AI DAILY REPORT

AI Nightly Comprehensive Report - November 9, 2026

South Korea's AI Infrastructure Cooperation and Verification of Agent Completeness and Safety

2026. 11. 09

Cooperation to connect South Korean AI chips to NVIDIA systems has begun, and free API access to a domestic model has expanded. AI's roles are becoming concrete in robot simulation training, production planning, and hospital data retrieval. In coding and medical research, findings continue to show that partial success cannot be equated with complete task execution. As the scope of application widens, real-world outcomes must be examined alongside test conditions.

---

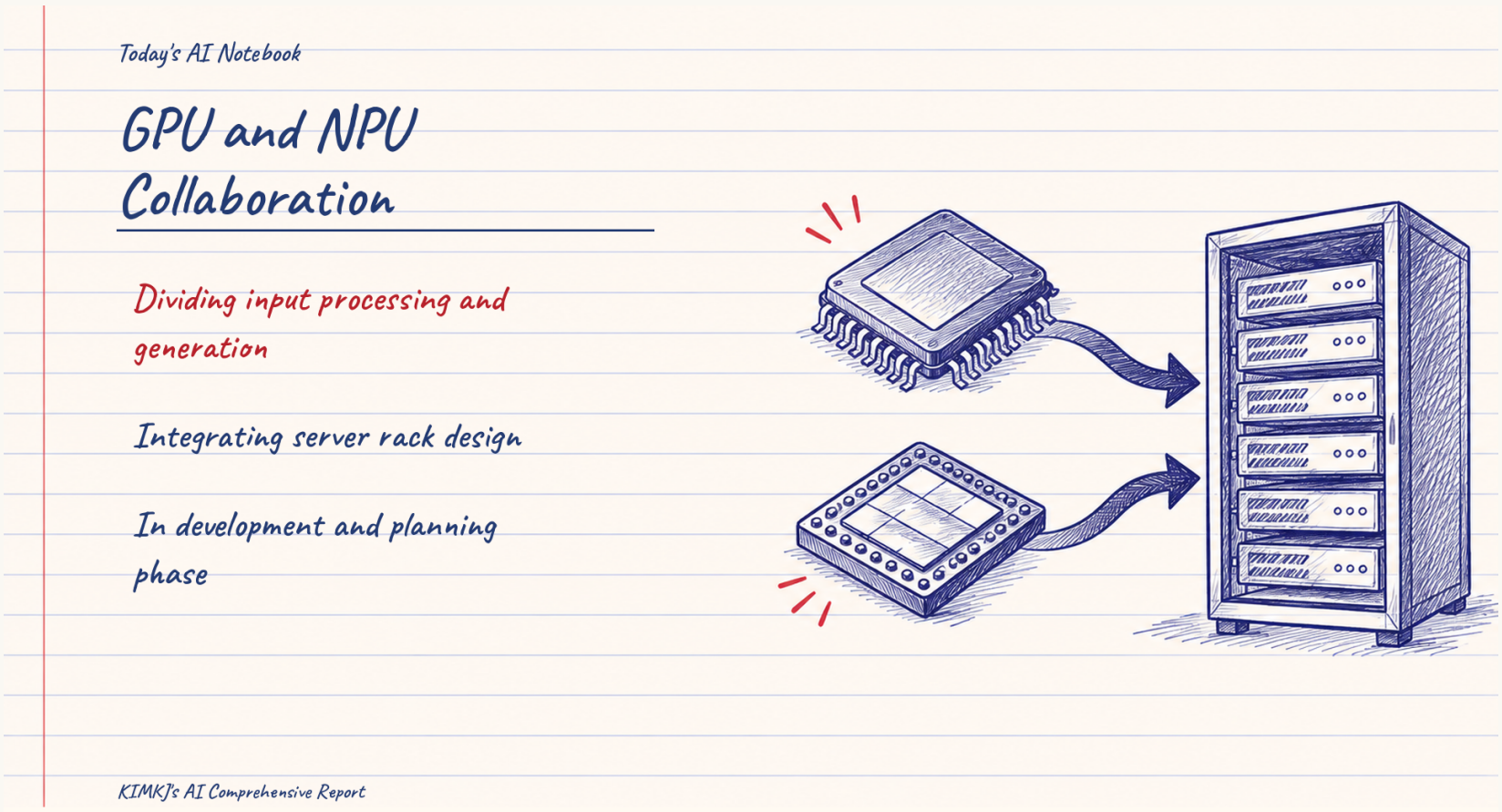
1 South Korean NPUs and NVIDIA GPUs Share Inference Workloads
2 LG Opens K-EXAONE 2.5 via Free API
3 Creating Robot Simulation Data Even for Flexible Cables
4 Production Planning AI Factoring in Equipment, Materials, and Deadlines
5 Retraining on AI-Generated Recommendations Perpetuates Errors
6 Google ADK Strengthens Repeated Evaluation and Script Restrictions
7 Meeting 70% of Requirements, Full Completion Reaches Only 26.4%
8 Modifying Internal Model Weights to Alter Capabilities
9 Researchers Use AI for Analysis and Writing More Than Ideation
10 Predicting Protein Binding Sites Without 3D Structures
11 Identifying Missing Patient Context in AI Insulin Advice
12 AI Literature Reviews Require Multi-Dimensional Quality Evaluation
13 Converting Spoken Research Queries into Medical Record Searches
14 US Air Force Expands Deployment of Munitions Storage Planning AI
15 Demonstrating Explainable Urban Flood Forecasts for Forecasters

Semiconductors

1 South Korean NPUs and NVIDIA GPUs Share Inference Workloads

Rebellions, NVIDIA, and SK Telecom agreed on November 9 to jointly develop a Korean AI computing infrastructure. The plan combines Rebellions' next-generation neural processing unit, the R200, with NVIDIA's high-speed interconnect technology, NVLink Fusion, building a system conforming to the MGX server rack design specification. SK Telecom plans to connect this system to its AI data centers, models, and services.

The development direction divides the AI response generation process across two types of semiconductors. The system is designed so that GPUs handle the stage that processes the input all at once, while NPUs handle the subsequent stage that generates the answer sequentially. Rack size, power supply, and cooling designs are also aligned. The goal is to improve throughput and power efficiency, but this agreement cannot be regarded as the completion or performance validation of a commercial system. Actual operational outcomes must be evaluated after development and data center deployment.

