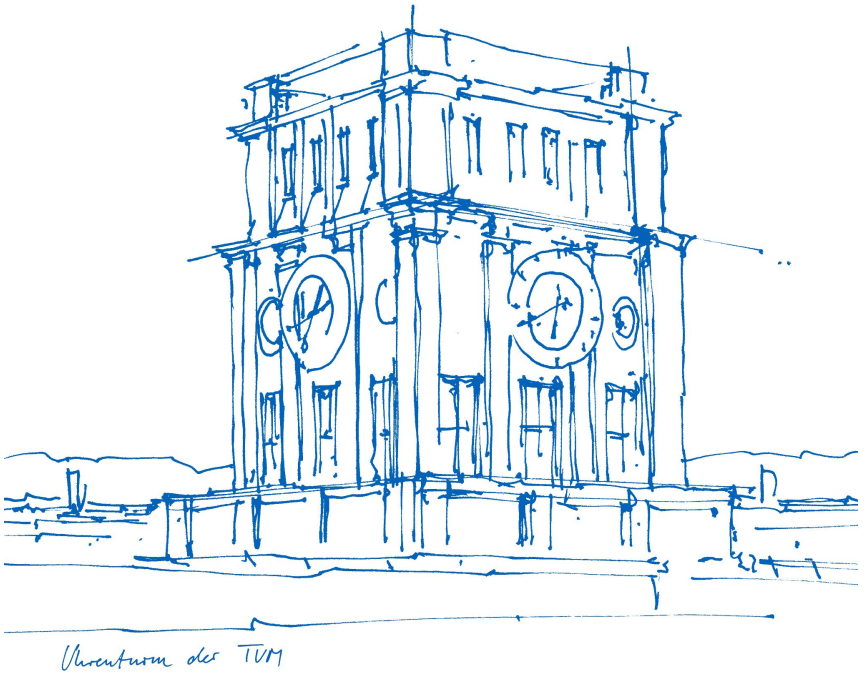
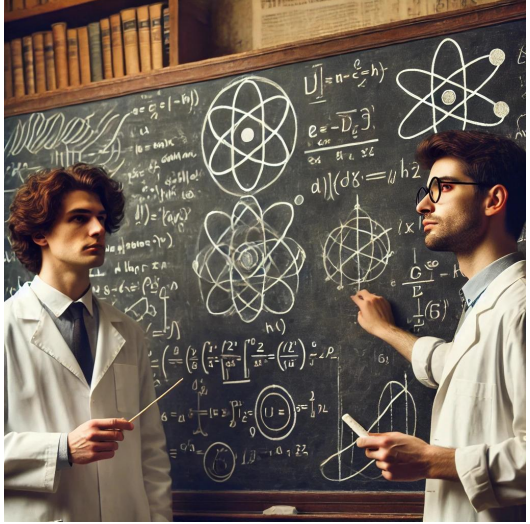
AI Alignment: Problem of Diversity & Physics

