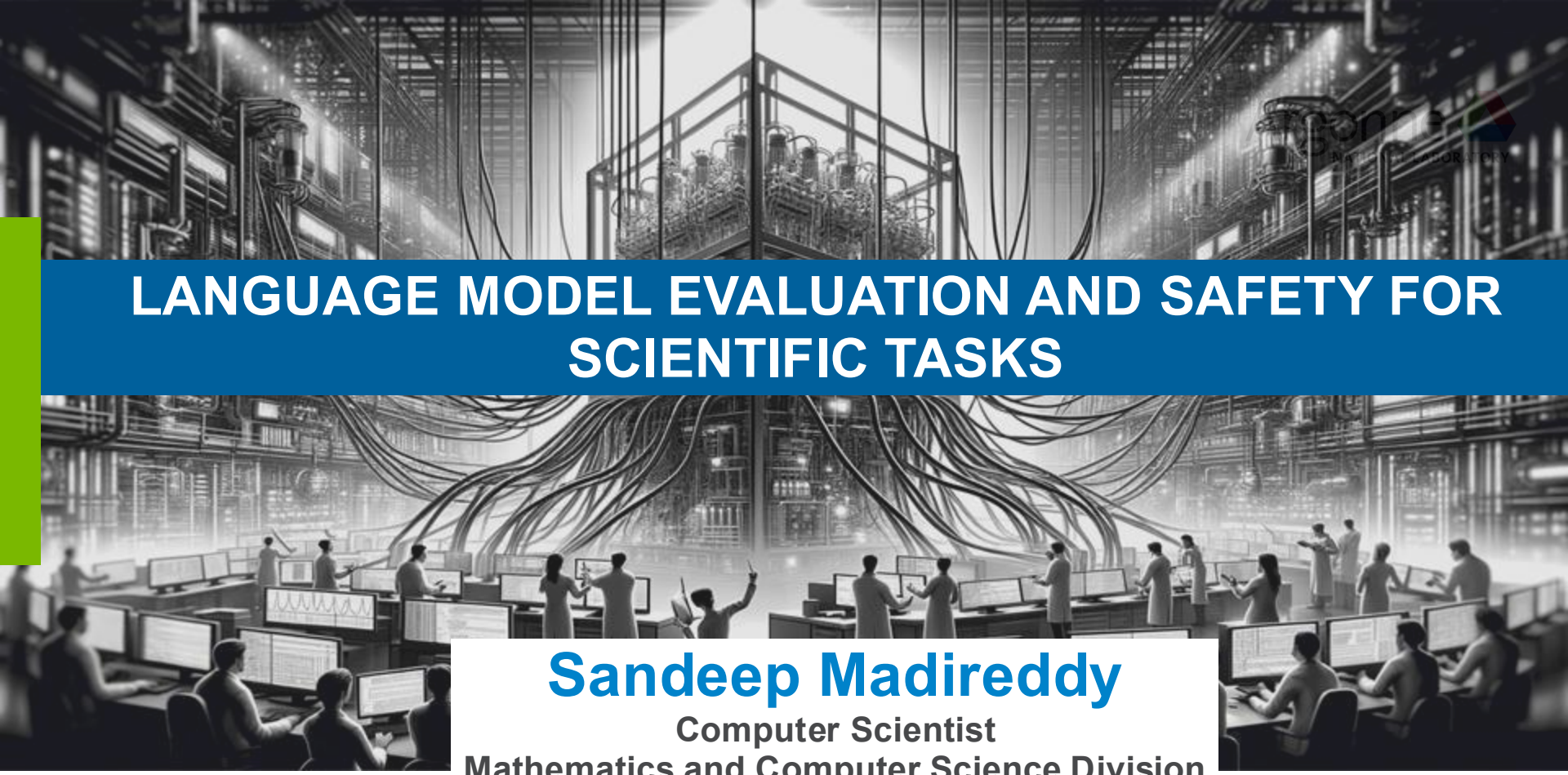


LANGUAGE MODEL EVALUATION AND SAFETY FOR SCIENTIFIC TASKS

Sandeep Madireddy

Computer Scientist
Mathematics and Computer Science Division
Argonne National laboratory

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

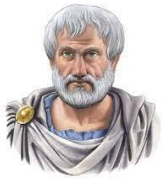
Slides credits: Ian Foster, Franck Cappello, Rick Stevens, Bo Li, AuroraGPT evaluation team

ATPESC 2025

THE PROGRESSION OF THE SCIENTIFIC METHOD

Increasing speed, automation, and scale

Accelerated Discovery



Empirical Science

1st Paradigm

Observations

Experimentation

Theoretical Science

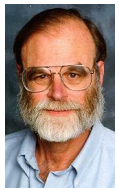
2nd Paradigm

Scientific laws in physics, biology, chemistry, etc.

Computational Science

3rd Paradigm

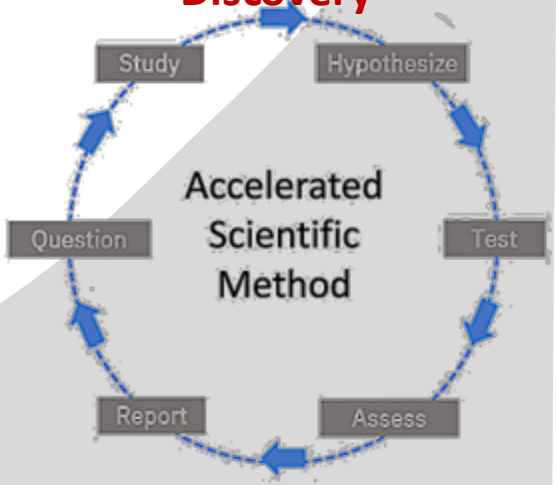
- Simulations
- Molecular dynamics
- Mechanistic models



Big Data-driven Science

4th Paradigm

- Big data, machine learning
- Patterns, anomalies
- Visualization

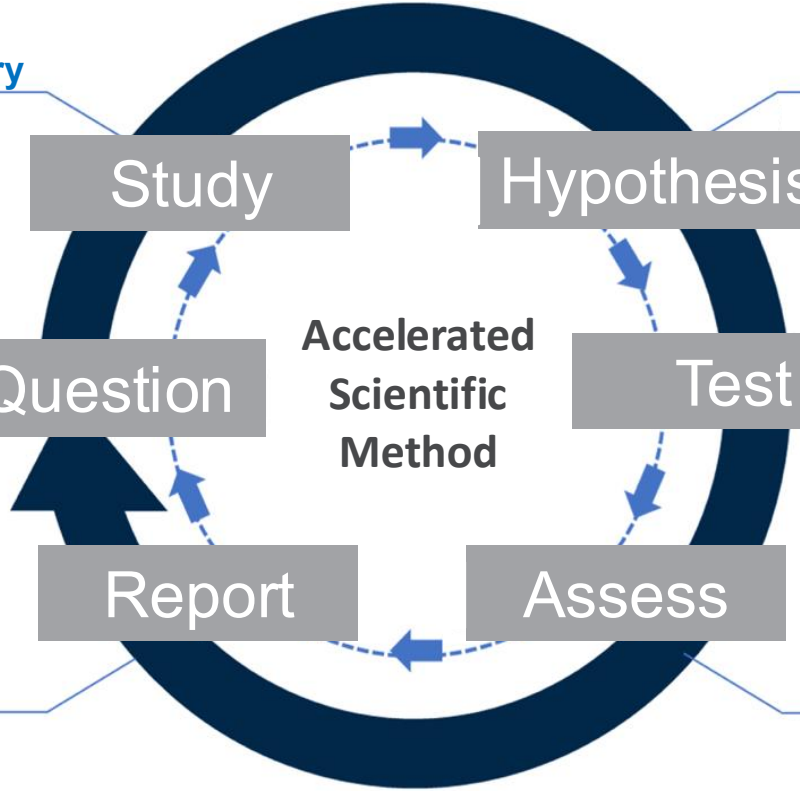


- **Scientific knowledge at scale**
- **AI-generated hypotheses**
- **Autonomous testing**

1600s 1950s 2000s 2020s

https://doi.org/10.1038/s41524-022-00765-z

ACCELERATING DISCOVERY



Search capabilities
Literature search, preliminary Data collection, Experiment, Simulation, Observation

Reasoning capabilities
Devise a research plan
Problem solving
Generate hypothesis
Propose innovative directions

Start here
With a research question

Agentic Capabilities
Experimental design
Launch Simulation, Experiment, Observation Campaigns. Self Driving Labs

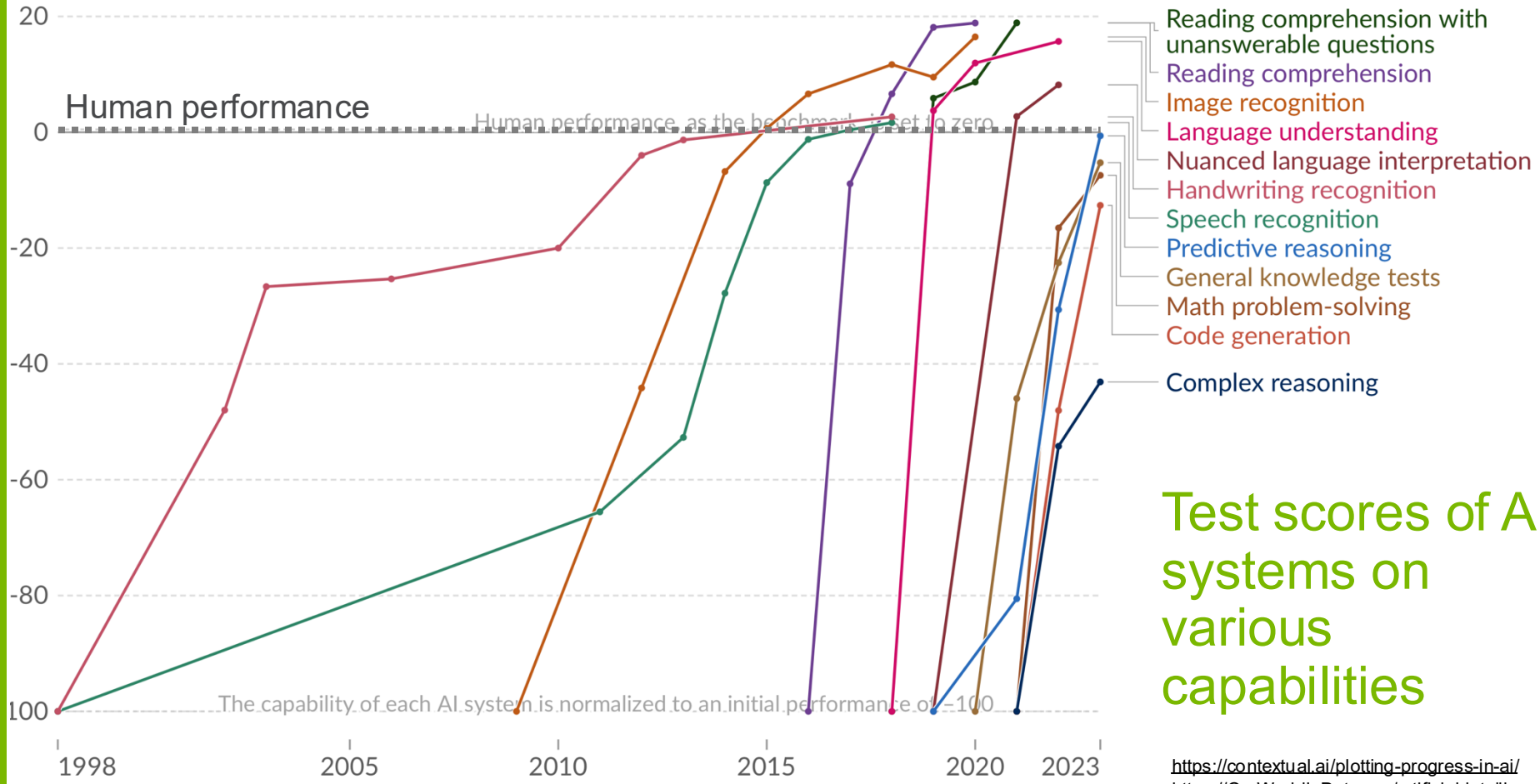
Reporting capabilities
Report generation (multi-modal)
Result Explanation
Conclusion generation

Analysis capabilities
Data (regression, correlation, etc.)
Result verifications, UQ
Hypothesis validation

An ideal research will help in all these steps

https://doi.org/10.1038/s41524-022-00765-z

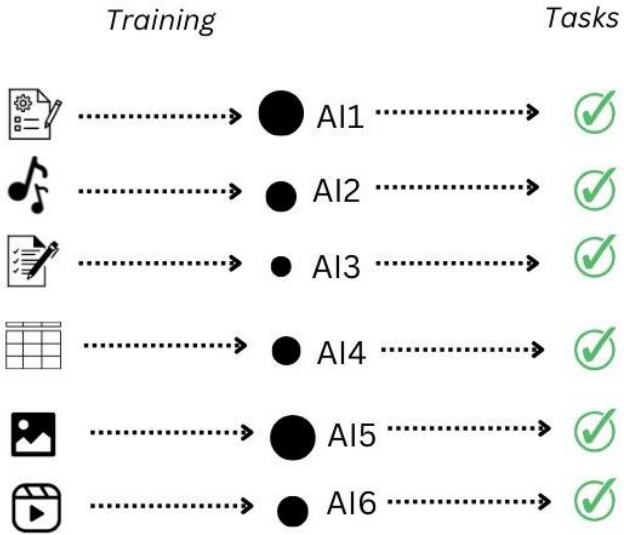
AI system capabilities are increasing rapidly



