

Single Day Event

Applying RAG in a Regulated
Environment: Lessons from RAGFlow
and LLM Deployment

Jiahui Zhu,
Hengrui Pharmaceuticals



Background and motivation

01

In the statistical analysis department,
knowledge and code are highly fragmented.

Spend more time looking for information rather than analyzing it

02

Existing retrieval methods are inefficient

03

Cloud LLM models limited by compliance environment constraints

- Data leakage risk
- Lack of traceability

We have knowledge, but no “interface of knowledge”





Background and motivation



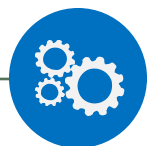
Build a internal knowledge system

Systematically collect and organize statistical and programming knowledge scattered across various projects and documents to form a unified and structured knowledge base



Intelligent retrieval

Leverage the RAGFlow platform combined with locally LLM to implement semantic-based intelligent question answering



Compliance in a regulated environment

Ensure that data does not leave the intranet and output results are traceable

We hope our team can finally communicate with knowledge



Collection and classification of documents



The "source" that determines the quality of knowledge



Classification structure determines retrieval accuracy

Correct classification allows the model to think in the appropriate knowledge pool, improve the accuracy of the answer

Providing structured input for embedding generation and indexing, make retrieval faster and more relevant

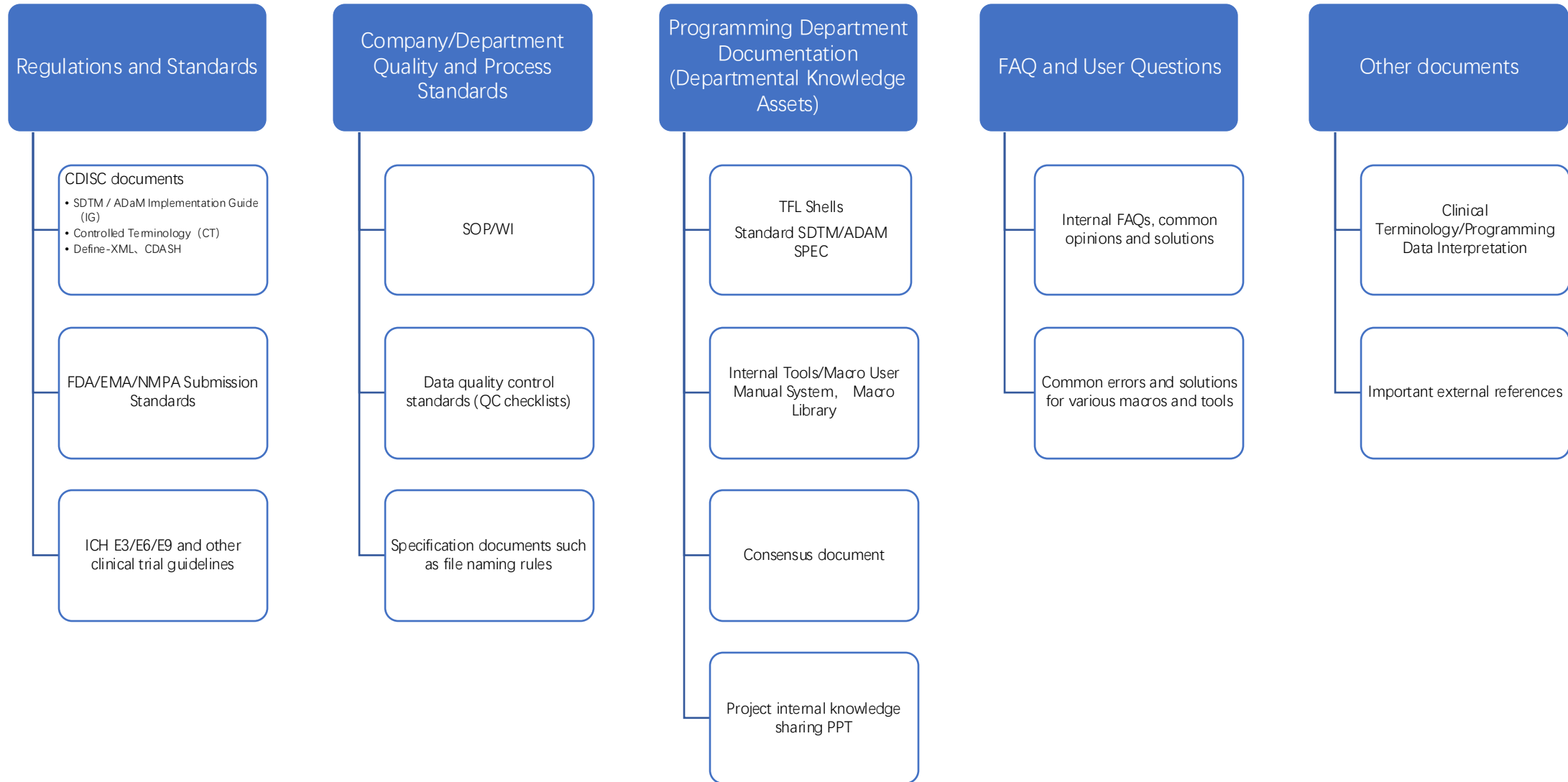


Reduce maintenance costs

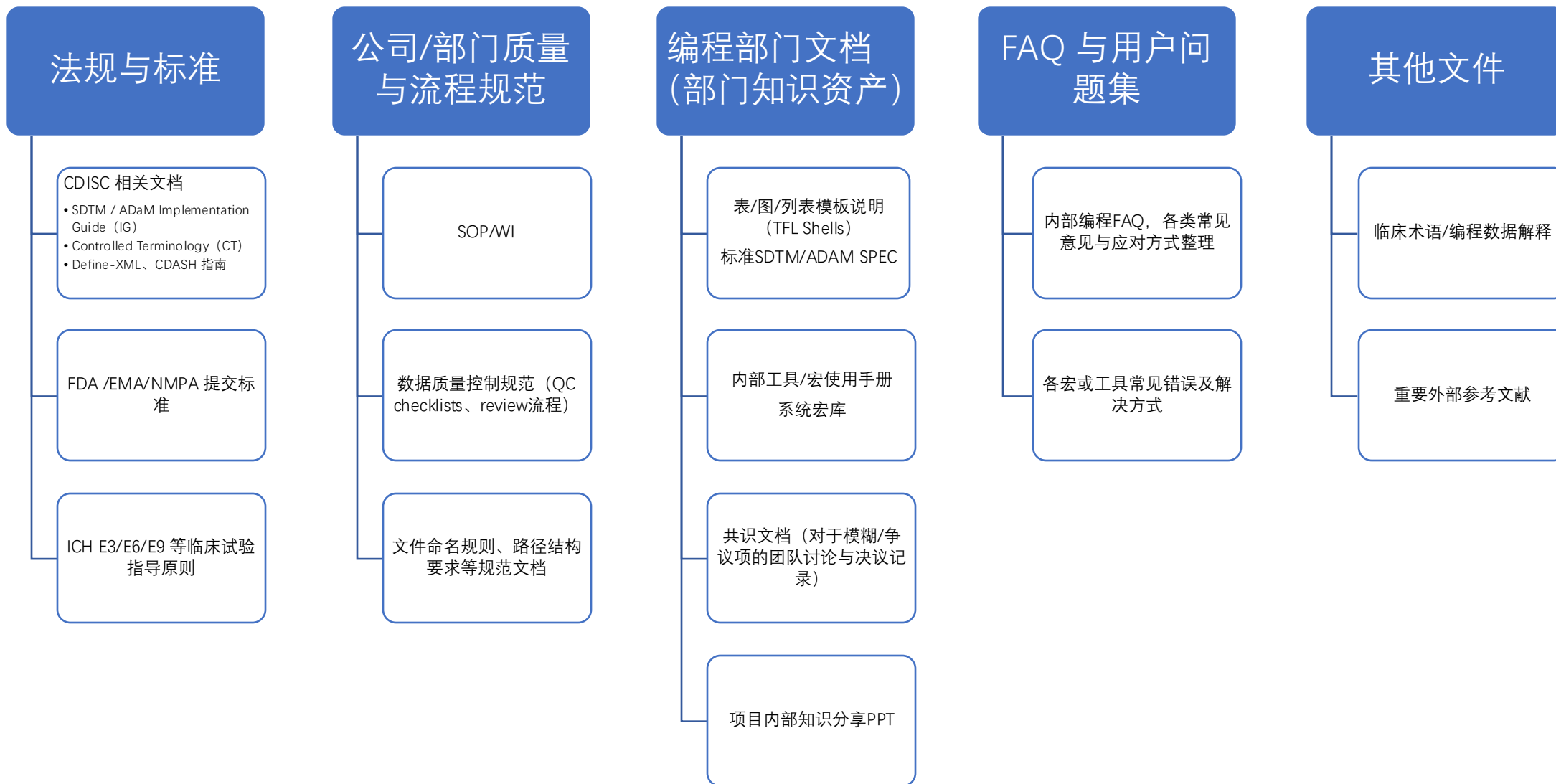
