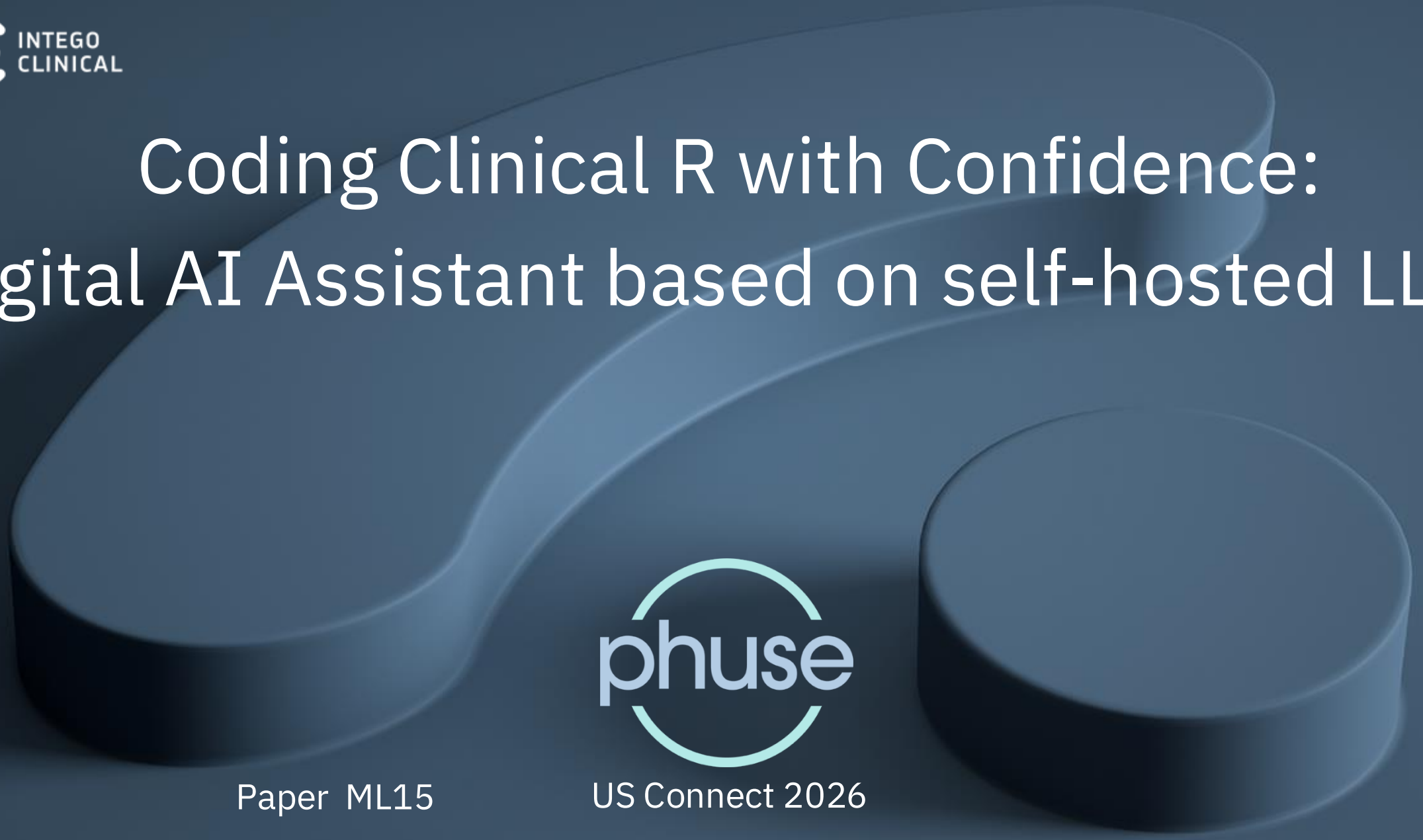
INTEGO CLINICAL

Coding Clinical R with Confidence: Digital AI Assistant based on self-hosted LLMs