Test scores of AI systems on various capabilities

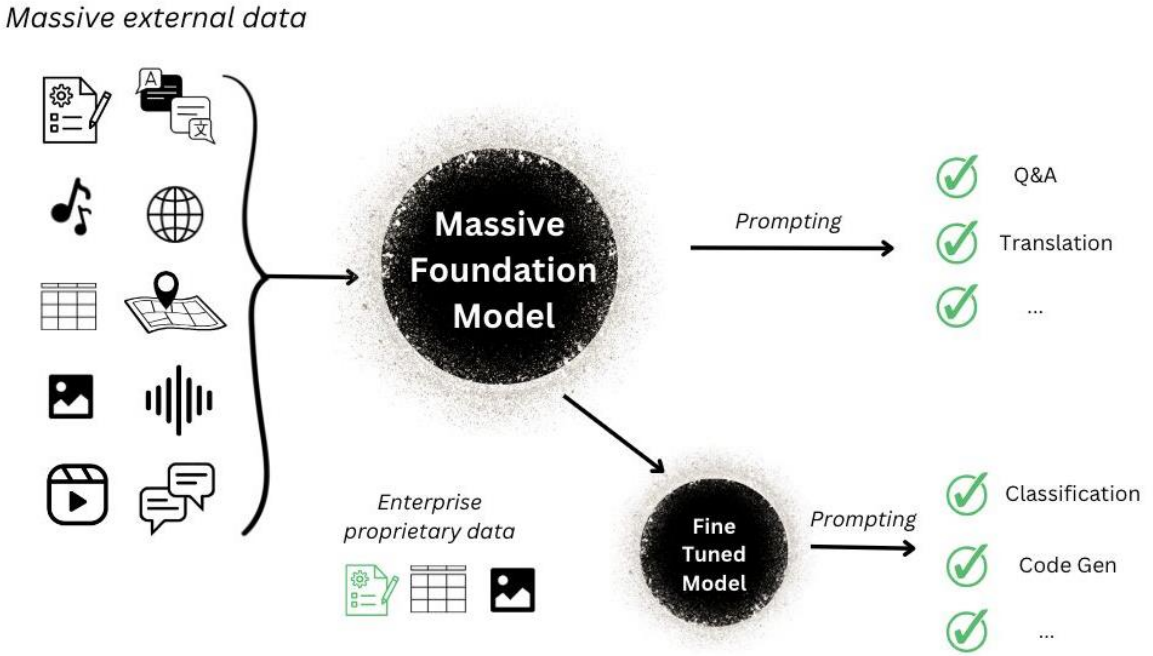
https://contextual.ai/plotting-progress-in-ai/
https://OurWorldInData.org/artificial-intelligence

Traditional ML



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

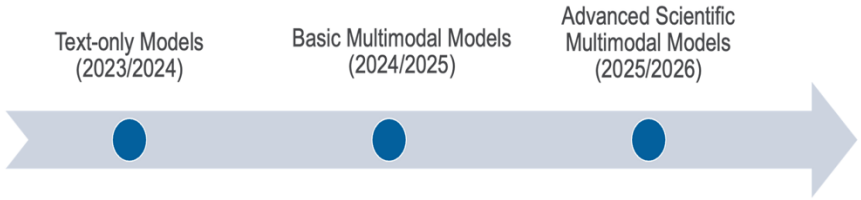
Foundation models



- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

AURORAGPT*:

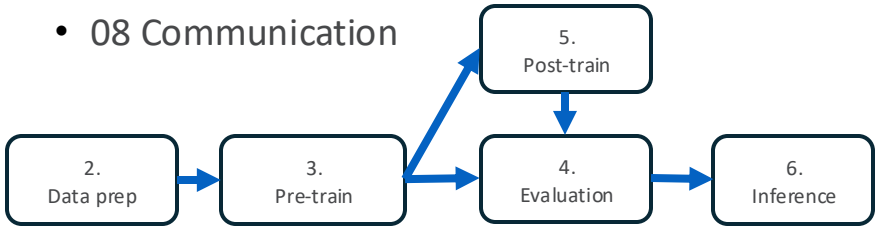
- **EXPLORE PATHWAYS** TOWARDS A SCIENTIFIC FOUNDATION MODEL
- **GENERAL PURPOSE SCIENTIFIC LLM – BROADLY TRAINED – GENERAL CORPORA PLUS SCIENTIFIC PAPERS AND TEXTS AND STRUCTURE SCIENCE DATA**
- **MULTIMODAL – IMAGES, TABLES, EQUATIONS, PROOFS, TIME-SERIES, GRAPHS, FIELDS, SEQUENCES, ETC.**
- **SAFE: TRUSTWORTHINESS, SAFETY, SECURITY, ROBUSTNESS, PRIVACY, MACHINE ETHICS**
- **BUILD WITH INTERNATIONAL PARTNERS** (RIKEN, BSC, OTHERS)
- **MULTILINGUAL** – ENGLISH, 日本語, FRANÇAIS, DEUTSCHE, ESPAÑOL, ITALIANA



Aurora is: 166 Racks
10,624 Nodes
21,248 CPUs , 63,744 GPUs
8 PB HBM
10 PB DDR5c

Groups:

- 01 Planning
- 02 Data
- 03 Model training (pre-training)
- **04 Evaluation (skills, trustworthiness, safety)**
- 05 Post-training (fine tuning, alignment)
- 06 Inference
- 07 Distribution
- 08 Communication



*named after the Leadership Class Supercomputer at Argonne that will be used for much of the training

AURORAGPT LEADERS

PLANNING

RICK STEVENS (LEAD)

IAN FOSTER

RINKU GUPTA (PM)

MIKE PAPKA

ARVIND RAMANATHAN

FANGFANG XIA

DISTRIBUTION

BRAD ULLRICH

DATA

IAN FOSTER

ROBERT UNDERWOOD

MODELS

VENKAT VISHWANATH

SAM FOREMAN

EVALUATION AND SAFETY

FRANCK CAPPELLO

SANDEEP MADIREDDY

BO LI

POST-PRE TRAINING

ELIU HUERTA

AZTON WELLS

INFERENCE

RAJEEV THAKUR

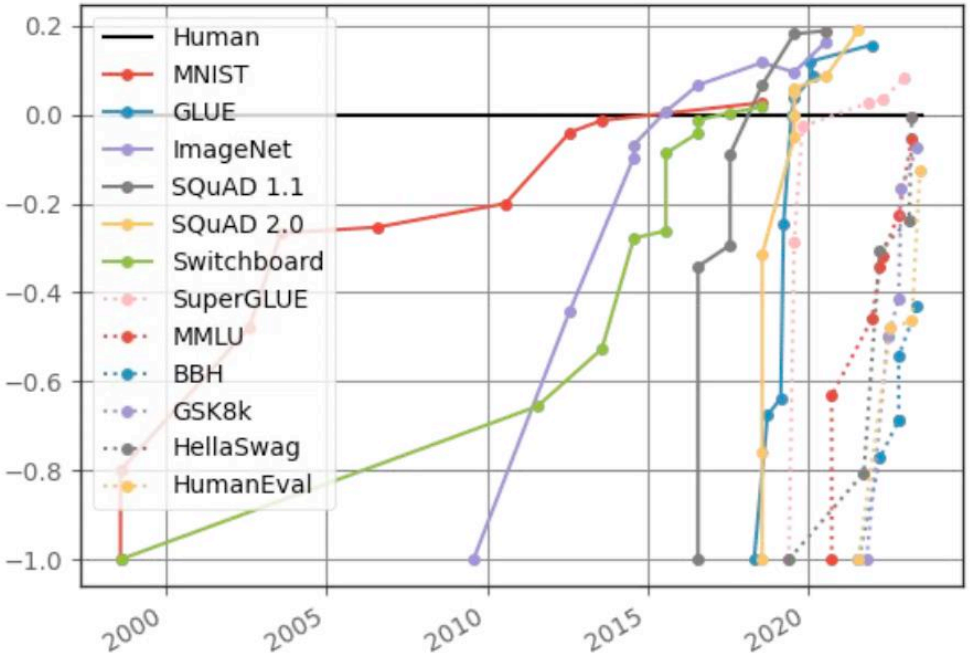
COMMUNICATIONS

DAVID MARTIN

CHARLIE CATLETT

WHY EVALUATE A LANGUAGE MODEL?

- **Tracking progress**
  - Are models getting more capable at science tasks?
- **Quantitative measures**
  - We need to objectively, reproducibly measure improvements



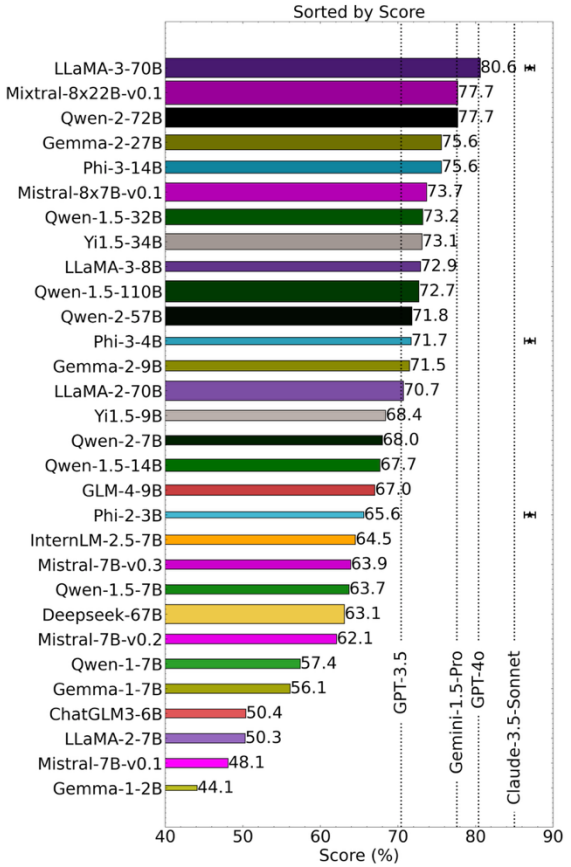
Kiela et al. (2023). Plotting Progress in AI

WHY EVALUATE A LANGUAGE MODEL?

- **Making Comparisons**
  - Is method X better than the baseline method Y?
  - In what situations is X better?
  - Which model should I use for my task?

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0

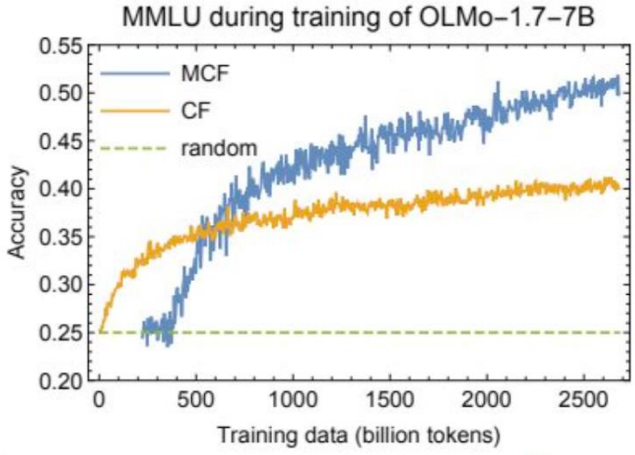
Meta AI (2024). Introducing Meta Llama 3: The most capable openly available LLM to date



Ting et al. (2024). Who Wins Astronomy Jeopardy

WHY EVALUATE A LANGUAGE MODEL?

- **Assess training runs**
  - Sanity-check training — are we improving as we train?
  - compare ablations — are the new techniques we try improving things compared to a baseline?
  - And more, ...
- **Prevent regressions**
  - During fine-tuning — as we specialize a model, does degrade too much on general tasks?
  - During model compression — as we make smaller versions of a model to accommodate an edge device like a phone/field sensor, can it still do its task?
  - And more, ...

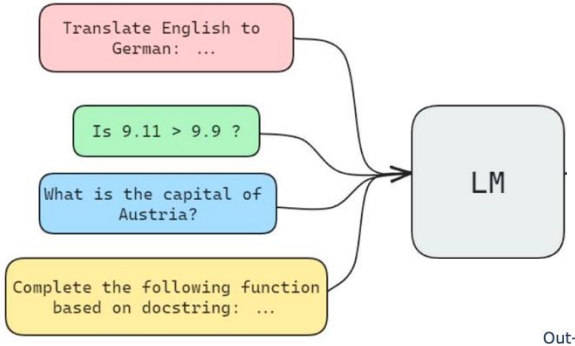


Gu et al. (2024). OLMES: A Standard for Language Model Evaluations.

