Longitudinal Drift in Large-Model Judgments of Scientific Abstracts

J. R. Whitfield, Amara Nwosu, Tomás Beltrán, Wei Zhang
Symposium on Reproducible Evaluation (SRE 2026), Held in Reykjavík
Marsh Institute of Technology · Cordillera Research Council

Background

Automated judges are widely used to score generated text, yet their stability across model versions is rarely audited.

Methods

We replayed 3,000 abstract-quality judgments across four model releases spanning eighteen months, holding rubric and inputs fixed.

Results

Mean score drifted by 0.41 points on a five-point scale, with disagreement concentrated in methodology criteria. Inter-version agreement fell to 0.62 kappa.

Conclusion

Evaluation pipelines must pin judge versions and re-baseline on upgrade; we release our replay harness to support this practice.

Index terms — LLM-as-judge, evaluation drift, reproducibility, version pinning