Agentic systems for assisting in SMH workflows

Srini Venkatramanan

SMH 2026 Annual Meeting

Baltimore, Maryland

Sep 23, 2026

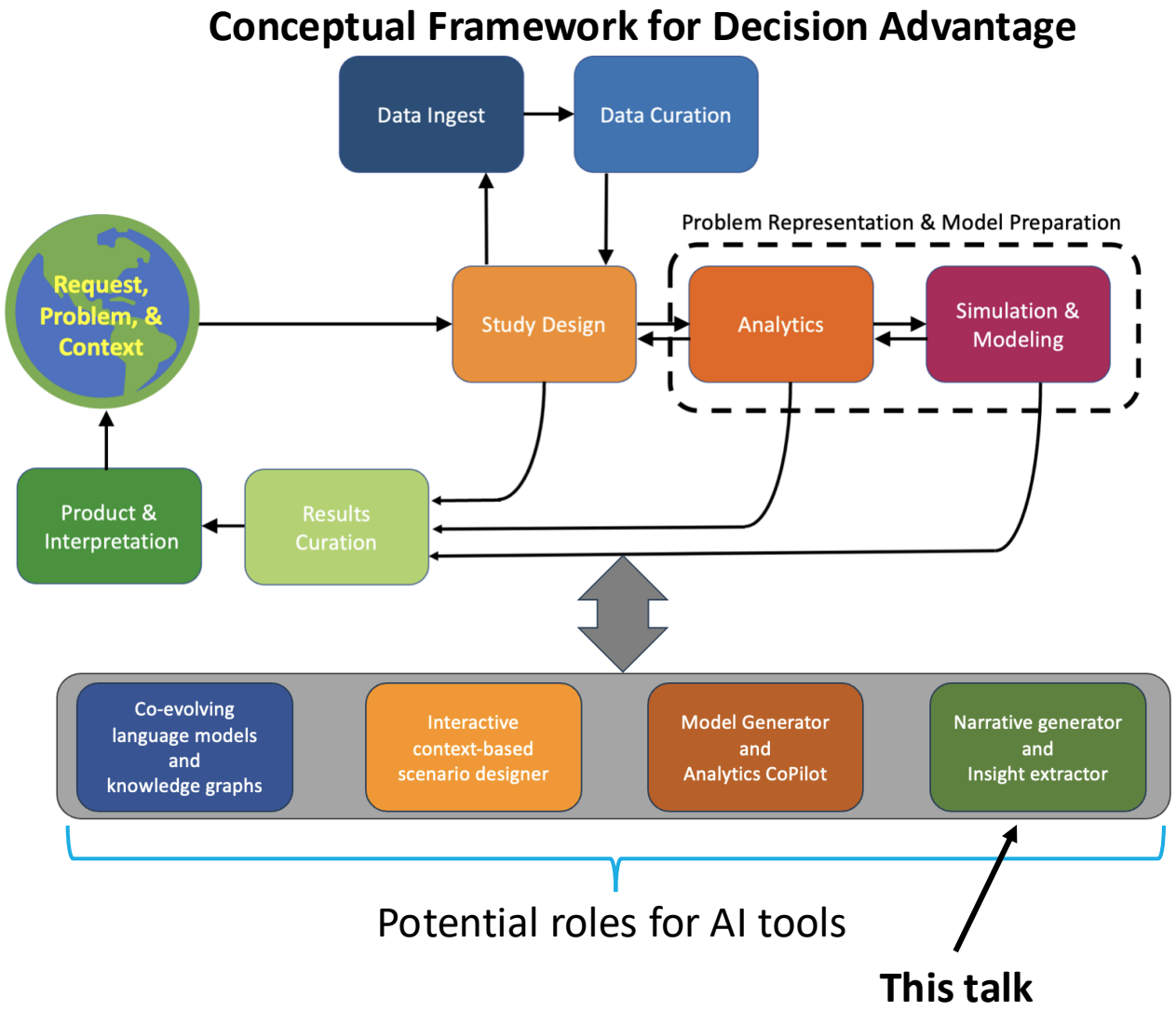


https://biocomplexity.virginia.edu

Extended Agency

- Most accepted form of agency in multi-agent systems: Planning Theory of Agency
  - Ability to choose your own goals and make effective plans to achieve them
  - Computationally operationalized as **Belief-Desire-Intention (BDI)** model
  - Shared agency emerges from structures of interconnected planning agency.