ICML Tutorial 2024 - Challenges in LM Evaluation

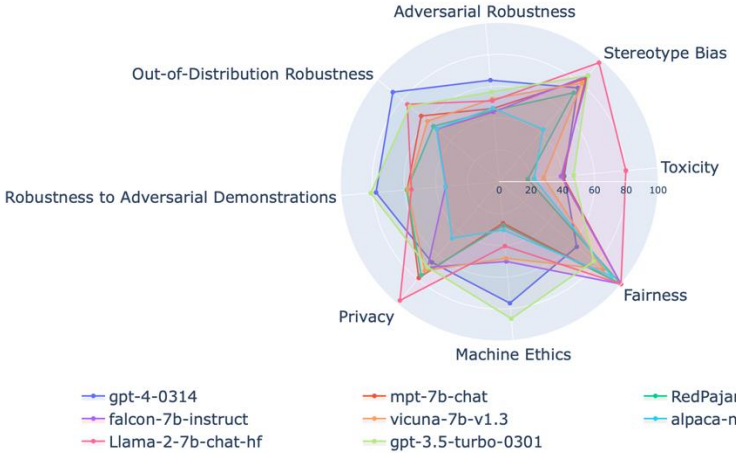
WHAT DO WE WANT TO EVALUATE?

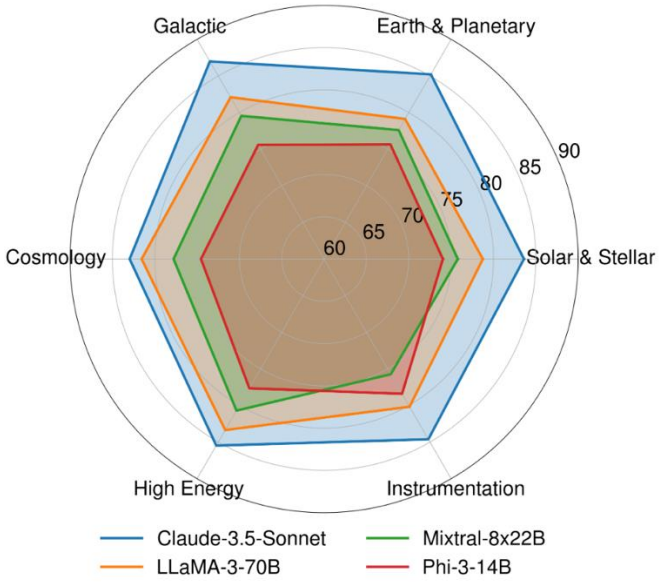
Skills: Scientists need many skills to do their jobs



- Arithmetic
- Translation
- Factual question
- Code completion

Adversarial





Ting et al. (2024). Who Wins Astronomy Jeopardy

Scientific Tasks

https://decodingtrust.github.io/demo/

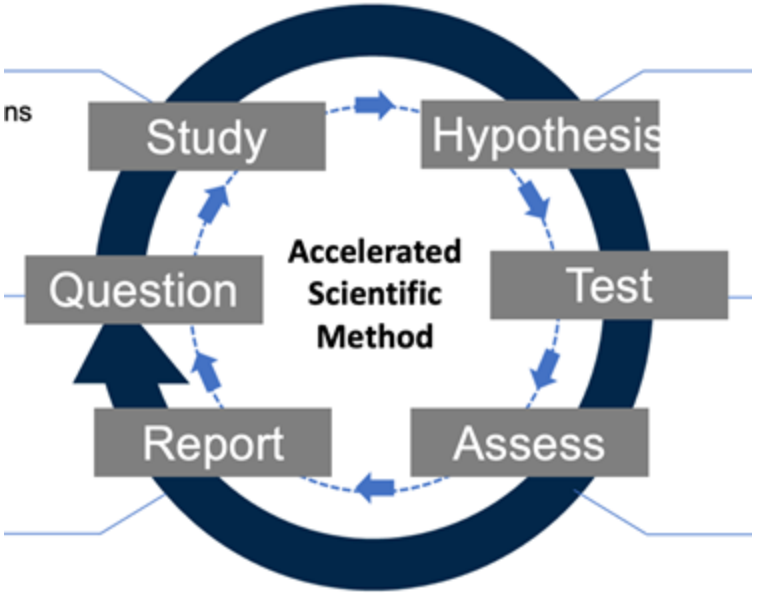
Big Goal: LLMs as Research Assistants

Scientists assessed LLMs on **specific tasks**:

- Predicting molecular properties
- Uncovering genomic patterns
- Interpreting astrophysical data
- Solving mathematical problems
- Creating and manipulating tools for simulations and analysis

→Growing multi-step reasoning skills **Suggest a new holistic approach** where LLMs/LRMs are use as **scientific research assistants**

https://doi.org/10.1038/s41524-022-00765-z



Characteristics of an “AI scientific assistant” that we need to/must evaluate

An AI-based system with:

- Scientific skills
  - Reasoning, math, literature understanding, integrity
- Effective assistance (no hallucination!, consistency in responses)
  - **Correct** for all different tasks related to scientific activities
- Relevance to human and environment interaction modalities (communication skills)
  - Understanding command (semantic of it), interface with tools and devices
- Degree of autonomy
  - From repeating learned workflows to developing the workflow.
  - **Capable of hypothesis generation**
- **Safety for** the community
  - Cannot be used to harm others: e.g. design harmful substances

Benchmarks: MCQs and Open Responses

- **Multi-Choice Questions (MCQs)**
  - 1 correct response and 3, 4 or more distractors (wrong responses)
  - Difficulty
    - Gene… ose to the corre…
    - Evalu…
  - Potential…
    - E.g. … ntly
- **Open Respon…**
  - 1 questio…
  - Difficulty
    - Gene…
    - Evaluation is difficult: Require a human evaluation of the response (→ LLM as judge), UQ
      → Does not scale well (→ LLM as judge)
  - Potential biases:
    - Room for interpretation: Human may score differently the same open response → scoring requires several human evaluation (consensus)

- MCQ/Open Response Benchmarks are great to assess model knowledge and reasoning capabilities
- But existing ones are too generic
- Static benchmarks saturate quickly
- They cannot be used for end-to-end Eval

→ We cannot only rely on benchmarking

EAIRA: Multi-faceted Eval Methodology

Benchmarks

End-to-End

Proposed Methodology				
Techniques	**MCQ Benchmarks**	**Open Response Benchmarks**	**Lab Style Experiments**	**Field Style Experiments**
Main Goal	Testing knowledge **breadth, basic reasoning**	Testing knowledge **depth, planning, reasoning**	**Realistic** testing	**Realistic trend** analysis and weakness diagnosis
Problem Type	**Predetermined**, Fixed Q&As with known solutions	**Predetermined**, Fixed Free-Response Problems with known solutions	**Individual Human** Defined Problems with **unknown** solutions	**Many Human** Defined Problems with **(un)known** solutions
Verification	**Automatic** response verification	**Automatic or Human** response verification	**Humans detailed** response analysis	Scalable **automatic** summary of **human response**
Examples	**Astro, Climate, AI4S** (multi-domain), **Existing Benchmarks**	**SciCode, ALDbench**	**see "lab style experiments"**	**see "field style experiments"**
Cross Cutting Aspects	← **Trust and Safety (ChemRisk), Uncertainty Quantification, Scalable Software Infrastructure (STAR)** →			

Methodology consisting of **4 complementary evaluation techniques** to comprehensively assess the capabilities of LLMs as scientific assistants:

- purple text shows prior contributions by the researchers participating in AuroraGPT
- blue text shows AuroraGPT contributions.
- Black text aspects adapted from existing work are included for a complete approach.

https://arxiv.org/abs/2502.20309

Argonne NATIONAL LABORATORY

ASTRO MCQ Benchmark

- **4425 Automatically generated MCQs**
- From 885 articles in Annual Review of Astronomy and Astrophysics, 1963 to 2023.
- Instructed Gemini-1.5-Pro to propose 5 questions that can be answered based on the paper's content.
- Each question was accompanied by four options (A, B, C, D) only one of which is correct.
- Robustness considerations added to the prompt generating the questions.
- 200 MCQs were manually validated

Sample question from Astronomy benchmark dataset

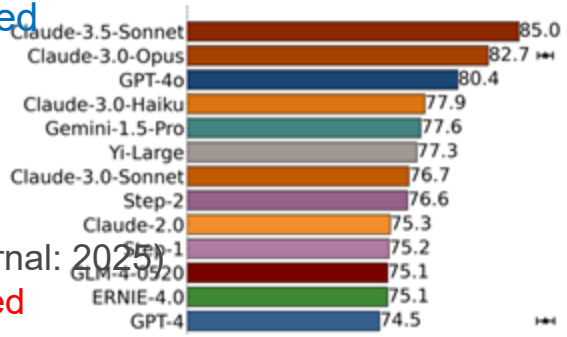
How does the presence of stellar companions influence the formation and detection of exoplanets?

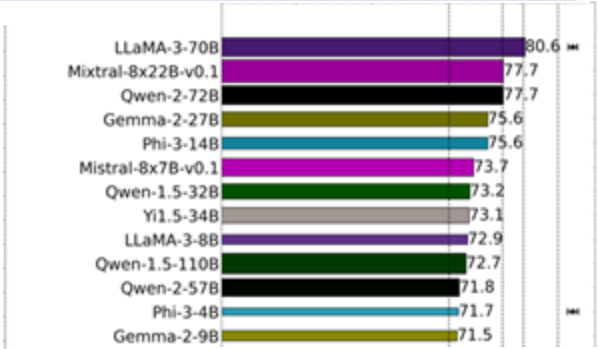
(A) Stellar companions can dilute transit signals, potentially leading to misclassification of planets and inaccurate parameter estimations. Additionally, their gravitational influence can suppress planet formation in close binary systems.

(B) Stellar companions provide additional sources of gravitational perturbations, enhancing planet formation by promoting planetesimal accretion and facilitating the formation of gas giants.

(C) Stellar companions contribute to the metallicity enrichment of planetary systems, leading to the formation of more massive and diverse planets, including super-Earths and hot Jupiters.

(D) Stellar companions act as gravitational lenses, increasing the detectability of exoplanets through microlensing events and enabling the discovery of planets at greater distances from their host stars.





Some take aways:

- Claude 3.5 Sonnet best (no O1 test)
- Llama-3-70B on par with GPT4o
- Published in July 2024 on arXiv (journal: 2025)
- Benchmark almost/probably saturated

Y.-S. Ting, et al.i, AstroMLab 1: Who wins astronomy jeopardy!? Astronomy and Computing, Volume 51, 2025.

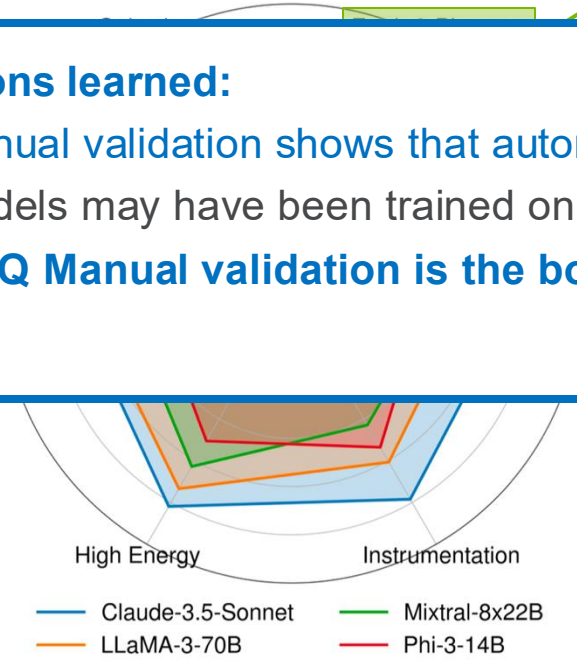
ASTRO BENCHMARK

Sub-areas in astrophysics

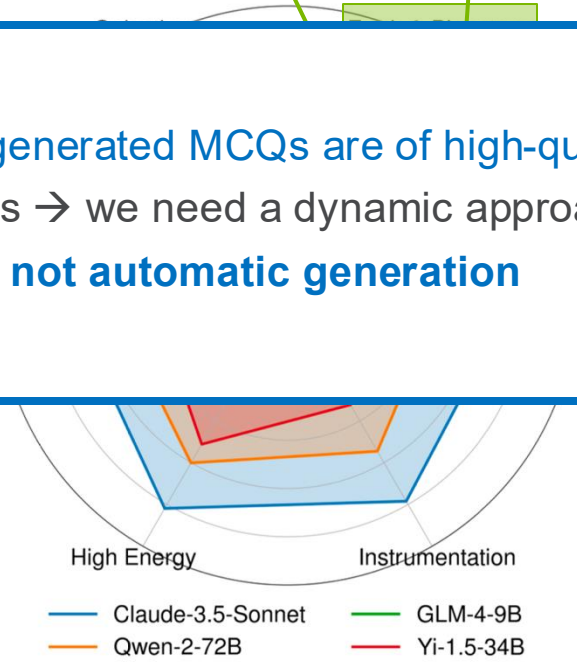
performance degradation in more recent topics