Subsequent updates and reconstructions of the knowledge base all rely on the initial classification system. Establishing consistent directories, labels, and naming conventions can greatly reduce long-term maintenance costs.



>>> Collection and classification of documents



>>> Collection and classification of documents

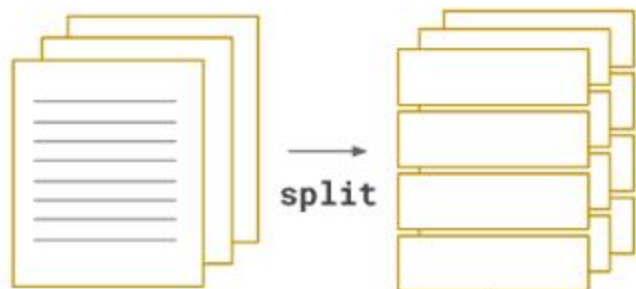


Some experience / tips

- Clarify scope and priorities
- Version control
 - For multi-version files, only the latest valid version + historical retention mark is retained, or the historical version files are stored in cold storage
- Desensitization strategy be implemented at the collection stage
- Ensuring parsability
 - Prioritize the collection of machine-readable (text-based) documents. txt, word, csv take precedence over PDF
- Document Responsibility / Traceability
 - Determine who is the data owner and who is the knowledge base administrator
 - Every time a file is updated, merged, or replaced, it must be written into a log

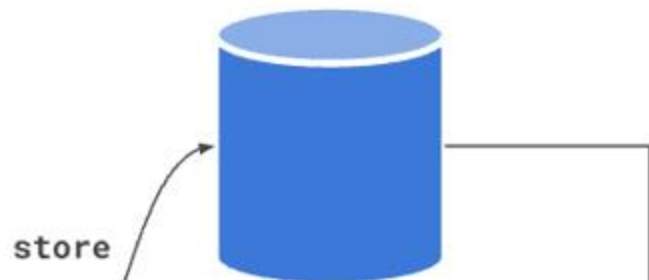
INDEXING VM1: RAGFLOW

Documents



VM3

VectorStore



RETRIEVAL & GENERATION

Embedding model

query

query embedding

Retriever

What are some recent highlights of Together AI?

[0.0234, 1.2356, -0.7812, 0.9941, ...]

augment

Generative model

VM2: DEEPSEEK V3/R1

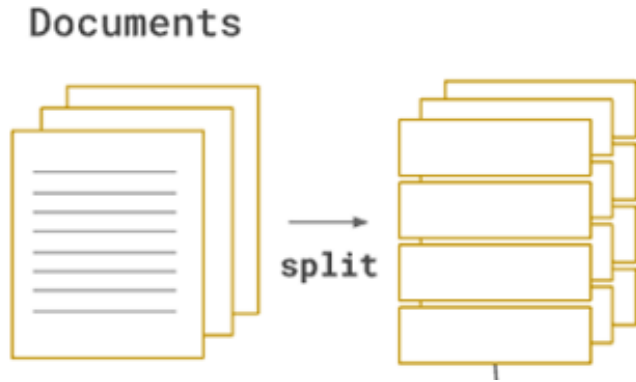
Retrieved context

Together AI provides the fastest cloud platform for building ...
Together AI recently released their StripedHyena models, which ...
Together AI is also the creator of the RedPajama Datasets...
⋮

K8s Cluster

containerd

Document Chunking/Embedding



Why split?