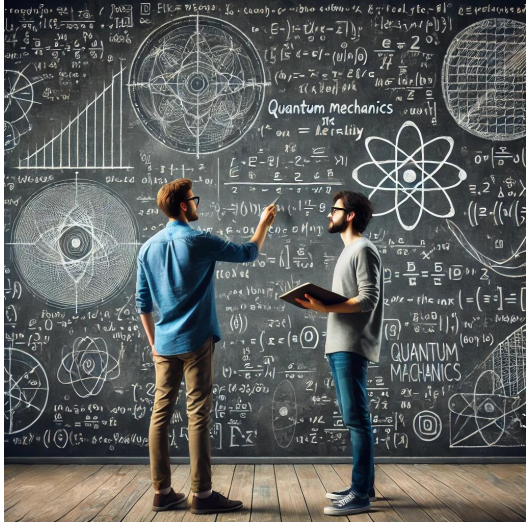
presented by

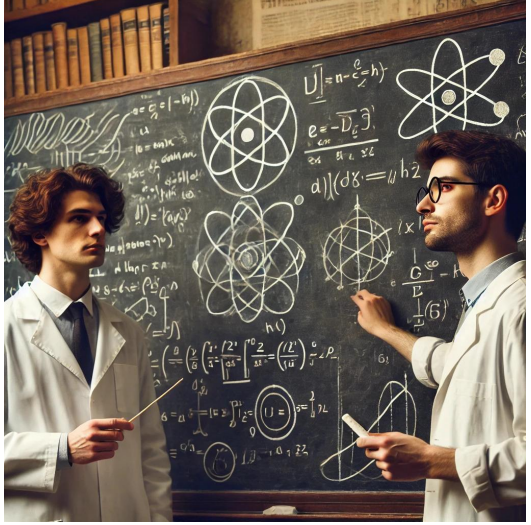
Philipp Alexander Kreer

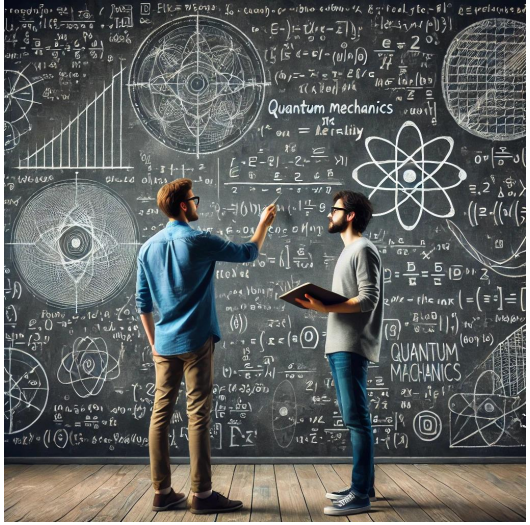
Technical University of Munich
APART Research

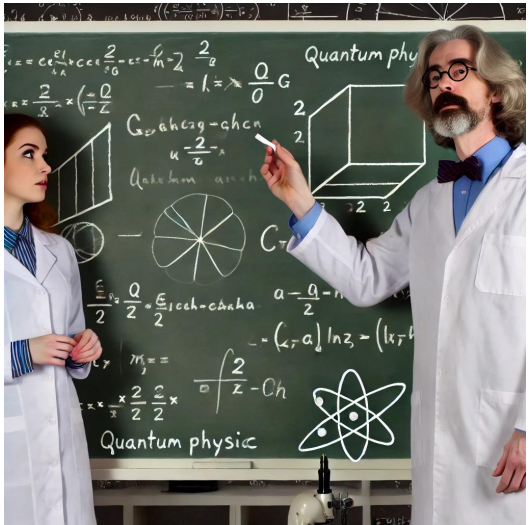





Quantum mechanics
QUANTUM
MECHANICS




Quantum mechanics
QUANTUM
MECHANICS


Quantum physic

Diversity in GPT4o?

❌ Gender equality
❌ Ethnic equality
❌ Gender neutrality
🤔 Why do they wear white coats?

