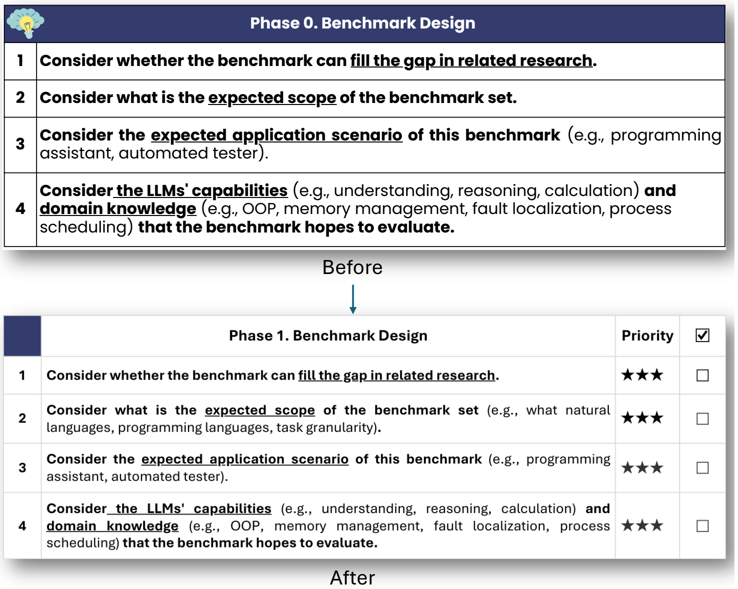
Explanation Of Revisions PDF

We made the following major changes in response to our previous round of reviews.

I. Establish clear indicators showing the generalizability and priority of each guideline.

According to previous suggestion, we revisited the 55 guidelines again, enriching the guidelines by adding two new dimensions as follows. The major updates lie in Sections 3.2, 4, and Appendix A and F.

- *Priority* -- indicating the importance. The more stars, the more important:
  - ★★★ (Highly recommended)
  - ★★ (Recommended)
  - ★ (Optional)
- *Generalizability* -- indicating whether it can be generalized to other types of benchmarks:
  - ✓ (Can generalize to other types of benchmarks)
  - ☐ (Hard to generalize to other types of benchmarks)



	Phase 0. Benchmark Design
1	**Consider whether the benchmark can fill the gap in related research.**
2	**Consider what is the expected scope of the benchmark set.**
3	**Consider the expected application scenario of this benchmark** (e.g., programming assistant, automated tester).
4	**Consider the LLMs' capabilities** (e.g., understanding, reasoning, calculation) **and domain knowledge** (e.g., OOP, memory management, fault localization, process scheduling) **that the benchmark hopes to evaluate.**

Before

↓

	Phase 1. Benchmark Design	**Priority**	☑
1	**Consider whether the benchmark can fill the gap in related research.**	★★★	☐
2	**Consider what is the expected scope of the benchmark set** (e.g., what natural languages, programming languages, task granularity)**.**	★★★	☐
3	**Consider the expected application scenario of this benchmark** (e.g., programming assistant, automated tester).	★★★	☐
4	**Consider the LLMs' capabilities** (e.g., understanding, reasoning, calculation) **and domain knowledge** (e.g., OOP, memory management, fault localization, process scheduling) **that the benchmark hopes to evaluate.**	★★★	☐

After

Figure 1Illustration of How2Bench Checklist Changes

II. Emphasize the novelty.

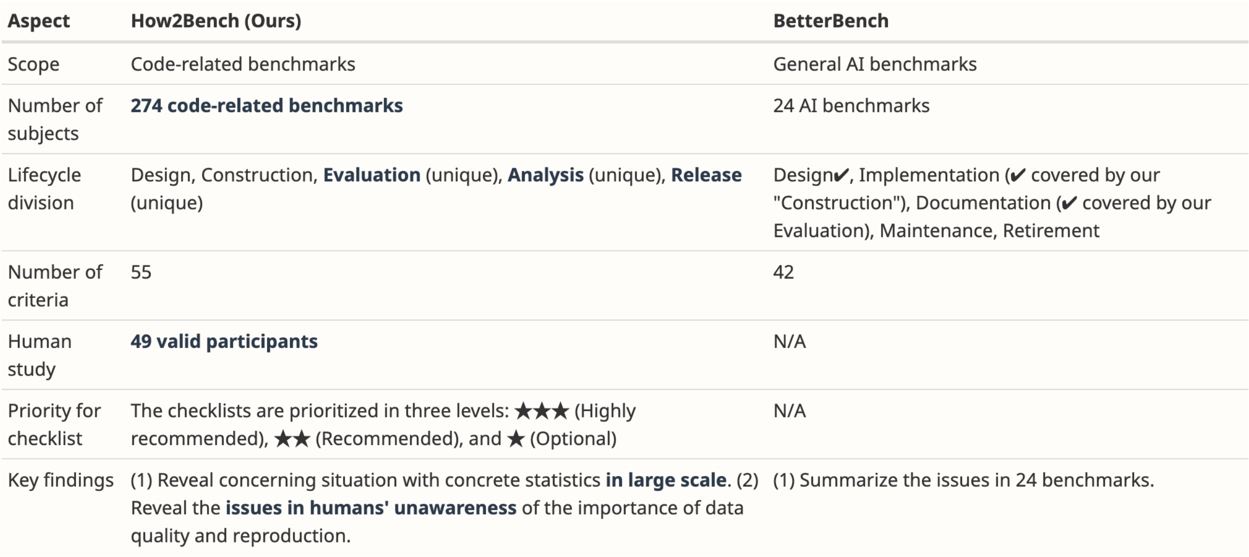
According to previous suggestion, we emphasize our novelty as follows. The major updates lies in Sections 2.

