

When One Benchmark Hides Many Truths: Mixture-IRT and LLM Difficulty Prediction

Enis Dogan¹, Burhan Ogut²

¹National Center for Education Statistics ²American Institutes for Research

Abstract

Studies that validate LLM-predicted item difficulty conventionally benchmark predictions against a single set of item parameter estimates fitted to the full examinee sample, treating the ordering of item difficulty as homogeneous across the population. We test that assumption using responses from 14,501 examinees to a 29-item national university-entrance mathematics examination, together with difficulty predictions from a locally hosted multimodal LLM (gemma4:26b) elicited through blind, position-counterbalanced pairwise comparison and aggregated via a Bradley–Terry model. A three-class mixture Rasch model shows latent classes that differ in proficiency, intended collegiate field, and instructional track. LLM predictions align most closely with the largest, highest-proficiency class ($\rho = .69$) and markedly less closely with the two lower-proficiency classes ($\rho = .48$ and $.49$), while the conventional single-population benchmark ($\rho = .63$) sits between these values and conceals the spread entirely. Because a judgment that one item is harder than another presumes a solution process (Mislevy & Verhelst, 1990), these predictions may encode an examinee model that is broadly correct without being universal. We recommend that LLM difficulty-prediction studies examine whether latent classes exist in their validation samples and, where they do, report alignment separately by latent class.

1 Introduction

A growing literature uses machine learning and large language models (LLMs) to predict the empirical difficulty of test items, supporting item development, pre-calibration, and assessment design. Several studies presented at 2025 AIME-Con report encouraging results (e.g., Mojoyinola et al., 2025; Adegoke et al., 2025; Han et al., 2025; Li et al., 2025), as does a growing wider literature (AlKhuzayy et al., 2024; Kolesnikova et al., 2026).

The prevailing validation convention in this literature benchmarks predictions against a single set of empirical difficulty estimates, obtained by fitting one item response theory (IRT) model to a full examinee sample and treated as the “ground truth”. This convention embeds a strong and

rarely examined assumption that the ordering of item difficulty is homogeneous across the population.

If the examinee population comprises qualitatively distinct latent subgroups for whom the item-difficulty orderings differ, consistent with subgroups employing different solution strategies, a benchmark computed against a single set of item parameter estimates averages over that heterogeneity. Predictions may then align well with the difficulty estimates of some subgroups and poorly with those of others, and such a benchmark will both understate the best attainable alignment and obscure systematic misalignment for particular subpopulations.

This heterogeneity has particular significance for AI-based difficulty prediction. A judgment that one item is harder than another implies something about the work a test-taker must do to answer it, and so presumes a solution process. Yet, test-takers employ a variety of strategies in responding to items. For example, Mislevy and Verhelst (1990) constructed an example in which items relatively hard under one solution strategy are relatively easy under another. If an LLM's predictions encode a particular model of how examinees approach items, that model may be broadly correct without being universal, aligning with the subpopulations whose approach it resembles and diverging from those it does not.

Using item responses from a national university-entrance mathematics examination ($N = 14,501$) and difficulty predictions from a multimodal LLM elicited through blind pairwise comparison, we ask: (1) Does a multi-class latent structure exist in a real high-stakes dataset, and is it substantively interpretable? (2) Do the item parameters differ enough across classes to matter for differential validation? (3) Does the alignment between LLM predictions and empirical difficulty vary by latent class, and if so, which class does the LLM most closely track? We find that alignment varies substantially across classes and that the aggregate benchmark conceals this variation.

2 Related Work

2.1 Predicting Item Difficulty from Item Content

Model-based prediction of item difficulty from item content predates the current generation of language models. Early approaches relied on readability indices, item metadata, and hand-coded cognitive features, whereas contemporary work replaces or augments these with *learned* text representations. For instance, Han et al. (2025) fine-tuned four transformer architectures on Grade 6 ELA items, reaching correlations near .70 with empirical difficulty. Li et al. (2025) compared fine-tuned small language models against larger general-purpose LLMs on medical licensure items and found the smaller fine-tuned models often stronger, and related work extends this line to international large-scale assessment (Mojoyinola et al., 2025; Adegoke et al., 2025). Across this literature, the reported outcome is a benchmark computed against a single set of item parameter estimates fitted to the full sample.

