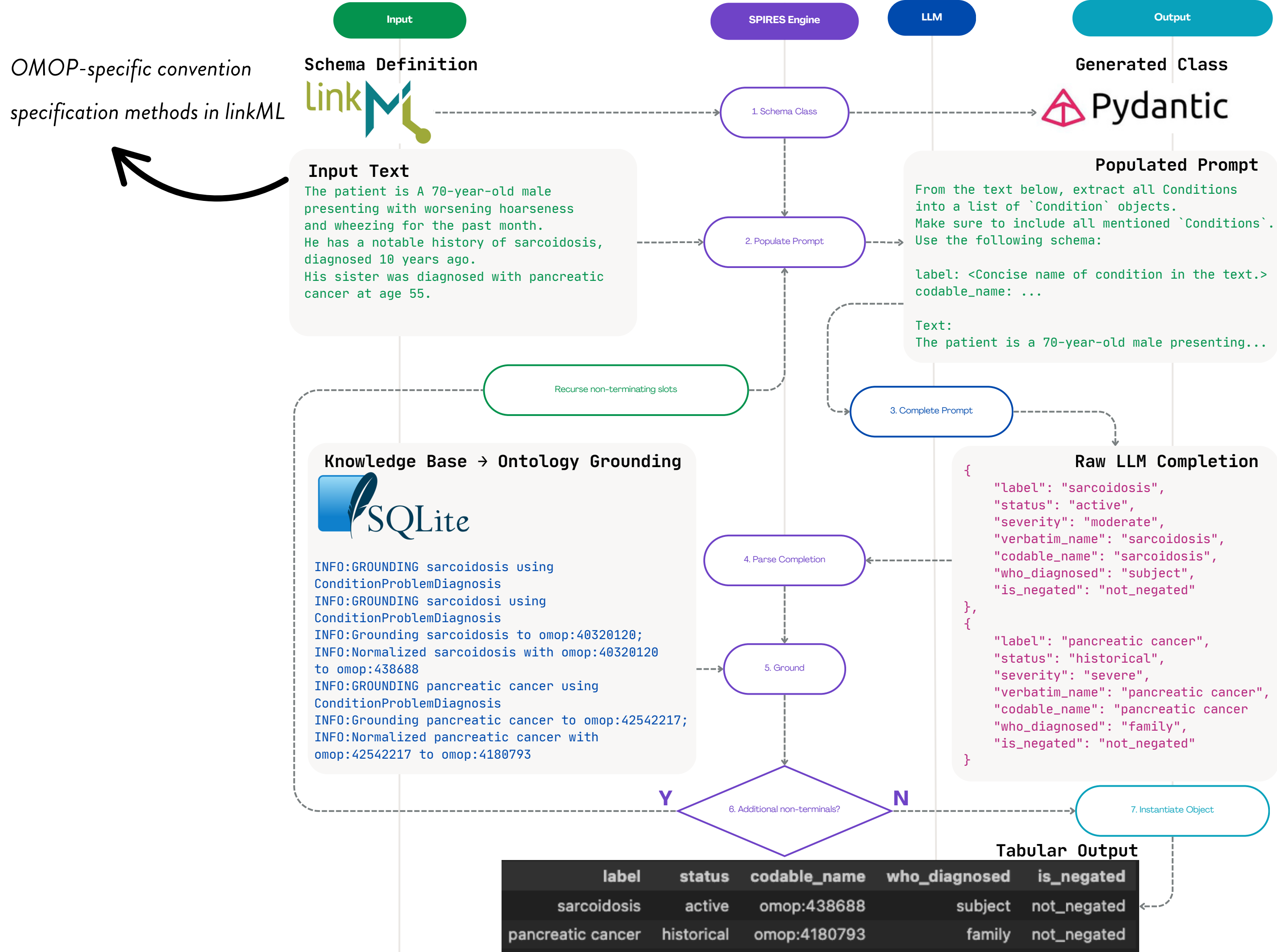
OntoGPT is a Python package for extracting structured information from text with large language models (LLMs), *instruction prompts*, and ontology-based grounding. It works well with OpenAI's GPT models as well as a selection of other LLMs. OntoGPT's output can be used for general-purpose natural language tasks (e.g., named entity recognition and relation extraction), summarization, knowledge base and knowledge graph construction, and more.

Methods

The primary extraction method currently implemented in OntoGPT is SPIRES:

- SPIRES: *Structured Prompt Interrogation and Recursive Extraction of Semantics*
- A Zero-shot learning (ZSL) approach to extracting nested semantic structures from text
- This approach takes two inputs - 1) LinkML schema 2) free text, and outputs knowledge in a structure conformant with the supplied schema in JSON, YAML, RDF or OWL formats
- Uses OpenAI GPT models through their API, or one of a variety of LLMs on your local machine





```
id: https://athena.ohdsi.org/search-terms/terms#person
name: person
title: OMOP CDM Person
description: >--
  Template for OMOP CDM Person class, which represents an individual in the OMOP Common Data Model.
license: https://creativecommons.org/publicdomain/zero/1.0/
prefixes:
  linkml: https://w3id.org/linkml/
  omop: https://athena.ohdsi.org/search-terms/terms#
  owl: "http://www.w3.org/2002/07/owl#"
  rdf: "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  rdfs: "http://www.w3.org/2000/01/rdf-schema#"
imports:
  - linkml:types
  - cdm_vocabulary
classes:
  People:
    tree_root: true
    attributes:
      persons:
        multivalued: true
        inlined_as_list: true
        range: Person
  Person:
    attributes:
      person_id:
        identifier: true
        description: Unique identifier for the person in the OMOP CDM.
        range: integer
      gender_concept_id:
        description: Concept ID representing biological sex at birth of the Person.
        range: standard_gender_domain
        required: true
      year_of_birth:
        description: Year of birth of the Person.
        range: integer
        required: true
      month_of_birth:
        description: Month of birth of the Person.
        range: integer
        minimum_value: 1
        maximum_value: 12
        required: false
      day_of_birth:
        description: Day of birth of the Person.
        range: integer
        minimum_value: 1
        maximum_value: 31
        required: false
enums:
  standard_gender_domain:
    description: >--
      Standard concepts constrained to domain 'Gender' in OMOP CDM.
    reachable_from:
      source_ontology: omop
      source_nodes:
        - omop:omop#2 # Gender domain
```

Generated Class



Populated Prompt

From the text below, extract all Conditions into a list of `Condition` objects. Make sure to include all mentioned `Conditions`. Use the following schema:

label: <Concise name of condition in the text.>
codable_name: ...

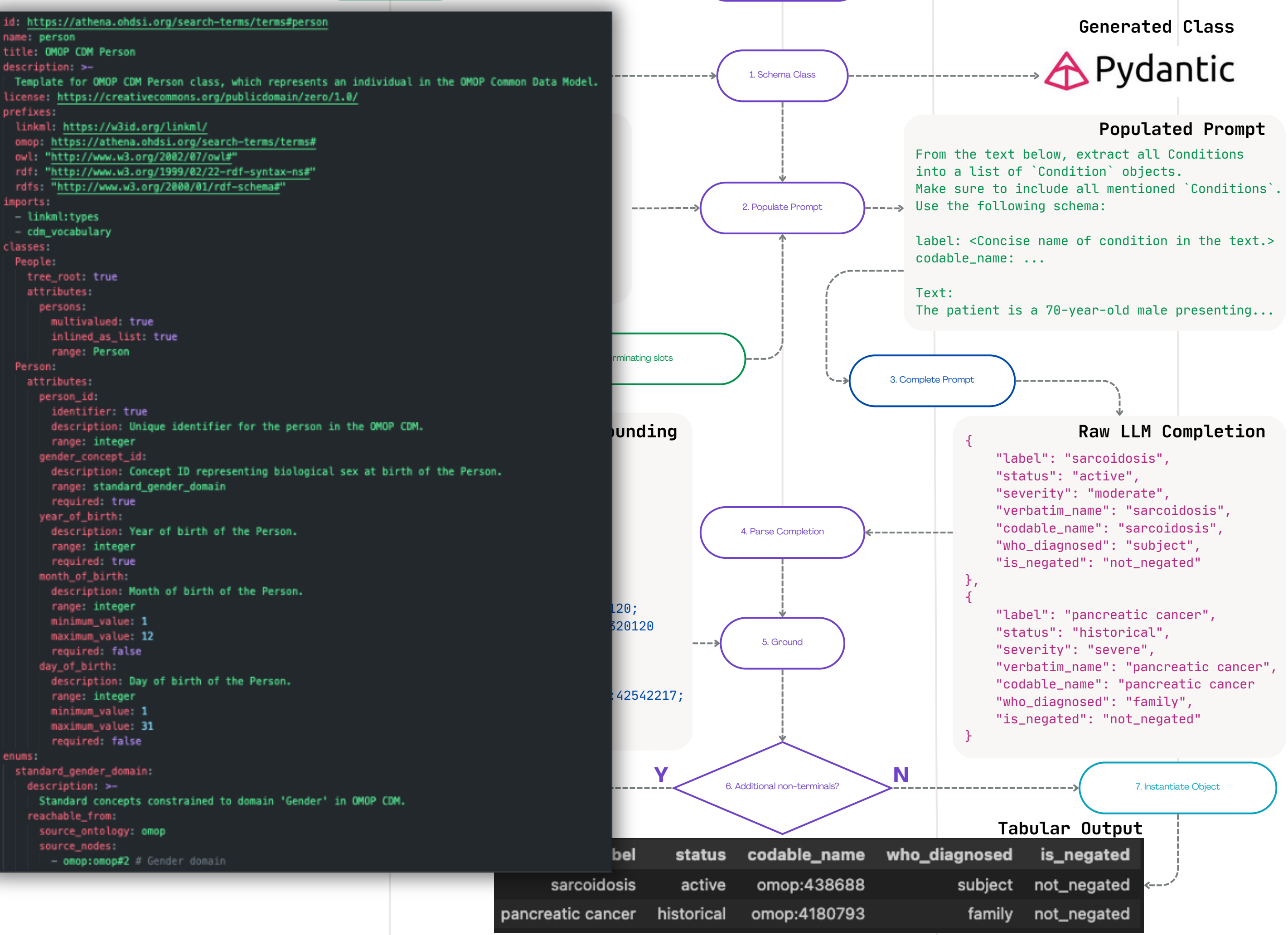
Text:
The patient is a 70-year-old male presenting...

Raw LLM Completion

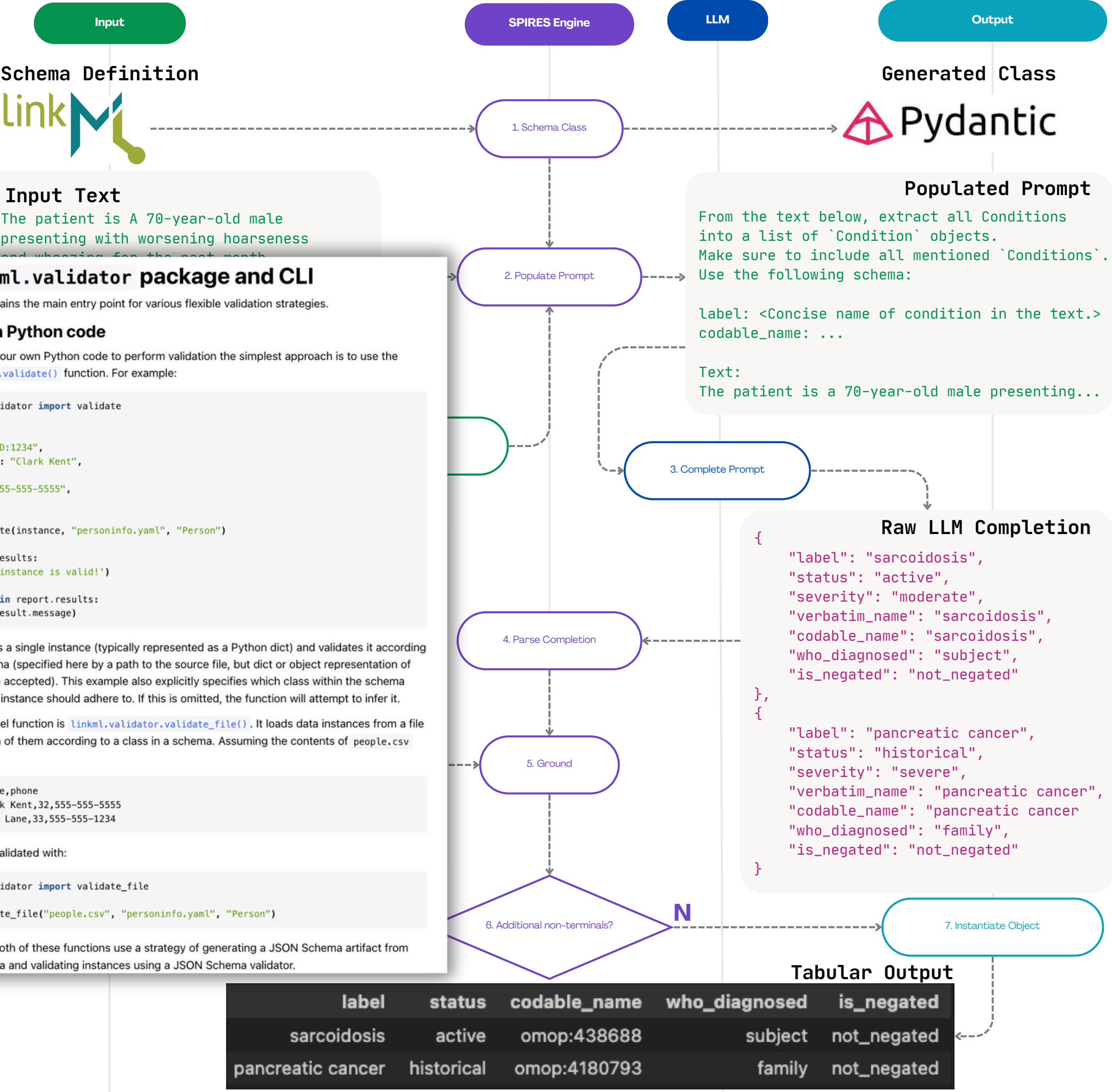
```
{
  "label": "sarcoidosis",
  "status": "active",
  "severity": "moderate",
  "verbatim_name": "sarcoidosis",
  "codable_name": "sarcoidosis",
  "who_diagnosed": "subject",
  "is_negated": "not_negated"
},
{
  "label": "pancreatic cancer",
  "status": "historical",
  "severity": "severe",
  "verbatim_name": "pancreatic cancer",
  "codable_name": "pancreatic cancer",
  "who_diagnosed": "family",
  "is_negated": "not_negated"
}
```

Tabular Output

label	status	codable_name	who_diagnosed	is_negated
sarcoidosis	active	omop:438688	subject	not_negated
pancreatic cancer	historical	omop:4180793	family	not_negated



OMOP-specific convention
specification methods in linkML



LinkM

linkml documentation

Q Search

CONTENTS:

LinkML at a glance

Quick Install Guide

Tutorial

LinkML Schemas

Working with Data

Converting between different representations

Data Validation

Working with RDF and LinkML

CSVs and Tabular Data

Python

Working with data in SQL Databases

Generators

How-to Guides

Examples of use

The LinkML Ecosystem

LinkML specification

Contribution Guidelines

FAQ

DEVELOPERS GUIDE:

CLI

The linkml.validator package and CLI

This package contains the main entry point for various flexible validation strategies.

Validation in Python code

If you are writing your own Python code to perform validation the simplest approach is to use the `linkml.validator.validate()` function. For example:

```
from linkml.validator import validate

instance = {
    "id": "ORCID:1234",
    "full_name": "Clark Kent",
    "age": 32,
    "phone": "555-555-5555",
}

report = validate(instance, "personinfo.yaml", "Person")

if not report.results:
    print('The instance is valid!')
else:
    for result in report.results:
        print(result.message)
```

This function takes a single instance (typically represented as a Python dict) and validates it according to the given schema (specified here by a path to the source file, but dict or object representation of the schema is also accepted). This example also explicitly specifies which class within the schema (`Person`) the data instance should adhere to. If this is omitted, the function will attempt to infer it.

The other high-level function is `linkml.validator.validate_file()`. It loads data instances from a file and validates each of them according to a class in a schema. Assuming the contents of `people.csv` look like:

```
id,full_name,age,phone
ORCID:1234,Clark Kent,32,555-555-5555
ORCID:5678,Lois Lane,33,555-555-1234
```

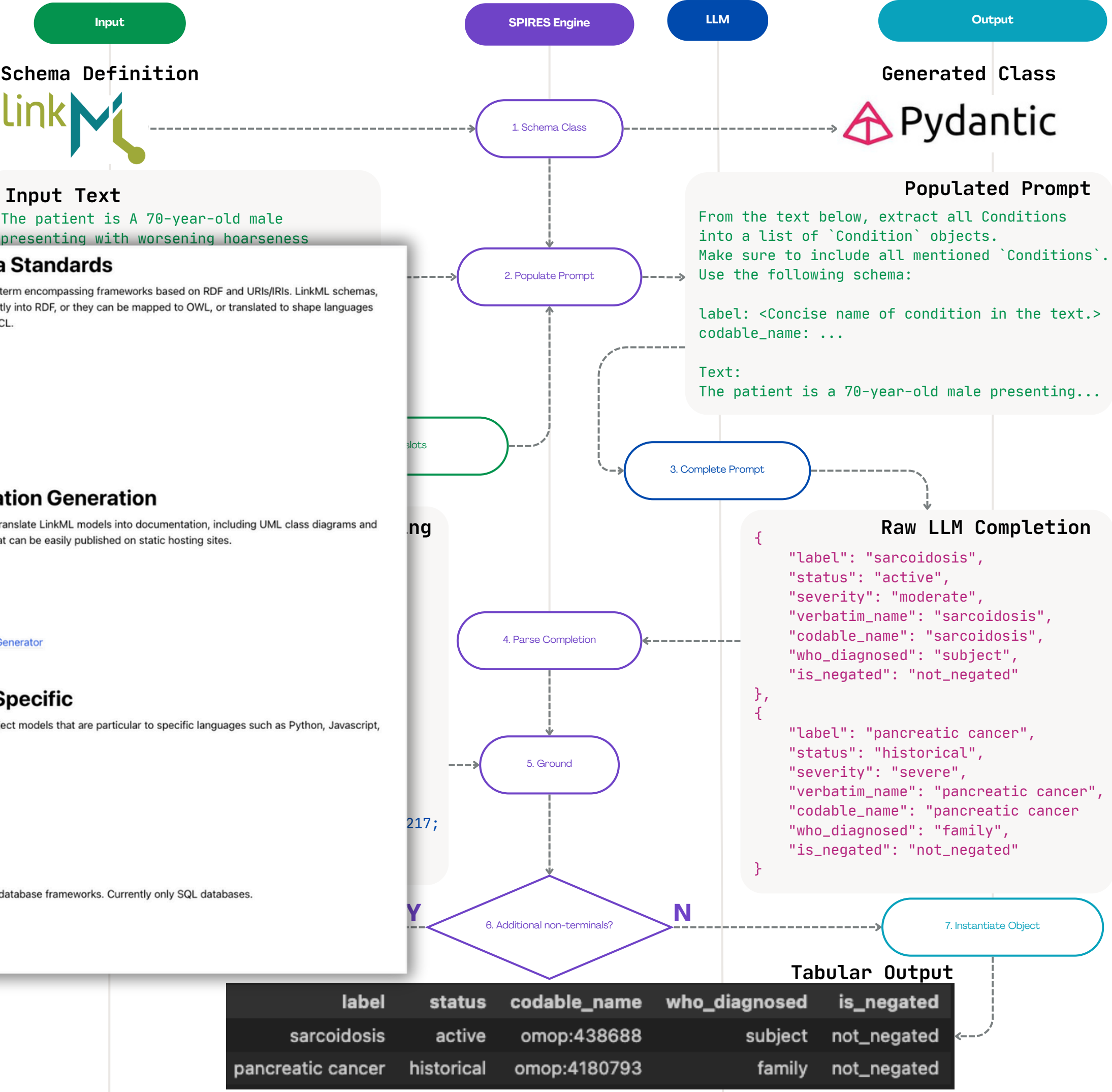
Each row can be validated with:

```
from linkml.validator import validate_file

report = validate_file("people.csv", "personinfo.yaml", "Person")
```

Under the hood, both of these functions use a strategy of generating a JSON Schema artifact from the LinkML schema and validating instances using a JSON Schema validator.

OMOP-specific convention
specification methods in linkML



linkml documentation

Q Search

CONTENTS:

LinkML at a glance

Quick Install Guide

Tutorial

LinkML Schemas

Working with Data

Generators

- JSON Schema
- ProtoBuf
- GraphQL
- JSON-LD Contexts
- JSON-LD
- RDF
- SPARQL
- ShEx
- SHACL
- OWL
- Markdown
- Documentation
- ER Diagrams
- YUML
- PlantUML Diagram Generator
- Project Generator
- Python
- Pydantic

Linked Data Standards

Linked Data is a broad term encompassing frameworks based on RDF and URIs/IRIs. LinkML schemas, can be translated directly into RDF, or they can be mapped to OWL, or translated to shape languages such as ShEx and SHACL.

CONTENTS:

- JSON-LD Contexts
- JSON-LD
- RDF
- SPARQL
- ShEx
- SHACL
- OWL

Documentation Generation

These generators will translate LinkML models into documentation, including UML class diagrams and markdown websites that can be easily published on static hosting sites.

CONTENTS:

- Markdown
- Documentation
- ER Diagrams
- YUML
- PlantUML Diagram Generator
- Project Generator

Language Specific

These will generate object models that are particular to specific languages such as Python, Javascript, or Java.

