

Large Language Models as Medical Writers

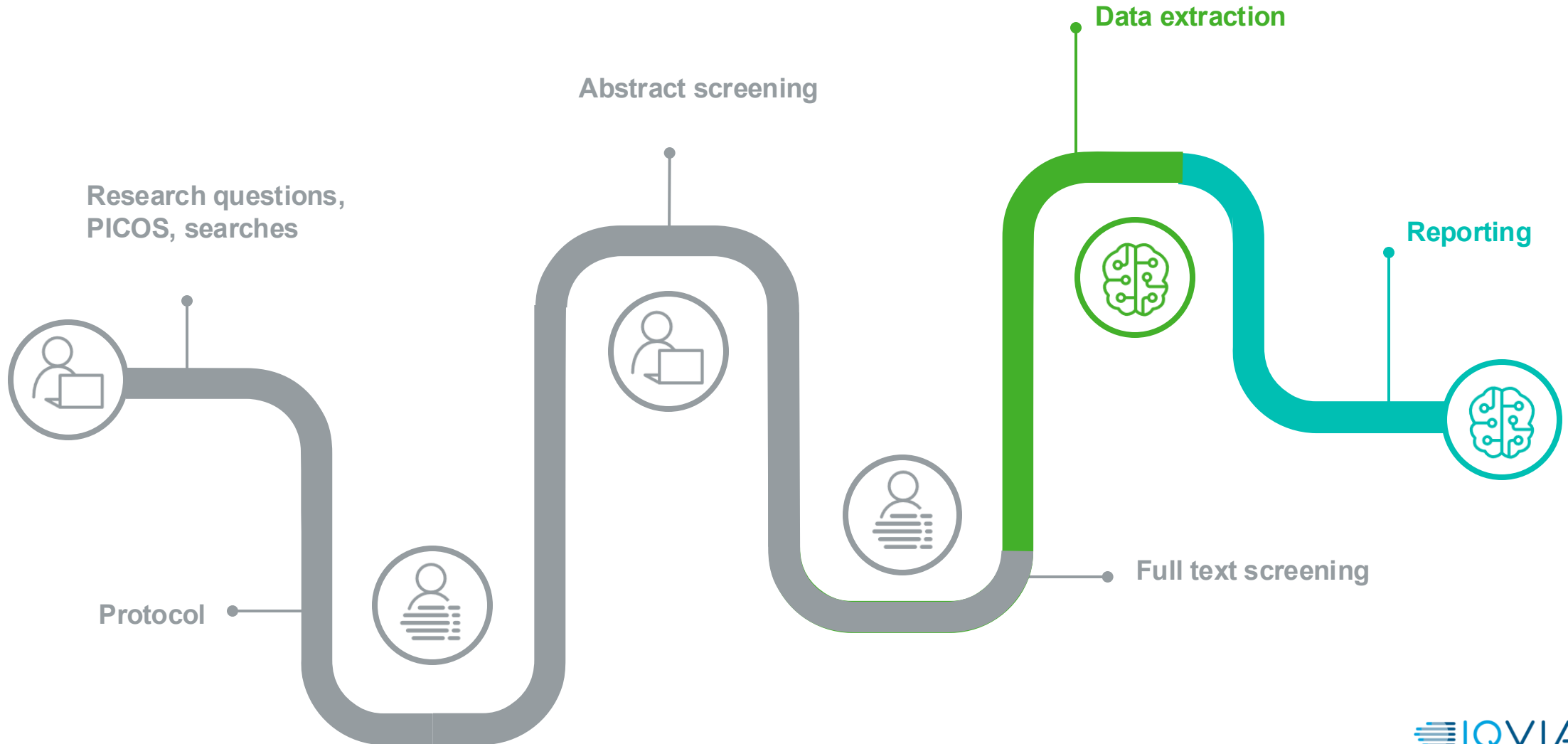
Challenges and Solutions

Ioannis Reklos, Alexandrina Lambova, Kiril Matev, Ketevan Rtveladze,
Julia Gallinaro, Ines Guerra

Introduction

- **Systematic Literature Reviews (SLRs) & HTA dossiers**
 - Cornerstone in evidence generation → underpin reimbursement, pricing, and regulatory decisions
 - Extremely **resource-intensive**: months of expert time, large multi-disciplinary teams

SLRs are composed of multiple steps and today's focus is on automation of data extraction and reporting



