LIFTING THE BENCHMARK ICEBERG WITH ITEM-RESPONSE THEORY

Mara Schilling-Wilhelmi
FSU Jena [*]

Nawaf Alampara
FSU Jena [†]

Kevin Maik Jablonka
FSU Jena, CEEC Jena, JCSM, HIPOLE Jena [‡]
mail@kjablonka.com

1 INTRODUCTION

Progress in chemistry and materials science increasingly relies on machine-learning models, making it essential to evaluate their performance rigorously (**?**). In the past, most such models have been built for a specific task. Measuring performance, in this case, is relatively easy as one can measure, for example, how the model performs on the task it has been trained on for a holdout set sampled from the same distribution. So-called foundation models (Bommasani et al., 2021), however, are trained to be of general purpose with the ability to solve many different tasks. This makes evaluating their performance difficult as one can not test them on any possible task. Instead, practitioners frequently use benchmarks that are compilations of exam-like questions in multiple-choice style. Often, these questions are not systematically created to measure a well-defined concept but instead collected from existing sources (which can introduce biases as the McNamara Fallacy describes (Basler, 2009)). Even though the individual questions define what the benchmark measures (i.e., we perform a rather pragmatic instead of representational measurement (Hand, 2016)), practitioners often summarize the performance of a model on a given benchmark in one (or a few numbers) to facilitate comparison between different models. This number is most commonly derived by simply averaging the scores obtained for different items in the benchmark. Clearly, not every question in such a benchmark corpus is created equally, and some might be more difficult than others or more suited to distinguish between models than others. In addition, it is not clear why the operation of an average should even be well-defined across a diverse set of questions. Thus, a simple average over the scores on individual items in a benchmark is an arbitrary choice as any other weighting (Banerjee et al., 2024; Binette & Reiter, 2024). Given that the performance of models on leaderboards created for commonly used benchmarks relies on many hidden decisions, this is very concerning. This is further exacerbated by the fact that metrics are random variables but are commonly reported only as point estimates without associated error bars (Miller, 2024).

Importantly, the choice of the aggregation of individual scores is only one of many decisions designers of benchmarks take and that are often taken for granted. In this work, we show the impact of those "hidden choices" in what we call the "benchmark iceberg" — where a lot of the variance in benchmark results is caused by implementation details. To improve transparency and comparability, we propose leveraging item-response theory (IRT) (Van der Linden & Hambleton, 2015), a well-established statistical framework from psychometrics. By adapting IRT, we transform opaque benchmarks (and associated leaderboards) into more rigorous measurement tools, making hidden assumptions explicit and providing error estimates.

[*] Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany.

[†] Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany.

[‡] Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany. Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany. Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany.

The challenge of benchmark reliability is particularly acute in materials science, where model performance directly influences real-world experimental design and resource allocation decisions. While chemistry and materials science benchmarks like ChemBench and MaCBench provide valuable starting points for evaluating foundation models in this domain, the hidden implementation choices in these benchmarks significantly impact how we perceive model capabilities. Our work addresses this gap by providing a statistical framework that can enhance the transparency and reliability of materials-focused benchmarks.

Concretely, we make the following contributions:

- **Demonstration of the impact of the benchmark iceberg**: Using two relevant benchmarks from the chemical sciences — ChemBench (Mirza et al., 2024) and MaCBench (Alampara et al., 2024) — we demonstrate that small changes in the implementation of a benchmark — that are often not communicated by users and developers of benchmarks — can fundamentally change the conclusions one might draw from a benchmark.
- **Implementation of item-response theory for more transparent and systematic analysis of benchmark results**: Building on item-response theory, we propose an analysis protocol that more clearly communicates assumptions and systematically analyzes the performance across multiple items in a benchmark, yielding capability metrics for each model and difficulties for each question. Thanks to our probabilistic implementation, we also obtain error bars. We envision that this will lead to more systematic benchmarking in machine learning for materials science.