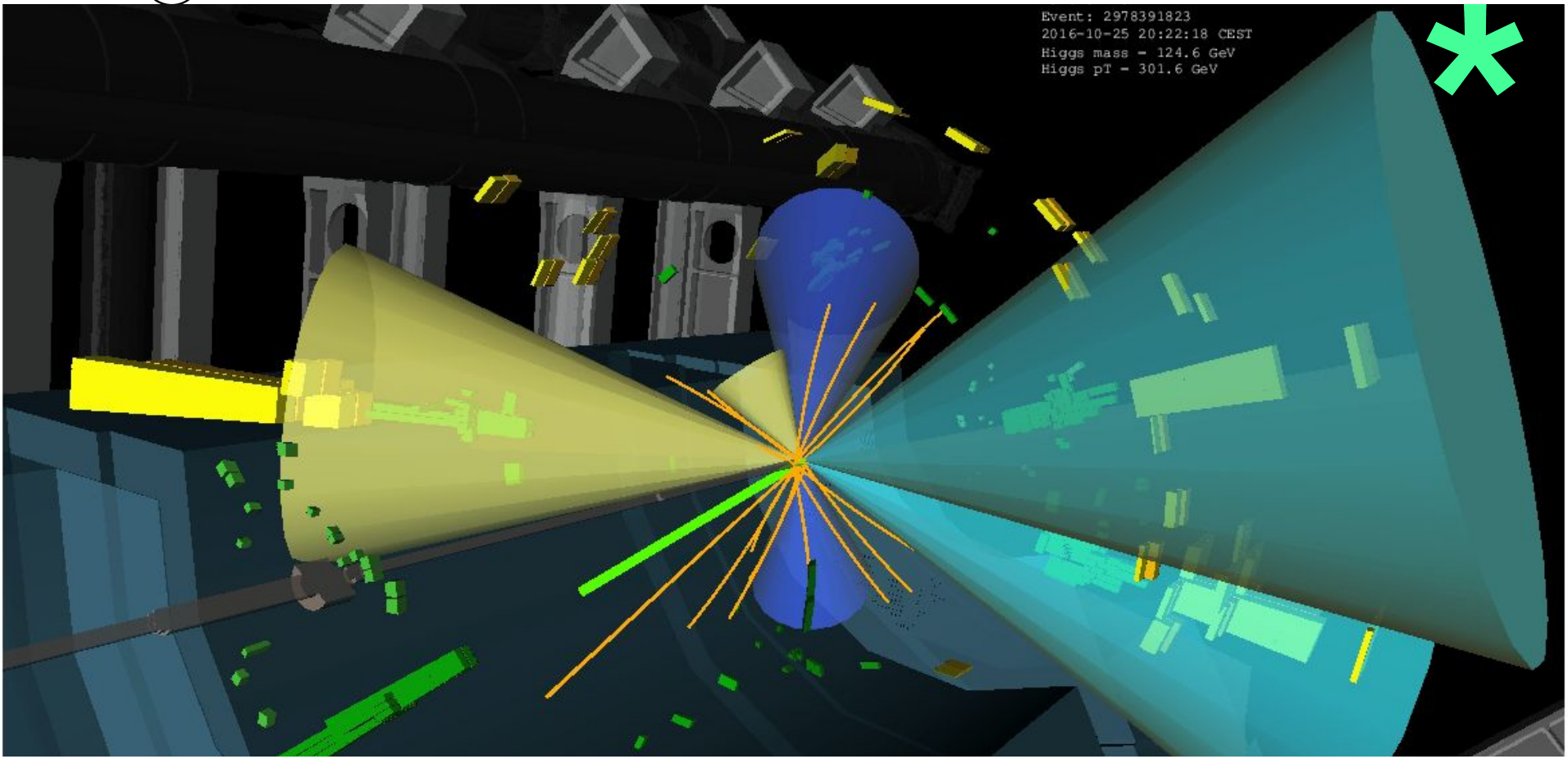
→ “AI Models are biased” (Ananya, 2024)

The pope according to Google’s Gemini (Brown, 2024)

AI Alignment Problem