2.2 Pairwise Comparison as an Elicitation Format

A parallel line of work elicits difficulty judgments comparatively rather than absolutely, drawing on the comparative-judgment tradition in educational measurement (Bramley, 2010; Crompvoets et al., 2020). Kolesnikova et al. (2026) provide the most direct precedent, comparing absolute judgment against pairwise comparison across six domains of primary-school mathematics and aggregating pairwise judgments via the Bradley–Terry model (Bradley & Terry, 1952). They found pairwise elicitation generally produced stronger alignment with empirical difficulty, and that few-shot prompting conferred little benefit. Their design randomized item order within each pair to mitigate order effects, though the magnitude of those effects was not reported. The present study adopts pairwise elicitation and additionally counterbalances rather than randomizes order, which permits the position-consistency diagnostic in Section 4.3.

2.3 Population Heterogeneity in Item Difficulty

Mixture item response models relax the assumption that a single set of item parameters can be obtained for an entire examinee

population. Rost (1990) integrated latent class analysis with the Rasch model, allowing distinct classes each with its own item-difficulty vector. Mislevy and Verhelst (1990) framed the same idea substantively, modeling responses when examinees employ different solution strategies. Bolt et al. (2002) applied mixture Rasch models to test speededness, showing that latent classes can capture interpretable differences in how subgroups engage with items rather than mere noise.

2.4 The Gap

These two literatures have not been brought into contact. Difficulty-prediction studies validate against a population-aggregate benchmark, but the mixture-IRT research establishes that such a benchmark can average over substantively distinct subpopulations. If distinct latent classes do exist, benchmarking predictions against a single set of item parameter estimates would be akin to a weighted average of subgroup-specific alignments that may differ considerably. No existing study, to our knowledge, reports LLM-difficulty alignment separately by latent class.

3 Methods

3.1 Analytic Logic

Our analysis proceeds in three linked stages. First, we establish a single-population (“one-class”) Rasch difficulty ordering, replicating the convention used throughout the LLM difficulty-prediction literature. Second, we relax the homogeneity assumption using mixture Rasch modeling. Model selection and covariate profiling assess whether the resulting classes are statistically supported and substantively interpretable rather than merely fit artifacts. Third, we generate an independent, LLM-derived difficulty ordering and evaluate its agreement with both the aggregate benchmark and the class-specific difficulty estimates.

3.2 Instrument

Data are item-level responses and the original item content from the mathematics subtest of a 2004 national university-entrance examination administered in an OECD-member country. The original mathematics subtest comprised 44 items covering numbers and operations, algebra, word problems, and geometry. Nine trailing items with 20% or more not-reached responses were

removed to address speededness, leaving 35. One item with a proportion-correct of .98 was removed for lack of discriminating information, yielding 34. An exploratory factor analysis (tetrachoric correlations, promax rotation) identified five coordinate-geometry items loading on a secondary factor; these too were removed, yielding a content-coherent set of 29 items suitable for unidimensional IRT modeling.

3.3 Sample

The analytic sample ($N = 14,501$) was 43.2% female and 56.8% male. By instructional track, 61.3% of students were enrolled in a mathematics–science program, 33.3% in a mathematics–social studies program, and 5.4% in a social-sciences program. Examinees also indicated the collegiate field they intended to apply to: a large majority (68.2%) reported interest in a STEM-related field, while the remainder indicated fields that admit students based on an equally weighted mathematics–social studies exam score.

3.4 Mixture Rasch Modeling

We fit a one-class Rasch model as a baseline and replication of prior literature, then two-, three-, and four-class mixture Rasch models using the *psychomix* package in R (raschmix, saturated score model; Frick et al., 2012).