LLM has an input limit and cannot understand dozens of pages of reports at a time

The entire document is too large to generate a single vector, and precise semantic matching is impossible.

Support traceability and reference:

"Results are from CSR Section 5.3 Table 5.3.1"

- The goal of segmentation is **not** "the smaller the better"
An ideal chunk is one that maintain semantic integrity and improve retrieval relevance
- Different types of files require different segmentation methods

Document Chunking Strategy

"General"

Supported file formats include MD, MDX, DOCX, XLSX, XLS (Excel 97-2003), PPT, PDF, TXT, JPEG, JPG, PNG, TIF, GIF, CSV, JSON, EML, and HTML.

The system uses a visual detection model to segment continuous text into multiple segments.

Step 1: Split into pieces.

- ¹ About the company
- ² Dolor enim eu tortor urna sed duis nulla. Aliquam vestibulum, nulla odio nisl vitasusae.
- ³ In aliquet pellentesque aenean hac vestibulum turpis mi bibendum npexat dixeacauism.
- ⁴ Tempor integer aliquam in vitae malesuada fringilla. Elit nisi in eleifend sed nisi. Pulvinar at orci, proin imperdiet commodo consectetur convallis risus.

Step 2: Merge into a chunk.

About the company
Dolor enim eu tortor urna sed duis nulla. Aliquam vestibulum, nulla odio nisl vitasusae.
In aliquet pellentesque aenean hac vestibulum turpis mi bibendum npexat dixeacauism.
Tempor integer aliquam in vitae malesuada fringilla. Elit nisi in eleifend sed nisi.
Ipsum at orci, proin imperdiet commodo consectetur convallis risus.

Document Chunking Strategy

"Table"

Supports [XLSX](#) and [CSV/TXT](#) file formats.

For CSV or txt files that column delimiter is TAB.The first row must contain column headers.Column headers must be meaningful terms for LLM to understand.

Each row in the table will be treated as a chunk.

	A	B	C	D	E	F	G
	Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition
1	C141657		No	10-Meter Walk/Run Functional Test Test Code	TENMW1TC	10-Meter Walk/Run Functional Test Test Code	10-Meter Walk/Run test code.
2	C174106	C141657		10-Meter Walk/Run Functional Test Test Code	TENMW101	TENMW1-Was Walk/Run Performed	10-Meter Walk/Run - Was the 10-meter walk/run performed?
3	C141700	C141657		10-Meter Walk/Run Functional Test Test Code	TENMW102	TENMW1-Time to Walk/Run 10 Meters	10-Meter Walk/Run - If yes, time to walk or run 10 meters.
4	C147592	C141657		10-Meter Walk/Run Functional Test Test Code	TENMW103	TENMW1-Wear Orthoses	10-Meter Walk/Run - If yes, did subject wear orthoses?
5	C141701	C141657		10-Meter Walk/Run Functional Test Test Code	TENMW104	TENMW1-Test Grade	10-Meter Walk/Run - Test grade.
6	C141656		No	10-Meter Walk/Run Functional Test Test Name	TENMW1TN	10-Meter Walk/Run Functional Test Test Name	10-Meter Walk/Run test name.
7	C141701	C141656		10-Meter Walk/Run Functional Test Test Name	TENMW1-Test Grade	TENMW1-Test Grade	10-Meter Walk/Run - Test grade.
8	C141700	C141656		10-Meter Walk/Run Functional Test Test Name	TENMW1-Time to Walk/Run 10 Meters	TENMW1-Time to Walk/Run 10 Meters	10-Meter Walk/Run - If yes, time to walk or run 10 meters.
9	C174106	C141656		10-Meter Walk/Run Functional Test Test Name	TENMW1-Was Walk/Run Performed	TENMW1-Was Walk/Run Performed	10-Meter Walk/Run - Was the 10-meter walk/run performed?
10	C147592	C141656		10-Meter Walk/Run Functional Test Test Name	TENMW1-Wear Orthoses	TENMW1-Wear Orthoses	10-Meter Walk/Run - If yes, did subject wear orthoses?
11	C141663		No	4-Stair Ascend Functional Test Test Code	A4STR1TC	4-Stair Ascend Functional Test Test Code	4-Stair Ascend test code.
12	C174103	C141663		4-Stair Ascend Functional Test Test Code	A4STR101	A4STR1-Was 4-Stair Ascend Performed	4-Stair Ascend - Was the 4-stair ascend performed?
13	C141706	C141663		4-Stair Ascend Functional Test Test Code	A4STR102	A4STR1-Time to Do 4-Stair Ascend	4-Stair Ascend - If yes, time taken to do 4-stair ascend.
14	C147590	C141663		4-Stair Ascend Functional Test Test Code	A4STR103	A4STR1-Wear Orthoses	4-Stair Ascend - If yes, did subject wear orthoses?
15	C141707	C141663		4-Stair Ascend Functional Test Test Code	A4STR104	A4STR1-Test Grade	4-Stair Ascend - Test grade.
16	C141662		No	4-Stair Ascend Functional Test Test Name	A4STR1TN	4-Stair Ascend Functional Test Test Name	4-Stair Ascend test name.
17	C141707	C141662		4-Stair Ascend Functional Test Test Name	A4STR1-Test Grade	A4STR1-Test Grade	4-Stair Ascend - Test grade.
18	C141706	C141662		4-Stair Ascend Functional Test Test Name	A4STR1-Time to Do 4-Stair Ascend	A4STR1-Time to Do 4-Stair Ascend	4-Stair Ascend - If yes, time taken to do 4-stair ascend.
19	C174103	C141662		4-Stair Ascend Functional Test Test Name	A4STR1-Was 4-Stair Ascend Performed	A4STR1-Was 4-Stair Ascend Performed	4-Stair Ascend - Was the 4-stair ascend performed?
20	C147590	C141662		4-Stair Ascend Functional Test Test Name	A4STR1-Wear Orthoses	A4STR1-Wear Orthoses	4-Stair Ascend - If yes, did subject wear orthoses?
21	C141661		No	4-Stair Descend Functional Test Test Code	D4STR1TC	4-Stair Descend Functional Test Test Code	4-Stair Descend test code.
22	C174104	C141661		4-Stair Descend Functional Test Test Code	D4STR101	D4STR1-Was 4-Stair Descend Performed	4-Stair Descend - Was the 4-stair descend performed?
23	C141704	C141661		4-Stair Descend Functional Test Test Code	D4STR102	D4STR1-Time to Do 4-Stair Descend	4-Stair Descend - If yes, time taken to do 4-stair descend.
24	C147591	C141661		4-Stair Descend Functional Test Test Code	D4STR103	D4STR1-Wear Orthoses	4-Stair Descend - If yes, did subject wear orthoses?
25	C141705	C141661		4-Stair Descend Functional Test Test Code	D4STR104	D4STR1-Test Grade	4-Stair Descend - Test grade.
26	C141660		No	4-Stair Descend Functional Test Test Name	D4STR1TN	4-Stair Descend Functional Test Test Name	4-Stair Descend test name.