- Tools extend agency by extending:
  - our cognitive abilities, and
  - the range of actions



Michael Bratman. Intention, Plans, and Practical Reason. Harvard University Press, Cambridge, 1987.
Michael E. Bratman. Shared Agency: A Planning Theory of Acting Together. Oxford University Press, Oxford, 2013.
S. Swarup, "Agency in the Age of AI", https://arxiv.org/abs/2502.00648

EpiNarrate: Agentic Generation of Grounded Narratives from Epidemiological Scenario Projections

Rituparna Datta, Srini Venkatramanan, Bryan L. Lewis, Yiqi Su, Harry Hochheiser, Lucie Contamin, Parantapa Bhattacharya, Naren Ramakrishnan and Anil Vullikanti

University of Virginia · Virginia Tech · University of Pittsburgh

https://arxiv.org/abs/2607.15544

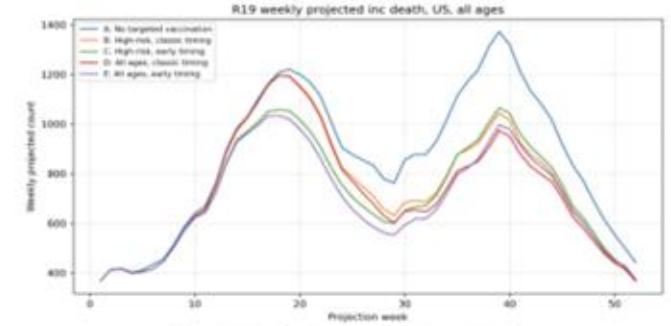
Motivation

SMH combines multi-model projections under alternative intervention assumptions. Experts translate these outputs into reports that compare strategies, quantify projected changes and identify vulnerable groups.

Why direct prompting is fragile

Large inputs force selection or compression. LLMs may mix scenario labels, compare different populations, miscalculate changes, or omit important findings despite fluent wording.

Input Data



Output (*EpiNarrate*)

- When combining these axes, shifting from high-risk/classic timing to a high-risk/early timing strategy (SD_010) reduces cumulative deaths by 2.3% (0-130), whereas moving from high-risk/classic timing to an all-ages/early timing strategy (SD_009) is projected to result in a 2.6% increase in cumulative deaths (0-130).

- In the baseline scenario with no targeted vaccination (Scenario A), we project **47,310 (95% PI: N/A) deaths** and **735,589 (95% PI: N/A) hospitalizations**. In the most optimistic intervention scenario (Scenario E: all ages targeted with early timing), the burden is reduced to **38,081 deaths** and **608,963 hospitalizations**.)

- If vaccination is restricted to high-risk groups → expanding eligibility to all ages is projected to provide an additional 2.2%–4.5% reduction across cumulative and weekly outcomes. Under classic timing, this corresponds to 29,407 fewer cumulative hospitalizations and 1,115 fewer cumulative deaths; with early timing,

...

Desired output

A concise narrative that preserves scenario assumptions, covers key comparisons, and links each quantitative claim to source values with an explicit calculation.

Input data, preprocessing and the reporting objective

Each SMH round defines a new scenario design and a multidimensional set of weekly ensemble projections.

Input component	Content and role
Scenario descriptions	Free-text assumptions on eligibility, timing, immune escape and other policies
Ensemble projections	Weekly hospitalization/death quantiles indexed by scenario, age and jurisdiction
Reference narratives	Expert national summaries used for reference, from previous rounds

