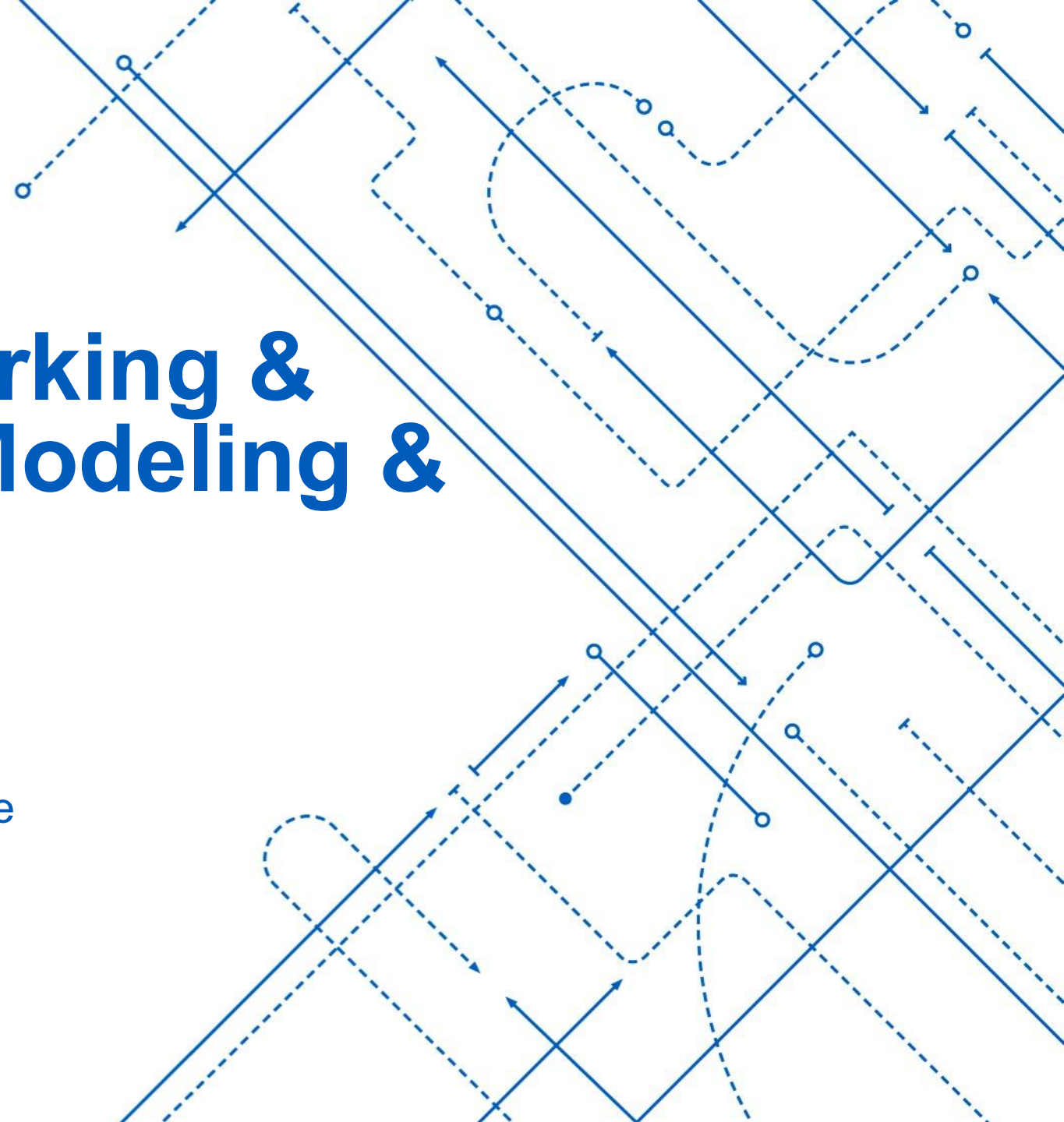
BEAMS: Benchmarking & Evaluating AI for Modeling & Simulation

Sara Metcalf and Billy Schoenberg

International System Dynamics Conference
Delft, The Netherlands | July 20-24, 2026

