



Haystack Search Relevance Conference  
23 April 2024

# Generative AI Search and Summarization Testing using Human SME Techniques

Speaker: Doug Rosenoff

Ed Chatelain

Sreenath Kottapuzhackal

Ashwin Mohanraj

Jess Awasung

# Who is LexisNexis?

- Part of RELX Corporation (Elsevier, Reed Exhibitions)
- Large International Legal Publishing and Technology Company
  - Research, Productivity, Analytics, Management & Marketing Tools for Legal Professionals
- Mission: **Advance the rule of law around the world**
- Leader in online publishing since 1973
- Leader in adoption and use of ML and AI since 2019

# LexisNexis: Content on the Spindle

**307m****Dockets & documents**

LexisNexis content includes more than 307 million court dockets and documents, over 168 million patent documents, 4.75 million State Trial Orders, and 1.5 million jury verdict and settlement documents.

**3.8pb****Global content**

Our global content collection contains roughly 3.8 petabytes of data, which is 790x the size of Wikipedia.

**138b****Documents and records**

Our global legal and news database contains 138 billion documents and records with 2.2 million new legal documents added daily.

**39k****Premium sources**

Nexis news and business content includes over 39,000 premium sources in 50 languages, covering more than 180 countries. It has data including 540 million company profiles with a content archive that dates back 45 years.

**65k****News articles**

In 2022, Law360 produced over 65,000 news and analysis articles. Legal analytics tool Lex Machina has normalized over 127 million counsel mentions and over 134 million party mentions since 2016.

**152m****Patent documents**

PatentSight includes objective ratings of the innovative strength (Patent Asset Index) of more than 152 million patent documents from more than 100 countries.

# LN Gen AI Goal

Use Gen AI methods to provide customers with trustworthy tools that simplify their work processes and improve their efficiency

# Initial Gen AI Use Cases for Legal Practitioners

- Conversational Search
- Editorial Content Summarization
- Customer Document Upload and Summarization
- Document Drafting

# Definitions

- *SME* = subject matter expert
- *HRT* = human relevance testing
- *Ask* = Generative AI equivalent to Search
- *Response* = analogous to SERP
- *System Prompt* = crafted prompt that describes common features and expectations in output
- *User Prompt* = end-user provided prompt initiating AI activity

