

1 Datasheet for the CS-Bench Dataset

2 Paper ID: 1729

3 This is a datasheet for the CS-Bench dataset in Submission 1729. The format of this datasheet was
4 introduced in [1] and consolidates the motivation, creation process, composition, and intended uses
5 of our dataset as a series of questions and answers.

6 The dataset can be accessed via the following GitHub and Hugging Face links:

7 <https://github.com/csbench/csbench/tree/main/Dataset>

8 <https://huggingface.co/datasets/CS-Bench/CS-Bench>.

9 We present the Croissant metadata in [https://github.com/csbench/csbench/blob/main/
10 Dataset/croissant.json](https://github.com/csbench/csbench/blob/main/Dataset/croissant.json).

11 The experimental code revolving around CS-Bench can be accessed at [https://github.com/
12 csbench/csbench/](https://github.com/csbench/csbench/).

13 The authors statement that they bear all responsibility in case of violation of rights, etc., and
14 confirmation of the data license.

15 1 Motivation

16 **Q1. For what purpose was the dataset created?** *Was there a specific task in mind? Was there
17 a specific gap that needed to be filled? Please provide a description.*

18 Understanding the performance of Large Language Models (LLMs) in computer science
19 (CS) is fundamental to the research and application of LLMs within the field. However,
20 existing benchmarks only consider computer science as a minor category within science
21 and engineering, lacking specialized evaluation and in-deep analysis tailored specifically
22 to computer science. Moreover, considering the intersection of computer science and code
23 programming, mathematics, and reasoning abilities, we have grounds to believe that cross-
24 capability research and analysis in CS can effectively propel the comprehensive development
25 of the LLM community.

26 **Q2. Who created the dataset (e.g., which team, research group) and on behalf of which
27 entity (e.g., company, institution, organization)?**

28 The authors of this work, the PRIS-NLP group from Beijing University of Posts and
29 Telecommunications in China, created this dataset.

30 **Q3. Who funded the creation of the dataset?** *If there is an associated grant, please provide
31 the name of the grantor and the grant name and number.*

32 There is no associated grant or funding which has been used to create the CS-Bench.

33 **Q4. Any other comments?**

34 No.

35 2 Composition

36 **Q5. What do the instances that comprise the dataset represent (e.g., documents, photos,
37 people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings;
38 people and interactions between them; nodes and edges)? Please provide a description.*

39 Our dataset consists of computer science problems represented in natural language. Each
40 instance contains a computer science question, its corresponding domain, the type of task,
41 the type of capability required, and the correct answer. Additionally, we provide answer
42 explanations for reasoning-type questions. Detailed examples can be found in Appendix
43 C.3.

44 **Q6. How many instances are there in total (of each type, if appropriate)?**

45 CS-Bench is a bilingual assessment benchmark supporting both Chinese and English evalua-
46 tions, containing a total of 4838 entries, where Chinese and English questions correspond
47 one-to-one, each comprising 2419 instances. CS-Bench is divided into four domains: Op-
48 erating Systems (OS) contains 1082 instances, Data Structures and Algorithms (DSA)
49 contains 1198 instances, Computer Networks (CN) contains 1314 instances, and Computer
50 Organization (CO) contains 1244 instances. Detailed information can be found in Appendix
51 C.1.

52 **Q7. Does the dataset contain all possible instances or is it a sample (not necessarily random)**
53 **of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the*
54 *sample representative of the larger set (e.g., geographic coverage)? If so, please describe*
55 *how this representativeness was validated/verified. If it is not representative of the larger set,*
56 *please describe why not (e.g., to cover a more diverse range of instances, because instances*
57 *were withheld or unavailable).*

58 Given that computer science is a vast and ever-evolving field, CS-Bench cannot cover all
59 areas of CS. CS-Bench encompasses 26 subfields within the four key domains of CS and
60 primarily includes computer science problems from university courses.

61 **Q8. What data does each instance consist of?** *“Raw” data (e.g., unprocessed text or images)*
62 *or features? In either case, please provide a description.*

63 Each instance contains a computer science question along with its answer. The instances are
64 provided in both Chinese and English, and are formatted as text strings in JSON.

65 **Q9. Is there a label or target associated with each instance?** *If so, please provide a description.*

66 Each instance has multiple labels, including domain, fine-grained subfield, task format, and
67 whether it belongs to knowledge-type or reasoning-type categories.

68 **Q10. Is any information missing from individual instances?** *If so, please provide a description,*
69 *explaining why this information is missing (e.g., because it was unavailable). This does not*
70 *include intentionally removed information, but might include, e.g., redacted text.*

71 No.

72 **Q11. Are relationships between individual instances made explicit (e.g., users’ movie ratings,**
73 **social network links)?** *If so, please describe how these relationships are made explicit.*

74 Each instance is independent and not related to other instances.

75 **Q12. Are there recommended data splits (e.g., training, development/validation, testing)?** *If*
76 *so, please provide a description of these splits, explaining the rationale behind them.*

77 CS-Bench is primarily used to evaluate LLMs’ performance in computer science and analyze
78 cross-capabilities of LLMs, so it does not include training data. In CS-Bench, 10% of the
79 data is randomly selected for validation to verify LLMs’ effectiveness during training, while
80 the remaining 90% is used for evaluating LLMs.