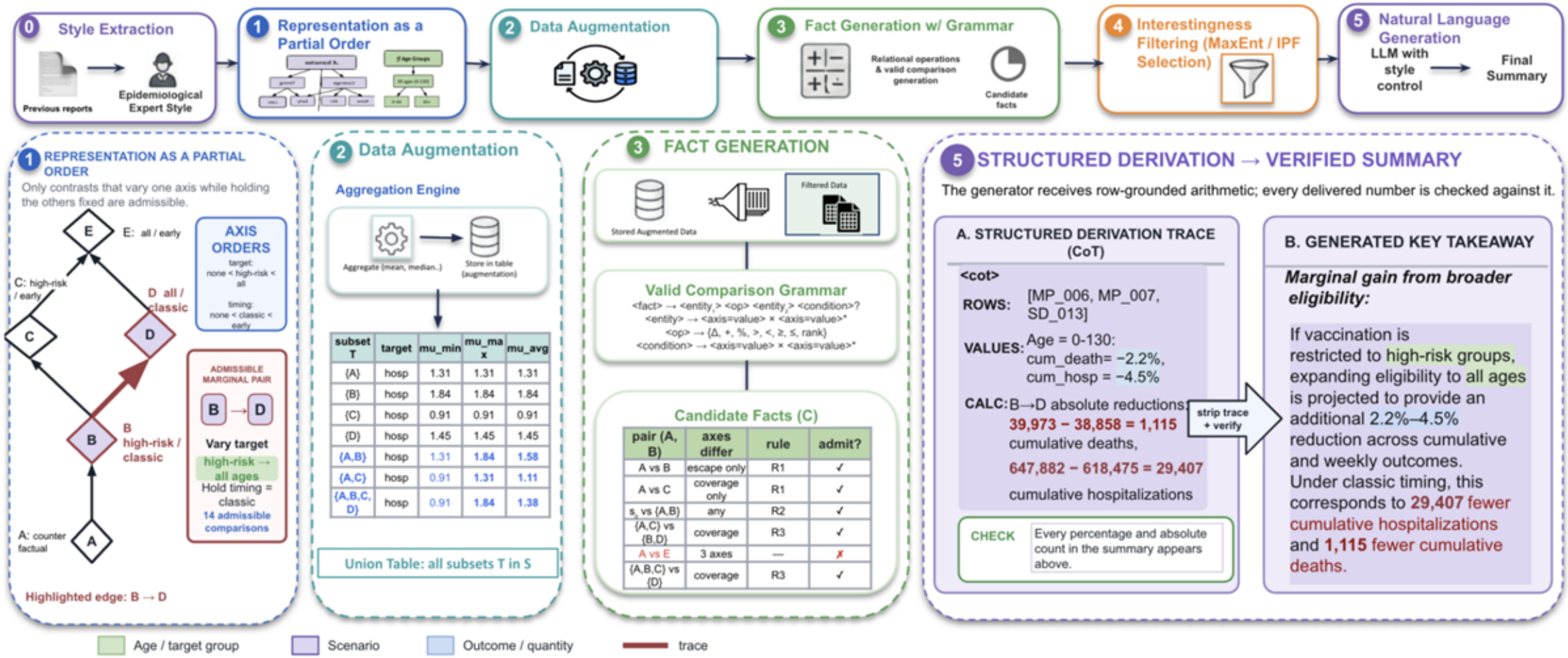
Preprocessing

Map columns to a canonical schema and summarize time series into compact statistics. Retain peak information, cumulative burden and the relevant quantile summaries.

Scale and constraints

Illustrative size: 5 scenarios × 3 outcomes × 4 groups × 50 regions × 52 weeks × 23 quantiles ≈ 3.6 million values. Compression must preserve comparison scope and uncertainty.

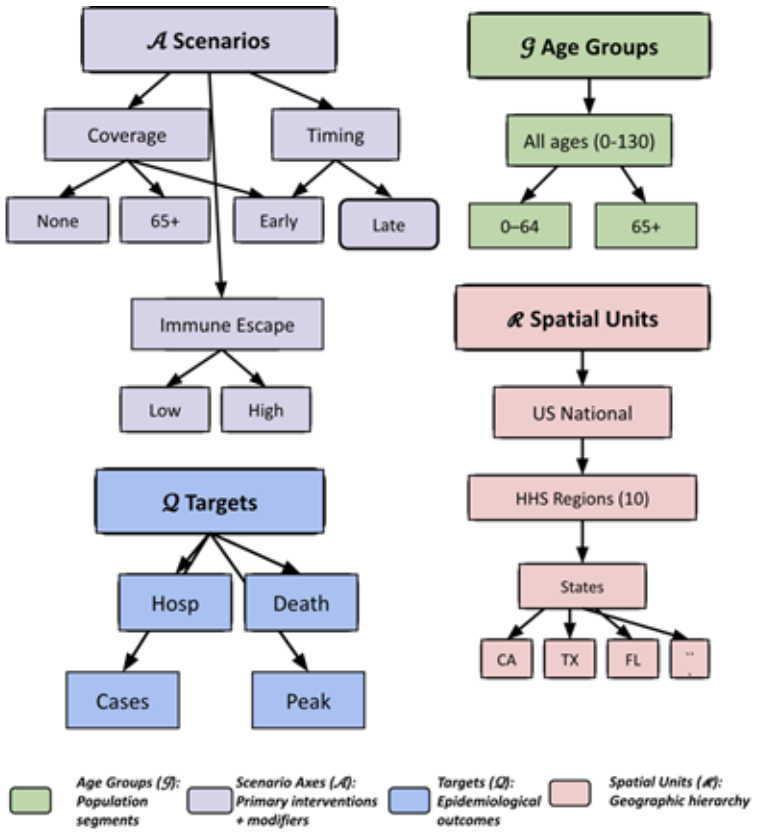
Architecture of EpiNarrate



Agentic components infer axes, schemas and reporting conventions. Structured operations aggregate data and compute admissible contrasts. The LLM narrates selected facts with an audit record.

STEP 1: Scenario decomposition and partial-order representation

Agents extract explicit axes from scenario descriptions and distinguish primary interventions from modifiers.