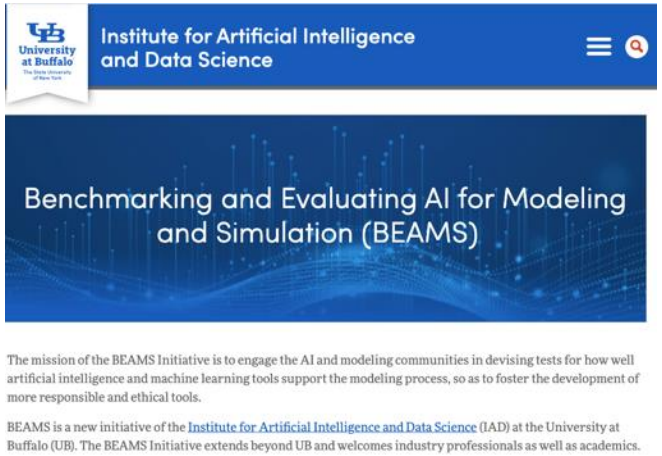
University at Buffalo
Institute for Artificial Intelligence and Data Science



Overview

- Introduction
- BEAMS Infrastructure
- Evaluation Design Principles
- Evaluations by Engine Type
- Evaluation Results
- Implications
- Conclusions

BEAMS Initiative



bit.ly/BEAMSinitiative

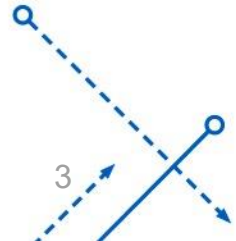
Introduction

Mission of the BEAMS Initiative

- To engage the AI and modeling communities in devising tests for how well AI and ML tools support the modeling process, so as *to foster the development of more responsible and ethical AI tools* for modeling and simulation through open collaboration on benchmarks

Questions motivating the initiative

- How do we drive LLM builders to make responsible and ethical AI tools that suit the modeling and simulation community?
- How do we start to measure the performance of AI tools for modeling and simulation tasks?



University at Buffalo
Institute for Artificial Intelligence and Data Science

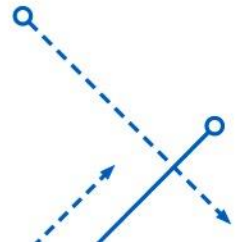
BEAMS Infrastructure

Open Source Platform

- sd-ai project on GitHub
  - Interface between modeling software client and external LLMs
    - Request handler
    - Structured output response format
  - Repository for developing and sharing
    - AI-enabled modeling tools
      - Engines
      - Agentic tools
    - Evals
      - Evaluation of AI-enabled tools in conjunction with external LLMs

Open Collaboration

- Working Groups
  - Steering group
    - Identifies and prioritizes evaluations
  - Technical group
    - Implements prioritized evaluations
- Ad hoc subgroups
  - Agent-based modeling extensions
  - Consideration of end-use cases
- Annual Open Forums
- Online discussion board
  - groups.io/g/sd-ai

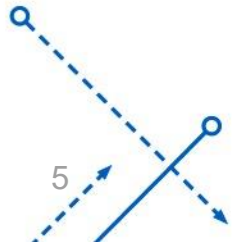


Framework for Evaluation Design

- Principles
  - How should AI-enabled automation improve dynamic modeling?
- Goals
  - What aspects of AI-enabled tools and their outputs can be measured and evaluated to spur their development in a way that improves dynamic modeling?
- Tests
  - How do we measure those aspects?

BEAMS Principles for AI Evaluation

1. Do no harm
2. Develop tools that help people
3. Increase access to the modeling process for all
4. Complement human abilities, don't substitute for the human
5. Provide information without bias, stereotypes, or generalizations
6. Work with the modeling process, don't bypass it
7. Deliver high quality models
8. Use appropriate information in building models



From Principles to Evaluations

Principles	Goals	Tests
7. Deliver high quality models	Accurately translate causal information; Eliminate hallucinated information in generated models	Causal Translation
6. Work with the modeling process, don't bypass it	Maintain interaction in the modeling process	Model Iteration
1. Do no harm; 2. Develop tools that help people; 7. Deliver high quality models; 8. Use appropriate information in building models	Present reasoning in terms of feedback and delays; Eliminate hallucinated information in generated models	Causal Reasoning
2. Develop tools that help people	Adhere to the user's intent	Conformance
3. Increase access to the modeling process for all; 2. Develop tools that help people	Help stakeholders understand the dynamics of a model; Present reasoning in terms of feedback and delays	Model Behavior Explanation
6. Work with the modeling process, don't bypass it; 4. Complement human abilities, don't substitute for the human	Maintain interaction in the modeling process; Present a logical and understandable approach that can be replicated	Suggested Model Building Steps
4. Complement human abilities, don't substitute for the human	Critique model and provide feedback to modeler	Suggested Model Fixes