切片结果

查看用于嵌入和召回的切片段落。

全文 省略

搜索

选择所有

☐ Code:C52756; Codelist Code:C74456; Codelist Name:Anatomical Location; CDISC Submission Value:SKIN OF THE NECK; CDISC Definition:The integument that covers the neck.; NCI Preferred Term:Neck Skin

☐ Code:C164047; Codelist Code:C74456; Codelist Name:Anatomical Location; CDISC Submission Value:SKIN OF THE LOWER LIMB; CDISC Definition:The integument that covers the lower limb.; NCI Preferred Term:Skin of the Lower Extremity

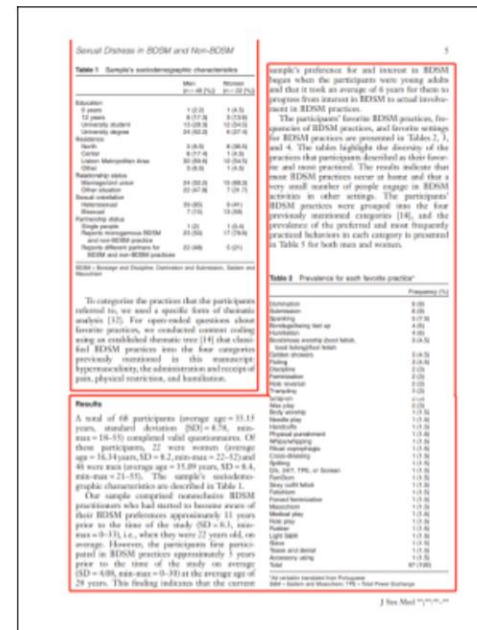
☐ Code:C12291; Codelist Code:C74456; Codelist Name:Anatomical Location; CDISC Submission Value:SKIN OF THE LIP; CDISC Definition:The skin and underlying tissue of the portion of the lip that contains hair.; NCI Preferred Term:Skin of the Lip

☐ Code:C161380; Codelist Code:C74456; Codelist Name:Anatomical Location; CDISC Submission Value:SKIN OF THE INGUINAL REGION; CDISC Definition:The integument that covers the inguinal region.; NCI Preferred Term:Skin of the Inguinal Region

"Paper"

Only PDF files are supported.

The paper will be segmented into its sections, allowing the LLM to better summarize the content of the relevant chapters in the paper and produce more comprehensive answers.



Document Chunking/Embedding

Size: 12 KB Uploaded Time: 25/08/2025 22:41:49

```
%if %nrbquote(&STUDYID_VAR) = %then %do;
  %put;
  %put
=====;
  %put %str(ERR)%str(OR): The parameter STUDYID_VAR was not provided.
;
  %put %str(ERR)%str(OR): Since this is a required parameter
;
  %put %str(ERR)%str(OR): Program will abort due to a fatal %str(ERR)%str(OR).
;
  %put
=====;
  %put;
  %goto src_t_bimo_csl_t2_exit;
%end;
%if %u_var_exist (dsin = _adsl ,varin = &STUDYID_VAR ,dispval = 1) eq 0 %then
%do;
  %put;
  %put
=====;
  %put %str(ERR)%str(OR): Variable %kuptime(&STUDYID_VAR) does not exist in
data _ADSL. ;
  %put %str(ERR)%str(OR): Program will abort due to a fatal %str(ERR)%str(OR).
;
  %put
=====;
  %put;
  %goto src_t_bimo_csl_t2_exit;
%end;
```