phuse

Paper ML15

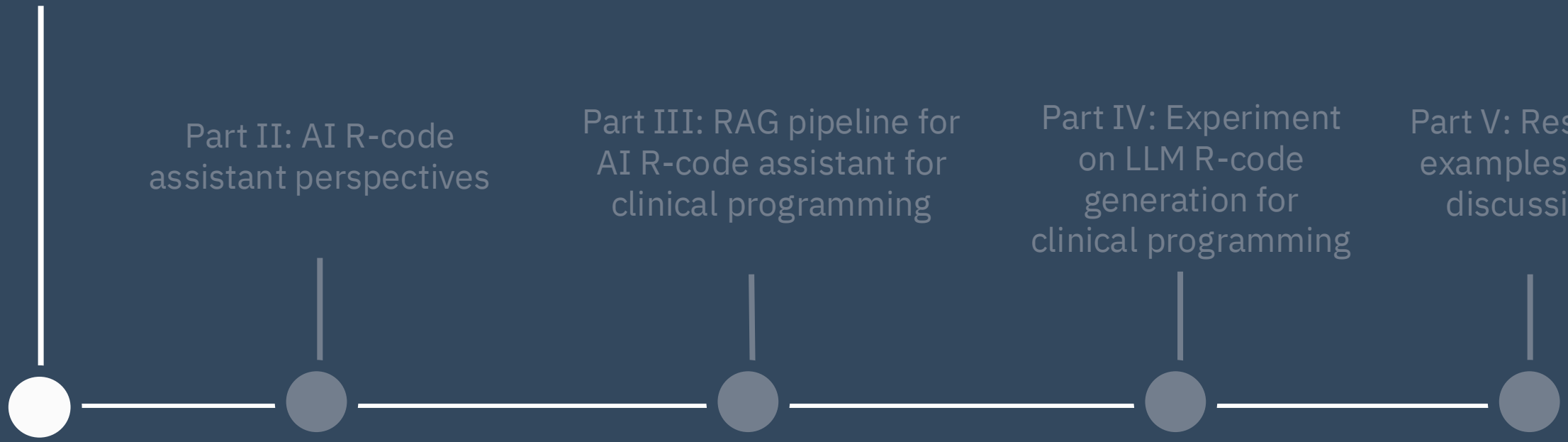
US Connect 2026

INTEGO CLINICAL

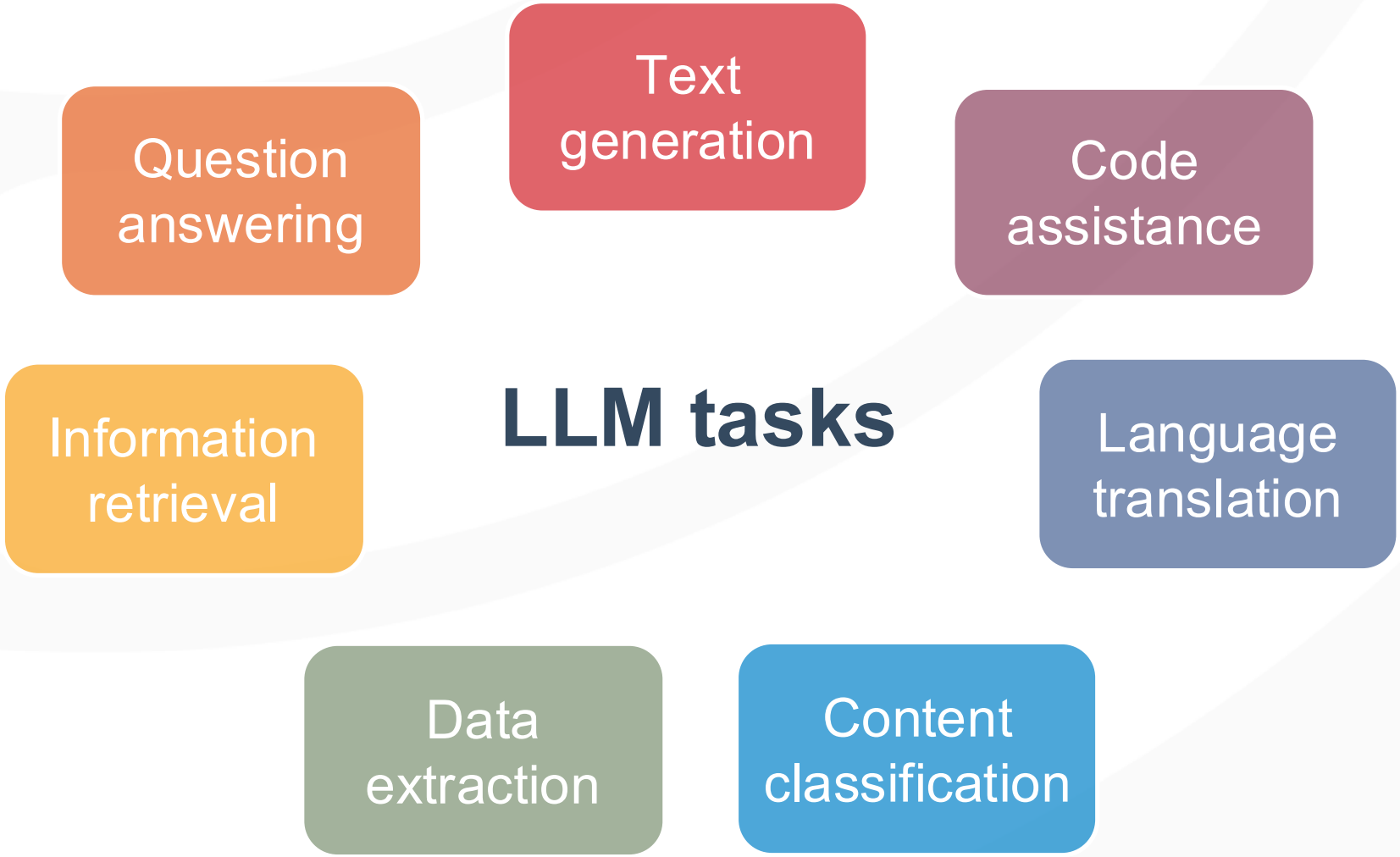
Today we'll cover:

I. Overview of Large Language Models
II. AI R-code assistant perspectives
III. RAG pipeline for AI R-code assistant for clinical programming
IV. Experiment on LLM R-code generation for clinical programming
V. Results, examples, and discussion

