Trending Papers

by AK and the research community

Search any paper with AI

Daily | Weekly | Monthly | Trending Papers

Qwen3-Omni Technical Report

We present Qwen3-Omni, a single multimodal model that, for the first time, maintains state-of-the-art performance across text, image, audio, and video without any degradation relative to single-modal counterparts. Qwen3-Omni matches the performance of same-sized single-modal models within the Qwen series and excels particularly on audio tasks. Across…

38 authors · Published on Sep 22, 2025

Upvote 107 | GitHub 2.11k | arXiv Page

WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent

Web agents such as Deep Research have demonstrated superhuman cognitive abilities, capable of solving highly challenging information-seeking problems. However, most research remains primarily text-centric, overlooking visual information in the real world. This makes multimodal Deep Research highly challenging, as such agents require much stronger…

14 authors · Published on Aug 8, 2025

Upvote 127 | GitHub 14.7k | arXiv Page

WebShaper: Agentically Data Synthesizing via Information-Seeking Formalization

The advent of Large Language Model (LLM)-powered agents has revolutionized artificial intelligence by enabling solutions to complex, open-ended tasks through web-based information-seeking (IS) capabilities. The scarcity of high-quality training data has limited the development of IS agents. Existing approaches typically adopt an information-driven…

13 authors · Published on Jul 21, 2025

Upvote 56 | GitHub 14.7k | arXiv Page

WebDancer: Towards Autonomous Information Seeking Agency

Addressing intricate real-world problems necessitates in-depth information seeking and multi-step reasoning. Recent progress in agentic systems, exemplified by Deep Research, underscores the potential for autonomous multi-step research. In this work, we present a cohesive paradigm for building end-to-end agentic information seeking agents from a data-…

12 authors · Published on May 29, 2025

Upvote 31 | GitHub 14.7k | arXiv Page

ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization

Large Language Model (LLM)-based web agents demonstrate strong performance on knowledge-intensive tasks but are hindered by context window limitations in paradigms like ReAct. Complex queries involving multiple entities, intertwined relationships, and high uncertainty demand extensive search cycles that rapidly exhaust context budgets before reaching…

14 authors · Published on Sep 17, 2025

Upvote 70 | GitHub 14.7k | arXiv Page

WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning

Transcending human cognitive limitations represents a critical frontier in LLM training. Proprietary agentic systems like DeepResearch have demonstrated superhuman capabilities on extremely complex information-seeking benchmarks such as BrowseComp, a feat previously unattainable. We posit that their success hinges on a sophisticated reasoning pattern…

17 authors · Published on Sep 17, 2025

Upvote 83 | GitHub 14.7k | arXiv Page

Scaling Agents via Continual Pre-training

Large language models (LLMs) have evolved into agentic systems capable of autonomous tool use and multi-step reasoning for complex problem-solving. However, post-training approaches building upon general-purpose foundation models consistently underperform in agentic tasks, particularly in open-source implementations. We identify the root cause: t…

22 authors · Published on Sep 17, 2025

Upvote 103 | GitHub 14.7k | arXiv Page

WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research

This paper tackles open-ended deep research (OEDR), a complex challenge where AI agents must synthesize vast web-scale information into insightful reports. Current approaches are plagued by dual-fold limitations: static research pipelines that decouple planning from evidence acquisition and one-shot generation paradigms that easily suffer from long-context…

12 authors · Published on Sep 17, 2025

Upvote 100 | GitHub 14.7k | arXiv Page

WebSailor: Navigating Super-human Reasoning for Web Agent

Transcending human cognitive limitations represents a critical frontier in LLM training. Proprietary agentic systems like DeepResearch have demonstrated superhuman capabilities on extremely complex information-seeking benchmarks such as BrowseComp, a feat previously unattainable. We posit that their success hinges on a sophisticated reasoning pattern…

19 authors · Published on Jul 3, 2025

Upvote 118 | GitHub 14.7k | arXiv Page

PaddleOCR 3.0 Technical Report

This technical report introduces PaddleOCR 3.0, an Apache-licensed open-source toolkit for OCR and document parsing. To address the growing demand for document understanding in the era of large language models, PaddleOCR 3.0 presents three major solutions: (1) PP-OCRv5 for multilingual text recognition, (2) PP-StructureV3 for hierarchical document parsin…

19 authors · Published on Jul 8, 2025

Upvote 8 | GitHub 56k | arXiv Page

Lyra: Generative 3D Scene Reconstruction via Video Diffusion Model Self-Distillation

The ability to generate virtual environments is crucial for applications ranging from gaming to physical AI domains such as robotics, autonomous driving, and industrial AI. Current learning-based 3D reconstruction methods rely on the availability of captured real-world multi-view data, which is not always readily available. Recent advancements in video diffusio…

13 authors · Published on Sep 24, 2025

Upvote 18 | GitHub 272 | arXiv Page

SciReasoner: Laying the Scientific Reasoning Ground Across Disciplines

We present a scientific reasoning foundation model that aligns natural language with heterogeneous scientific representations. The model is pretrained on a 206B-token corpus spanning scientific text, pure sequences, and sequence-text pairs, then aligned via SFT on 40M instructions, annealed cold-start bootstrapping to elicit long-form chain-of-thought, an…

32 authors · Published on Sep 26, 2025

Upvote 63 | GitHub 33 | arXiv Page

Tree Search for LLM Agent Reinforcement Learning

Recent advances in reinforcement learning (RL) have significantly enhanced the agentic capabilities of large language models (LLMs). In long-term and multi-turn agent tasks, existing approaches driven solely by outcome rewards often suffer from the problem of sparse supervision. To address the challenge, we propose Tree-based Group Relative Policy Optimizati…

