Manuscript Suggestions

AI Benchmark Carpentry and Democratization

Renato Umeton

Total Suggestions	High	Moderate	Low
79	**12**	**27**	**40**

#	Sev.	Category	Manuscript Text	Suggested Edit	Reason
HIGH SEVERITY SUGGESTIONS (12)					
1	HIGH	Factual Error Table VI	‘‘Frontier[160] 2021 Hybrid CPU/GPU’’ **Page 24, Table VI**	Change year from “2021” to “2022”	*Frontier became operational and topped TOP500 in May 2022, not 2021.*
2	HIGH	Contradiction Taxonomy	‘‘We distinguish two different data sets: static and dynamic’’ ...followed by discussion of ‘‘living datasets’’ **Page 7**	Revise to: “We distinguish three categories of datasets: static, dynamic, and living”	*Text claims two types but introduces and extensively discusses a third type.*
3	HIGH	Missing Section Structure	‘‘In Section IV, we summarize... In Section VI, we define activities...’’ **Page 4**	Add: “In Section V, we address sharing benchmarks and FAIR principles.”	*Section V is omitted from paper roadmap despite being in Table of Contents.*
4	HIGH	Duplicate Ref References	[20]: ‘‘W.-C. Feng and K. W. Cameron...’’ AND [167]: ‘‘W.-c. Feng and K. Cameron...’’ **References**	Consolidate references [20] and [167] into single entry.	*Same Green500 paper cited twice with minor formatting differences.*
5	HIGH	Incomplete Ref References	‘‘[75] TBD, Aime, [Online accessed 2025-06-24]’’ **Page 34**	Complete with actual authors and publication details.	*“TBD” as authors indicates unfinished reference.*
6	HIGH	Incomplete Contributions	‘‘x Gary Mazzaferro garym@oedata.com TBD’’ **Page 45, Appendix B**	Complete contribution description or remove author.	*Author contribution marked “TBD” is inappropriate for submission.*
7	HIGH	Spec Gap Formalization	‘‘In more complex situations... we may use W instead of T’’ **Page 7**	Formally incorporate W into specification as B = (I, D, T—W, M, C, R).	*Workflows introduced but never integrated into formal definition.*
8	HIGH	No Citation Claims	‘‘Large language models... known for their memorization of static benchmarks’’ **Pages 2–3**	Add citation(s) on benchmark contamination or data leakage in LLMs.	*Central claim lacks supporting reference.*
9	HIGH	Set Error Formalization	‘‘Cc | c ∈ {B, I, D, T, M, R, A}’’ **Page 6**	Change to “Cc — c ∈ {B, I, D, T, M, R}”	*A is part of T = (A, P), not a top-level component.*
10	HIGH	Citation Mismatch References	‘‘MLCommons [24] provides one of the most comprehensive...’’ **Page 9**	Change [24] to proper MLCommons organizational reference.	*Reference [24] is about MLPerf HPC, not MLCommons organization.*
11	HIGH	Undefined Formalization	‘‘min{Bi(..., M, ...)(Sj) | ∀jM(Sj)}’’ **Page 6**	Define Sj explicitly and clarify minimization expression.	*Sj undefined; notation is ambiguous and potentially malformed.*
12	HIGH	Contradiction Cataloging	‘‘we have catalogued... all MLCommons benchmarks that have a result submission’’ **Page 9**	Revise to include planned/proposed benchmarks in description.	*Tables include “planned,” “in development,” and GPT-5/6 projections.*
MODERATE SEVERITY SUGGESTIONS (27)					
13	MOD	Typo Grammar	‘‘a metric that is to bi minimized’’ **Page 6**	Change “bi” to “be”	*Typographical error in formal specification.*
14	MOD	Typo Grammar	‘‘denote a benchamrk with a fixed metric’’ **Page 6**	Change “benchamrk” to “benchmark”	*Typographical error in key definition.*

Contd on next page...