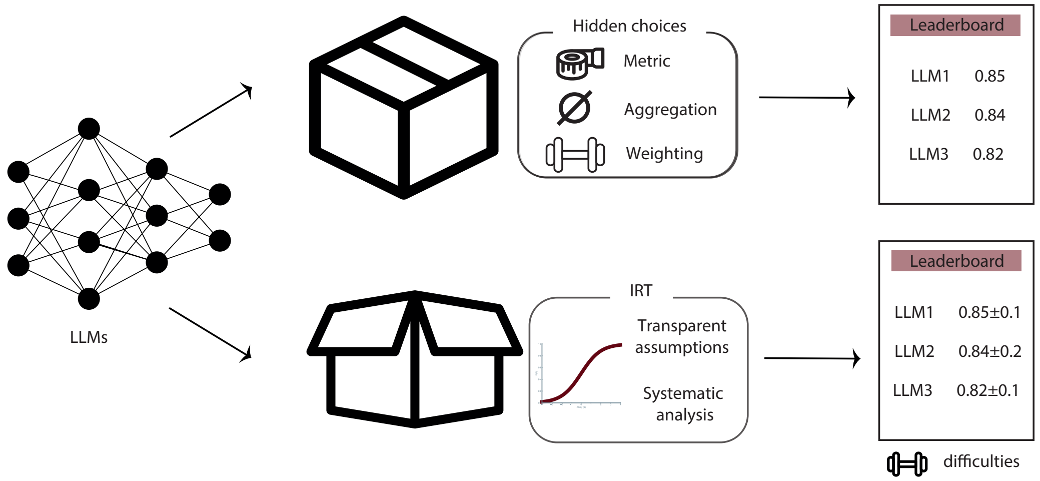

Figure 1: **Comparison of the current scoring workflow and the IRT analysis.** The current scoring workflow involves various choices of metrics, aggregation methods, and weightings, many of which remain implicit. In contrast, IRT scoring makes these assumptions transparent and enables a systematic analysis. Additionally, it not only produces a score but also provides error bars and difficulty estimates for all questions.

2 Related Work

LLM Benchmarks Traditional NLP benchmarks, which focused on specific, discrete tasks (Wang et al., 2018), no longer adequately measure the capabilities of current LLMs. These models now demonstrate broad knowledge and reasoning abilities across multiple domains. Current evaluation methods address reasoning, knowledge, reliability, and safety. New evaluation frameworks include MMLU (Hendrycks et al., 2020), which tests performance across 57 tasks from mathematics to law, along with specialized benchmarks for specific cognitive abilities. Other benchmarks even focus on tasks that are graduate-level and "Google proof" to adequately measure the performance of leading models (Rein et al., 2023). The challenge of evaluating LLMs is further complicated by their increased use in various domains such as coding, finance, and science. In science, broad benchmarks such as SciKnowEval (Feng et al., 2024), SciBench (Wang et al., 2023), JEEBench (Arora et al., 2023), SciQ (Welbl et al., 2017) have been proposed alongside domain specific benchmarks have been proposed such as ChemBench (Mirza et al., 2024), MaCBench (Alampara et al., 2024), LabSafety Bench (Zhou et al., 2024), LAB-Bench (Laurent et al., 2024), MaScQA for the chemical sciences (Zaki et al., 2024). All of these benchmarks rely on exam-like questions and aggregate scores to a final score.

Item Response Theory Item Response Theory (IRT) emerged from psychometrics as a statistical framework for measuring latent traits (such as intelligence) through test responses (e.g., on an exam). Originally developed for educational assessment, IRT provides the statistical foundation for many standardized tests and psychological measurements (Hori et al., 2020; Toland, 2013; Lovelace & Brickman, 2013).

At its core, IRT addresses the challenge of measuring characteristics that cannot be directly observed, such as ability, attitude, or personality traits. Instead, these latent traits must be inferred through responses to items or questions in tests or surveys. The fundamental approach of IRT is to model the probability of specific responses by examining the relationship between an examinee's ability and various item characteristics, such as difficulty and discrimination, via a so-called item-response curve. Mattos et al. (2021) have been using ideas from IRT to compare benchmark suites and could conclude using this methodology that the benchmarks they analyzed are not suitable to discriminate between relevant algorithms.

To our knowledge, IRT has not been used to systematically assess LLM benchmarks and our work lays the foundation for more systematic analyses of future benchmarks.

3 Methods

3.1 Benchmark

For our analysis, we use two recently proposed benchmarks for chemistry and materials science.

ChemBench ChemBench (Mirza et al., 2024) is a compilation of exam-like questions measuring models' knowledge and reasoning abilities across diverse chemical topics. While ChemBench includes both multiple-choice questions and questions that expect a floating-point number as the output, we focused our analysis only on the multiple-choice questions for greater clarity. However, the approach can also be generalized to the other questions. ChemBench also includes questions about chemical preference (Choung et al., 2023). However, since these measure quite different capabilities than the rest of the corpus, we excluded them from the analysis in this work.

MaCBench MaCBench (Alampara et al., 2024) is a compilation of multi-modal chemistry and material science questions measuring knowledge across key steps of the scientific process, like data extraction, experiments, and data analysis. Just like with ChemBench, we focus our analysis on the multiple-choice questions of MaCBench. We also excluded all ablation questions for our analysis.

