Haystack 2024 • Synthetic Raters in SME Testing

Enhancing Generative AI Evaluation with Synthetic Raters

D. Rosenoff, E. Chatelain, V. Ganguru, S. Kottapuzhackal, E. Miracle, & A. Mohanraj

LexisNexis

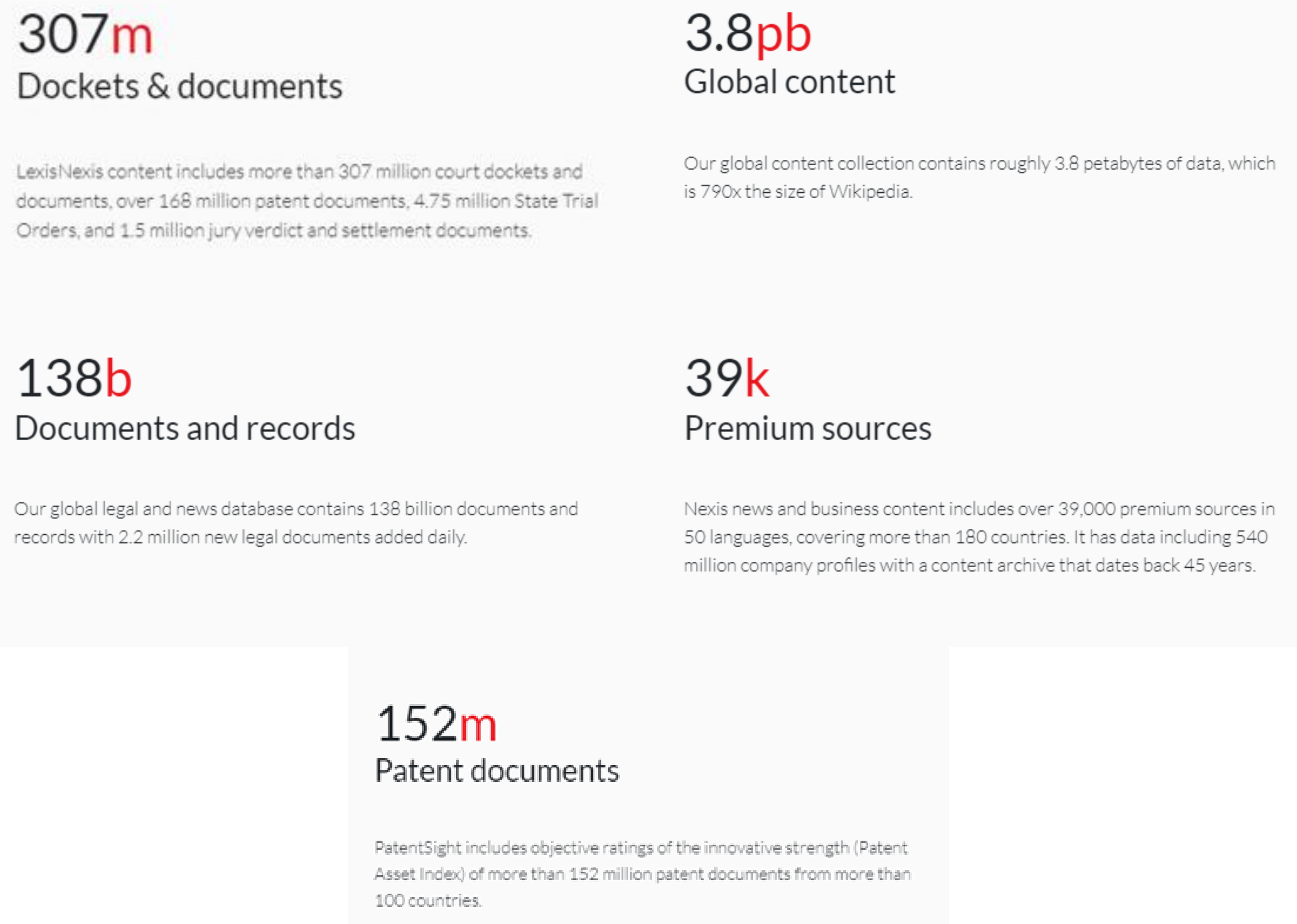
Haystack Search & AI Conference

Charlottesville, VA • 24 April 2025

Who is LexisNexis?

- Part of RELX Corporation (LN, Elsevier, Reed Exhibitions, Risk)
- 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.

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.

Problem Introduction

Definitions

SME = subject matter expert

HRT = human relevance testing

Ask = Generative AI equivalent to Search

Response = analogous to determinative SERP

System Prompt = crafted prompt that describes evaluation criteria and expectations in Gen AI output

Metric Prompt = crafted prompt rubric that describes how evaluate a fragment of text for a particular metric within a Gen AI output

User Prompt = end-user provided prompt initiating a Gen AI service request

Gen AI Use Cases for Legal Practitioners

- Intelligent Search
- Editorial Content Summarization
- Document Drafting
- Customer Document Summarization
- Taxonomy Classification and Evaluation

STF 8 (defined)

STF is an internally developed framework for testing both determinative and generative AI search* testing. STF provides fast and frequent tools for evaluation and testing of large, diverse corpora against multiple metrics, specific to the operation type and results to be generated.

STF is based on using multiple human SME raters evaluating results against a specific rubric. Ratings are then used to compute appropriate statistical metrics for relevance, precision, recall, and more specific properties of a given type of input operation.

A key foundation of STF is that the HRT/SME testing meta is the same, regardless of determinative or generative AI evaluation.

This observation led to some interesting work in 2024 and 2025...

Generative AI moved the Cheese

- After Years of Search tuning (Boolean, NL, Reranking, and ML) of Search against industry standard metrics like hDCG, RBO, p(k), etc., Gen AI moved the cheese.