*Chunking method ⓘ

General

PDF parser ⓘ

DeepDoc

*Embedding model ⓘ

 bge-m3

Recommended
chunk size



640

*Delimiter for text ⓘ

``%macro`` ``%mend;`` ``data `` ``proc `` ``run;`` ``quit;``

Page rank ⓘ



0

Auto-keyword ⓘ



10

Auto-question ⓘ

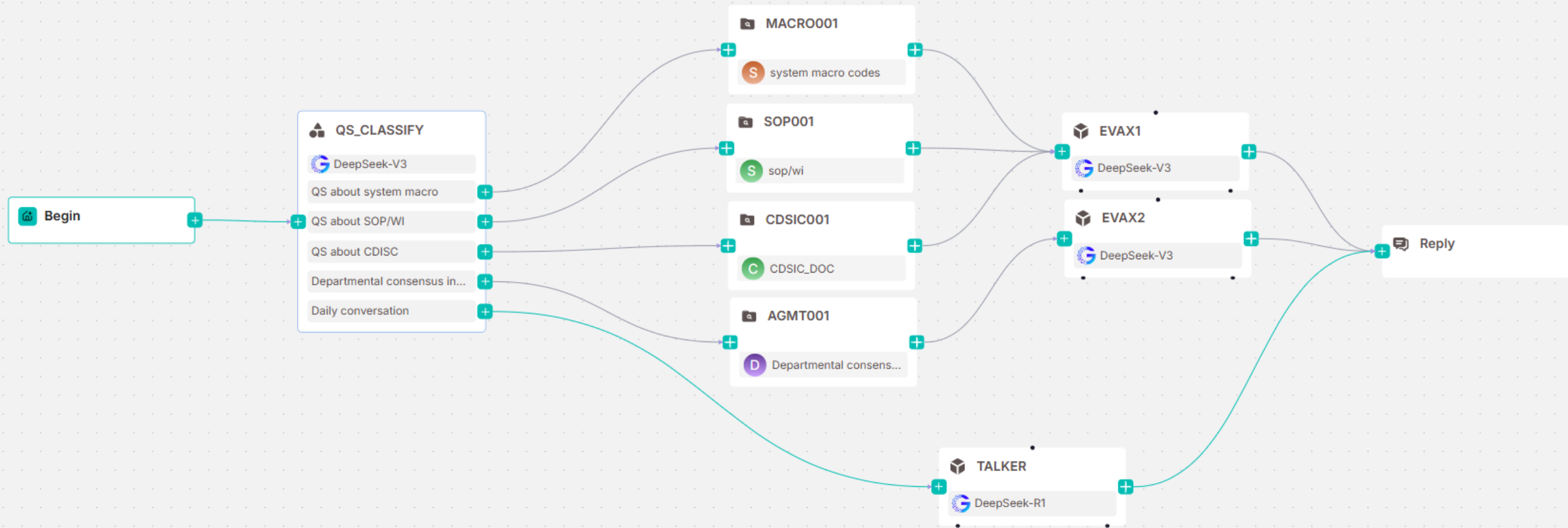


0

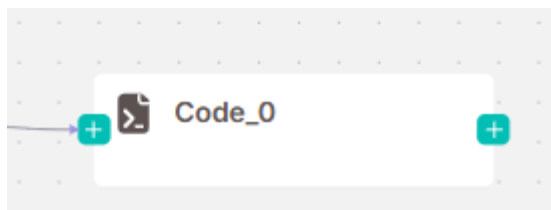
Excel to HTML ⓘ



➤➤➤ RAGFLOW Design



➤➤➤ RAGFLOW Design



ID: Code_0

Write your custom Python or Javascript logic.

Input

arg1

select value

×

arg2

select value

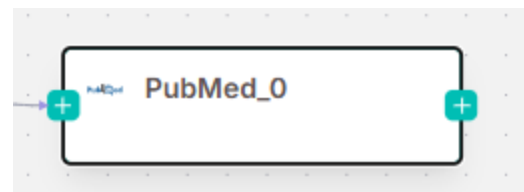
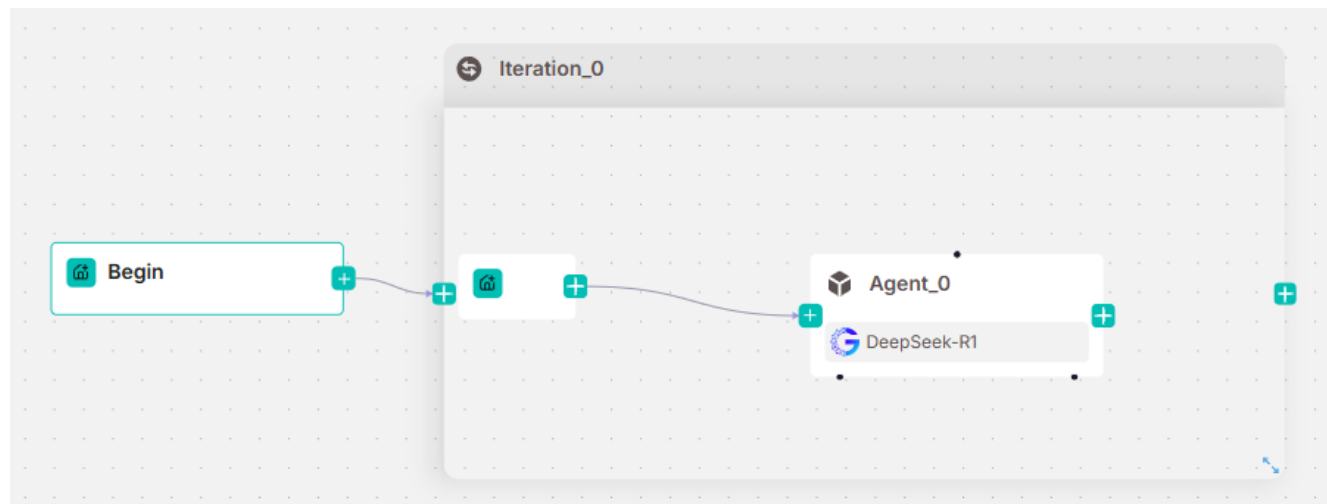
×

+ Add variable

Code

python ▾

```
1 def main(arg1: str, arg2: str) -> str:
2     return f"result: {arg1 + arg2}"
3
```