In both ChemBench and MaCBench, a binary metric ("all correct") is used to classify whether a given answer is entirely correct.

3.2 Item response theory

For evaluating LLM benchmarks, we employ the two-parameter logistic (2PL) model that is commonly used in IRT studies:

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \tag{1}$$

Two parameters characterize each benchmark task:

- The discrimination parameter (a_i) indicates how effectively a task distinguishes between different levels of LLM capability

- The difficulty parameter (b_i) represents the ability level needed for a 50% probability of success

This model is particularly suitable for LLM evaluation as it captures both the varying complexity of benchmark tasks and their power to differentiate between model capabilities. The IRT framework enables us to estimate these parameters independently of the specific LLM sample, providing robust task characterization across different evaluation contexts. By placing benchmarks on a common scale, the method allows for meaningful comparisons across diverse datasets and tasks. Additionally, it identifies tasks that effectively differentiate between models with varying performance levels, thereby enhancing evaluation precision. Furthermore, adaptive testing methods can be employed to design more efficient benchmarks, reducing the time and resources required for comprehensive evaluation.

While traditional implementations of IRT often rely on maximum likelihood estimation, we adopt a fully Bayesian approach to the 2-parameter Logistic (2PL) model, leveraging probabilistic programming (Patil et al., 2010) to enhance interpretability and flexibility and to obtain error bars.

4 Results and Discussion

4.1 Arbitrariness of benchmarks due to "hidden choices"

To measure the sensitivity of major benchmarks to implementation details, we computed new rankings for the ChemBench benchmarks by compiling the results after performing different, sensible implementation changes. In particular, we test the sensitivity to

- **Score aggregation:** As an alternative to simple averaging, we employ different weighted averages.
- **Score choice:** The choice of the ChemBench developers to focus their evaluation on the binary "all correct" metric is a meaningful one. However, there are other meaningful choices, such as the use of Hamming loss, precision, or recall.

In Figure 2 we show the sensitivity of the ranking of models to those changes. The parallel coordinates plot shows how the ranking (higher is better) for different models (colored lines) changes under the definition of the score and its aggregation (x-axis).

We find that not only do the relative performance gaps between models change when the metric definition is changed, but the ranking of models also changes. For instance, from the perspective of recall, Galactica is the best model, whereas it is the worst from most other perspectives. However, we find that even under the same metric, some rankings change (e.g., Gemini-Pro vs. GPT-4) when we change the weighting of different questions in the aggregation.

4.2 Item response theory for analysis of benchmark results

As a first attempt to address some of these limitations, we implemented the commonly used 2-parameter logistic model from IRT (Equation (1)) as a probabilistic model (see Appendix A.2 for details).

One of the main outputs of fitting an IRT model to the benchmark results is the assignment of capabilities θ for each model. In Figure 5, we compare those scores to the one obtained with the current ChemBench implementation.

The first observable difference is that we can plot error bars thanks to our Bayesian implementation. Those error bars highlight that we cannot statistically distinguish the difference in capability between some models (e.g., o1-preview and Claude-3.5-Sonnet). We show a similar analysis for MaCBench in Figure 8, where most models are statistically indistinguishable from each other with the benchmark.

While there are changes in relative performance estimates, there are no rank reversals in the overall scores. However, if we analyze the performance per topic, this is no longer the case, and we can observe several rank reversals (see Figure 4 in the Appendix) — for example, changing our assessment of which model performs best in safety-related questions.