Round 19 active scenarios

Target group	Classic	Early
High-risk	B	C
All eligible	D	E

Primary axis: vaccination target (high-risk or all eligible).
Modifier axis: classic or early campaign timing.
Scenario A supplies the shared counterfactual.

B vs D changes eligibility only. B vs C changes timing only. B vs E changes both and cannot isolate either individual effect.

Left: general dimensions across SMH rounds; not all axes vary in Round 19.

STEP 2: Dataset augmentation and a worked numerical comparison

Augmented rows describe individual scenarios and scenario subsets, enabling matched comparisons at different granularities.

Subset	Shared / varying axes	Hospitalization statistic
B	High-risk; classic timing	647,882
D	All eligible; classic timing	618,475
{B, C}	High-risk; timing varies	642,903 (mean of member point estimates)
{D, E}	All eligible; timing varies	613,719 (mean of member point estimates)

Union-table construction

Five scenarios yield $2^5 - 1 = 31$ nonempty subsets. Across three reporting age groups, this gives 93 subset–age combinations. Shared and union axis values describe each subset.

Trace B vs D

647,882 − 618,475 = 29,407 fewer hospitalizations.
29,407 / 647,882 × 100 ≈ 4.54% lower relative to B.
National, all ages, same projection horizon.

STEP 3: Comparison grammar: which contrasts are admissible?

Comparison validity depends on scenario assumptions, not simply on whether two numerical rows are available.

Rule	Constraint	Round 19 example
R1: marginal pair	Change one axis; hold other axes fixed	B vs D: eligibility at classic timing B vs C: timing for high-risk groups
R2: grouped counterfactual	Baseline vs an intervention pool; variation within the pool is confined to modifiers	A vs {B, C}: high-risk intervention pooled over timing
R3: balanced pools	Compare pools matched on all non-varying axes	{B, C} vs {D, E}: eligibility, balanced on classic/early timing

Common scope

Keep outcome, age group, geographic unit and projection horizon aligned. Each emitted comparison receives an identifier that can be cited by the narrative.

STEP 5: Narration: evidence selection, interpretation and realization

The generator receives computed comparisons and reporting conventions instead of having to infer every fact from raw data.

Stage	Narration task	Retained evidence
Observe	Describe selected values, contrasts and regional patterns	Comparison IDs, raw values, residuals and aggregates
Reflect	Choose 2–3 relevant findings and explain their relevance	A concise finding and a grounded interpretation
Synthesize	Write an executive summary with explicit scenario labels	Clean prose plus the supporting derivation record

Style and coverage control

Prior expert reports guide naming, quantitative reporting and uncertainty language. A checklist tracks comparison IDs cited in ROWS. Missing IDs can trigger another generation pass.

What the evidence record contains

ROWS identifies comparisons. VALUES stores exact quantities. CALC shows derived arithmetic or ranges. CHECK records consistency with the intended scope.

A numerical claim and its evidence record

Round 19: matched eligibility comparisons, holding campaign timing fixed within each pair.

“Expanding booster eligibility from high-risk people to all ages provides an incremental 2.2%–2.8% reduction in cumulative deaths.”

Evidence field	Supporting information
ROWS	MP_002 and MP_003: matched eligibility comparisons
VALUES	All-age cumulative deaths: −2.8% and −2.2%
CHECK	Keep outcome, population and comparison scope explicit

The range spans matched comparisons, not a prediction interval. Auditability allows reviewers to check the numerical statement but does not prove every interpretation correct.

Broader benchmark: fidelity, claim count and coverage

M1 measures factual fidelity, where “100%/30” denotes 100% fidelity over 30 extracted numerical claims; M2 and M3 measure coverage and style, respectively. Bold denotes the highest score for each component within a round

Model	R19			R18			R17		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
Decoder-based / LLM baselines									
LIDA	38%/8	0.17	0.50	20%/10	0.00	0.42	25%/8	0.00	0.57
Single-shot CoT	0%/4	0.00	0.43	12%/8	0.00	0.48	83%/6	0.50	0.38
Multi-stage CoT	60%/5	0.42	0.45	0%/5	0.00	0.46	40%/5	0.00	0.50
Plan-then-Generate	0%/6	0.33	0.48	20%/5	0.00	0.46	20%/5	0.00	0.60
DeepAnalyze-8B	0%/21	0.42	0.67	0%/30	0.44	0.71	7%/28	0.06	0.58
DeepAnalyze$_{Gemma}$	81%/37	0.50	0.72	**100%**/22	0.32	0.43	73%/18	0.25	0.65
InsightBench$_{GPT4o}$	100%/18	0.07	0.78	100%/16	0.00	0.76	80%/20	0.25	0.59
GPT 5.5	**100%**/13	0.35	**0.79**	88%/28	0.40	**0.86**	**100%**/15	0.28	**0.85**
Claude Opus 4.6	0.9%/11	0.38	0.09	60%/32	0.38	0.80	75%/23	0.25	0.82
Encoder–decoder baselines									
T5-Large	33%/3	0.33	0.34	67%/3	0.00	0.34	100%/1	0.00	0.48
BART	0%/1	0.00	0.34	0%	0.00	0.06	0%/1	0.00	0.38
Ours									
EpiNarrate	**100%**/30	**0.78**	0.73	**100%**/36	**0.71**	0.81	**100%**/26	**0.62**	0.79