CONTENTS:

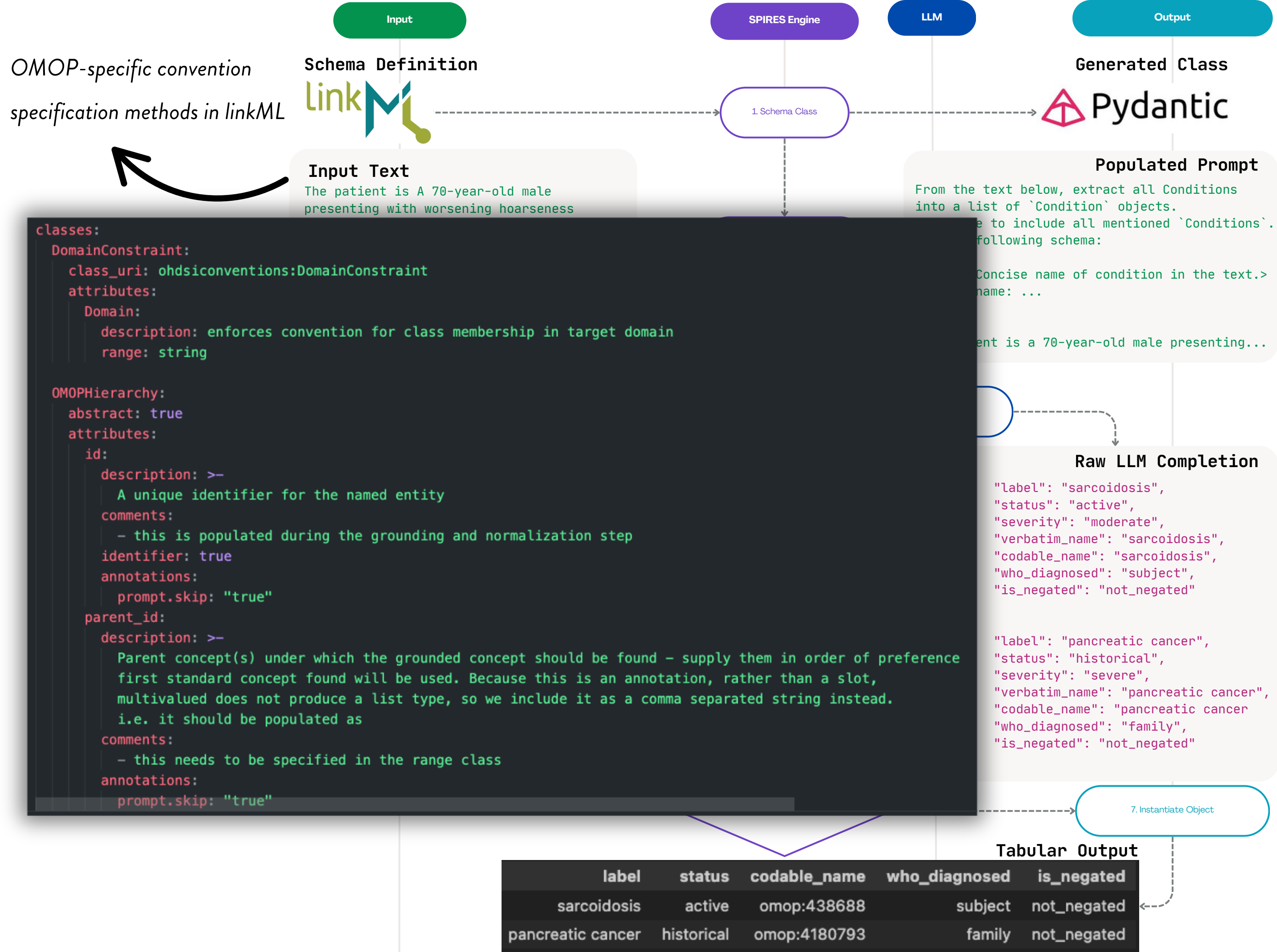
- Python
- Pydantic
- Java
- Typescript

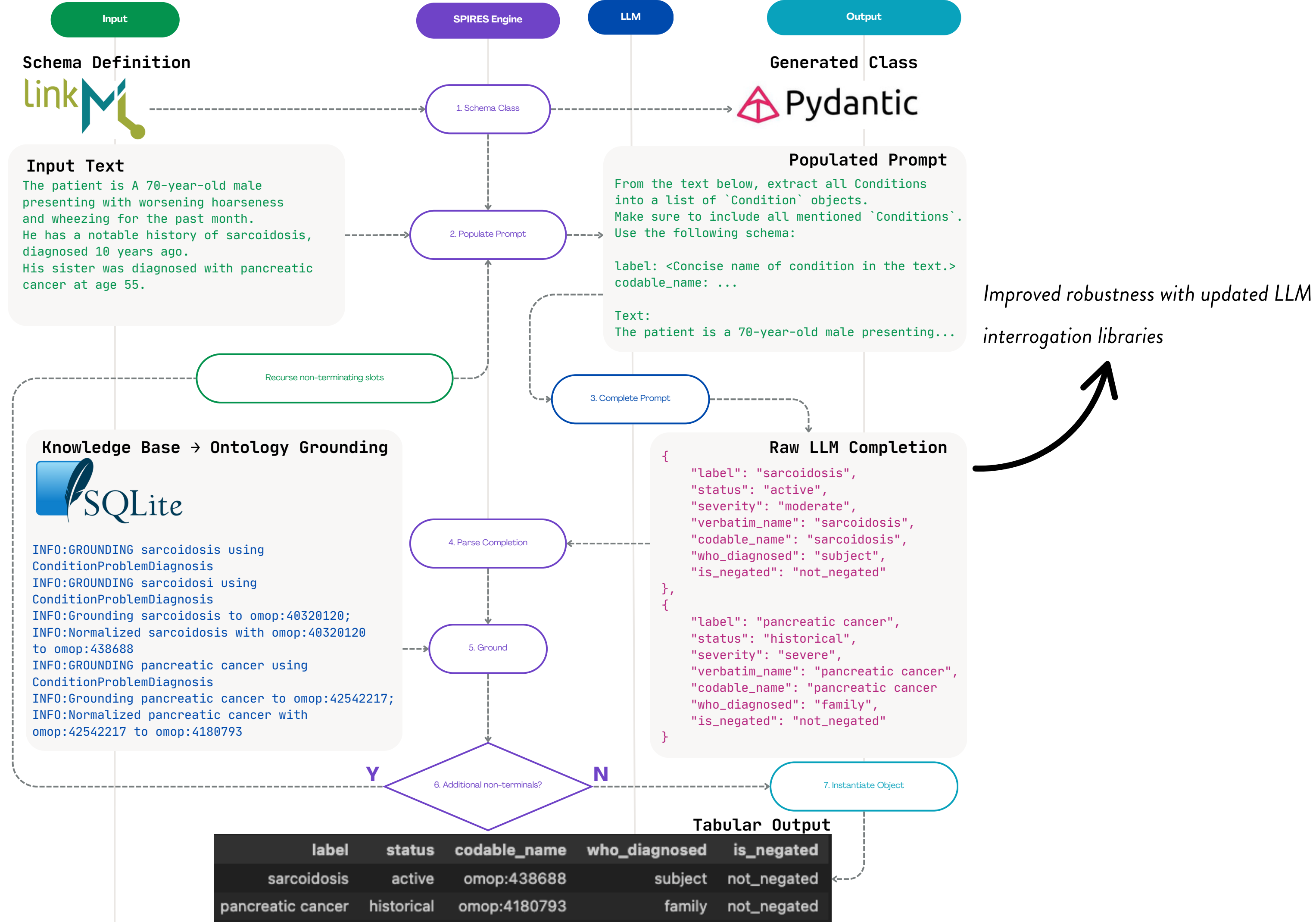
Database

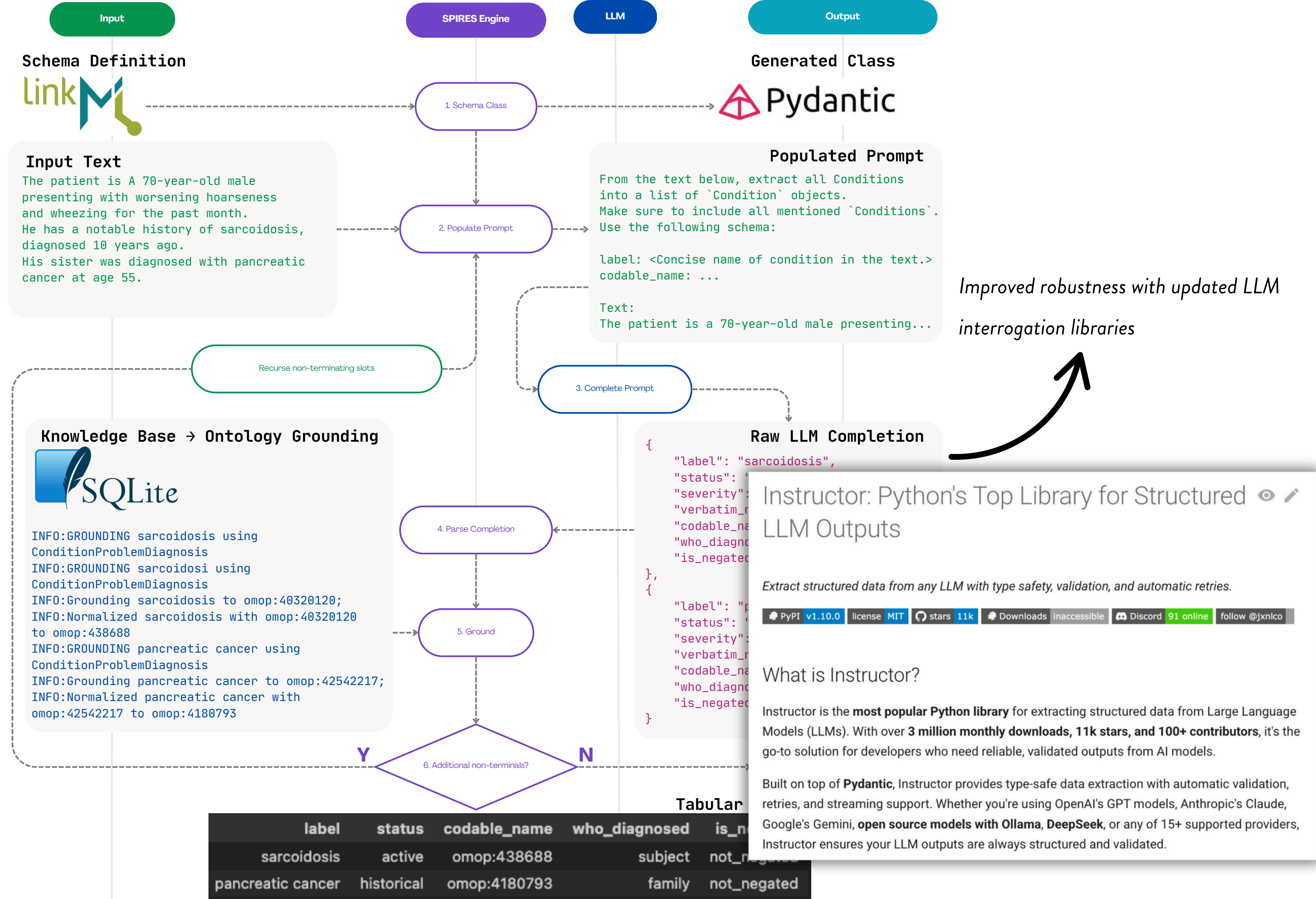
Generators specific to database frameworks. Currently only SQL databases.

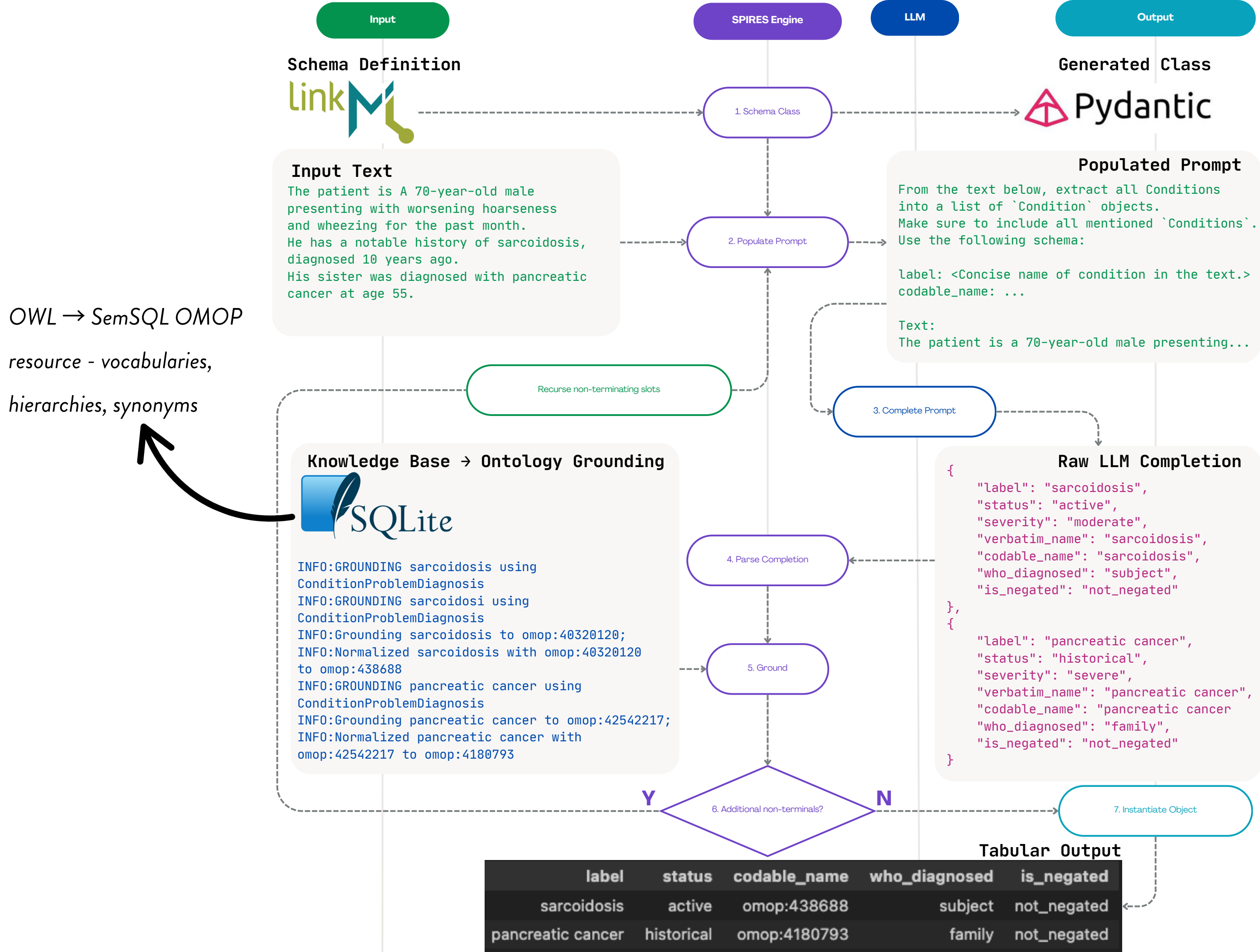
CONTENTS:

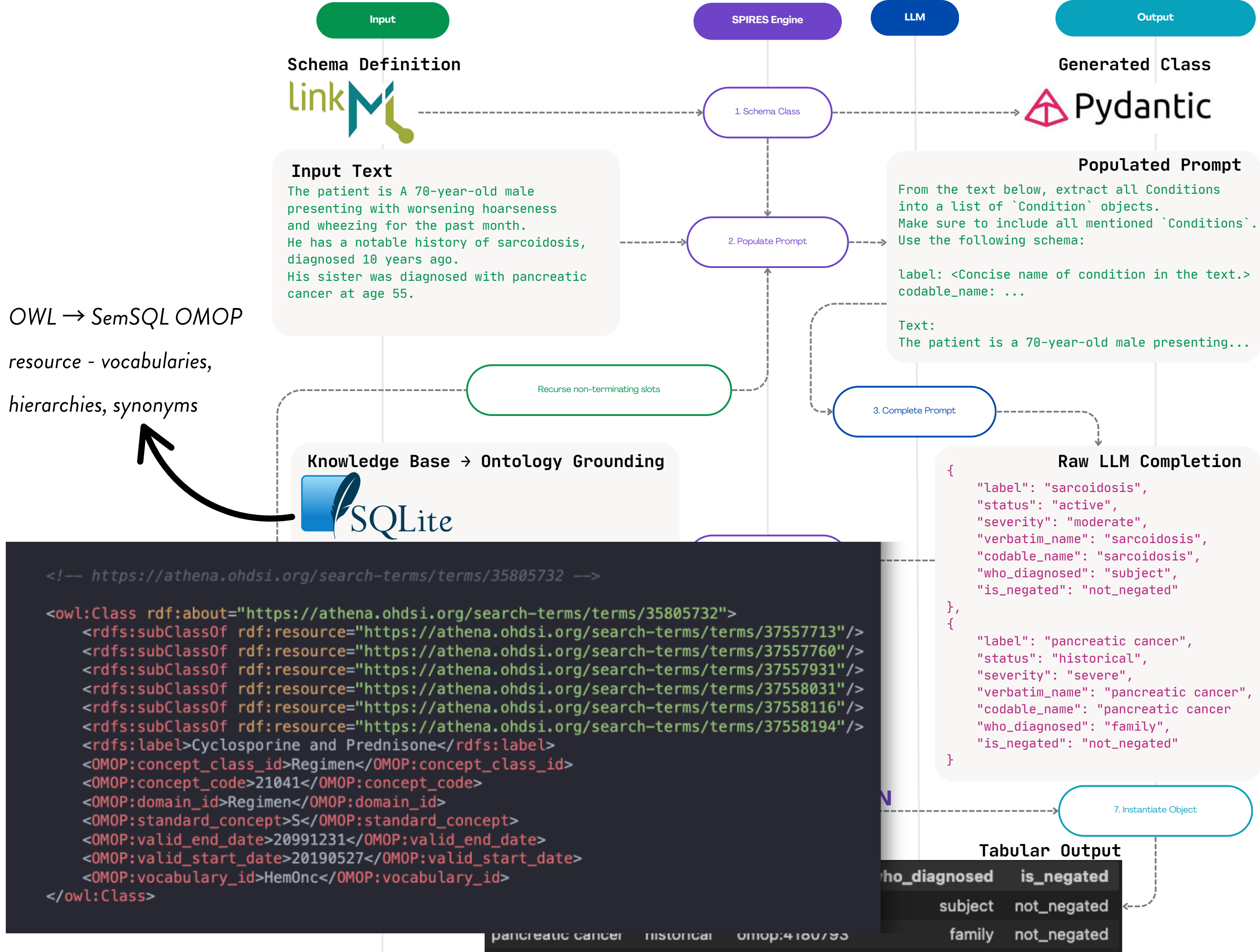
- SQL DDL
- SQLAlchemy

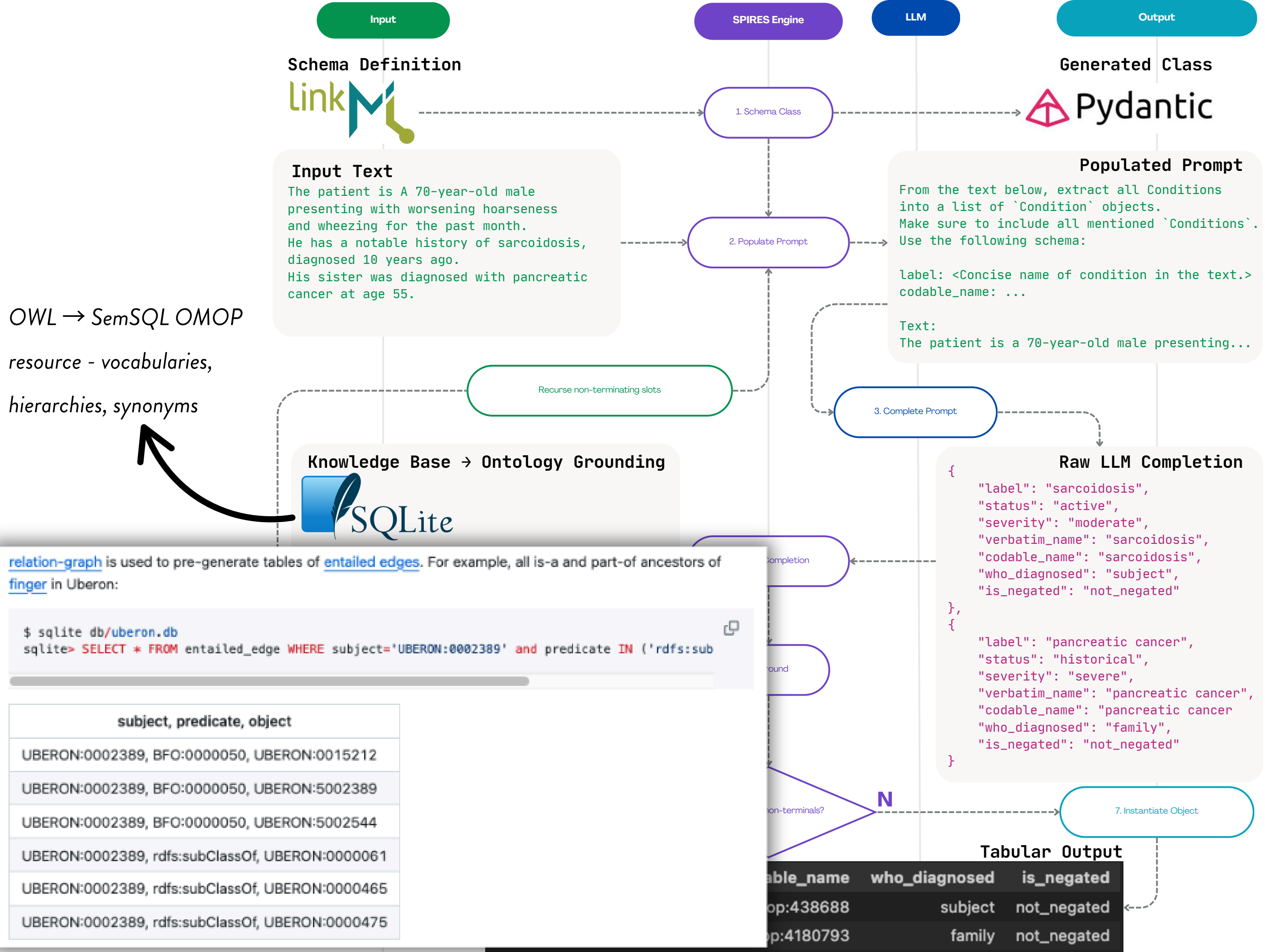


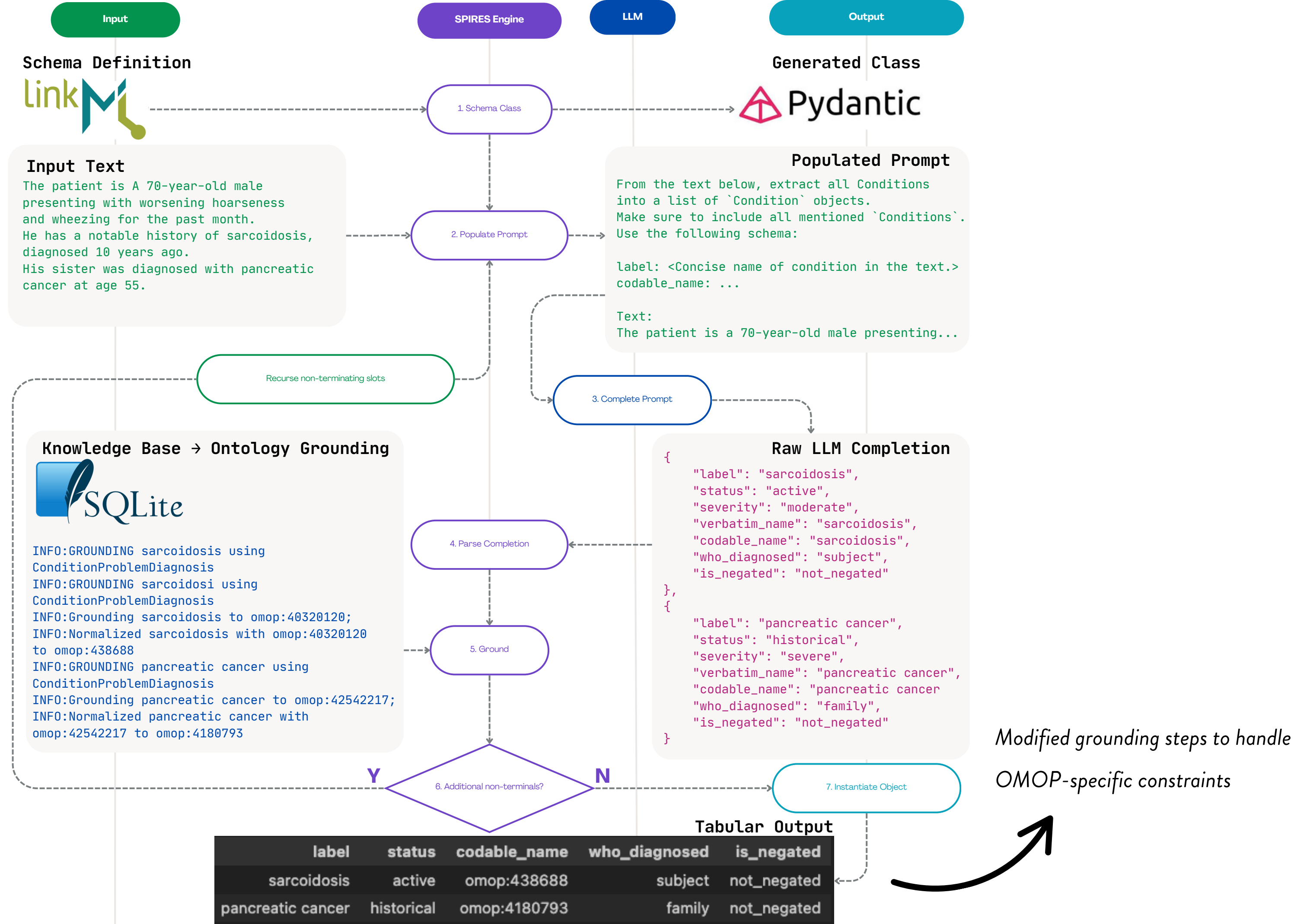


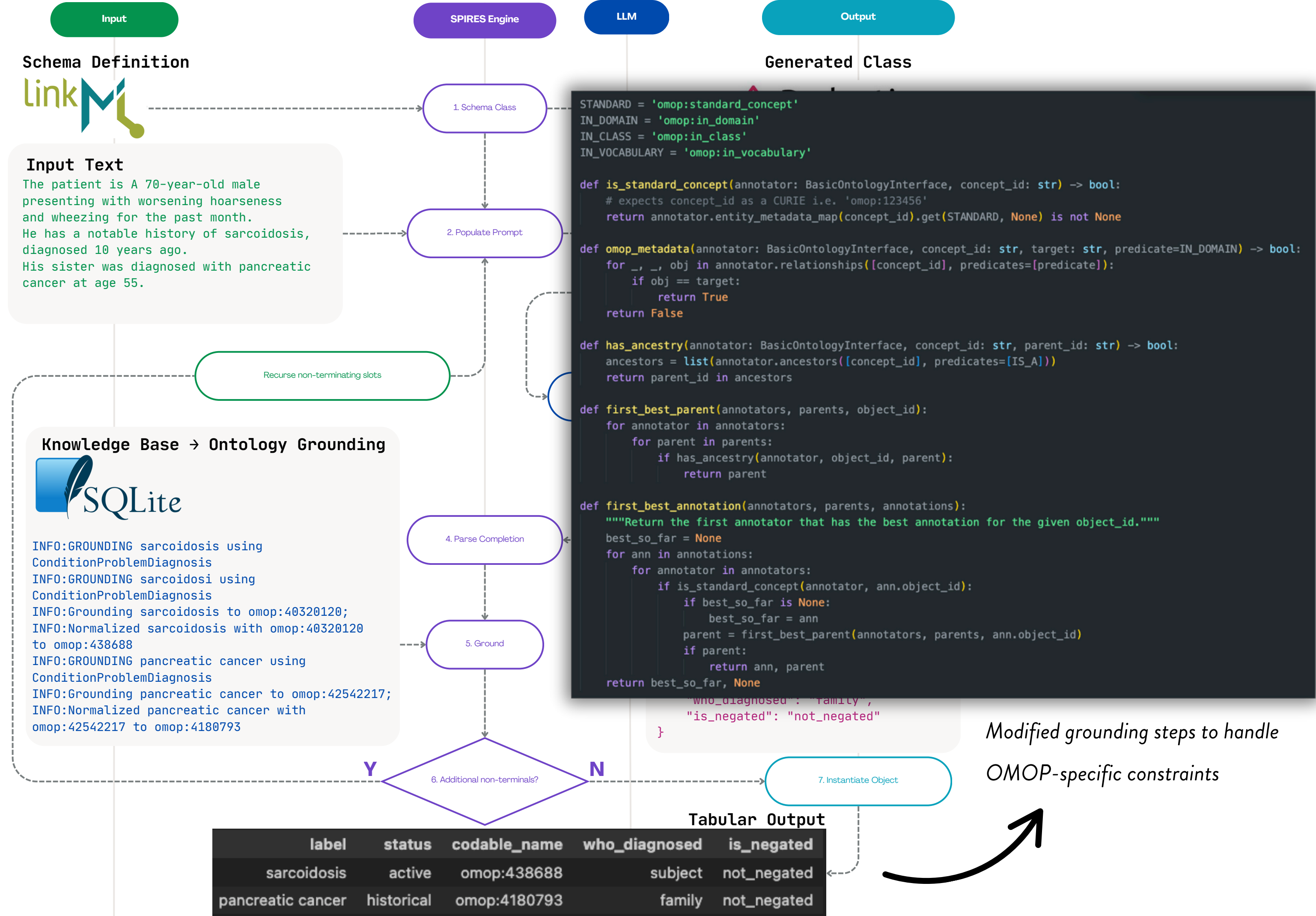












INPUT	llama3 label	IDs	evaluation	usagi label	IDs	evaluation	medspacy label	IDs	evaluation
<i>rectum + pelvis</i>	Pelvic region;Rectum structure	4044034;4144541	correct	PEL	4123163	invalid			no match
<i>(l) lung</i>	Lung structure	4213162	correct	Lung structure	4213162	correct	Entire lung	4111459	correct
<i>ph i prostate+sv</i>	Prostatic structure	4165732	valid	Prostatic structure	4165732	correct			no match
<i>glottis</i>	Glottis structure	4047227	correct	Glottis structure	4047227	correct	Entire glottis	4131315	correct
<i>l chest wall</i>	Chest wall structure	4193513	correct	Chest wall structure	4193513	correct	Entire chest wall	4109932	correct
<i>lt femur</i>	Bone structure of femur	4323581	correct	Bone structure of femur	4323581	correct	Entire bone of femur	37115374	correct
<i>ph1 prostate</i>	Prostatic structure	4165732	correct	Prostatic structure	4165732	correct	Entire prostate	4110208	correct
<i>l/s spine</i>	lumbar	4045660	correct	Structure of vertebral column	4227378	valid	Entire vertebral column	4185891	valid
<i>ph2 prostate bed</i>	Prostatic structure	4165732	correct	Prostatic structure	4165732	correct	Entire prostate	4110208	correct
<i>distal oesophagus</i>	Esophageal structure	4140098	correct	Esophageal structure	4140098	correct	Esophageal structure	4140098	correct
<i>rtbreast+scf+imc+sib</i>	Breast structure	4298444	valid	Breast structure	4298444	correct			no match
<i>prostate + pelvis</i>	Prostatic structure;Pelvic region	4165732;4044034	correct	PEL	4123163	invalid	Entire prostate, Entire pelvis	4110208, 4041832	correct
<i>rt nasal ala</i>			no match	Lateral nasal artery	37157433	invalid			no match
<i>t11-l3</i>	thoracic;lumbar	4047490;4045660	correct	ST11	4159026	invalid	Level of the eleventh thoracic vertebra	4134469	near match
<i>t9-l3</i>	thoracic	4047490	correct	T9-T10 rotator thoracis	4077547	invalid			no match
<i>upper pelvis</i>	Pelvic region	4044034	correct	PEL	4123163	invalid	Entire pelvis	4041832	correct
<i>rt parietal</i>	Brain structure	4133034	correct	Structure of left parietal bone	37158682	near match			no match
<i>right pelvis</i>	Pelvic region	4044034	correct	Structure of right renal pelvis	4184440	invalid	Entire pelvis	4041832	correct
<i>(r) breast/low axilla</i>	Breast structure;Axillary region structure	4298444;4238919	correct	Axillary region structure	4238919	correct	Entire breast	4108283	correct
<i>thyroid</i>	Thyroid structure	4321375	correct	Thyroid structure	4321375	correct	Thyroid structure	4321375	correct


```


classes:
  Region:
    tree_root: true
    attributes:
      label:
        description: >-
          The name of the radiation therapy region verbatim as it appears in the text.
        range: string
      # give the model two chances to ground the region
      # some models show preference for more or less specificity
      # need to codify disambiguation - closest match?
      location:
        description: >-
          Target location of the radiation therapy region without modifiers.
          Remove modifiers like radiation technique, relative location, laterality,
          leaving just the target location.
        range: BodySite
      body_site:
        description: >-
          Specific body site or organ mentioned in the radiation therapy region.
          This should be a list of each discrete anatomical site.
          Do not use abbreviations or acronyms.
        range: BodySite
        required: false
        multivalued: true
      laterality:
        description: >-
          The laterality of the radiation therapy region, if this is mentioned.
          This should be "left", "right", "bilateral", or "na" if not specified.
          It is often specified as an abbreviation like 'lt', 'rt', (l), (r), r, l etc.
        range: Laterality
        required: false

```

```











BodySite:
  is_a: OMOPHierarchy
  id_prefixes:
    - omop
  annotations:
    annotators: 'OMOP_OWL/ohdsi_test.db'
    # body site, body organ - in order of preference
    parent_id: omop:4190005, omop:4240671
  Laterality:
    is_a: OMOPEnum
    attributes:
      concept_name:
        range: LateralityEnum
    annotations:
      meaning: concept_id
enums:
  LateralityEnum:
    permissible_values:
      left:
        meaning: omop:45883143
      right:
        meaning: omop:45881626
      bilateral:
        meaning: omop:21498852
      na:
        description: The radiation therapy region does not specify laterality

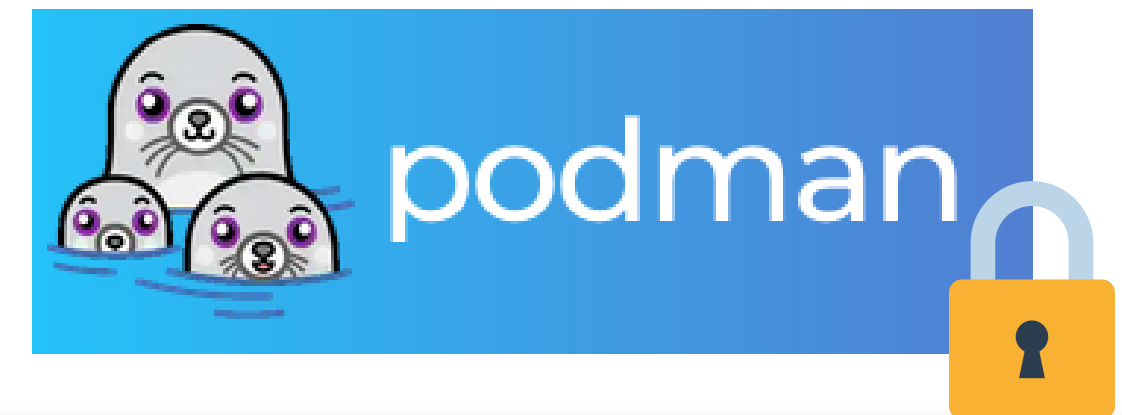
```

 **grounding_dockerised_mwe** Private


Edit Pins Watch 0

main 1 Branch 0 Tags Add file Code

 gkennos gguf dl script dc02b4f · 15 hours ago 18 Commits		
 docker	no do not try build vllm on an m4...	yesterday
 image-service-demo	rest of image functionality	last week
 inference-wrapper	no do not try build vllm on an m4...	yesterday
 omop-spires-demo	works with podman now but not ramalama yet	2 days ago
 populate_containers	gguf dl script	15 hours ago
 .dockerignore	works with podman now but not ramalama yet	2 days ago
 .gitignore	gguf dl script	15 hours ago
 DONTREADME.md	no do not try build vllm on an m4...	yesterday
 README.md	no do not try build vllm on an m4...	yesterday



README Code of conduct Contributing MIT license Security



[RamaLama](#) strives to make working with AI simple, straightforward, and familiar by using OCI containers.

Description

RamaLama is an open-source tool that simplifies the local use and serving of AI models for inference from any source through the familiar approach of containers. It allows engineers to use container-centric development patterns and benefits to extend to AI use cases.

RamaLama eliminates the need to configure the host system by instead pulling a container image specific to the GPUs discovered on the host system, and allowing you to work with various models and platforms.

SPIRES paper



OntoGPT Library



OMOP-links



Semantic SQL



Ramalama



linkML docs



Python Instructor



These slides



More detailed write-up



georgina.kennedy@unsw.edu.au



MAP-CARE: Enhancing Cross-Lingual Medical Intervention Terms Analysis Through LLM-supported Semantic Embeddings

University Hospital of Bern, Bern University, Switzerland, 26.08.2025

Hugo Guillen-Ramirez, Karen Triep, Christoph Gaudet-Blavignac, Baljit Phull, Guido Beldi, Olga Endrich




Use Case Surgical and Interventional Procedures

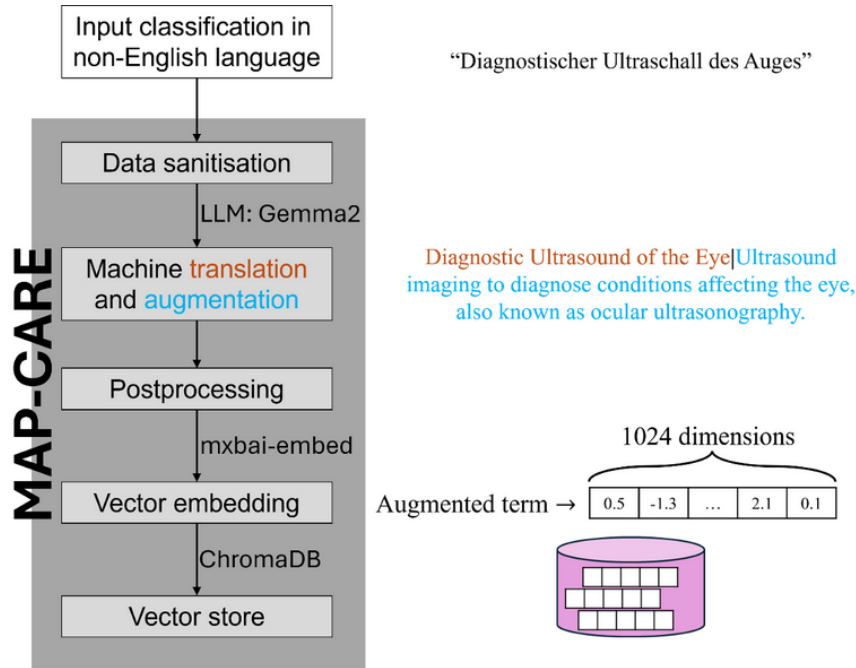
- Classifications of medical procedures are often country- or healthcare-system-specific
- Available in local language only
- Multiaxial, includes abbreviations, free-text descriptors, multiple interpretations possible
- Semantic mapping or key-word search are challenging or misleading
- Better Interoperability needed!

Use case Operation & Procedures Catalogue Switzerland & Germany

- CHOP Catalogue, Swiss development of the ICD-9 procedure catalogue
Available in German, French, and Italian
- OPS Operation Procedures Classification Germany
Available in German
- CHOP and OPS are not interoperable

Herzklappenrekonstruktion (35.J1-)		Xenograft, nahtfrei	
35.F	Klappenersatz durch Herzklappenprothese oder klappentragende Gefäßprothese	35.F1.41	Aortenklappenersatz durch selbstexpandierendes Xenograft, nahtfrei, über vollständige Sternotomie
Kod. eb.: Verwendung eines Embolie-Protektionssystems (39.E1)		35.F1.42	Aortenklappenersatz durch selbstexpandierendes Xenograft, nahtfrei, über minimalinvasive Thorakotomie (partielle obere Sternotomie, laterale Thorakotomie)
 Korrektur von Endokardklappendefekten (AV-Kanal) mit Prothese, mit Bio-Graft oder sonstige (35.S4, 35.G3, 35.73)		35.F1.43	Aortenklappenersatz durch selbstexpandierendes Xenograft, nahtfrei, endovaskulärer Zugang
Beachte: Codes mit der Bezeichnung «endovaskulärer Zugang» beinhalten die transfemorale oder transaortale kathetergesteuerten Eingriffe.		35.F1.44	Aortenklappenersatz durch selbstexpandierendes Xenograft, nahtfrei, transapikal
35.F1	Aortenklappenersatz durch Herzklappenprothese oder klappentragende Gefäßprothese	35.F1.45	Aortenklappenersatz durch selbstexpandierendes Xenograft, nahtfrei, transapikal, mit Verwendung eines perkutanen apikalen Zugangs- und Verschlusssystems
35.F1.0	Detail der Subkategorie 35.F1	35.F1.5	Aortenklappenersatz durch mechanische Prothese
35.F1.00	Aortenklappenersatz durch Herzklappenprothese oder klappentragende Gefäßprothese, n.n.bez.		
35.F1.09	Aortenklappenersatz durch		

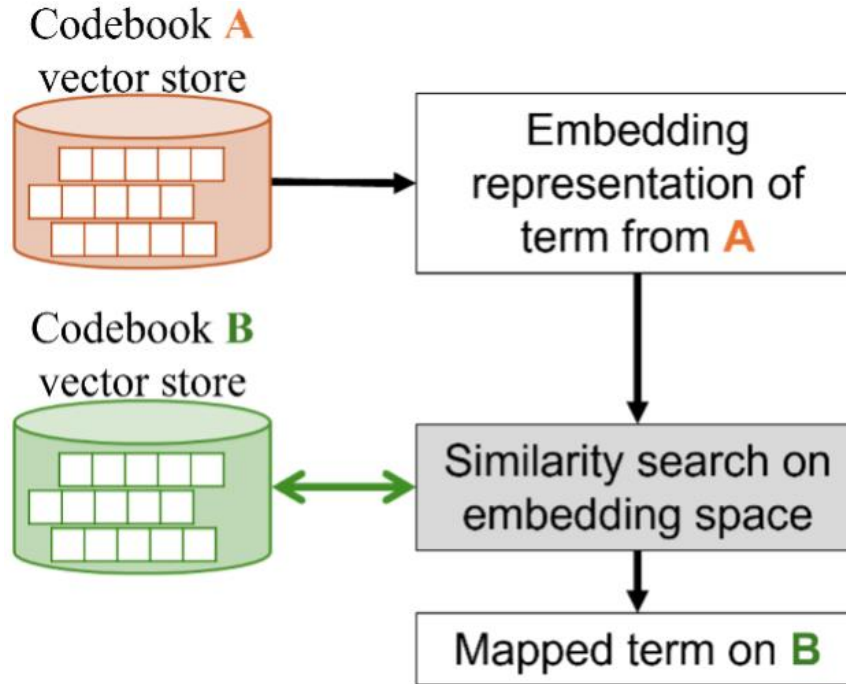
MAP-CARE



MAP-CARE workflow for multilingual classification integration

Hugo Guillen-Ramirez, Karen Triep, Christophe Gaudet-Blavignac et al.
 MAP-CARE: Enhancing Cross-Lingual Medical Intervention Terms Analysis Through LLM-supported Semantic Embeddings, 13 June 2025, PREPRINT available at Research Square
<https://doi.org/10.21203/rs.3.rs-6848278/v1>