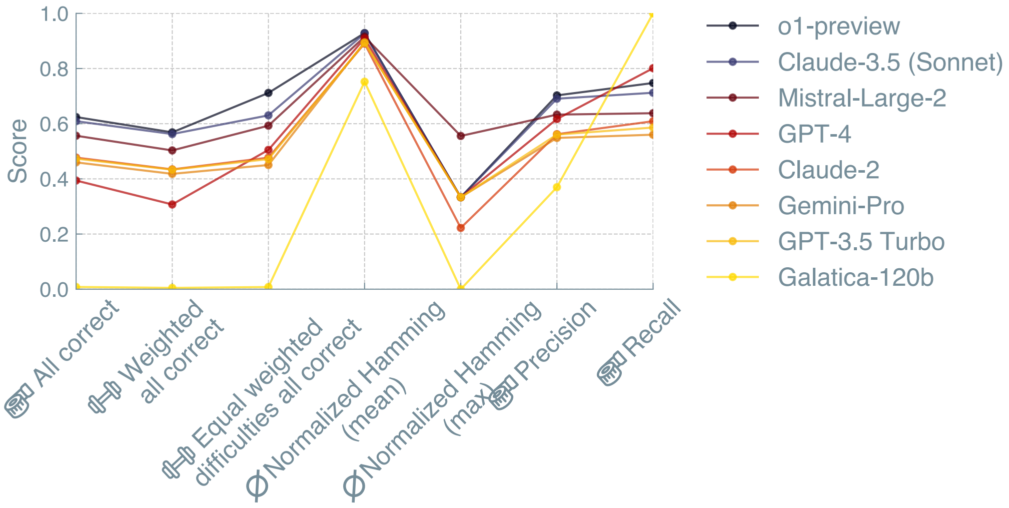

Figure 2: **ChemBench ranking based on different scoring metrics.** All metrics are a sum, weighted sum, or maximum values over all multiple-choice questions. The weighted sums are calculated by taking the manually rated difficulty (basic, immediate, and advanced) of the question into account. For equal weighting, all categories are weighted equally, regardless of the number of questions. The metric all correct is a binary metric indicating if a given answer is completely correct. For normalized Hamming (max), the normalized maximum value of the Hamming loss of each model was taken.

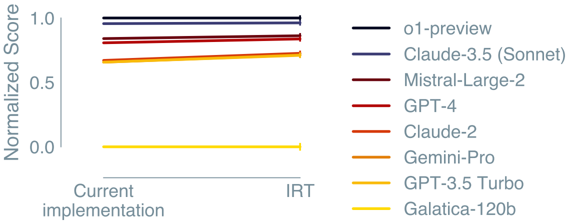

Figure 3: **Default scoring pipeline in ChemBench vs. IRT result.** Comparison between calculated ability (θ) and the current implementation (average all correct score). The error bars indicate the standard deviation of the posterior.

We compared the IRT model to a simple averaging (as it is currently done) using the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010; Gelman et al., 2013). The IRT model demonstrated a substantially better fit (ΔWAIC = 7776.3 ± 107.2, though the high posterior variance of log predictive densities suggests some instability in these estimates (Vehtari et al., 2017)), indicating that it is a better choice for analyzing the benchmark results.

4.3 Implications for Materials Science Benchmarking

The materials science community faces unique benchmarking challenges due to the multi-scale nature of materials properties, the diversity of experimental validation approaches, and the need to balance accuracy with computational efficiency. Our analysis of ChemBench and MaCBench reveals that current benchmarking approaches may provide an incomplete picture of model capabilities specifically relevant to materials discovery workflows.

For instance, the large error bars in our IRT analysis of MaCBench (Figure 8) suggest that current benchmarks may not offer sufficient statistical power to differentiate between models for materials-specific tasks.

The variation in model ranking across different chemistry sub-domains (Figure 4) further suggests that materials benchmarks must better account for domain-specific capabilities rather than averaging across diverse task types.

5 Limitations and Future Work

Our current implementation of IRT handles binary outcomes well but needs extension in three key directions. First, we need to incorporate partial credit and floating-point answers to capture the full spectrum of scientific responses. Second, we must develop models that can handle multiple metrics simultaneously — a crucial feature for modern benchmarks that measure diverse aspects of performance. Third, we should extend the framework to other popular benchmarks and package it as an easy-to-use tool for the broader community. These extensions are all technically feasible and build naturally on the foundation we have established. The mathematical framework of IRT is flexible enough to accommodate these improvements, and existing work in educational testing provides clear pathways for implementation. By tackling these challenges systematically, we can create a comprehensive evaluation framework that serves the needs of the entire scientific language model community.

6 Conclusions

Model benchmarks evolved from simple accuracy metrics on narrow tasks to comprehensive test suites attempting to capture the expanding capabilities of language models. Yet, with this evolution came complexity — and opacity. Our analysis revealed how seemingly minor implementation choices in benchmarks can dramatically alter conclusions about model performance, threatening to reduce our evaluation frameworks to mere illusions of progress.

Our work represents an important step toward more rigorous model evaluation. Using IRT, we move beyond the current paradigm of arbitrary scoring choices and point estimates. IRT not only quantifies model capabilities with confidence intervals but also makes hidden assumptions explicit, helping to reveal and address the benchmark iceberg.

To improve future benchmarking practices, it is crucial to explicitly document metric choices, aggregation methods, and scoring principles while assessing the impact of implementation decisions. Integrating uncertainty quantification, as demonstrated through our probabilistic implementation of IRT, should become standard practice. A more systematic and transparent approach will lead to more meaningful model comparisons and better-informed decision-making in model selection and development.

We hope that this work lays part of the foundation for a systematic science of model evaluations. By making benchmark assumptions explicit and providing statistical rigor to performance comparisons, we can move beyond simple leaderboards toward a deeper understanding of model capabilities. This transition from arbitrary scoring to principled evaluation frameworks is essential as language models increasingly support scientific discovery and decision-making.