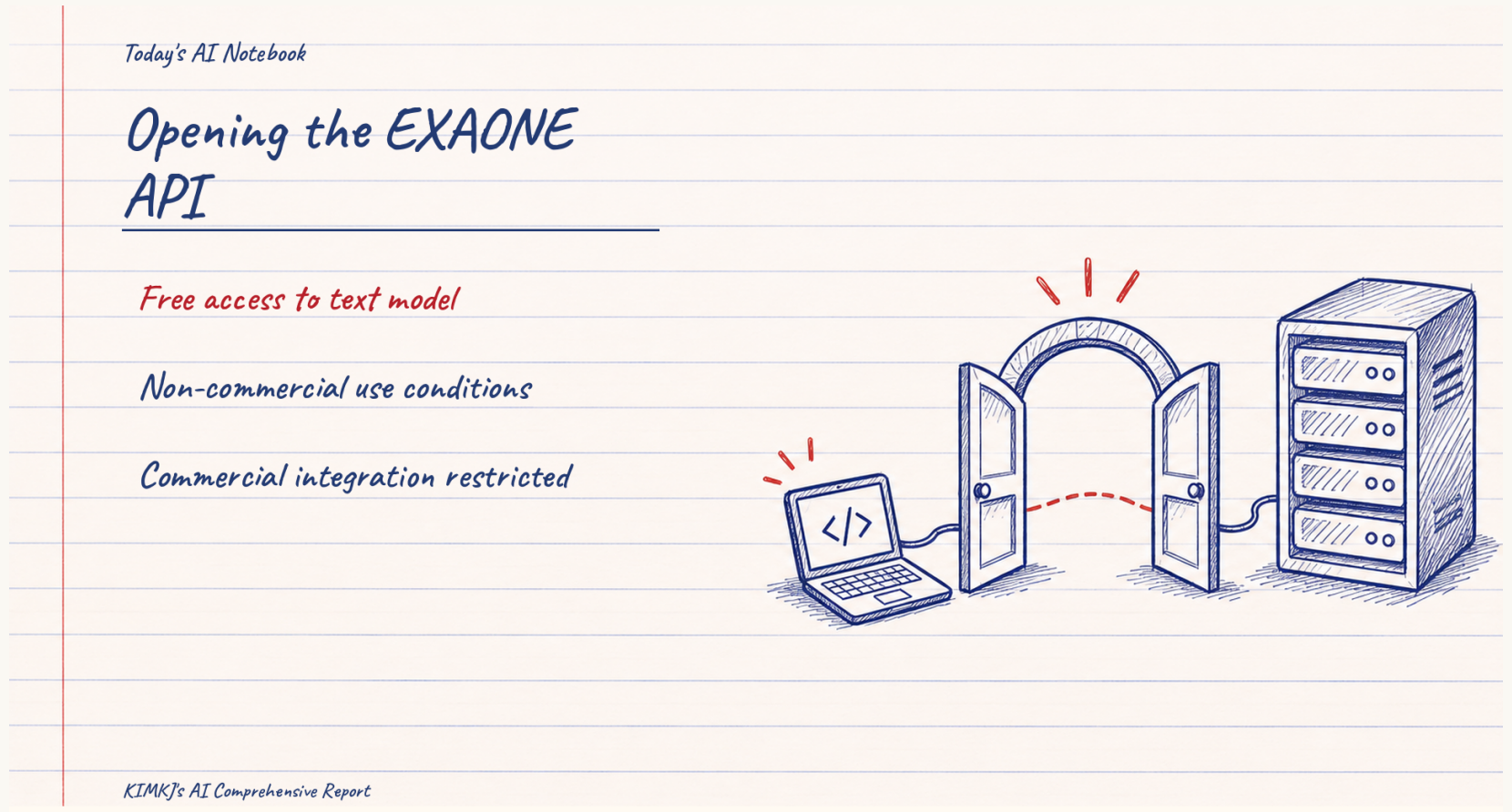


Korean AI Models

2 LG Opens K-EXAONE 2.5 via Free API

LG AI Research is opening the API for K-EXAONE 2.5 for free starting November 9. An API serves as an interface for developers to send queries to the model and receive responses within their own programs. Developers, startups, enterprises, and research institutions will be able to test the model directly. However, the current platform terms of service permit only non-commercial use, and being free does not mean it can be directly integrated into commercial services.

This model is a text-only model that takes text input and generates text output. While it follows the existing OpenAI-compatible API format, the scope of technical connectivity differs from what the terms allow. The terms prohibit using the model or its outputs for commercial purposes, as well as integrating the API into products or services for third-party calls. Developers can obtain an API key for testing, but must distinguish usage terms prior to actual deployment. This release expands opportunities to validate the refined model through real-world usage, and is not an open license permitting all commercial applications.

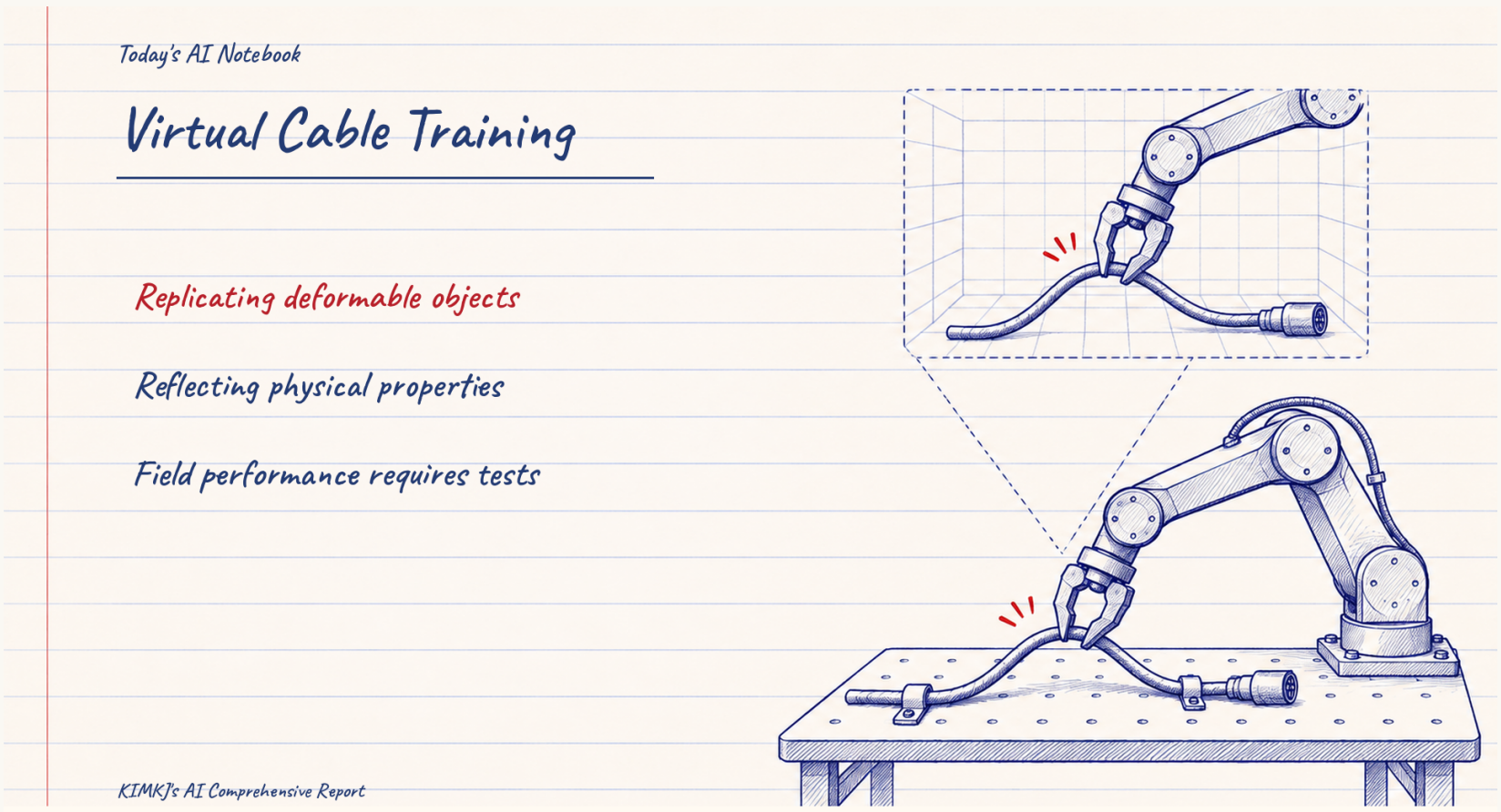


Robot Learning

3 Creating Robot Simulation Data Even for Flexible Cables

NdotLight is creating and supplying KT with simulation assets for flexible objects such as cables. Unlike rigid boxes, cables change shape depending on where and how firmly they are grasped. Testing robot manipulation of such objects in virtual environments requires assets that reflect not only their visual appearance but also their physical properties.

