ScienceToStartup Weekly Benchmark - 2026-05-25

Methodology: v2
Artifact ID: live-benchmark:2026-05-25:b538f9d3ca1ab1a4

1. Towards Human-Like Interactive Speech Recognition With Agentic Correction and Semantic Evaluation
2. GPIC: A Giant Permissive Image Corpus for Visual Generation (2605.30341v1) - composite 90.5
3. Before the Shutter: Aesthetic and Actionable Portrait Photography Planning in 3D Scenes (2605.30318v
4. MedCase-Structured: A Text-to-FHIR Dataset for Benchmarking Diagnostic Reasoning in Clinically Rea
5. mcp-proto-okn: Natural-language access to open scientific knowledge graphs through the Model Contex
6. Qwen-VLA: Unifying Vision-Language-Action Modeling across Tasks, Environments, and Robot Embod
7. Loong: A Human-Like Long Document Translation Agent with Observe-and-Act Adaptive Context Selec
8. PhyGenHOI: Physically-Aware 4D Generation of Dynamic Human-Object Interactions (2605.30268v1) -
9. How LoRA Remembers? A Parametric Memory Law for LLM Finetuning (2605.30260v1) - composite 80
10. Automating Low-Risk Code Review at Meta: RADAR, Risk Calibration, and Review Efficiency (2605.30
11. CalArena: A Large-Scale Post-Hoc Calibration Benchmark (2605.30188v1) - composite 80
12. No More K-means:Single-Stage Sparse Coding for Efficient Multi-Vector Retrieval (2605.30120v1) - co
13. PokerSkill: LLMs Can Play Expert-Level Poker without Training or Solvers (2605.30094v1) - composite
14. Selective QA over Conflicting Multi-Source Personal Memory: A Diagnostic Testbed and Method Comp
15. REPOT: Recoverable Program-of-Thought via Checkpoint Repair (2605.30052v1) - composite 80