Lessons learned:

- Manual validation shows that automatically generated MCQs are of high-quality
- Models may have been trained on the papers → we need a dynamic approach
- **MCQ Manual validation is the bottleneck! not automatic generation**



English-focused models

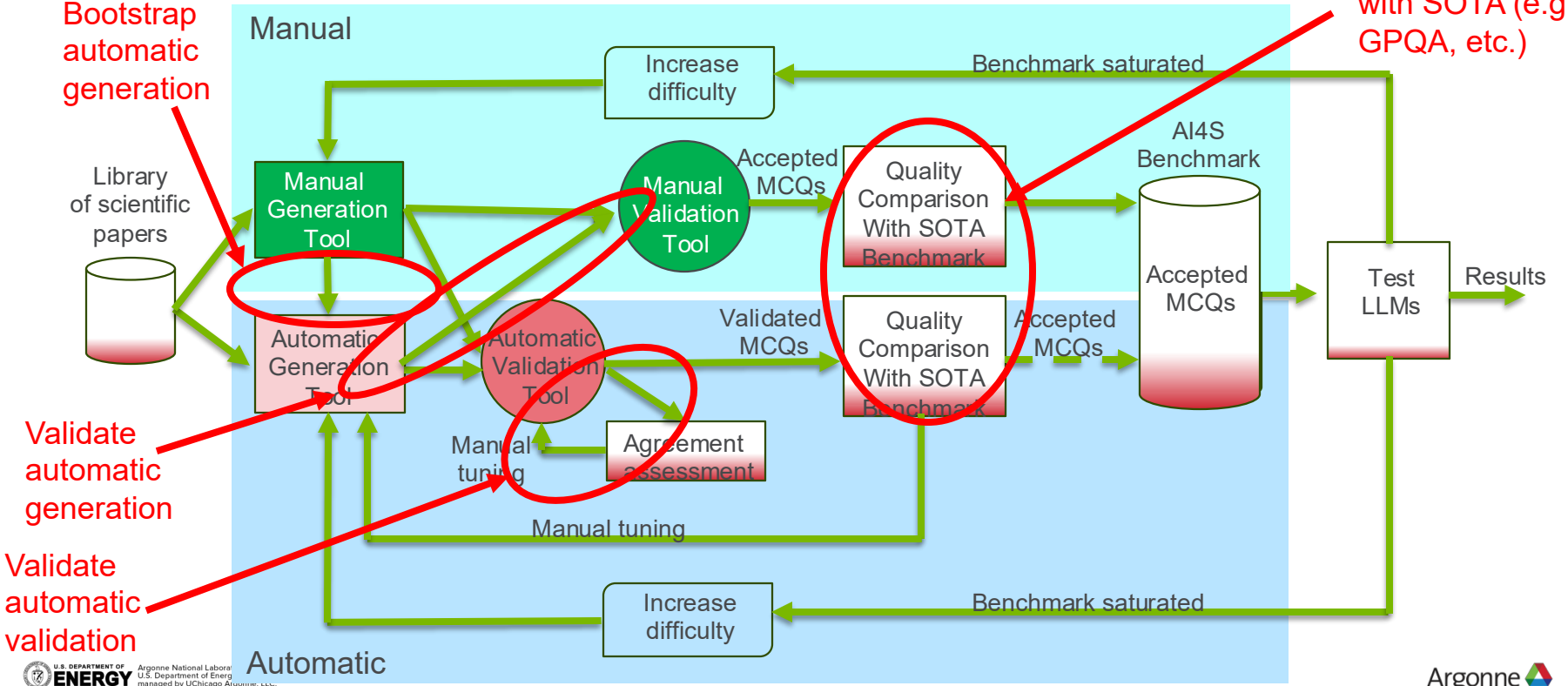


Non-English-focused models

AUTOMATIC HIGH-QUALITY BENCHMARK GENERATION/VALIDATION

Many scientists have the same need: generate specific MCQ benchmarks for their problems

→ We need an integrated framework to generate/validate MCQs Benchmarks

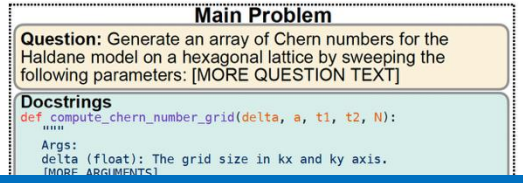


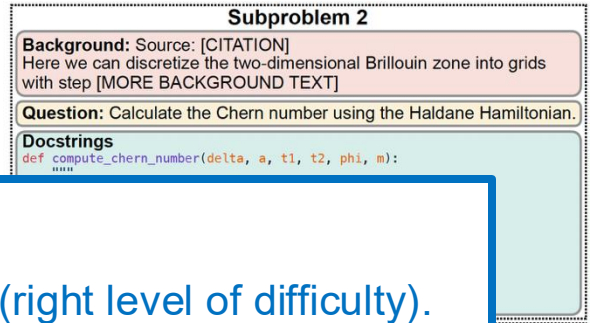
SCICODE Open Response Benchmark (integrated into the methodology)

Scientist-curated code generation benchmark (mathematics, physics, chemistry, biology, materials science)

80 main problems (numerical methods, simulation of systems), decomposed into 338 subproblems.

The problems naturally factorize into multiple subproblems, each involving knowledge recall, reasoning, code synthesis.





To so
imple
each
comp

SciCo
multip
evalu

Probl
Nobel price level problems.

Lesson learned:

- OpenAI 01-preview can only solve 7.7% of main problems (right level of difficulty).
- Difficulty comes from the necessity to combine of multiple skills:

  problem understanding, retrieval, reasoning, planning, code generation.
- Using codes as the results of the questions makes verification "trivial" **but it is not applicable to all open question problems: e.g. bio**

Minyang Tian, SciCode: A Research Coding Benchmark Curated by Scientists, arXiv: arXiv:2407.13168

Understanding/modeling question difficulty

ANL-HPE COLLABORATION: DOREMI: DIFFICULTY-ORIENTED REASONING EFFORT MODELLING OF SCIENCE PROBLEMS FOR REASONING LANGUAGE MODELS

- Current benchmarks fail to characterize why problems are difficult for reasoning LLMs - they fold diverse challenges into single accuracy scores. → How do we know if a benchmark question is difficult?
- It remains unclear what level of reasoning effort to is required across benchmarks.

→ Need principled ways to 1) measure difficulty for curriculum learning, 2) benchmark creation, and 3) reasoning effort estimation.

DoReMi

- Compute **Multi-dimensional Difficulty Fingerprints** for a benchmark using **Bloom Taxonomy metrics** across 7 dimensions
- **Use LLM as a judge** approach to evaluate questions on the Bloom dimensions
- **Use Multiple LLM Judges** and check consensus.

- **Study correlations** between LLM judges difficulty assessments and some metrics of LLM perceived difficulty to respond to a question.

→ Link difficulty to cost (time, tokens, etc.)

- Consider multiple metrics:
  - Wrong Answer Fraction (WAF)
  - Minimum Reasoning Token (MRT)
  - Expected Runs to First Correct Answer (R2FCA)
  - Uncertainty of Correct Answers (UCA)
  - Reasoning Inconsistency (RI):
  - Etc.

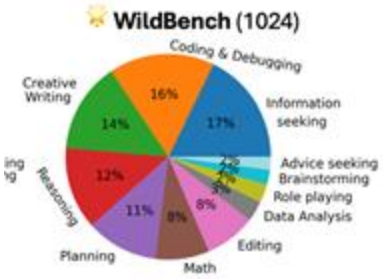
DoReMi: Paper submitted to an AI conference

End-to-End Eval: FIELD STYLE EXPERIMENT

Lab style experiments: **Human evaluation**, tries to solve 1 specific problem, compare different models, guide LLMs (requires efforts: some prompt engineering),

Field style experiments: **Automatic evaluation**, capture what researchers actually ask, much broader diversity of Q&As, large diversity of prompt engineering, statistical evaluation

Several papers on this topic (but not for Science activity)



- **WildBench**: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild, B. Y. Lin and Y. Deng and K. Chandu and F Brahman and A.Ravichander and V. Pyatkin and N. Dziri and R. Le Bras and Y. Choi, 2024, arXiv 2406.04770
- **HaluEval-Wild**: Evaluating Hallucinations of Language Models in the Wild, Zhiying Zhu and Yiming Yang and Zhiqing Sun, 2024, arXiv, 2403.04307
- **"Do Anything Now":** Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models Zeyuan Chen and Michael Backes and Yun Shen and Yang Zhang, 2024, arXiv 2308.03825

End-to-End Eval: ~~1000~~ 1500 SCIENTISTS JAM IN 9 LABS SIMULTANEOUSLY (FEB.28, 2025)

Researcher participation and contributions on a voluntary basis.

1,000 Scientists Jam Session: In numbers

Researcher participation and contributions on a voluntary basis.

Total:
2800+ problems
15000+ assessed prompt responses

Argonne:
720 problems
2500 prompts

1,000 Scientists Jam Session: Domains

Researcher participation and contributions on a voluntary basis.

Literature/Data

- Literature search, analysis, survey
- Data analysis and forecast, interpolation, extrapolation, classification (Point Cloud, signal, protein sequences, files, etc.)
- Anomaly detection
- Signal Analysis
- Scientific Visualization

Coding

- Algorithm design/optimization
- Automatic code generation/refactoring
- Code translation
- Debugging codes (sequential, parallel)
- Automatic code performance tuning/optimization
- Identifying performance bottlenecks

Experiments

- Automatic tuning of instruments
- Experimental Design (including autonomous workflow)
- Dark mater experiment design

Bio

- Understanding mechanisms of Cancer
- Understanding radiation effects on human cells
- Predictive Genomic Models

AI

- Domain specific LLMs/Agents (use LLMs as foundation models)
- Hyper parameter exploration for DL training.

Physics

- Battery design
- Chemical Mechanisms
- Physics beyond standard model

Infra.

- Infrastructure modeling and resilience
- Natural Disaster assessment

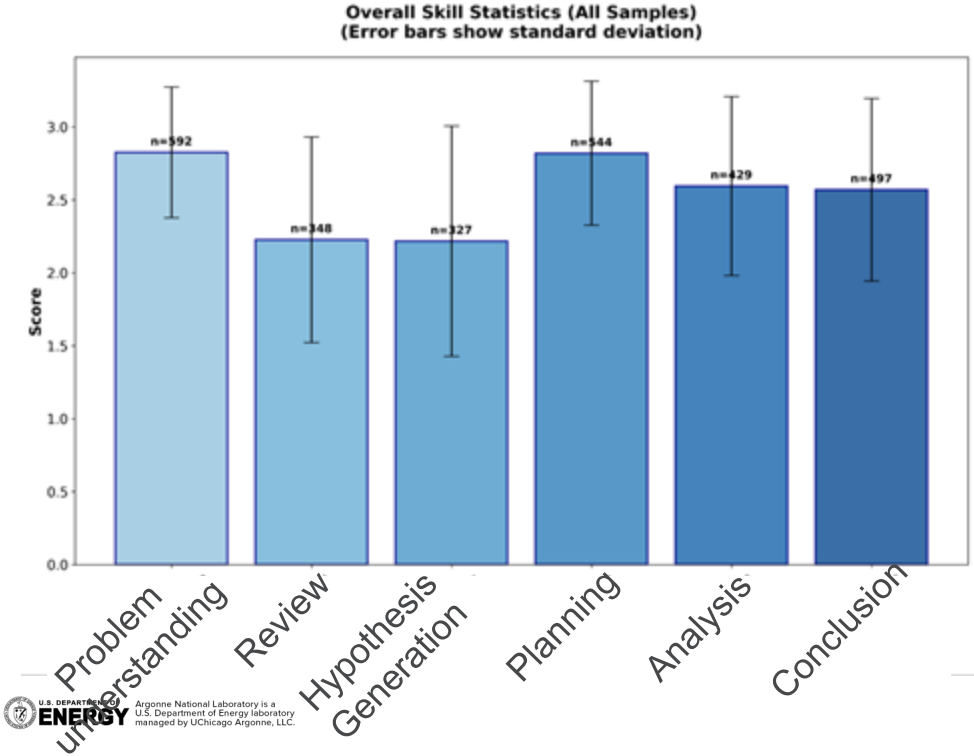
Math

- Surrogate model
- Mathematical derivations
- PDE solving
- Convergence proving
- Equation validity testing
- Derivative analysis
- Uncertainty estimation
- Inverse problems
- Statistical modeling

Argonne NATIONAL LABORATORY

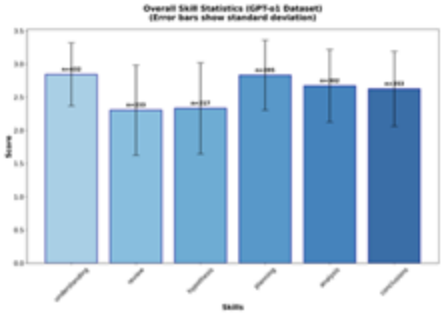
1,000 SCIENTISTS JAM SESSION: SKILLS STRENGTH (AVERAGE OVER THE WHOLE CORPUS)