- The determinative relevance evaluation metrics were no longer adaptable to the prompt response model.

- While the HRT/SME model meta remained viable, new Metrics were needed to measure relevance.

At Haystack 2024: 1st Generation AI Metrics

In 2025, we added a few more and divided them into primary/direct metrics and secondary/derived metrics.

Metric	Computation	Description
relevance	Primary	Prompt concentrates on the relevance of the AI response
accuracy	Primary	Prompt concentrates on the accuracy of the AI response
completeness	Primary	Prompt concentrates on the completeness of the AI response
fluency	Primary	Prompt concentrates on the fluency of the AI response
unsupportedstatements	Primary	Prompt concentrates on the unsupported legal statements in the AI response
document citation	Primary	For each citation, evaluates the relevancy of the Full document against the user prompt
authority	Secondary	Generated based on the citation ratings. Minimum(all citation ratings) for a prompt
misattributions	Secondary	If the authority rating is 1 (poor), then misattribution is true, else false
hallucinations	Secondary	If the accuracy rating is 1 (poor), then hallucinations is true, else false
usefulness	Secondary	Mean(Relevance, Authority, Accuracy, Completeness, Fluency)
localterminology	New*	Newly added in STF, yet to support in synthetic rating*

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

Other Ratings Outputs: User Comments & Error Notes
All metrics must be completed to submit a rating

Gen AI SME relevance ratings require alot

- More metric criteria / rating (1 to 11 metrics)
- Longer results to rate (not just cites to other documents)
- More complex results to rate (primary and secondary results)
- More depth of evaluation per rating (more personal judgement)
- Need for ratings increasing exponentially (new paradigm = new level of trust requirement → lots of testing to confirm)

Gen AI ASK time/rating and rating costs have risen substantially

Last year at Haystack…

David Fisher, Scott Stults, Jeremy Hudson (Moodys), and Rene Kriegler talked about the notion of using LLMs instead of *humans to help measure the quality of text generated by LLM or RAG systems. Hmmmm… That sounds vaguely familiar to STF* ⚒ 🪨

- What are the goals, benefits, and use cases of using LLM-derived raters?
- What does an LLM-derived rater look like, architecture- and code-wise?
- What does a Gen AI rater need to do, aside from ratings?
- How do we train a Gen AI rater?
- How accurate is a Gen AI rater compared to a human?
- How would we frame Gen AI raters for use at industrial scale and speed?
- What new things can Gen AI raters do?
- What do we name a Gen AI rater?

Most Importantly, When can we get started?

Synthetic Raters

A programmatic Gen AI construct that mimics SME rating behavior

STF Synthetic Rater Goals & Benefits

Goals

- Significantly reduce time and effort required for HRT studies
- Provide Neutral, Pluggable, Scaleable SR tooling for multiple Gen AI and non-AI services
- Like for like test tooling for SR-to-SR, SR-human & human-human comparison results
- Leverage as much as possible of STF Code base for provisioning SR Framework
- Provide all generic and custom metrics for Gen-AI and non-Gen AI testing
- Leverage extensive previous HRT studies to provide SR gold data training and tuning data
- Provide production and data science level SR testing tool with flexible inputs and outputs

Desired Benefits

- Tunable and Repeatable quantitative testing, comparison, and regression tooling using standardized metrics
- Low-cost extension of existing well understood and widely used internal test framework
- Lower bar for widespread testing with synthetic raters across multiple services and content types
- Leverage experience of large business units to assist smaller ones
- Provision automated, sophisticated out-of-the-box analysis tools at Enterprise scale
- High ROI

Potential SR Use Cases for Product Rating Evaluations

- Run Synthetic Rater Evaluation Job ***Before*** Human Rating
  - Prioritize overall ratings strategy / Quick review of large prompt sets to evaluate for anomalous rating results
  - Depending on results:
    - Spend less human review time / skip human ratings for Great Results (especially for small budget studies / LBUs)
    - Spend less human review time / skip human ratings for Poor results (especially for small budget studies / LBUs)
    - Spend less human review time / skip human ratings for very accurate metrics
  - Generate LLM Synthetic Rater comments and explanations
    - Expose on demand to Admins and / or Raters
    - Always expose to Admins and / or Raters
    - Expose Synthetic Rater Rating on Demand (with or w/o comments) via conversational interface

How to keep SR ratings from becoming a crutch for honest human ratings?

- Run Synthetic Rater Evaluation Job ***After*** Human Rating
  - Evaluate Individual Rating Results
  - Compare Human Rater Results to Synthetic
- Run Comparison and/or Regression tests across raters, ratings, job metrics
  - Human to Human (Data Quality, Ratings Spread/Std Deviation)
  - Human to Synthetic (Correlation Coefficient, Ratings Spread/Std Deviation, IRR)
  - Synthetic to Synthetic (Correlation Coefficient Agreement across Synthetic Raters, Ratings Spread/Std Deviation, IRR)
- Use Synthetic Rater Evaluation Results for Rater Training
  - Rating Quality Evaluation
  - Data Quality Evaluation

Potential SR Use Cases for Data Science

- Test and tune LLMs, System & Metric Prompts – easy setup, execute, and compare with prior results
- Arena style SR competitive evaluation* taking advantage of strengths of different LLMs to generate statistical averages
- Extensible to handle LLM settings (e.g. temperature, creativity, triggers, stored procedures etc.) in Synthetic Rater Policies Structure
- Versioned Synthetic Raters so specific SRs can be tracked, reused, & extended
- Synthetic Rater API is technology neutral API, so they can share STF can use LLMs, system prompts, metric prompts, and policies developed by other groups, internal and external organizations
- Full Comparison and Regression testing tools for evaluation and analysis at individual rating, rater pool, metric, and job levels
- Synthetic Raters extensible to Gen-AI services (Summarization, Taxonomy, Drafting) and non-Gen AI services (Search, Editorial Evaluation etc.)