In preparing the dataset for the IRT estimation, skipped items — those left blank before the examinee's final recorded response — were treated as incorrect, on the assumption that the examinee saw the item and elected not to answer. Not-reached items, those following the last item attempted, were treated as not-presented rather than incorrect throughout estimation, since coding them incorrect could produce a spurious speededness class rather than a substantive ability class.

Each multi-class solution was estimated from 200 independent random starts executed in parallel with an EM iteration cap of 500, retaining the highest-log-likelihood solution to guard against local maxima. Model enumeration considered AIC, BIC, marginal fit improvement, convergence behavior across starts, classification certainty, and substantive interpretability. Class membership was profiled against intended collegiate field, instructional track, and sex, and against response accuracy, defined as the

percentage of attempted items answered correctly and the number of items not reached.

3.5 LLM Difficulty Prediction

Item difficulty was independently estimated via a blind, stateless pairwise-comparison pipeline using a local multimodal LLM (gemma4:26b, via Ollama). Each item was represented as an image crop of the original question, preserving the original text, mathematical notation, and any figures, so that the model judged items in their native visual form rather than a transcribed rendering. Two items included figures and one included a table.

Difficulty predictions were elicited as a complete round-robin of pairwise comparisons among the 29 analytic items. Each unordered pair was presented twice, once in each presentation order, at temperature 0, yielding 812 independent judgments ($406 \text{ unordered pairs} \times 2 \text{ orders}$). Each call presented two item images labeled ITEM A and ITEM B and required a forced binary verdict (ties disallowed) with a one-clause rationale, returned as constrained JSON. The model's chain-of-thought reasoning mode was disabled, so each response consisted of the verdict and rationale alone. Calls carried no conversation history, so each judgment is independent of every other and position bias can be assessed directly. These 812 judgments form the basis for all analyses reported in Section 4.

Judgments were aggregated into a continuous latent difficulty scale using the Bradley–Terry model (Bradley & Terry, 1952), under which the probability that item i is judged harder than item j is $\sigma(\theta_i - \theta_j)$, where θ is the item's difficulty on a log-odds scale and σ is the logistic function. Parameters were estimated by maximum likelihood over the 812 comparisons, with the scale identified by fixing one item's θ at zero. One item lost every comparison in which it appeared, for which no finite maximum-likelihood estimate exists; a small ridge penalty ($\lambda = 0.05$) on θ was applied to keep all estimates finite. The resulting scale was compared with empirical difficulty from the one-class baseline and each class of the three-class solution using Spearman's ρ and Pearson's r .

4 Results

4.1 Model Selection

Estimation was stable for the one-, two-, and three-class solutions. The three-class solution

was retained based on convergence behavior, classification quality, decelerating marginal fit improvement, and the substantive coherence of the resulting classes (Section 4.2). Fit improved monotonically across the one-, two-, and three-class solutions. BIC fell from 406,064.00 at one class to 401,661.43 at two, a gain of 4,403 points, and to 400,006.09 at three, a gain of 1,655. Classification quality was adequate and stable: mean maximum posterior probability was .907 at two classes and .857 at three, with relative entropy of .688 and .687 respectively, indicating that the third class did not degrade the certainty with which examinees are assigned.

The four-class solution was sought but eventually abandoned because it showed estimation instability and produced degenerate estimates. Although the additional class had a mixing proportion of 16.29%, multiple score-model parameters were estimated at $-\infty$, indicating that the solution was not identifiable in a substantively useful way.