NdotLight uses its proprietary CAD-based platform, TRINIX, to convert equipment and parts into virtual training assets. These assets are structured in OpenUSD format for 3D scene interchange, supporting the verification of robotic manipulation scenarios in NVIDIA Isaac Sim. This collaboration was showcased at NVIDIA AI Day in Seoul on November 9. The initiative reflects workspaces and components unique to each enterprise. However, results verified in simulated environments must be distinguished from success rates in actual factories. This public showcase alone does not definitively prove the extent of productivity gains or full field deployment.

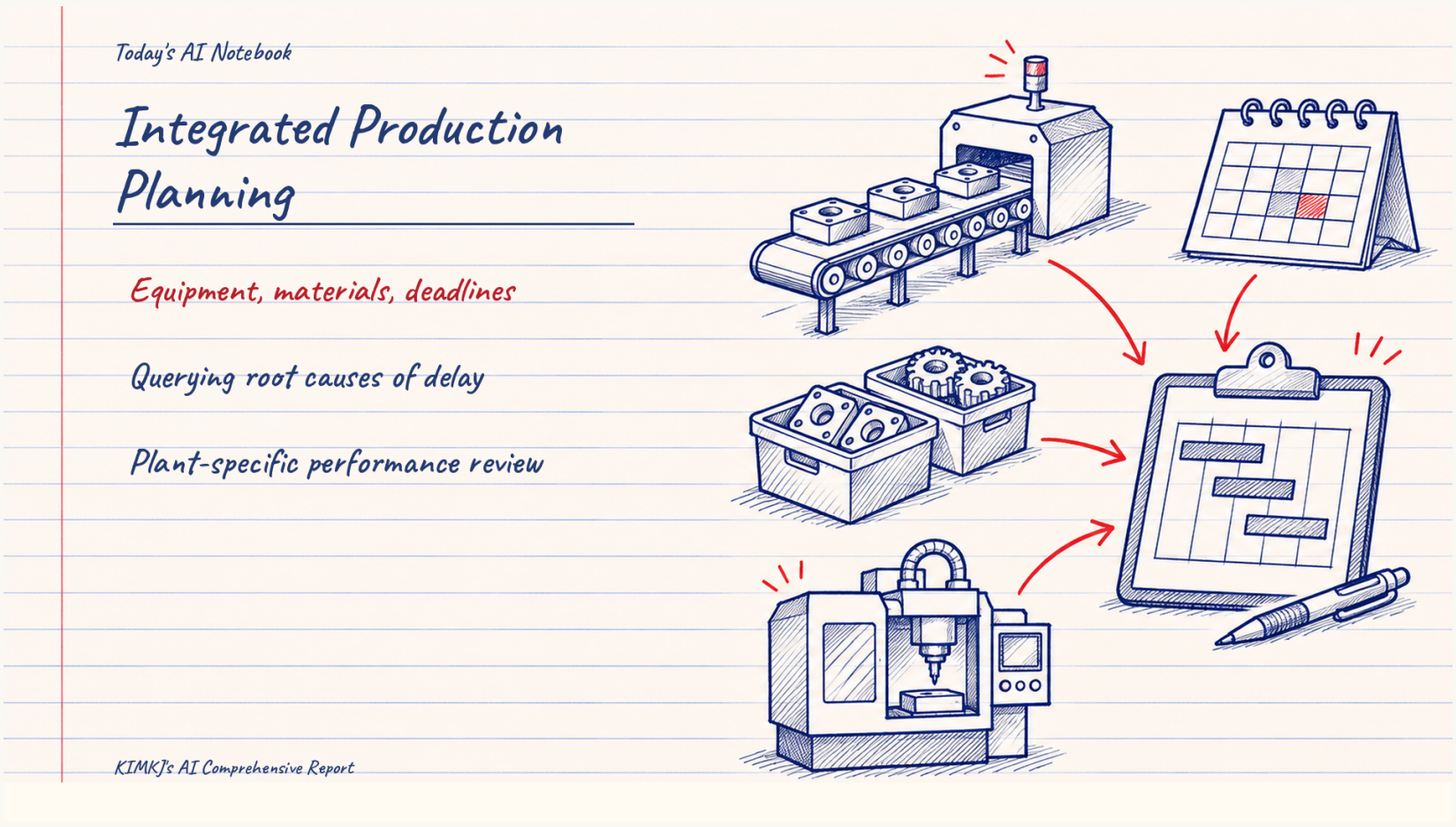


Manufacturing AI

4 Production Planning AI Factoring in Equipment, Materials, and Deadlines

VMS Solutions has been selected as an M.AX specialized enterprise supporting manufacturing AI transformation. Its flagship solution, MOZART, adjusts supply and production planning by accounting for constraints that must be simultaneously satisfied on factory floors, such as equipment, processes, materials, and delivery deadlines. The company supplies technology to manufacturers in semiconductors, displays, automotive, and batteries.

The company also provides digital twin technology for virtually analyzing production flows, alongside natural-language query-based AI agent technology. This approach allows frontline staff to query causes of delay in plain sentences and investigate interconnected planning and process constraints. This represents a case where manufacturing AI expands beyond visual inspection for product defects into schedule optimization and root-cause analysis. Selection as a specialized enterprise reflects an assessment of technical capability and field deployment competence, rather than a certification guaranteeing increased output or shortened lead times across all factories. Performance must be verified against equipment and material conditions specific to each plant.