# HRT Testing Process

Admins & Customers



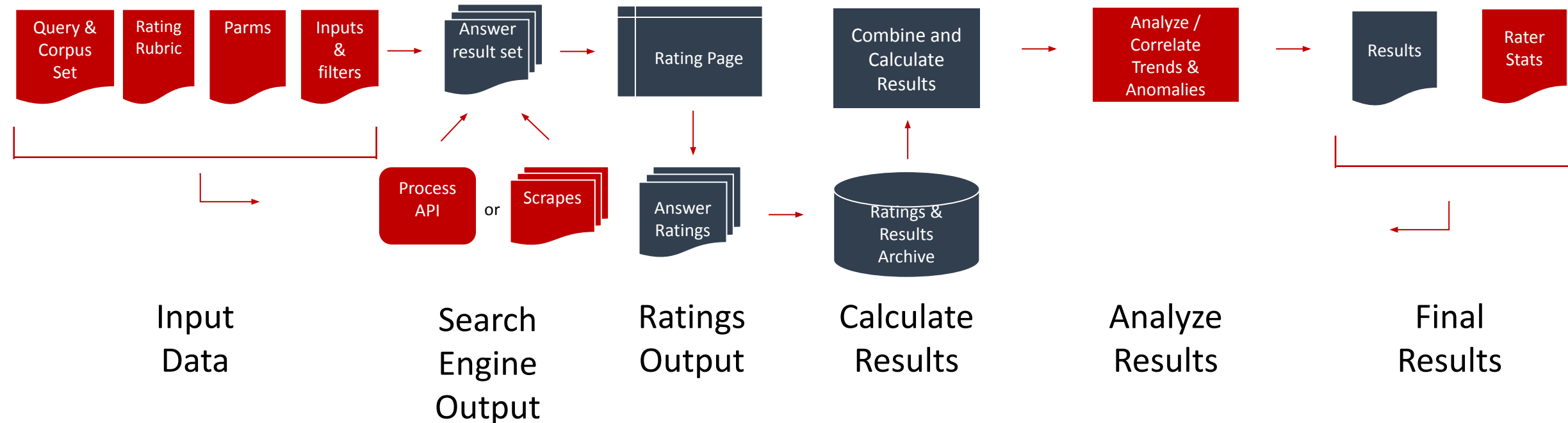
Raters



Analysts/Admins



## HRT Testing Framework



# Gen AI SME Testing Process

Admins & Customers



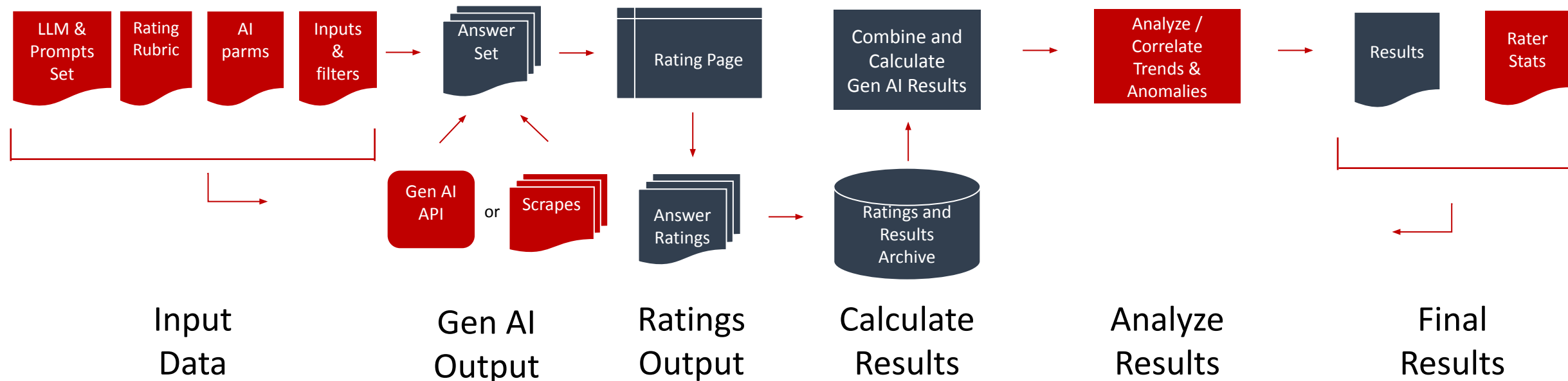
Raters



Analysts/Admins



## SME Testing Framework





Standard HRT  
& Gen AI SME  
evaluations use  
*identical* meta-  
processes

# Initial Gen AI SME Testing Goals

## Overall:

1. Evaluation of Gen AI responses against multiple metrics
2. Standardize usage of common Gen AI metrics across company
3. Comparison of Gen AI responses using different LLMs, system prompts, user prompts, and combinations
4. Drive macro improvement of Gen AI responses within corpora by content type & language

## Ask:

5. Data-driven decision-making using comparison & regression methods

## Summarization:

6. Comparison of Gen AI summaries against existing editorial summaries
7. Explore extension of Gen AI methods to editorial processing

# Basic AI Metric

**Simple Response Quality Metric (1-4)** Contributing factors *rated as a gestalt and relies on rater to integrate into a single value.*

- Useful
- No Hallucinations Response does not contain information that cannot be verified
- Relevant
- Accurate *and possibly other quality aspects*
- Authoritative
- Comprehensive
- Fluent

*1 = poor  
2 = fair  
3 = good  
4 = great*

- *Other ratings outputs: user comments & error notes, & citation ratings & comments*
- *Useful for reconnaissance, quick, & cheap evaluations*

# Advanced AI Metric Ratings

## Multiple Specific Metrics; one for each Rating attribute

- **Relevant (1-4)** Answers the user prompt
- **Authoritative (1-4)** Citations should be valid and support the legal statements in the response
- **Accurate (1-4)** Response contains information that is true and consonant with input
- **Comprehensive (1-4)** Response addresses every part of what is requested by user
- **Fluency (1-4)** Response is appropriate format, style, and tone
- **Useful (1-4)** Overall summary rating
- **Hallucinations (Y/N)** Response does not contain information that cannot be verified.

1 = poor  
2 = fair  
3 = good  
4 = great

- *Other ratings outputs: user comments & error notes, citation ratings & comments*
- *All metrics must be complete before rating is complete*



# Generative AI Ask Testing using SME Evaluation Methods

# SME Gen AI Ask Testing Use Cases

1. *Fixed LLM, Fixed Corpora, Fixed System Prompt, User Prompts -> Response* } MVP 1.0 Use Case
2. System prompt & LLM combinations, Fixed Corpora, User Prompts -> Response
3. Conversational User Prompts to Ask Result Ratings
4. Detection and Elimination of Hallucinations
5. Broad and Deep testing by Content Type, Document Type, Document Section
6. Boolean & NLP vs. Generative Ask Comparison Ratings\*
7. Boolean & NLP vs. Generative Citation List Comparison\*

# Generative AI SME Ask Testing Process

Admins & Customers



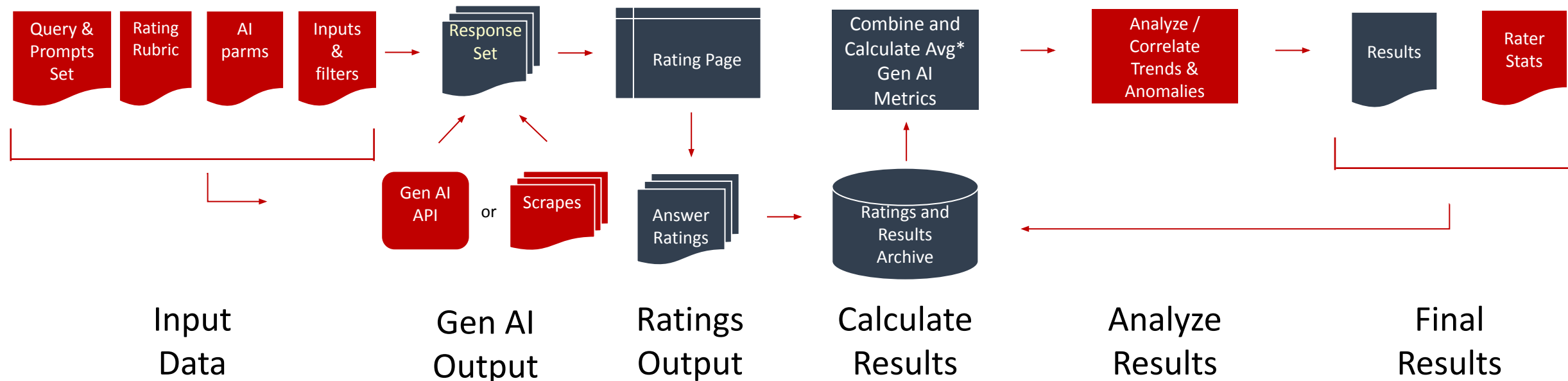
Raters



Analysts/Admins



## SME Testing Framework



# Gen AI SME Ask: Ratings Page

STF Home

LEGALKOTTAPS1 - Report a Problem

**Prompt:** what do i need to bring to an attorney for them to draft my will

Relevance:

Accuracy:

Fluency:

Usefulness:

Authority:

Complete:

1-4 Numeric Rating

Hallucinations:

Binary Rating

## AI Search Response

While the provided information does not specifically list what you need to bring to an attorney for drafting your will, it does emphasize the importance of promptly providing the attorney with necessary information after the initial client interview. The attorney will then establish a timetable for sending the initial draft, following up with the client, and signing the final estate plan documents [1]. To ensure a smooth process, it is advisable to gather and bring relevant personal and financial information, including details about your assets, liabilities, and beneficiaries, to the initial meeting with the attorney. This will help the attorney draft a will that accurately reflects your wishes and estate planning goals.

	Ratings	Citations	Comments
1.	<input type="text"/>	California Wills and Trusts	Comments (optional)
2.	<input type="text"/>	Wis. Stat. § 802.05	Comments (optional)

Comments (optional)

Submit and get next rating.

Submit and quit rating.

Cancel

1-4 Rating

# Gen AI SME Ask Output Metrics

- Results For each prompt / response pair (Individual Prompt Results)
  - Individual ratings for each metric\* from each rater
  - Representative (mode || average) ratings for each metric for prompt/response pair
- Results For the entire job (Cross-Dataset Summary Results)
  - Ratings Results (%'age of prompt/response ratings completed)
  - Average ratings for each metric (except hallucinations; given in job-wide %'age)
  - Major Reference Precision (agreement of top two ratings for each metric)
  - Minor Reference Precision (agreement of top three ratings for each metric)
- Rater Completion Metrics (Time to Rate, IRR, etc)

# Gen AI SME Ask: Dataset Results.a

SEARCH TEST FRAMEWORK

LEXISNEXIS

Return to Result List

Refresh

Download

Pause Rating

Email Raters

Ratings Complete

Promote to Named

Delete Results

Summary

Execution Results

Select All

Clear All

Default

Summary Averages

☒ Test Data Info

☒ HRT Test Data Info

☒ HRT AI Rating Scores

- ☒ Usefulness
- ☒ Relevance
- ☒ Authority
- ☒ Accuracy
- ☒ Comprehensiveness
- ☒ Fluency
- ☒ Hallucinations
- ☒ Major reference precision
- ☒ Minor reference precision

☒ HRT AI Rating Information

☒ Reviewer Information

☒ Error Information

☒ Job Info

☐ Job Timing Info

VIEW OPTIONS

Test Data Information

Test Name	Creator Name	Description	Test Type	Message for Rater	HRT Status	Scrape Name	Rater Pool	Product	Environment	HLCT	Hide Citation Ratings	Query Set Name	Persistence Type
-----------	--------------	-------------	-----------	-------------------	------------	-------------	------------	---------	-------------	------	-----------------------	----------------	------------------

HRT Test Data Information

Query count	Estimated total docs to rate	Ratings per document	Estimated total ratings needed	Max document review %	Instructions
70	70	1	70	40%	N/A

HRT AI Search Rating Information

	Usefulness	Relevance	Authority	Accuracy	Comprehensiveness	Fluency	Hallucinations Present
Avg. Score	2.42	2.81	2.6	2.98	2.79	3.04	3.85%
4 - Great (queries)	9 (17.31%)	13 (25%)	7 (13.46%)	13 (25%)	6 (11.54%)	12 (23.08%)	Pass: 50 (96.15%)
3 - Good (queries)	15 (28.85%)	25 (48.08%)	23 (44.23%)	29 (55.77%)	34 (65.39%)	33 (63.46%)	
2 - Fair (queries)	17 (32.69%)	5 (9.62%)	17 (32.69%)	6 (11.54%)	7 (13.46%)	4 (7.69%)	
1 - Poor (queries)	11 (21.15%)	9 (17.31%)	5 (9.62%)	4 (7.69%)	5 (9.62%)	3 (5.77%)	Fail: 2 (3.85%)
Major reference precision	24 (34.29%)	38 (54.29%)	30 (42.86%)	42 (60%)	40 (57.14%)	45 (64.29%)	
Minor reference precision	41 (58.57%)	43 (61.43%)	47 (67.14%)	48 (68.57%)	47 (67.14%)	49 (70%)	

HRT Rating Information

Query Counts	
Total New Queries to be Rated for this HRT	70
Queries with 1 Ratings Completed	54
Total Queries Completed	
Query Ratings	
Additional Ratings to be Completed	16
Total Ratings Completed in Job	54
Total Ratings Required in Job	70

# Gen AI SME Ask: Dataset Results.b

HRT AI Search Rating Information

	Usefulness	Relevance	Authority	Accuracy	Comprehensiveness	Fluency	Hallucinations Present
Avg. Score	4	3.75	3.6	3.88	3.88	3.75	0%
4 - Great (queries)	4 (100%)	3 (75%)	2 (50%)	3 (75%)	3 (75%)	3 (75%)	Pass: 4 (100%)
3 - Good (queries)	0	1 (25%)	2 (50%)	1 (25%)	1 (25%)	1 (25%)	
2 - Fair (queries)	0	0	0	0	0	0	
1 - Poor (queries)	0	0	0	0	0	0	Fail: 0
Major reference precision ⓘ	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)	
Minor reference precision ⓘ	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)	