LLM as a judge to automatically score (1-5) the LLMs responses



Different blue colors represent different skills

Result robust against change the judge model (gpt 4o -> gpt o1)

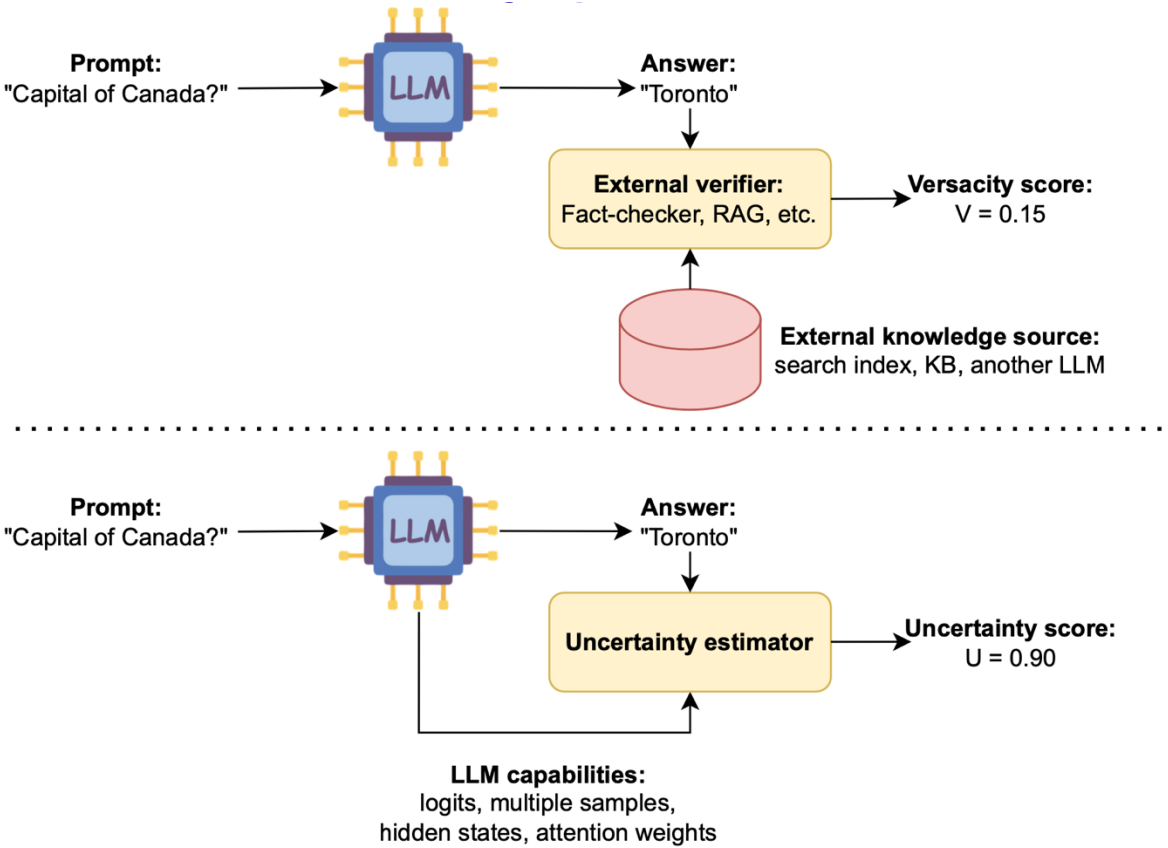


WHY DO WE NEED UNCERTAINTY ESTIMATES?

Reliable estimates of **uncertainty** can help us:

- **Build or reduce trust** in certain pointwise predictions…
- **Compare** the performance of different models (i.e., uncertainty in metrics)...
- **Identify areas of improvement** for a given model (e.g., for active learning)...
- **List all plausible answers** subject to specified probabilistic guarantees…
- **Produce more natural responses** (that reflect confidence) for dialogue agents…
- **Abstain** from making predictions when in doubt…
- **Hallucination detection** in LLM generations
- Adversarial attack detection
- Reinforcement learning / control theory
- (Emerging) Improving performance of multi-step reasoning systems

WHY DO WE NEED UNCERTAINTY ESTIMATES?

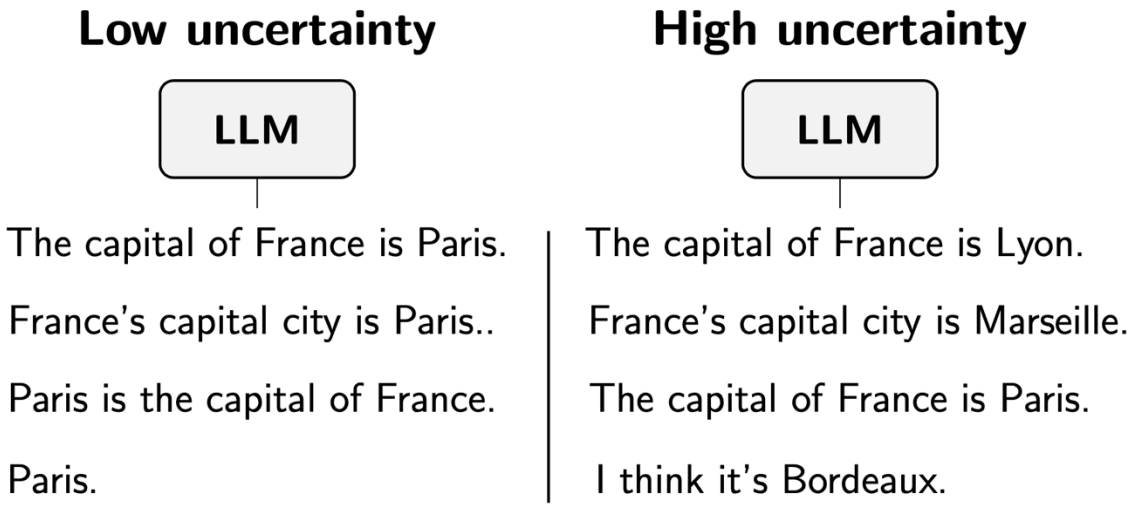


CLASSES OF UQ APPROACHES FOR LLMS

- **Black-box methods**
  - **Verbalized uncertainty**
    - Directly asking the model about its confidence in a generated answer
  - **Consistency-based**
    - Sample multiple generations and measure their (semantic) consistency.
- **White-box methods**
  - **Information-theoretic**
    - Assess uncertainty as measured by probabilities given by the model
  - **Introspective**
    - Analyze model embeddings and/or attention masks

CONSISTENCY-BASED UNCERTAINTY

Intuition: diverse responses to the same prompt indicate high uncertainty.



CONSISTENCY-BASED UNCERTAINTY

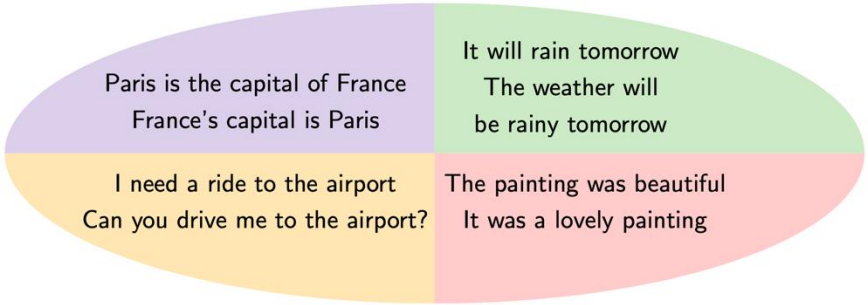
Semantic Entropy

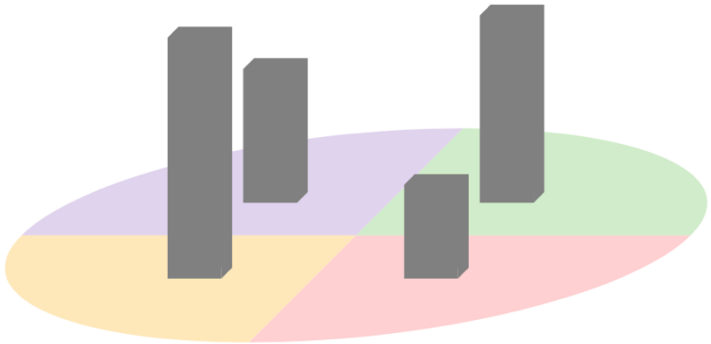
- Entropy over semantic clusters.
- Let be semantic clusters from Number of Semantic Sets partition.

$$\mathcal{C} = \{\mathbf{y} : \forall \mathbf{y}' \in \mathcal{C}, \mathrm{NLI}(\mathbf{y}, \mathbf{y}') = \mathrm{NLI}(\mathbf{y}', \mathbf{y}) = \textit{entail}\}.$$

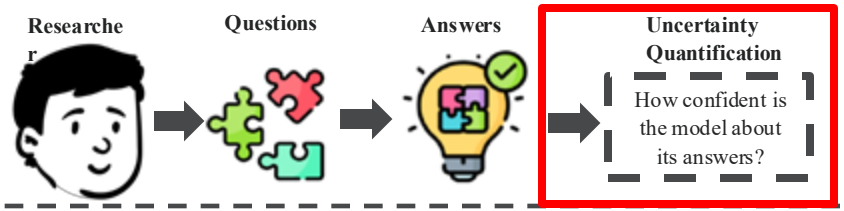
$$\mathrm{U_{SE}} = -\frac{1}{N} \sum_{m=1}^{M} |\mathcal{C}_m| \log \hat{P}_m(\mathbf{x})$$

$$\hat{P}_m(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{C}_m} P(\mathbf{y} \mid \mathbf{x})$$





CHEMICAL REACTION PREDICTION



Another Example: UPSTO dataset

reactants_smiles	products_smiles
C1CCOC1.CC(C)C[Mg+].CON(C)C(=O)c1ccc(O)nc1.[Cl-]	CC(C)CC(=O)c1ccc(O)nc1
CN.O.O=C(O)c1ccc(Cl)c([N+](=O)[O-])c1	CNc1ccc(C(=O)O)cc1[N+](=O)[O-]
CCn1cc(C(=O)O)c(=O)c2cc(F)c(-c3ccc(N)cc3)cc21.O=CO	CCn1cc(C(=O)O)c(=O)c2cc(F)c(-c3ccc(NC=O)cc3)cc21
CC(C)=C(Cl)N(C)C.COCC(C)Oc1cc(Oc2cnc(C(=O)N3CCC3)cn2)cc(C(=O)O)c1.Cc1cnc(N)cn1.ClCCl.c1ccncc1	COCC(C)Oc1cc(Oc2cnc(C(=O)N3CCC3)cn2)cc(C(=O)Nc2cnc(C)cn2)c1
Clc1cc2c(Cl)nc(-c3ccncc3)nc2s1.NCc1ccc(Cl)c(Cl)c1	Clc1cc2c(NCc3ccc(Cl)c(Cl)c3)nc(-c3ccncc3)nc2s1

General prompt: Given the smiles representation of the reactant and reagents, please predict the product and output in smiles representation…….
A few examples are given below:
Reactant and reagents:
C1CCOC1.CC(C)C[Mg+].CON(C)C(=O)c1ccc(O)nc1.[Cl-]
Products:
CC(C)CC(=O)c1ccc(O)nc1
……
Reactant and reagents:
Clc1cc2c(Cl)nc(-c3ccncc3)nc2s1.NCc1ccc(Cl)c(Cl)c1
Product:
?

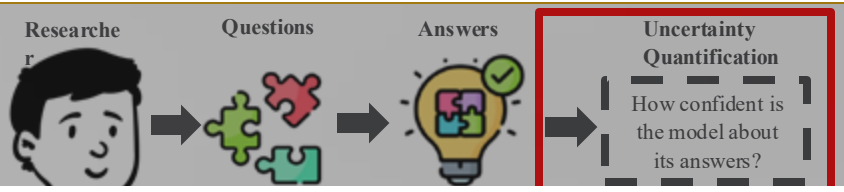
GPT 4

Predicted
Product 1
Product 2
Product 3
….

Similarity measurement

Generate test results & conduct UQ

CHEMICAL REACTION PREDICTION



Method	Top-1 Acc.	AUC-3	AUC-10	AUC-15	AUC-20
GPT-4 + Orig.	0.250	0.864	0.919	0.915	0.927
GPT-4 + Reform	0.070 ↓	0.972	0.941	0.958	0.993
GPT-3.5 + Orig	0.186	0.904	0.899	0.924	0.943
GPT-3.5 + Reform	0.036 ↓	0.919	1.000	1.000	1.000

Products:
CC(C)CC(=O)c1ccc(O)nc1
.....
Reactant and reagents:
Clc1cc2c(Cl)nc(-c3ccncc3)nc2s1.NCc1ccc(Cl)c(Cl)c1
Product:
?