#	Sev.	Category	Manuscript Text	Suggested Edit	Reason
15	MOD	Typo Grammar	`‘‘to actually fond the best algorithmic solution’’` **Page 6**	Change “fond” to “find”	*Typographical error in key sentence.*
16	MOD	Duplicate Grammar	`‘‘limited limited root access on many HPC systems’’` **Page 21**	Remove one “limited”	*Duplicate word, likely copy-paste error.*
17	MOD	Missing Plural Grammar	`‘‘on HPC system we find that’’` **Page 21**	Change “system” to “systems”	*Missing plural form.*
18	MOD	Incomplete Grammar	`‘‘such as data carpentry [9, 10] Together, this includes:’’` **Page 5**	Add period and verb: “...data carpentry [9, 10]. Together, these efforts include:”	*Grammatically incomplete; missing verb and period.*
19	MOD	Redundancy Grammar	`‘‘skills akin analogous to a hammer’’` **Page 5**	Use either “akin to” OR “analogous to”	*“Akin” and “analogous” are synonyms; using both is redundant.*
20	MOD	Undefined Terminology	`‘‘as formalized by CASP’’` **Page 7**	Add: “CASP (Critical Assessment of protein Structure Prediction)”	*CASP undefined; not in Appendix A.*
21	MOD	Notation Formalization	`‘‘c = (c, Cc)	c ∈ {B, I, D, T, M, R, A}’’` **Page 6**	Use different variable: “x = (x, Cx) — x ∈ ...”	*Using “c” on both sides creates confusion.*
22	MOD	Historical Accuracy	`‘‘a global community effort has sprung up since 2018 [8]’’` **Page 5**	Clarify: The Carpentries merger was 2018; Software Carpentry dates to 1998.	*Software Carpentry existed before 2018; statement is misleading.*
23	MOD	Misleading Data	`Table VI: Titan ‘‘27’’ PF Peak Performance` **Page 24**	Add footnote: “Peak theoretical; Linpack ratings differ”	*Titan’s Linpack was 17.59 PF; 27 PF is theoretical peak.*
24	MOD	Temporal Consistency	`‘‘As of Oct 1, 2025, we find 106 entries’’` **Page 8**	Establish consistent “as of” date for all time-sensitive data.	*Paper mixes data from different time points without anchoring.*
25	MOD	Contradictor Consistency	`‘‘static... is to be preferred’’ vs. ‘‘continuous adaptive benchmarking frameworks’’` **Pages 7, 3**	Add: “Static preferred when applicable; dynamic essential for evolving domains.”	*Tension between advocacy for dynamic and preference for static.*
26	MOD	Imprecise Accuracy	`‘‘gem5 currently focuses its support on AMD GPUs’’` **Page 27**	Revise: “...emphasizes AMD GPUs, with additional ARM GPU support [212]”	*Text later mentions ARM support, contradicting exclusive AMD focus.*
27	MOD	Naming Terminology	`‘‘MLCommons Science & HPC Working Group’’ vs. ‘‘MLCommons Science Working Group’’` **Throughout**	Standardize to one official name.	*Inconsistent naming may confuse readers.*
28	MOD	Grammar Acknowledgments	`‘‘grammar of selected section’’` **Page 30**	Change to “selected sections”	*Missing plural form.*
29	MOD	Acronym Terminology	`‘‘FAIR: Apply the fair principle’’` **Page 29**	Change to “FAIR principles” (capitalized, plural)	*Should be capitalized acronym; four principles exist.*
30	MOD	Citation References	`Table V caption cites [151,149,152,153,154,155] but entry uses [156,149]` **Page 24**	Reconcile caption and in-table citations.	*Reference [156] in table but not in caption list.*
31	MOD	Unverified Technical	`‘‘Accel-Sim... (Volta through Blackwell) GPUs’’` **Page 27**	Verify Blackwell support or revise to “Volta through Hopper/Ada”	*Blackwell is very recent (2024); support should be verified.*
32	MOD	Context Clarity	`‘‘Two of them are the Compute Coordinator and the Experiment Executer’’` **Page 21**	Add: “Two workflow frameworks developed by the authors [79] are...”	*Tools appear without adequate context.*
33	MOD	Statistics Clarity	`‘‘8% on average (max 22%) with outliers up to 1.5× slower’’` **Page 23**	Clarify relationship between 22% max and 1.5× outliers.	*Relationship between statistics is confusing.*

Contd on next page...