4.2 Class Characterization

The three latent classes formed a clear, monotone proficiency gradient. Because not-reached responses were treated as missing, classes are distinguished primarily by response accuracy (correctness) on attempted items rather than by number of items attempted. Class 2 was the largest and highest-performing ($n = 8,887$; 61.3%; 73.7% correct on attempted items), Class 1 was intermediate in proficiency ($n = 3,648$; 25.2%; 46.3% correct on attempted items), and Class 3 was smallest and lowest-performing ($n = 1,966$; 13.6%; 35.4% correct on attempted items). Against a chance rate of roughly 20% for five-option items, even the lowest class performed above chance, so the gradient is one of proficiency rather than a contrast between engaged and guessing examinees. Relative entropy was .687, with a mean maximum posterior class probability of .857. Table 1 reports the covariate profile.

	Class 2	Class 1	Class 3
% of sample	61.3	25.2	13.6
% correct (attempted items)	73.7	46.3	35.4
Mean items not reached	0.03	0.85	2.61
Collegiate field (%): STEM	92.5	35.7	18.4
Female (%)	36.7	61.8	38.3
Track (%):			
Mathematics-Science	79.6	36.5	24.7
Mathematics-Social Studies	20.1	56.2	50.6
Social Studies	0.3	7.4	24.7

Table 1. Covariate profile by latent class, ordered by proficiency.

The intended collegiate field separated the classes most sharply ($V = .67$), with STEM field at 92.5% in Class 2, against 35.7% in Class 1 and 18.4% in Class 3, $\chi^2(2) = 6,442.8$, $p < .001$. Examinees also differed by instructional track by latent class ($V = .39$), with mathematics–science enrollment falling from 79.6% in Class 2 to 36.5% in Class 1 and 24.7% in Class 3, while social-studies enrollment rose from 0.3% to 7.4% to 24.7%, $\chi^2(4) = 4,306.6$, $p < .001$.

Sex was significantly but far more weakly associated with class membership ($V = .22$) and, importantly, was not monotone in proficiency: the intermediate Class 1 was majority female (61.8%), whereas both the highest-performing Class 2 (36.7% female) and the lowest-performing Class 3 (38.3% female) were majority male, $\chi^2(2) = 689.1$, $p < .001$.

Non-response followed the proficiency gradient. Class 2 examinees reached essentially the whole set (mean 0.03 items not reached), Class 1 substantially less so (mean 0.85 items not reached), and Class 3 least (mean 2.61 items not reached). The two lower-proficiency classes thus differed from the highest not only in number of correct answers but in test-taking endurance under time pressure, consistent with the speededness mechanism documented by Bolt et al. (2002).

4.3 Pairwise Comparisons and Bradley–Terry Scale

Two diagnostics bear on the quality of the resulting scale. The first is internal fit: how well the fitted Bradley–Terry scale reproduces the judgments from which it was estimated. The scale correctly predicted 750 of the 812 judgments (92.36%), indicating that the LLM's judgments were well summarized by a single latent difficulty dimension.

The second diagnostic is comparison accuracy: for each of the 812 judgments, whether the item the LLM selected as harder is in fact the harder of the pair under a given empirical benchmark. Unlike internal fit, this is a validity measure and varies by benchmark; it is reported alongside the correlations in Section 4.4. Because it is computed judgment by judgment rather than from the aggregated scale, it provides a check on alignment that does not depend on the Bradley–Terry aggregation or on the choice of correlation coefficient.

Position consistency is a third property of the judgment set. Of the 406 unordered pairs tested in both presentation orders, the LLM's judgment reversed with position in 54 pairs (13.3%) (judging item A as more difficult when A is presented first, then B as more difficult when B is presented first). Order sensitivity of this kind is a documented property of LLM judges rather than a peculiarity of this pipeline: position bias is a recognized failure mode in the LLM-as-a-judge literature (Zheng et al., 2023; Shi et al., 2024). Reversals here were not evenly distributed across items. The two most reversal-prone items appeared in 11 and 10 reversed pairs respectively, together accounting for 19.4% of all reversals, while five of the 29 items appeared in one or zero reversals. The most reversal-prone item had an empirical difficulty near the center of the scale.