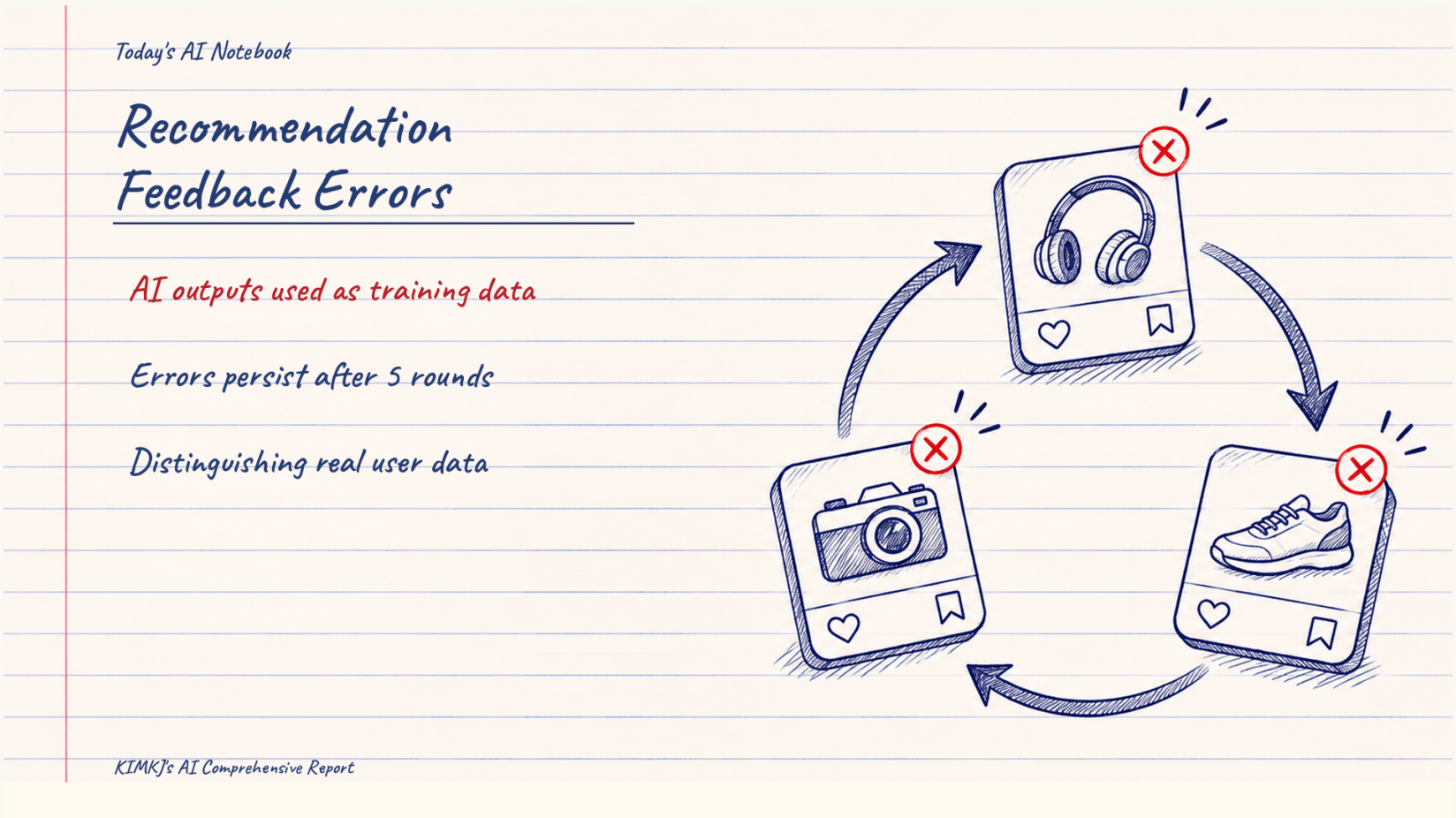


Recommendation AI

5 Retraining on AI-Generated Recommendations Perpetuates Errors

A research team led by Professor Yeon-Chang Lee at UNIST has developed EchoTrace to track errors and biases in large language model-based recommendation systems. It is a diagnostic tool that examines processes such as directly generating recommendation lists, creating synthetic user profiles, and generating scarce training data. It inspects loops where AI-generated outputs become inputs for subsequent training.

In experiments where the research team had AI directly generate recommendation lists, the rate of recommending items not present in candidate pools was 4.07% for movies, 7.40% for books, and 13.14% for stores. Under the synthetic profile approach, 93.16% of occupations inferred from movie usage data differed from the original user profiles. Even after repeating the retraining loop on recommendation outputs five times, initial errors did not disappear, and in some cases bias toward popular items intensified. These results were obtained from the specific datasets and systems tested. They cannot be generalized as error rates across all recommendation services. They highlight the problem that unless AI-generated records are distinguished from data gathered from real users, initial assumptions can become entrenched through iterative training.

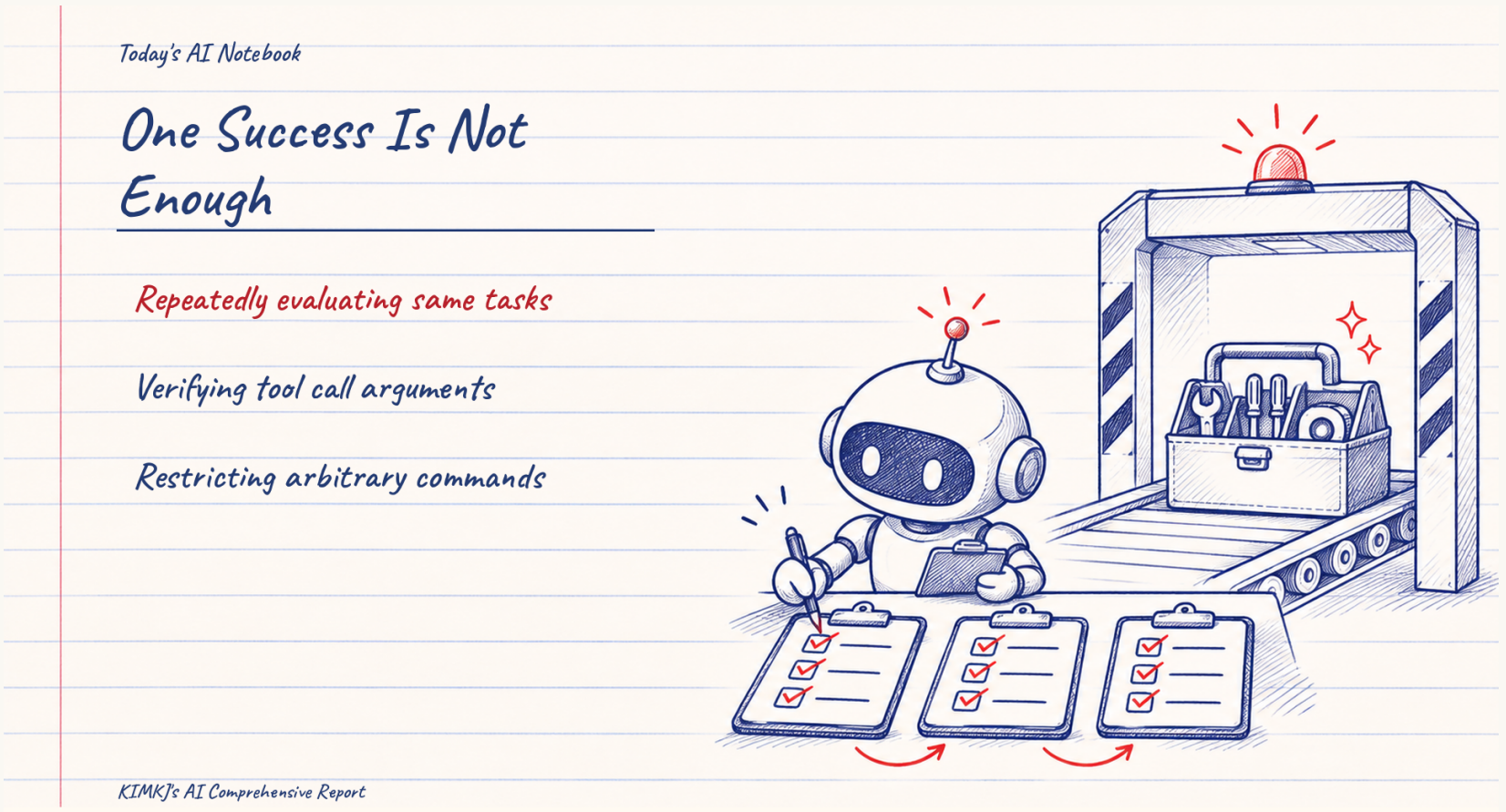


Agent Development

6 Google ADK Strengthens Repeated Evaluation and Script Restrictions

Google's agent development toolkit, ADK 2.14.0, has added functionality to aggregate scores and pass rates by running the same evaluation benchmark multiple times. Since AI can produce different outcomes for identical requests, this helps prevent judging performance based on a single success. Arguments passed to tool calls can also be inspected via exact match, numerical tolerance, and semantic comparison.

For enterprise integration, functionality was introduced to deploy agents across Microsoft Teams chats and channels. Security changes are also present. In environment mode, the skill execution tool now uses validated script paths and structured arguments instead of accepting free-form shell commands. This measure aims to reduce the risk of malicious instructions within documents leading to arbitrary command execution. The update also includes breaking changes that may not be compatible with existing code. As the data types for certain tool arguments and session storage formats have changed, developers must independently check legacy program operations and rollback feasibility.



Coding Evaluation

7 Meeting 70% of Requirements, Full Completion Reaches Only 26.4%

Co-Design Bench, an evaluation benchmark for coding agents, tests models by converting real user app modification requests into tasks. Researchers examined 37,463 deduplicated requests from 26,314 user sessions to construct 273 tasks. The benchmark evaluates not only UI layout but also whether actual functionality works and whether multi-file modifications remain consistent.

Even the best result among the 10 models tested in the same execution environment passed an average of 69.9% of requirements per task, yet tasks satisfying all conditions reached only 26.4%. This highlights the gap between partially plausible results and end-to-end completion. Evaluation inspects source code and browser behavior, followed by another AI judge that cross-examines and verifies pass decisions. Performance declined when multiple types of modifications overlapped.