#	Sev.	Category	Manuscript Text	Suggested Edit	Reason
34	MOD	Methodology Completeness	‘‘106 entries on arXiv... 2,490 entries for Google Scholar’’ **Page 8**	Add methodology: search terms, date ranges, explanation of difference.	*Numbers provided without methodology; unreproducible.*
35	MOD	Citation References	‘‘Trillion Parameter Consortium [1]’’ **Page 3**	Add proper TPC reference separate from AuroraGPT presentation.	*Reference [1] is about AuroraGPT, not TPC itself.*
36	MOD	Missing Abbreviations	Appendix A omits: CASP, BSP, SKU, Pfac, PIT **Pages 42–44**	Add missing abbreviations to Appendix A.	*Several acronyms used in paper not defined in appendix.*
37	MOD	Imbalance Structure	Section V <1 page; Section IV ∼20 pages **Pages 8–28**	Expand Section V or integrate into another section.	*Extreme section length imbalance.*
38	MOD	Notation Formalization	‘‘C’’ for Constraints but text uses CB, Cc, CI, CD **Page 6**	Add explicit definition of relationship between C and subscripted variants.	*Relationship between C and Cx never explicitly stated.*
39	MOD	Missing Formalization	B = (I, D, T, M, C, R) has no temporal component **Page 6**	Consider: B = (I, D, T, M, C, R, V) where V = Version/Timestamp.	*Paper emphasizes evolution but spec lacks temporal component.*
LOW SEVERITY SUGGESTIONS (40)					
40	LOW	Markers Formatting	‘‘x Vijay Janapa Reddi’’, ‘‘x Marco Colombo...’’ **Pages 44–45**	Remove “x” markers or explain meaning.	*Unexplained editing artifacts.*
41	LOW	Grouping Consistency	‘‘x Marco Colombo, Benjamin Hawks, and Nhan Tran have worked...’’ **Page 45**	Separate into individual entries per author.	*Inconsistent with other individual listings.*
42	LOW	Imprecise Clarity	‘‘Executing the same queries in Google Scholar’’ **Page 8**	Change to “equivalent queries”	*Different interfaces; “same” is technically inaccurate.*
43	LOW	Conflation Terminology	MLCommons and MLPerf used interchangeably **Throughout**	Add: “MLPerf is the benchmark suite maintained by MLCommons”	*Organization vs. benchmark suite distinction unclear.*
44	LOW	Truncation Tables	Table I: ‘‘Data Center C (NVIDIA B20...’’ **Pages 10–14**	Expand or use abbreviation scheme with legend.	*Truncation reduces table utility.*
45	LOW	Cross-Ref Completeness	‘‘we have provided in Table V the energy required’’ **Page 24**	Add: “...in Table V and visualized in Figure 1”	*Table and figure show same data; both should be referenced.*
46	LOW	Speculative Data	Table V: ‘‘GPT-5 >60,000 (estimated)’’ **Page 24**	Add caveat: “GPT-5/6 values are speculative projections”	*Unreleased models alongside measured data may mislead.*
47	LOW	Redundant References	[26] and [27] both point to MLCommons benchmark collection **References**	Consolidate or differentiate purposes.	*Both reference same GitHub resource.*
48	LOW	Informal References	[22] and [23] cite arXiv.org and Google Scholar **References**	Mention in text instead of numbered references.	*Citing search engines as references is unconventional.*
49	LOW	Format References	‘‘arXiv preprint arXiv:XXXX’’ vs. ‘‘arXiv: XXXX’’ **References**	Standardize all arXiv citation formats.	*Inconsistent formatting.*
50	LOW	Vague/Specific Consistency	‘‘households’’ vs. ‘‘130 homes’’ **Page 24**	Use consistent quantification throughout.	*Vague then specific is stylistically inconsistent.*
51	LOW	Missing Column Tables	‘‘Tables I and II... a brief note’’ **Page 9**	Add Notes column or remove from description.	*Mentioned “Notes” column not visible.*
52	LOW	Illegible Figures	Table III: Radar charts in ‘‘Ratings’’ column **Pages 16–20**	Increase size or add numerical ratings.	*Charts too small to read.*
53	LOW	Title Case Formatting	‘‘Sharing benchmarks’’ vs. ‘‘Towards a formal specification’’ **Contents**	Standardize capitalization convention.	*Mixed conventions across section titles.*

Contd on next page...