81 **Q13. Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please*
82 *provide a description.*

83 We performed manual cross-checking on the dataset to minimize the possibility of errors.
84 For fill-in-the-blank and open-ended questions where the answers may not be unique, we
85 provide one correct reference answer.

86 **Q14. Is the dataset self-contained, or does it link to or otherwise rely on external resources**
87 **(e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are*
88 *there guarantees that they will exist, and remain constant, over time; b) are there official*
89 *archival versions of the complete dataset (i.e., including the external resources as they*
90 *existed at the time the dataset was created); c) are there any restrictions (e.g., licenses,*
91 *fees) associated with any of the external resources that might apply to a dataset consumer?*
92 *Please provide descriptions of all external resources and any restrictions associated with*
93 *them, as well as links or other access points, as appropriate.*

94 The dataset is self-contained.

- 95 **Q15. Does the dataset contain data that might be considered confidential (e.g., data that is**
96 **protected by legal privilege or by doctor– patient confidentiality, data that includes the**
97 **content of individuals’ non-public communications)?** *If so, please provide a description.*
98 **No.**
- 99 **Q16. Does the dataset contain data that, if viewed directly, might be offensive, insulting,**
100 **threatening, or might otherwise cause anxiety?** *If so, please describe why.*
101 **No.**
- 102 **Q17. Does the dataset relate to people?**
103 **No.**
- 104 **Q18. Any other comments?**
105 **No.**

106 **3 Collection Process**

- 107 **Q19. How was the data associated with each instance acquired?** *Was the data directly*
108 *observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or*
109 *indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses*
110 *for age or language)? If the data was reported by subjects or indirectly inferred/derived*
111 *from other data, was the data validated/verified? If so, please describe how.*
112 **Our data mainly comes from three sources:** 1. Public channels providing computer science-
113 related questions, such as professional exams and practice tests (e.g., <https://github.com/CodePanda66/CSPostgraduate-408>, <https://github.com/ddy-ddy/cs-408>
114). 2. Knowledge-based questions manually extracted and adapted by professionals
115 from various academic-permitted CS-related blog posts (e.g., <https://www.wikipedia.org/>, <https://www.cnblogs.com/>, <https://www.csdn.net/>, <https://zhuanlan.zhihu.com/>). 3. Questions constructed from non-public teaching materials and exam papers
116 authorized by the author’s affiliated institution, namely Beijing University of Posts and
117 Telecommunications.
118
- 119 **Q20. What mechanisms or procedures were used to collect the data (e.g., hardware appara-**
120 **tuses or sensors, manual human curation, software programs, software APIs)?** *How*
121 *were these mechanisms or procedures validated?*
122 **For resources sourced from the internet, we manually extract the knowledge points and**
123 **questions. For physical materials, we use Optical Character Recognition (OCR) to obtain**
124 **the data. We then ensure the accuracy of the collected data through manual cross-checking.**
125
- 126 **Q21. If the dataset is a sample from a larger set, what was the sampling strategy (e.g.,**
127 **deterministic, probabilistic with specific sampling probabilities)?**
128 **No.**
129
- 130 **Q22. Who was involved in the data collection process (e.g., students, crowdworkers, contrac-**
131 **tors) and how were they compensated (e.g., how much were crowdworkers paid)?**
132 **The data collection was carried out by a team of five students with bachelor’s degrees in**
133 **computer science. We paid each of them an hourly wage of 50 CNY (approximately 7 USD),**
134 **which is higher than the local (Beijing, China.) minimum hourly wage standard of 26.4**
135 **CNY (approximately 3.7 USD).**
- 136 **Q23. Over what timeframe was the data collected?** *Does this timeframe match the creation*
137 *timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?*
138 *If not, please describe the timeframe in which the data associated with the instances was*
139 *created.*
140 **Both the data collection and the dataset creation were completed in January 2024.**
- 141 **Q24. Were any ethical review processes conducted (e.g., by an institutional review board)?**
142 *If so, please provide a description of these review processes, including the outcomes, as well*
143 *as a link or other access point to any supporting documentation.*

144

No.

145

Q25. Any other comments?

146

No.

147

4 Preprocessing/cleaning/labeling

148

Q26. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

149

150

151

152

153

154

155

Based on whether the questions require in-depth reasoning and computation, we label each question as either knowledge-type or reasoning-type. Additionally, we tag each instance with domain and task type labels. For English data, we used GPT-4 to translate the Chinese instances into English instances, followed by manual verification.

156

Q27. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

157

158

159

No.

160

Q28. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

161

162

163

164

In data processing, we used tools including OCR (https://developers.weixin.qq.com/doc/offiaccount/Intelligent_Interface/OCR.html) and GPT-4 (<https://openai.com/index/gpt-4>)

165

Q29. Any other comments?

166

No.

167

5 Uses

168

Q30. Has the dataset been used for any tasks already? If so, please provide a description.