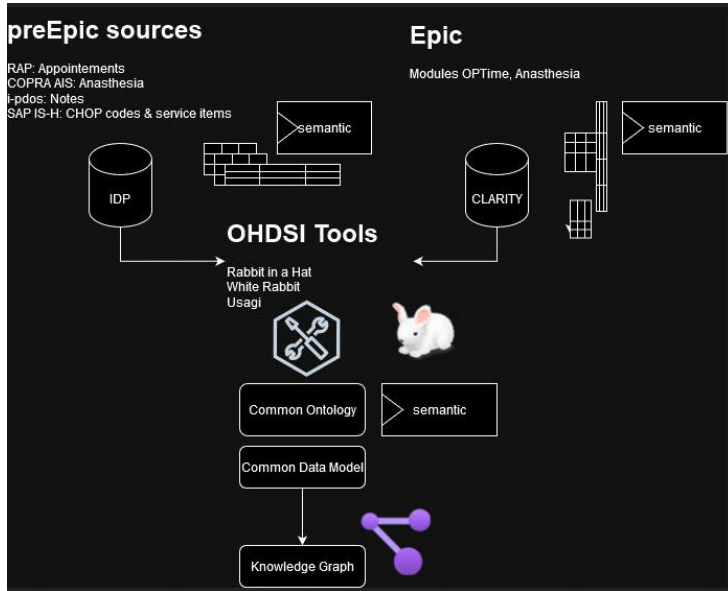
Embedding space



Loading of catalogues

Searchable in English

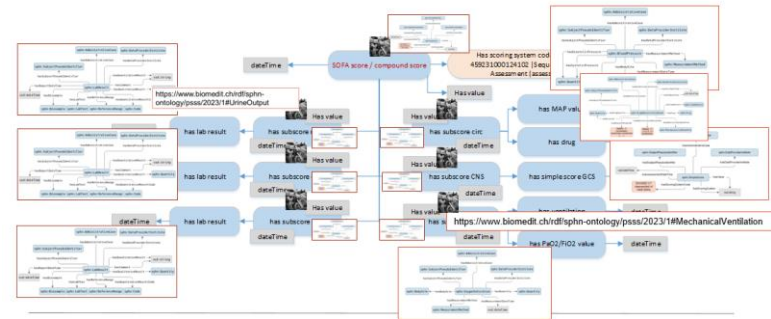
Retrieving
Swiss CHOP and OPS
Germany terms from the
embedding space



Simple mappings: Expression in SNOMED CT, LOINC
Complex mappings:

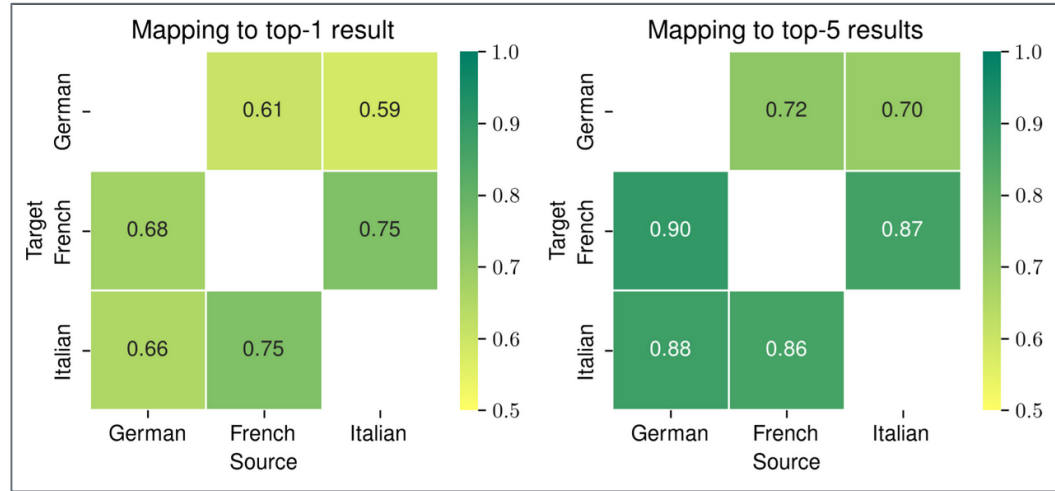
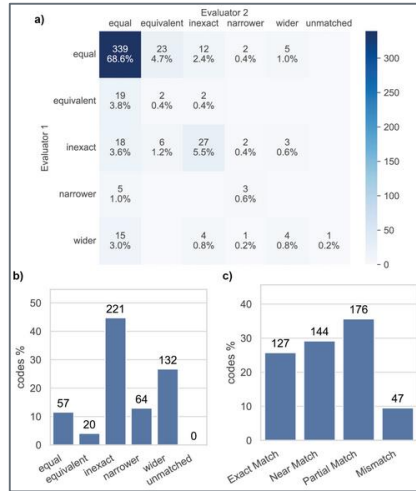
Translation of **surgical procedures** (CHOP) to **SNOMED CT**

The Clinical Data Model is used as a conceptual link between the different data realities such as preEpic & Epic Data



Results

multilingual (German, French, Italian, English) and cross-system integration (CHOP & OPS) of medical procedural data



Thank you

Insel Gruppe AG, Kommunikation und Marketing, Freiburgstrasse 18, CH-3010 Bern

A decorative footer consisting of several overlapping triangles in shades of grey, pink, green, and blue.

Using LLM to create concept sets for epidemiological studies

Joel Swerdel, Dmytro Dymshyts,
Martijn Schuemie, Anna Ostropolets,
Azza Shoaibi, Erica Voss, and Frank DeFalco

Johnson & Johnson
Innovative Medicine

Steps in the Process

1. Create a list of the suggested concept(s) and its descendants
2. Add in all the recommended concepts from PHOEBE
3. Pass the list to LLM for analysis
4. Parse through list of LLM results
5. Create a list of the concepts selected by the LLM and their descendants
6. Pass the new list to the LLM for analysis
7. Parse through list of LLM results
8. Create a final list of all the selected concepts
9. Post the concept set to ATLAS

Calling the LLM Function

- We developed a process using R and R Studio

```
executeLlmConceptCreate <- function(  
  conceptName, #the clinical condition
```

```
  originalConceptList, #one or more concepts that gives a high-level idea of the condition
```

```
  excludedConditions = "none",    #whether any health conditions should be explicitly excluded
```

```
  conceptSetName = conceptName,    #what to name the concept set when posted to ATLAS
```

```
  outputDirectory      #where to put all the artifacts from the process
```

```
)
```

Prompt (partial)

You will estimate the proportion of patients with Right ventricular failure with sepsis who also have Right heart failure, selecting a proportion category:

1. ****Rare:**** Less than 10% of patients with Right ventricular failure with sepsis have Right heart failure
2. ****Common:**** 10% to less than 90% of patients with Right ventricular failure with sepsis have Right heart failure
3. ****Very Common:**** 90% to 95% of patients with Right ventricular failure with sepsis have Right heart failure
4. ****Ubiquitous:**** greater than 95% of patients with Right ventricular failure with sepsis have Right heart failure

Current Status

- Results look promising
- Have tested in about 100 health conditions
 - Including chronic , acute , and symptomatic conditions
 - Concept sets are similar to those produced by experts
 - LLM rationale has brought additional knowledge to many conditions providing evidence for including concepts that would have not previously been included

Anamnézis

Család

tbc: neg. DM: anyai nagypapa HT: neg. Stroke: neg.

AMI: neg. Daganat: anyai nagymama (agy) Elme: apai nagyanya (vasc. dementia)

Dohányzás: neg. Alkohol: <1 E/nap

Gyerekkorban visszatérő felső légúti infekciók, otitis mediák és tüdőgyulladások voltak. 1983-ban tonsillectomia. 1987-ben hepatitis A fertőzése zajlott. 2003-ban pleuropneumonia. 2010. februárban spasztikus bronchitis miatt kezelték. Azóta ismert zsírmáj. 2011. októberben mko pneumonia. 2013. januárban jo.-i pneumonia, majd mononucleosis infectiosa fertőzés zajlott (postinfectiós neuronitis). 2011. júniusban derült ki euthyreoticus struma nodosa (15 mm-es cysta FNAB: struma nod.). 2013. február óta ismert hyperlipidemia, 2-es típusú diabetes mellitus és hyperurikemia (ugyanekkor transzamináz emelkedés miatt részletes vizsgálatok történtek neg. eredménnyel). 2013. novemberben aranyeres vérzés miatt vizsgálták. 2014. júniusban atipusos angina háttérében ISZB nem igazolódott. 2015. augusztus óta metformint szed. 2015. novemberben végtagi bőrelváltozások excisiója (hyst.: psoriasis). 2015. december óta ismert microalbuminuria (52 mg/l). 2016. októberben mko sarokcsont sarkantyú. 2017. szeptemberben HbA1c 7.9% volt. 2021. júniusban atipusos mellkasi panaszok háttérében kardiológiai ok nem igazolódott. 2022. májusban LV-SI lumboischialgia miatt fizioterápiás kezelés. 2022. augusztusban fract. prox. phal. dig. V. ped. 1.d. miatt konz. kezelés. 2023. októberben mellkas CT-n tüdőkbén enyhe fokú, enyhe kiterjedésű bronchitis - bronchiolitis, tracheat min. diszlokáló bal pajzsmirigy lebeny (benne meszesedés), steatosus hepatis látszott. 2024. januárban gastroenteritis miatt obs.

Jelen Panaszok:

Ezetimibe-t még nem kezdte el. Dereka most rendben van, de válla és talpa is panaszos. Hypoglykémia: 0/hó.

Orthostasis: neg., angina pect.: neg., szimpt. neuropathia: neg., claudicatio: neg., láb: sebzés: nincs, ADP: mko. jól tap.

Gyógyszerérzékenység: sebtapaszkó, statin (laboreltérése volt). Icterus: nem volt. Transzfúzió: nem volt.

Státusz

RR: 122/81 Hgmm, P: 86/min, Ts: 146.5 (-4) kg, Tm: 190 cm.

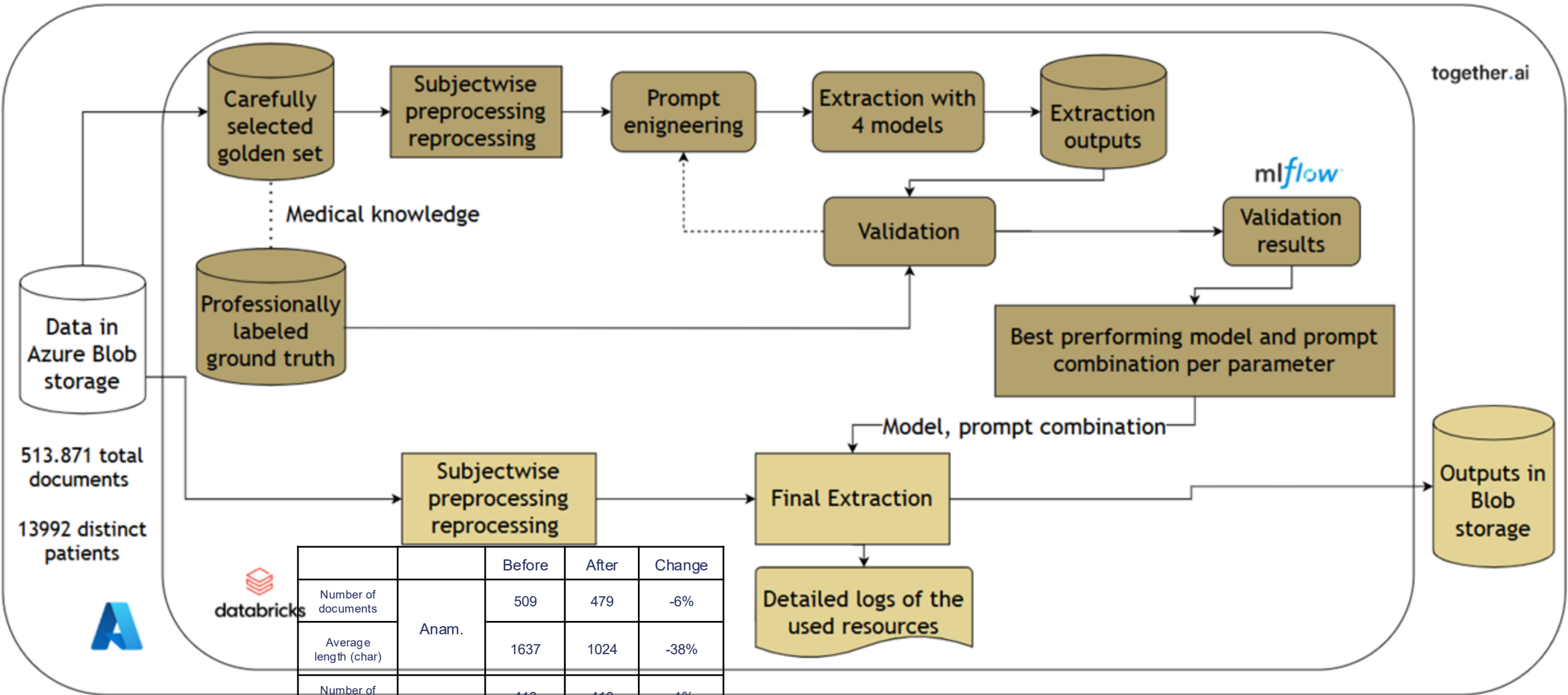
Overall goal: pilot project for developing a tool for extracting data from free text for dAlbetes (Horizon2020) project aiming to train AI model for diabetes treatment.

P S	P Diastolic	Heart Rate (BPM)	BMI (kg/m ²)	Smoking Status
122	81	86	40.6	Negative

- misspelling, abbreviations
- english/latin/hungarian
- no exact location in text
- extremely redundant
- context is important

Complete workflow

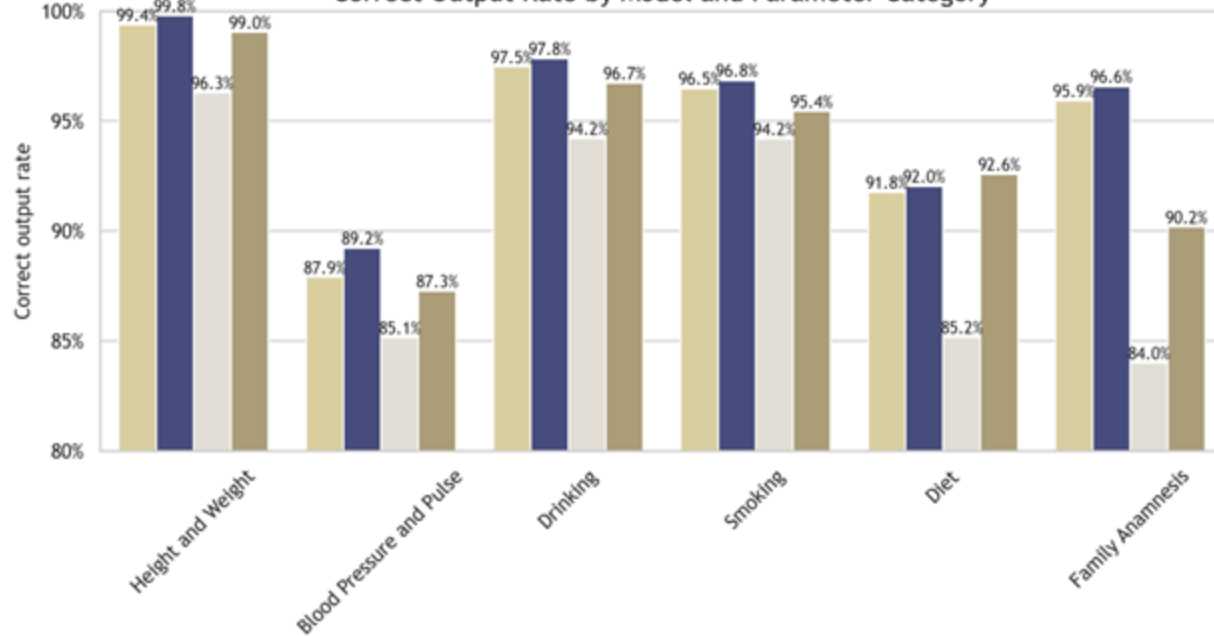
- Development and validation round
- Production round



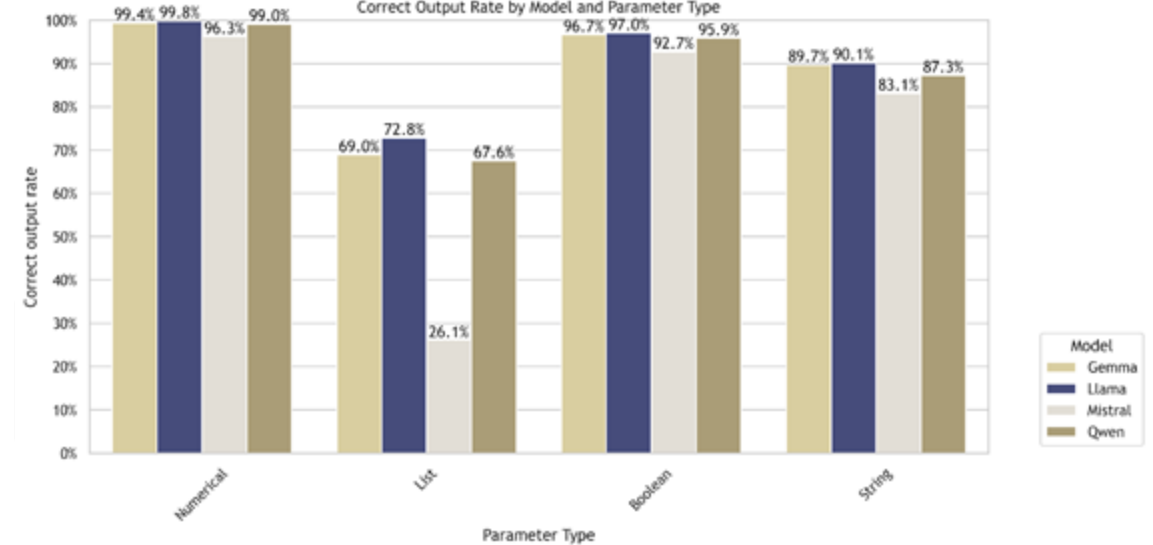
		Before	After	Change
Number of documents	Anam.	509	479	-6%
Average length (char)		1637	1024	-38%
Number of documents		413	410	-1%

Results - Aggregated correctness of the models

Correct Output Rate by Model and Parameter Category



Correct Output Rate by Model and Parameter Type



Results - Final Model Selection

Our final choice for the majority of extraction tasks was the **Qwen**, driven by its superior cost-effectiveness:

- it delivers performance comparable to Llama at approximately one-third of the inference cost
- offering long-term availability within the TogetherAI framework.

We strategically decided to use the Llama model for the specific task of extracting family medical history to ensure optimal performance.



Parameter Groups	Estimated Number of Accurate, Non-empty data
Body measurements	21267
Cardivascular parameters	236745
Family history	19308
Smoking status	28264
Drinking status	8115
Diet	12312



Automatic LOINC Document Ontology Named Entity Recognition for Clinical Note Titles

Jianlin Shi

VA Informatics and Computing Infrastructure (VINCI) Resource Center

August 26, 2025

Motivation



- Challenge: Unstandardized clinical note titles hinder secondary use and research
- Need: Standardized note classification for reliable phenotyping and cohort selection
- LOINC DO: Facilitates harmonization across EHRs and supports OMOP CDM integration

What are LOINC DO Axes?



- **LOINC Document Ontology (DO)** provides a standardized way to categorize clinical documents.
- Each document is described using five key axes:
 - **Kind of Document (KoD):**
 - The general type of document (e.g., Note, Report, Consult)
 - **Setting:**
 - Where the document was created (e.g., Inpatient, Outpatient, Emergency)
 - **Type of Service (ToS):**
 - The clinical service or activity (e.g., Surgery, Therapy, Evaluation)
 - **Role:**
 - The role of the author or intended audience (e.g., Physician, Nurse, Patient)
 - **Subject Matter Domain (SMD):**
 - The clinical specialty or domain (e.g., Cardiology, Nursing, Radiology)

Methods



- VA CDW: 1,000 annotated note titles, high-quality gold standard
- Annotation schema aligns with LOINC DO axes (KoD, Setting, ToS, Role, SMD)
- Models: BERT (fine-tuned), Qwen2.5, Tulu3 (open-source, no fine-tuning)
- Evaluation: Token-level lenient F1 score



Model	Precision	Recall	F1 Score
BERT	0.840	0.831	0.836
Qwen2.5	0.706	0.819	0.758
Tulu3	0.683	0.792	0.731

Implication for OHDSI



- NLP can automate mapping to standard vocabularies, supporting OMOP note_type_concept_id
- LLMs offer adaptability, potential for cross-institutional generalizability
- Error analysis: Acronym ambiguity, ontology limitations, need for metadata/content

Limitations & Future Directions



- LLMs not fine-tuned for the task; limited generalizability
- Only note titles used; content and metadata needed for full context
- Next steps: Integrate note content/metadata, test more models, map to exact LOINC codes, enhance ontology for OHDSI use cases

Acknowledgements



- Funding support: Put VA Data to Work for Veterans (VA ORD 24-D4V-02).
- Authors: Annie E. BOWLES^{a,b,1}, Qiwei GAN^{a,b}, Elizabeth HANCHROW^a, Scott DUVALL^{a,b}, Patrick R. ALBA^{a,b}, and Jianlin SHI^{a,b}
 - ^a *VA Salt Lake City Health Care System*
 - ^b *Department of Internal Medicine, University of Utah Medical School, Salt Lake City, UT, USA*
- Other VINCI NLP team members: Mengke Hu, Johnathon Stanley
- This publication does not represent the views of the Department of Veterans Affairs or the United States Government.