4.4 Differential LLM Alignment Across Classes

Figures 1 and 2 report the agreement between the LLM's Bradley–Terry scale and empirical difficulty (β) from the single-population benchmark and each latent class: Figure 1 gives rank correlations, and Figure 2 gives comparison accuracy computed directly from the 812 individual judgments, where a comparison counts as accurate if the item the LLM judged harder is the one with the higher empirical difficulty estimate under that benchmark.

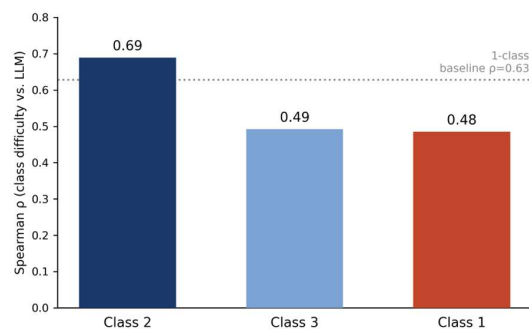


Figure 1. Spearman correlation between empirical item difficulty and LLM-predicted difficulty, by latent class. The dotted line shows the single-benchmark (one-class) correlation.

The LLM aligns most closely with the highest-proficiency class (Class 2: $\rho = .69$, $r = .63$, 72.91% accuracy in pairwise comparisons) and least closely with the intermediate and lowest-proficiency classes (Class 1: $\rho = .48$, $r = .45$, 65.02% accuracy; Class 3: $\rho = .49$, $r = .52$, 65.27% accuracy), a spread of .20 in rank correlation, corroborated by both the Pearson correlation and a roughly eight-percentage-point gap in comparison accuracy. Because accuracy is computed from individual pairwise judgments rather than the aggregated scale, the pattern is not an artifact of the correlation metric or of Bradley–Terry aggregation.

Critically, the single-population benchmark ($\rho = .63$, $r = .57$, 69.95% accuracy) — the figure a conventional validation would report — sits between these values on both metrics, concealing that alignment with the two lower-proficiency classes, together 38.8% of the population, is just over two-thirds as strong by rank correlation as alignment with the highest-performing class. Against a chance rate of 50% for forced binary comparisons, the accuracy gap is a larger share of the attainable signal than it first appears: the LLM's judgments are correct 22.9 percentage points above chance for the highest-proficiency class but only about 15 points above chance for the two lower classes, roughly a third less. This is the central empirical finding motivating our argument against single-benchmark validation.

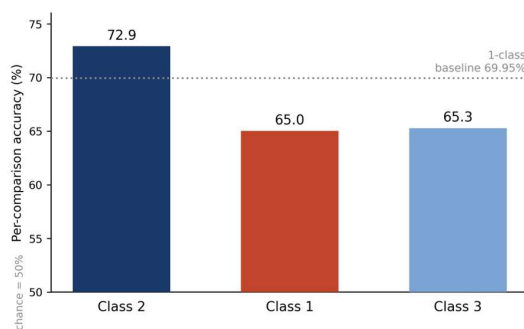


Figure 2. Comparison accuracy of LLM difficulty judgments against empirical item difficulty, by latent class (ordered by proficiency). The dotted line shows the single-benchmark (one-class) accuracy. The axis begins at the 50% chance rate for forced binary comparisons.

One nuance the three-class solution makes visible is that alignment is not monotone in proficiency. The intermediate class ($p = .48$) is tracked marginally less well than the lowest class ($p = .49$), and the two are indistinguishable at this item count. The finding is best characterized as the LLM tracking the highest-proficiency majority class well and both lower-proficiency classes comparably less well, rather than as a smooth decline across the ability range.