HRT Rating Information

Query Counts		
	Total New Queries to be Rated for this HRT	4
	Queries with 1 Ratings Completed	0
	Queries with 2 Ratings Completed	0
	Queries with 3 Ratings Completed	4
	Total Queries Completed	
Query Ratings		
	Additional Ratings to be Completed	0
	Total Ratings Completed in Job	12
	Total Ratings Required in Job	12
Errors and Exclusions		
	Total Errors Reported	1 Errors
	Total Queries Excluded	0

Reviewer	Instructions	Number of Ratings	Number Ratings Excluded	Number of Errors	Average Time of Review	Average Reviewer Score	Two Points From Mode	Three Points From Mode
	N/A	0 (0%)	0	0		N/A	N/A	N/A
	N/A	0 (0%)	0	0		N/A	N/A	N/A
	N/A	4 (33%)	0	0	1 Minute 4 Seconds	N/A	N/A	N/A
	N/A	4 (33%)	0	1	23 Seconds	N/A	N/A	N/A
	N/A	4 (33%)	0	0	16 Seconds	N/A	N/A	N/A
	N/A	0 (0%)	0	0		N/A	N/A	N/A
	N/A	0 (0%)	0	0		N/A	N/A	N/A

Job Information

# Gen AI SME Ask: Individual Results

Return to Result List

Refresh

Edit Ratings

Edits Complete

Download

Promote to Named

Delete Results

Summary

Execution Results

Select All

Clear All

Default

Execution Scores

☒ AI Search Response
 ☒ HRT AI Rating Scores
 

☒ Usefulness
 ☒ Relevance
 ☒ Authority
 ☒ Accuracy
 ☒ Comprehensiveness
 ☒ Fluency
 ☒ Hallucinations

☒ Ratings
 

☒ Citations
 ☒ Citation Ratings

VIEW OPTIONS

HRT Execution Information

Query GUID	Query Term	Results Count	Errors Count
2023_AI_001	Can you explain a situation where a lack of "consideration" could make a contract unenforceable?	1	0

AI Search Response

A contract can be unenforceable due to a lack of consideration if one party does not provide anything of value in exchange for the other party's promise or performance. For example, if A promises to give B \$100 in exchange for nothing, there is no consideration from B, so the contract is not enforceable *Bradley v. Dhybrid Sys., LLC Bavelis v. Doukas (In re Bavelis) Alessi Equip., Inc. v. Am. Piledriving Equip., Inc.* . Or if A and B sign an employment agreement stating B will work for A, but the agreement does not require A to pay B, there is no consideration from A, making the contract unenforceable *Alessi Equip., Inc. v. Am. Piledriving Equip., Inc.* . The key is that both sides must provide consideration - something of legal value - for a contract to be enforceable. If one side provides no consideration at all, the contract fails for lack of consideration.

Usefulness	Relevance	Authority	Accuracy	Comprehensiveness	Fluency	Hallucinations Present
4.00	4.00	3.00	4.00	4.00	4.00	0%

Rater	Usefulness	Relevance	Authority	Accuracy	Comprehensiveness	Fluency	Hallucinations	Time	Reviewed	Reason	Comment
	4 - Great	3 - Good	3 - Good	4 - Great	4 - Great	4 - Great	Pass	13 Seconds	11/8/23, 12:39 PM		
	4 - Great	4 - Great	3 - Good	3 - Good	4 - Great	4 - Great	Pass	14 Seconds	11/8/23, 12:36 PM		
	4 - Great	4 - Great	4 - Great	4 - Great	4 - Great	4 - Great	Pass	14 Seconds	11/8/23, 12:35 PM		

Citation	Mode (or Avg. Score)	Title
1	4.00	Fain Family First Pōship v. EOG Res., Inc.

Rater	Citation Score
	4 - Great
	4 - Great
	4 - Great

Citation	Mode (or Avg. Score)	Title
2	4.00	Collings v. Sidhartan (In re KSRP, Ltd.)

Rater	Citation Score
	4 - Great
	4 - Great
	4 - Great

Citation	Mode (or Avg. Score)	Title
3	4.00	Hoffman v. Geico Ins. Co.

Rater	Citation Score
	4 - Great
	4 - Great
	4 - Great

Representative rating values



# Gen AI Summarization Testing using SME Evaluation Methods

# Gen AI SME Summarization Testing Use Cases

Rate and Analyze:

- 1: *Document to Summary Comparison (MVP 1.0)*
- 2: *Editorial vs Generative AI Summary Comparison*
- 3: *Summary Recommendation Content Ingest*
- 4: Editorial vs Gen AI Summary Comparison\*
- 5: Gen AI Summary Comparison\* (same user prompt, different LLM **and/or** system prompts)
- 6: Citation Comparison\* (same user prompt, different LLM/versions **and/or** system prompts)
- 7: Automated Regression Series Results?
- 8: Complex / Combinatorial Documents