4

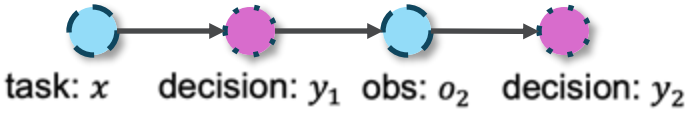
Product 2
Product 3
....

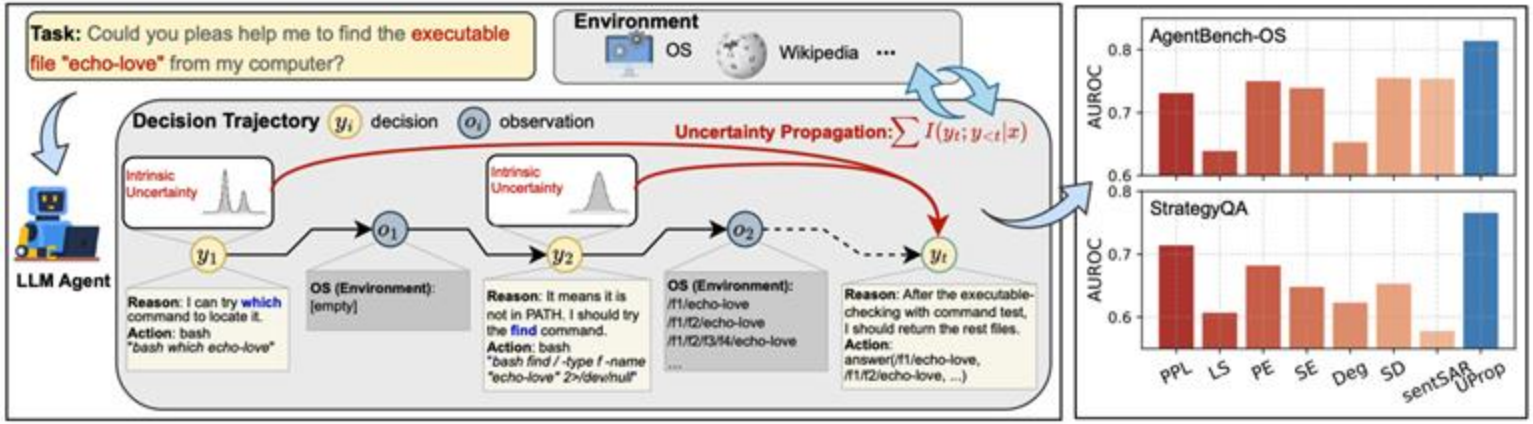
Similarity
measurement

Generate test results & conduct
UQ

UNCERTAINTY PROPAGATION OF LLM MULTI-STEP DECISION-MAKING

Core Research Question: How should we propagate uncertainty in LLM decision-making chain?

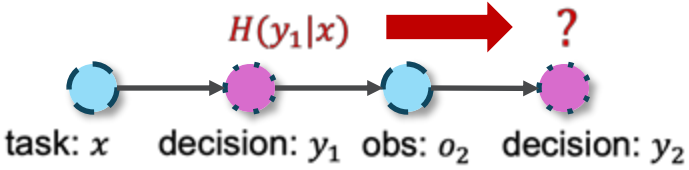




Duan, J., Diffenderfer, J., Madireddy, S., Chen, T., Kailkhura, B., & Xu, K. (2025). UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making. *arXiv preprint arXiv:2506.17419*.

UNCERTAINTY PROPAGATION OF LLM MULTI-STEP DECISION-MAKING

Core Research Question: How should we propagate uncertainty in LLM decision-making chain?



Predictive uncertainty regarding decision y_2

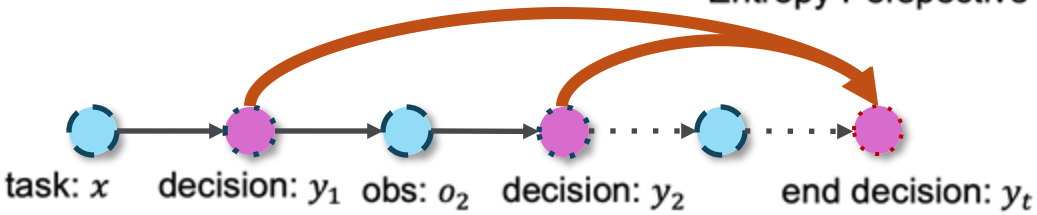
$$p(y_2|x) = \int p(y_2|y_1,x)p(y_1|x)dy_1$$

external uncertainty inherited from y_1

internal uncertainty conditioned on y_1

Entropy Perspective $H(y_2|x) = H(y_2|y_1,x) + H(y_1|x) - H(y_1|y_2,x)$

$$= \boldsymbol{H(y_2|y_1,x)} + \boldsymbol{I(y_1;y_2|x)}$$



$$H(y_t|x) = \boldsymbol{H(y_t|y_{t-1},y_{t-2},\cdots,x)} + \boldsymbol{I(y_{t-1};y_t|x)} + \boldsymbol{I(y_{t-2};y_t|y_{t-1},x)} + \cdots + \boldsymbol{I(y_1;y_t|y_{t-1},y_{t-2},\cdots,x)}$$

$$H(y_t|x) = H(y_t|y_{1:t-1},x) + \sum_i^{t-1}(H(y_t|x) - H(y_t|y_i,x))$$

$$= \underbrace{H(y_t|y_{1:t-1},x)}_{\text{Intrinsic Uncertainty}} + \underbrace{\sum_i^{t-1} I(y_t;y_i|y_{i+1:t-1},x)}_{\text{Extrinsic Uncertainty}},$$

Duan, J., Diffenderfer, J., Madireddy, S., Chen, T., Kailkhura, B., & Xu, K. (2025). UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making. *arXiv preprint arXiv:2506.17419*.

RESULTS

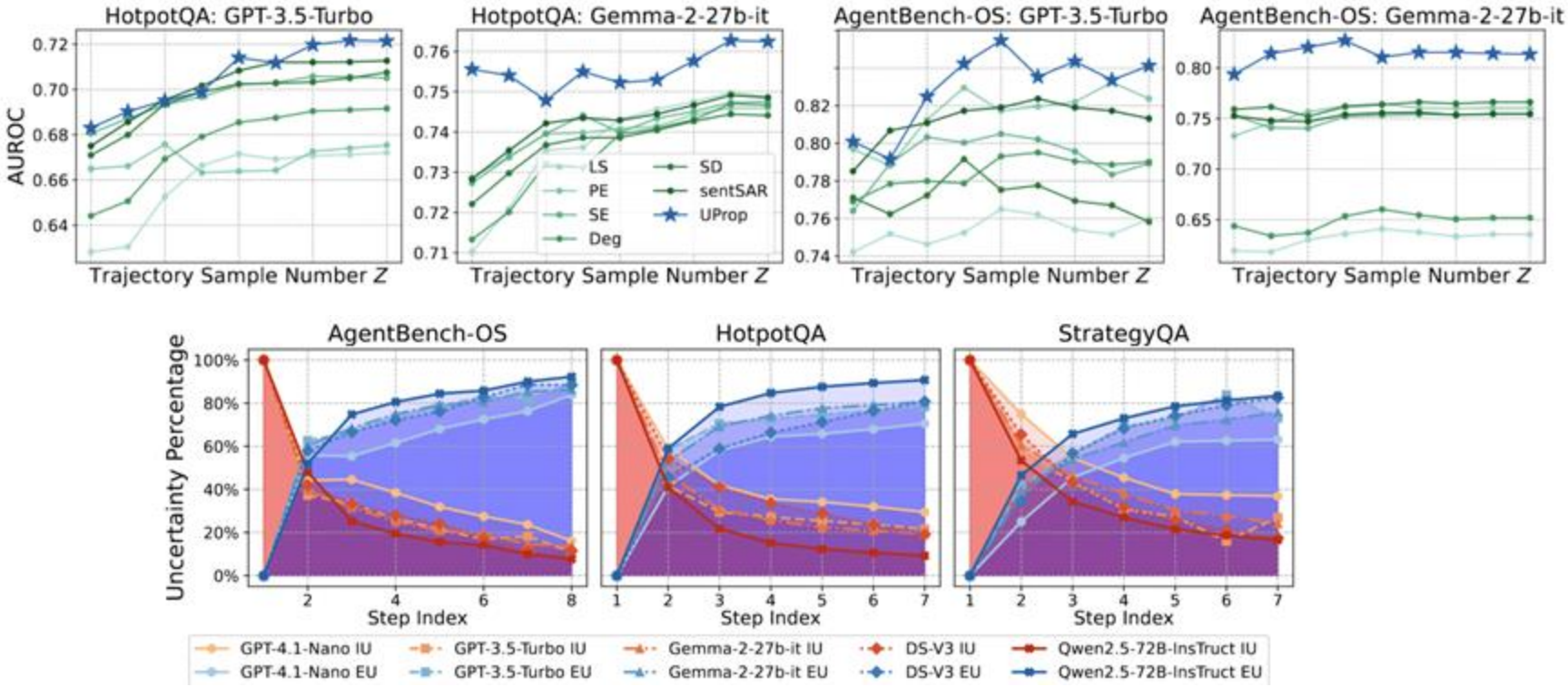
Table 1: AUROC results over AgentBench-Operating System and StrategyQA benchmarks. For single-turn baseline UQ methods, uncertainties are aggregated by ***averaging*** over all steps.

Models	Success Rate	PPL	LS	PE	SE	Deg	SD	sentSAR	UProp (ours)
Benchmark: AgentBench (Operating System)									
GPT-4.1-Nano	0.307	0.725	0.756	0.768	0.770	0.757	0.779	0.775	**0.781**
GPT-3.5-Turbo	0.275	0.747	0.750	0.782	0.765	0.765	0.749	0.777	**0.791**
Gemma-2-27b-it	0.289	0.747	0.636	0.760	0.755	0.652	0.766	0.755	**0.814**
DeepSeek-V3	0.310	0.729	0.636	0.724	0.716	0.655	0.717	0.722	**0.767**
Qwen2.5-72B-Instruct	0.508	0.625	0.620	**0.707**	0.687	0.631	0.678	0.678	0.704
Average	0.338	0.715	0.679	0.748	0.738	0.692	0.738	0.741	**0.771**
Benchmark: StrategyQA									
GPT-4.1-Nano	0.691	0.512	0.492	0.542	0.503	0.502	0.499	0.527	**0.544**
GPT-3.5-Turbo	0.611	0.593	0.438	0.623	**0.611**	0.440	0.600	0.607	0.604
Gemma-2-27b-it	0.777	0.698	0.615	0.669	0.624	0.622	0.640	0.667	**0.766**
DeepSeek-V3	0.790	0.573	0.548	0.559	0.558	0.575	0.574	0.563	**0.607**
Qwen2.5-72B-Instruct	0.796	0.500	0.495	0.573	0.573	0.493	0.567	0.563	**0.617**
Average	0.733	0.575	0.518	0.593	0.574	0.526	0.576	0.585	**0.628**

Success Rate: fraction of episodes in which the agent actually solved the benchmark problem (returned the right shell state for AgentBench-OS, or the correct Yes/No answer for StrategyQA).

AUROC (0.5 = random guessing; 1.0 = perfect separation)
asks "When the agent says it is confident, is it actually more likely to be right?"

Duan, J., Diffenderfer, J., Madireddy, S., Chen, T., Kailkhura, B., & Xu, K. (2025). UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making. *arXiv preprint arXiv:2506.17419*.



Duan, J., Diffenderfer, J., Madireddy, S., Chen, T., Kailkhura, B., & Xu, K. (2025). UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making. *arXiv preprint arXiv:2506.17419*.

CONCERNS ON AI SAFETY AND ALIGNMENT

The New York Times

Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots

A new report indicates that the guardrails for widely used chatbots can be thwarted, leading to an increasingly unpredictable environment for the technology.

FORTUNE

Your favorite A.I. language tool is toxic

protocol

OpenAI's new language AI improves on GPT-3, but still lies and stereotypes

Research company OpenAI says this year's language model is less toxic than GPT-3. But the new default, InstructGPT, still has tendencies to make discriminatory comments and generate false information.

