

THE DURABILITY CURVE

THE EVIDENCE FILE · Nº 01

The Sources

Every study behind “Your AI Grader Is Only as Good as Its Answer Key”, with a working link.

arXiv 2503.05061 · Krumdick et al.

COLM 2026

No Free Labels

arXiv 2507.02856 · Chandak et al.

PREPRINT · 2025

Answer Matching Outperforms Multiple Choice

arXiv 2306.05685 · Zheng et al.

NEURIPS 2023

Judging LLM-as-a-Judge

arXiv 2310.08491 · Kim et al.

ICLR 2024

Prometheus

arXiv 2410.12784 · Tan et al.

ICLR 2025

JudgeBench

18 STUDIES · 2 PRACTITIONER GUIDES

EVERY TITLE AND ADDRESS IS A LINK · PREPRINTS ARE MARKED

THE DURABILITY CURVE

harryfloyd.substack.com

FREE · COMPANION PDF

THE STUDIES

-
- 01** **COLM 2026** **ARXIV 2503.05061**
No Free Labels
Krumdick et al. · Graders with and without a key
arxiv.org/abs/2503.05061
-
- 02** **PREPRINT 2025** **ARXIV 2507.02856**
Answer Matching Outperforms Multiple Choice
Chandak et al. · Matching answers to a reference
arxiv.org/abs/2507.02856
-
- 03** **NEURIPS 2023** **ARXIV 2306.05685**
Judging LLM-as-a-Judge
Zheng et al. · GPT-4 against experts; reference-guided grading
arxiv.org/abs/2306.05685
-
- 04** **ICLR 2024** **ARXIV 2310.08491**
Prometheus
Kim et al. · A grader that uses a reference
arxiv.org/abs/2310.08491
-
- 05** **ICLR 2025** **ARXIV 2410.12784**
JudgeBench
Tan et al. · Graders on hard answer pairs
arxiv.org/abs/2410.12784
-
- 06** **PREPRINT 2026** **ARXIV 2606.19544**
Reliability without Validity
Norman, Rivera, Hughes · 2026 graders: agreement and order effects
arxiv.org/abs/2606.19544
-
- 07** **ICML 2026** **ARXIV 2602.20629**
QEDBench
Gonzalez et al. · Proof graders against experts
arxiv.org/abs/2602.20629
-
- 08** **AUTOMATED SOFTWARE ENGINEERING · 2026** **ARXIV 2603.00539**
Are LLMs Reliable Code Reviewers?
Jin and Chen · Code review against a written spec
arxiv.org/abs/2603.00539
-

-
- 09** **ACL 2025** **ARXIV 2501.10970**
The Alternative Annotator Test
Calderon, Reichart, Dror · Can a grader stand in for an annotator?
arxiv.org/abs/2501.10970
-
- 10** **ICML 2024** **ARXIV 2403.04132**
Chatbot Arena
Chiang et al. · GPT-4 against expert relabels
arxiv.org/abs/2403.04132
-
- 11** **NEURIPS 2025** **ARXIV 2507.01001**
SciArena
Zhao et al. · Science answers from the literature
arxiv.org/abs/2507.01001
-
- 12** **PREPRINT 2026** **ARXIV 2609.12822**
Scaling Clinical Judgment
Buckley et al. · Graders against physicians
arxiv.org/abs/2609.12822
-
- 13** **PREPRINT OPENAI · 2025** **ARXIV 2505.08775**
HealthBench
Arora et al. · A grader with physician-written criteria
arxiv.org/abs/2505.08775
-
- 14** **TMLR 2026** **ARXIV 2604.23178**
Judging the Judges: A Systematic Evaluation of Bias Mitigation Strategies in LLM-as-a-Judge Pipelines
Soumik · Markdown and other style biases
arxiv.org/abs/2604.23178
-
- 15** **ICML 2026** **ARXIV 2601.22548**
Are LLM Evaluators Really Narcissists?
Roytburg et al. · Self-preference, controlled for difficulty
arxiv.org/abs/2601.22548
-
- 16** **PREPRINT 2026** **ARXIV 2609.12002**
Can We Trust LLM Judges
Zhang et al. · Graders on their own wrong answers
arxiv.org/abs/2609.12002
-

- 17** **PREPRINT** 2024 ARXIV 2404.18796
Replacing Judges with Juries
Verga et al. · Panels of smaller graders
arxiv.org/abs/2404.18796

- 18** **ICLR 2025** ARXIV 2410.13341
LLM as Judge Won't Beat Twice the Data
Dorner, Nastl, Hardt · Correcting graders with human labels
arxiv.org/abs/2410.13341

ALSO CITED

- 19** **PRACTITIONER GUIDE**
Using LLM-as-a-Judge For Evaluation: A Complete Guide
Hamel Husain · The expert-agreement method the Judge Check builds on
hamel.dev/blog/posts/llm-judge/
- 20** **PRACTITIONER GUIDE**
Evaluating the Effectiveness of LLM-Evaluators (aka LLM-as-Judge)
Eugene Yan · A review covering reference-based grading and order effects
eugeneyan.com/writing/llm-evaluators/

READ THE PIECE

Your AI Grader Is Only as Good as Its Answer Key

On Substack · harryfloyd.substack.com/p/your-ai-grader-answer-key

On the site · durabilitycurve.com/blog/your-ai-grader-answer-key



SCAN TO READ