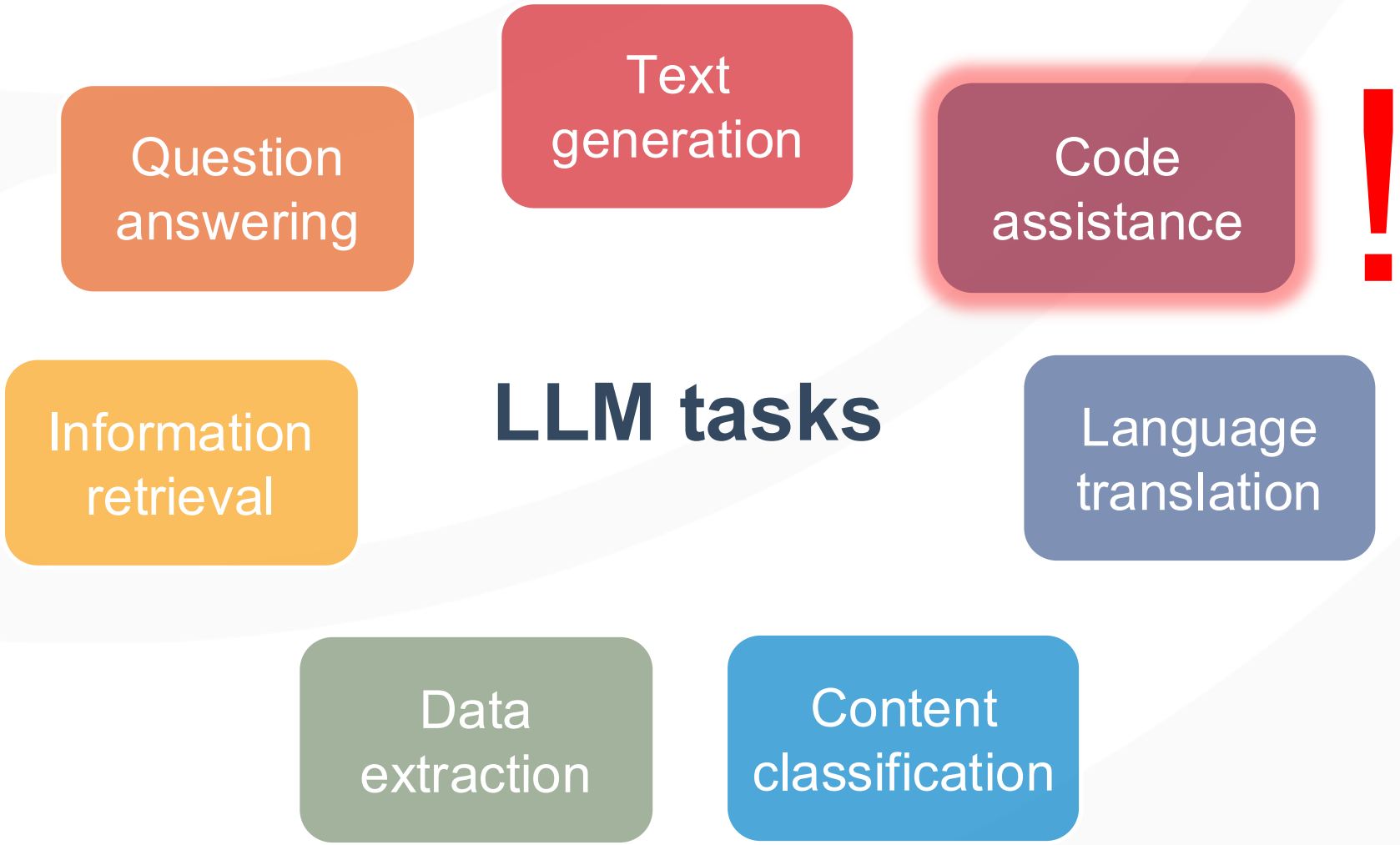
Part I: Overview of Large Language Models



Large Language Models (LLMs):
AI Models for Natural Language Processing



Large Language Models (LLMs):
AI Models for Natural Language Processing



INTEGO CLINICAL

CLOSED-SOURCE MODELS vs OPEN-SOURCE MODELS

CLOSED-SOURCE MODELS

Advantages:
- Performance
- Safety and ethics
- Support and Infrastructure

Limitations:
- High usage costs
- Limited customization options
- Data privacy concerns
- Black box architecture

Examples:
- GPT-4o mini (OpenAI)
- Claude 3.5 (Anthropic)
- Gemini 2.0 (Google)

OPEN-SOURCE MODELS

Advantages:
- Full access to mode architecture
- Free to use and deploy independent of external services
- No external data sharing

Limitations:
- Lower performance
- High resource consumption
- Requires technical expertise
- Limited support options

Examples:
- Llama 3 (Meta)
- Phi-4 (Microsoft)
- Gemma 2 (Google)

Local LLM Deployment Solutions

Ollama

- user-friendly tool for managing and deploying LLMs
- runs on macOS®, Windows®, and Linux®
- offers both CPU and GPU acceleration

LM Studio

- desktop application for Windows and macOS
- streamlines the process of downloading and running local LLMs

Hugging Face

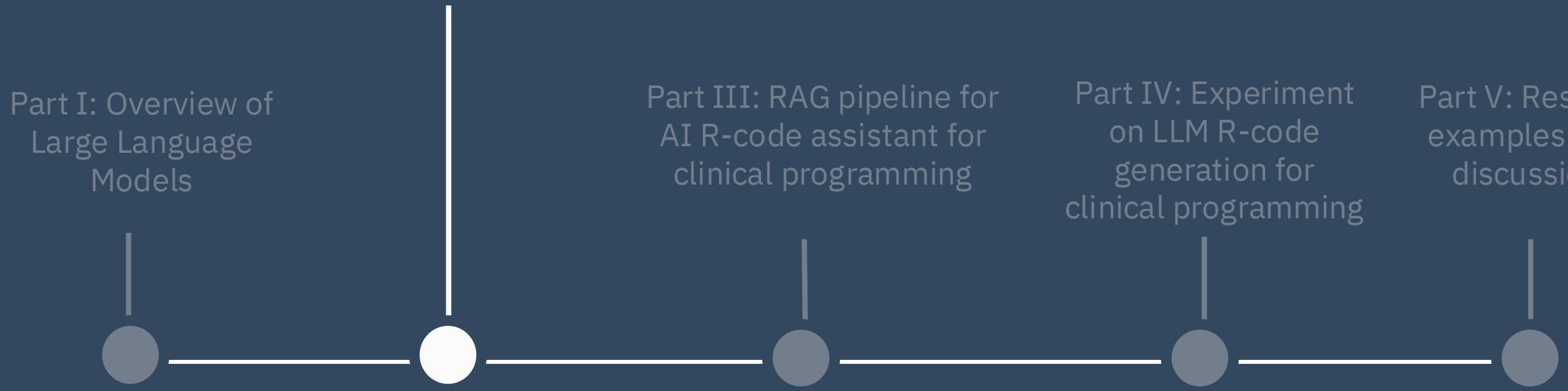
- leading platform for open-source AI and machine learning
- provides tools, models and datasets for developers and researchers