Evaluations of Qualitative AI Tools

Causal Translation [eval: qualitativeTranslation]

- Evaluates AI tool's ability to create causal relationships from plain language

Model Iteration [eval: qualitativeIteration]

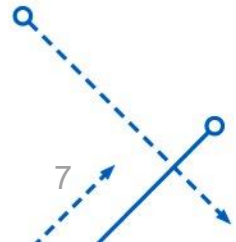
- Evaluates AI tool's ability to add new causal relationships to existing CLDs

Causal Reasoning [eval: qualitativeCausalReasoning]

- Evaluates whether AI-generated CLDs include key variables and causal mechanisms that domain experts consider essential

Conformance [eval: conformance]

- Evaluates AI tool's ability to follow user instructions



Evaluations of Quantitative AI Tools

Causal Translation [eval: quantitativeTranslation]

- Evaluates AI tool's ability to create structured representations from plain language

Model Iteration [eval: quantitativeIteration]

- Evaluates AI tool's ability to add new relationships to existing stock-flow models

Causal Reasoning [eval: quantitativeCausalReasoning]

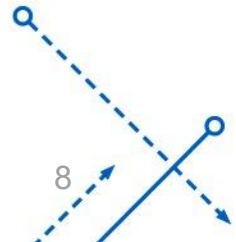
- Evaluates whether AI-generated models include key stocks, flows, and causal mechanisms that domain experts consider essential

Conformance [eval: conformance]

- Evaluates AI tool's ability to follow user instructions

Suggested Model Fixes [eval: quantitativeErrorFixing]

- Evaluates AI tool's ability to identify and fix model formulation errors



Evaluations of AI Discussion Tools

Model Behavior Explanation [eval: feedbackExplanation]

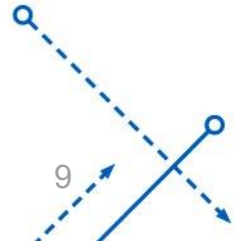
- Evaluates whether AI tool accurately explains dynamic behavior of models

Suggested Model Building Steps [eval: modelBuildingSteps]

- Evaluates whether AI tool generates appropriate steps to build a system dynamics model given a problem statement and background knowledge

Suggested Model Fixes [eval: errorFixingSuggestions]

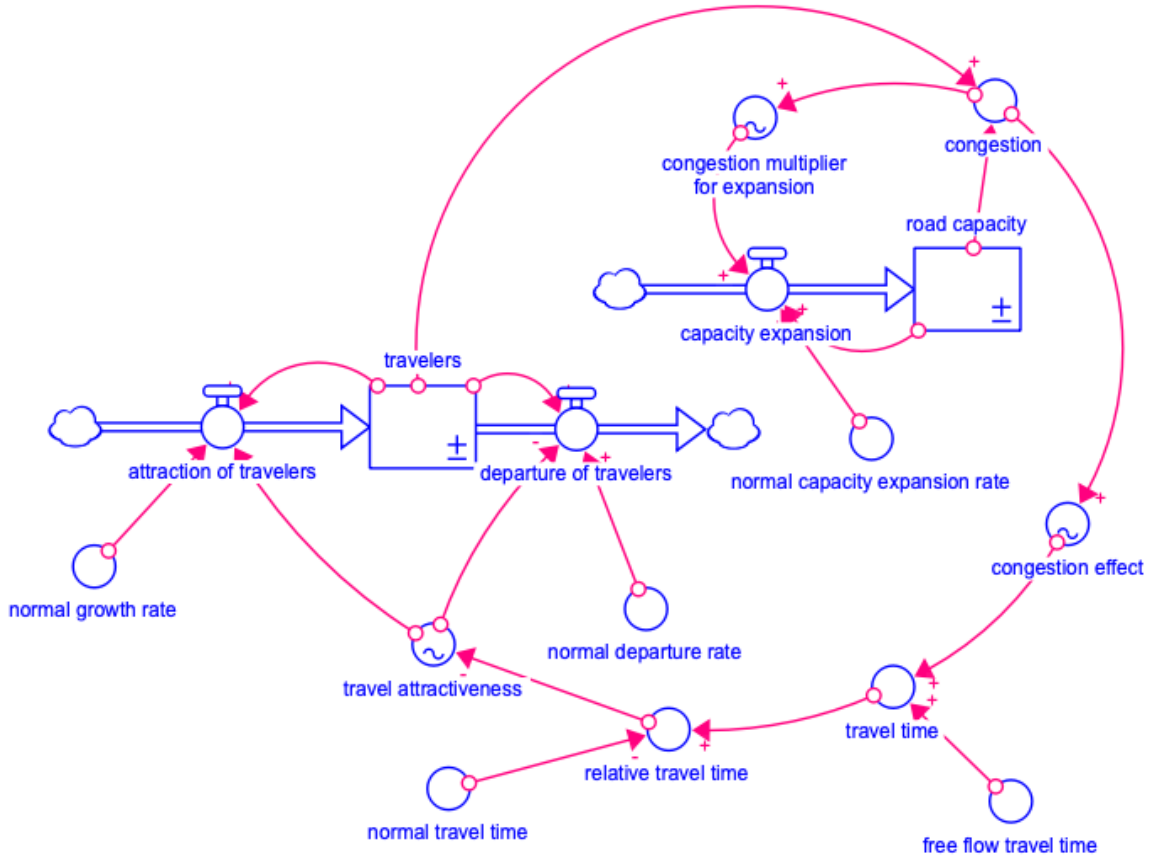
- Evaluates whether AI tool can identify and explain model formulation errors