Introduction

Report writing for SLRs and HTA dossiers

1. Extraction of variables/information

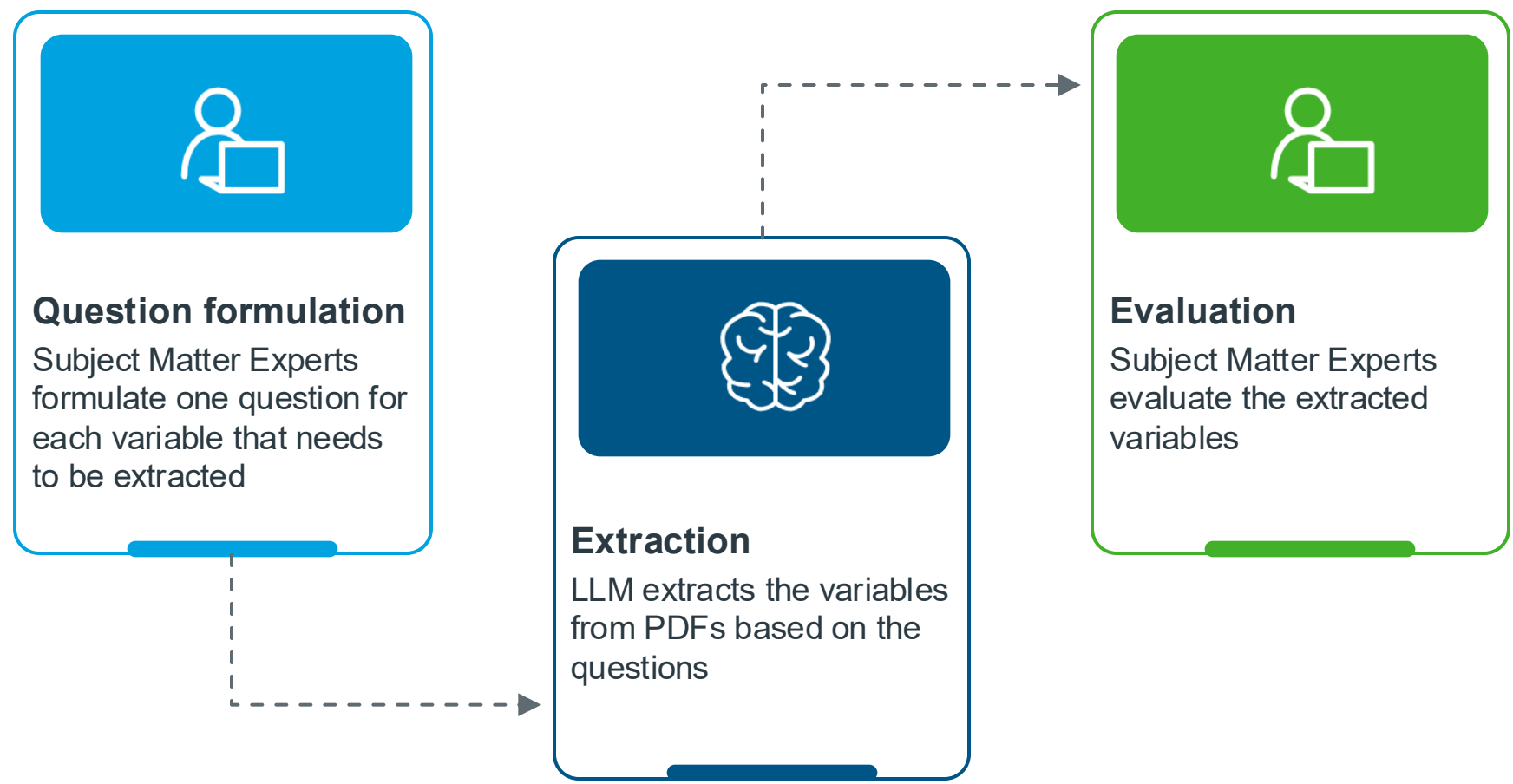
- Can be contained in:
 - › Text
 - › Tables
 - › Figures
- May be extracted from:
 - › digital documents
 - › scanned documents

2. Report writing

- Tone
- Focus
- Vocabulary/language
- Formatting

Goal: Automation using LLMs

Initial Extraction Pipeline



Topic	Total Questions	Not Working	% Not Working
Model Inputs	9	6	66.67%
Model Outcomes	32	9	28.13%