Enhancing LLM performance

	Fine-tuning	RAG*
Essence	Retrains the model with specific area data	Uses external database for retrieving contexts
Tasks	Adaptation for specific area Tuning the generation style	Knowledge-based generation Document summarization
Requires	Powerful hardware Large training dataset	External database Tuning the retriever parameters

*Retrieval-augmented generation

PART II: AI R-CODE ASSISTANT PERSPECTIVES



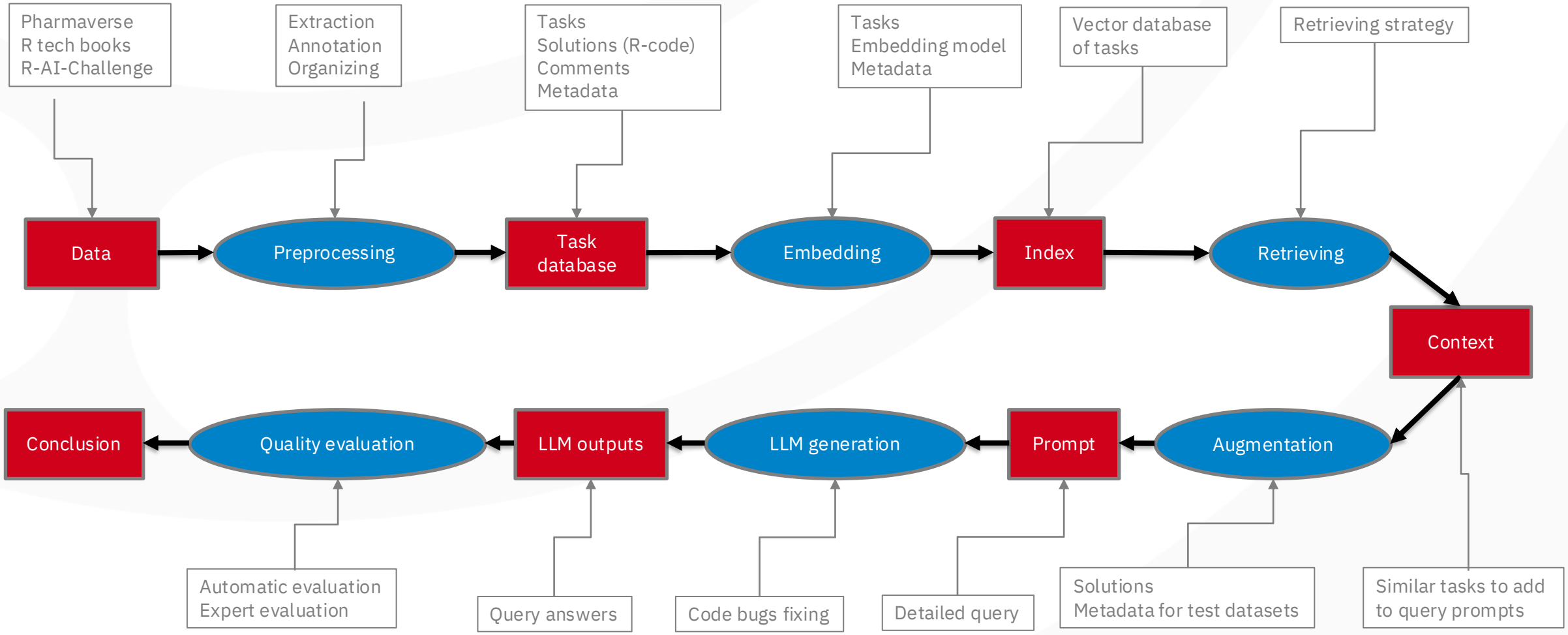
INTEGO CLINICAL

Training transform, “humans out, AI in”:

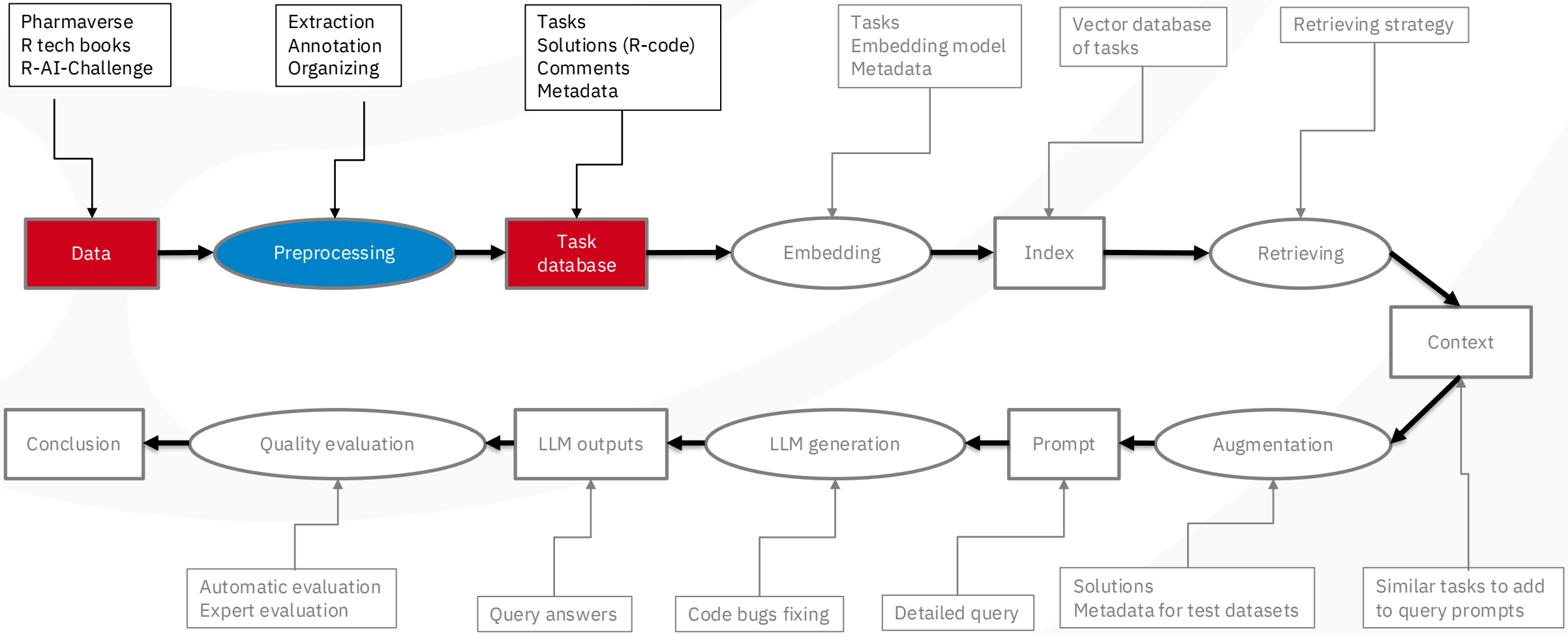
- Industry is migrating to R
- Request for the fast re-training of specialists
- In this re-training, some assistance (AI) is required
- Fast prototyping of solutions
- Any AI assistant should be domain specific

