The sparsity problem in benchmarks

Author: Benchmarks' Last Exam Lab 2026-11-10

Terminal-Bench-Science (TBS) is a benchmark, a set of tasks used to assess AI models. Its v0.1.0 release divides **70 tasks** into **20 fields**. Chemistry has two tasks, both about the three-dimensional shape of molecules and related geometric constraints. These tasks assess a model's ability to handle that kind of molecular computation. Assessing its ability to analyse chemical experiments or design reactions requires tasks that examine those abilities.

Names such as **Science and Last Exam** can suggest an assessment of an entire field. Readers may then treat a high score as evidence of broad ability. We examined the research content of all 70 TBS tasks to identify which topics have tasks and which do not.

This article has the same problem: it has a broad title, yet it examines only one typical benchmark. We all want to tell a story, but we hope this article also helps you see some of the deeper issues.

TBS pursues difficulty for its own sake

According to our internal assessment and what we have learned about TBS's development, contributors were repeatedly asked to make tasks harder. The standard was **whether a task could stump the strongest model available at the time**. New models kept appearing during development. Once a new model solved an existing task, that task could cease to meet the requirement, and maintainers would ask its contributor to raise the difficulty again. Earlier work could have to be discarded, leaving many contributors exhausted by repeated revisions. In our view, many of these changes pursued **difficulty for its own sake**, moving tasks away from real scientific problems. The central obstacles in many tasks would not arise in actual scientific discovery.

Large gaps remain within fields

TBS has 20 fields, and every field has tasks. To judge how thoroughly TBS tests a field, we also need to identify the specific research topics its tasks cover.

ANZSRC stands for the Australian and New Zealand Standard Research Classification. Its **Fields of Research** classification divides broad disciplines, such as chemistry and biology, into more specific research topics.

We selected **591 specific research topics** for the 20 TBS fields from this classification. Each task counts towards the topic that best describes its **main research problem**. A topic counts as covered if it has at least one task. Under these rules, TBS covers **53 topics, or about 9.0%**. The 591 topics are the comparison set for this analysis. TBS has not promised to test every topic in that set.

The fields in Figure 1 also have different numbers of topics in the comparison set. Operations research and optimisation has two, while medicine and health has 168. Coverage percentages depend on the classification and the topics selected for comparison. The percentages alone cannot tell us which field TBS tests more thoroughly.