5 Conclusions and Limitations

This study showed that a single, population-aggregate correlation between LLM-predicted and empirical item difficulty can mask substantial heterogeneity across an examinee population's latent subgroups. Using a three-class mixture Rasch model of a national mathematics examination, we found that a locally hosted multimodal LLM's pairwise judgments aligned strongly with the highest-proficiency class ($p = .69$) but considerably less well with the intermediate and lowest-proficiency classes ($p = .48$ and $.49$), a gap the conventional benchmark ($p = .63$) obscures. The pattern held at the level of both aggregated rank correlation and individual pairwise judgments, and the classes were externally corroborated by intended collegiate field, instructional track, and not-reached behavior.

One plausible account for the differential alignment is that difficulty for high-proficiency examinees is closely tied to intrinsic item complexity — solution steps, abstraction,

conceptual depth an expert-style judge would readily recognize — while difficulty for lower-proficiency examinees is shaped by factors an LLM cannot access from item text or image alone: guessing, differential content exposure and other factors. A second, compatible account follows from Mislevy and Verhelst (1990): if the classes differ partly in solution strategy, then perhaps the LLM's predictions encode a particular model of how an examinee approaches an item, valid for the subpopulation whose approach it resembles.

The practical implication is clear. A single population-level benchmark cannot establish that a predictor is equally trustworthy for every subgroup a testing program serves. We recommend that LLM difficulty-prediction studies examine whether latent classes exist in their validation samples and, where they do, report alignment separately by latent class. Several limitations qualify these conclusions:

- **Item and sample exclusion criteria.** The item reduction and the removal of examinees with extreme scores or excessive missingness were principled but potentially consequential; since not-reached responses and extreme scores may relate to latent-class formation, these decisions could shape the composition of the lowest-proficiency class specifically. No formal sensitivity analysis varying these thresholds was conducted.
- **Single LLM, single pass.** Findings rest on a single LLM at one temperature in a single pass; the 13.3% reversal rate indicates instability even within that pass, and generalization to other LLMs is untested.
- **Single instrument and population.** All findings come from one mathematics subtest. The results may not generalize elsewhere.

Future work should extend this analysis to additional LLMs, instruments, and populations, and should pair class-specific alignment reporting with the identifiability diagnostics this study found necessary.

References

- Adegoke, H. O., Du, Y., & Dwyer, A. (2025). Medical item difficulty prediction using machine learning. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con), Works in Progress* (pp. 185–190). National Council on Measurement in Education.
- AlKhuyaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3), 862–914.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39(4), 331–348.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324–345.
- Bramley, T. (2010, June). Locating objects on a latent trait using Rasch analysis of experts' judgments. Paper presented at the conference “Probabilistic Models for Measurement in Education, Psychology, Social Science and Health,” Copenhagen, Denmark.
- Bramley, T. (2010, June). Locating objects on a latent trait using Rasch analysis of experts' judgments. In conference “Probabilistic Models for Measurement in Education, Psychology, Social Science and Health”, Copenhagen, Denmark]. Retrieved February (Vol. 20, p. 2026).
- Crompvoets, E. A., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, 45(3), 316–338.
- Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, 48(7), 1–25.
- Han, S., Rijmen, F., Boykin, A. A., & Lottridge, S. (2025). Leveraging fine-tuned large language models in item parameter prediction. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, Vol. 1: Full Papers (pp. 250–264). National Council on Measurement in Education.
- Kolesnikova, D., Fedyanin, K., Hofman, A. D., Brinkhuis, M. J., & Bolsinova, M. (2026). Estimating Item Difficulty with Large Language Models as Experts. *arXiv preprint arXiv:2605.18562*.
- Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item difficulty modeling using fine-tuned small and large language models. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, Vol. 3: Coordinated Session Papers (pp. 48–55). National Council on Measurement in Education.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2), 195–215.
- Mojoyinola, M., Kehinde, O. J., & Tang, J. (2025). Enhancing item difficulty prediction in large-scale assessment with large language model. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, Vol. 2: Works in Progress (pp. 218–222). National Council on Measurement in Education.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282.
- Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2025, December). Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (pp. 292–314).
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. In *Advances in Neural Information Processing Systems* (Vol. 36). Track on Datasets and Benchmarks.