Extraction Issues using Large Language Models

1. Tables

- Table fragmentation
- Header/cell misrecognition

2. Text

- Prompt brittleness
- Retrieval failure

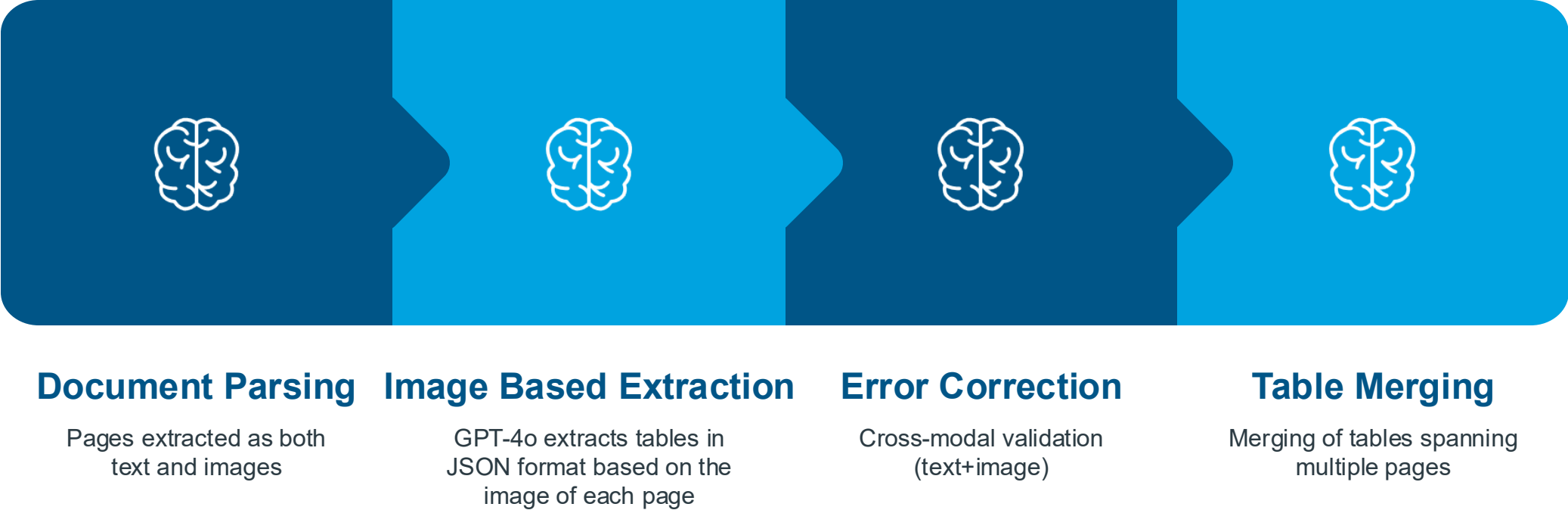
Our Approach

Extraction Issues

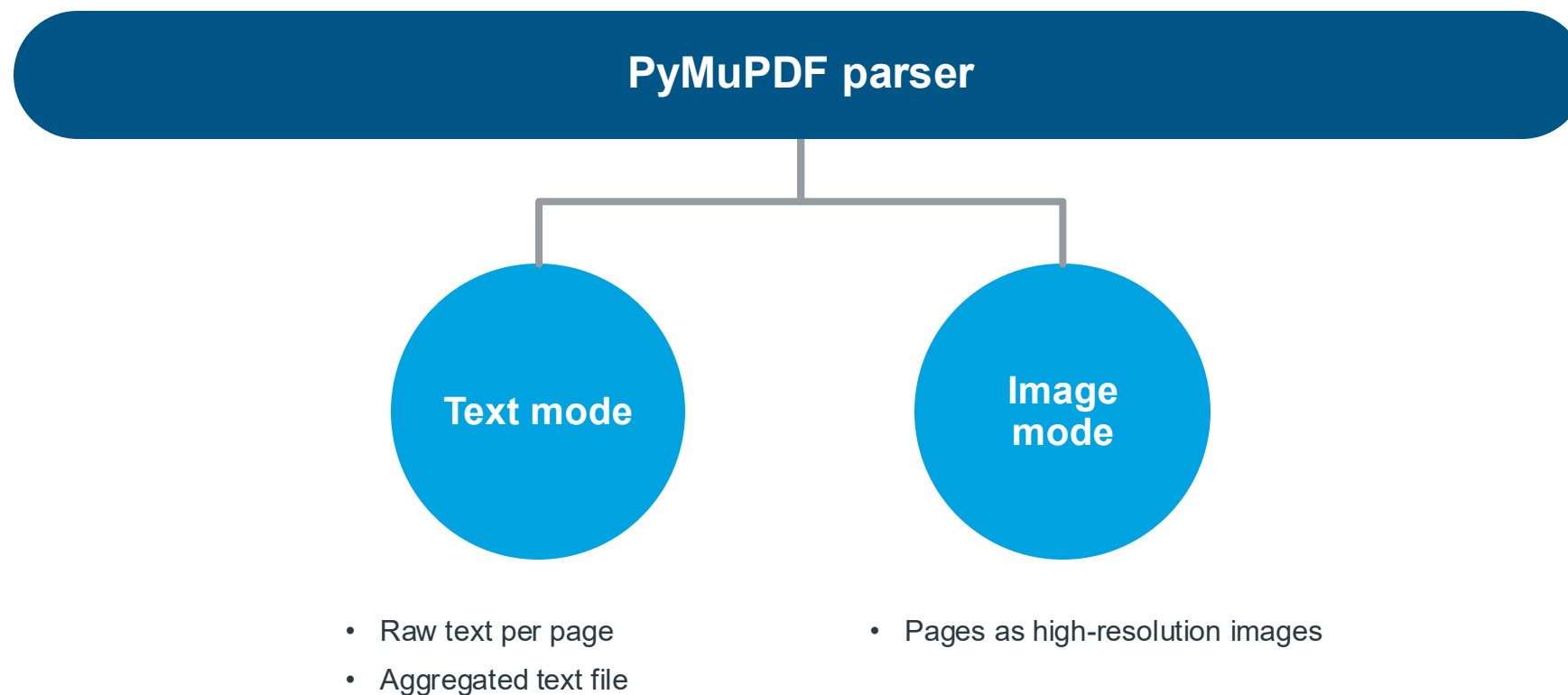
1. Tables → Automated Multi-Modal Table Extraction Pipeline