Tests by Evaluation and Engine Type

	Engine Type		
Evaluation Type	**Qualitative**	**Quantitative**	**Discussion**
Causal Translation	24	9	
Model Iteration	8	9	
Causal Reasoning	3	3	
Conformance	18	18	
Model Behavior Explanation			6
Suggested Model Building Steps			4
Suggested Model Fixes		7	7
Total Tests	**53**	**46**	**17**

Example Causal Reasoning Test



sd-ai URL https://ai.iseesystems.com

Agent | Basic

I'm trying to understand the key causal mechanisms that drive this system's behavior over time.

Background Knowledge

The system involves road infrastructure, traveler behavior, and traffic volume dynamics.

Use these variable names:
road capacity, travelers, travel time, congestion.

AI Assistant (Prompt)

What can I do for you?

Submit

AI Reasoning

You instructed me to...

Please give me a quantitative model that captures the key causal processes described in the background information.

Here is what I have done...

Induced Demand Dynamics in Transportation

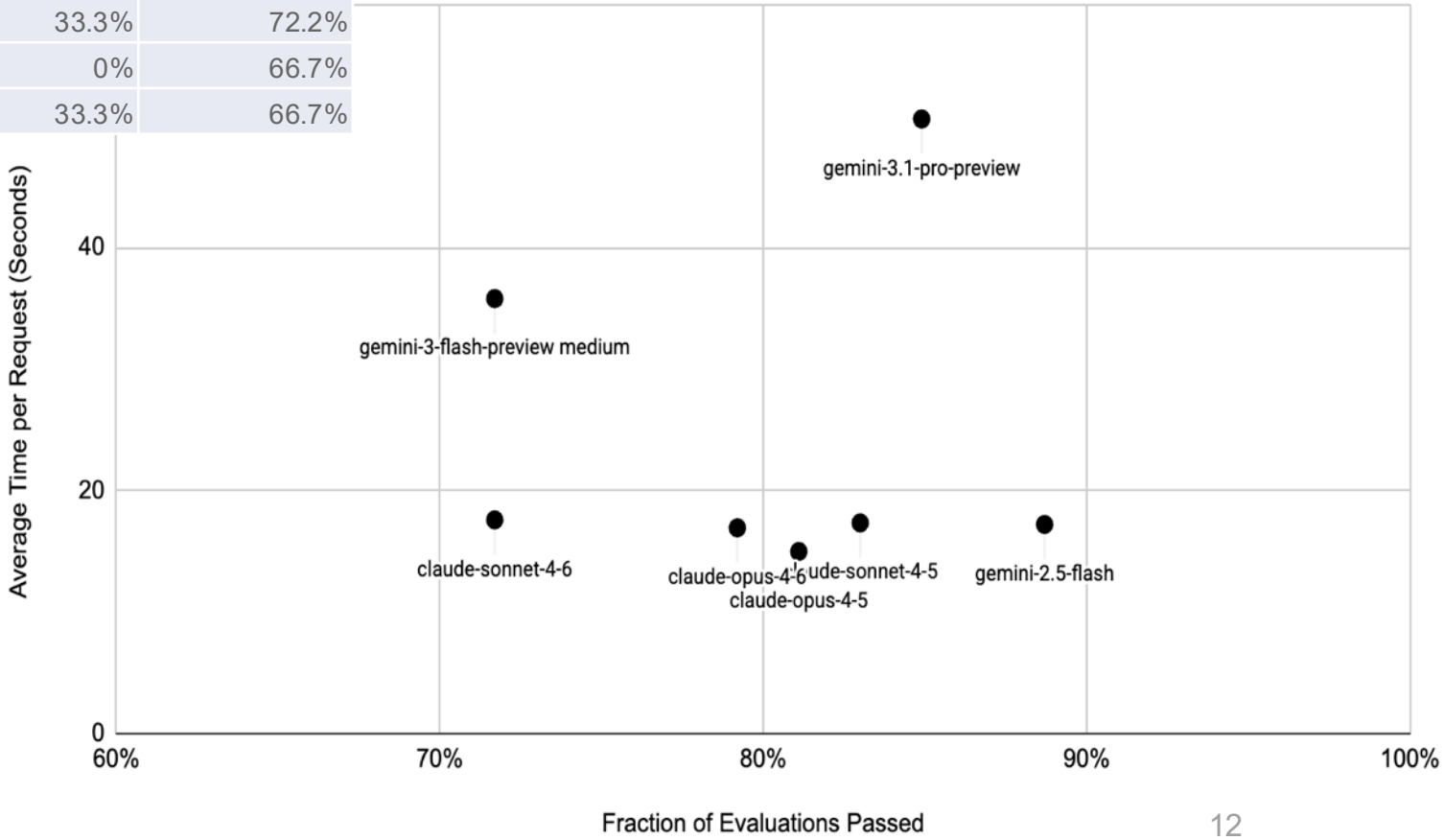
This model represents the classic system archetype of 'Fixes that Fail' applied to traffic infrastructure. The system contains two main feedback loops. A reinforcing loop occurs as new road capacity is added, which initially reduces congestion and travel times, driving up travel attractiveness. This attractiveness prompts an inflow of

University at Buffalo
Institute for Artificial Intelligence and Data Science

Qualitative Tool Evaluation Results

LLM coupled with Qualitative Engine	Overall Score	Causal Translation	Model Iteration	Causal Reasoning	Conformance
gemini-2.5-flash	88.7%	100%	87.5%	33.3%	83.3%
gemini-3.1-pro-preview	84.9%	100%	100%	0%	72.2%
claude-sonnet-4-5	83%	95.8%	62.5%	33.3%	83.3%
claude-opus-4-5	81.1%	87.5%	75%	0%	88.9%
claude-opus-4-6	79.2%	95.8%	62.5%	33.3%	72.2%
gemini-3-flash-preview medium	71.7%	87.5%	62.5%	0%	66.7%
claude-sonnet-4-6	71.7%	83.3%	62.5%	33.3%	66.7%