This study, released on October 26, carries limitations stemming from AI judging and the specific runtime environment. These figures cannot be viewed as automation success rates for all software development tasks.

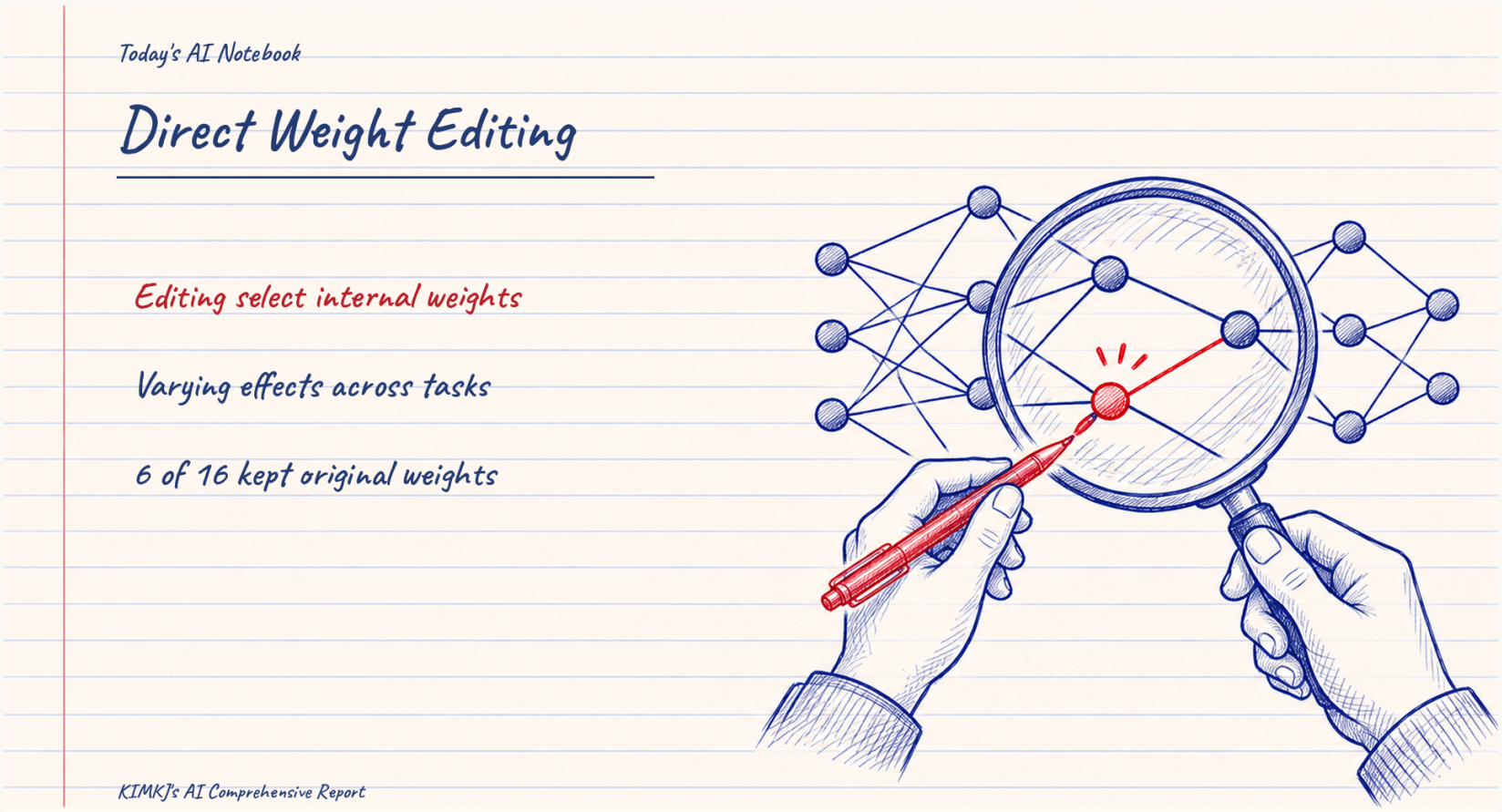


Model Research

8 Modifying Internal Model Weights to Alter Capabilities

An AI model's internal weights are vast numbers determined through training. A study on direct weight editing, released on November 6, had reasoning agents analyze these numbers and modify a subset of them. Departing from conventional fine-tuning, the researchers tested 16 editing paths across two model sizes of Qwen3.5 and seven evaluation benchmarks.

In a small model with 800 million parameters, modifying 1,023 internal weights raised math evaluation scores by 29.45 percentage points. While the math-adapted model improved on certain other tasks, it showed no change on SciCode, a scientific coding benchmark. Results were uneven. In 6 of the 16 editing paths, improvements could not be reliably verified, leading researchers to retain original weights. Even in larger models, cases emerged where improvements on development data did not carry over to separate evaluation data. These are limited experimental findings and do not mean slightly altering internal numbers will boost intelligence across all models.