"How can we create agents that behave in accordance with the user's intentions?" (Leike et al., 2018)

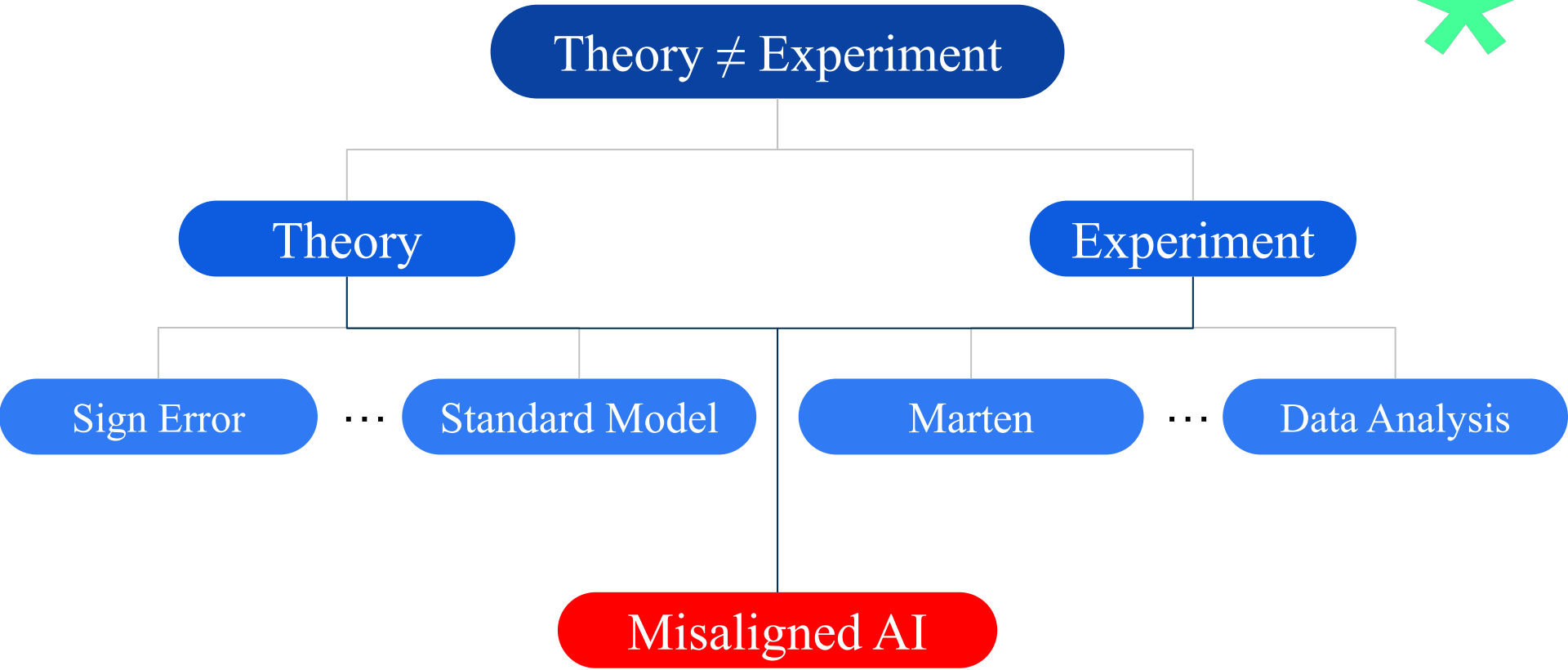
AI @ LHC (ATLAS 2024, Winterhalder et al., 2024)