Synthetic Raters Requirements Wishlist

Minimal process disruption

Simple, clean, & homogenous fitment into existing SME testing framework

Synthetic rater architectural requirements:

- A (blind) fixed name (e.g. Alice3 or Bob42)
- an assigned RAG/LLM
- an assigned system prompt
- Policies structure
- Versioning information

INSIGHT

Synthetic Raters are neurodivergent Human Rater Peers

Multiple SRs assignable to a (SR or HR+SR) rater pool (multiple raters per ratings job)

Tunable at system and metric levels

Generate the same ratings types as a human rater

Generate rater stats like a human rater

Support Multiple Rating Types and Services (Gen AI, Determinative, Search, Taxonomy, Summary, etc)

Non-Functional:

- Scaleable across multiple services, content types, products, and environments
- De/Re-hydrate methods
- Neutral architecture and APIs
- Cloud native
- Scaleable / very high performance
- Downloadable results files

Synthetic Rater Architecture

- **LLM or related structure**
- **Rater System Prompt**
- **Policies**
  - Rater Metric Prompts
  - Settings
  - Parameters & Extensions
- **Metadata**

  LLM Information (+ Version)

  Synthetic Rater Prompt (+ Version)

  Policy Information (+ Version)

  Synthetic Rater Information (+ Version)

  Name: Bob

  Target Service: ASK

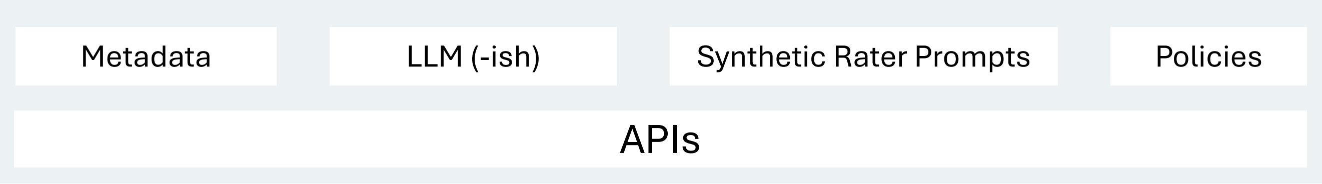
  Target Content Type: US Statutes

  Birthday: 20250115

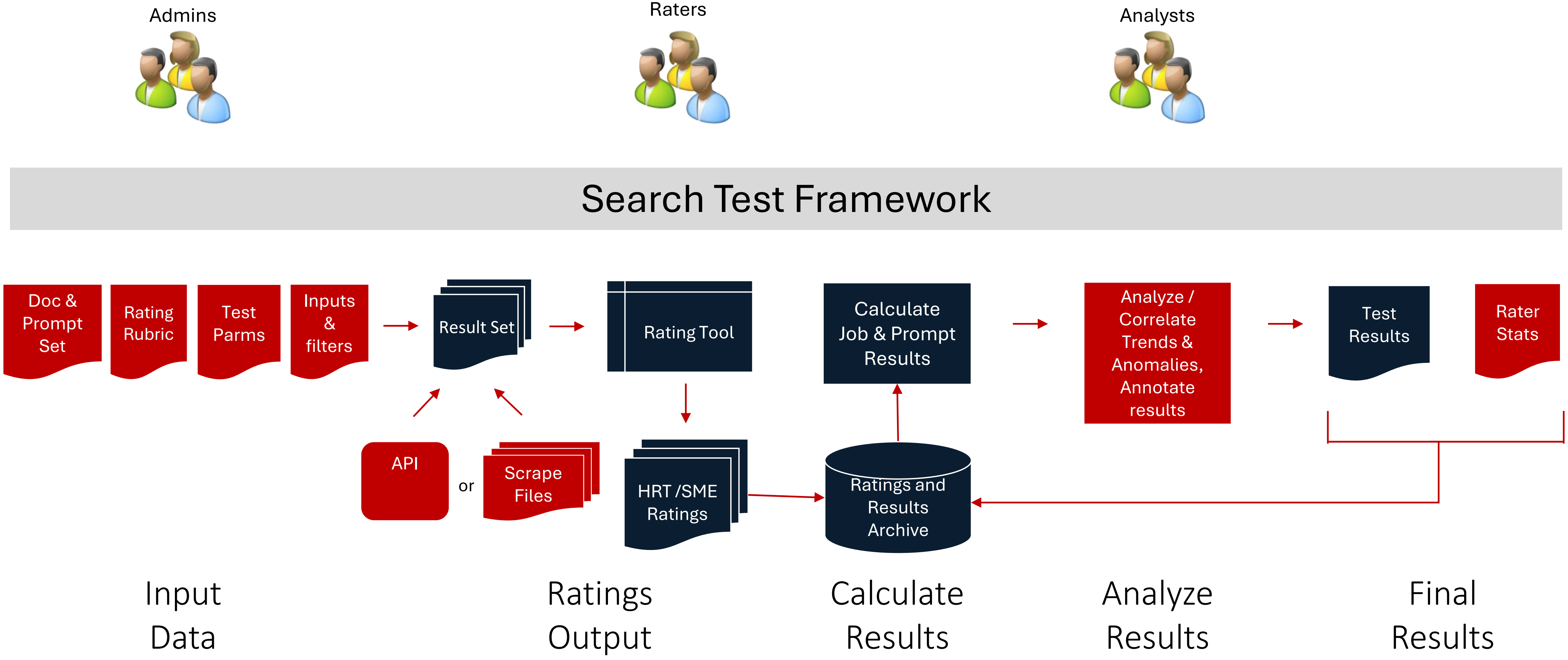
  Last Used Date: 20250331/13:08Z
- **APIs**