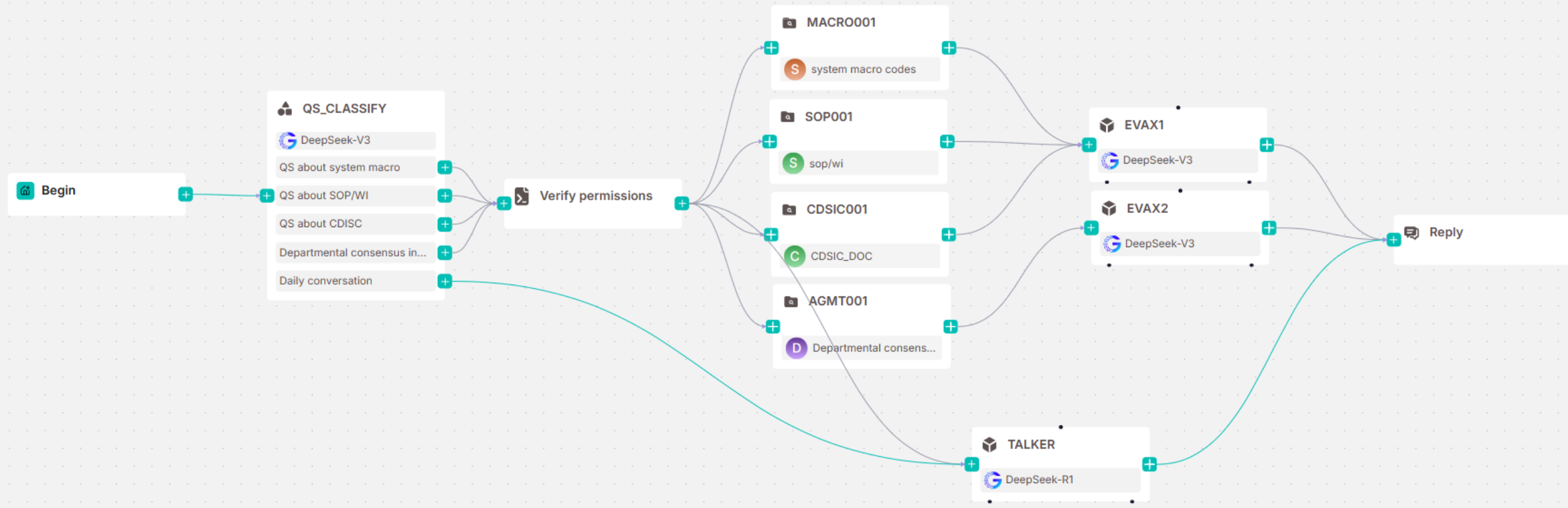


A component that searches from <https://pubmed.ncbi.nlm.nih.gov/>, allowing you to specify the number of search results using TopN. It supplements the existing knowledge bases.

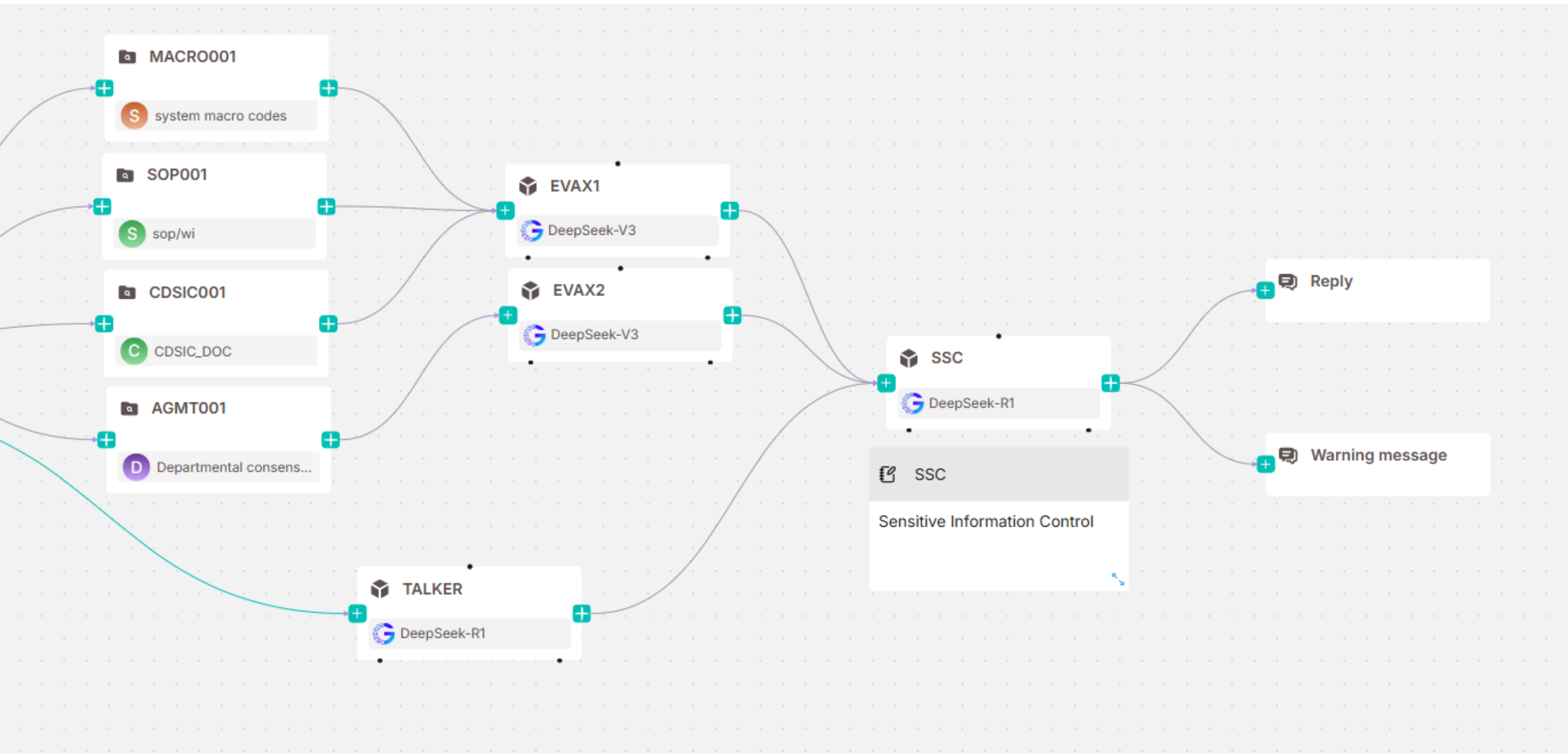
Query variables ⓘ

sys.query

➤➤➤ RAGFLOW Design



➤➤➤ RAGFLOW Design





根据SDTMIG_v3.4规范的相关内容，针对受试者在家自行给药时是否应在EX域中强制补充VISIT/VISITNUM的问题，总结如下：

基本原则

当给药行为完全发生在临床访视期间（如诊所注射）时，EX域中应包含VISITNUM作为时间变量，同时可补充VISITDY和VISIT①。

若给药行为不局限于临床访视时间范围内（如受试者在家服药），则**不应**在EX域中包含VISITNUM，因为EX域用于记录治疗暴露的时序，而非药物分发时序①。

替代方案

对于非访视期间的给药，应通过以下变量记录时序：

使用EXSTDTC/EXENDTC记录实际给药起止日期时间①

采用时间点变量（如--TPT, --TPTNUM, --ELTM, --TPTREF）表示相对于参考事件（如首次给药）的时间①

特殊情况处理

若协议要求基于非访视事件（如不良事件阈值）触发给药，可通过以下方式记录：

在EC域中记录计划/实际给药事件（使用ECMOOD区分）①

通过RELTYPE在RELREC中建立EC与EX域的关联①①



某试验允许受试者在家自行给药（不一定发生在临床访视期间）。是否应该在 EX 中强行补充 VISIT/VISITNUM?