LLM judges: structured frontier variants as Judge

LLM-as-a-judge results are reported as correctness/coverage/style/total, with respective maxima of 40/30/30/100. Bold denotes the highest score for each component within a round, underlining marks EpiNarrate as runner-up.

Model	R18		R19	
	Opus judge (4.8)	GPT judge (5.6)	Opus judge (4.8)	GPT judge (5.6)
EpiNarrate$_{gpt}$	36/22/19/77	**40**/23/17/<u>80</u>	36/22/20/78	**40**/24/18/<u>82</u>
EpiNarrate$_{opus}$	39/**28**/21/**88**	**40**/**26**/12/78	39/**29**/22/**90**	**40**/**26**/12/78
EpiNarrate	36/20/**24**/80	38/12/24/74	**40**/21/**24**/<u>85</u>	**40**/21/18/79
GPT 5.5	**40**/23/22/85	34/23/26/**83**	38/18/21/77	**40**/23/**27**/**90**
Claude Opus 4.6	33/17/20/70	30/17/**27**/74	39/6/16/61	24/3/21/48

LLM as a judge (Claude Opus 4.8)

Summary	M1 · Correctness /40	M2 · Coverage /30	M3 · Style /30	Total /100
National Report Round 20	**38** — All 12 averted figures match the computation table within rounding; absolute BAU burden (240k hosp / 11.5k deaths) exactly parquet-verifiable; correctly labels figures as prospective (Jun 2026–Jun 2027). Minor: two-wave peak timing claims unverifiable from cumulative data.	**28** — Broadest: absolute anchor, all scenario contrasts, both marginal steps, seasonality, VE=55%, immune-escape/waning/variant caveats, state heterogeneity, calibration design. Gap: no explicit D-v-B / E-v-C coverage contrasts.	**26** — Reference exec-summary register: bolded lead-ins, CIs in prose, assumptions quarantined into notes. Reads as a complete summary.	**92**
EpiNarrate	**38** — 14/14 exact against the computation table (percentages, absolutes, and 95% paired intervals to the decimal); cleanest fidelity of any summary. Minor: averted figures carry no period qualifier under the full-span framing.	**22** — Full frequency×coverage grid + unique D-v-B / E-v-C coverage contrasts and the "higher coverage amplifies frequency benefit" synthesis. Misses absolute anchor, two-wave timing, VE=55%, state heterogeneity.	**24** — Terse parallel bullets with counts + paired intervals; good register match. Held back by repetitive templating across seven near-identical bullets.	**84**

LLM as a judge (GPT 5.6 sol)

Summary	M1 · Correctness /40	M2 · Coverage /30	M3 · Style /30	Total /100
National Report Round 20	**40** — All numerical claims, including the prospective BAU burden, scenario differences, percentages, and uncertainty intervals, are faithful to the ensemble data after rounding.	**30** — Covers every reference information unit: absolute burden, major and marginal scenario comparisons, temporal patterns, VE assumptions, uncertainty, calibration, state heterogeneity, and limitations.	**29** — Clear and well-organized executive-summary narrative with appropriate context and caveats. Only a minor deduction for a few dense sentences and bullets.	**99**
EpiNarrate	**40** — Every reported hospitalization and death difference, percentage, and paired interval agrees with the computation table. No unsupported numerical claim was identified.	**18** — fully covers the C-vs-B and D-vs-A units, and partially covers B-vs-A and E-vs-D because one outcome is omitted. It does not report the absolute prospective BAU burden. It nevertheless includes useful D-vs-B and E-vs-C contrasts absent from the national report.	**25** — Concise, auditable, and easy to verify, with consistent presentation of estimates and intervals. Repetitive sentence structures and limited narrative synthesis prevent a higher score.	**83**

Limitations

Boundary	Implication for interpretation and use
Approximate metrics	M1 depends on LLM-based extraction/verification; M3 is an indirect similarity measure
Interestingness	A large MaxEnt residual can require more context and may not merit a narrative claim
Human oversight	Experts review scenario assumptions, uncertainty, interpretation and final public health language