  Access and Sharing (within STF)

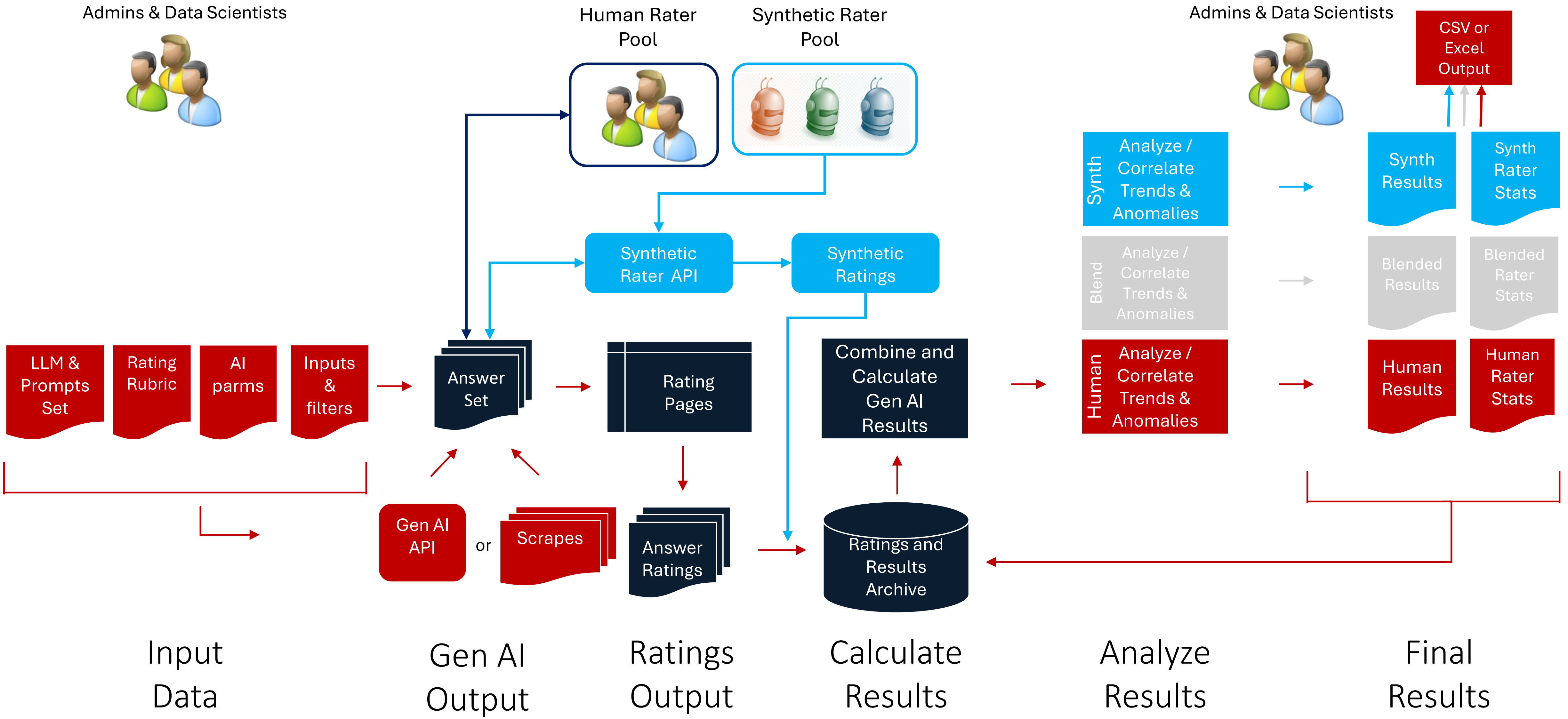
  Read/Write/Edit/Delete



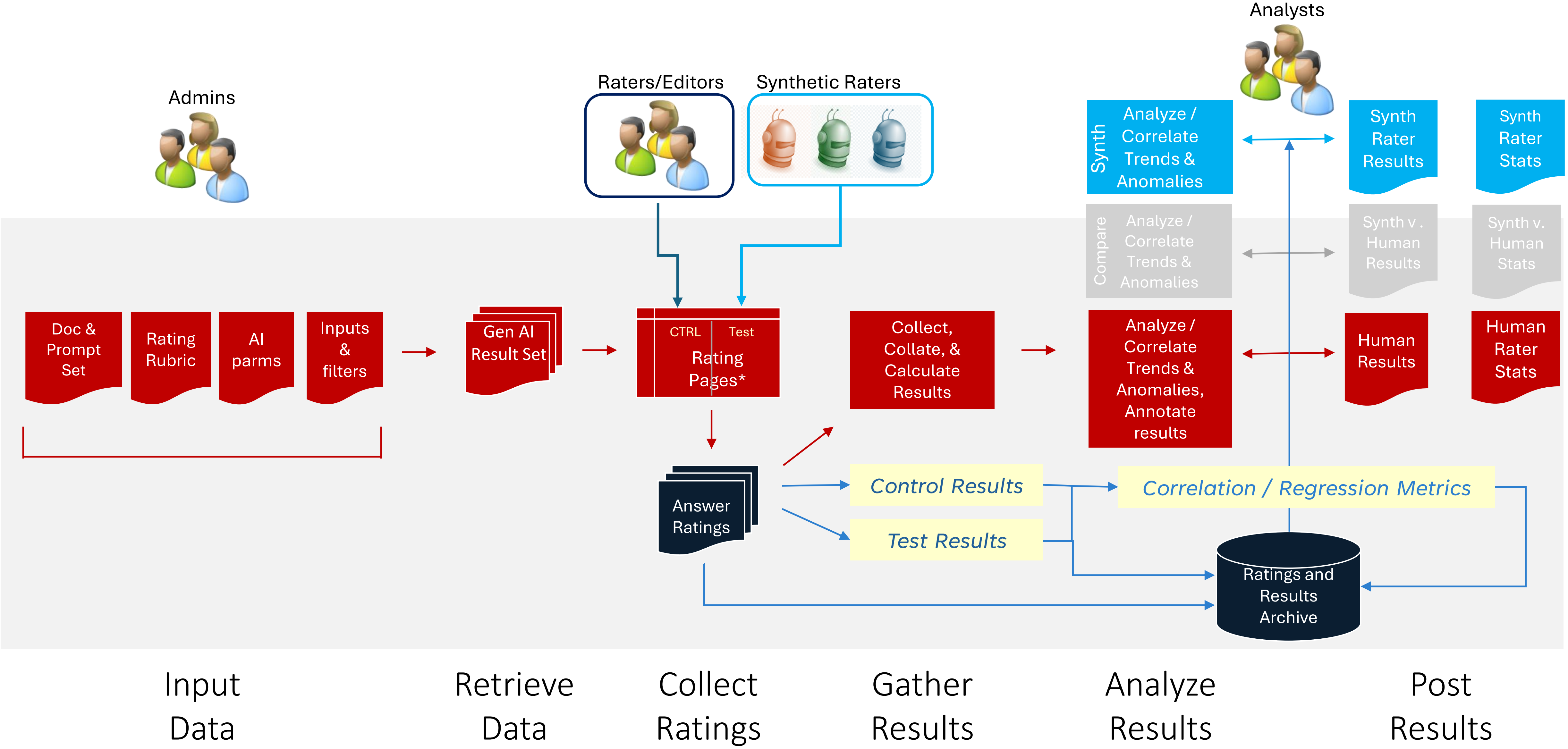
Generic HRT / SME Testing Process



Gen AI Testing Process with Synthetic Raters



Synthetic Rater Comparison / Regression Process



** May Involve Multiple Ratings Jobs*

Meaningful Synthetic Rater Comparisons

1. Metrics

   Correlation Coefficient

   Major and Minor Precision

   Standard Deviation Overlap

   Inter-Rater Reliability (κ)*

2. Synthetic Rater to Synthetic Rater (Same Config)

   understand statistical result variability

3. Synthetic Rater to Synthetic Rater (Different Config)

   understand behavior and optimization

4. Above Comparisons 1-3 for Human Rater v. Synthetic Rater

   Human (herds) are always correct, right?

Correlation Coefficient
$-1.0 \leq CC \leq 1.0$
1.0 = Perfect Correlation
0.0 = Random (no) Correlation
-1.0 = Opposite Correlation

Major Precision
Results agree: Great or Food

Minor Precision
Results Agree: Great or Good or Fair

** Randolph's Free-marginal multi-rater kappa (2005)*

Haystack 2024 • Synthetic Raters in SME Testing

Results

Results (-ish) and Learnings

Preliminary & Early Days

Changing Daily

(Highly) dependent on Data Quality (data set size, rater experience)

Highly dependent on type and amount of training

Primary metrics only

Some metrics more problematic than others

Still evaluating Synthetic Rater metrics

Still evaluating Human to Synthetic Rater metrics

LLM Testing

Synthetic Rater Name	LLM Used
Bob	mistral_7b_instruct_v0_1
Milo	llama_v3_3_70b_instruct
Claire	meta_llama_v3_1_405b_instruct_fp8

(Preliminary: 31 March 2025)