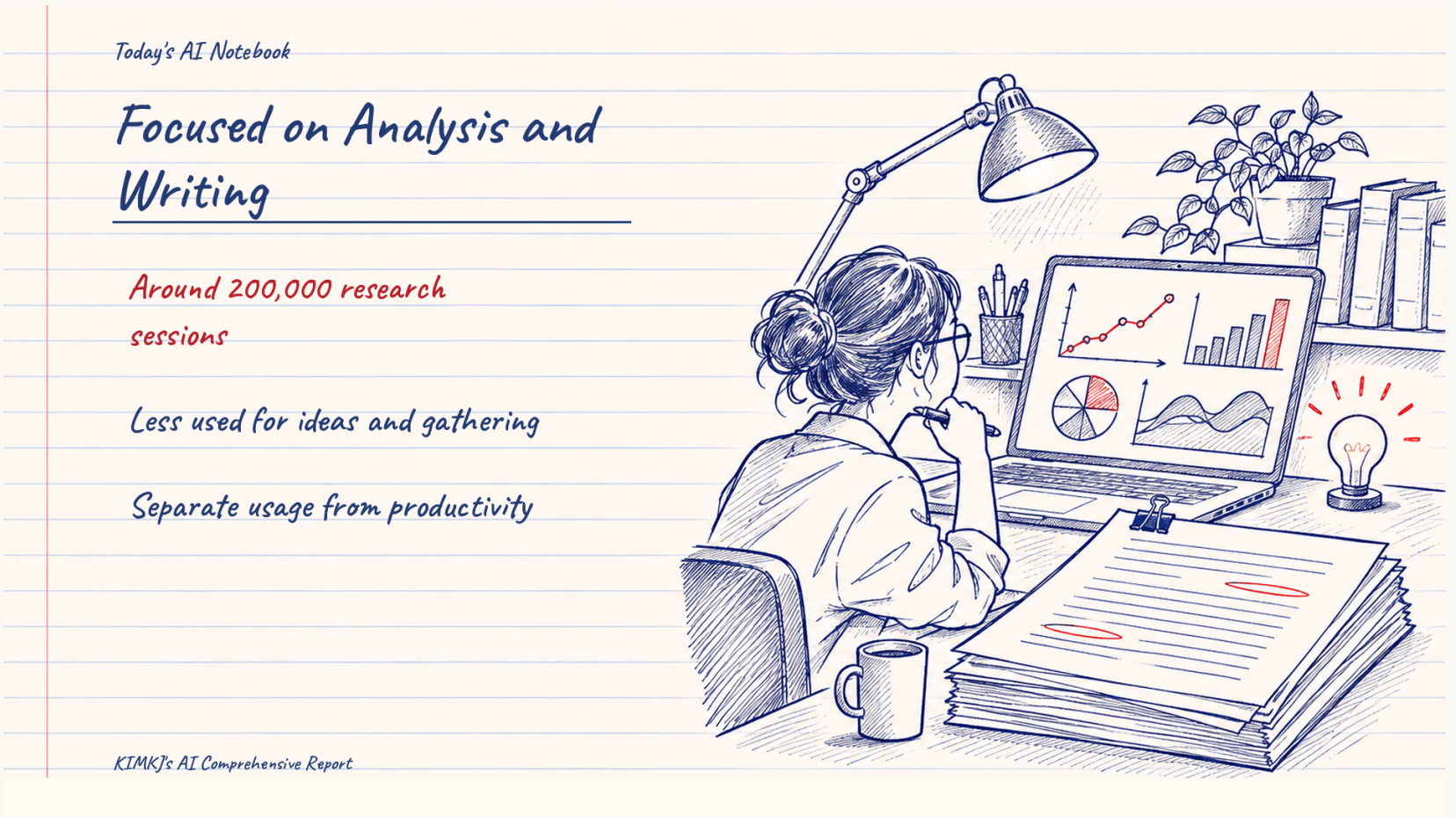


Scientific Research

9 Researchers Use AI for Analysis and Writing More Than Ideation

In a research usage analysis released by Anthropic on November 5, scientists' use of Claude was concentrated on data analysis and writing. The analysis examined 207,049 research-related sessions alongside a survey of 528 researchers. Data analysis accounted for 33.7% of Claude sessions and 48.1% of Claude Code sessions. In contrast, it was used relatively less for idea generation and data collection.

Survey respondents cited idea generation and data collection as critical to scientific discovery. Crucial tasks and the tasks frequently delegated to AI thus do not yet align. In science-specific tools, ideation tasks showed a higher share of usage. The analysis covered individual subscriber usage records, excluding enterprise and API usage, and AI classification was applied to categorize disciplines. These findings reflect a company analyzing the use of its own product. High usage volume alone does not prove a corresponding rise in productivity, and calculations regarding increased research output represent hypothetical scenarios rather than observed outcomes.

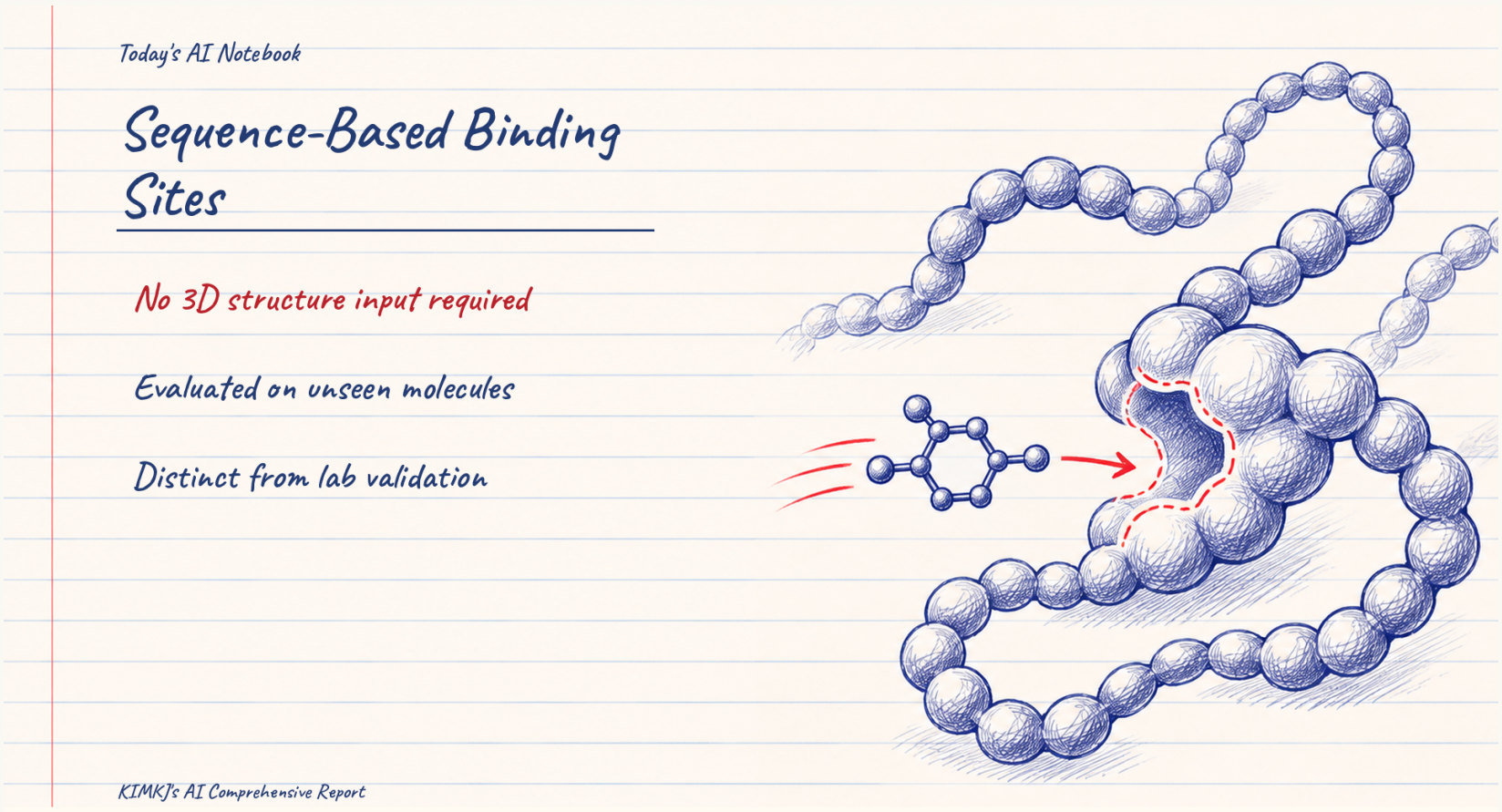


Drug Discovery

10 Predicting Protein Binding Sites Without 3D Structures

Identifying where proteins and drug candidate molecules bind is the starting point of drug discovery. A study revised and published on November 7 predicts binding sites by simultaneously reading protein amino acid sequences and molecular SMILES strings. Rather than directly taking 3D protein structures as inputs, the method combines protein language models with chemical language models.

Researchers evaluated performance across molecules with abundant training data, scarce data, and unseen molecules. On unseen molecules, the F1 score representing binding site prediction balance was 0.3109 when aggregated across all residues, and 0.2338 when averaging per-molecule values. Because these two values use different calculation methods, they should not be conflated as interchangeable metrics. Shuffling or randomizing molecular representations degraded performance. While this supports that the model utilizes actual molecular inputs, it does not demonstrate efficacy or binding in laboratory wet-lab experiments. This is a preprint study representing a computational screening method to assist drug candidate selection.



Medical AI Safety

11 Identifying Missing Patient Context in AI Insulin Advice

A study examining AI responses to caregivers inquiring about hyperglycemia in pediatric type 1 diabetes was published on November 8. The specific hazard examined by researchers was recommending insulin administration without explicit warnings to account for insulin on board. Responses were tested under varying contexts regarding patient meals, prior insulin doses, and treatment regimens.

These potential safety failures were observed across all four models tested via API and two models tested via chat interfaces. When contexts regarding patient meals, prior insulin administration, and treatment regimens were varied, the occurrence rates varied as well. This study evaluated generated responses and does not imply that actual patients suffered harm at identical rates. Because the preprint abstract does not provide specific failure rates per model, products cannot be ranked. The findings indicate that clinical dialogue systems must be verified not only for natural explanations but also for ensuring conditions essential to medical decision-making are not omitted.