To support this transition, we provide an open-source implementation of our IRT-based evaluation approach at github.com/lamalab-org/irt-on-bench. This package can be readily used to evaluate existing and future benchmarks, encouraging more robust, transparent, and reproducible model comparisons.

References

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv: 2411.16955*, 2024.

Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 7527–7543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.468. URL https://aclanthology.org/2023.emnlp-main.468/.

Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. The vulnerability of language model benchmarks: Do they accurately reflect true llm performance?, 2024. URL https://arxiv.org/abs/2412.03597.

M. H Basler. Utility of the mcnamara fallacy. *BMJ*, 339(aug04 3):b3141–b3141, August 2009. ISSN 1468-5833. doi: 10.1136/bmj.b3141. URL http://dx.doi.org/10.1136/bmj.b3141.

Olivier Binette and Jerome P. Reiter. Improving the validity and practical usefulness of ai/ml evaluations using an estimands framework, 2024. URL https://arxiv.org/abs/2406.10366.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Oh-Hyeon Choung, Riccardo Vianello, Marwin Segler, Nikolaus Stiefl, and José Jiménez-Luna. Extracting medicinal chemistry intuition via preference machine learning. *Nature Communications*, 14(1), October 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42242-1. URL http://dx.doi.org/10.1038/s41467-023-42242-1.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv: 2406.09098*, 2024.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24:997–1016, 2013. doi: 10.1007/s11222-013-9416-2. URL http://link.springer.com/article/10.1007/s11222-013-9416-2/fulltext.html.

D. J. Hand. *Measurement: a very short introduction*. Very short introductions. Oxford University Press, Oxford, first edition edition, 2016. ISBN 9780198779568.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Kazuki Hori, Hirotaka Fukuhara, and Tsuyoshi Yamada. Item response theory and its applications in educational measurement part i: Item response theory and its implementation in r. *WIREs Computational Statistics*, 14(2), October 2020. ISSN 1939-0068. doi: 10.1002/wics.1531. URL http://dx.doi.org/10.1002/wics.1531.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. Labbench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv: 2407.10362*, 2024.

Matthew Lovelace and Peggy Brickman. Best practices for measuring students' attitudes toward learning science. *CBE—Life Sciences Education*, 12(4):606–617, December 2013. ISSN 1931-7913. doi: 10.1187/cbe.12-11-0197. URL http://dx.doi.org/10.1187/cbe.12-11-0197.

David Issa Mattos, Lucas Ruud, Jan Bosch, and Helena Holmström Olsson. On the assessment of benchmark suites for algorithm comparison. *arXiv preprint arXiv: 2104.07381*, 2021.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv: 2411.00640*, 2024.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists?, 2024. URL `https://arxiv.org/abs/2404.01475`.

Anand Patil, David Huard, and Christopher J Fonnesbeck. Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1, 2010.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv: 2311.12022*, 2023.

Michael D. Toland. Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, 34(1):120–151, November 2013. ISSN 1552-5449. doi: 10.1177/0272431613511332. URL `http://dx.doi.org/10.1177/0272431613511332`.

Wim J Van der Linden and Ronald K Hambleton. *Handbook of item response theory*. CRC press, 2015.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018. URL `https://api.semanticscholar.org/CorpusID:5034059`.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv: 2307.10635*, 2023.

Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594, 2010. URL `http://jmlr.org/papers/v11/watanabe10a.html`.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *NUT@EMNLP*, 2017. doi: 10.18653/v1/W17-4413.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024. ISSN 2635-098X. doi: 10.1039/d3dd00188a. URL `http://dx.doi.org/10.1039/D3DD00188A`.

Yujun Zhou, Jingdong Yang, Kehan Guo, Pin-Yu Chen, Tian Gao, Werner Geyer, Nuno Moniz, Nitesh V Chawla, and Xiangliang Zhang. Labsafety bench: Benchmarking llms on safety issues in scientific labs. *arXiv preprint arXiv: 2410.14182*, 2024.

7 Acknowledgment

This work was supported by the Carl Zeiss Foundation. A grant from OpenPhilanthropy additionally supported parts of the work. In addition, M.S-W.'s work was supported by Intel and Merck via the AWASES programme. K.M.J. is part of the NFDI consortium FAIRmat funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 460197019.

