AI Can Publish at Top AI Conferences Now!

Apex Research Team/Apex Intelligence/research@apexin.ai

Webpage GitHub

Statement. *Our goal is to develop self-improving AI systems to unlock undiscovered discovery for the benefit of humanity. We recognize that submitting AI-generated manuscripts to real peer review consumes valuable reviewer time, and we sincerely apologize for this burden. We conducted this evaluation because unbiased and rigorous assessment by expert reviewers is essential for understanding whether AI can make meaningful contributions to scientific research. All AI-generated submissions will be withdrawn after the review process, and all manuscripts, experimental protocols, and reviews will be released publicly.*

Core message. **Apex Research** has now passed a Turing test for scientific research at top AI conferences. Among 34 fully AI-generated manuscripts submitted to ACL Rolling Review, 8 received average overall scores above 3.0, outperforming more than 88% of human submissions, and 2 achieved a score of 3.67, exceeding the competitive acceptance threshold and outperforming more than 99% of human papers. Evaluated as ordinary conference submissions, these papers achieved review scores comparable to human research, with an AI average of 2.60 and a human average of 2.63. AI systems that autonomously generate and validate scientific discoveries have the potential to greatly expand the frontiers of human knowledge. **The ChatGPT moment for AI scientists may have arrived.** Apex Research is developed by Apex Intelligence, a next-generation self-evolving foundation model company aiming to unlock undiscovered discovery.

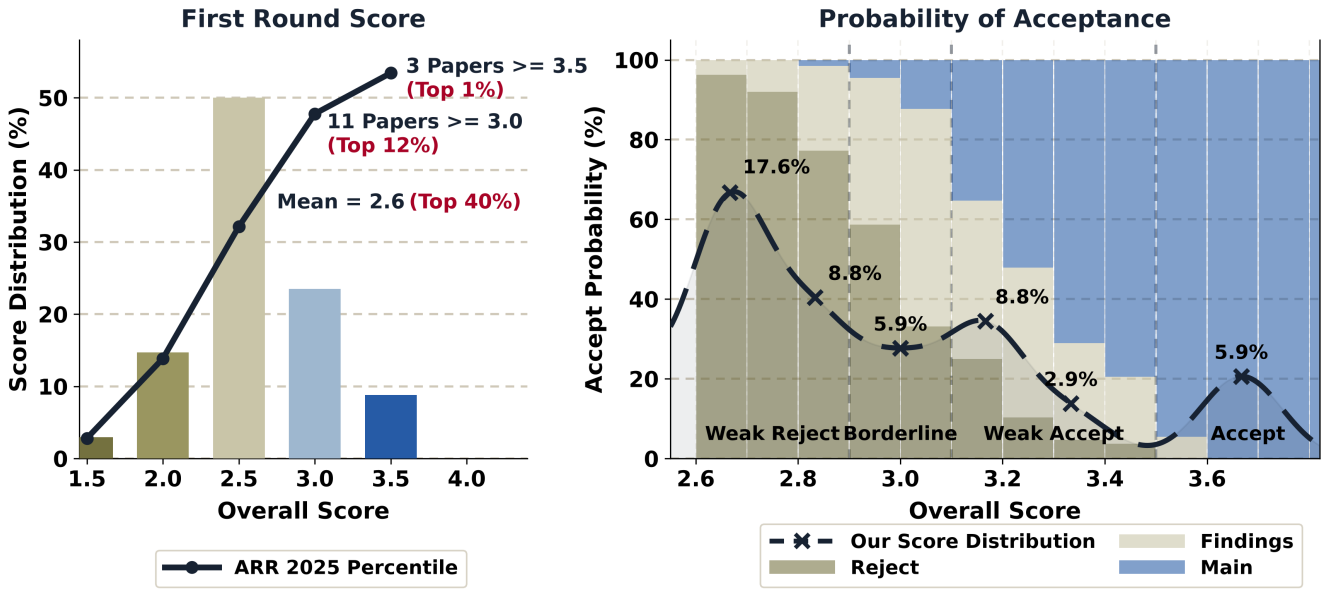


Figure 1. First round review scores and acceptance probability of all submissions. Our mean score of 2.6 is in the top 40% of ARR 2025 submissions, with 11 papers in the top 12% (score: 3.0) and 3 papers in the top 1% (score: 3.5). The right panel shows 32% of our submissions can be accepted as Main or Findings.

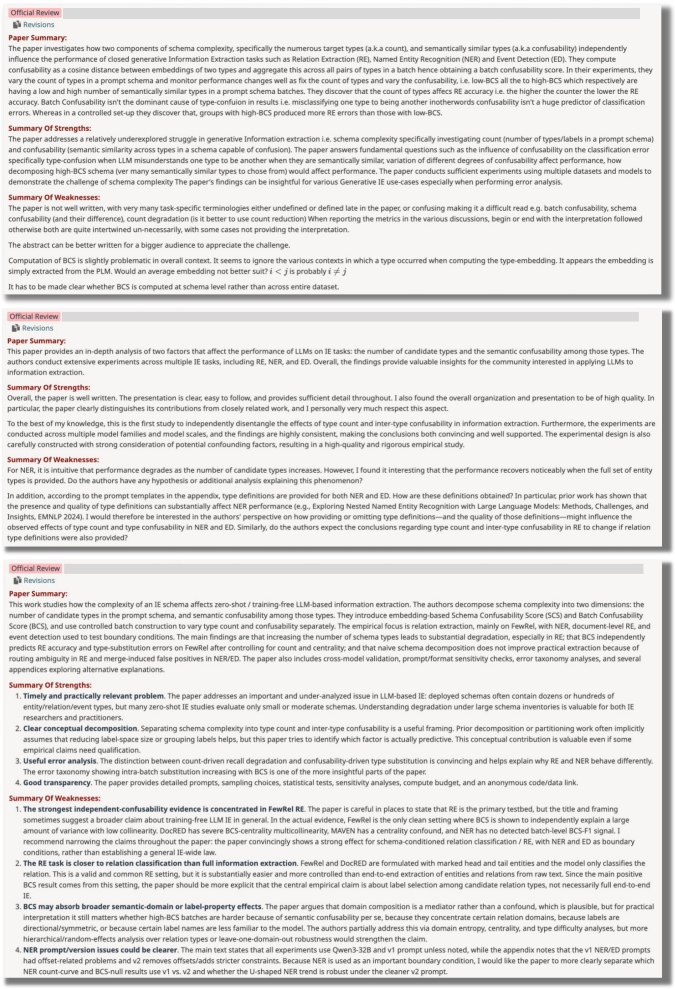
Apex Research Conducts 34 End-to-End Research Projects with Limited Non-Expert Human Assistance