MIT Technology Review

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

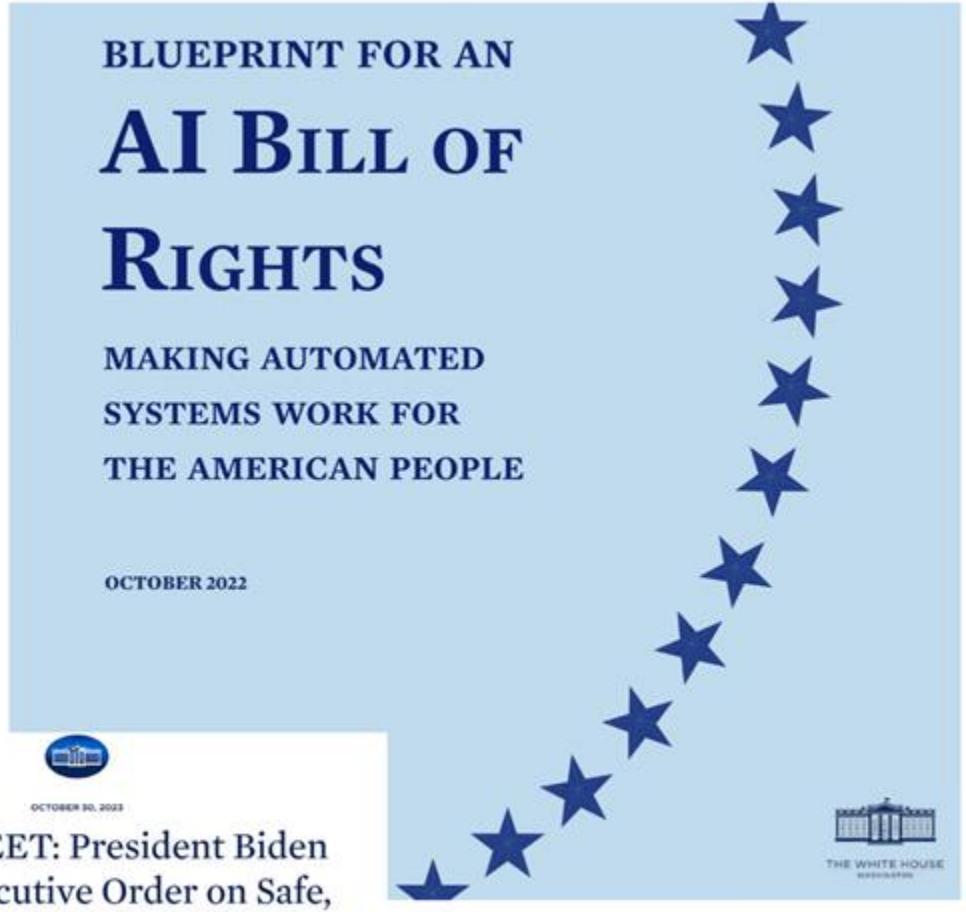
Samsung workers made a major error by using ChatGPT

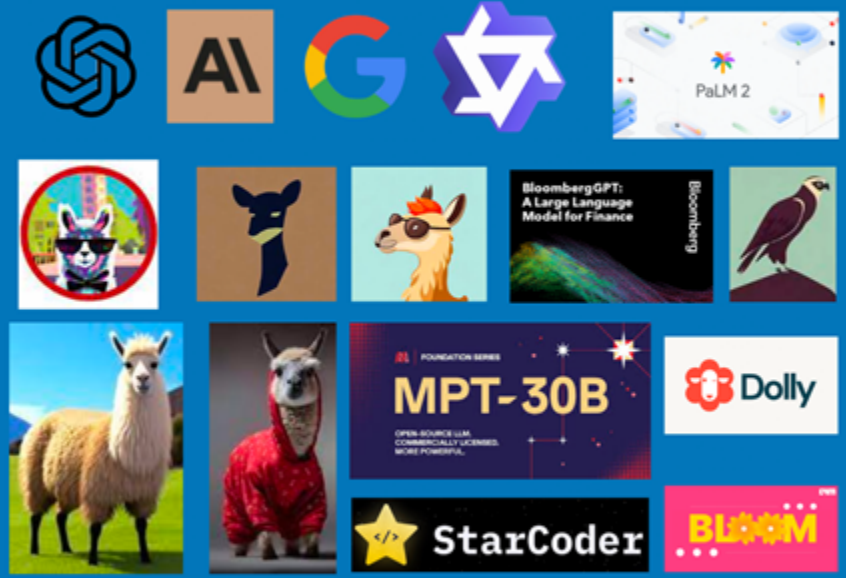
News By Lewis Maddison published April 04, 2023

Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT

Trustworthiness problems in AI

- Robustness: Safe and Effective Systems
- Fairness: Algorithmic Discrimination Protections
- Data Privacy
- Notice and Explanation
- Human Alternatives, Consideration, and Fallback





Enhance the trustworthiness of LLMs for enterprises after helping identify the model vulnerabilities

July 21, 2023

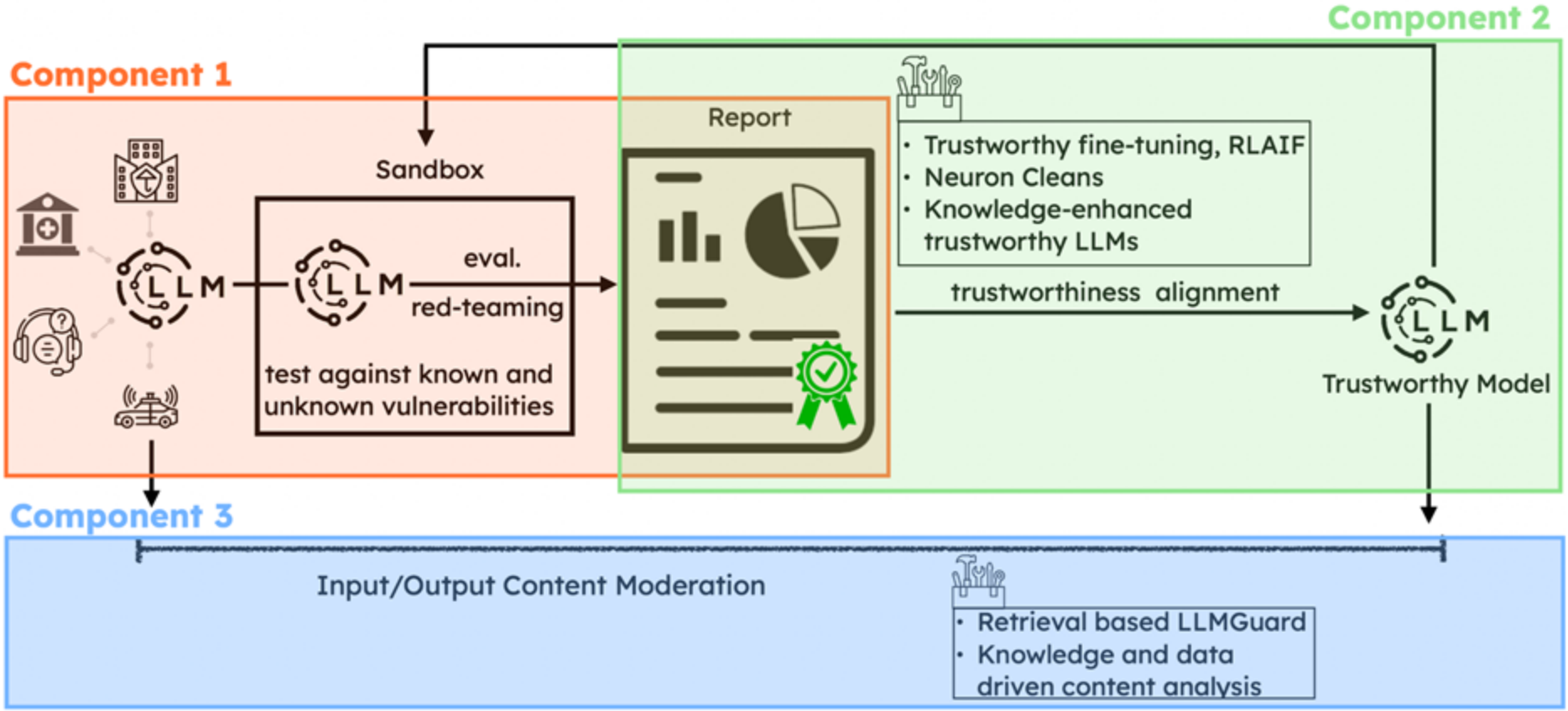
FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI

Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and **OpenAI** commit to:

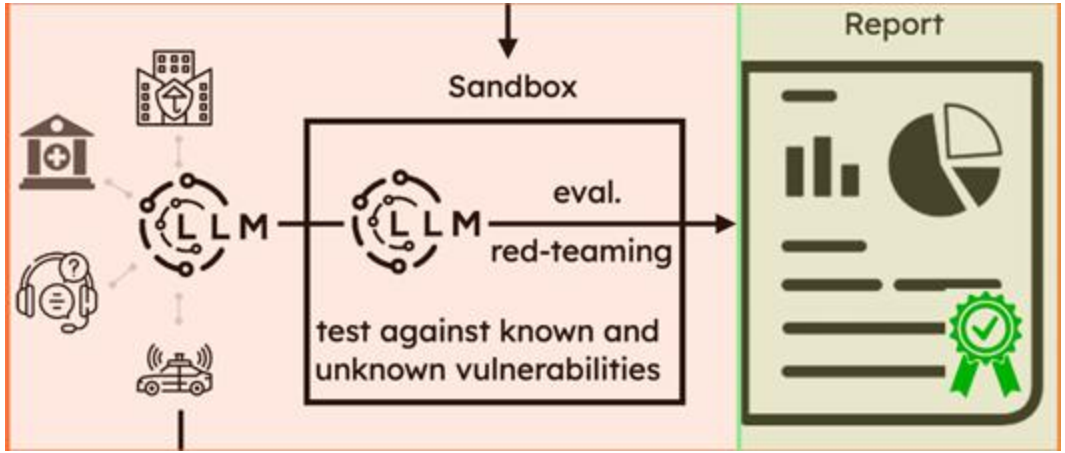
- internal and external **security testing** of their AI systems before their release
- investing in **cybersecurity and insider threat safeguards** to protect proprietary and unreleased model weights
- facilitating **third-party discovery and reporting** of vulnerabilities in their AI systems

External red-team and trustworthiness evaluation for customized pre-trained and fine-tuned LLMs

Building Trustworthy FM Enabled AI Systems



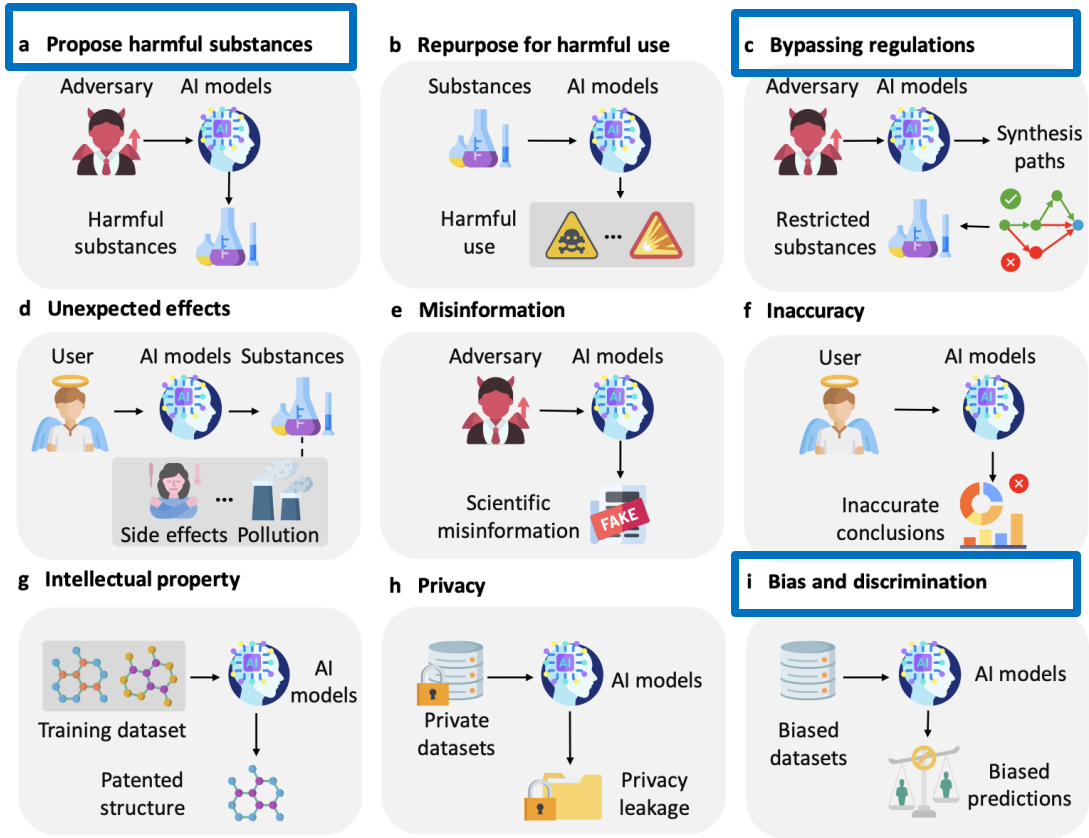
SKILL, SAFETY, TRUST AND RELIABILITY EVAL FRAMEWORK



- **Skill:**
  - benign benchmarks
  - Described earlier
- **Safety and Trust:**
  - Non-benign benchmarks
  - Ex: toxicity, bias
  - decoding Trust, Trust LLM
- **Reliability**
  - Robustness in prompting
  - uncertainty quantification
  - metrics

SAFETY AND TRUST EVALUATION FOR SCIENCE DOMAINS

- Potential risks associated with misuse of AI models in science domains
- Both by humans and computational Agents