Published @ <https://pubmed.ncbi.nlm.nih.gov/40775962/>



LLM magic enablement for OHDSI

Initial goal

Try things out,
see what can be done,
have som fun.

Hecate – Semantic Search

<https://hecate.pantheon-hds.com/>

<https://github.com/OHDSI/Hecate>

≡ Hecate

i have a hard time using standard concepts

Applied filters:

standard concept: Standard

clear filters

id	code	name	class	domain	validity	concept	vocabulary	score
1616679	98068-0	Difficulty communicating in usual language	Survey	Observation	Valid	Standard	LOINC	0.558
4111964	286366005	Difficulty using language	Clinical Finding	Condition	Valid	Standard	SNOMED	0.552
4127510	288614000	Difficulty using words	Clinical Finding	Condition	Valid	Standard	SNOMED	0.538
4124818	288748008	Difficulty making self understood	Clinical Finding	Condition	Valid	Standard	SNOMED	0.538
1146950	CDM6	concept.standard_concept	Field	Metadata	Valid	Standard	CDM	0.534
4083467	247657007	Poor grasp of abstract concepts	Clinical Finding	Observation	Valid	Standard	SNOMED	0.53
37396613	716489000	Difficulty using semantics in communication	Clinical Finding	Observation	Valid	Standard	SNOMED	0.522
4127380	288579009	Difficulty communicating	Clinical Finding	Condition	Valid	Standard	SNOMED	0.519

OHDSI / Hecate

Type the search

Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

Hecate

Public

10k files

Week

Year

Star

main

1 Branch

0 Tags

Go to file

Code

About

Recent commits

update standard concept

1 commit

2 days ago

10 Commits

api

update standard concept

2 days ago

autocomplete

add autocomplete code

last month

mcpe

consolidate all config

2 days ago

ui

update standard concept

2 days ago

api-gate

add extension logic for hecate, do no longer fil...

2 days ago

LICENCE

add licence

last month

README.md

add mcg server code

last month

No description, website, or topics provided

Readme

Apache-2.0 license

Activity

Custom properties

1 star

1 watching

0 forks

Report repository

Releases

No releases published

Create a new release

Hecate

No packages published

Publish your first package

Languages

But can it do this?

Hecate Model Context Protocol Server

Could you please provide me with the standard concepts from this dutch clinical note:

"patient komt binnen met pijn aan been waarschijnlijk gebroken tijdens het voetballen, doorverwezen naar de rad"

Standard Medical Concepts:

1. Pain in lower limb

- Concept ID: 4024561
- SNOMED Code: 10601006
- Domain: Condition
- This captures "pijn aan been" (leg pain)

2. Fracture of lower leg

- Concept ID: 4185758
- SNOMED Code: 414292006
- Domain: Condition
- This represents "waarschijnlijk gebroken" (probably broken)

3. Injury whilst playing football

- Concept ID: 42689705
- SNOMED Code: 1067311000000103
- Domain: Condition
- This captures the context "tijdens het voetballen" (during football/soccer)

4. Referral to

- Concept ID: 4208328
- SNOMED Code: 440379008
- Domain: Observation
- This represents "doorverwezen naar" (referred to)

5. Radiology service

- Concept ID: 4150863
- SNOMED Code: 310125001
- Domain: Observation
- This captures "rad" (radiology department)

These standard concepts provide a structured, coded representation of the key clinical information in the note: a patient presenting with leg pain, suspected fracture from a football injury, who was referred to radiology for further evaluation.

Hecate – Model Context Protocol Server

How can I identify iodine contrast or related procedures using only standard concepts?

Key Standard Concepts for Iodine Contrast

Contrast Media (Device Domain)

- **Iodinated contrast media** (4141354) - Primary concept for iodine-based contrast
- **Radiographic iodinated contrast medium** (4139488) - Specific for radiographic use
- **X-ray contrast media** (4146659, 4168823) - General X-ray contrast concepts
- **Water soluble contrast medium** (4130246) - Most iodine contrasts are water-soluble

Specific Iodine Formulations (Drug Domain)

- **iodine** (1780601) - Base ingredient concept
- Various concentration formulations like:
 - **iodine 410 MG/ML** (19100634)
 - **iodine Injectable Solution** (40054302)
 - Brand-specific concepts like **Iodine 350 MG/ML [Omnipaque]** (43161033)

Contrast-Enhanced Procedures (Procedure Domain)

CT Procedures with IV Contrast:

- **CT with intravenous contrast** (4013967) - General concept
- Specific anatomical regions:
 - **CT Head W contrast IV** (3002086)
 - **CT Chest W contrast IV** (3013610)
 - **CT Abdomen W contrast IV** (21492176)
 - **CT Pelvis W contrast IV** concepts
 - **CT angiography with contrast** (4303663)

MRI with Contrast:

- **MRI with contrast** (4198856) - General MRI contrast concept

General Imaging Procedures:

- **Diagnostic radiography with contrast media** (4100357)
- **Radiology with contrast** (44791413)
- **Imaging of liver with contrast** (1242802)

Contrast Administration Procedures

- **Injection of radicontrast material** (4140276) - Primary administration concept
- **Diagnostic radiography with contrast media by injection** (4180015)
- **Injection of contrast media for radiography by catheter** (4031987)

Search Strategy for Identifying Iodine Contrast Exposure

1. **Device Domain Search:** Look for patients with exposure to concepts 4141354, 4139488, or 4146659
2. **Procedure Domain Search:**
 - Search for any procedure containing "contrast IV" or "with contrast"
 - Use concept 4013967 for general CT with IV contrast
 - Use concept 4140276 for contrast injection procedures
3. **Drug Domain Search:** Look for iodine-containing medications (concept 1780601 and related)
4. **Combined Approach:** Use UNION queries across domains to capture all potential exposures

The most reliable approach is to focus on the high-level standard concepts like **"Iodinated contrast media" (4141354)** and **"CT with intravenous contrast" (4013967)**, then expand to more specific anatomical or procedural variants as needed for your research question.

Another tool...

Under the hood

MCP SERVER

API

DATABASE

EMBEDDINGS

EMBEDDING MODEL

LLM FUN

AGENTS

CONCEPT SET CREATION

CONCEPT EXTRACTION

AMAZING PROMPTS

MAPPING TOOLS

R PACKAGES

BEUATIFUL USER INTERFACES

Under the hood

MCP SERVER

<https://hecate.pantheon-hds.com/mcp/sse>

API

<https://hecate.pantheon-hds.com/openapi/#/>

DATABASE

PostgreSQL

Qdrant

EMBEDDINGS

<https://github.com/OHDSI/EmbeddingsLibrary>

EMBEDDING MODEL

open source?

proprietary

Can we move towards consolidation?
Only with some help.

Feedback
Suggestions
Input
Contributions
Constructive criticism

Anything to improve quality and create valuable community
capability

<https://github.com/OHDSI/EmbeddingsLibrary>
<https://github.com/OHDSI/Hecate>
rowan.parry@live.com

Let's meet each others Iurii (Yurri) Iurchenko



Domain Expertise in Healthcare & OMOP

- **4 years** in Healthcare Informatics, 2 years with OMOP CDM
- 2 years in Scientific Research & Biochemistry
- **Fellow**, American College of Health Data Management
- **OMOP training from EHDEN Academy**

Leadership & Product Management

- 3+ years in leadership and management roles
- Certified Professional Scrum Master™ I (PSM I)
- Speaker and Technical Trainer (audiences of 10–100 people)
- Mentored peers, mentor on ADPList (30+ sessions)

Data Analytics Expertise

- **12 years** in Big Data Engineering & Analytics
- Microsoft Certifications: DP-203, DP-600, DP-700

AI & Software Engineering Expertise

- 5 years building data solutions, applications, and pipelines
- 2+ years building AI tools; hackathon participation

AutoCohort

Empower the Healthcare Researchers with a Second AI Brain



by Iurii (Yurri) Iurchenko

Application Creator, OHDSI Open Source Contributor



Product

AutoCohort in Action

- Ask a question in plain English
- Approve the list of concepts
- AutoCohort builds and checks cohorts for you
- Run the analysis and save code and checks

1



Ask a question in plain English

2



Approve the list of concepts

3



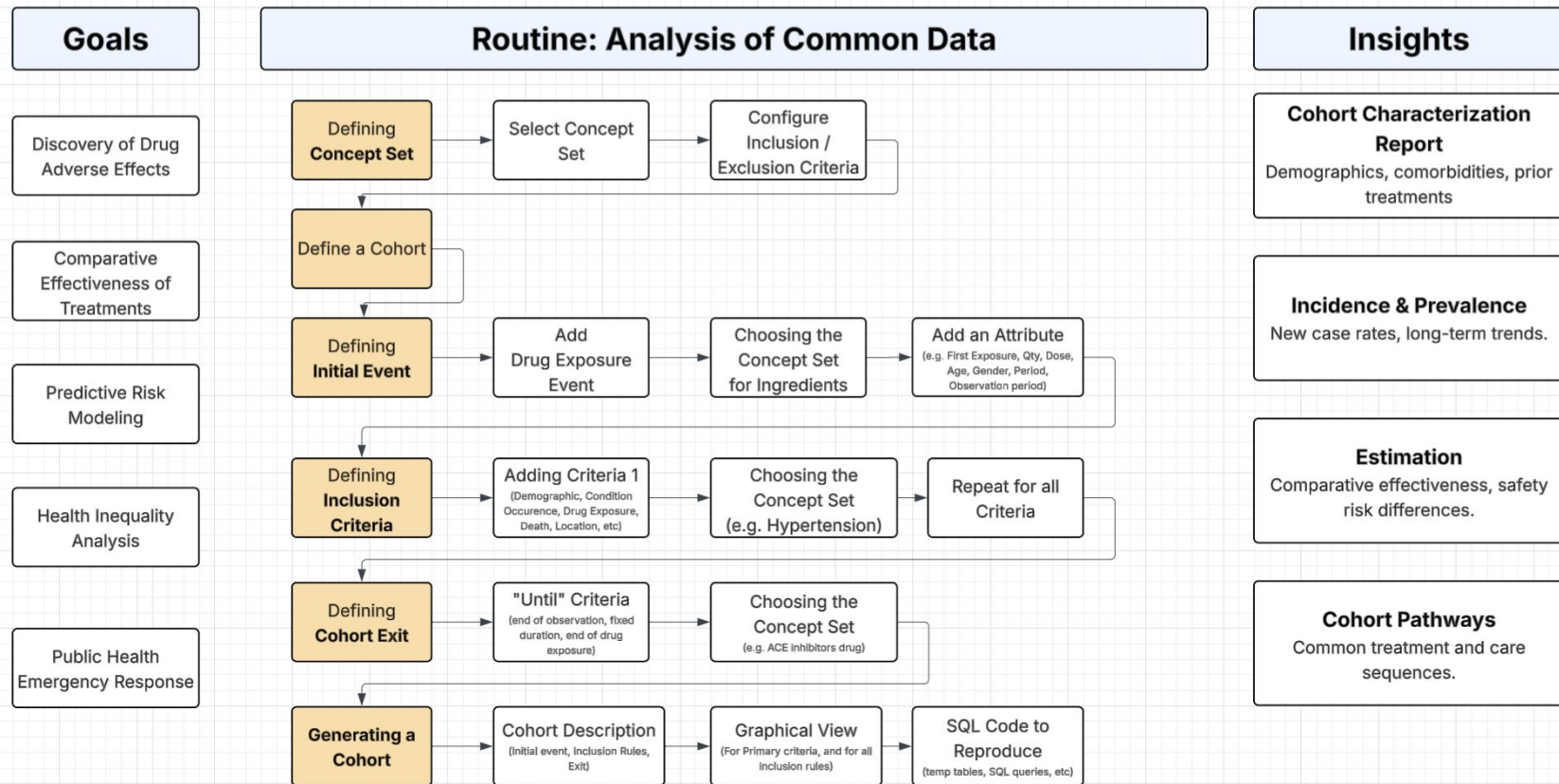
AutoCohorts builds and checks cohorts for you

4



Run the analysis and save code and checks

Life of a Researcher



New Life of a Researcher

Goals

Discovery of Drug
Adverse Effects

Comparative
Effectiveness of
Treatments

Predictive Risk
Modeling

Health Inequality
Analysis

Public Health
Emergency Response

Routine: Analysis of Common Data

Routine work - automated.

1. Goals set through simple, plain-language requests.
2. Results can be validated for maximum accuracy.
3. If anything is missing, the system will prompt you.

Insights

Cohort Characterization Report

Demographics, comorbidities, prior
treatments

Incidence & Prevalence

New case rates, long-term trends.

Estimation

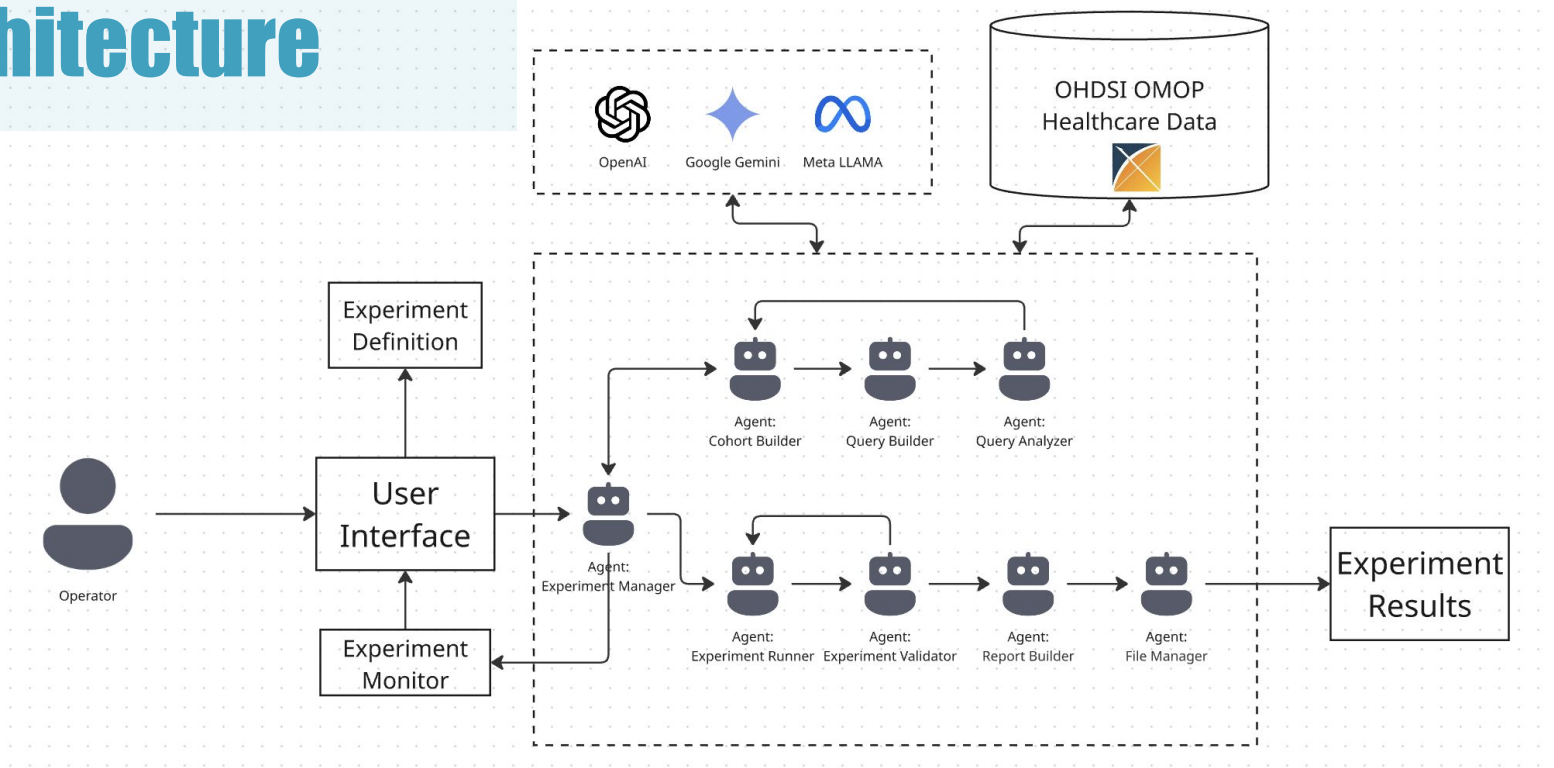
Comparative effectiveness, safety
risk differences.

Cohort Pathways

Common treatment and care
sequences.

Key Features & Functionalities:

Solutions Architecture



Key Features & Functionalities:

Technologies & Code Snippets

Database Connectors:

```
1 from sqlalchemy import create_engine
2 import os
3
4 def get_engine(db_type: str):
5     conns = {
6         "postgres": os.getenv("POSTGRES_CONN"),
7         "sqlserver": os.getenv("SQLSERVER_CONN"),
8         "oracle": os.getenv("ORACLE_CONN")
9     }
10     return create_engine(conns[db_type])
```

Example of a Query Tool:

```
1 from langchain.tools import tool
2 from db import get_engine
3
4 @tool
5 def query_postgres(sql: str) -> str:
6     engine = get_engine("postgres")
7     with engine.connect() as conn:
8         result = conn.execute(sql).fetchall()
9     return str(result)
```

Example of a Query Builder Agent:

```
1 from langchain.prompts import PromptTemplate
2
3 sql_prompt = PromptTemplate(
4     input_variables=["question"],
5     template="""
6 You are an expert healthcare data engineer and researcher working
7 with the OMOP Common Data Model (CDM) used by the OHDSI community.
8 Your task is to convert the following plain-language clinical question
9 into a SQL query that extracts a cohort of patients from an OMOP CDM-compliant database.
10
11 The query must:
12 - Use standard OMOP CDM table structure (e.g., person, condition_occurrence, drug_exposure, etc.)
13 - Include appropriate JOINS, WHERE clauses, concept IDs, and date logic where applicable
14 - Return distinct person_ids who satisfy the inclusion criteria
15 - Output MUST be ONLY a single valid SQL query
16 Plain-language question: {question}
17 """)
```

User Interface

OHDSI AutoCohort

Dashboard

Concept Sets

Cohort Builder

AI-powered Evidence
Builder

Monitor

Configuration

OHDSI AutoCohort

An intelligent automation platform that transforms how researchers interact with OMOP
Common Data Model, reducing manual effort while ensuring quality and compliance with
OHDSI standards.

OHDSI Collaborative

OMOP CDM 6.0+

AI-Powered

Desktop Application

Our Mission

Helping Healthcare Researchers with an AI-powered Second Brain That Automates their Workflow.

Key Features & Capabilities



OMOP Concept Management

AI-powered concept discovery and validation with comprehensive OMOP CDM integration



Automated Cohort Building

4-step workflow from concept selection to SQL generation and validation



AI-powered Evidence Building

Intelligent experiments that handle complex research workflows autonomously



Real-time Monitoring

Track automation progress with detailed execution monitoring and error handling



Quality Assurance

Built-in validation, quality checks, and compliance with OHDSI standards



Workflow Integration

Seamless integration with existing OHDSI tools and research pipelines

AI Assistant

Ask me anything about OHDSI workflows



Please describe the goal you want to achieve, such as building a concept set, defining a cohort, or starting to generate evidence.

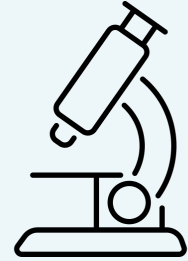
06:57 PM



Working UI Demo: May Be Tested Here:

1. Open your internet browser, preferably Chrome, and enter the url: <https://auto-cohort.figma.site/>
2. Test the interface
3. Send your feedback to the email:
4iurchenko@ieee.org

Disclaimer: This is solely a visual prototype to test user flow and interface implementation. For the functional product, please email me at **4iurchenko@ieee.org**



Appendix

Key Feature #1

AI-Powered Concept Sets Builder

OHDSI AutoCohort

Dashboard

Concept Sets

Cohort Builder

AI-powered Evidence
Builder

Monitor

Configuration

Concept Sets

Manage and organize medical concept definitions for your cohorts.

+ New Concept Set

Find Concept Sets

Input the description of the concept set using a plain language

Generate using AI

Browse OMOP Concepts

Search and browse concepts from the OMOP Common Data Model vocabulary.

Search by concept name, ID, or code...

All Domains

ID: 201826 Code: 44054006 SNOMED Condition Standard
Type 2 diabetes mellitus
Class: Clinical Finding

ID: 443238 Code: 38341003 SNOMED Condition Standard
Hypertensive disorder
Class: Clinical Finding

ID: 316866 Code: 398254007 SNOMED Condition Standard
Hypertensive disorder, systemic arterial
Class: Clinical Finding

ID: 134057 Code: 57054005 SNOMED Condition Standard
Acute myocardial infarction
Class: Clinical Finding

AI Assistant

Ask me anything about OHDSI workflows

Please describe the goal you want to achieve, such as building a concept set, defining a cohort, or starting to generate evidence.
06:57 PM

Ask me anything...

Key Feature #2

Automated Cohort Builder & Validator

OHDSI AutoCohort

Dashboard

Concept Sets

Cohort Builder

AI-powered Evidence Builder

Monitor

Configuration

Cohort Builder

Define patient populations using clinical criteria and concept sets.

Step 1
Concept Sets

Step 2
Criteria

Step 3
SQL

Step 4
Diagnostics

Cohort Information

Basic details about your cohort definition

Cohort Name

Enter cohort name...