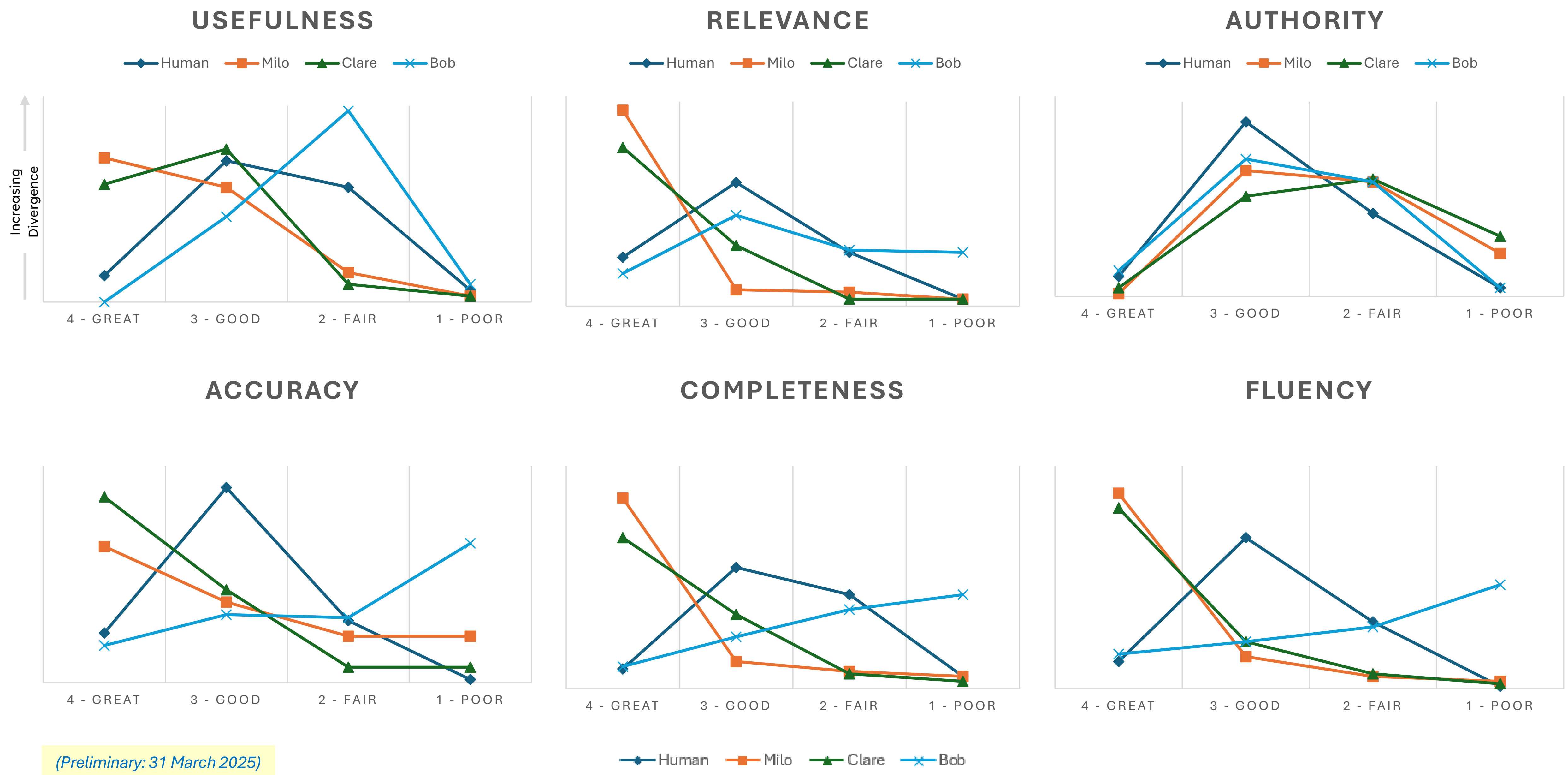
Performance: Fast

Efficiency: *Total job rating time reduced by > 99%*

Type	Number of Ratings	Total Review Time**	Average Time of Review/doc
Human	100 (100%)	c. 175 Hours	c. 1 hour 45 minutes
SR (Milo)	100 (100%)	c. 40 minutes	<30 Seconds

(Preliminary: 31 March 2025)

Metrics: Synthetic Rater LLM Rating Comparison



USEFULNESS

RELEVANCE

AUTHORITY

ACCURACY

COMPLETENESS

FLUENCY

(Preliminary: 31 March 2025)

Some Interesting Results for Synthetic Raters

Amazing performance baselines ... but will inevitably degrade

- More metrics
- More complex metrics
- Secondary / Reasoning metrics
- Augmented Information Requests and Analysis
- Dynamic Behavior and Parameter changes

Necessity of *well-curated, high quality* training data

- Exposes / Emphasizes that SRs are not quite to human skill levels (yet)
- Synthetic Raters will not directly replace humans as SMEs anytime soon
- No perfect IRRs, No perfect Comparisons
- Watch rating spreads / rating Standard Deviations as well as (average) values
- Understand your training data at a deep level : Garbage in / Garbage out (still)
- What are you training for? Relevance or Ratings Matching?

Necessity of budgeting in a time of explosive innovation

- Accuracy & Performance improve on daily to weekly timeframes
- Computing Cost Reductions do not keep pace with improvement rates
- Significant work remains to explore and exploit possibilities for use

Challenges, Next Steps & Conclusions

Synthetic Rater Challenges

- Selection of LLM(s)
- Optimizing Synthetic Rater System Prompt
- Optimizing Synthetic Rater Metric Prompt
- Synthetic Rater Smart Extension / API surfaces
- UI / UX Complexity: more metrics, deeper results, data cues
- Sufficient Curated Tuning Quality Data
- Multi-Level and Type Tuning: Single Shot, Two Shot, Multi-Shot
- Meaningful Metrics for measuring Ratings Quality
- Human Rating Quality preferred overall and for each metric (?)
- How close do Human Ratings & Synthetic Ratings need to be?

More Synthetic Rater Challenges 2

How much and when do you show Synthetic Rater results to Human Raters?

How to play to a Synthetic Rater’s AI strengths?

How do you compute a realistic ROI for Synthetic Raters?

Thought experiment: What about blending Human and Synthetic Rater Results?

- Mixed averages,
  - across all metrics or
  - by metric
- Is it a fair thing to do?

Immediate Next Steps:

- Continue Tuning ASK
  - Other LLMs
  - Refine Prompts
  - Explore content type dependency
- Extend to other Gen AI-based Evaluations (e.g. Taxonomy, Drafting, Summarization, etc.)
- Extend to Determinative Evaluations (HRT for Documents, HRT for SERPs)
- Explore, develop, code, and test
  - Contextual chunking to improve accuracy
  - Multi-Turn Gen AI service techniques
  - What's Missing / Comprehensiveness techniques
  - Parameter change and stored procedure models

Conclusions: The work is not done

1. Synthetic Raters are not a substitute for humans (they *are* neurodivergent rater peers)
2. Synthetic Rater Use cases are still surfacing
3. System prompt and metric prompts tuning is challenging
4. Document v. Contextual chunking (for multi-turn responses)
5. How to find *What's Missing* (citations, concepts, arguments, etc)
6. Metricate *What's Missing*
7. Can a Gen AI (consistently) check itself for binary quality metrics (e.g. hallucinations, misattribution, & unsupported statements)?