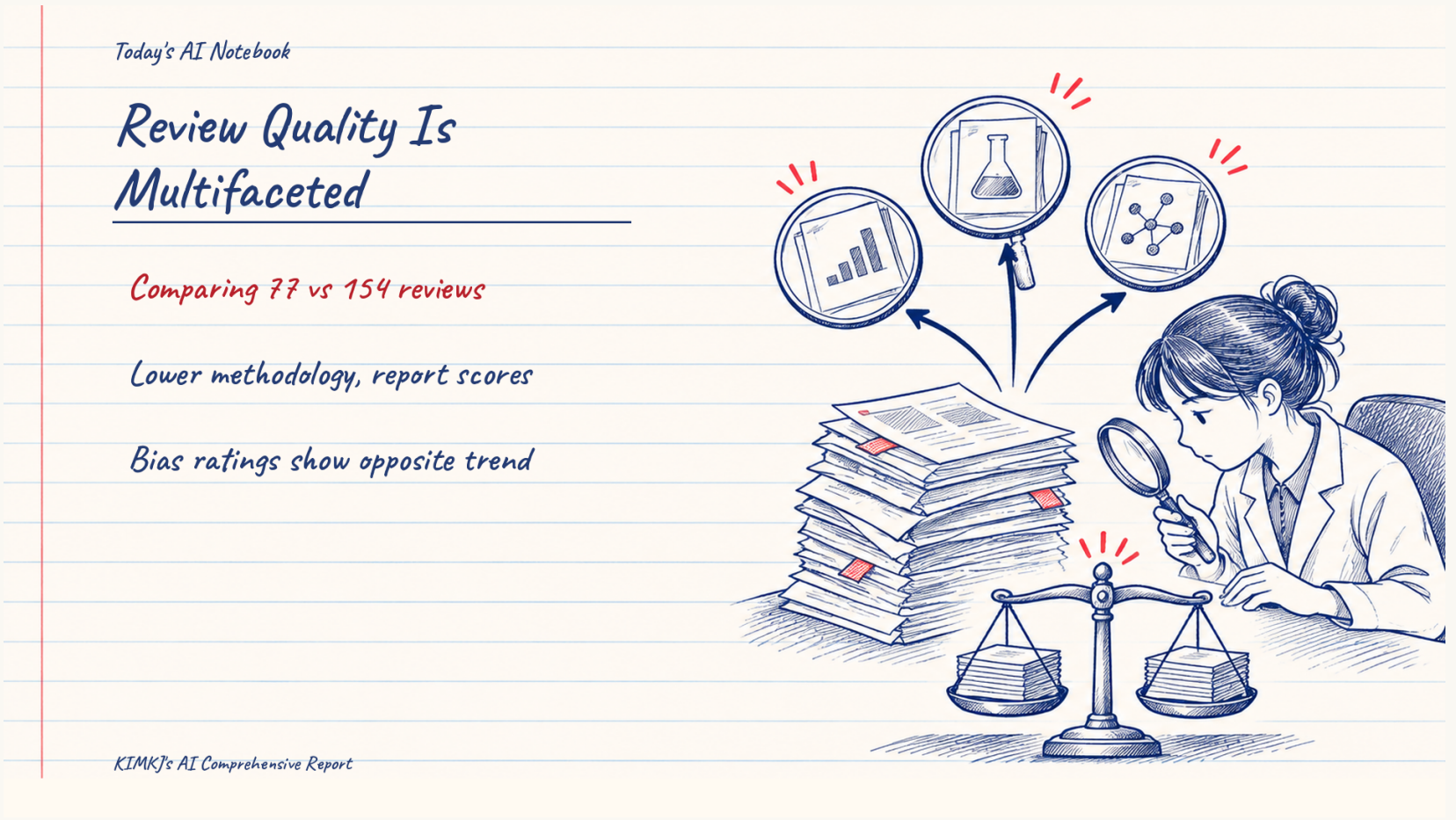


Research Quality

12 AI Literature Reviews Require Multi-Dimensional Quality Evaluation

The use of AI is growing in systematic reviews that synthesize multiple medical studies to draw clinical conclusions. A study published on November 8 compared 77 AI-assisted systematic reviews published between 2023 and 2025 with 154 reviews that did not report AI use. Methodological adherence, reporting completeness, and risk of bias were evaluated using distinct assessment tools.

In adjusted analyses, the AI-assisted group scored 1.64 points lower on methodological rigor, and its reporting adherence rate for main analyses was 4.7 percentage points lower. However, among papers evaluable for risk of bias, the proportion of low risk was 75.0% in the AI-assisted group compared to 65.6% in the comparison group, pointing in a different direction. These findings show that a single score cannot determine whether an entire study is good or bad. The control group may also contain undeclared AI use, and observational comparisons cannot establish AI as the causal factor behind differences. As a preprint, the study should be read as evidence supporting simultaneous examination of methodological rigor, reporting transparency, and bias.