- Table fragmentation
- Header/cell misrecognition

Table Extraction Pipeline overview

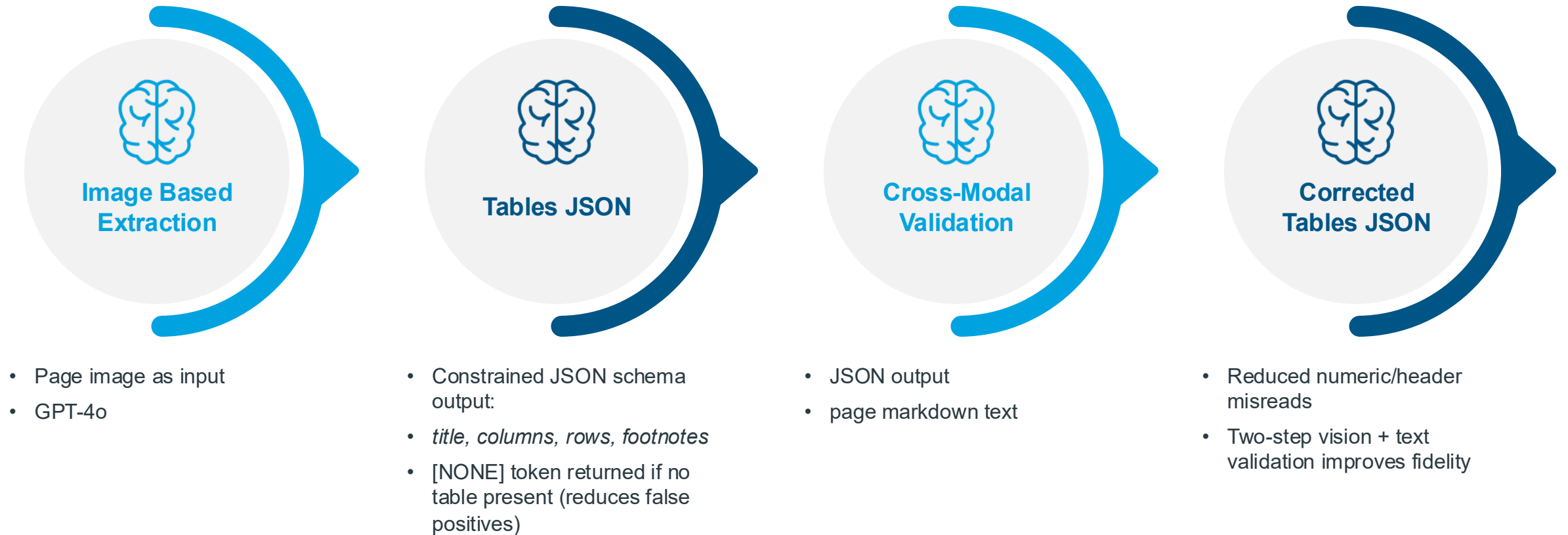


Automated Multi-Modal Table Extraction Pipeline – Step 1



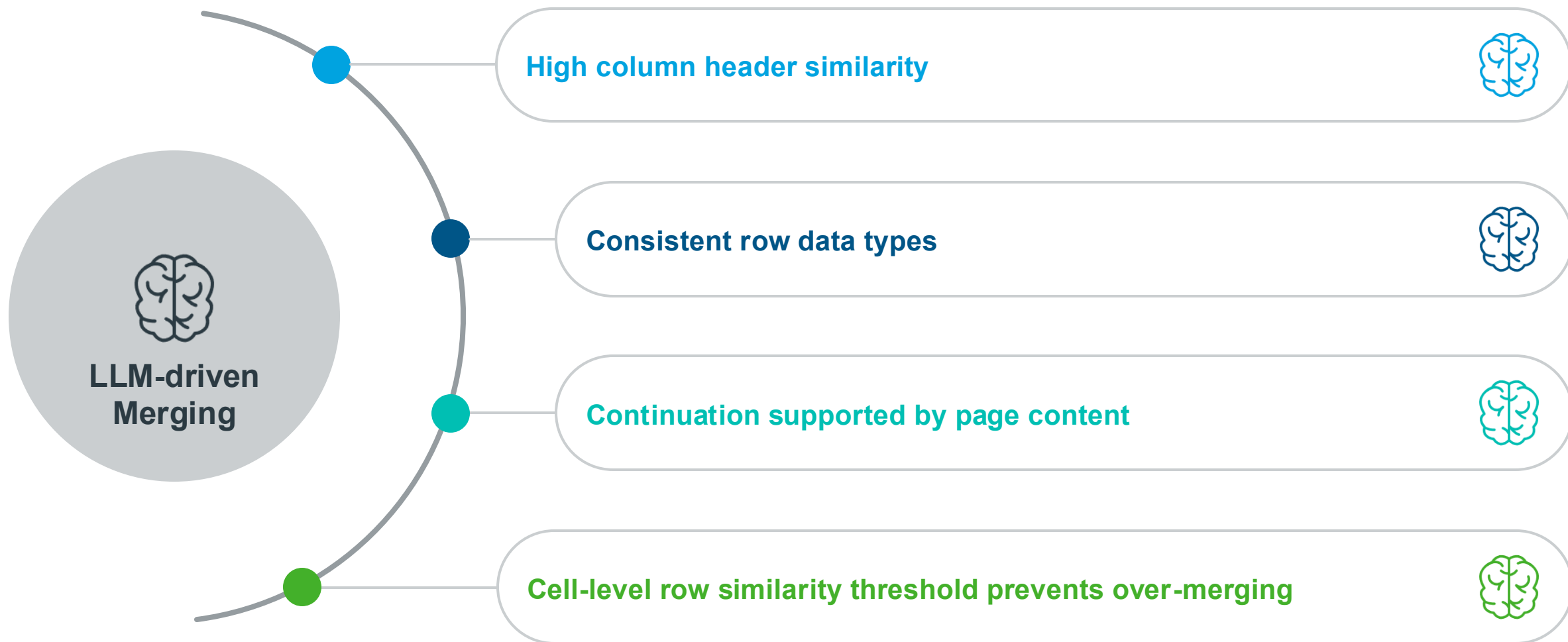
Automated Multi-Modal Table Extraction Pipeline – Step 2&3