PART III: RAG PIPELINE FOR AI R-CODE ASSISTANT FOR CLINICAL PROGRAMMING

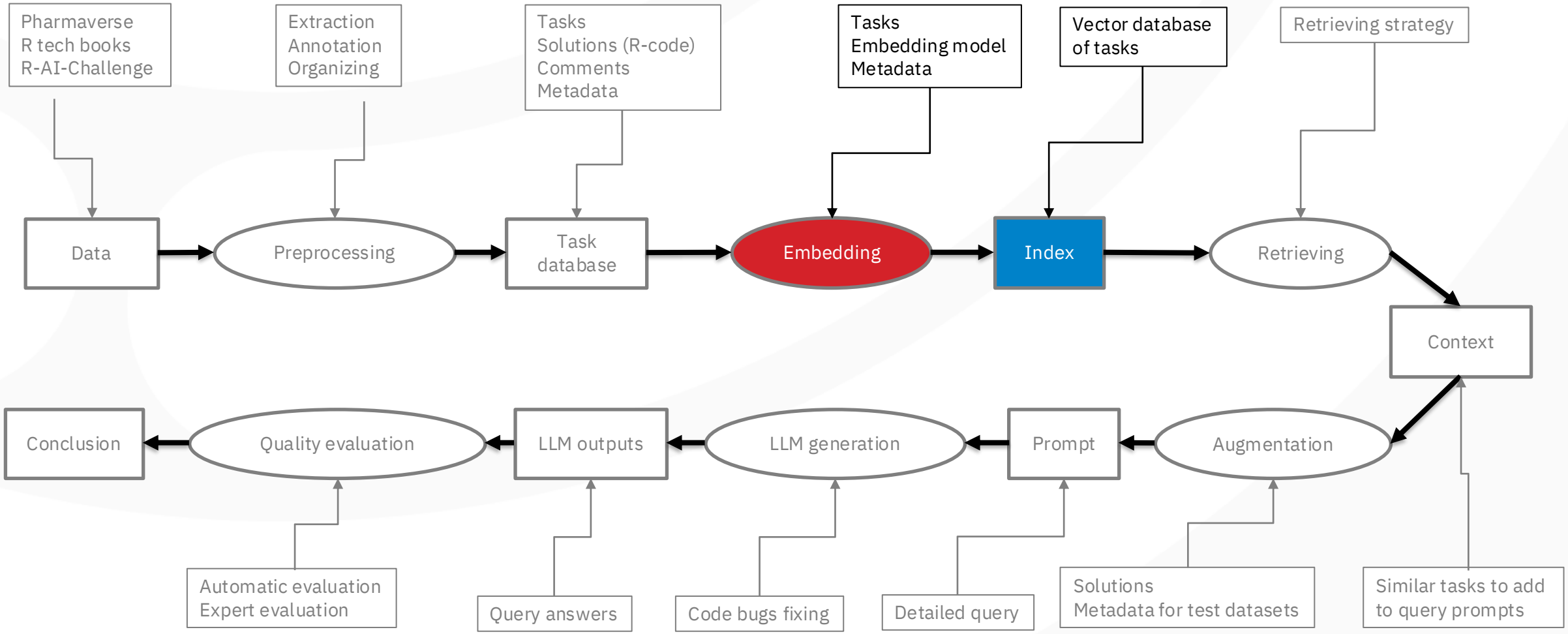
RAG pipeline for R-code assistant



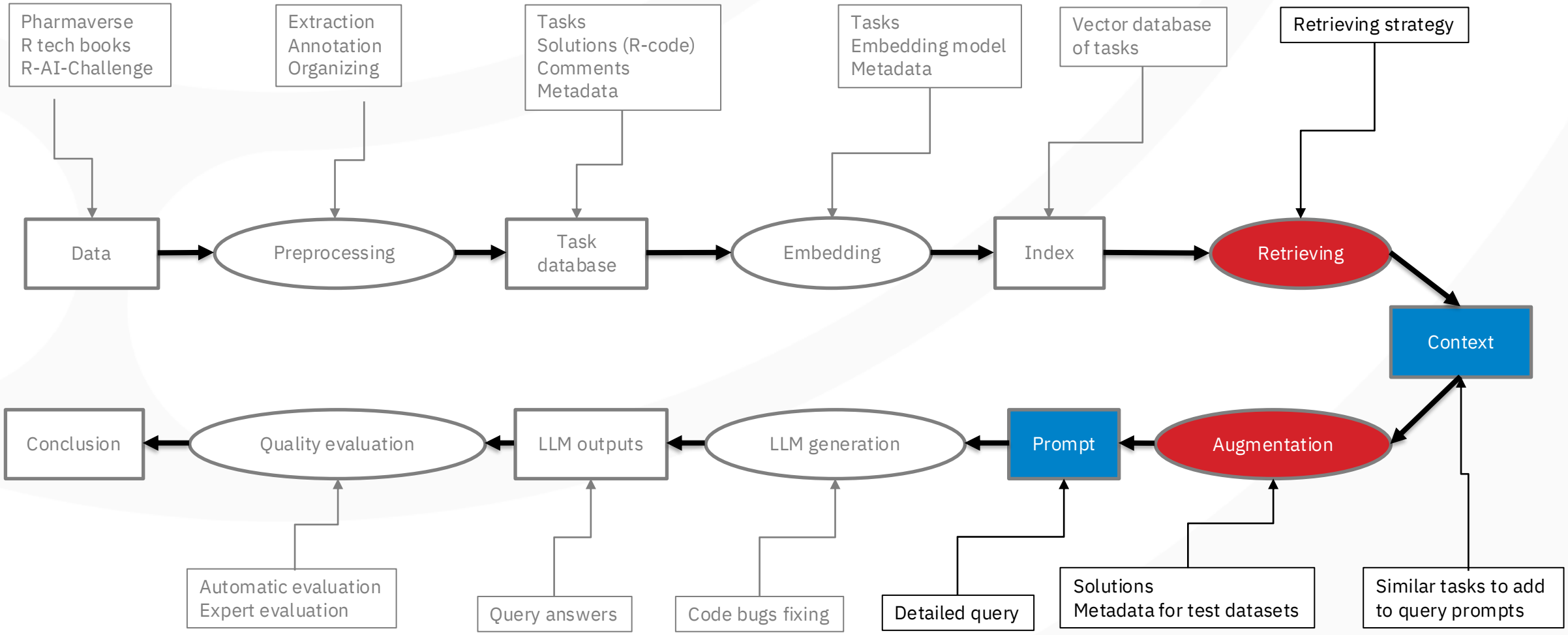
RAG pipeline for R-code assistant



RAG pipeline for R-code assistant



RAG pipeline for R-code assistant

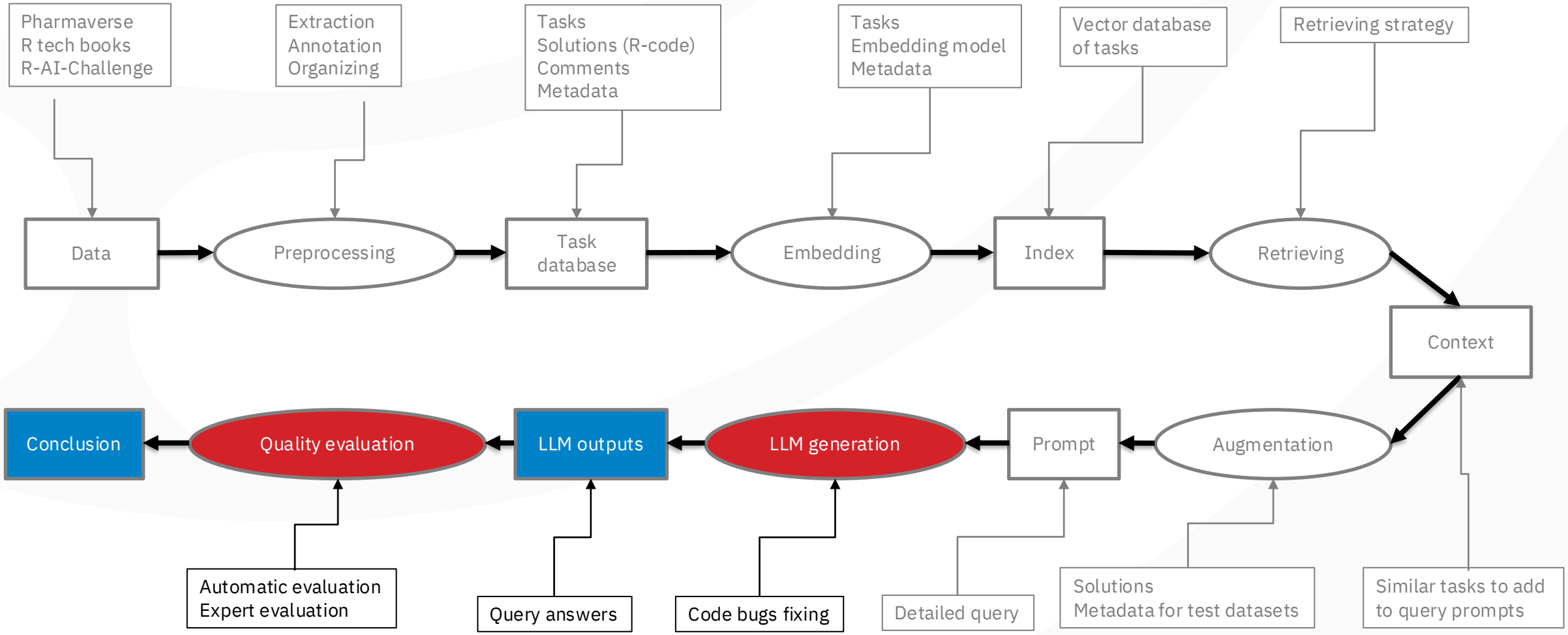


INTEGO CLINICAL

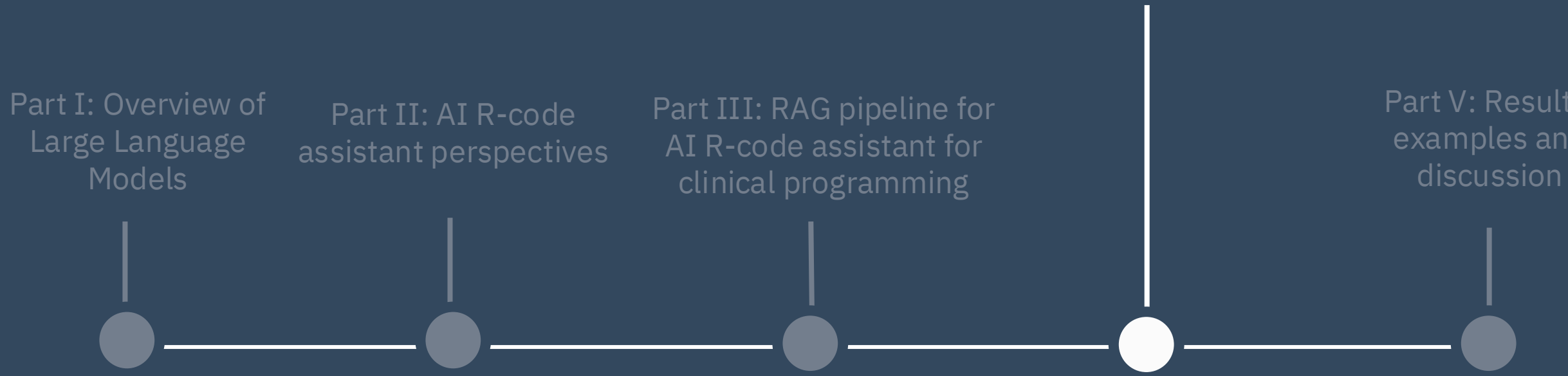
Prompt Engineering:
Essential elements shaping the model's response

Instruction:	• Defines the specific task or directive for the model
Behavior Control:	• Sets tone, style, constraints, and output length
Context:	• Provides background or external information to guide responses
Input Data:	• The core query or content for model processing
Examples:	• Provide reference responses to guide the model's output style and accuracy
Output Indicator:	• Specifies the expected response type or format

RAG pipeline for R-code assistant



PART IV: Experiment on LLM R-code generation for clinical programming



R knowledge base

OPEN-SOURCE DATABASE

Sources:	• Pharmaverse R-packages documentation • R tech books, tutorials, repositories