Explore the Qualitative Leaderboard:
ub-iad.github.io/sd-ai/#/leaderboard/cld

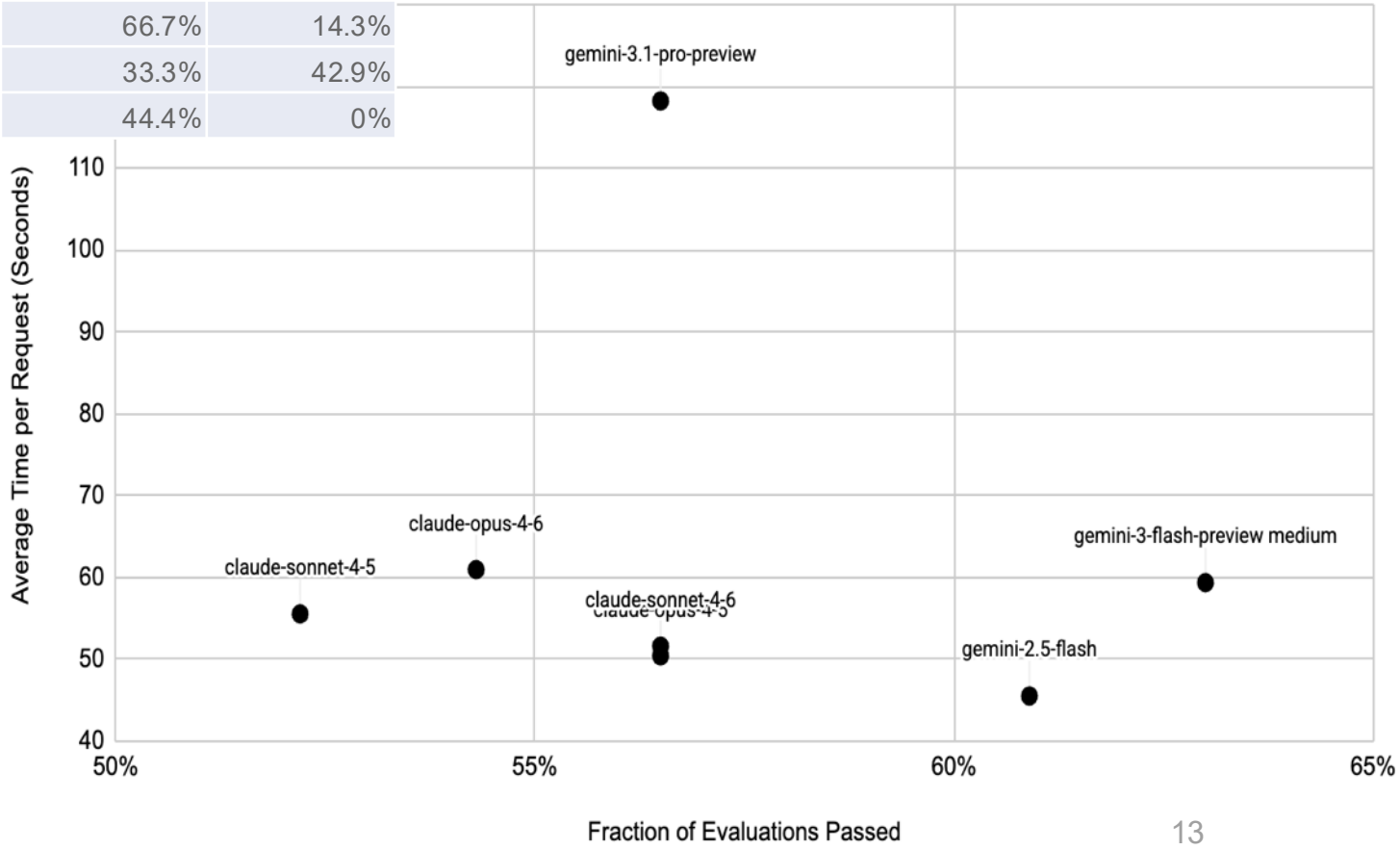


University at Buffalo
Institute for Artificial Intelligence and Data Science

Quantitative Tool Evaluation Results

LLM coupled with Quantitative Engine	Overall Score	Causal Translation	Model Iteration	Causal Reasoning	Conformance	Suggested Model Fixes
gemini-3-flash-preview medium	63%	77.8%	66.7%	66.7%	66.7%	28.6%
gemini-2.5-flash	60.9%	88.9%	77.8%	66.7%	44.4%	42.9%
gemini-3.1-pro-preview	56.5%	77.8%	88.9%	66.7%	50%	0%
claude-opus-4-5	56.5%	88.9%	77.8%	33.3%	44.4%	28.6%
claude-sonnet-4-6	56.5%	88.9%	44.4%	33.3%	66.7%	14.3%
claude-opus-4-6	54.3%	88.9%	77.8%	33.3%	33.3%	42.9%
claude-sonnet-4-5	52.2%	88.9%	77.8%	33.3%	44.4%	0%