LLM Table Extraction & Error Correction



Automated Multi-Modal Table Extraction Pipeline – Step 4

Table Merging Across Pages



Results

Goal: *Extract structured tables from complex scientific PDFs*

Evaluation

Subject Matter Experts based

- Model Inputs – 9 queries
- Model Outcomes – 32 queries
- Comparison scenarios:
 - Extraction prompt + PDF text
 - Extraction prompt + PDF text + Extracted tables

Topic	Total Questions	Not Working	Improved	Worsened	Unchanged	% Improvement
Model Inputs	9	6	2	1	3	16.67%
Model Outcomes	32	9	7	2	23	55.56%

$$\% \text{ improvement} = \frac{\# \text{ improved} - \# \text{ worsened}}{\# \text{ not working}} \times 100$$

Extraction Issues using Large Language Models

1. Tables

- Table fragmentation
- Header/cell misrecognition

2. Text

- Prompt brittleness
- Retrieval failure

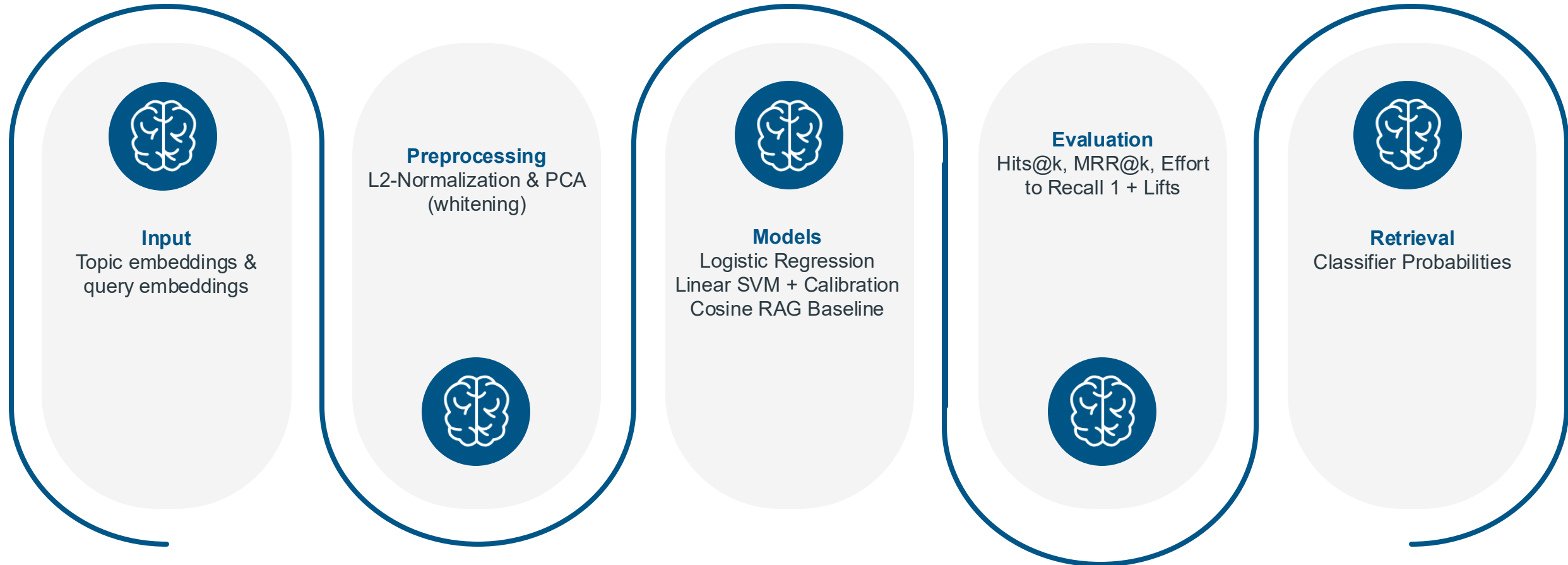
Our Approach

Extraction Issues

2. Text

- Prompt brittleness
- **Retrieval failure → Topic Classifiers to recognize relevant chunks**

Topic Classifiers & Retrieval Pipeline



Topic Classifiers

Goal: *Topic-level retrieval in clinical-style documents*

Training Data:

- ~25 records, each with 3 topics:
 - *Intervention, Patient characteristics, Objective*
- Dataset $\mathcal{D} = \{(x_j, y_j, g_j)\}_{j=1}^n$
 - $x_j \in R^d$ is the topic embedding
 - $y_j \in \mathcal{Y}$ is the topic label
 - $g_j \in \{1, \dots, G\}$ is a group identifier (prevent cross-document leakage in validation)

Query dataset:

- one embedding per class $q_c \rightarrow$ fixed across folds

Intervention	Patient Characteristics	Objective
What is the intervention in the provided documents?	What are the patient characteristics in the provided documents?	What is the objective of the studies?