**Key  
Use  
Cases**

# Basic Gen AI SME Summarization Testing Process

Admins & Customers



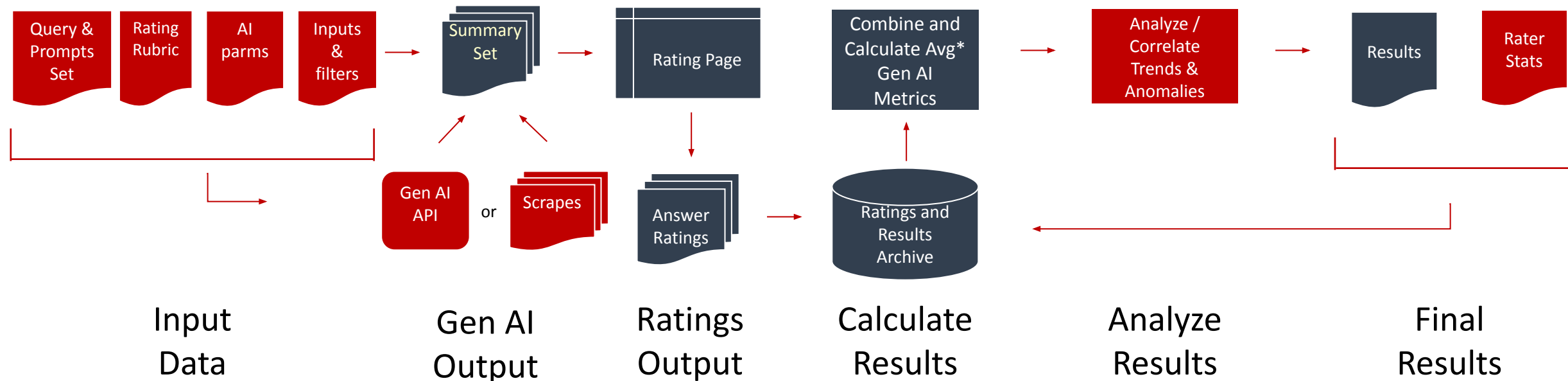
Raters



Analysts/Admins



## SME Testing Framework



GENAI TESTING @ LEXISNEXIS

## Summary Text

# Gen AI SME Summarization: Rating.b

## Document Text (cont'd)

17 (1st Dept 2019.)

Various discovery disputes have arisen between the parties in litigating plaintiff's remaining claims, which have led in turn to the four motions at issue in this decision.

Motion sequence 005 is a motion by defendants to quash a subpoena served by plaintiff on nonparty Capital One, N.A., seeking bank records for defendants' bank accounts with Capital One; and seeking to suppress any information already obtained from Capital One pursuant to the subpoena. Motion Sequence 006 is a motion by plaintiff to quash documentary and testimonial subpoenas served by defendants on plaintiff's personal lawyer, [Eileen Caulfield Schwab](#) ▼, Esq., and her law office. On motion sequence 007, defendants seek to quash further bank-records subpoenas served by plaintiffs on Capital One and nonparty Bank of America, NA, and to suppress any information produced to plaintiffs by the banks. Motion sequence 008 is defendants' motion to quash a nonparty subpoena served by plaintiff on [Parker Pohl LLP](#) ▼, which served as counsel in a related action.

Additionally, in connection with [\*3] the subpoenas that are the subject of motion sequence 007, defendants pointed out that those subpoenas listed Ms. Sabharwal's unredacted Social Security number

## Summary Text (cont'd)

- Bank records would not shed light on jewelry value, and payment amounts are not disputed.
- Attorney-client privilege shields some documents sought from plaintiff's attorney.
- Plaintiff's attorney is a fact witness regarding jewelry appraisal process.
- Information sought from defendants' prior counsel unlikely to lead to admissible evidence.
- Certain deposition questions improperly objected to by defendants' counsel.

### Procedural Outcome

Plaintiff's and defendants' motions to quash various subpoenas granted in part and denied in part. Plaintiff's request to compel deposition answers granted in part.

### Parties and How the Case Affects Them

Plaintiff Shelley Rubin - Allowed limited additional discovery from defendants and plaintiff's attorney.

Defendants Nisha Sabharwal and Vastra Inc. - Prevented from obtaining some privileged information but allowed other discovery. Must provide additional deposition testimony.



## Citations in Summary

	Citations	Comments
1.	<a href="#">Rubin v. Sabharwal</a>	1.

Comments (optional)

## Evaluation Comments

Submit and get next rating.

Submit and quit rating.

Cancel

# Gen AI SME Summarization Results

- Results For each prompt / summary pair ([Individual Prompt Results](#)):
  - Individual ratings for each metric\* from each rater (including citations)
  - Representative (mode,average) ratings for each metric for prompt/summary pair
- Results For the entire job ([Dataset Results](#)):
  - Summary of Ratings done (%'age of prompt/summary ratings completed)
  - Representative Ratings for each metric (except hallucinations; given in job-wide %'age)
  - Representative Citation Ratings
  - Major Reference Precision (agreement of top two ratings for each metric)
  - Minor Reference Precision (agreement of top three ratings for each metric)
  - Underperforming Precision (queries with delta of >2 rating points for each metric)
- Rater Completion Metrics ([Ave Rating Time, IRR, etc](#))

# Gen AI SME Summarization: Dataset Results.a

Summary

Execution Results

Select All

Clear All

Default

Summary Averages

☒ Test Data Info

☒ HRT Test Data Info

☒ HRT AI Rating Scores

☒ Usefulness

☒ Relevance

☒ Authority

☒ Accuracy

☒ Complete

☒ Fluency

☒ Hallucinations

☒ Major reference precision

☒ Minor reference precision

☒ HRT AI Rating Information

☒ Reviewer Information

☒ Error Information

☒ Job Info

☐ Job Timing Info

VIEW OPTIONS

### Test Data Information

Test Name	Creator Name	Description	Test Type	Message for Rater	HRT Status	Scrape Name	Rater Pool	Product	Environment	HLCT	Hide Citation Ratings	Query Set Name	Persistence Type

### HRT Test Data Information

Query count	Estimated total docs to rate	Ratings per document	Estimated total ratings needed	Max document review %	Instructions
1	1	1	1	100%	N/A