A Appendix

A.1 Advantages of IRT-based analysis

Our Bayesian 2PL implementation provides three key advantages over simple averaging:

1. Probabilistic modeling explicitly accounts for measurement uncertainty through $P(X_{ij} = 1|\theta_j) = \text{logit}^{-1}(\alpha_i(\theta_j - \beta_i))$, where model ability θ_j and item parameters (α_i, β_i) are jointly estimated
2. Invariant comparisons enable meaningful question difficulty estimates (β_i) that are independent of specific model samples
3. Discrimination-aware weighting automatically emphasizes questions that best differentiate model capabilities (α_i parameters)

Note that IRT also offers the possibility for adaptive benchmark design: That is, the creation of small benchmarks that show maximum discrimination between models (and/or lowest uncertainty in the capability estimates).

A.2 Model specification

Our hierarchical 2PL model adopts a fully Bayesian approach using probabilistic programming, offering flexibility and robust uncertainty quantification.

Individual Responses For each model n attempting task j, the response is:

$$x_{nj} \sim \text{Bernoulli}(p_{nj}), \quad p_{nj} = \text{logistic}(\alpha_j(\theta_n - \beta_j))$$

where θ_n represents model ability, β_j the task difficulty, and α_j the task discrimination.

Model Parameters These priors reflect the natural variability of models and tasks, ensuring stable estimates even with sparse data.

Model Abilities: $\theta_n \sim N(0, \sigma_\theta^2)$
Task Parameters: $\alpha_j \sim N^+(\mu_\alpha, \sigma_\alpha^2), \quad \beta_j \sim N(\mu_\beta, \sigma_\beta^2)$

Hyperpriors Hyperpriors govern the variability in model abilities and task parameters, enabling the model to adapt dynamically to the data.

$$\sigma_\theta \sim \text{HalfNormal}(0.5), \quad \mu_\alpha \sim N(1, 0.1^2), \quad \sigma_\alpha \sim \text{HalfNormal}(0.2)$$
$$\mu_\beta \sim N(0, 0.5^2), \quad \sigma_\beta \sim \text{HalfNormal}(0.5)$$

Sampling Our Hamiltonian Monte Carlo implementation used:

- 4 chains with 2,000 tune + 2,000 sample iterations
- Target acceptance rate 0.99 for improved high-dimensional sampling

A.3 Scores per topic

A.4 Question difficulty assessment

An important piece of information in benchmarks is how difficult the different elements (e.g., questions) are. Most commonly, as in ChemBench and MaCBench, this annotation is performed manually and thus subjective and noisy. Interestingly, the fit of an IRT model to the benchmark results also yields a fully data-driven difficulty assignment per question.

It is thus interesting to compare the manually assigned difficulties to the data-driven ones.