Explore the Quantitative Leaderboard:
ub-iad.github.io/sd-ai/#/leaderboard/sfd

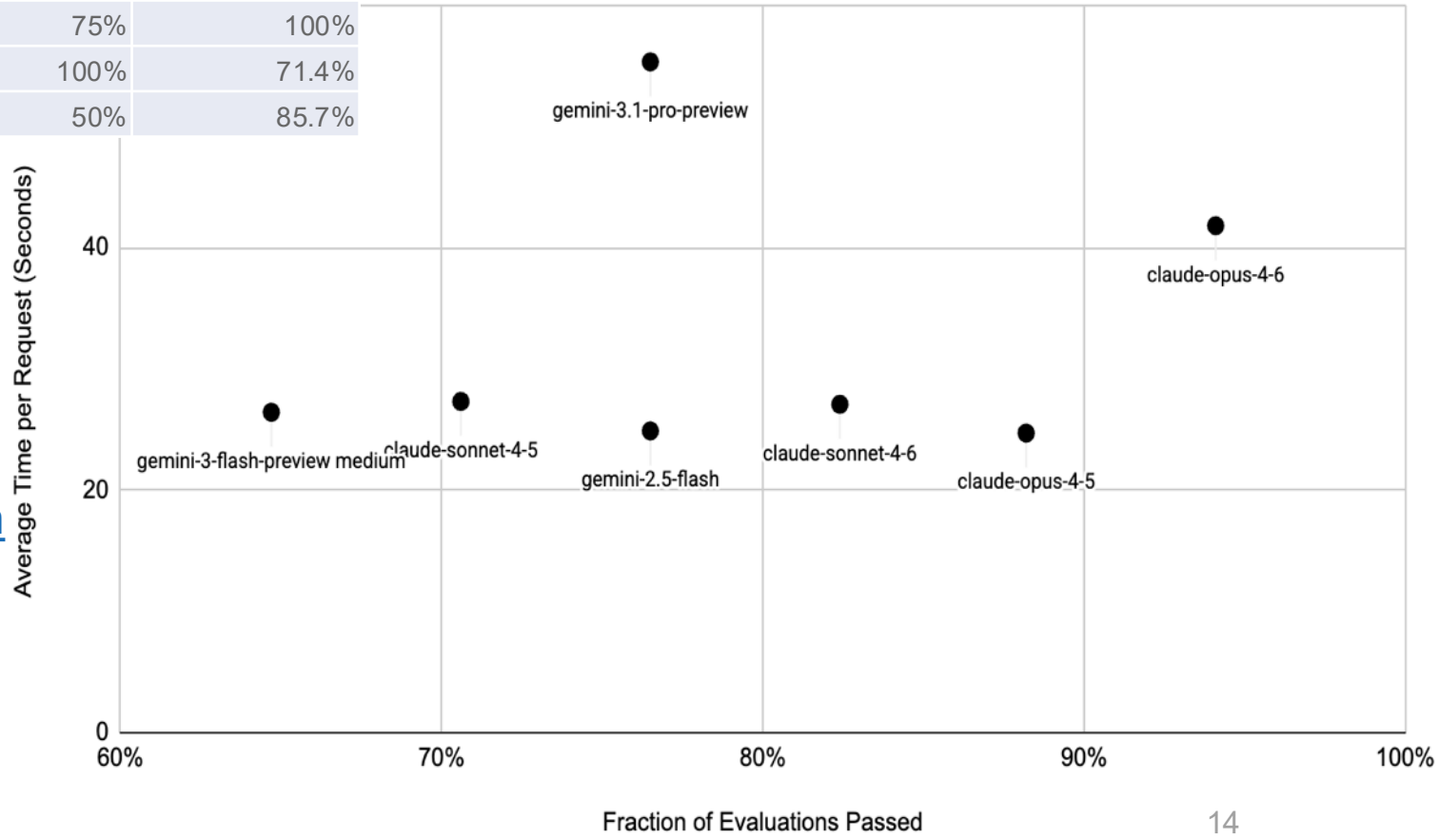


University at Buffalo
Institute for Artificial Intelligence and Data Science

Discussion Tool Evaluation Results

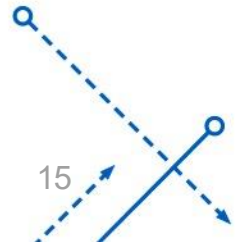
LLM coupled with Seldon Engine	Overall Score	Model Behavior Explanation	Suggested Model Building Steps	Suggested Model Fixes
claude-opus-4-6	94.1%	83.3%	100%	100%
claude-opus-4-5	88.2%	66.7%	100%	100%
claude-sonnet-4-6	82.4%	83.3%	100%	71.4%
gemini-2.5-flash	76.5%	66.7%	75%	85.7%
gemini-3.1-pro-preview	76.5%	50%	75%	100%
claude-sonnet-4-5	70.6%	50%	100%	71.4%
gemini-3-flash-preview medium	64.7%	50%	50%	85.7%