Further evaluation

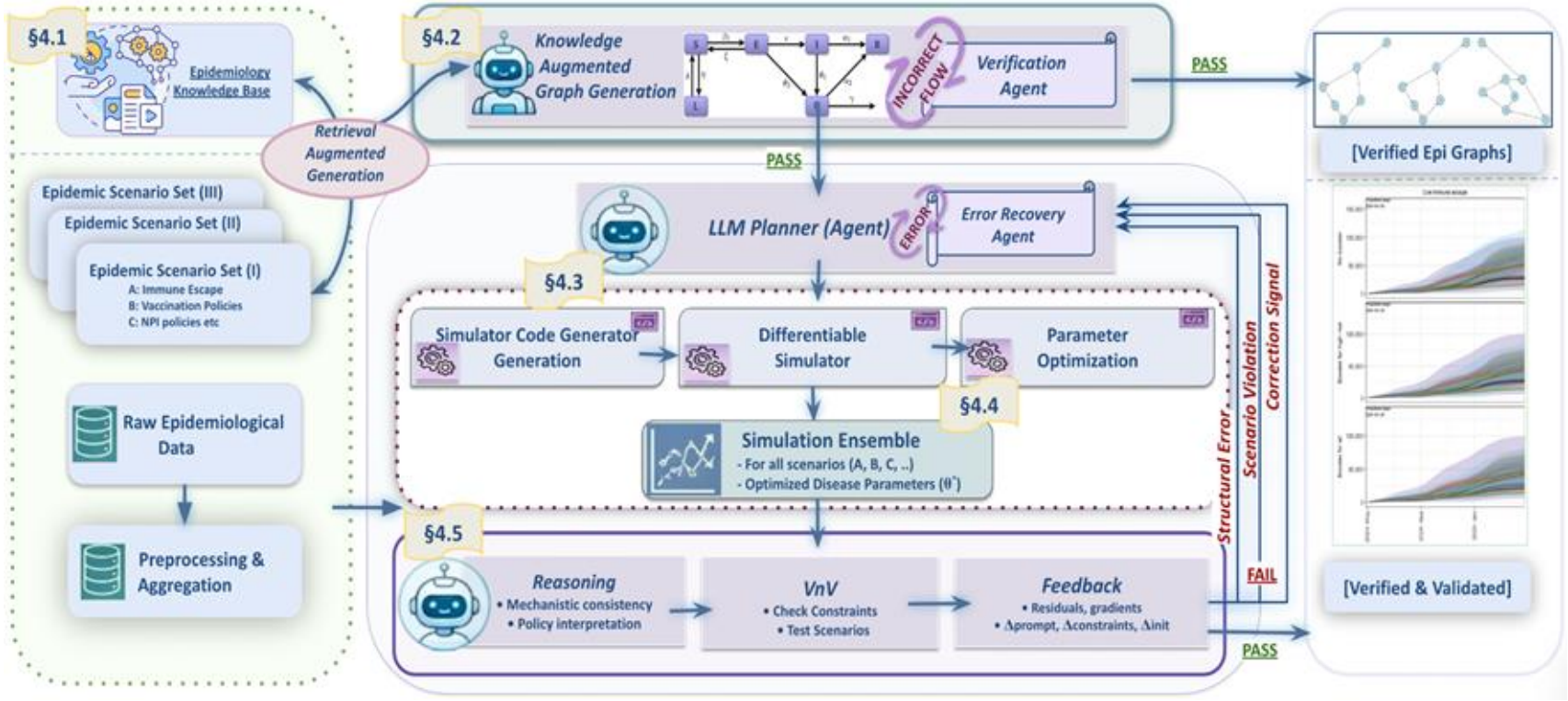
Priorities include additional pathogens, different scenario designs, repeated runs and stronger expert assessment.

Treat the system as a drafting and prioritization assistant.