#	Sev.	Category	Manuscript Text	Suggested Edit	Reason
54	LOW	Awkward Grammar	‘‘democratize AI benchmarking in the design-space exploration’’ **Page 28**	Change to “by enabling design-space exploration”	*“in” should be “by enabling”.*
55	LOW	Tangential Focus	‘‘extended reality (XR) interfaces for immersive exploration...’’ **Page 28**	Condense or connect explicitly to benchmarking.	*XR discussion peripheral to main focus.*
56	LOW	Email Formatting	Some authors have inline emails, most do not **Appendix B**	Standardize: include all or remove all.	*Inconsistent contact information.*
57	LOW	Unsupported Accuracy	‘‘Google Scholar does not include all entries from arXiv, but... most’’ **Page 8**	Cite source or remove claim.	*Coverage claim unsupported.*
58	LOW	Undefined Completeness	‘‘four types of monitoring: (a)...(b)...(c)...(d)...’’ **Page 21**	Add brief definitions of each type.	*Types listed but never elaborated.*
59	LOW	Superlative Accuracy	‘‘one of the most comprehensive and standardized ecosystems’’ **Page 9**	Soften to “a comprehensive and standardized ecosystem”	*Superlative lacks comparative evidence.*
60	LOW	Redundant ID References	‘‘arXiv preprint arXiv:2511.05614, 2025. arXiv: 2511.05614’’ **Ref [25]**	Remove duplicate arXiv ID.	*ID listed twice in same entry.*
61	LOW	Formatting Curriculum	Curriculum bullets have inconsistent punctuation **Pages 28–29**	Standardize colon usage, capitalization, structure.	*Mixing styles reduces consistency.*
62	LOW	Redundancy Style	‘‘in order to work towards a formal definition’’ **Page 4**	Simplify: “to establish a formal definition”	*“work towards” is redundant phrasing.*
63	LOW	Grammar Style	‘‘And can be simply written in general as’’ **Page 6**	Change to: “This can be written in general form as”	*Starting with “And” is informal; sentence fragment.*
64	LOW	Redundancy Style	‘‘we try to identify the minimum min{...}’’ **Page 6**	Change to “we compute min{...}”	*“identify the minimum min” is redundant.*
65	LOW	Comma Grammar	‘‘In many cases however we still have’’ **Page 7**	Add commas: “In many cases, however, we still have”	*Missing commas around parenthetical.*
66	LOW	Confusing Clarity	‘‘simulation of real-time data while using a static dataset that is simulating...’’ **Page 7**	Simplify: “Living datasets may use real-time data or simulate updates using static data.”	*Circular phrasing is hard to parse.*
67	LOW	Vague Clarity	‘‘fostering community capacity to host independent infrastructure’’ **Page 28**	Clarify: “...to host benchmark repositories and leaderboards independently”	*“Independent infrastructure” is vague.*
68	LOW	Awkward Grammar	‘‘non-computer scientists would improve the use and development of the software’’ **Page 5**	Revise: “could improve their use and development of needed software”	*Current phrasing suggests they improve software itself.*
69	LOW	Spelling Consistency	‘‘democratization’’ (US) vs ‘‘data-centre’’ (UK) vs ‘‘Data Center’’ (US) **Throughout**	Standardize to one spelling convention.	*Mixed American/British spellings.*
70	LOW	Hyphenation Style	‘‘multi-mega-watt data centers’’ **Page 23**	Change to “multi-megawatt data centers”	*Extra hyphen is non-standard.*
71	LOW	Parallel Grammar	‘‘applied to analyze..., used as..., and to perform’’ **Page 28**	Use parallel verb forms throughout list.	*Inconsistent forms break parallel structure.*
72	LOW	Incomplete Examples	‘‘tasks include classification, translation, reasoning...’’ **Page 7**	Expand to include segmentation, object detection, recommendation from tables.	*Tables include many more task types not mentioned.*
73	LOW	Missing Legend	Table III radar charts show 1--5 scale not in table **Pages 16–20**	Add legend explaining 1–5 scale.	*Scale described in text but not visible in table.*
74	LOW	Title Accuracy	Table IV: ‘‘Profiling Tools... for Deep Learning and AI’’ **Page 22**	Change to “for HPC, Deep Learning, and AI Workloads”	*Table includes general HPC profilers.*

Contd on next page...

#	Sev.	Category	Manuscript Text	Suggested Edit	Reason
75	**LOW**	Citation Quality	`‘‘130 homes in the US [150]’’` **Page 24**	Cite U.S. EIA for household energy data.	*World Economic Forum not most authoritative source.*
76	**LOW**	Incomplete Hardware	`Some Table I entries just say ‘‘Data Center’’` **Pages 10–14**	Complete specifications or use “Data Center (various)”	*Vague entries reduce table utility.*
77	**LOW**	Notation Mismatch	`Figure 2 vs text: spelled-out vs abbreviated (Pfac/PIT)` **Page 25**	Use consistent notation throughout.	*Figure and text use different notation for same formula.*
78	**LOW**	Missing Defs Appendix	`Appendix A uses Pfac and PIT but never defines them` **Pages 42–43**	Add: “Pfac = Facility Power; PIT = IT Equipment Power”	*Abbreviations in definitions rely on undefined terms.*
79	**LOW**	Cross-Ref Completeness	`Section V on sharing doesn’t mention containerization` **Page 28**	Add: “Containerization (Section IV-C2) further supports reproducible sharing.”	*Containerization discussed earlier but not linked to sharing.*