First, we distinguished our How2Bench from BetterBench in Section 2.2:

*(Section 2.2, lines 195 - 209) Recently, BetterBench (Reuel et al., 2024) is a concurrent work assessing the AI benchmarks against 46 criteria. Then, it scored 24 AI benchmarks in various domains and ranked them. BetterBench differs from this paper in several key aspects: **scope***

*(general benchmarks vs. code-related benchmarks), **lifecycle division** (it addresses benchmark retirement, while How2Bench focuses on benchmark evaluation, analysis, and release), and **objectives** (scoring benchmarks vs. offering comprehensive guidelines for future benchmark development). Additionally, the study in this paper was conducted on **a much larger scale (24 vs. 274 benchmarks)**, statistically highlighting the prevalent issues in existing benchmarks.*

Second, to better show the difference, we compared with BetterBench in the following aspects:



Aspect	**How2Bench (Ours)**	**BetterBench**
Scope	Code-related benchmarks	General AI benchmarks
Number of subjects	**274 code-related benchmarks**	24 AI benchmarks
Lifecycle division	Design, Construction, **Evaluation** (unique), **Analysis** (unique), **Release** (unique)	Design✔, Implementation (✔ covered by our "Construction"), Documentation (✔ covered by our Evaluation), Maintenance, Retirement
Number of criteria	55	42
Human study	**49 valid participants**	N/A
Priority for checklist	The checklists are prioritized in three levels: ★★★ (Highly recommended), ★★ (Recommended), and ★ (Optional)	N/A
Key findings	(1) Reveal concerning situation with concrete statistics **in large scale**. (2) Reveal the **issues in humans' unawareness** of the importance of data quality and reproduction.	(1) Summarize the issues in 24 benchmarks.

Third, unlike BetterBench (Reuel et al., 2024), which proposed guidelines and scored existing benchmarks, we went beyond that. We not only characterized the severe situation quantitatively on a large scale (274 vs 24), but also delved deeper into understanding the underlying root causes. Therefore, we conducted a human study, interviewing 50 experienced practitioners to explore whether discrepancies in human cognition and behavior are synchronized. Consequently, we discovered that the issues of poor reproducibility and low data quality in the current benchmarking landscape are, in fact, due to gaps in awareness among individuals. For instance, 16% of respondents were unaware of potential noise in the data, and 40% did not recognize the importance of providing necessary information for reproducibility.

III. Emphasize the significance and the contribution.

According to the suggestions, we emphasize the significance and contribution of this work in the presentation with more concrete statements. <u>The major updates lie in Sections 1, 2 and 5.</u>

- ***Non-trivial efforts and comprehensive statistics***: We conducted a thorough survey over 274 benchmarks, interviewed 50 participants (49 valid), and closely examined more than 1 thousand subjects from 30 benchmarks in five tasks in order to characterize the current situation quantitatively. The comprehensive statistics over 274 benchmarks alone is a non-trivial contribution, let alone the guideline proposal/refinement/explanation and the human study.

- ***In-depth analysis***: Unlike BetterBench (Reuel et al., 2024) which proposed guidelines and scored existing benchmarks, we went beyond that. After characterizing the severe situation quantitatively in a large scale (274 vs 24), we aim to delve deeper into understanding the underlying root causes. Therefore, we conducted a human study, interviewing 50 experienced practitioners to explore whether discrepancies in human cognition and behavior are synchronized. Consequently, we discovered that the issues of poor reproducibility and low data quality in the current benchmarking landscape are, in fact, due to gaps in awareness among individuals. For instance, 16% of respondents were unaware of potential noise in the data, and 40% did not recognize the importance of providing necessary information for reproducibility.
- ***Revealing issues that have not been revealed before***: We reported six issues that have not been reported before (Figure 17, Figure 20, Figure 22, Figure 25, Figure 45, and Figure 46). If we haven't closely dive into these subjects in the benchmarks, how can we identify these previously unreported issues via "simple basic features"?