169

170

171

In this paper, CS-Bench has been used to evaluate the performance of mainstream LLMs in computer science and to explore the relationships between the capabilities of LLMs (math, code, CS).

172

Q31. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

173

174

175

Future work citing the CS-Bench dataset will be listed by citation trackers such as Google Scholar and Semantic Scholar.

176

Q32. What (other) tasks could the dataset be used for?

177

178

179

CS-Bench can also be used to validate the ability of AI tools, not limited to LLMs, to solve CS tasks. Additionally, we anticipate that CS-Bench can be used as training data for fine-tuning LLMs.

180

Q33. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

181

182

183

184

185

186

No.

187

Q34. Are there tasks for which the dataset should not be used? If so, please provide a description.

188

189

No.

190 **Q35. Any other comments?**

191 No.

192 **6 Distribution**

193 **Q36. Will the dataset be distributed to third parties outside of the entity (e.g., company,**
194 **institution, organization) on behalf of which the dataset was created? If so, please**
195 **provide a description.**

196 Yes, the dataset is freely and publicly available and accessible.

197 **Q37. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does**
198 **the dataset have a digital object identifier (DOI)?**

199 We distribute our datasets on GitHub and Hugging Face, as follows: [https://](https://github.com/csbench/csbench/tree/main/Dataset)
200 github.com/csbench/csbench/tree/main/Dataset, [https://huggingface.co/](https://huggingface.co/datasets/CS-Bench/CS-Bench)
201 [datasets/CS-Bench/CS-Bench](https://huggingface.co/datasets/CS-Bench/CS-Bench).

202 **Q38. When will the dataset be distributed?**

203 The dataset has been available since June 12, 2024.

204 **Q39. Will the dataset be distributed under a copyright or other intellectual property (IP)**
205 **license, and/or under applicable terms of use (ToU)? If so, please describe this license**
206 **and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant**
207 **licensing terms or ToU, as well as any fees associated with these restrictions.**

208 The dataset is licensed under CCBY-NC 4.0.

209 **Q40. Have any third parties imposed IP-based or other restrictions on the data associated**
210 **with the instances? If so, please describe these restrictions, and provide a link or other**
211 **access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees**
212 **associated with these restrictions.**

213 No.

214 **Q41. Do any export controls or other regulatory restrictions apply to the dataset or to**
215 **individual instances? If so, please describe these restrictions, and provide a link or other**
216 **access point to, or otherwise reproduce, any supporting documentation.**

217 No.

218 **Q42. Any other comments?**

219 No.

220 **7 Maintenance**

221 **Q43. Who will be supporting/hosting/maintaining the dataset?**

222 The dataset will be maintained by the PRIS-NLP group at Beijing University of Posts and
223 Telecommunications in China.

224 **Q44. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

225 The maintainers can be contacted via email: csbench2024@gmail.com, [songxi-](mailto:songxi-aoshuai@bupt.edu.cn)
226 [aoshuai@bupt.edu.cn](mailto:songxi-aoshuai@bupt.edu.cn), dmx@bupt.edu.cn

227 **Q45. Is there an erratum? If so, please provide a link or other access point.**

228 Currently, there is no erratum. If errors are encountered, the dataset will be updated with a
229 fresh version. They will all be provided in the same github repository.

230 **Q46. Will the dataset be updated (e.g., to correct labeling errors, addnew instances, delete in-**
231 **stances)? If so, please describe how often, bywhom, and how updates will be communicated**
232 **to dataset consumers(e.g., mailing list, GitHub)?**

233 Yes, future changes will be documented in the README file of the GitHub repository:
234 <https://github.com/csbench/csbench>. Differences in single files can be tracked in
235 the Git history.

236 **Q47. If the dataset relates to people, are there applicable limits on the retention of the data**
237 **associated with the instances (e.g., were the individuals in question told that their data**
238 **would be retained for a fixed period of time and then deleted)?** *If so, please describe*
239 *these limits and explain how they will be enforced.*

240 **No.**

241 **Q48. Will older versions of the dataset continue to be supported/hosted/maintained? If so,**
242 **please describe how. If not, please describe how its obsolescence will be communicated**
243 **to dataset consumers. If others want to extend/augment/build on/contribute to the**
244 **dataset, is there a mechanism for them to do so?** *If so, please provide a description.*
245 *Will these contributions be validated/verified? If so, please describe how. If not, why not? Is*
246 *there a process for communicating/distributing these contributions to dataset consumers? If*
247 *so, please provide a description.*

248 **Yes, older versions of the benchmark will be maintained in the GitHub repository:** <https://github.com/csbench/csbench>.
249

250 **Q49. If others want to extend/augment/build on/contribute to the dataset, is there a mech-**
251 **anism for them to do so?** *If so, please provide a description. Will these contributions*
252 *be validated/verified? If so, please describe how. If not, why not? Is there a process for*
253 *communicating/distributing these contributions to dataset consumers? If so, please provide*
254 *a description.*

255 **Any potential contributors are strongly encouraged to our dataset through contacting the**
256 **authors of the paper.**

257 **Q50. Any other comments?**

258 **No.**

259 **References**

260 [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
261 Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*,
262 64(12):86–92, 2021.