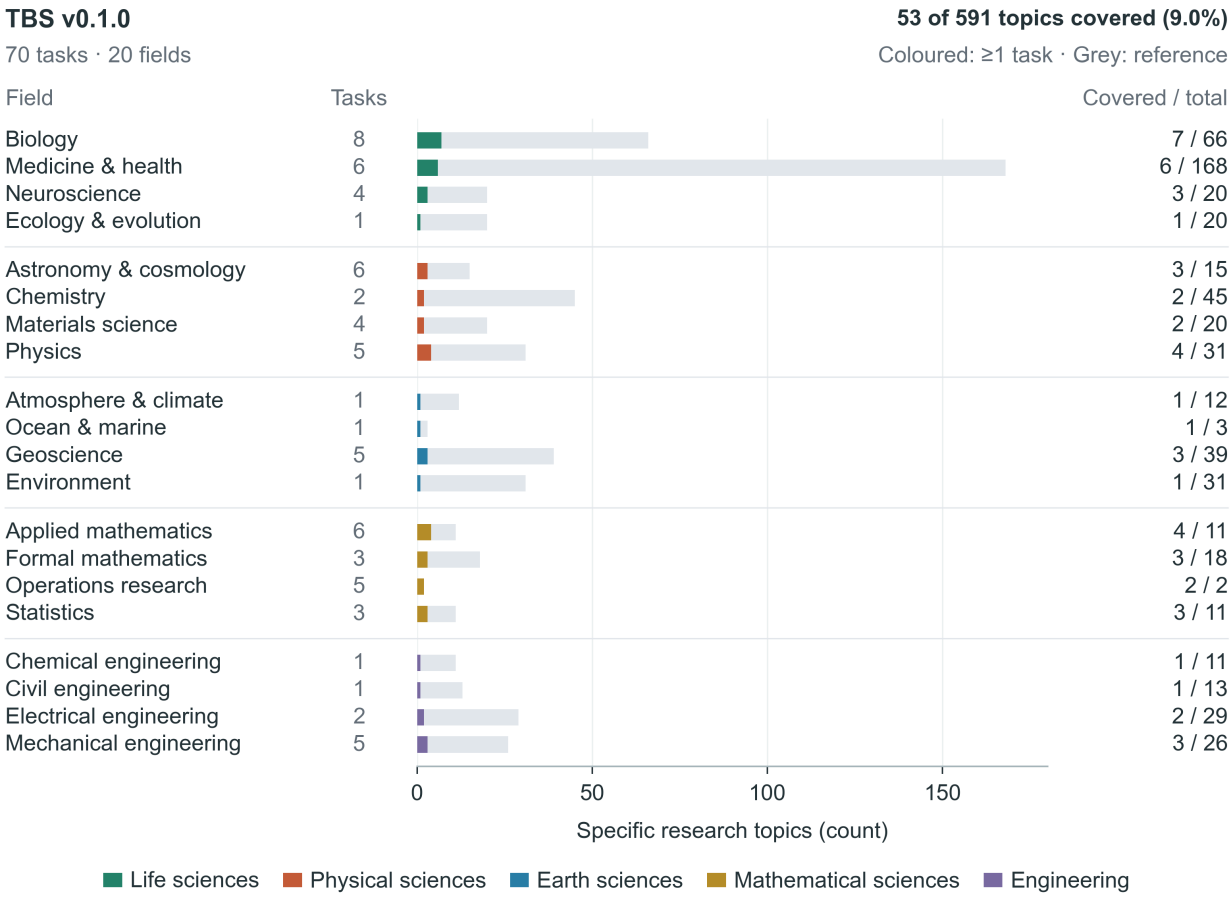


Figure 1 | Specific research topics covered by TBS. Grey bars show the number of topics selected for comparison in each field. Coloured bars show how many have at least one task. Task counts appear on the left, with covered and total topic counts on the right. Two tasks fall outside the selected set of 591 topics.

A small task set cannot represent an entire field

Of the 20 fields in TBS, **eight have only one or two tasks, or 40% of the fields**. Six have just one task. For these fields, the assessment provides evidence about how a model performs on those one or two tasks.

A model may complete a task because it is familiar with the tools or data involved. New tasks are needed to test whether the model can still succeed with different **data, given information, or objectives**. Even a difficult task, or one with several scientific steps, cannot replace those tests.

Task counts limit coverage

Of the 53 specific research topics covered by TBS, **41 have only one task, or 77.4%**.

Under the rule that each task counts towards one main research topic, 70 tasks can cover at most 70 topics. Even if every task tests a different topic, those tasks can cover only **11.8%** of the selected 591 topics. The current coverage of 9.0% therefore reflects both the choice of tasks and the size of the task set.

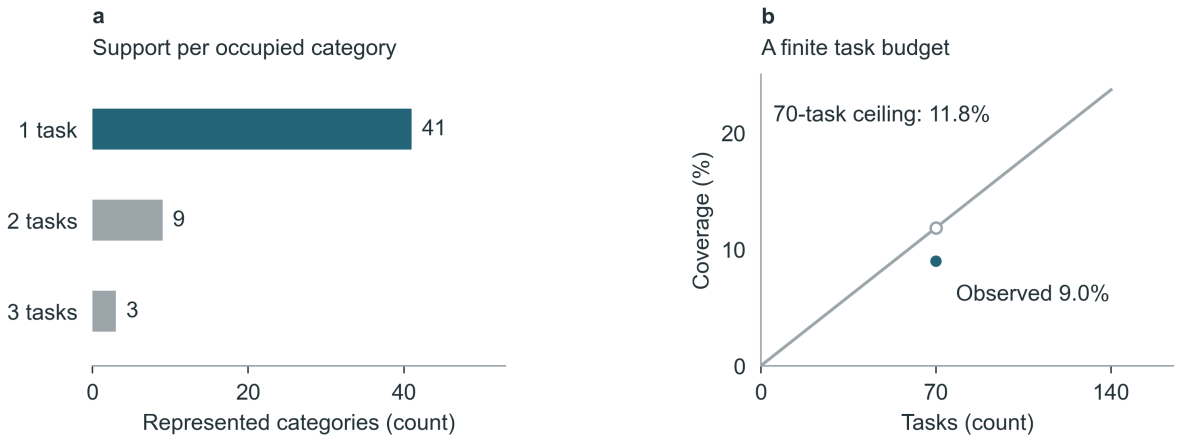


Figure 2 | Task counts limit coverage. Panel a counts topics with one, two, or three tasks among the 53 topics covered by TBS. Panel b shows the coverage ceiling when each task counts towards one main topic. The filled point marks the current coverage of TBS.