EpiAgent: An Agentic Framework for Epidemiological Modeling

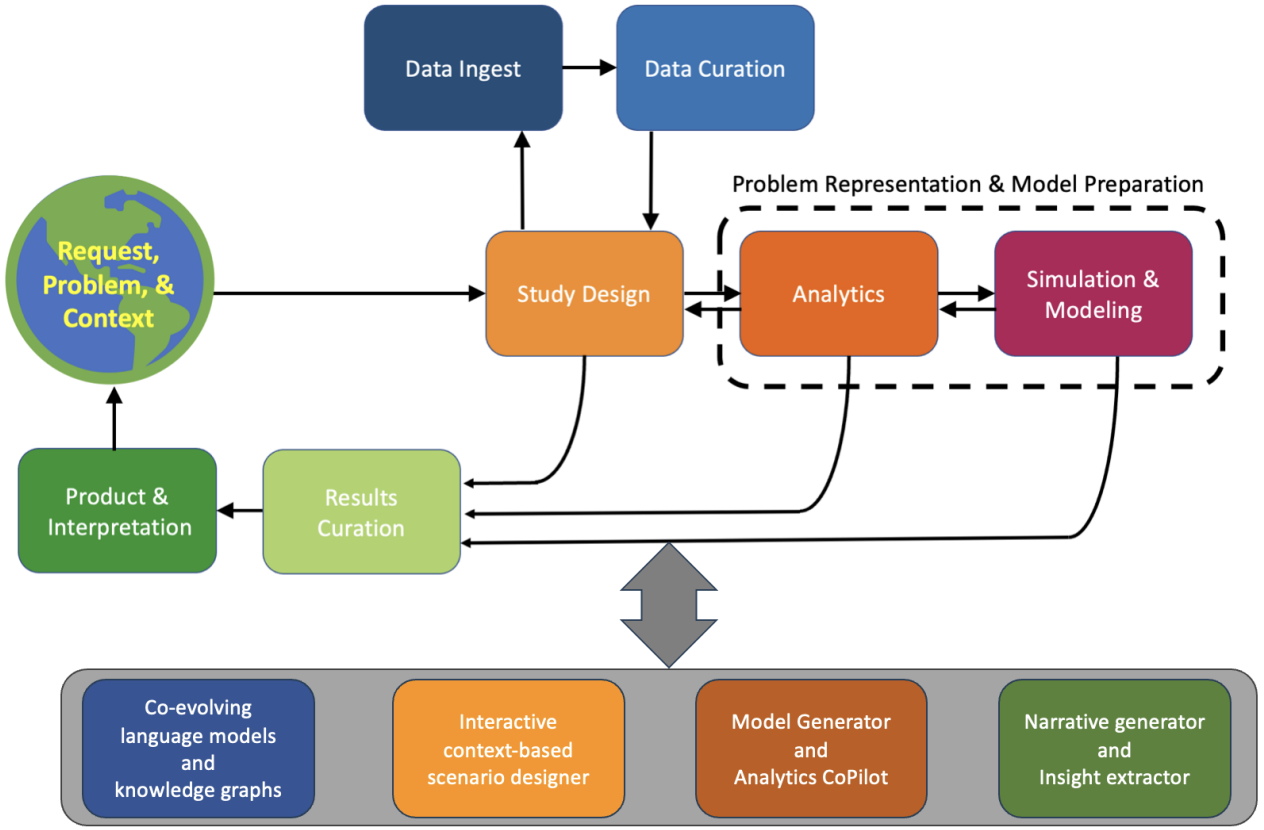
R. Datta · Z. Guan · B. Espinoza · Y. Su · P. Pitre · S. Venkatramanan · N. Ramakrishnan · A. Vullikanti
University of Virginia · Biocomplexity Institute · Virginia Tech

Presented at International Conference on Machine Learning (ICML) 2026



Breakout session: AI4SMH: Discussion Topics

- What **other ways** can AI systems assist in the SMH process?
- How do we get around the challenge of **validation in real-time**?
- What are the implications for **training the next generation**?
- How to mitigate the risks of making such tools **broadly available**?



What role does an epi modeler* play ~~in the year 2040~~?

during the next outbreak?

**if not extinct*

Appendix

Narrative bullets linked to comparison records

The original audit trail connects clean summary text with comparison IDs, exact values, calculations and checks.