Apex Research autonomously generated research ideas and completed the full research workflow for 34 projects, including literature grounding, method design, implementation, experiments, result analysis, figure and table preparation, LaTeX writing, self-review, self-judgment, and iterative self-improvement, with all manuscripts submitted to ACL ARR 2026 May cycle through the standard conference review process. Among the 34 projects, 32 used GPU resources, consuming 1,286 allocated GPU-days in total, with an average of 2.4 GPUs and 38 GPU-days per project. Human participation remained limited, contributing only 0.38% of all human-plus-AI message turns. Human interactions consisted of general guidance rather than domain expertise, including checking progress, requesting the latest manuscript version, identifying formatting issues such as page limits or overlapping table entries, and instructing the system to refine the paper layout.

Performance in Real-world Peer Review

Figure 1 summarizes the review scores of all 34 ARR submissions. The results represent a clear breakthrough. Rather than producing a few isolated successes, Apex Research consistently generated papers that achieved competitive performance under real conference peer review. Overall, according to the data of the year 2025, **17 papers outperformed 60% of human submissions, eight outperformed 88%, and the two highest-scoring papers outperformed more than 99%** of the ARR 2025 reference distribution. Together, these results show that the system can reliably produce high-quality research papers rather than relying on a handful of exceptional cases. According to the historical data of 2024 ACL ARR, 11 papers are highly likely to be accepted by ACL and EMNLP as Main or Findings papers.

One representative paper (Figure 2) received an average review score of **3.67**, with reviewer **Overall Assessments of 4.0/3.5/3.5**, **Soundness Scores of 3.5/4.0/3.5**, **Excitement Scores of 4.0/4.5/3.0**, and **Confidences of 3.0/4.0/3.0**. Reviewers consistently recognized both its novelty and technical rigor. One reviewer described it as *"the first study to independently disentangle the effects of type count and inter-type confusability in information extraction,"* and further concluded that *"the findings are highly consistent, making the conclusions both convincing and well supported."* Collectively, the reviews praised its **conceptual novelty**, **rigorous experimental design**, and **practical relevance**, reinforcing the broader finding that **Apex Research is capable of repeatedly generating research that is recognized by expert reviewers for its originality and technical rigor.**



Reviewer 1 : The paper addresses a relatively **underexplored struggle in generative Information extraction...** The paper conducts **sufficient experiments** using **multiple datasets and models** to **demonstrate the challenge** of schema complexity... The paper's findings can be **insightful** for...

Reviewer 2 : To the best of my knowledge, **this is the first study** to... The experimental design is also **carefully constructed with strong consideration of potential confounding factors,** resulting in a **high-quality and rigorous empirical study.**

Reviewer 3 : **Timely and practically relevant problem.** The paper **addresses an important and under-analyzed issue** in... Understanding degradation under large schema inventories **is valuable for both IE researchers and practitioners.**

Clear conceptual decomposition. Separating schema complexity into type count and inter-type confusability **is a useful framing...**

Useful error analysis... The error taxonomy showing intra-batch substitution increasing with BCS **is one of the more insightful parts of the paper.**

Figure 2. Representative paper case with consistently high review scores and remarks from all three reviewers. The paper received an average review score of 3.67, with reviewer Overall Assessments of 4.0/3.5/3.5, Soundness Scores of 3.5/4.0/3.5, Excitement Scores of 4.0/4.5/3.0, and Confidences of 3.0/4.0/3.0.

Implications and Open Questions of Apex Research

The emergence of Apex Research raises several fundamental questions for the future of scientific discovery:

1. What roles will AI scientists play across academia, industrial R&D, and domain-specific scientific discovery?
2. As AI scientists become increasingly capable, what unique strengths and responsibilities will remain for human researchers?
3. How should education and talent development evolve in the era of AI scientists? What knowledge, skills, and training paradigms will future researchers need to effectively collaborate with increasingly autonomous AI research systems?
4. If research papers are no longer a sufficient measure of scientific ability, what new metrics should be used to evaluate researchers and AI scientists?
5. Will the research paper remain the primary medium for communicating scientific discoveries, or will more AI- and agent-native forms of scientific communication emerge?
6. How can we improve the creativity and innovation capabilities of AI scientists to enable truly groundbreaking discoveries rather than incremental advances?
7. As AI scientists dramatically accelerate the pace of research, how can we reliably evaluate the novelty, validity, reproducibility, and long-term impact of the growing volume of scientific discoveries?
8. What ethical, legal, and governance frameworks are needed for AI-driven scientific discovery? How should issues such as intellectual property ownership, authorship attribution, AI contribution disclosure, research equity, scientific integrity, and research security be addressed?