Topic Classifiers – Preprocessing

Goal: Topic-level retrieval in clinical-style documents

- All models in **scikit-learn Pipelines** (no leakage)

- **Normalization:**

- Unit normalized for cosine geometry:

- $\tilde{x} = \frac{x}{\|x\|_2}$

- $\widetilde{q_c} = \frac{q_c}{\|q_c\|_2}$

- **SafePCA** (fold-safe, with whitening):

- $k_{fit} = \min\{k, n_{train}, d\}$

- Candidate $k \in \{8, 16\}$

- Whitening reduces variance in small-n settings

- **Baseline:** cosine similarity in raw embedding space (no PCA) – RAG default

Topic Classifiers – Models and Validation

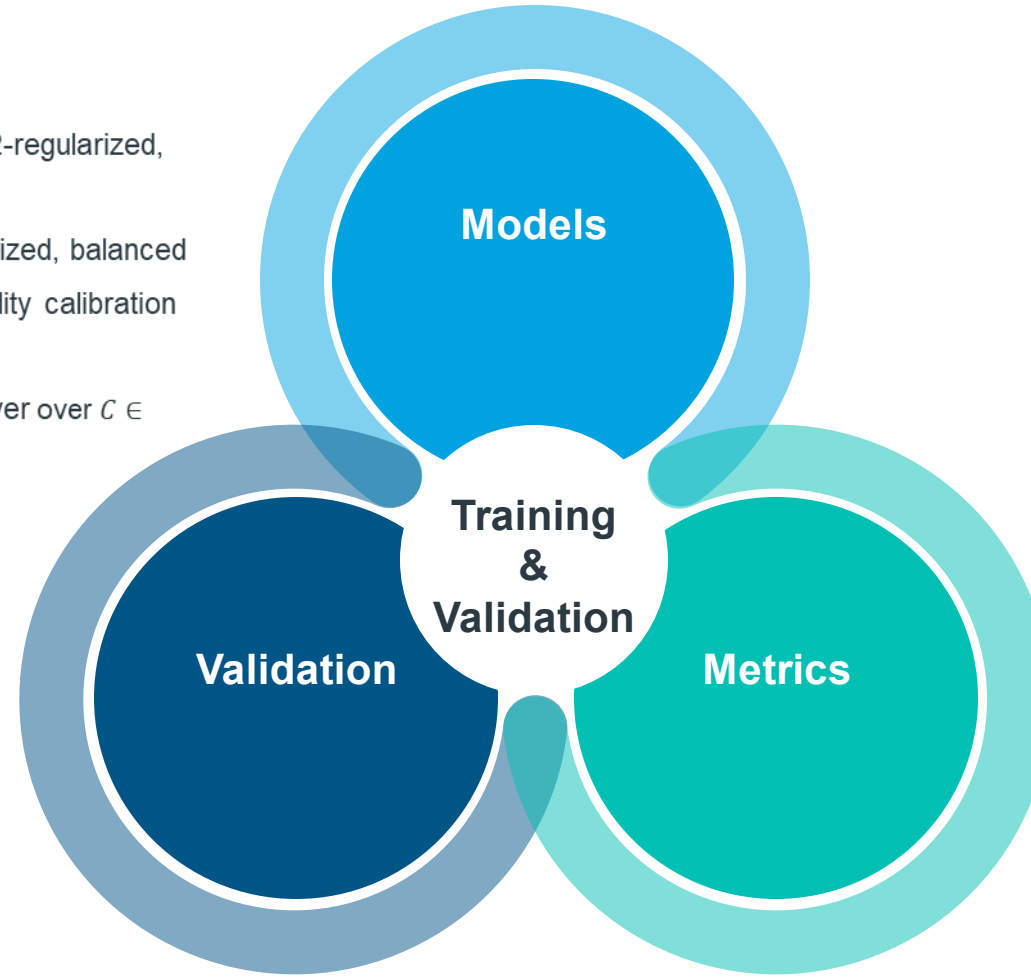
Goal: Topic-level retrieval in clinical-style documents

Linear

- Logistic Regression (L2-regularized, balanced class weights)
- Linear SVM (L2-regularized, balanced class weights) + probability calibration (Platt, 3-fold)
- Regularization tuned over over $C \in \{0.05, 0.1, 0.5, 1.0\}$.

Nested group-aware CV

- Outer: GroupKFold (k=5) at document level
- Inner: GroupKFold (k=4) for hyperparameter tuning