Preprocessing:	• Code extraction • “Task-solution” dataset • Missing data LLM generation
Size and details:	• Approx. 7000 chunks • Mostly one-line code • General purpose

CLOSED-SOURCE DATABASE

Sources:	• Internal company R-AI-Challenge
Preprocessing:	• High-quality solutions selection by R-experts • “Task-solution” dataset with metadata
Size and details:	• Approx. 100 tasks and 250 solutions • Mostly complex solutions • Domain specified

INTEGO CLINICAL

R-AI-Challenge

A coding challenge within *Intego Clinical* with:

- Approx. 100 tasks of 3 complexity levels formulated by senior statistical programmers
- Task addressed SDTM, ADaM, and TLF
- Special platform to submit solutions
- Review and scoring by R-experts

- Result: **≈250 high quality solutions**

INTEGO CLINICAL

Model selection: generation

HumanEval (OpenAI, 2021; Python)

- execution-based testing
- **functional correctness** over textual similarity
- pass only if **all unit tests** pass

HumanEval-R

- R-language adaptation
- hosted on Hugging Face
- 161 tasks
- a natural language problem description
- R function signature (name + arguments)
- automated R unit tests

Model selection: generation

Metric: pass@k success if **at least one** of *k* generated solutions **passes all unit tests**

LLM	pass@1	runtime
Open-weight models		
gpt-oss:20b	0.714	2h 46m
codellama:13b-instruct	0.255	2h 25m
starcoder2:instruct	0.379	1h 27m
wizardcoder:33b-v1.1	0.416	4h 5m
codestral:22b	0.398	2h 16m
Closed-weight models		
o3	0.739	
claude-sonnet-4	0.72	

Model selection: embedding

Custom retrieval benchmark

- chunked corpus: pharmaverse R package documentation
- 200 LLM-generated questions: 1 question per chunk

Retrieval setup

- vector database per embedding model
- return top-3 chunks

Model selection: embedding

Metrics

MRR: higher when the relevant chunk ranks closer to **#1**

Hit Rate (Hit@3): relevant chunk appears in the **top 3** results

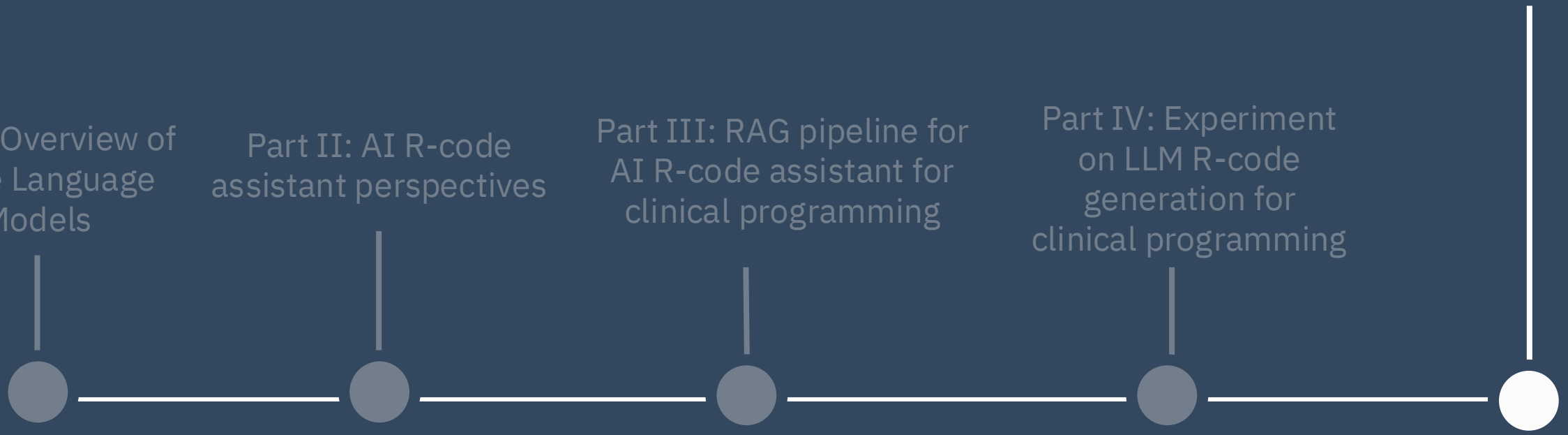
Embedding model	Runtime	MRR	Hit Rate
nomic-embed-text	43.33 s	0.952	0.980
jina-embeddings-v2-base-code	47.92 s	0.934	0.965
granite-embedding	30.56 s	0.970	0.990
oscardp96/medcpt-query	41.34 s	0.538	0.620
mxbai-embed-large	57.96 s	0.960	0.980
embeddinggemma	68.16 s	0.932	0.960