### HRT AI Summary Rating Scores [HRT AI Metrics Interpretation](#)

	Usefulness	Relevance	Authority	Accuracy	Complete	Fluency	Hallucinations Present
Avg. Score	3	4	2	2	4	3	0%
4 - Great (queries)	0	1 (100%)	0	0	1 (100%)	0	Pass: 1 (100%)
3 - Good (queries)	1 (100%)	0	0	0	0	1 (100%)	
2 - Fair (queries)	0	0	1 (100%)	1 (100%)	0	0	
1 - Poor (queries)	0	0	0	0	0	0	Fail: 0
Major reference precision	1 (100%)	1 (100%)	0	0	1 (100%)	1 (100%)	
Minor reference precision	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	

### HRT Rating Information

Query Counts		
	Total New Queries to be Rated for this HRT	1
	Queries with 1 Ratings Completed	1
	Total Queries Completed	
Query Ratings		
	Additional Ratings to be Completed	0
	Total Ratings Completed in Job	1
	Total Ratings Required in Job	1

# Gen AI SME Summarization: Dataset Results.b

Reviewer	Instructions	Number of Ratings	Number Ratings Excluded	Number of Errors	Average Time of Review	Average Reviewer Score	Two Points From Mode	Three Points From Mode
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0
					10:30:00	5.0	5.0	5.0

## HRT Error Summary

Problem Category	Count
Other (please specify)	0
Cannot access the document (Restricted or Out of Plan message).	0
The instructions are not clear.	1
There is a technical problem with this task (please describe).	0
I lack the expertise for this task.	0

## Annotations

### Observations

LEGAL:MOHANRAX 10/24/23, 9:27 AM
test
LEGAL:MOHANRAX 10/24/23, 9:27 AM
test
LEGAL:MOHANRAX 10/24/23, 9:27 AM
test observation

### Conclusions

LEGAL:MOHANRAX 10/24/23, 9:27 AM
test
LEGAL:MOHANRAX 10/24/23, 9:27 AM
test
LEGAL:MOHANRAX 10/24/23, 9:27 AM
test Conclusions

### Entry Logs

Changed By	Changed On	Comments

## Job Information

Job Status	Job Priority	Job Created	Job Submit	Job Start	Job End	Search Engine Version	Search Engine Build	Tags
Completed	Normal	10/23/23, 6:01 AM	10/23/23, 6:01 AM	10/23/23, 6:02 AM	10/23/23, 6:04 AM	N/A	N/A	Test 3 rating Sum STF annoauthor:MOHANRAX

# Gen AI SME Summarization: Individual Result

SummaryExecution Results

Select AllClear AllDefault

Execution Scores

- AI Summary Response
- HRT AI Rating Scores
  - Usefulness
  - Relevance
  - Authority
  - Accuracy
  - Complete
  - Fluency
  - Hallucinations
  - Ratings
  - Citations

VIEW OPTIONS

HRT Execution InformationHRT AI Metrics Interpretation

Query GUID	Query Term	Results Count	Errors Count
2023_AI_QG_001	3S4X-CYB0-003B-S4J4-00000-00	1	0

AI Summary Response

Here is a summary of Roe v. Wade, 410 U.S. 113

**Procedural Posture**

Personal liberty under the Due Process Clause extended to woman's abortion decision, but was subject to regulation aimed at compelling state interests of maternal health and human life potentiality; abortion laws that did not distinguish between early and later abortions and limited legal justification to saving mother's life were unconstitutional.

**Overview**

Plaintiffs and intervenor appealed directly to the instant Court on the injunctive rulings. The State cross-appealed from the declaratory judgment. The Court affirmed the judgment, holding that abortion was within the scope of the personal liberty guaranteed by the Due Process Clause. This right was not absolute, but could be regulated by narrowly drawn legislation aimed at vindicating legitimate, compelling state interests in the mother's health and safety and the potentiality of human life. The former became compelling, and was thus grounds for regulation after the first trimester of pregnancy, beyond which the state could regulate abortion to preserve and protect maternal health. The latter became compelling at viability, upon which a state could proscribe abortion except to preserve the mother's life or health. The Texas statutes made no distinction between abortions performed early in pregnancy and those performed later, and it limited the legal justification for the procedure to a single reason --saving the mother's life -- so it could not survive the constitutional attack. This conclusion made it unnecessary for the Court to consider the doctor's vagueness challenge.

**Outcome**

The judgment of the district court as to the doctor's intervention was reversed, and the doctor's complaint in intervention was dismissed. In all other respects, the judgment of the district court was

Usefulness	Relevance	Authority	Accuracy	Complete	Fluency	Hallucinations Present
2.50	2.00	2.00	4.00	1.50	3.00	50%

	Rater	Usefulness	Relevance	Authority	Accuracy	Complete	Fluency	Hallucinations	Time	Reviewed	Reason	Comment
	stfautomation1@lexisnexis.com	2 - Fair	1 - Poor	2 - Fair	4 - Great	1 - Poor	3 - Good	Fail	12 Seconds	10/24/23, 9:17 AM		
Excluded	stfautomation2@lexisnexis.com	4 - Great	2 - Fair	1 - Poor	3 - Good	3 - Good	4 - Great	Fail	12 Seconds	10/24/23, 9:17 AM		
	ashwinkumar.mohanraj@lexisnexis.com	3 - Good	3 - Good	2 - Fair	4 - Great	2 - Fair	3 - Good	Pass	15 Seconds	10/23/23, 8:19 AM		test comments

