MASTER'S RESEARCH INTERNSHIP (M2)

Statistical Inference for Model Benchmarking

Olivier Cappé | **Nicolas Schreuder**

CNRS | **DI ENS, ENS – PSL** | **Centre Sciences des Données (CSD)**

The rapid development of successive generations of models that are in competition with each other raises the question of reliably evaluating their respective merits from an external point of view. The question is particularly acute for large language models (LLMs) for which many benchmarks have been developed to compare them on aspects of performance and safety, from answering questions correctly to refusing harmful requests. Their average scores provide a useful summary, but two models with similar scores may succeed on different questions, and a small lead may not persist on another comparable set [1]. Adding questions may improve the precision of a comparison, but increases computation and may add little information when questions are redundant.

Item response theory (IRT), originally developed for educational and psychological testing, connects these aims of reliable comparison and efficient evaluation by modelling individual responses in terms of model abilities and question characteristics [2]. In its *two-parameter logistic (2PL)* form, $Y_{ij} = 1$ if model i meets the evaluation criterion on question j, and $Y_{ij} = 0$ otherwise:

$$\Pr(Y_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{1}{1 + \exp\big[-a_j(\theta_i - b_j)\big]}, \qquad a_j > 0.$$

Here θ_i describes model i's ability on the aspect being assessed, and the difficulty b_j is the ability level giving a 50% probability of success. The discrimination a_j governs how sharply this probability increases with ability, allowing questions to differ in how well they distinguish models with similar abilities.

Fitting this model to the responses of several models allows questions to be compared by the information they provide. Empirical studies have used IRT to predict full-benchmark scores from shorter tests [3]; on safety benchmarks, it has also helped distinguish refusal behaviour from truthfulness and detect some cases where models were instructed to underperform [4].

A key issue is that, in contrast to more traditional use of the IRT model, typical LLM comparisons may contain thousands of questions but only a few dozen evaluated models. Each model supplies many answers, yet each question's difficulty and discrimination must be inferred from relatively few responses. A precise average score can therefore coexist with uncertain question parameters, and studies document difficulties with item estimation and model rankings in this setting [5].

Objectives. The internship will study how accurately abilities and question characteristics can be estimated in this regime, and how to quantify their uncertainty for model comparison. Building on existing IRT theory, including finite-sample results for simpler models [6], it will combine theoretical analysis with the assessment and design of estimation methods. Simulations and benchmark experiments will examine their accuracy and sensitivity to modelling assumptions. If time permits, open problems related to benchmark compression and sequential evaluation will also be considered.

Candidate profile. M2 student in applied mathematics, statistics or machine learning with a solid foundation in probability and statistics and an interest in pursuing a PhD at DI ENS / CSD.

Contact. Interested candidates should write to olivier.cappe@ens.fr and nicolas.schreuder@ens.fr with the subject `[model benchmarking internship]`.

References

[1] Drew Keller, Kweku Kwegyir-Aggrey, Ryan Steed, Anita K. Rao, Julia L. Sharp, and A. Stevie Bergman. Expanding the AI evaluation toolbox with statistical models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 56704–56739. PMLR, 2026. URL `https://proceedings.mlr.press/v306/keller26a.html`.

[2] Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Item response theory—a statistical framework for educational and psychological measurement. *Statistical Science*, 40(2):167–194, 2025. doi: 10.1214/23-STS896.

[3] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR, 2024. URL `https://proceedings.mlr.press/v235/maia-polo24a.html`.

[4] Joshua Fonseca Rivera, Neil Shah, David Demitri Africa, and Konstantinos Voudouris. Item response theory for AI safety. arXiv preprint arXiv:2608.05086, 2026. URL `https://arxiv.org/abs/2608.05086`.

[5] Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? arXiv preprint arXiv:2607.15190, 2026. URL `https://arxiv.org/abs/2607.15190`.

[6] Yuepeng Yang and Cong Ma. Random pairing MLE for estimation of item parameters in Rasch model. *Journal of the American Statistical Association*, 121(554):1683–1694, 2026. doi: 10.1080/01621459.2025.2582873.