AI in Theoretical Particle Physics

- Polylog simplification (Dersy et al., 2024)
- Bootstrapping amplitudes N=4 SYM (Cai et al., 2024)
- Amplitude regression (Spinner et al., 2024)
- Simplification of amplitudes (Cheung et al., 2024)
- Integration of Feynman integrals (Calisto et al., 2024)

Alignment Problem in Particle Physics



Science of Deep Learning

Science of Deep Learning

Singular Learning Theory = Bayesian Learning
+ Theoretical Physics
+ Symmetries
(+ Algebraic Geometry)

(Watanabe, 2009)

Science of Deep Learning

Singular Learning Theory = Bayesian Learning
+ Theoretical Physics
+ Symmetries
(+ Algebraic Geometry)

(Watanabe, 2009)

Right theory of deep learning?

Test Singular Learning Theory

Sandbagging = Strategic Underperformance of AI Systems

(Järviniemi et al., 2024; van der Weij et al., 2024)

Intuitive picture:

1. Sandbagging = shallow property
2. Add noise = destroy Sandbagging behavior

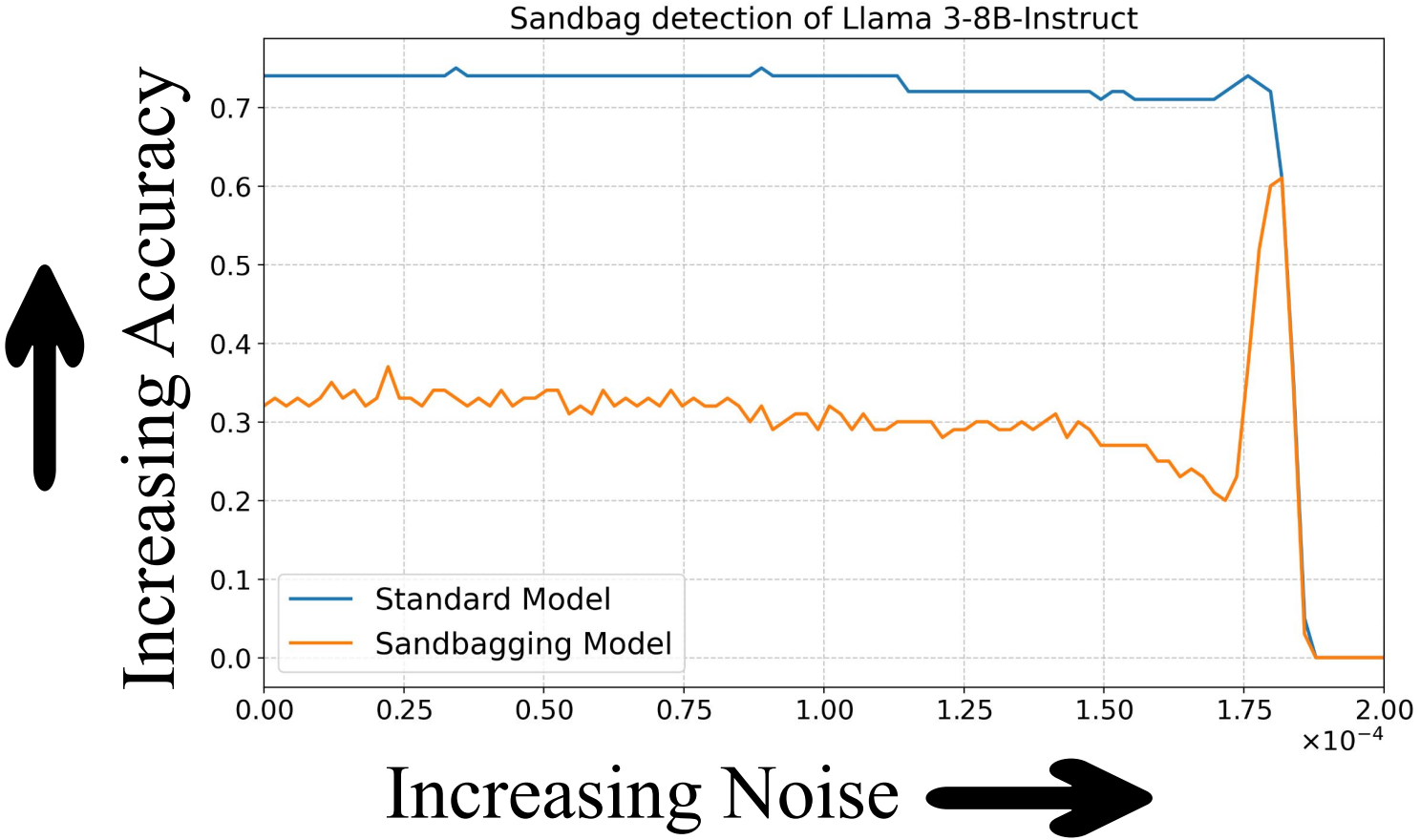
Experimental Setup

1. Large Language Model e.g. Llama 3 8b
2. Bench mark e.g. tinyMMLU
3. Initiate Deceptive Mode e.g. prompting or fine-tuning
4. Draw random numbers from Gaussian distribution
5. Add noise to model parameters
6. Measure accuracy in dependence of noise

