AltRA-Test: when is an LLM judge statistically good enough to replace human annotators?

22 November 2026 · 7 min read · blogs.iictp.org/altra-test-llm-judges-replace-human-annotators

Short answer: AltRA-Test, the Alternative Annotator Test with Resampled Annotations, is a statistical test from an arXiv preprint dated 19 November 2026 that checks whether an LLM judge can replace human annotators when different items were labelled by different people. It declares a replacement supported when the inter-annotator agreement of a mixed human-and-LLM panel is statistically non-inferior to an all-human panel. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- AltRA-Test, from an arXiv preprint dated 19 November 2026, checks statistically whether an LLM judge can replace human annotators.
- It compares agreement of a mixed human-and-LLM panel against an all-human panel, and about 100 items often suffice.
- Passing means non-inferior agreement, not correctness, so collect a small human-labelled sample for your own task before trusting an LLM judge.

Read online, listen to it, or change the text size: https://blogs.iictp.org/altra-test-llm-judges-replace-human-annotators

What problem does it solve?

LLMs are increasingly used as judges and graders. The question is when that substitution is statistically justified. The earlier Alternative Annotator Test (Alt-Test) was designed for dense settings in which all items are labelled by the same annotators; many large annotation projects use different subsets of annotators per item.

How it works

- A bootstrap resamples items and draws k human annotations for each one.
- One of those human annotations is replaced with the LLM's annotation, producing paired panels: all-human and mixed human-LLM.
- The test checks whether the mixed panel's inter-annotator agreement (IAA) is non-inferior to the all-human panel's. If so, the LLM preserves the signal the humans share.
- The authors say they establish that it is a consistent test of asymptotic level alpha.

Results reported by the authors

- They tested 18 LLMs and 5 prompting techniques on 20 datasets covering text, image, audio and video.
- About 100 items often suffice for the test.
- In simulations, where the truth is known, AltRA-Test made more correct replacement decisions than existing baselines.
- In real-world data, LLMs passed the test in about half of the setups, depending mainly on the task.

Limitations and open questions

- Preprint, not peer reviewed; only the abstract was read.
- Passing means non-inferior agreement on the tested items, not that the LLM is correct. If humans share a bias, so will the LLM.
- The result depends on annotation quality and task; passing on one dataset does not transfer.
- Non-inferiority margins and how they are chosen were not read.

How to apply the idea

1. Collect a small human-annotated sample for your task, with at least k annotations per item.
2. Run your LLM judge on the same items and compute IAA with and without one human replaced.
3. Test non-inferiority rather than comparing averages, and report an interval.
4. Re-run when the prompt, model or data changes.

What it means for teams using LLM evaluation

LLM-as-a-judge numbers are only as good as the check that they track human judgement. This is a recipe for that check that does not need the same annotators on every item. See how to evaluate an LLM application.

Frequently asked questions

- **How many items do I need?** The authors say about 100 often suffice.
- **Do LLMs usually pass?** In about half of their real-world setups, depending mainly on the task.
- **Does passing mean the judge is accurate?** It means it preserves the agreement the humans share.

Related reading

- How to evaluate an LLM application
- LLMOps: running AI in production
- AI Practitioner

Sources and currency

Sources were read on 22 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Omar N., Dror R., Gueta A., Gal A., Reichart R., Calderon N. AltRA-Test: An Alternative Annotator Test for LLM Judges Without Shared Annotators. arXiv:2611.26073. 19 November 2026. arxiv.org/abs/2611.26073.

Published by IICTP, an independent technical credential body, on 22 November 2026. The current version of this post is at https://blogs.iictp.org/altra-test-llm-judges-replace-human-annotators. More guides: blogs.iictp.org.