Classification metrics

- macro-F1
- Accuracy
- Precision
- Recall

Topic Classifiers – Retrieval Scenarios

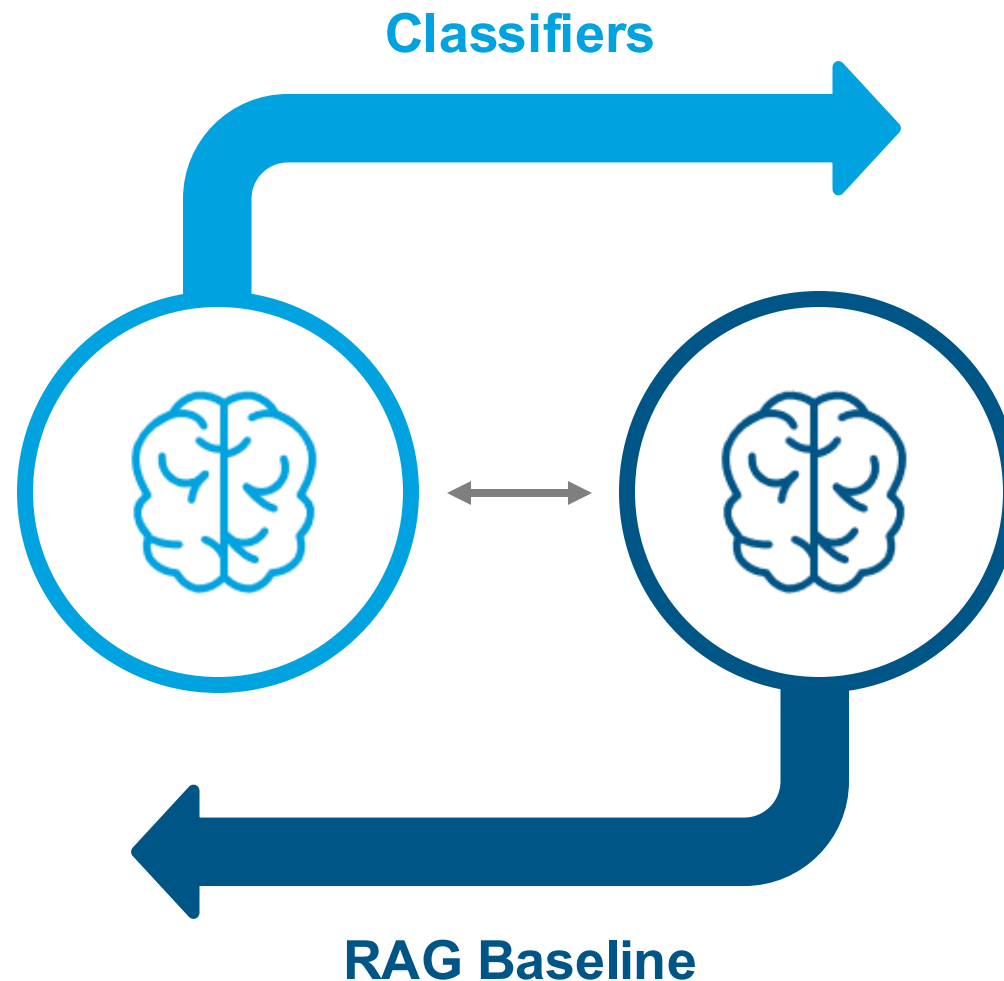
Goal: Topic-level retrieval in clinical-style documents

Score

$$s_{cos}(\tilde{x}; c) = \tilde{q}_c^\top \tilde{x}$$

Ranking

Similarity to query embedding



Score

$$s_{clf}(x; c) = p(y = c \mid x; \theta^*)$$

θ^* : The optimal model parameters

Ranking

Probabilities from LR or calibrated SVM

Preprocessing

Incorporates SafePCA + normalization

Topic Classifiers – Evaluation Metrics & Lift Computation

Goal: Topic-level retrieval in clinical-style documents

Metrics per (fold, class):

- Hits@k (k = 1,3,5)
- MRR@k (earliest relevant item, truncated at k)
- Effort to Full Recall $k_{\text{recall}=1}$

Lift vs cosine baseline:

- Hits@k Lift = $\frac{\text{hits}@k_{clf}}{\text{hits}@k_{cos}}$

- MRR@k Lift = $\frac{\text{MRR}@k_{clf}}{\text{MRR}@k_{cos}}$

- Effort-to-Recall Lift = $\frac{k_{cos}}{k_{clf}}$

Macro summaries: mean of per-(fold,class) ratios

- Treats each class and fold equally

Results – Model Comparison

Goal: Topic-level retrieval in clinical-style documents

Model	Hits@1/Lift	Hits@3/Lift	Hits@5/Lift	MRR@1/Lift	MRR@3/Lift	MRR@5/Lift	Effort@Recall 1/Lift
SVM	1.0 / 1.0	3.0 / 1.57	4.67 / 1.97	1.0 / 1.0	1.0 / 1.21	1.0 / 1.4	5.6 / 1.95
LR	1.0 / 1.0	3.0 / 1.57	4.67 / 1.96	1.0 / 1.0	1.0 / 1.21	1.0 / 1.4	5.7 / 1.91
Cosine (RAG)	0.8	2.1	3.07	0.8	0.86	0.87	10.5

Results – Topic Level Comparison

Goal: Topic-level retrieval in clinical-style documents

Model	Label	Hits@1	Hits@3	Hits@5	MRR@1	MRR@3	MRR@5	Effort@Recall1
SVM	Intervention	1.0	3.0	4.6	1.0	1.0	1.0	5.6
SVM	Objective	1.0	3.0	4.8	1.0	1.0	1.0	5.2
SVM	Patient characteristics	1.0	3.0	4.6	1.0	1.0	1.0	6.0
LR	Intervention	1.0	3.0	4.4	1.0	1.0	1.0	6.0
LR	Objective	1.0	3.0	4.8	1.0	1.0	1.0	5.2
LR	Patient characteristics	1.0	3.0	4.8	1.0	1.0	1.0	5.8
Cosine (RAG)	Intervention	0.4	1.2	2.2	0.4	0.57	0.62	13.0
Cosine (RAG)	Objective	1.0	2.2	2.6	1.0	1.0	1.0	12.8
Cosine (RAG)	Patient characteristics	1.0	3.0	4.4	1.0	1.0	1.0	5.8