Citation	Title
1	Roe v. Wade

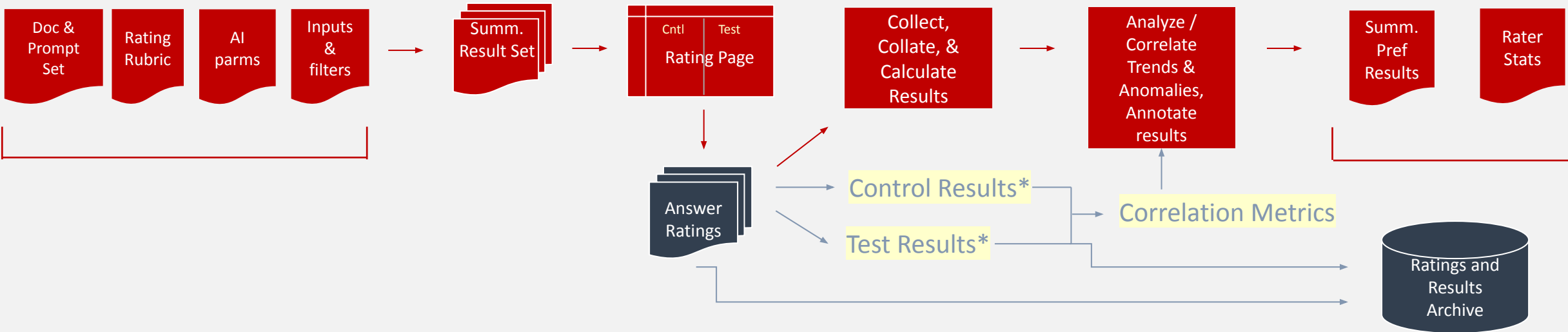
Query GUID	Query Term	Results Count	Errors Count
2023_AI_QG_002	55XW-HX71-F04G-G03T-00000-00	1	0