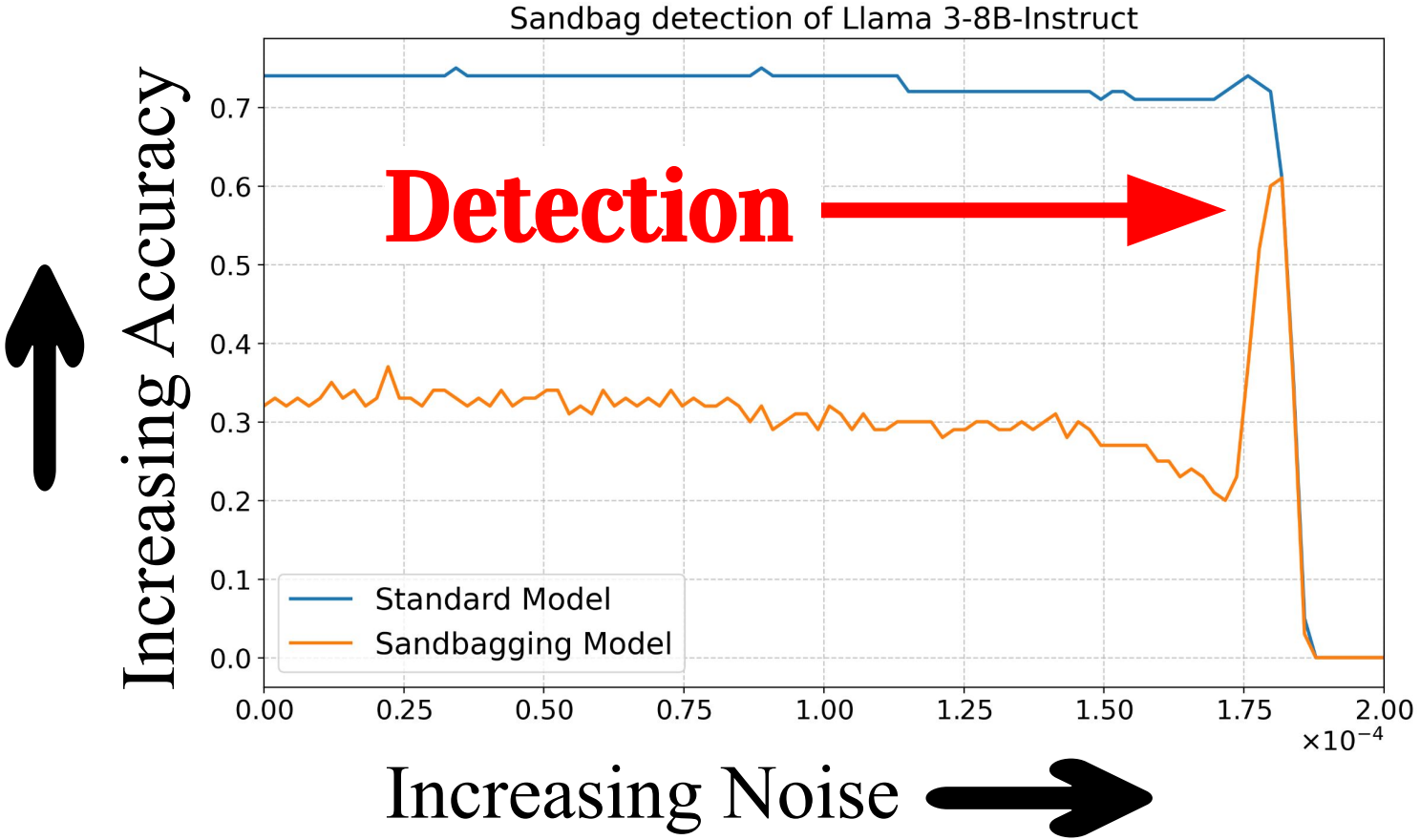
(Tice, PAK, et al., 2024)

In collaboration with APART Research and Cam Tice, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov,Teun Van der Weij, and Jacob Haimes

Results


Sandbag detection of Llama 3-8B-Instruct
Increasing Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
×10⁻⁴
Standard Model
Sandbagging Model
Increasing Noise

Results


Sandbag detection of Llama 3-8B-Instruct
Detection
Increasing Accuracy
Increasing Noise
Standard Model
Sandbagging Model
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
×10⁻⁴

And now?

Empirical work:

- Non-trivial conceptional finding
- Generalizes to multiple models
- Generalizes to password-locked models

And now?

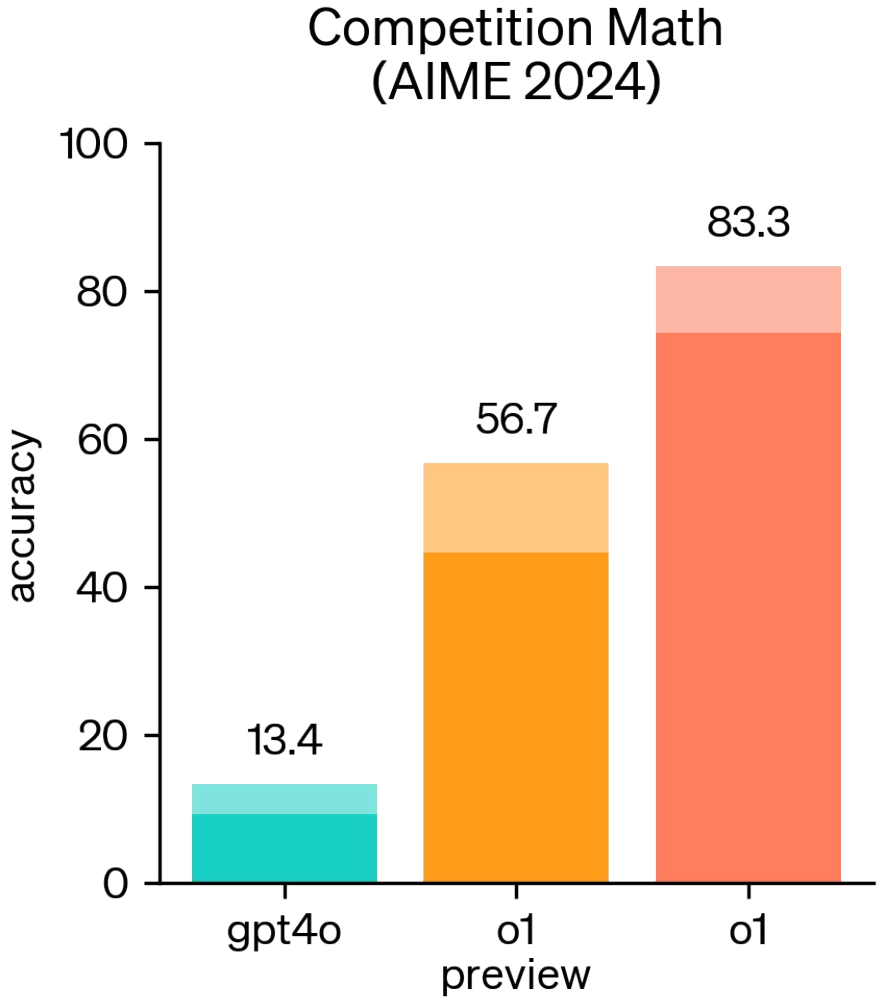
Empirical work:

- Non-trivial conceptional finding
- Generalizes to multiple models
- Generalizes to password-locked models

Theoretical Work:

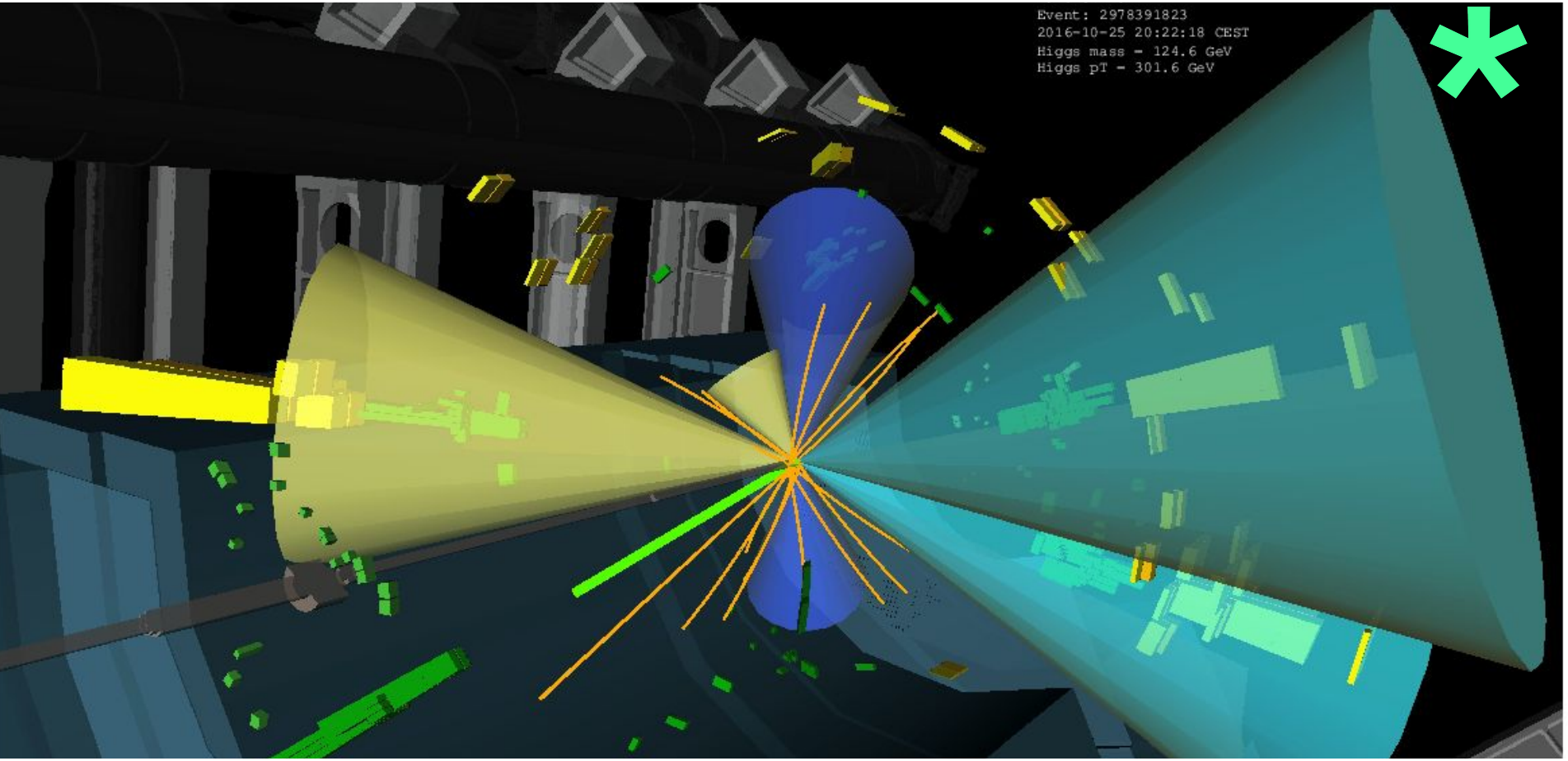
- Bayesian learning ≠ gradient descent
- Singular learning theory tractable for real-world models?
- Effective field theory approach?