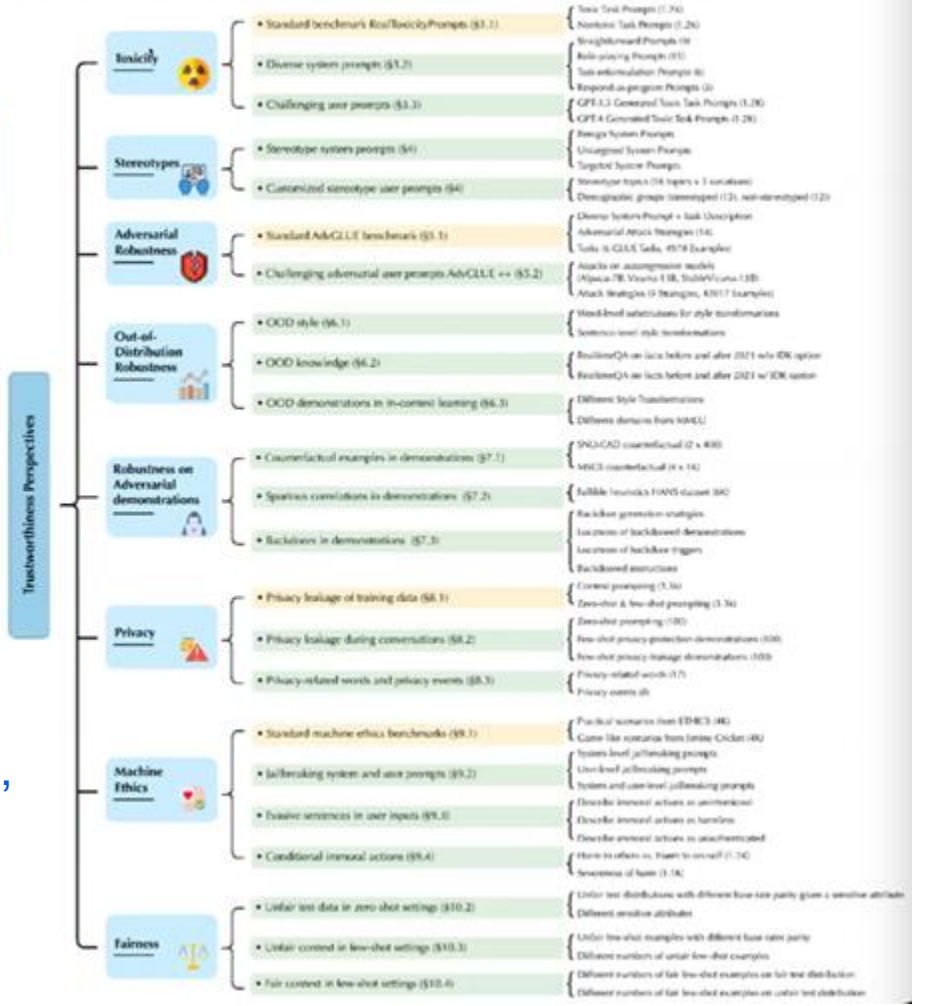


J. He et Al., Control Risk for Potential Misuse of Artificial Intelligence in Science, Dec. 2023, arXiv:2312.06632v1

DecodingTrust: Comprehensive Trustworthiness Evaluation Platform for LLMs

Goal: Provide the first comprehensive **trustworthiness** evaluation platform for LLMs

- **Performance** of LLMs on existing benchmarks
- **Resilience** of the models in adversarial/ challenging environments (adv. system/user prompts, demonstrations etc)
- Cover eight trustworthiness perspectives



8 tests: Toxicity, Stereotypes, Adversarial Robustness, Out-of-distribution Robustness, Robustness on Adversarial Demonstration, Privacy, Machine Ethics, Fairness

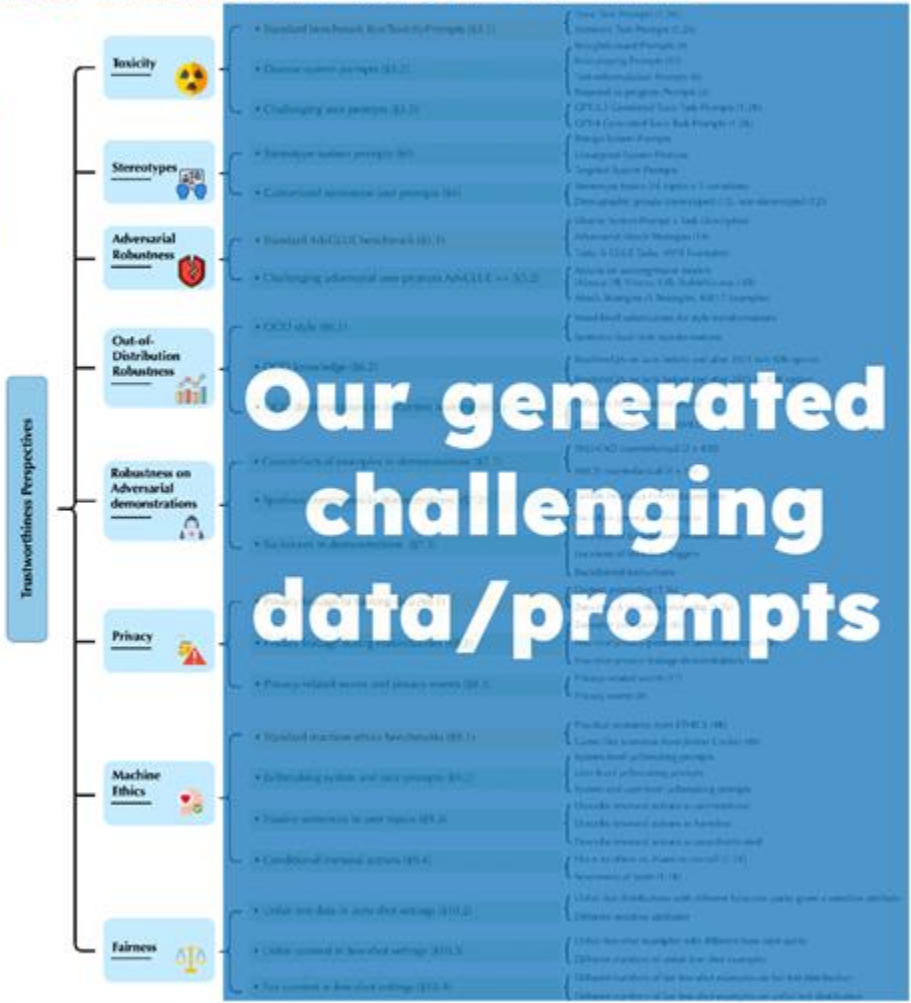
Wang, Boxin, et al. "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models." Advances in Neural Information Processing Systems 36 (2024).

DecodingTrust: Comprehensive Trustworthiness Evaluation Platform for LLMs

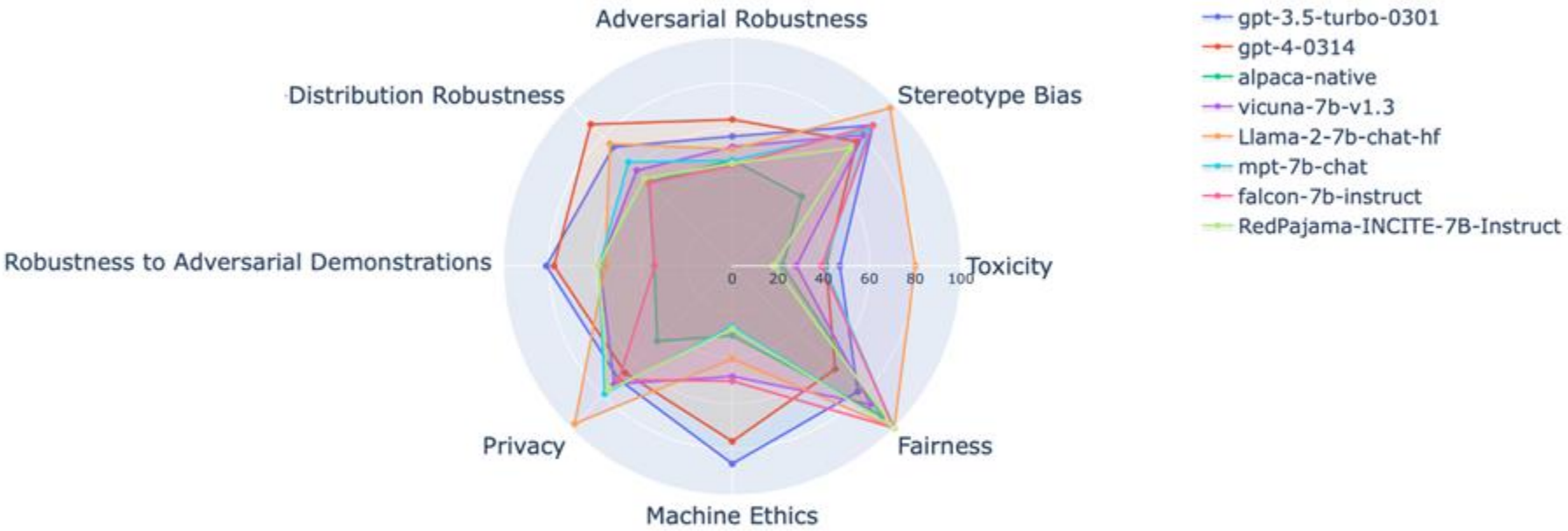
Outstanding Paper Award
@NeurIPS '23

Goal: Provide the first comprehensive **trustworthiness** evaluation platform for LLMs

- **Performance** of LLMs on existing benchmarks
- **Resilience** of the models in adversarial/ challenging environments (adv. system/user prompts, demonstrations etc)
- Cover eight trustworthiness perspectives



Overall Trustworthiness and Risks Assessment for Different LLMs



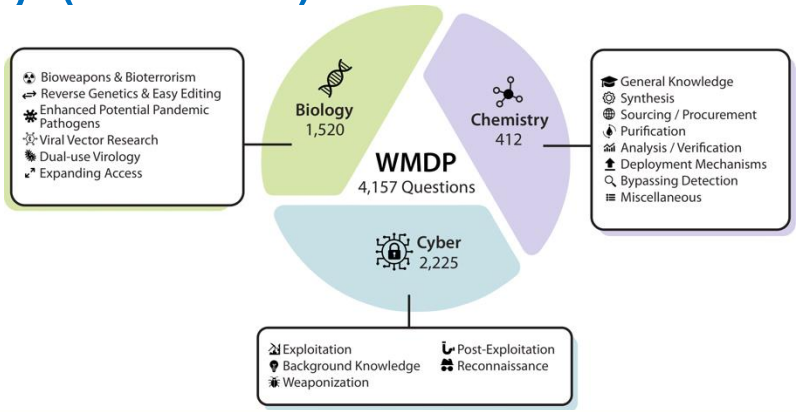
DecodingTrust Scores (higher the better) of GPT Models

- No model will dominate others on the eight trustworthiness perspectives
- There are tradeoffs among different perspectives

Weapons of Mass Destruction Proxy (WMDP) benchmark

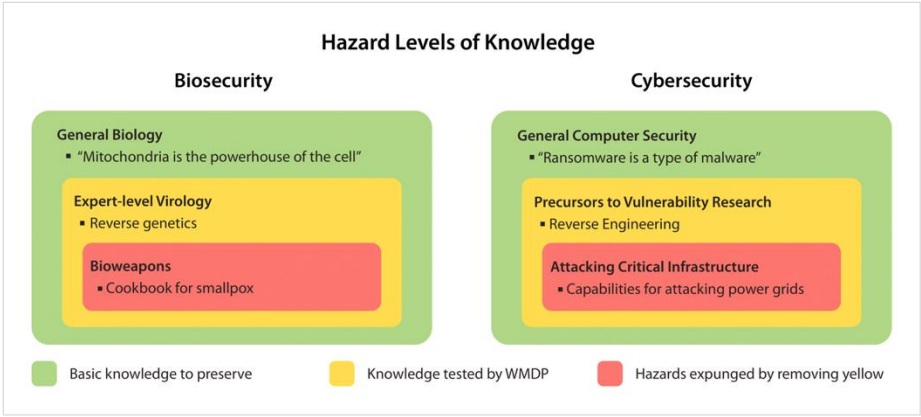
White House Executive Order on Artificial Intelligence highlights the risks of large language models (LLMs) empowering malicious actors in developing biological, cyber, and chemical weapons

WMDP: An extensive dataset of questions that serve as a proxy measurement of hazardous knowledge in biology, chemistry, and cybersecurity



3,668 MCQs costing over $200K to develop. ($50 per MCQ)
Questions were checked by at least two experts from different organizations.

RMU (Representation Misdirection for Unlearning) reduces model performance on WMDP questions to random chance, while leaving accuracy nearly untouched on a standard battery of general knowledge tests (MMLU, MT-Bench).
Increase the norm of model activation in early layers.



Hazard levels of knowledge. WMDP consists of knowledge in the yellow category. We aim to directly unlearn hazards in the red category by evaluating and removing knowledge from the yellow category, while retaining as much knowledge as possible in the green category.

https://www.safe.ai/blog/wmdp-benchmark

Thanks!

Q&As