Final summary	COT analysis from Augmented Table -> Filtered Table
* **Intervention Analysis:** Targeting high-risk populations alone reduces cumulative deaths by 16.6% (0-130) [GC_006]; targeting all ages alone reduces cumulative deaths by 18.7% (0-130) [GC_005]. Regarding timing, using classic immunization timing alone reduces cumulative deaths by 16.7% (0-130) [GC_007], while shifting to early immunization timing alone reduces cumulative deaths by 18.6% (0-130) [GC_008]. When combining these axes, shifting from high-risk/classic timing to a high-risk/early timing strategy (SD_010) reduces cumulative deaths by 2.3% (0-130), whereas moving from high-risk/classic timing to an all-ages/early timing strategy (SD_009) is projected to result in a 2.6% increase in cumulative deaths (0-130). * **Marginal Gains:** Expanding the booster target from high-risk populations to all ages provides an incremental reduction in cumulative deaths of 2.2% to 2.8% (0-130) [MP_002, MP_003]. Similarly, shifting the vaccination campaign from classic to early timing provides a marginal benefit, reducing cumulative deaths by 2.0% to 2.6% (0-130) [MP_001, MP_004].	<cot> ROWS: [GC_006, GC_005, GC_007, GC_008, SD_010, SD_009] VALUES: [GC_006: -16.6, -12.6, -14.6, -12.1; GC_005: -18.7, -16.6, -17.5, -15.4; GC_007: -16.7, -13.9, -14.3, -13.0; GC_008: -18.6, -15.2, -17.8, -14.6; SD_010: -2.3, -1.5, -4.0, -1.8; SD_009: +2.6, +4.8, +3.5, +3.9] CALC: N/A CHECK: All comparison IDs for intervention axes are present. </cot> * **Intervention Analysis:** Targeting high-risk populations alone reduces cumulative deaths by 16.6% (0-130) [GC_006]; targeting all ages alone reduces cumulative deaths by 18.7% (0-130) [GC_005]. Regarding timing, using classic immunization timing alone reduces cumulative deaths by 16.7% (0-130) [GC_007], while shifting to early immunization timing alone reduces cumulative deaths by 18.6% (0-130) [GC_008]. When combining these axes, shifting from high-risk/classic timing to a high-risk/early timing strategy (SD_010) reduces cumulative deaths by 2.3% (0-130), whereas moving from high-risk/classic timing to an all-ages/early timing strategy (SD_009) is projected to result in a 2. <cot> ROWS: [MP_001, MP_004, MP_002, MP_003] VALUES: [MP_001: -2.6, -0.6, -3.9; MP_004: -2.0, -1.4, -3.6; MP_002: -2.8, -5.1, -3.1; MP_003: -2.2, -5.9, -2.7] CALC: Range for timing (MP_001/004): -2.0% to -2.6% (deaths); Range for targeting (MP_002/003): -2.2% to -2.8% (deaths). CHECK: All marginal pair IDs are present. </cot> * **Marginal Gains:** Expanding the booster target from high-risk populations to all ages provides an incremental reduction in cumulative deaths of 2.2% to 2.8% (0-130) [MP_002, MP_003]. Similarly, shifting the vaccination campaign from classic to early timing provides a marginal benefit, reducing cumulative deaths by 2.0% to 2.6% (0-130) [MP_001, MP_004].

These indicate filtered row from augmented table

Maximum-entropy selection identifies additional geographic findings

National summaries can miss state-level departures from expected regional and metric patterns.

Operation	Purpose
Normalize fields	Express absolute burden and averted counts per 100,000 population before comparing states
Fit a baseline with IPF	Match specified HHS regional and field-category / outcome marginals
Rank residuals	Prioritize state–metric cells with large departures from the fitted expectation
Control repetition	Cap selections per field and deduplicate complementary age-share findings

Selection diagnostics

For the top 15 cells, the analysis uses at most 3 selections per field. Selection entropy describes whether findings are spread across states rather than concentrated in one location.

Experimental setup and baseline families

Evaluation uses SMH rounds 17, 18 and 19, with expert-written national summaries as references.

Configuration	Setting
Backbone	Gemma 4 31B, without fine-tuning
Generation	Temperature 0.0 for reasoning; 0.2–0.4 for generation
Execution	Reported runtime 3–5 minutes per round; prompts/responses logged to JSONL

Prompting and planning baselines

Schema-only LIDA; direct GPT 5.5 and Claude Opus 4.6; single-shot and multi-stage CoT; a reimplemented Plan-then-Generate cascade.

Other comparisons and fairness

DeepAnalyze and InsightBench represent agentic analysis. T5-Large and BART run zero-shot because only three paired rounds are available. Use the same backbone where the method permits.

Direct LLM baselines receive a stratified subsample of the hypercube. Input representation and context limits affect the comparison.

Evaluation: fidelity, coverage and reference similarity

A useful narrative should be numerically supported and informative. No single metric captures both.

Metric	Definition	How to interpret it
M1: numerical fidelity	N verified / N extracted claims	Report the denominator. A high score with very few claims can be uninformative.
M2: reference coverage	(N covered + 0.5 × N partial) / N reference units	Measures recovery of the expert report. New valid findings may not appear in the reference.
M3: embedding similarity	Cosine similarity of generated and reference embeddings	An indirect similarity measure; it does not independently establish factual correctness.

Verification is approximate

M1 relies on LLM-based claim extraction and checking.

Complementary assessments

The paper also reports LLM-judge correctness /40, coverage /30 and style /30, plus expert review of MaxEnt findings. These answer different questions.

Ablations distinguish numerical grounding from comparison coverage

Round 19 component removals hold the remaining pipeline fixed.

Round 19 variant	Fidelity / claims	Coverage
Full EpiNarrate	100% / 30	0.78
Without CoT	70% / 46	0.52
Without partial order	100% / 27	0.28

Without the derivation record

M1 falls from 100% to 70% even as claims increase from 30 to 46. Coverage falls from 0.78 to 0.52. Additional numbers are less reliably grounded.

Without the partial order

M1 remains 100% over 27 claims, but M2 falls to 0.28. Correct individual statements can still omit most relevant scenario comparisons.