Outlook



(OpenAI, 2024)

Outlook


Event: 2978391823
2016-10-25 20:22:18 CEST
Higgs mass = 124.6 GeV
Higgs pT = 301.6 GeV

Conclusion

1. Solving AI alignment is crucial to ensuring equal diversity and unlocking AI's potential in particle physics.
2. Rich interaction between particle physics and AI Alignment research

PostDoc - April 2025

- AI solutions for Particle Physics
- Explainable AI
- AI Alignment

Thank you!

https://github.com/camtice/SandbagDetect

Sources:

Ananya. (2024, March). *AI image generators often give racist and sexist results: Can they be fixed? Nature, 627*(8005), 722–725. https://doi.org/10.1038/d41586-024-00674-9

Brown, E. N. (2024, June). *The Great Black Pope and Asian Nazi Debacle of 2024*. *Reason*. https://reason.com/2024/05/28/the-great-black-pope-and-asian-nazi-debacle-of-2024/

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction*. arXiv preprint arXiv:1811.07871.

ATLAS Collaboration. (2024). *Higgs Boson Observations in ATLAS Experiment*. Retrieved September 27, 2024, from https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/HIGG-2020-24/

Winterhalder, R., et al. (2024). *A living review of machine learning for particle physics*. GitHub. Retrieved August 7, 2024, from https://github.com/iml-wg/HEPML-LivingReview

Spinner, J., Bresó, V., de Haan, P., Plehn, T., Thaler, J., & Brehmer, J. (2024). *Lorentz-equivariant geometric algebra transformers for high-energy physics*. arXiv preprint arXiv:2405.14806.

Cai, T., Merz, G. W., Charton, F., Nolte, N., Wilhelm, M., Cranmer, K., & Dixon, L. J. (2024). *Transforming the bootstrap: Using transformers to compute scattering amplitudes in planar N = 4 Super Yang-Mills theory*. arXiv preprint arXiv:2405.06107.

Dersy, A., Schwartz, M. D., & Zhang, X. (2024). *Simplifying polylogarithms with machine learning. International Journal of Data Science and Mathematical Sciences, 1*(2), 135–179. https://doi.org/10.1142/S2810939223500028

Source II

Cheung, C., Dersy, A., & Schwartz, M. D. (2024). *Learning the simplicity of scattering amplitudes*. arXiv. arXiv:2408.04720.

Calisto, F., Moodie, R. & Zoia, S. (2024). *Learning Feynman integrals from differential equations with neural networks*. J. High Energ. Phys. 2024, 124.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory* (Cambridge Monographs on Applied and Computational Mathematics). Cambridge University Press.

Tice, C., Kreer, P. A., Ryzhenkov, F., Helm-Burger, N., & Shahani, P. S. (2024). *Sandbag detection through model degradation*. Research submission to the **Deception Detection Hackathon: Preventing AI deception** research sprint hosted by **Apart Research**. https://apartresearch.com

Järviniemi, O., & Hubinger, E. (2024). *Uncovering deceptive tendencies in language models: A simulated company AI assistant*. arXiv. https://arxiv.org/abs/2405.01576

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). *AI Sandbagging: Language models can strategically underperform on evaluations*. arXiv. https://arxiv.org/abs/2406.07358

OpenAI. (2024, September 12). *Learning to reason with LLMs*. https://openai.com/index/learning-to-reason-with-llms/