Explore the Discussion Leaderboard:
ub-iad.github.io/sd-ai/#/leaderboard/discussion



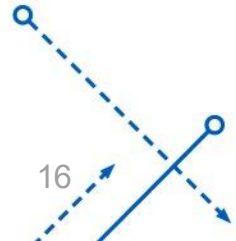
Implications for Engine Use

- Front-load described structure for qualitative tasks
  - Provide explicit variable lists, hypothesized feedback loops, and constraints
- Scaffold quantitative model building
  - Decompose into micro-tasks, small additions, and use tools like Seldon to feed criticisms of generated models back into the LLM for targeted fixes
- Leverage discussion engines like Seldon as intermediary
  - Use model behavior explanation to produce interpretable narratives, suggested modeling steps to drive model construction, and suggested fixes to pre-screen before automated model editing
- Optimize for speed under constraints
  - Prefer LLMs that meet a known accuracy threshold quickly
- Adopt per-task LLM routing
  - Given rough equivalence among overall top performers despite task-specific variability, select the LLMs with the best results for each BEAMS evaluation type



Conclusions

- BEAMS evaluations demonstrate that AI-enabled modeling tools perform better at discussion and basic qualitative tasks than causal reasoning and quantitative error fixing
- AI-enabled modeling tool performance assessed through BEAMS evaluations reveals variability across LLMs coupled with engines
- No single LLM dominates across qualitative, quantitative, and discussion engine types
- Some tradeoffs exist between speed and accuracy of performance
- Ongoing efforts of the BEAMS initiative aim to create benchmarks that address bias by considering alternative perspectives and human-centered use cases

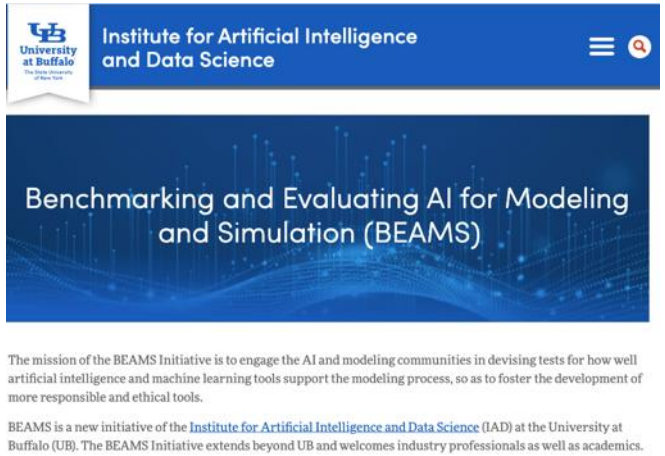


University at Buffalo
Institute for Artificial Intelligence and Data Science

This is an Open Invitation to Join Us

Questions?

BEAMS Initiative



bit.ly/BEAMSinitiative

Contact us:

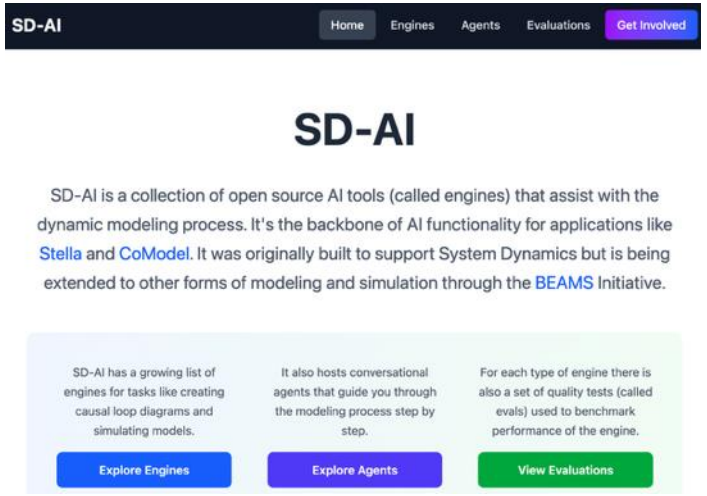
Sara Metcalf
smetcalf@buffalo.edu

Billy Schoenberg
bschoenberg@iseesystems.com

Join Our Discussion Board

groups.io/g/sd-ai

Open sd-ai Project



ub-iad.github.io/sd-ai/