Description

Describe the purpose and criteria of this cohort...

Select Concept Sets

Choose the concept sets that will be used to define your cohort.

☐ Type 2 Diabetes Mellitus 45 concepts
ICD-10 and SNOMED codes for Type 2 diabetes

☐ Hypertension 32 concepts
Blood pressure related conditions

AI Assistant

Ask me anything about OHDSI workflows

Please describe the goal you want to achieve, such as building a concept set, defining a cohort, or starting to generate evidence.
06:57 PM

Ask me anything...

Key Feature #3

Automated Evidence Builder

OHDSI AutoCohort

Dashboard

Concept Sets

Cohort Builder

AI-powered Evidence Builder

Monitor

Configuration

AI-powered Evidence Builder

Create and configure AI-powered automation experiments for your OHDSI workflows.

Cohort Definition

Select an existing cohort or define a new one with custom criteria and outcome measures

Select Existing Cohort

Define New Cohort

Select Cohort

Choose an existing cohort...

Outcome Definition

Define the clinical outcomes and time parameters for tracking patient events in this experiment.

Events or Conditions to Track

Select outcomes to track...

Select multiple events to track as outcomes for this experiment

Outcome Measurement Start Point

Outcome Measurement End Point

Cohort entry date

1 year after start

☐ Apply specific time window constraints

Experiment Settings

Configure the automation experiment behavior

Experiment Name

My Cohort Experiment

Execution Mode

Manual Trigger

Priority Level

Medium

- ☒ Email notifications
- ☒ Detailed logging
- ☐ Auto-retry on failure

Configuration Status

- ☐ Cohort Defined
- ☐ Outcomes Selected
- ☒ Output Path Set
- ☒ Experiment Configuration

AI Assistant

Ask me anything about OHDSI workflows



Please describe the goal you want to achieve, such as building a concept set, defining a cohort, or starting to generate evidence.

06:57 PM

Ask me anything...





Leveraging LLM to populate OMOP Oncology CDM from the EHR

Subin Kim

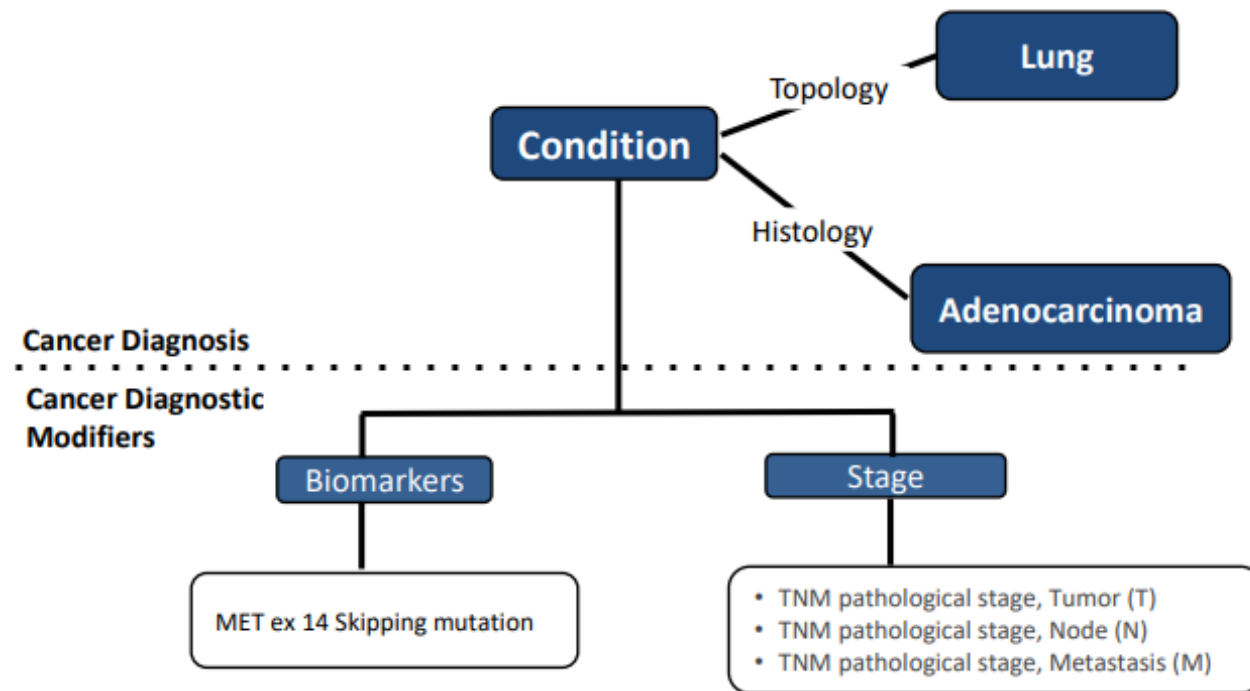


Challenges of EHR for Oncology CDM

- Cancer-specific data is **unstructured** in EHR
 - Challenges to standardize clinical data into Oncology CDM

Cancer Modifiers

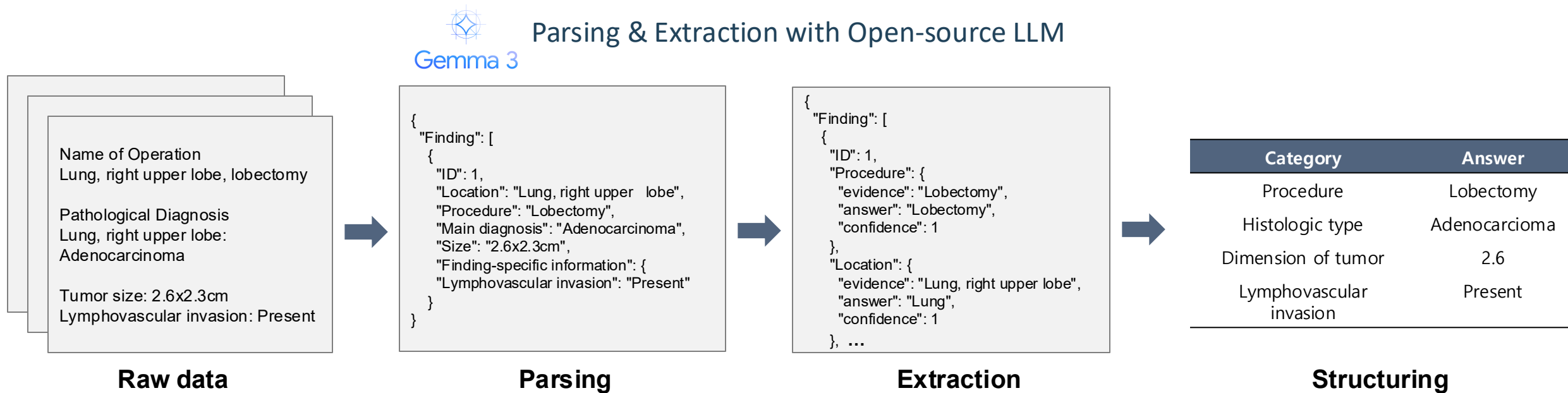
- Staging/Grading
- Topography
- Histological pattern
- Dimension
- Extension/Invasion
- Metastasis
- Margin
- Biomarker





NLP pipeline

- We are developing strategy to extract the cancer information from pathology report using state-of-the-art LLM





NLP pipeline

- **Step 1: Parsing**

- Classify the entire clinical text into four domains: Finding, Lymph node, Biomarker, Others

Name of Operation
Lung, right upper lobe, lobectomy

Pathological Diagnosis
Lung, right upper lobe:
Adenocarcinoma

Tumor size: 2.6x2.3cm
Lymphovascular invasion: Present

Raw data



```
{
  "Finding": [
    {
      "ID": 1,
      "Location": "Lung, right upper lobe",
      "Procedure": "Lobectomy",
      "Main diagnosis": "Adenocarcinoma",
      "Size": "2.6x2.3cm",
      "Finding-specific information": {
        "Lymphovascular invasion": "Present"
      }
    }
  ]
}
```

Parsing



```
{
  "Finding": [
    {
      "ID": 1,
      "Procedure": {
        "evidence": "Lobectomy",
        "answer": "Lobectomy",
        "confidence": 1
      },
      "Location": {
        "evidence": "Lung, right upper lobe",
        "answer": "Lung",
        "confidence": 1
      }
    }, ...
  ]
}
```

Extraction



Category	Answer
Procedure	Lobectomy
Histologic type	Adenocarcioma
Dimension of tumor	2.6
Lymphovascular invasion	Present

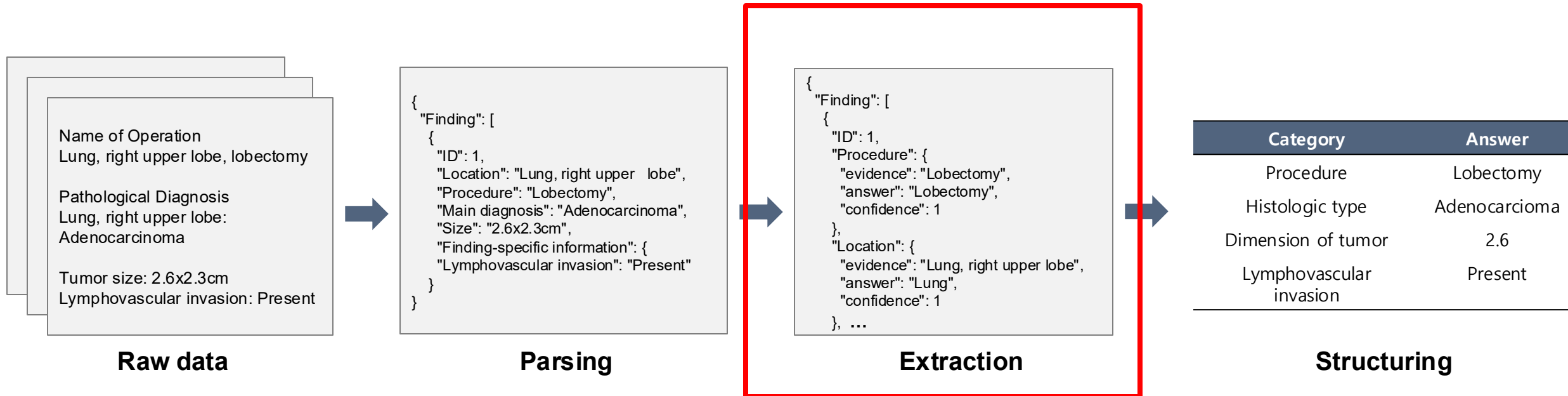
Structuring



NLP pipeline

- **Step 2: Extraction**


- Extract cancer-specific modifiers and convert into JSON format





NLP pipeline

- **CAP protocols** are used to define the variables that need to be extracted

 COLLEGE of AMERICAN
PATHOLOGISTS

Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Colon and Rectum

Version: 4.3.1.0
Protocol Posting Date: June 2024
CAP Laboratory Accreditation Program Protocol Required Use Date: September 2024
The changes included in this current protocol version do not affect the prior accreditation date.
For accreditation purposes, this protocol should be used for the following procedures AND tumor types:

Procedure	Description
Colectomy	Includes specimens designated total, partial, or segmental resection
Rectal Resection	Includes specimens designated low anterior resection or abdominoperineal resection

Tumor Type	Description
Carcinoma	Invasive carcinomas including small cell and large cell (poorly differentiated) neuroendocrine carcinoma

This protocol is NOT required for accreditation purposes for the following:


Procedure
Primary resection specimen with no residual cancer (e.g., following neoadjuvant therapy)
Cytologic specimens

The following should NOT be reported using this protocol:

Procedure
Excisional biopsy (polypectomy)(consider the Colon Excisional Biopsy protocol)
Endoscopic mucosal resection
Endoscopic mucosal dissection
Transanal disk excision

Tumor Type
Well-differentiated neuroendocrine tumors (consider the Colorectal NET protocol)
Lymphoma (consider the Precursor and Mature Lymphoid Malignancies protocol)
Sarcoma (consider the Soft Tissue protocol)

Authors
Dhanpat Jain, MD*; William V. Chopp, MD*; Rondell P. Graham, MBBS*; Yue Xue, MD, PhD*.
With guidance from the CAP Cancer and CAP Pathology Electronic Reporting Committees.
* Denotes primary author.

 COLLEGE of AMERICAN
PATHOLOGISTS

Protocol for the Examination of Resection Specimens from Patients with Primary Non-Small Cell Carcinoma, Small Cell Carcinoma, or Carcinoid Tumor of the Lung

Version: 5.0.0.0
Protocol Posting Date: December 2024
CAP Laboratory Accreditation Program Protocol Required Use Date: September 2025
The changes included in this current protocol version affect accreditation requirements. The new deadline for implementing this protocol version is reflected in the above accreditation date.

For accreditation purposes, this protocol should be used for the following procedures AND tumor types:

Procedure	Description
Resection	Includes pneumonectomy, lobectomy, segmentectomy, and wedge resection

Tumor Type	Description
Carcinoma	Includes non-small cell carcinoma, small cell carcinoma, and carcinoid tumor of the lung


This protocol is NOT required for accreditation purposes for the following:

Procedure
Biopsy
Primary resection specimen with no residual cancer (e.g., following neoadjuvant therapy)
Cytologic specimens

The following tumor types should NOT be reported using this protocol:

Tumor Type
Mesothelioma (consider the Diffuse Pleural Mesothelioma protocol)
Lymphoma (consider the Precursor and Mature Lymphoid Malignancies protocol)
Sarcoma (consider the Soft Tissue protocol)

Version Contributors
Cancer Committee Authors: Frank Schneider, MD*, Kirtae Raparia, MD, FCAP*
Other Expert Contributors: Kelly J. Butnor, MD, Mary Beth Beasley, MD, Sanja Dacic, MD, PhD
* Denotes primary author.

 COLLEGE of AMERICAN
PATHOLOGISTS

Protocol for the Examination of Resection Specimens from Patients with Invasive Carcinoma of the Breast

Version: 4.10.0.0
Protocol Posting Date: June 2024
CAP Laboratory Accreditation Program Protocol Required Use Date: March 2025
The changes included in this current protocol version affect accreditation requirements. The new deadline for implementing this protocol version is reflected in the above accreditation date.
For accreditation purposes, this protocol should be used for the following procedures AND tumor types:

Procedure	Description
Excision less than total mastectomy	Includes specimens designated excision, segmental resection, lumpectomy, quadrantectomy, and segmental or partial mastectomy, with or without axillary contents
Total Mastectomy	Includes skin-sparing and nipple-sparing mastectomy, with or without axillary contents

Tumor Type	Description
Invasive breast carcinoma of any type, with or without ductal carcinoma in situ (DCIS)	Includes invasive and microinvasive carcinomas

This protocol is NOT required for accreditation purposes for the following:

Procedure
Needle or skin biopsy
Primary resection specimen with no residual cancer (e.g., following neoadjuvant therapy)
Additional excision performed after the definitive resection (e.g., re-excision of surgical margins)
Cytologic specimens

The following tumor types should NOT be reported using this protocol:

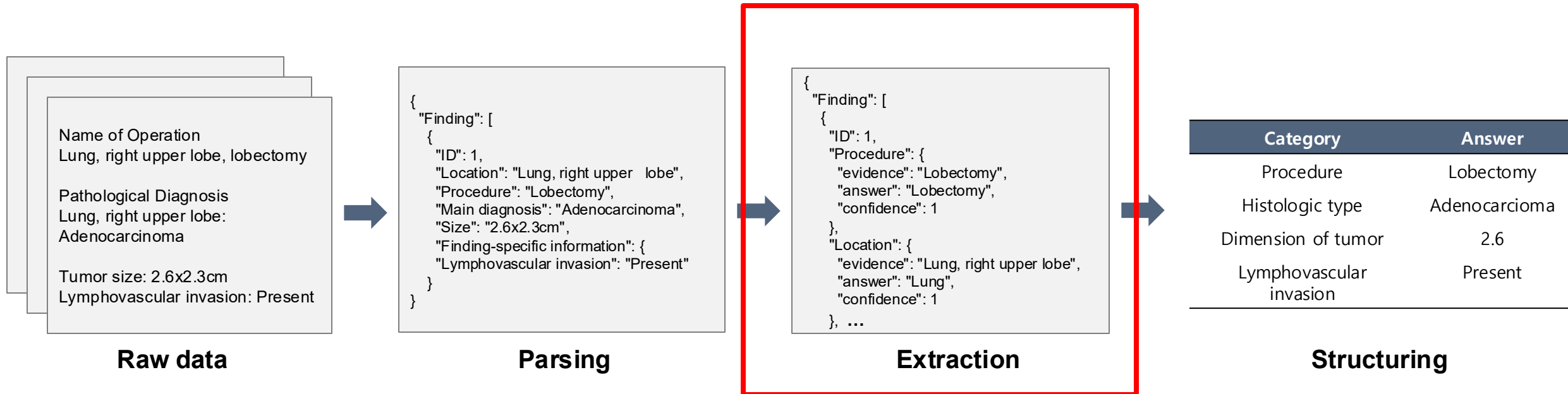
Tumor Type
Ductal carcinoma in situ without invasive carcinoma (consider the Breast DCIS Resection protocol)
Paget disease of the nipple without invasive carcinoma (consider the Breast DCIS Resection protocol)
Incapaculated or solid papillary carcinoma without invasion (consider the Breast DCIS Resection protocol)
Phyllodes tumor (consider the Phyllodes tumor protocol)
Lymphoma (consider the Precursor and Mature Lymphoid Malignancies protocol)
Sarcoma (consider the Soft Tissue protocol)



NLP pipeline

- **Step 2: Extraction**

- Extract cancer-specific modifiers and convert into JSON format

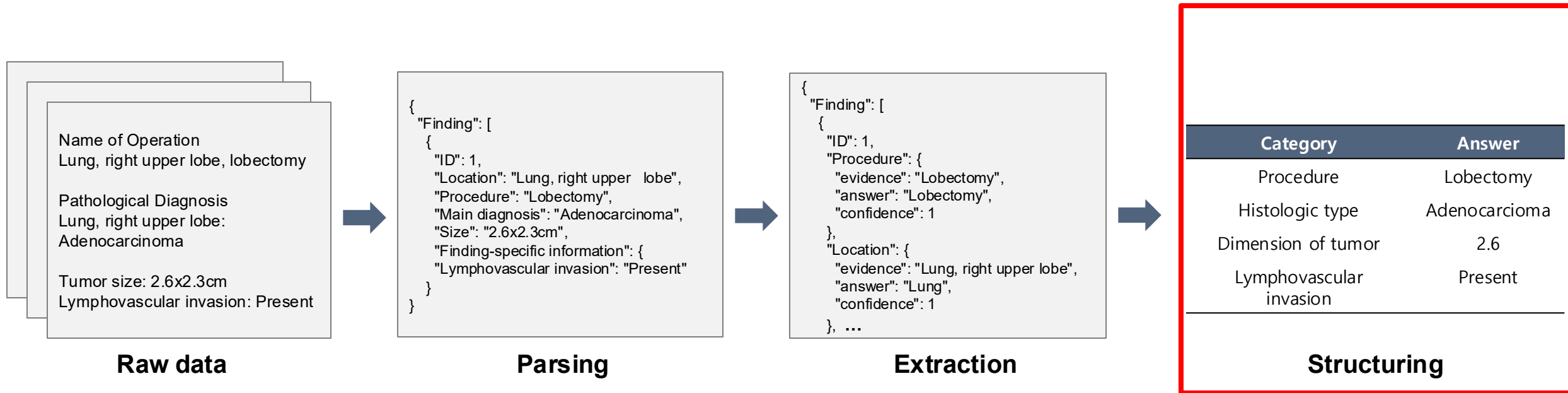




NLP pipeline

- **Step 3: Structuring**

- Convert data from JSON format to a structured tabular format





NLP pipeline

- Preliminary Result

Type	No. of reports	No. of category	Accuracy (%)	Precision (%)	Recall (%)
Colorectum	100	1,637	98.4	98.8	99.6
Breast	100	2,614	96.5	96.7	99.9
Lung	100	1,528	93.8	94.1	99.7



Mapping extracted values to OMOP CDM

Extracted data will be **converted to OMOP CDM**

Category	Answer
Procedure	Lobectomy
Histologic type	Adenocarcinoma
Dimension of tumor	2.6
Lymphovascular invasion	Present

Structuring



Category	Answer	Concept ID
Procedure	Lobectomy	4054047
Histologic type	Adenocarcinoma	37152526
Dimension of tumor	2.6	36768664
Lymphovascular invasion	Present	36768891

Standardization



Mapping extracted values to OMOP CDM

- Determine vocabulary for each cancer modifier

Category	Domain	Vocabulary	Category	Domain	Vocabulary
Diagnosis	Condition	ICD-O-3	Resection margin	Measurement	Cancer Modifier
Topography	Measurement	Cancer Modifier	Lymph node metastasis	Measurement	SNOMED
	Measurement	Cancer Modifier		Observation	SNOMED
Histology	Observation	SNOMED	Biomarker	Measurement	OMOP Genomic
Grade	Observation	SNOMED			LOINC
Dimension	Measurement	Cancer Modifier			NAACCR
Invasion	Measurement	Cancer Modifier	T stage	Measurement	Cancer Modifier
			N stage	Measurement	Cancer Modifier

Category	source_value_1	measurement_concept_id			
		domain_id	vocabulary_1	class_id_1	concept_id
Dimension	<the greatest dimmension of tumor in centimeters (cm)>	Measurement	Cancer Modifier	Dimension	36768255
Depth of invasion	Intramucosal carcinoma (in situ)	Measurement	Cancer Modifier	Histopattern	36769623
Depth of invasion	Invades submucosa	Measurement	Cancer Modifier	Extension/Invasion	36768886
Depth of invasion	Invades into muscularis propria	Measurement	Cancer Modifier	Extension/Invasion	36769076
Depth of invasion	Invades through muscularis propria into the pericolonc or perirectal tissue	Measurement	Cancer Modifier	Extension/Invasion	36769648
Depth of invasion	Invades through the visceral peritoneum	Measurement	Cancer Modifier	Extension/Invasion	36769563
Depth of invasion	Directly invades or adheres to adjacent structures	Measurement	Cancer Modifier	Extension/Invasion	36770430
Resection margin	Free from carcinoma	Measurement	Cancer Modifier	Margin	36770153
Resection margin	Involved by carcinoma	Measurement	Cancer Modifier	Margin	36768316
Perineural invasion	Present	Measurement	Cancer Modifier	Extension/Invasion	36768846