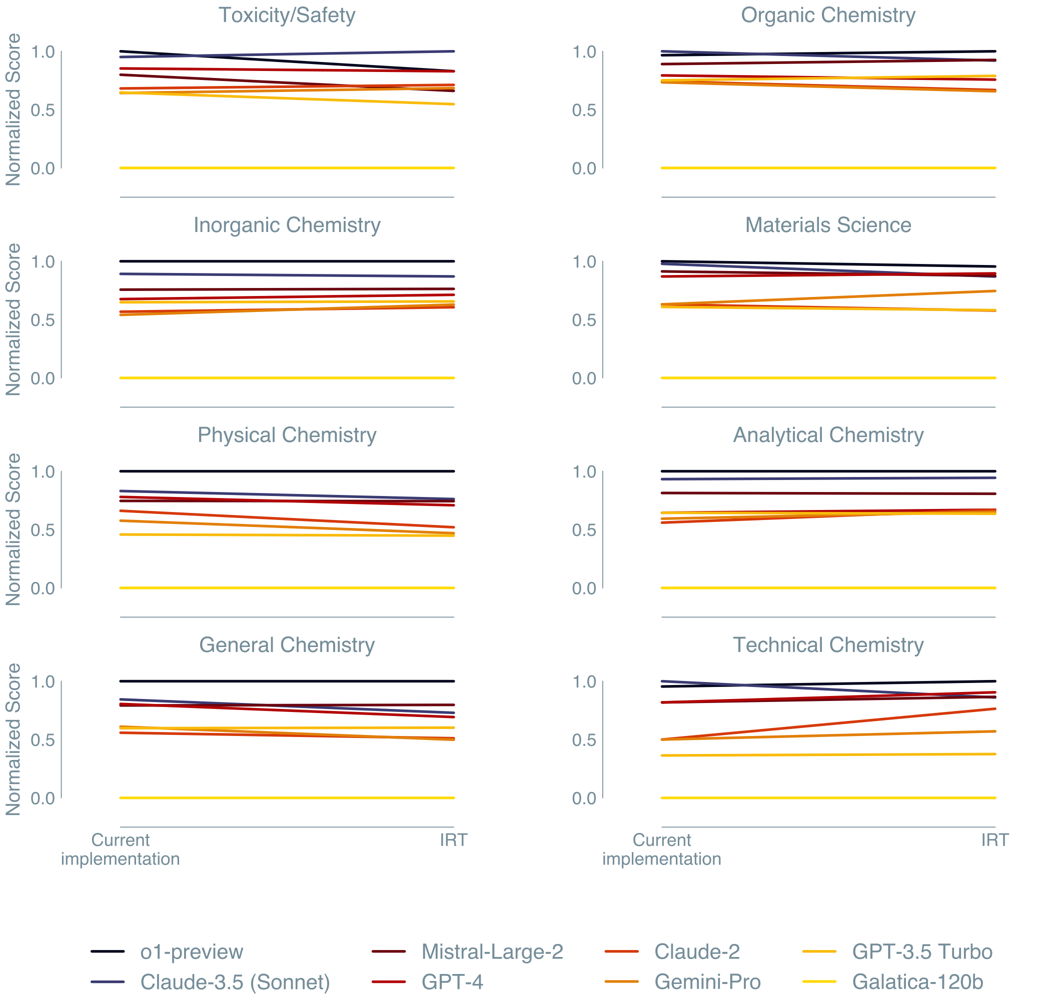

Figure 4: **IRT and all correct scores separate on each topic ChemBench:** Comparison between calculated ability (θ) and the current implementation (average all correct score). The error bars indicate the standard deviation of the posterior.

In Figure 6, we show the distribution of difficulties β for different manually assigned difficulty classes. We can observe that the data-driven difficulties do not agree with the manually assigned ones. The low correlation between the human-assigned difficulties and the data-driven ones might be due to different factors:

- **Curation bias:** human raters might overweight surface features like technical vocabulary while underestimating models' pattern recognition capabilities on "complex" questions
- **Different reasoning patterns:** models might struggle with questions humans consider intermediate due to atypical reasoning patterns
- **Expertise projection bias:** Human raters overestimate difficulty of "basic" concepts that models handle reliably

A.5 MaCBench

Figure 8 and Figure 9 show the results of an IRT analysis on the MaCBench benchmark. We note the large error bars that make it statistically impossible to distinguish the capabilities of many of the analyzed models.