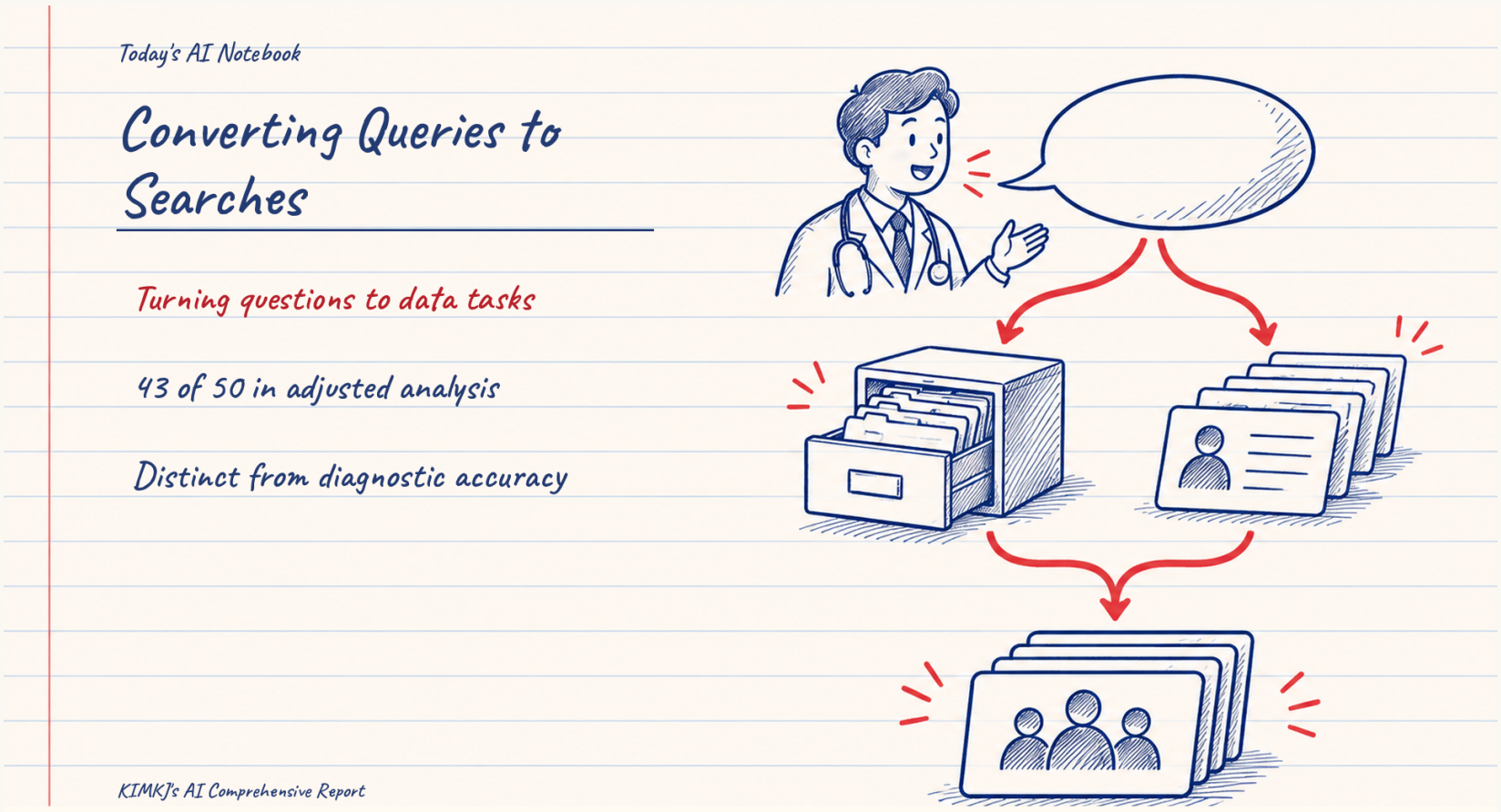


Clinical Research Automation

13 Converting Spoken Research Queries into Medical Record Searches

Experiments have emerged where an AI agent searches electronic health records when a clinical researcher describes a target patient cohort in natural language. The system breaks down natural-language requests into data-specific tasks, constructing and executing database queries. It identifies cohorts matching research criteria by parsing both tabular structured records and clinicians' unstructured clinical notes.

A preprint study published on November 6 compared 50 pediatric orthopedic queries against results from human data analysts. Among these, 43 queries were included in the analysis adjusted through manual review. The median adjusted positive predictive value was 0.93, sensitivity over the combined cohort retrieved by either method was 0.92, and the F1 score was 0.90. The evaluation scope of sensitivity must not be extrapolated to all hospital patients. While results show promise for cohort identification tasks, they do not validate diagnostic accuracy or performance across other clinical specialties. The figures also do not reveal how much time or cost was saved.

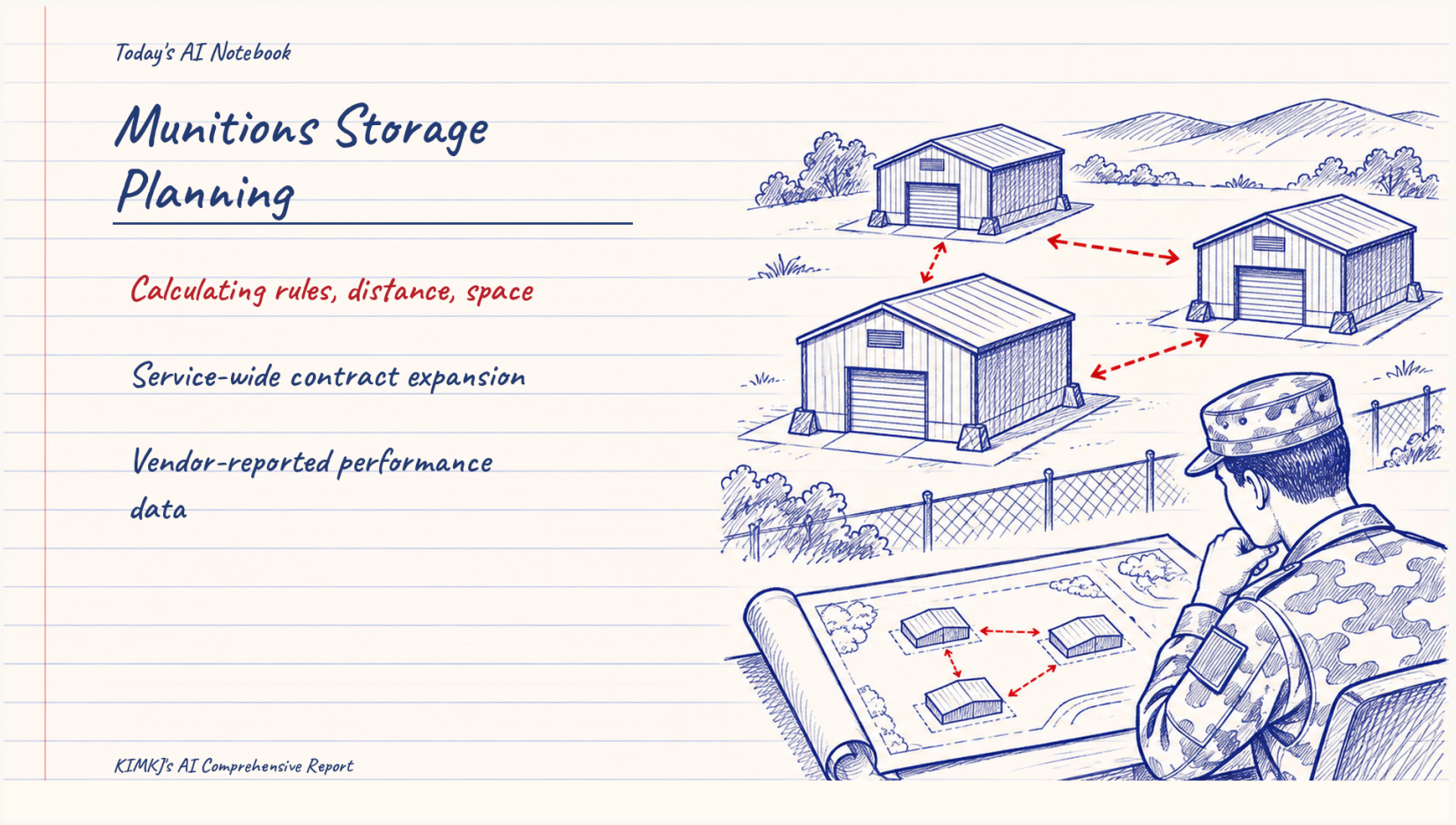


Military AI

14 US Air Force Expands Deployment of Munitions Storage Planning AI

Virtualitics announced on November 3 that its contract for the U.S. Air Force's ammunition storage planning system, A-MSP, has been expanded service-wide. The system calculates ammunition storage regulations, safety separation distances, and facility layout constraints concurrently. It supports logistics operations determining where and how to store munitions, representing a domain distinct from weapon targeting or automated engagement.

According to the company's announcement, the prior operational footprint encompassed seven major commands and hundreds of users. Planning time was reduced by 70%, while manual processes slated for replacement were estimated to take over 50,000 hours annually. At Air Combat Command, re-evaluating existing storage spaces also helped curb expenditures on new facilities. These figures rely on vendor-reported operational metrics and estimates. The contract expansion itself does not imply that deployment and operation across all bases are complete, nor can individual results be extrapolated as uniform savings across the entire Air Force. Aligning plans while complying with storage regulations remains central to this application.



Urban Safety

15 Demonstrating Explainable Urban Flood Forecasts for Forecasters

The Han River Flood Control Office and the Korea Institute of Science and Technology Information (KISTI) launched a partnership in August to advance urban flood forecasting around Gangnam Station and Sindaebang Station. The plan integrates high-resolution rainfall analysis, rapid numerical simulations utilizing high-performance computing, and AI that detects flooding from CCTV feeds. The initiative validates field applicability using past torrential rain events and observational records.

This includes explainable artificial intelligence (XAI) that presents the rationale behind specific AI predictions. The goal is to assist forecasters in examining the underlying reasoning alongside alert outputs. It represents an approach to bridge AI forecasts with human professional judgment in disaster management operations. The details examined here concern an agreement and demonstration initiatives, rather than a newly finalized system completed in November. Quantitative metrics regarding accuracy gains or false alarm reductions have not yet been provided. The technological integration plan must be evaluated separately from actual reductions in flood damage.