1	PROCEDURE_OCCURRENCE
PK	procedure_occurrence_id
FK	person_id
FK	procedure_concept_id
	procedure_date
	procedure_datetime
	procedure_end_date
	procedure_end_datetime
FK	procedure_type_concept_id
FK	modifier_concept_id
	quantity
FK	provider_id
FK	visit_occurrence_id
FK	visit_detail_id
	procedure_source_value
FK	procedure_source_concept_id
	modifier_source_value

2	NOTE
PK	note_id
FK	person_id
	note_date
	note_datetime
FK	note_type_concept_id
FK	note_class_concept_id
	note_title
	note_text
FK	encoding_concept_id
FK	language_concept_id
FK	provider_id
FK	visit_occurrence_id
FK	visit_detail_id
	note_source_value
	note_event_id
FK	note_event_field_concept_id

4	NOTE_NLP
PK	note_nlp_id
FK	note_id
FK	section_concept_id
	snippet
	"offset"
	lexical_variant
FK	note_nlp_concept_id
FK	note_nlp_source_concept_id
	nlp_system
	nlp_date
	nlp_datetime
	term_exists
	term_temporal
	term_modifiers

5	MEASUREMENT
	measurement_id
FK	person_id
FK	measurement_concept_id
	measurement_date
	measurement_datetime
	measurement_time
FK	measurement_type_concept_id
FK	operator_concept_id
	value_as_number
FK	value_as_concept_id
FK	unit_concept_id
	range_low
	range_high
FK	provider_id
FK	visit_occurrence_id
FK	visit_detail_id
	measurement_source_value
FK	measurement_source_concept_id
	unit_source_value
FK	unit_source_concept_id
	value_source_value
	measurement_event_id
FK	meas_event_field_concept_id

3		Prompt	
AlsUnitNo	PthoNo	description	Domain
2700517	SS1932958	<p>Colon, sigmoid, laparoscopic anterior resection</p> <p>***Pathological Diagnosis***</p> <p>Status post endoscopic mucosal dissection (see S019-27461)</p> <p>Main diagnosis: Adenocarcinoma, moderately differentiated, residual, microscopic</p> <p>Location: Sigmoid colon</p> <p>Size: 0.2x0.2cm</p> <p>Depth of invasion: Inadequate, through the muscularis propria into pericolorectal tissues (pT3)</p> <p>Reaction margin</p> <p>Proximal: Free of carcinoma (safety margin: 5.0cm)</p> <p>Distal: Free of carcinoma (safety margin: 5.0cm)</p> <p>Circumferential: Free of carcinoma (safety margin: 1.0cm)</p> <p>Separately sent, proximal and distal: Free of carcinoma</p> <p>Lymph node, regional (L7): Metastasis in 3 out of 7 regional lymph nodes (pN1a)</p> <p>Lymphovascular invasion: Not identified</p> <p>Perineural invasion: Not identified</p> <p>[Additional Report]</p> <p>***Commentary***</p>	Finding
			General
			Lymph Node
ID	Category	Values	Details
1	Location	answer: Sigmoid colon evidence: Sigmoid colon confidence: 1	
1	Histologic type	answer: Adenocarcinoma, not otherwise specified evidence: Adenocarcinoma, moderately differentiated, residual, microscopic confidence: 1	
1	Dimension of tumor	answer: 0.2 evidence: Size: 0.2x0.2cm confidence: 1	
1	Lymphovascular invasion	answer: Not identified evidence: Lymphovascular invasion: Not identified confidence: 1	
1	KRAS mutation	answer: Mutation detected: KRAS G12S (GGT>AGT) Mutation evidence: KRAS mutation (Pyrosequencing): KRAS G12S (GGT>AGT) Mutation confidence: 1	
1	Location	answer: Regional, NOS evidence: Location: Regional confidence: 1	
1	Number of metastasis node	answer: 1 evidence: Number of metastasis node: 1 confidence: 1	
1	Number of examined node	answer: 7 evidence: Number of examined node: 7 confidence: 1	

NLP (32858)

Frameworks for Trustworthy and Explainable Use of LLMs in Healthcare

- Developed a confidence-linked and uncertainty-based staged framework for phenotype validation using LLMs
- Built an ontology-driven framework for standardizing radiology procedure terminology using LLMs and RAG

Sumin Lee,
*Department of Biomedical Systems Informatics,
Yonsei University College of Medicine*

Severance



Phenotype algorithm: An operational definition that identifies specific diseases using not only diagnostic codes but also relevant clinical characteristics.

- Rigorous validation is required to ensure that the definition accurately captures the patients of true clinical interest
- Gold Standard : Manual Chart review → accurate but time-consuming, resource-intensive, subjective
- LLMs can automate and scale validation → but black-box models, hallucinations, lack of uncertainty awareness

JOURNAL ARTICLE

Confidence-linked and uncertainty-based staged framework for phenotype validation using large language models

[Get access >](#)

Sumin Lee, BS , Hyeok-Hee Lee, MD, PhD , Hokyou Lee, MD, PhD , Kyu Sun Yum, MD , Jang-Hyun Baek, MD, PhD , Jaewon Khil, PhD , Jaeyong Lee, MD , Sojung Shin, MS , Minsung Cho, BSPH , Na Yeon Ahn, BSN, RN ... [Show more](#)

Journal of the American Medical Informatics Association, Volume 32, Issue 8, August 2025,

Pages 1320–1327, <https://doi.org/10.1093/jamia/ocaf099>

Published: 17 June 2025 **Article history** ▼

We propose a staged framework “CLUES” using LLMs

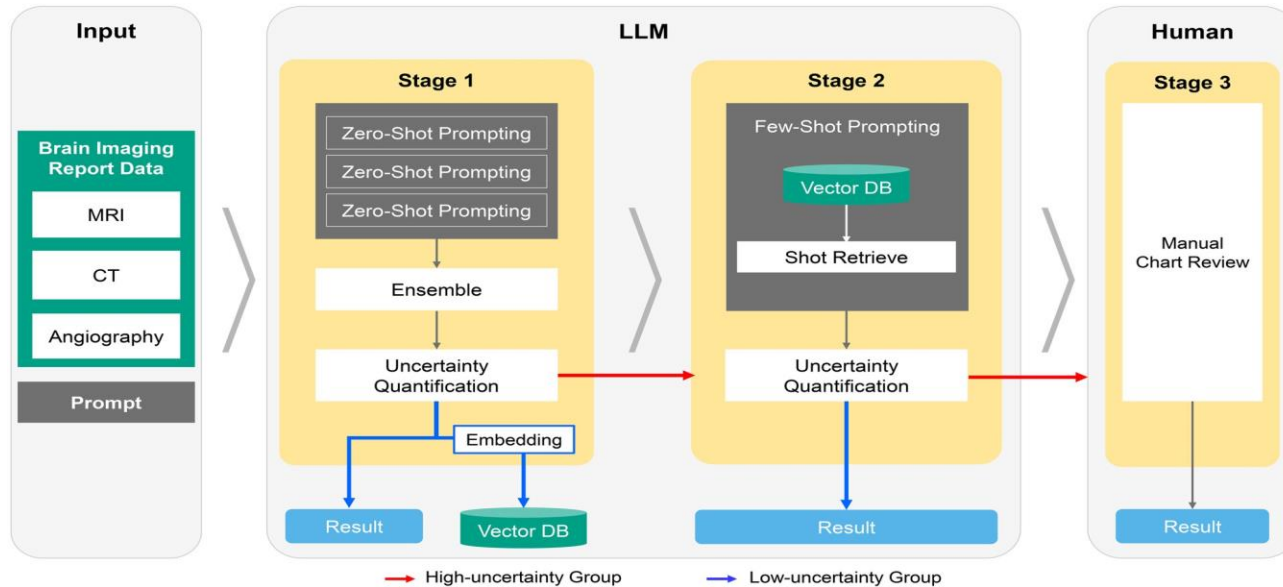
- to support large-scale chart review by inferring stroke probability from radiology reports
- quantifying uncertainty and applying prompt engineering for reliable validation

Data source & Study population

- Imaging reports(CT, MRI, angiography) of stroke patients from 24 hospitals in Korea (2011–2022) with expert-validated labels

CLUES Framework: At each stage, response uncertainty is quantified, and only results with high reliability are adopted

- Stage 1** (LLM initial inference) : Zero-Shot + Ensemble
- Stage 2** (Refinement with additional context): RAG + Few-shot
- Stage 3** (Final human adjudication): Manual chart review



Overview of the CLUES framework

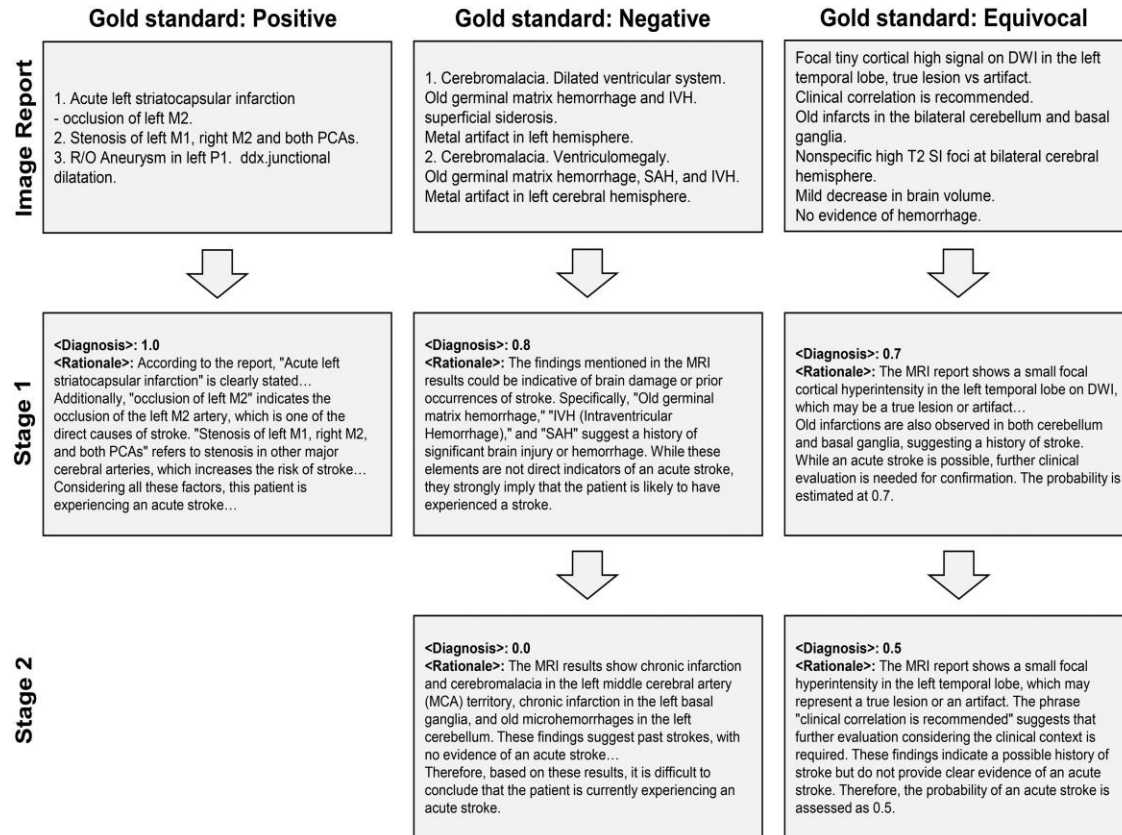
LLM explicit confidence value p (probability of the patient having a stroke)

$$X = \begin{cases} 0 \text{ (negative) with prob } 1 - p \\ 1 \text{ (positive) with prob } p \end{cases}$$

Entropy of X is calculated by:

$$H(X) = -p \log p - (1 - p) \log (1 - p) = H(p)$$

- By setting the entropy threshold at the median for each stage, the proportion of cases requiring manual chart review (Stage 3) was reduced to 25%
- At all stages, the the low-uncertainty group consistently outperformed the high-uncertainty group
- Performance improved progressively as cases advanced through the staged framework.



Stage 1 (n=1072)						
Uncertainty group	Gold standard	No. of patients	Precision	Recall	F1-score	Weighted F1-score
Low uncertainty (n=507)	Negative	89	0.99	0.82	0.90	0.94
	Equivocal	8	0.00	0.00	0.00	
	Positive	410	0.94	1.00	0.97	
High uncertainty (n=565)	Negative	358	0.96	0.45	0.61	0.57
	Equivocal	34	0.00	0.00	0.00	
	Positive	173	0.42	0.97	0.59	

Stage 2 (n=565)						
Uncertainty group	Gold standard	No. of patients	Precision	Recall	F1-score	Weighted F1-score
Low uncertainty (n=280)	Negative	144	0.97	0.76	0.85	0.82
	Equivocal	8	0.00	0.00	0.00	
	Positive	128	0.74	0.98	0.84	
High uncertainty (n=285)	Negative	214	0.93	0.52	0.66	0.58
	Equivocal	26	0.50	0.15	0.24	
	Positive	45	0.26	0.91	0.41	

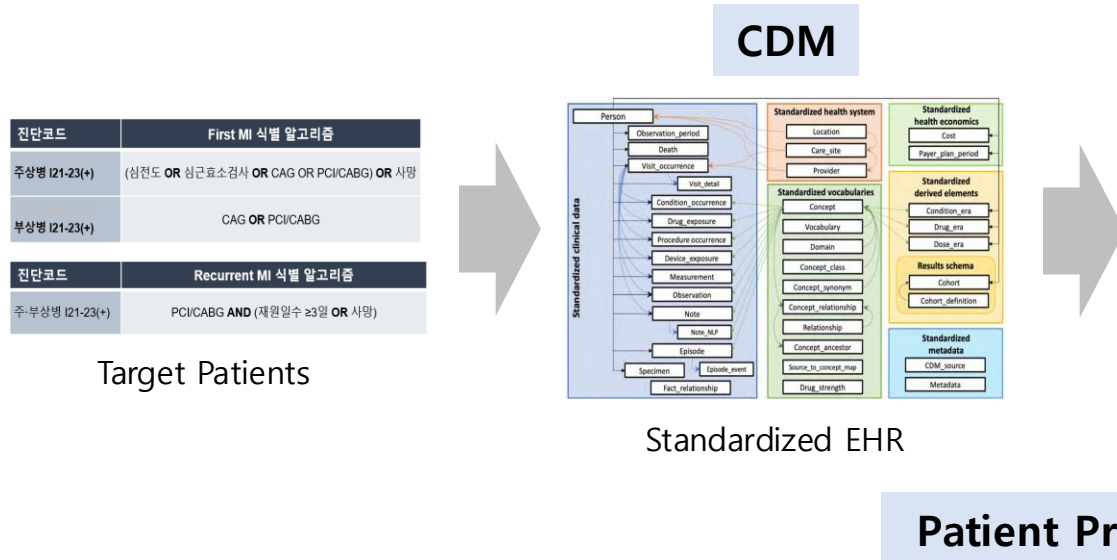
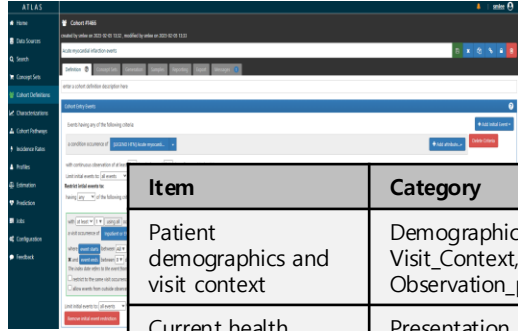
Performance by stage adoption in the CLUES framework excluding equivocal labels (n=1030).

Adopted stage	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score	Kappa
Stage 1	0.787	0.988	0.526	0.731	0.971	0.840	0.543
Stages 1 and 2	0.845	0.986	0.660	0.791	0.974	0.878	0.671
Stages 1, 2, and 3	0.947	0.993	0.886	0.920	0.990	0.955	0.890

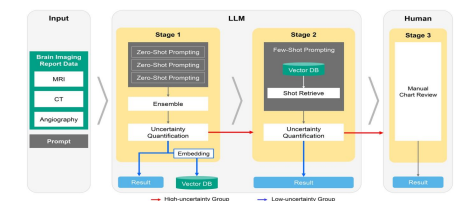
Future Direction: Expand to diverse diseases and datasets

- Develop additional automation strategies leveraging CDM
- Apply the CLUES to patient-level profiles generated using KEEPER (Ostropolets et al., JAMIA, 2024, Schuemie et al., NPJ Digit Med, 2025)

Patient profiles automatically generated using CDM

Item	Category
Patient demographics and visit context	Demographics, Visit_Context, Observation_period
Current health status and diagnostic process	Presentation, Comorbidities, Symptoms, Diagnostic_procedures, Measurements, Alternative_diagnosis
Prior disease and treatment history	Prior_disease, Prior_drugs, Prior_treatment_procedures
Outcomes and prognosis after treatment	After_disease, After_drugs, After_treatment_procedures, Death



Apply the CLUES to patient profile review for disease adjudication

Institutions use **custom imaging procedure codes**, limiting consistency in phenotyping and data retrieval.

- While semantic standardization can address this, manual mapping is time-consuming and irreproducible.
- The LOINC/RSNA Radiology Playbook offers radiology-specific granularity by encoding attributes like modality, anatomy, and contrast—surpassing general vocabularies such as EDI (claim code), SNOMED.

Local Code term	EDI Code term	SNOMED Code term	LOINC-RadLex Code term		
			LOINC Code term	RadLex PartType	RadLex Part term
CT Research Liver – LBW (contrast)	RC4018 [Abdominal CT (contrast)]	429862006 [CT of liver with contrast]	24815-3 [CT Liver W contrast IV]	modality	CT
				anatomic location	Abdomen
					Liver
				pharmaceutical	Contrast
		419394006 [CT of abdomen and pelvis]	36813-4 [CT Abdomen and Pelvis W contrast IV]	IV	
				timing	W
CT Abdomen + Pelvis (contrast)				modality	CT
				anatomic location	Abdomen
					Pelvis
				pharmaceutical	Contrast
				IV	
CT Research liver + pelvis_RPP (contrast)				timing	W

Example of radiology procedure codes standardization


NOW AVAILABLE: LOINC Version 2.80 Hotfix is here [Log In](#) [Sign Up](#)

LOINC
From Regenstrief

[Learn](#) [Content](#) [Downloads](#) [Community](#)

HOME ► CONTENT

Download LOINC



DOWNLOAD

Loinc_2.80.zip

FREE LOINC LOGIN REQUIRED

The LOINC database is available for free download. In addition to the LOINC table files, this zip archive also includes a number of accessory files which can prove useful in your implementation or use of LOINC.

Released: 2025-02-26
Size: 76MB
MD5: e4c5f4ec8844d79abed07a0e3493d1a

All documentation is available in the [LOINC Knowledge Base](#).

LOINC/RSNA Playbook from LOINC 2.80 (2025-02-26 released)

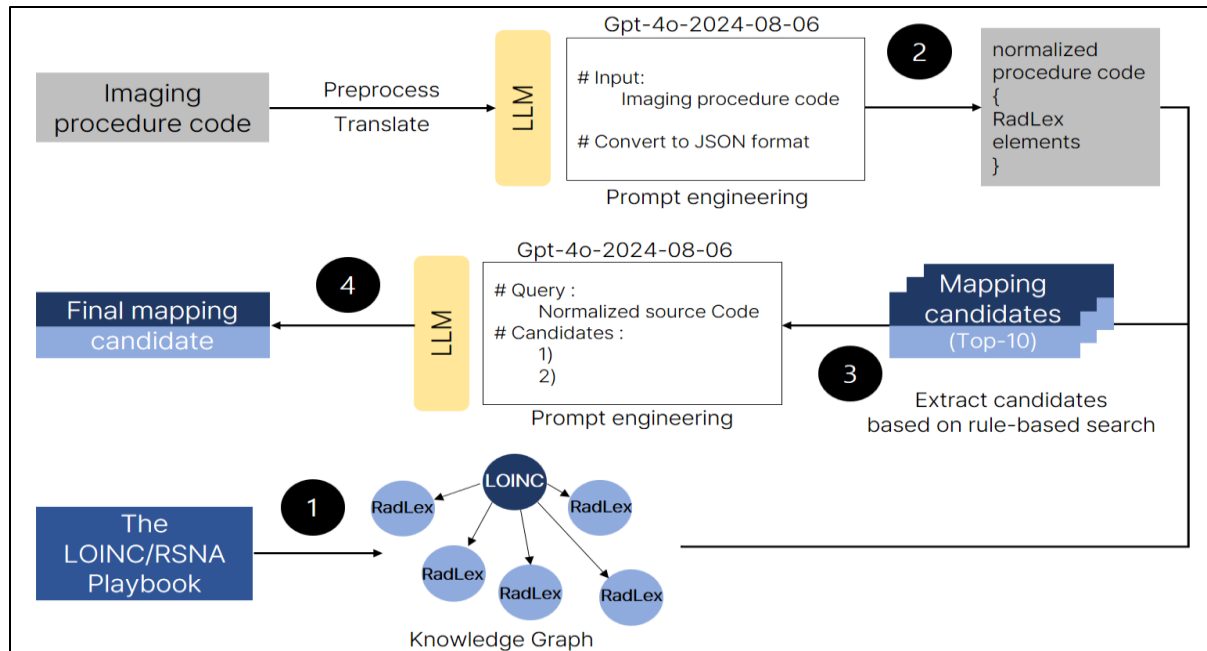
We propose an automated framework “ARKE” using LLMs and RAG

- to map logical imaging procedure terms to LOINC-RadLex codes
- Leveraging ontology to ensure semantically consistent standardization

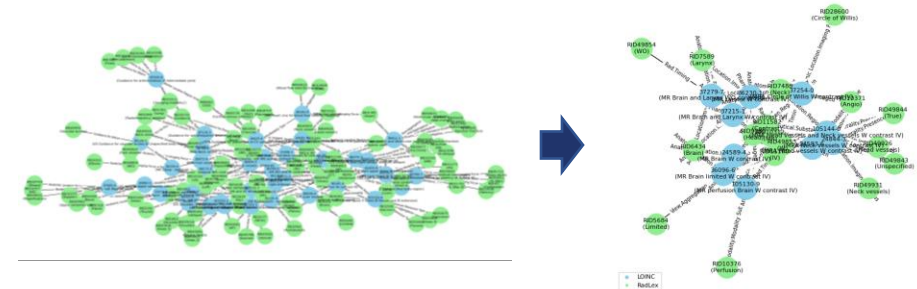
Data source : 2,126 local imaging procedure terms collected from YUHS, a tertiary hospital in South Korea, with the LOINC/RSNA Radiology Playbook (LOINC 2.80) used as the reference standard.