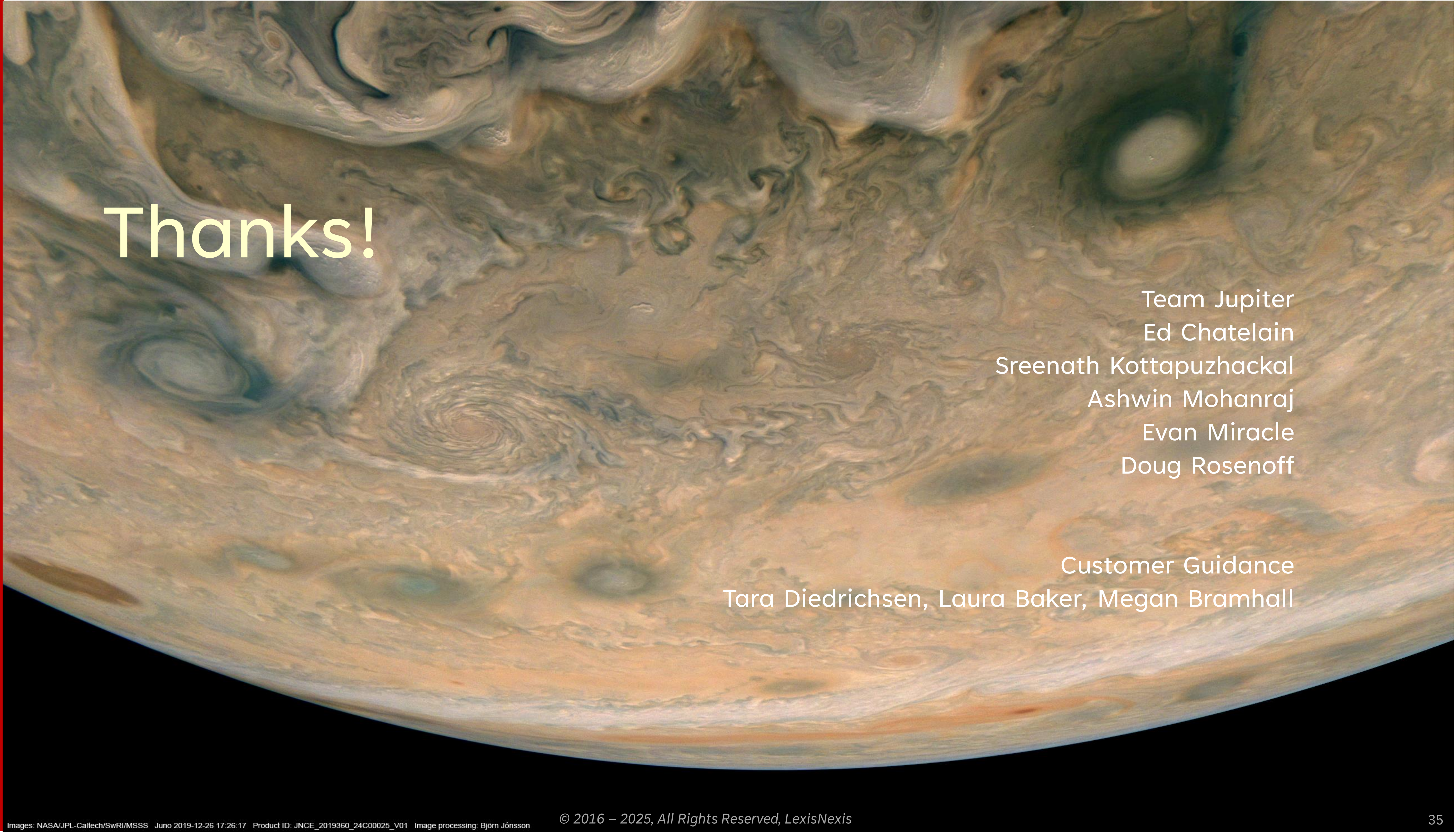
Questions?

JOIN US IN SHAPING A MORE JUST WORLD

Around the globe, LexisNexis employees are connected by the desire to shape a better world where the rule of law increases peace, prosperity, and justice. Everything that we do as a commercial business advances the rule of law. A career with LexisNexis can help you make a difference in the communities where we live and work. Join us on our journey today!