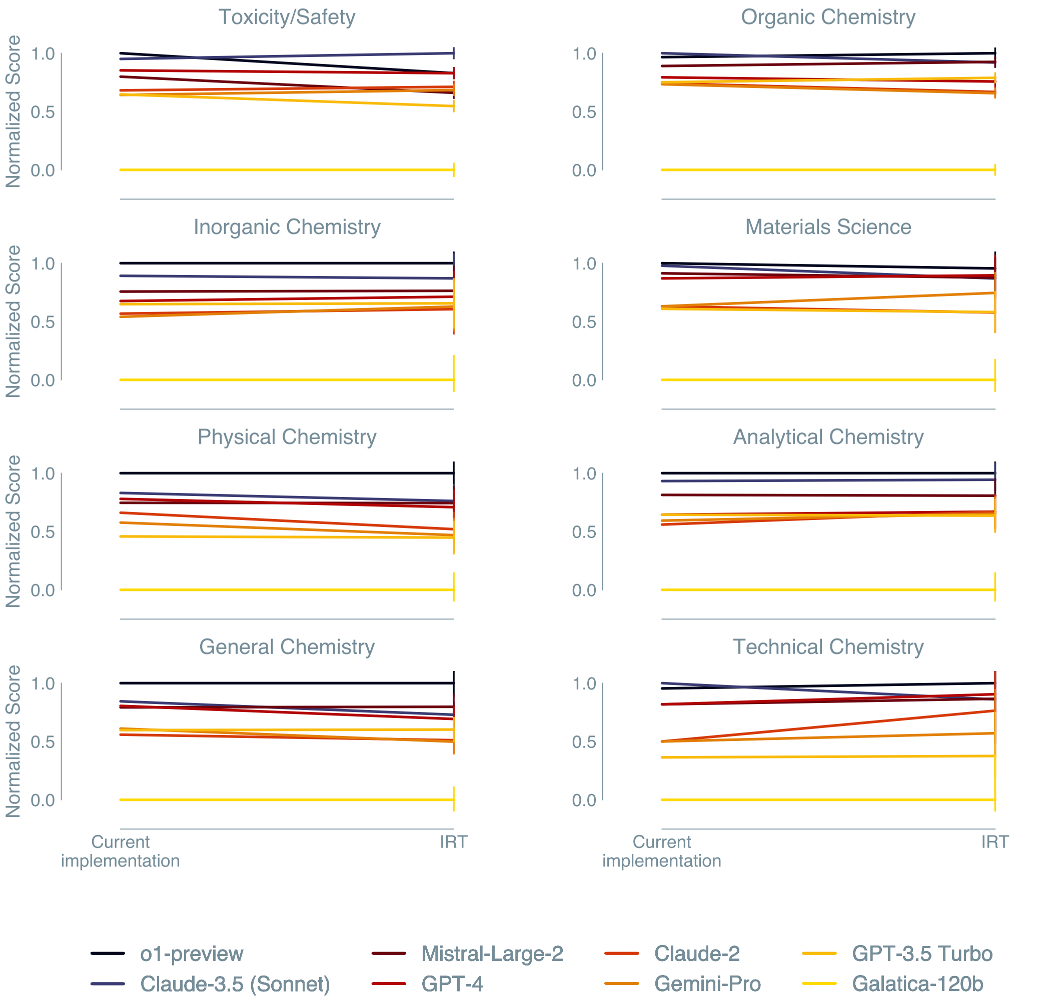

Figure 5: **IRT and all correct scores separate on each topic ChemBench:** Comparison between calculated ability (θ) and the current implementation (average all correct score).

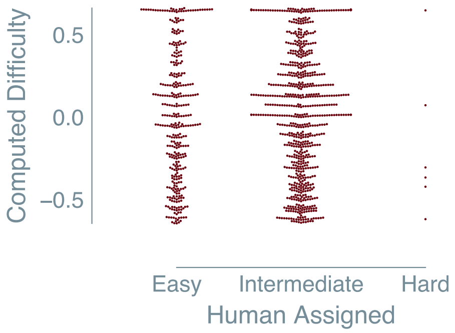

Figure 6: **ChemBench difficulty assignment comparison.** Comparison between manually assigned difficulties and difficulty computed with IRT.

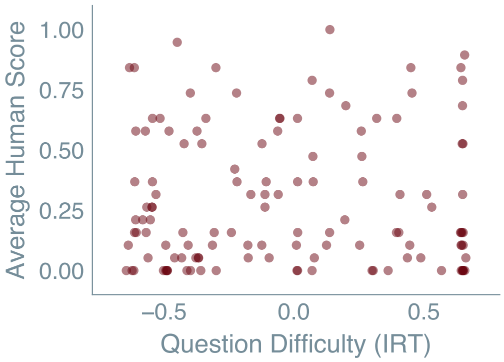

Figure 7: **ChemBench comparison of difficulty computed with IRT and human scores:** Comparison between average scores (all correct) over all humans compared to the computed difficulties with IRT.

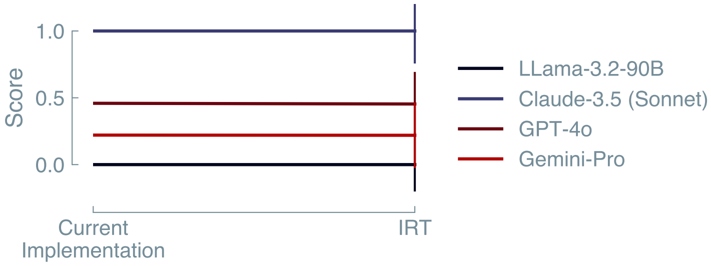

Figure 8: **IRT and fraction correct scores on MaCBench:** Comparison between calculated ability (θ) and the current implementation (average all correct score).

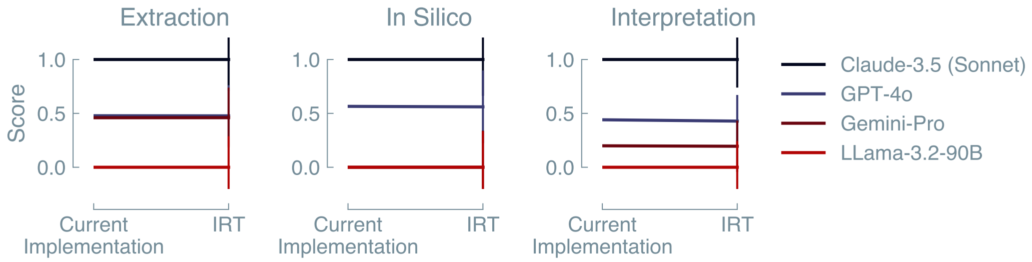

Figure 9: **IRT and all correct scores for three main classes of tasks in MaCBench:** Comparison between calculated ability (θ) and the current implementation (average all correct score). Both scores are normalized using min-max scaling.