ARKE Framework

- 1 Construct Knowledge Graph from the LOINC/RSNA Radiology Playbook
- 2 Convert imaging procedure terms into JSON format (LLM)
- 3 Retrieve Top-10 candidates using rule-based matching (RAG)
- 4 Select One (LLM)



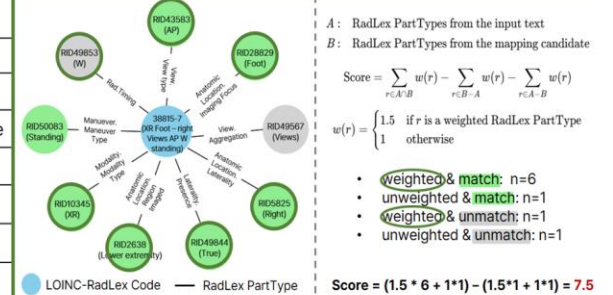
Workflow of ARKE framework



Category	Weighted RadLex PartType Name
Modality	Modality.Modality Type
	Modality.Modality SubType
Anatomic Location	Anatomic Location.Region Imaged
	Anatomic Location.Imaging Focus
	Anatomic Location.Laterality.Presence
	Anatomic Location.Laterality
View	View.View Type
Timing	Timing
Pharmaceutical	Pharmaceutical.Substance Given

Input text: Foot Standing AP Rt

Top-1 candidate: 38815-7 [XR Foot - right Views AP W standing]



A : RadLex PartTypes from the input text
 B : RadLex PartTypes from the mapping candidate

$$\text{Score} = \sum_{r \in A \cap B} w(r) - \sum_{r \in B - A} w(r) - \sum_{r \in A - B} w(r)$$

$$w(r) = \begin{cases} 1.5 & \text{if } r \text{ is a weighted RadLex PartType} \\ 1 & \text{otherwise} \end{cases}$$

- weighted & match: n=6
- unweighted & match: n=1
- weighted & unmatched: n=1
- unweighted & unmatched: n=1

Score = (1.5 * 6 + 1*1) - (1.5*1 + 1*1) = 7.5



- ARKE effectively maps imaging procedure terms to LOINC–RadLex codes, often exceeding the silver reference
- LOINC–RadLex outperforms direct LOINC mapping
- Hit rate improves with larger N → Top-1 alone is insufficient (importance of prompt engineering)
- Weighted matching shows best performance (62% accuracy, Top-5 hit rate 0.68)

Input Text	Silver Reference	Final Mapping Candidate
US Breast Gold insertion	US Guidance for placement of needle in Breast	US Guidance for needle localization of Breast
Rib cage view Both Oblique	XR Chest Right oblique and Left oblique	XR Ribs anterior - bilateral Views
MRI Brain Limited study non contrast	MR Guidance for stereotactic localization of Brain-- WO contrast	MR Brain limited WO contrast
GYN US Routine General Doppler Limited	US.doppler Abdomen and Pelvis	US.doppler Pelvis vessels limited
GYN US Routine General Doppler	US.doppler Abdomen and Pelvis	US.doppler Pelvis vessels
Cryosurgical Ablation of Liver	US Guidance for ablation of tissue of Liver	Guidance for cryoablation of Liver
15T MRI Wholespine noncontrast diffusion	MR Spine WO contrast	MR Cervical and thoracic and lumbar spine WO contrast

Method	Final candidate	Top-1	Top-3	Top-5	Top-10		
	Accuracy	Accuracy	Hit Rate	Hit Rate	Hit Rate	MRR	NDCG@10
Direct LOINC mapping with cosine similarity	0.5359	0.2785	0.5021	0.5949	0.7511	0.4170	0.4964
LOINC-RadLex mapping with jaccard similarity	0.6160	0.4641	0.6245	0.6624	0.7679	0.5560	0.6065
LOINC-RadLex mapping with simple overlap	0.6245	0.4810	0.6118	0.6624	0.7553	0.5628	0.6086
LOINC-RadLex mapping with weighted match	0.6245	0.4768	0.6203	0.6835	0.7722	0.5626	0.6123
LOINC-RadLex mapping with f1-score	0.6203	0.4641	0.6245	0.6624	0.7679	0.5560	0.6065



Future Direction

- Generalization & Scalability
 - Expand mapping across multiple health systems and procedure lists
 - Extend beyond radiology (e.g., SNOMED, ICD-10-PCS)
- Framework Refinement
 - Enhance weighted-match attribute selection
 - Apply chain-of-thought prompting for better interpretability
- Evaluation & Validation
 - Broaden reviewer studies with inter-rater agreement metrics
 - Perform error analysis to address ambiguous or difficult cases
- Collaboration & Community
 - Collaborate with OHDSI Vocabulary WG for standardization
 - Leverage prior work on ICD-10PCS ↔ SNOMED mapping

Thank you

Sumin Lee (lsm0801@yuhs.ac)
Seng Chan You (chandryou@yuhs.ac)

From ATLAS to Strategus: LLM-Powered Automated Study Design with THESEUS

Hanjae Kim,* Min Seong Kim, Seng Chan You

* Department of Biomedical Systems Informatics,
Yonsei University College of Medicine

26 August 2025

Severance

ATLAS

- A web-based tool for designing observational studies with graphical user interface (GUI).
- Researchers without programming knowledge can design and execute analyses.



ATLAS

Home

Data Sources

Search

Concept Sets

Cohort Definitions

Characterizations

Cohort Pathways

Incidence Rates

Profiles

Estimation

Prediction

Jobs


Configuration

Feedback

Apache 2.0

open source software

provided by

OHDSI

join the journey

Cohort #1770710

New users of ACE inhibitors as first-line monotherapy for hypertension

Definition ? Concept Sets Generation Reporting Export Messages 3

enter a cohort definition description here

Cohort Entry Events ?

Events having any of the following criteria:

+ Add Initial Event

a drug exposure of ACE inhibitors + Add attribute... Delete Criteria

✗ for the first time in the person's history

with continuous observation of at least 365 days before and 0 days after event index date

Limit initial events to: earliest event per person.


Restrict initial events

Inclusion Criteria ?

New inclusion criteria

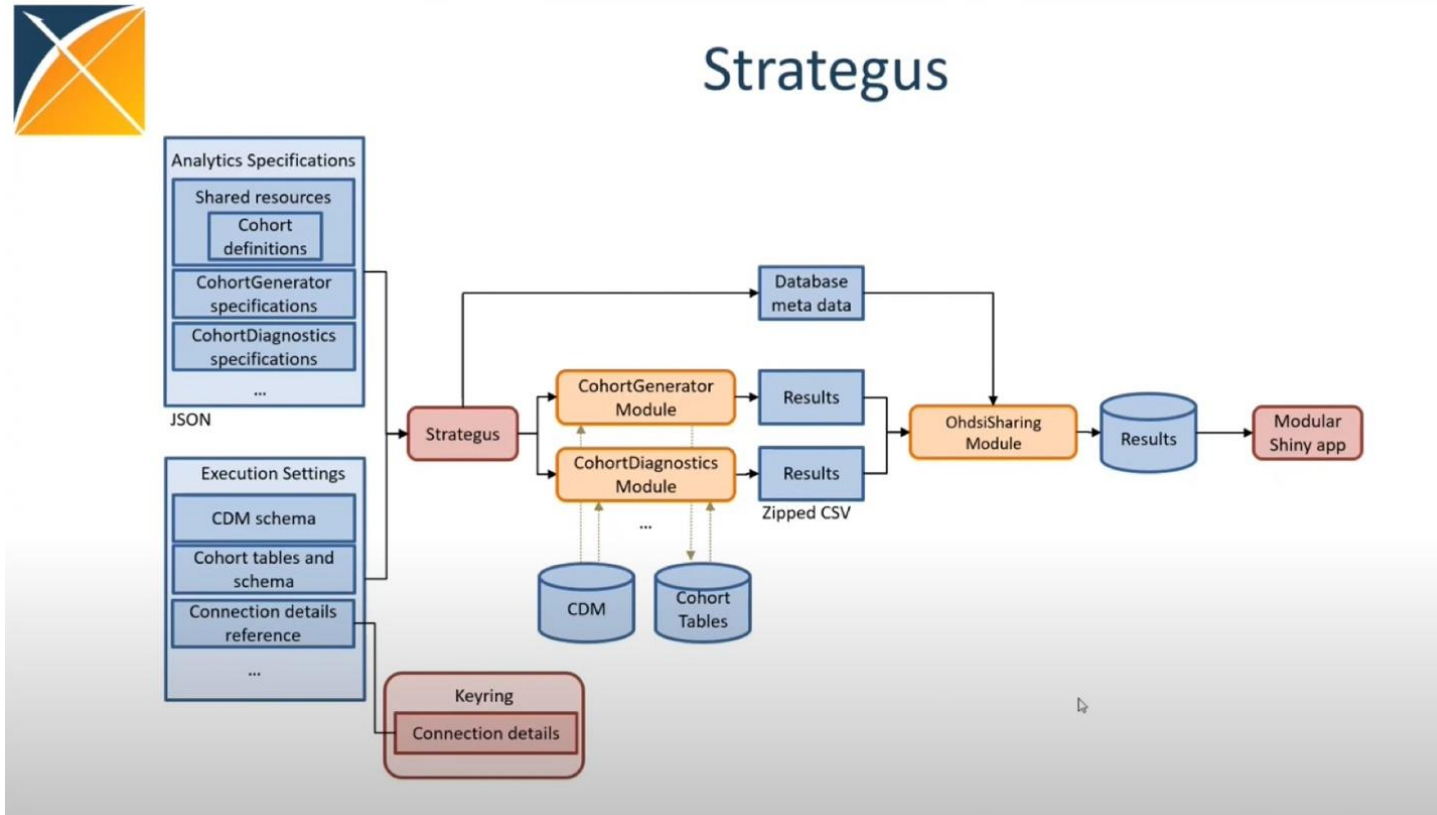
1. has hypertension diagnosis in 1 yr prior to treatment

2. Has no prior antihypertensive drug exposures in medical

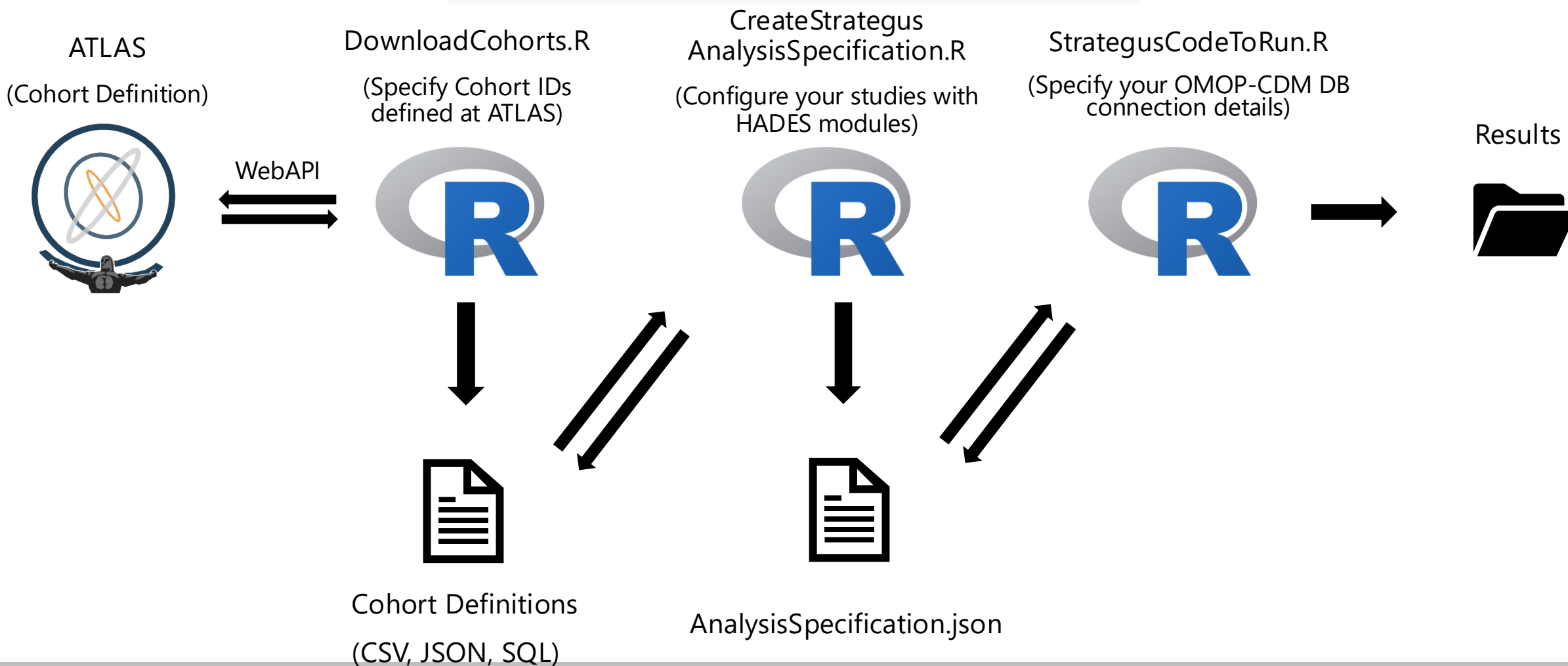
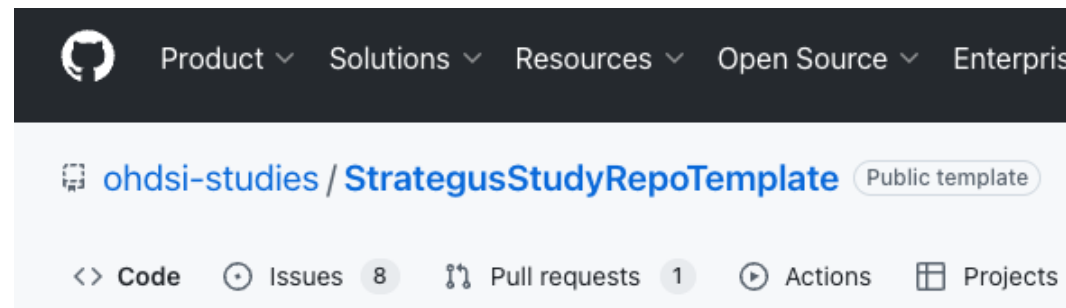
 YONSEI
UNIVERSITY

Strategus

- R package for coordinating and executing analyses using OHDSI HADES modules.
- A researcher can create a JSON document that specifies study designs and distribute it to other researchers for network studies.



Strategus Template



Limitation of Strategus



ATLAS provides GUI tools for researchers to define cohorts without programming expertise

VS

Strategus

R-CMD-check passing codecov 98%

Strategus is part of [HADES](#).

Strategus requires knowledge of R and HADES modules



What if researchers could create Strategus scripts with a simple click of a button in ATLAS?

THESEUS: Text-guided Health-study Estimation and Specification Engine Using Strategus

- A prototype app modeled on the 'population-level estimation' tab of ATLAS.
- Key Features (Powered by LLM):
 1. Users can input study design in free text, which is automatically reflected in the ATLAS UI
 2. Users can convert their study design into Strategus Script (CreateAnalysisSpecification.R)

The screenshot displays the 'New Population Level Estimation Analysis - Comparative Cohort Analysis' form. At the top, there is a 'Study Name' input field. Below this, the 'Comparisons' section is expanded, showing three sub-sections: 'Cohort Settings', 'Negative Control Concept Set', and 'Covariate Selection', each with a right-pointing arrow. The 'Analysis Settings' section is also expanded, showing 'Study Population' and 'Time At Risk'. The 'Time At Risk' section includes an 'Add' button and a 'Remove' button. Below these, there is a 'Description' input field, a 'Time-at-risk Start' input field with a dropdown menu, and a 'Time-at-risk End' input field with a dropdown menu. At the bottom, there are two buttons: 'Text to UI Settings' and 'Generate Strategus R code'.

<https://theseus2.vercel.app/>

THESEUS: Text-guided Health-study Estimation and Specification Engine Using Strategus

Copy the generated R script into 'CreateStrategusAnalysisSpecification.R' in the Strategus study repository template and execute it.

Study Start & End Dates

Set the cohort index date range. Leave blank to use all time.

Add

Remove

LLM-powered Strategus R code generated

Uses an LLM to convert the current UI into Strategus-ready R code for analysis.

Convert to Strategus Code

LLM-generated Strategus Code

```
#####
# CreateStrategusAnalysisSpecification.R
#
# This script creates a Strategus analysis specification JSON for the
# "TicagrelorVsClopidogrel" study using the settings supplied in the
# Analysis Specifications. The script follows the OHDSI Strategus template
# structure and includes detailed annotations to explain how each JSON
# setting is mapped into Strategus/CohortMethod module specifications.
#####
# -----
# Shared resources & Cohort Definitions
# -----
# Base Atlas WebAPI URL used to export cohort definitions and concept sets.
# Modify this to your Atlas/WebAPI endpoint if needed.
```

Copy Code

Remove Overlapping Subjects

Remove subjects that are in both the target and comparator cohort?

keep all

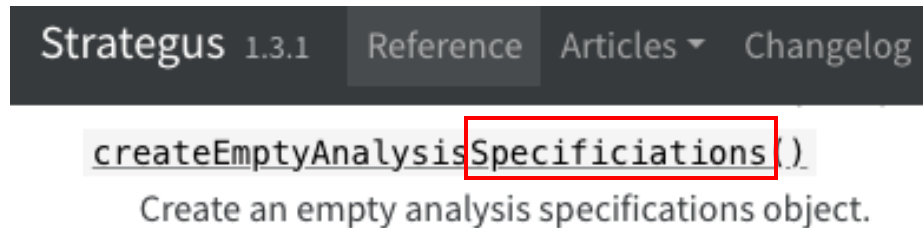


CreateStrategusAnalysisSpecification.R

```
24
25 # Export the cohort definitions for:
26 # - Target cohort (New Users of Ticagrelor) : cohort id 1788314
27 # - Comparator cohort (New Users of Clopidogrel) : cohort id 1788315
28 # - Outcome cohort(s) (NACE) : cohort id 1788319
29 #
30 # We request generateStats = TRUE so that cohort generation statistics are
31 # available if the CohortGenerator module is run.
32 cohortDefinitionSet <- ROHdsiWebApi::exportCohortDefinitionSet(
33   baseUrl = baseUrl,
34   cohortIds = c(
35     1788314, # Target: New Users of Ticagrelor
36     1788315, # Comparator: New Users of Clopidogrel
37     1788319 # Outcome: NACE
38   ),
39   generateStats = TRUE
40 )
41
42 # Re-number cohorts for internal use in the Strategus specification:
43 # - 1 = target
44 # - 2 = comparator
45 # - 3 = outcome
46 cohortDefinitionSet[cohortDefinitionSet$cohortId == 1788314,]$cohortId <- 1
47 cohortDefinitionSet[cohortDefinitionSet$cohortId == 1788315,]$cohortId <- 2
48 cohortDefinitionSet[cohortDefinitionSet$cohortId == 1788319,]$cohortId <- 3
49
50 # -----
51 # Negative control outcomes (concept set)
52 # -----
53 # The analysis specification lists a negative control concept set:
54 # concept set id = 1882685 ("Negative Controls - TicagrelorVsClopidogrel")
55 # We resolve the concept set into individual concepts and convert those into
56 # a cohort-like table that Strategus expects for negative control outcomes.
57 negativeControlOutcomeCohortSet <- ROHdsiWebApi::getConceptSetDefinition(
58   conceptSetId = 1882685,
59   baseUrl = baseUrl
60 ) %>%
61 ROHdsiWebApi::resolveConceptSet(baseUrl = baseUrl) %>%
62 ROHdsiWebApi::getConcepts(baseUrl = baseUrl) %>%
63 dplyr::rename(
64   outcomeConceptId = "conceptId",
65   cohortName = "conceptName"
66 ) %>%
67 # Assign cohortId values starting at 101 (so we don't collide with 1,2,3 used above)
68 dplyr::mutate(cohortId = dplyr::row_number() + 100) %>%
69 dplyr::select(cohortId, cohortName, outcomeConceptId)
70
```

Findings

- GPT-5 can convert natural language into ATLAS study configurations, but sometimes fails to completely understand complex analysis descriptions.
- As long as the analysis settings are provided, GPT-5 is excellent at generating the Strategus code.
 - The sole issue:
'createEmptyAnalysisSpecificiations()' is sometimes altered to 'createEmptyAnalysisSpecificiations()'





YONSEI UNIVERSITY
COLLEGE OF MEDICINE