Extraction and Generation Issues using Large Language Models

Extraction issues

2. Text

- Prompt brittleness

Generation issues

- Uncontrolled content expansion
- Retrieval failure

Our Approach

Extraction Issues

2. Text

- Prompt brittleness → Automatic generation of extraction prompt based on examples

Generation issues

- Uncontrolled content expansion → Structured outline-based generation
- Retrieval failure

Example Based Text Generation

Goal: LLM based extraction and generation of prompts

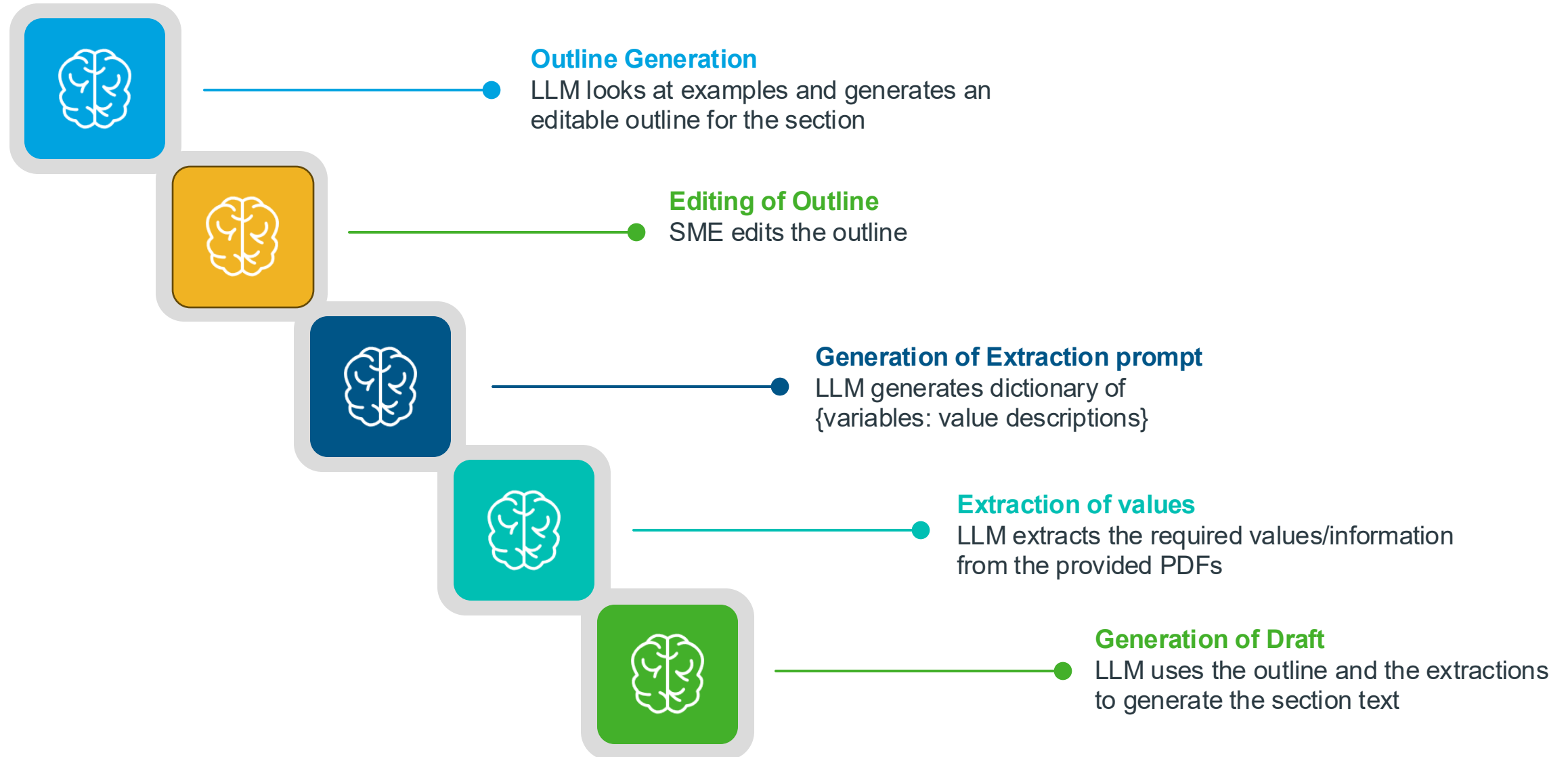
Uncontrolled content expansion

- Generate editable outline for sections based on examples
- Add instructions for tone, language, audience etc.

Prompt brittleness

- Remove the need for SMEs to specify extraction prompts
- Automated generation based on outline

Example Based Text Generation Pipeline



Conclusion

- **Structure-aware extraction** significantly improves parsing of fragmented, multi-page tables and mixed text–table evidence.
- **Topic-guided retrieval** mitigates classical RAG failure modes in both extraction and generation, enhancing evidence precision and recall.
- **Outline-based extraction** reduces SME workload and minimizes prompt brittleness.
- **Outline-constrained generation** enforces section scope, citation behavior, and stylistic consistency in regulatory narratives.
- The framework **reduces SME workload** while shifting expert effort from low-level tasks to high-value oversight.
- Future work will target **multimodal reasoning** to improve extraction accuracy, **section-level retrieval calibration** for evidence completeness, and **augmented outline granularity** to further close the gap with expert-authored content.

Thank you!

If you have any questions, please contact me at ioannis.reklos@iqvia.com