SCAN HERE TO LEARN MORE ABOUT OUR OPPORTUNITIES

For more information, email TalentAcquisition@lexisnexis.com



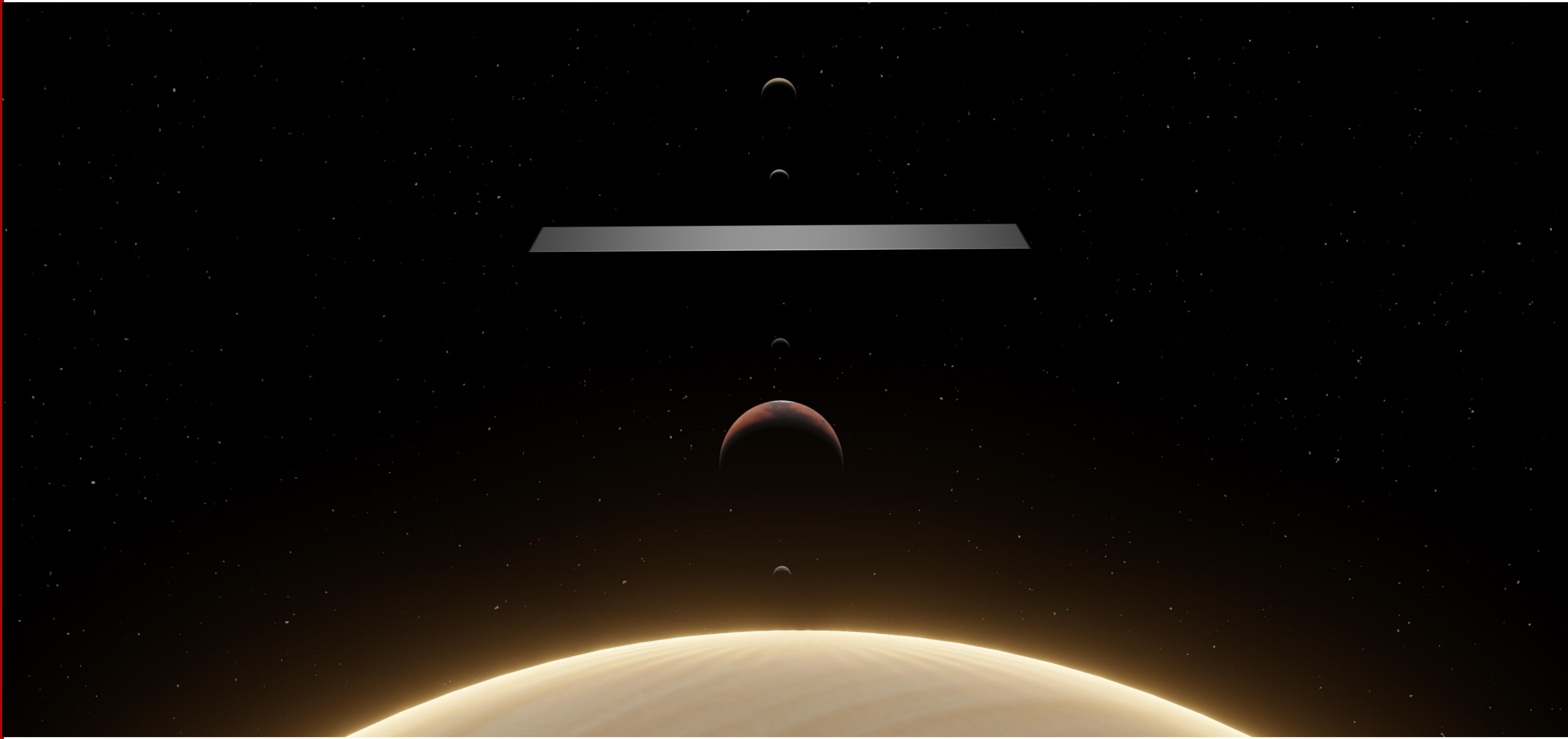
Thanks!

Team Jupiter
Ed Chatelain
Sreenath Kottapuzhackal
Ashwin Mohanraj
Evan Miracle
Doug Rosenoff

Customer Guidance
Tara Diedrichsen, Laura Baker, Megan Bramhall

Images: NASA/JPL-Caltech/SwRI/MSSS Juno 2019-12-26 17:26:17 Product ID: JNCE_2019360_24C00025_V01 Image processing: Björn Jónsson

Appendices



Author Biography

Ed Chatelain

edward.chatelain@lexisnexis.com

Ed is the Technical Lead for the Search Test Framework (STF), he has been working on STF since the project was started 7 years ago. Prior to that Ed worked at IBM for 29 years on a variety of interesting projects.

Author Biography

Vanaja Ganguru

Vanaja.Ganguru@lexisnexis.com

Vanaja is a Senior QA Engineer working on Search Test Framework. She has been with LexisNexis for nine years. Before joining LexisNexis, she completed her Master's degree in the United Kingdom and gained six years of QA experience in the financial domain.

Author Biography

Sreenath Kottapuzhackal

sreenath.kottapuzhackal@lexisnexis.com

Sreenath is a Senior software engineer working on the Search Test Framework. Before joining LexisNexis, he was a technology lead at Infosys and has worked on multiple LexisNexis projects for more than 15 years.

Author Biography

Evan Miracle

evan.miracle@lexisnexis.com

Evan Miracle lives in Raleigh, NC and has been with LexisNexis for 5 years. He is currently focused on customer behavior insight generation and evaluation of AI enabled products. He splits his free time between family, boardgame, and metal working.

Author Biography

Ashwin Mohanraj

Ashwin.Mohanraj@lexisnexis.com

Ashwin has over 16 years of experience in quality assurance and automation. His expertise spans the entire software development lifecycle, from requirements gathering to deployment, with a focus on automation and quality assurance, and lately, Generative AI.

Author/Speaker Biography

Doug Rosenoff

Email: douglas.rosenoff@lexisnexis.com / dtrosenoff@gmail.com

Doug is the Director of Global Product Search Testing Tools. He has worked for more than 28 years in electronic publishing and research, with domestic and international patents in Search Algorithms and Automatic Linking. He enjoys cats, photography, and music.

Abstract:

In the realm of Generative AI, human Subject Matter Experts (SMEs) are the gold standard for evaluating AI outputs across diverse domains such as medicine, law, and finance. However, the human evaluation process is resource-intensive, both in terms of time and cost. This presentation explores the innovative use of Generative AI-based Synthetic Raters as a cost-effective alternative for evaluating AI-generated content.

A Synthetic Rater is a composite of three elements: a trained Large Language Model (LLM) or similar AI construct, a set of system-level parameters (e.g., prompts), and metadata for identification and versioning. These components mirror those used in human rating processes, allowing for a seamless integration into existing evaluation frameworks. The primary distinction lies in the training and background differences between human and synthetic raters, which can be analyzed using comparison and regression tools within an SME rating framework.

Our research introduces a robust framework for SME-based evaluation that leverages both human and synthetic rater results. We conducted extensive tests using various LLMs and system prompts, comparing synthetic-to-synthetic and human-to-synthetic evaluations across multiple metrics. The findings reveal significant potential for synthetic raters to complement human evaluations, offering diverse perspectives and enhancing overall assessment quality.

This presentation will detail the common metrics employed, the testing methodologies, and the results of our evaluations. We will also explore practical use cases and propose innovative strategies for integrating human and synthetic ratings, ultimately paving the way for more efficient and scalable AI evaluation processes.