Log



Begin



QS_CLASSIFY 0.5034s



CDSIC001 20.704s



EVAX1 17.108s



Reply 17.103s



聊天

者在家自行给药时是否应在EX域中强制补充VISIT/VISITNUM的问题，总结如下：

基本原则

当给药行为完全发生在临床访视期间（如诊所注射）时，EX域中应包含VISITNUM作为时间变量，同时可补充VISITDY和VISIT[!]。



3. Categorization and grouping a. EXCAT and EXSCAT may be used when appropriate to categorize treatments into categories and subcategories. For example, if a study contains several active comparator medications, EXCAT may be set to "ACTIVE COMPARATOR". Such categorization may not be useful in all studies, so these variables are permissible. 4. Timing variables a. The timing of exposure to study treatment is captured by the start/end date and start/end time of each constant-dosing interval. If the subject is only exposed to study medication within a clinical encounter (e.g., if an injection is administered at the clinic), VISITNUM may be added to the domain as an additional timing variable. VISITDY and VISIT would then also be permissible qualifiers. However, if the beginning and end of a constant-dosing interval is not confined within the time limits of a clinical encounter (e.g., if a subject takes pills at home), then it is not appropriate to include VISITNUM in the EX domain. This is because EX is designed to capture the timing of exposure to treatment, not the



SDTMIG_v3.4.pdf

关键引用依据

"If the beginning and end of a constant-

dosing interval is not confined within the

Parameter Debugging

Change the parameters of the retrieval model and LLM to adjust *accuracy* and *response time*

Knowledge bases ⓘ

system macro co... ✕



Model

DeepSeek-V3 ▾

Creativity

Balance ▾

Temperature ⓘ



0.5

Top P ⓘ



0.85

Presence penalty ⓘ



0.2

Frequency penalty ⓘ



0.3

Max tokens ⓘ



4096

But How to find the balance point?

Advanced Settings

Similarity threshold ⓘ



0.2

Keyword similarity weight ⓘ



0.7

Top N ⓘ



8

Rerank model ⓘ

bge-reranker-v2-m3 ✕ ▾



Top-K ⓘ



1024

Parameter Debugging

Parameter

Accuracy Impact

Response Speed Impact

Top N

Larger Top N improves accuracy by considering more candidate chunks.

Higher Top N increases processing time due to more input for LLM.

Reranker

Enabling reranker significantly improves accuracy by re-ranking results.

Enabling reranker increases latency due to additional model inference.

keywords_similarity_weight

Higher weight improves accuracy in exact keyword matching scenarios.

Little impact on speed but can introduce noise if set too high.

vector_similarity_weight

Higher weight improves accuracy for semantic matching (synonyms, intent).

Little speed impact, but larger candidate sets may slow down processing.

similarity_threshold

Higher threshold reduces noise by ensuring only highly relevant chunks.

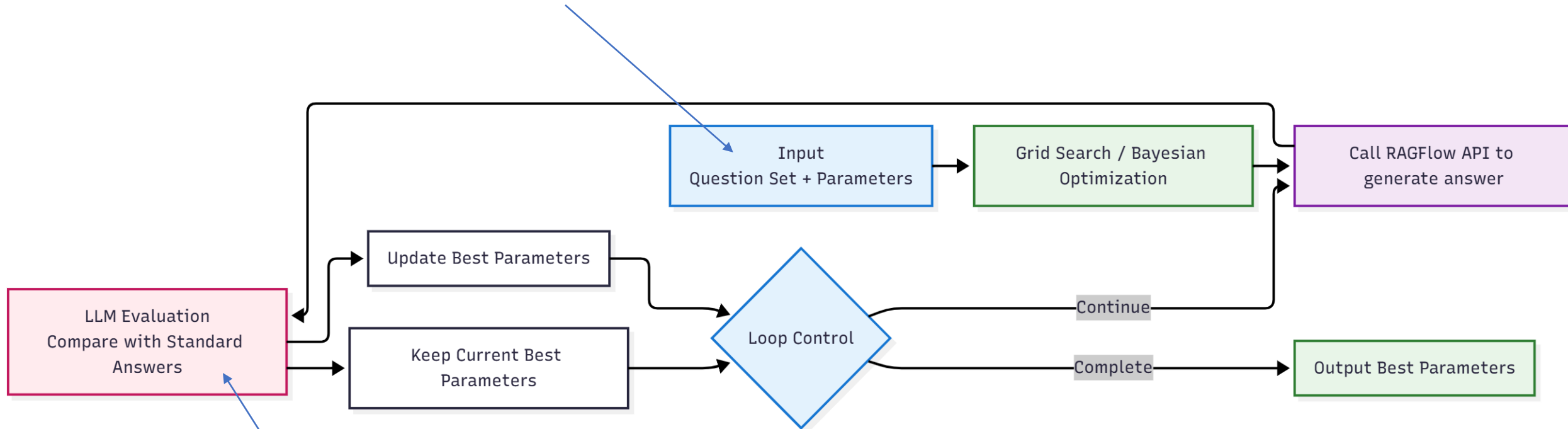
Lower thresholds increase candidate chunks, slightly affecting speed.

Use knowledge graph

Enabling improves accuracy by considering relationships between entities.

Significantly increases processing time due to extra steps for reasoning.

- Ground truth with 120 questions
- Tunable parameters for RAGFlow (e.g. chunk_size, retrieval_top_k, similarity_threshold, reranker, temperature, etc.)