# Gen AI SME Processes & By-Processes

# Gen AI SME Comparison Process



## Gen AI Test Framework



Input  
Data

Retrieve  
Data

Collect  
Ratings

Gather  
Results

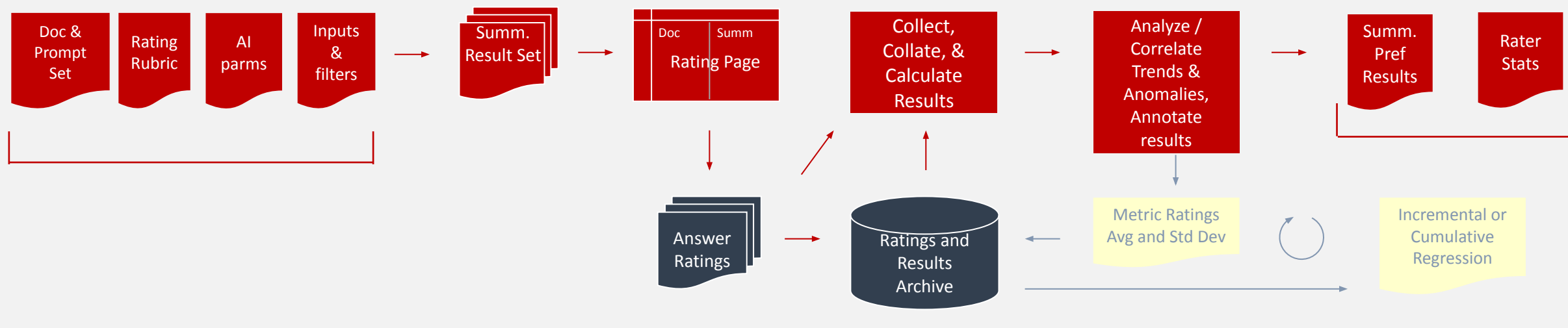
Analyze  
Results

Post  
Results

# Gen AI SME Rating Regression



## Gen AI Test Framework



Input  
Data

Retrieve  
Data

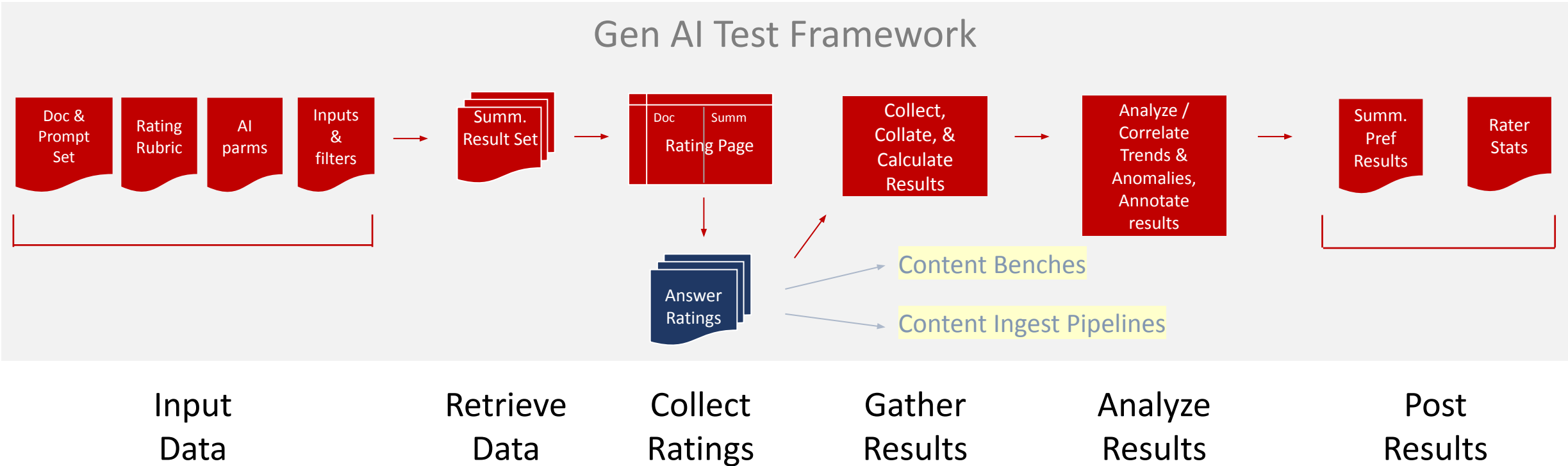
Collect  
Ratings

Gather  
Results

Analyze  
Results

Post  
Results

# Gen AI SME Content Ingest Process



# Gen AI SME Gold Data Processes

Admins



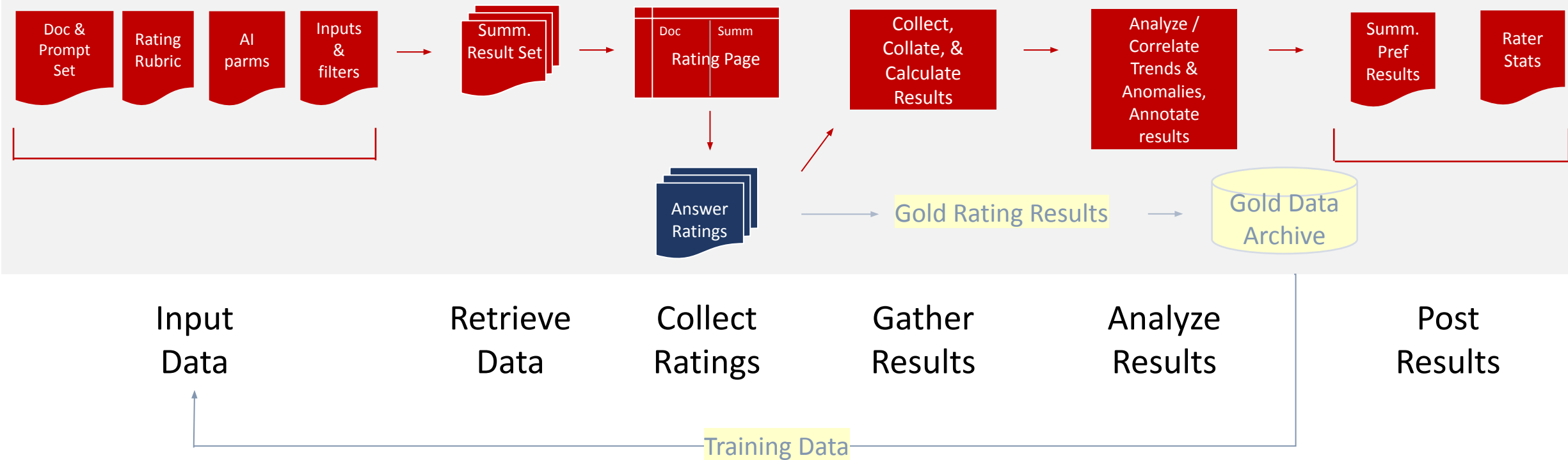
Raters/Editors



Analysts



## Gen AI Test Framework



# Lessons Learnt To Date\*

- SME Evaluation is a highly accurate tool for Generative AI evaluation but requires a scaleable tool to provide broad and deep quantitative testing.
- SME Evaluation, *as a process*, transfers easily and successfully from conventional to Generative AI evaluation.
- Generative AI ratings require much more SME expertise and take considerably longer.
- Generative AI as a testing problem is complex with both quantitative and qualitative aspects. Generic Gen AI metric definitions remain a difficult problem.
- Generative AI SME methods spin off useful (free) by-products.

# Outro: AI testing in RL is different\*

1. Conventional Search vs AI Ask responses
2. Editorial Summaries vs AI Summary results
3. Editorial vs AI generated results for content
4. Editorial vs AI generated results for customer operations
5. Combinatorial Testing (LLM, System Prompt, User Prompt)
6. Testing Multiple & Complex Document Inputs (e.g. Summarization)
7. Testing Multi-turn Conversational Inputs (e.g. Ask)
8. Metric vs Rating Comparison & Regression
9. Metricizing & Testing Generative AI is also a qualitative exercise:
  - *Intent Fulfillment*: is the result what you expected?
  - *Trust Maintenance*: hallucinations are a product of AI's nature
  - *Conversation Satisfaction*: conversations are not metrics
  - *Context Fit*: looser or tighter = more useful?

# Next Steps

- Additional use cases for Ask and Summarization testing
- Additional content types
- Additional products
- Gen AI SME based by-processes
- End Customer Feedback Processes

Also,

- Additional AI-specific tests and metrics
- Multi-document summarization
- Multi-Turn Conversational Ask
- Other Stuff ;^)

# Thanks

## **Co-authors / Jupiter Team**

Ed Chatelain

Sreenath Kottapuzhackal

Ashwin Mohanraj

Jess Awasung

## **Test Customers**

Tara Diedrichsen, Charm Bolarda, Eloise Luo, & Laura Baker



# Questions?

[douglas.rosenoff@lexisnexis.com](mailto:douglas.rosenoff@lexisnexis.com)

425.328.0085



# LexisNexis®

Interested in working with Gen AI at scale  
on applications with global scope and social impact?

**We are hiring.**

<https://www.lexisnexis.com/en-us/about-us/careers.page>

<https://www.elsevier.com/about/careers>

# Abstract

An internally developed Human Relevance Testing Framework was successfully modified to support Subject Matter Expert testing of generative AI Search results and summaries. These extensions to an existing testing framework have resulted in fast and frequent evaluation and testing of large, diverse corpora for gen AI search and summarization functions. This paper will discuss the modification of the framework from traditional Human Relevance Testing methods to generative AI SME methods, the testing processes, the metrics created in support of generative AI, and the outputs of the methodology. In addition, various other potential testing use cases based on this methodology will be covered, including comparison and regression methods and alternate product extensions, issues, and metrics.