Experiment design:

RAG vs RAG-free based on open-source database

RAG vis RAG-free based on closed-source database

PART V: Results, examples, and discussion



Open-source database RAG

RAG system evaluation goal

- apply *pharmaverse* package documentation to generate correct R code

Task design

- task description
- input tables as **metadata**
- new variables/columns to create (if applicable)

System prompt conventions

- inputs **already loaded**
- standardized output object: **`out_dataset`**

Ground truth

- canonical R code → reference dataset **`derived_dataset`**

Open-source database RAG

Automated test script

- load inputs
- R canonical code
- R LLM-generated code
- compare `derived dataset` vs `out dataset`

Generation model

- **gpt-oss-20b**

Retriever: hybrid retrieval

- dense (semantic): **granite-embedding** in Chroma
- sparse (lexical): **BM25**

Two test modes

- **No RAG** (no retrieved context)
- **RAG** (retrieved documentation context)

Open-source database RAG

No RAG behavior

- limited knowledge of *pharmaverse* packages (e.g., **admiral**, **sdtmval**, **sdtm.oak**, **tidyCDISC**)
- falls back to generic tools (**dplyr**)

RAG behavior

- better package/function selection
- needs **domain-specific conventions** for correct usage

Domain-specific conventions

- symbols for column names (variables)
- exprs() for multiple vars/expressions
- strings for values only

Case study 1. RAG vis RAG-free on R-AI-Challenge Data

Request:

Derive the Treatment-Emergent Flag in the ADAE dataset. Assume ASTDTM, AENDTM, TRTSDTM and TRTENDTM are available in ADAE.

Reference solution, RAG solution: admiral, automatic

```r
library(admiral)
library(dplyr)

adae_trtemfl <- adae %>%
  derive_var_trtemfl(
    trt_start_date = TRTSDTM,
    trt_end_date = TRTEDTM,
    end_window = 0
  )
```

gpt-oss solution: dplyr, manual

```r
library(dplyr)

# Derive the TE flag
adad <- adae %>%
  mutate(
    TEFLG = case_when(
      # Event starts during the treatment period
      !is.na(ASTDTM) & !is.na(TRTSDTM) & !is.na(TRTEDTM) &
        ASTDTM >= TRTSDTM & ASTDTM <= TRTEDTM ~ "Y",
      # Event starts outside the treatment period
      !is.na(ASTDTM) & !is.na(TRTSDTM) & !is.na(TRTEDTM) &
        (ASTDTM < TRTSDTM | ASTDTM > TRTEDTM) ~ "N",
      # Missing key dates - flag as missing
      TRUE ~ NA_character_
    )
  )
```

The Task

1. Identify baseline records (ABLFL = 'Y', last obs before dosing)
2. Derive CHG = AVAL – BASE, PCHG = % change
3. Summarize mean/SD by visit and treatment arm

Three Evaluation Conditions

① No RAG	② RAG	③ RAG+Meta
Base model only	R-AI-Challenge context	Context + dataset metadata

Scoring Breakdown

Dimension	① No-RAG	② RAG	③ RAG+Meta
		CDISC-compliant	*Best structure*
Cross-dataset reasoning	0	0	3
Baseline deduplication (defensive)	3	0	0
Post-baseline filter (clinical)	**0**	3	3
Standards derivation (admiral)	1	3	2
CDISC naming (ADT, PARAMCD)	0	3	3
Parameter scope (standard)	0	3	1
Code robustness	0	1	2
TOTAL	**4/21**	**13/21**	**14/21**

No-RAG included baseline rows (CHG=0) — resulting in extra output rows.

Critical Code Differences

❌ **File Reading**	`read_csv() for XPT → wrong`	`haven::read_xpt()` ✓
❌ **CHG Derivation**	`Manual mutate() → no validation`	`admiral::derive_var_chg()` ✓
🔴 **Post-Baseline Filter**	`NO FILTER → includes baseline (WRONG)`	`filter(ADT >= TRTSDT)` ✓
❌ **Variable Naming**	`STUDYDAY (WRONG)` `PARAM (label) → fragile`	`ADT` ✓ `PARAMCD (code)` ✓

Output Row Count Evidence:

No-RAG: 1296 rows (includes 59 baseline) | RAG: 594 rows (post-baseline only)

Presentation takeaways

- Self-hosted, open source LLMs play the key role in closed regulatory environments
- Joint LLM + embedding optimization is required for building R-code assistants.
- RAG is required for generation regulatory-compliant clinical R code.
- Expert data enables submission-ready generation.
- **R AI Challenge** is an experiment that is re-producible and scalable withing organizations willing to develop their own clinical R assistant

See our paper ML15 ***"Coding Clinical R with Confidence: Digital AI Assistant based on self-hosted LLMs"*** *for more details*

INTEGO CLINICAL

Questions?

Contact the authors:

Kostiantyn Drach kostya.drach@intego-group.com

Iryna Kotenko irina.kotenko@intego-group.com

Intego Group, LLC, 2300 Maitland Center Pkwy, Ste. 240,

Maitland, FL 32751

Phone: +1 (407) 512-1006

https://integoclinical.com/data-science