Design a Scoring Rules : e.g. $\text{Score} = 0.6 * \text{Accuracy score} - 0.4 * \log(\text{Response Time})$

Parameter Debugging

Question set

Category	Question Level	Question	Answer
SDTM-related	1	What are the length and character rules for --TESTCD?	≤ 8 characters; cannot start with a number; only letters, digits, and underscores are allowed.
SDTM-related	2	When should a non-standard variable be placed in SUPP-- instead of the main DOMAIN?	When the variable is not included in the domain's standard variable set but must still be submitted, it should be placed in SUPP-- (supplemental qualifiers).
SDTM-related	2	What are the main changes in the latest SDTMIG (v3.4)?	Major updates: added 4 new domains (e.g., BS, CP), removed one domain (MO), introduced a new variable --CLSIG, and significantly restructured several domains such as SV.
DEFINE-related	2	What are common values for Origin in Define.xml?	According to Define-XML v2.1: SDTM: Collected, Derived, Assigned, Protocol, Predecessor; ADaM: Derived, Assigned, Predecessor.
ADaM-related	1	When should the BASETYPE variable be used in an ADaM dataset?	When the same analysis parameter has multiple baseline definitions within one dataset, BASETYPE should be used to distinguish different baseline types.
SDTM-related	4	If a study allows subjects to self-administer medication at home (not necessarily during a clinic visit), should VISIT/VISITNUM be populated in EX?	No. Only when dosing clearly occurs during a clinical visit and is limited to that visit should VISIT/VISITNUM be filled in EX. If dosing is not restricted to visits, these variables should not be artificially added.
SDTM-related	3	If a lab record has both numeric and text results, e.g. LBORRES = '>1000' and conversion factor = 0.1, how should LBSTRESN and LBSTRESC be derived?	Leave LBSTRESN blank; in LBSTRESC, convert the numeric portion using the factor and keep the text part, e.g. '>100'.
SDTM-related	4	The CRF collects two tests: Troponin I and High Sensitivity Troponin I. How should these be represented in SDTM LB domain?	Both correspond to the same biological analyte, so map them as LBTESTCD = 'TROPONT', LBTEST = 'Troponin I'. Different sensitivities are distinguished via LBTMTHSN — set 'High Sensitivity' for high-sensitivity tests and leave blank for regular tests.
SDTM-related	4	In a clinical study, some lab tests are not performed (e.g., due to technical reasons). How should this be represented in SDTM? Is it correct to put 'NOT APPLICABLE' in LBSTRESC?	Do not enter 'NOT APPLICABLE' in LBSTRESC. Per SDTMIG v3.4, if a test is not performed, set LBSTAT = 'NOT DONE' and provide the reason (e.g. 'technical reason') in LBREASND.

Parameter Debugging

Question set

分类	问题等级	问题	答案
SDTM相关	1	--TESTCD 的长度与字符规则是什么？	≤8 个字符；不能以数字开头；只允许字母/数字/下划线。
SDTM相关	2	什么时候应该把非标准变量放入 SUPP-- 而不是主DOMAIN？	当该变量不在该域允许的标准变量集合中，但又需要提交时，放入 SUPP--（补充限定符）。
SDTM相关	2	最新版的SDTMIG (v3.4) 主要有什么变化？	主要变化包括：新增4个域（如 BS, CP），删除1个域（MO），引入新变量 --CLSIG，并对 SV 等域的结构进行了重大调整。
DEFINE相关	2	在 Define.xml 中，Origin 常见取值有哪些？	根据define-XML v2.1, SDTM: Collected、Derived、Assigned、Protocol、Predecessor ADaM: Derived、Assigned、Predecessor
ADAM相关	1	在 ADaM 数据集中，什么情况下需要使用 BASETYPE 变量？	当同一分析参数在同一数据集中存在多种基线定义时，应使用 BASETYPE 来标识不同的基线类型。
SDTM相关	4	某试验允许受试者在家自行给药（不一定发生在临床访视期间）。是否应该在 EX 中强行补充 VISIT/VISITNUM？	不应该强行补充。只有当给药明确发生在一次临床就诊/访视中并受其所限时，才在 EX 中填写 VISIT/VISITNUM；若给药并非限定于访视场景，就不应人为添加访视变量。
SDTM相关	3	如果一条实验室记录同时有数值结果和字符型文本结果，比如LBBORRES = ">1000"，转换系数factor=0.1，则如何衍生对应的lbbstresn和lbbstresc？	lbbstresn留空，lbbstresc中数值部分经过转换因子转换，同时保留字符部分为">100"
SDTM相关	4	CRF 中收集了两个检测项：肌钙蛋白 I（Troponin I）和高敏肌钙蛋白 I（High Sensitivity Troponin I），在 SDTM LB 域中应如何呈现？	在 SDTM 中，这两项检测都属于同一个生物学指标，因此都应映射为LBTESTCD="TROPONI"、LBTEST="Troponin I"。两者区别在于检测方法的灵敏度，应通过变量 LBTMTHSN 进行标识，高敏检测填写为 "HIGH SENSITIVITY"，而常规检测则留空。
SDTM相关	4	在一项临床试验中，研究者希望在 LB（实验室结果）中表示某些实验室指标的检测结果“不适用”，例如由于技术原因该项检测并未开展。在 SDTM 中应如何表示这种情况？是直接在 LBSTRESC 写上 "NOT APPLICABLE"，还是有推荐的做法？	不应在 LBSTRESC 中直接写入 "NOT APPLICABLE"。按照 SDTMIG v3.4 的指导，若测试未进行，应在 LBSTAT 赋值为 "NOT DONE"，并可在 LBREASND 中给出原因（如“技术原因”）。

➤➤➤ Future Directions

General-Purpose Models have weak understanding of domain-specific terms / abbreviations; prone to misinterpretation .

AE → Account Executive?

- **Introducing Domain-Specific Models**

└→ Fine-tune with internal corporate documents (SOPs, CSRs, regulatory templates) and Industry-related expertise.

- **Introducing human-in-the-loop**

"Is this recalled document correct/incorrect/irrelevant?"

└→ Update the weighting parameters for embedding similarity
Automatically adjust chunk splitting or indexing strategies

- **Quick knowledge base maintenance**

└→ Users quickly intervene in the knowledge base in the feedback system: Specify "authoritative documents"; manually fill in knowledge nodes on specific topics, etc.

Q&A

竺嘉辉

+86 18583509508

jiahui.zhu.jz53@hengrui.com



一简摸月

中国香港 沙田区



扫一扫上面的二维码图案，加我为朋友。



The Global Healthcare Data Science Community

Contact Channels

- 📞 UK +44 1843 609600
- ✉️ office@phuse.global
- 🌐 phuse.global

Social Media

- 🐦 [@phusetwitta](https://twitter.com/phusetwitta)
- 📘 [/phusebook](https://facebook.com/phusebook)
- 📺 [/phusetube](https://youtube.com/phusetube)
- 🌐 [/company/phuse](https://linkedin.com/company/phuse)