Research workflows focus on computation

TBS has 20 field labels, but tasks in different fields can still use similar **workflows**. We reviewed all 70 tasks' instructions, reference solutions, and grading code.

61 of the 70 official reference solutions use Python for calculations, or 87.1%. Some also use other languages or specialist software. Three depend on prior human annotations or image registration. The instructions for **22 tasks** require a Python program, module, or reproduction script. Python is the benchmark's most commonly used computational tool. Many TBS tasks are essentially software engineering work.

By main goal, 23 tasks focus on parameter estimation, model identification, or statistical inference. Another 11 focus on prediction, classification, or ranking, and 10 on image and signal processing. **Together they account for 44 tasks, or 62.9%.** These tasks typically provide data and a defined problem, then grade the resulting parameters, predictions, images, or code.

Three of the five mechanical-engineering tasks **locate damage** using waves travelling through a structure. Their reference solutions compare sensor recordings with healthy or reconstructed signals, calculate damage indicators along each path, and combine them into a location map. The reference comes from an excitation signal defined by a formula, measurements of an undamaged structure, or structural simulations.

Four tasks require mathematical proofs written and checked in **Lean**, a proof tool. Interactive tasks include **protein active learning**, which requires three rounds of experiment selection and updates predictions with each round's measurements.

Task quality is under scrutiny

Our Terminal-Bench-Science Review Report found defects in **34 of 70 tasks**, unresolved concerns in **five**, and no substantive defects in **31**. One medical task wrongly labelled two samples as inflamed, penalising correct answers based on the pathology records.

Horizon found two tasks that scored correct answers zero for unstated requirements. It counted 21 reviewed tasks after excluding seven with already public problems. A public issue also reported worse reference results after further route optimisation, with possible grading effects.

Counting only the 31 tasks with no substantive defects identified reduces coverage from 20 to 17 fields and from 53 to 29 specific topics.

The classification can guide training and evaluation

This subfield classification can help plan **the mix of training data**. During **pretraining**, models learn from large collections of text, code, and other material. Developers can use the categories to check whether that material matches the model's intended uses. During **post-training**, they refine the model with targeted tasks and feedback, and can organise those tasks by subfield and solution method. The proportions still need adjustment based on data quality and performance after training. For evaluation, the same classification can guide new task selection and results by subfield, helping explain changes in the overall score and whether an assessment matches a researcher's needs.

Conclusion

TBS has 70 tasks across 20 fields. Within the classification used here, those tasks cover a small share of the specific research topics. Most covered topics have just one task. The work also focuses on computation, and some tasks follow similar workflows.

With **reliable grading**, a high TBS score shows that a model performs well on this set of tasks. Extending that result to science as a whole requires testing whether the model can handle research topics and ways of working **outside the benchmark**.

Reference

[1] Terminal-Bench-Science v0.1.0 release inventory and task descriptions, revision f81afac.

[2] This analysis's classification and counts for all 70 tasks, using the Fields of Research component of ANZSRC 2020, revised 24 October 2025. Each task has one primary category; catch-all categories for work not classified elsewhere are excluded. The 53/591 and 77.4% figures and reference comparisons are results of this analysis, not official TBS coverage statistics.

[3] Benchmarks' Last Exam, Terminal-Bench-Science Review Report, 22 October 2026, updated 3 November 2026. Reviewed revision: f81afac.

[4] Benchmarks' Last Exam, review of solution methods across all 70 TBS v0.1.0 tasks, 10 November 2026, revision f81afac.

[5] Horizon, Preliminary Data Quality Index, Terminal-Bench Science entry. Accessed 10 November 2026.

[6] GitHub user afterquery-official, energy-routing issue #2709, official Terminal-Bench-Science repository, 8 November 2026.

[7] Xie et al., DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining, NeurIPS 2023.

[8] Lambert et al., Tulu 3: Pushing Frontiers in Open Language Model Post-Training, 2024, arXiv:2411.15124.

Contact Us

benchmarks-last-exam.com

Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.