6 authors · Published on Sep 25, 2025

Upvote 49 | GitHub 40 | arXiv Page

AgentScope 1.0: A Developer-Centric Framework for Building Agentic Applications

Driven by rapid advancements of Large Language Models (LLMs), agents are empowered to combine intrinsic knowledge with dynamic tool use, greatly enhancing their capacity to address real-world tasks. In line with such an evolution, AgentScope introduces major improvements in a new version (1.0), towards comprehensively supporting flexible and efficient…

23 authors · Published on Aug 22, 2025

Upvote 49 | GitHub 12.5k | arXiv Page

VLA-Adapter: An Effective Paradigm for Tiny-Scale Vision-Language-Action Model

Vision-Language-Action (VLA) models typically bridge the gap between perceptual and action spaces by pre-training a large-scale Vision-Language Model (VLM) on robotic data. While this approach greatly enhances performance, it also incurs significant training costs. In this paper, we investigate how to effectively bridge vision-language (VL) representations to acti…

16 authors · Published on Sep 11, 2025

Upvoted 214 | GitHub 424 | arXiv Page

LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models

Efficient fine-tuning is vital for adapting large language models (LLMs) to downstream tasks. However, it requires non-trivial efforts to implement these methods on different models. We present LlamaFactory, a unified framework that integrates a suite of cutting-edge efficient training methods. It allows users to flexibly customize the fine-tuning of 100+ LLMs without…

5 authors · Published on Mar 20, 2024

Upvote 145 | GitHub 59.4k | arXiv Page

Qwen-Image Technical Report

We present Qwen-Image, an image generation foundation model in the Qwen series that achieves significant advances in complex text rendering and precise image editing. To address the challenges of complex text rendering, we design a comprehensive data pipeline that includes large-scale data collection, filtering, annotation, synthesis, and balancing. Moreover,…

39 authors · Published on Aug 4, 2025

Upvote 255 | GitHub 5.35k | arXiv Page

Paper2Agent: Reimagining Research Papers As Interactive and Reliable AI Agents

We introduce Paper2Agent, an automated framework that converts research papers into AI agents. Paper2Agent transforms research output from passive artifacts into active systems that can accelerate downstream use, adoption, and discovery. Conventional research papers require readers to invest substantial effort to understand and adapt a paper's code, data,…

4 authors · Published on Sep 9, 2025

Upvote 31 | GitHub 669 | arXiv Page

4KAgent: Agentic Any Image to 4K Super-Resolution

We present 4KAgent, a unified agentic super-resolution generalist system designed to universally upscale any image to 4K resolution (and even higher, if applied iteratively). Our system can transform images from extremely low resolutions with severe degradations, for example, highly distorted inputs at 256x256, into crystal-clear, photorealistic 4K outputs. 4KAgent…

13 authors · Published on Jul 10, 2025

Upvote 101 | GitHub 505 | arXiv Page

Mobile-Agent-v3: Fundamental Agents for GUI Automation

This paper introduces GUI-Owl, a foundational GUI agent model that achieves state-of-the-art performance among open-source end-to-end models on ten GUI benchmarks across desktop and mobile environments, covering grounding, question answering, planning, decision-making, and procedural knowledge. GUI-Owl-7B achieves 66.4 on AndroidWorld and…

15 authors · Published on Aug 21, 2025

Upvote 63 | GitHub 5.5k | arXiv Page

Look Before You Leap: A GUI-Critic-R1 Model for Pre-Operative Error Diagnosis in GUI Automation

In recent years, Multimodal Large Language Models (MLLMs) have been extensively utilized for multimodal reasoning tasks, including Graphical User Interface (GUI) automation. Unlike general offline multimodal tasks, GUI automation is executed in online interactive environments, necessitating step-by-step decision-making based on real-time status of the…

12 authors · Published on Jun 5, 2025

Upvote 18 | GitHub 5.9k | arXiv Page

PC-Agent: A Hierarchical Multi-Agent Collaboration Framework for Complex Task Automation on PC

In the field of MLLM-based GUI agents, compared to smartphones, the PC scenario not only features a more complex interactive environment, but also involves more intricate intra- and inter-app workflows. To address these issues, we propose a hierarchical agent framework named PC-Agent. Specifically, from the perception perspective, we devise an Active Percepti…

11 authors · Published on Feb 20, 2025

Upvote 29 | GitHub 5.9k | arXiv Page

UI-S1: Advancing GUI Automation via Semi-online Reinforcement Learning

Graphical User Interface (GUI) agents have demonstrated remarkable progress in automating complex user interface interactions through reinforcement learning. However, current approaches face a fundamental dilemma: offline RL enables stable training on pre-collected trajectories, but struggles with multi-step task execution for lack of trajectory-level rewar…

11 authors · Published on Sep 15, 2025

Upvote 46 | GitHub 5.9k | arXiv Page

MMR1: Enhancing Multimodal Reasoning with Variance-Aware Sampling and Open Resources

Large multimodal reasoning models have achieved rapid progress, but their advancement is constrained by two major limitations: the absence of open, large-scale, high-quality long chain-of-thought (CoT) data, and the instability of reinforcement learning (RL) algorithms in post-training. Group Relative Policy Optimization (GRPO), the standard framework for RL…

15 authors · Published on Sep 25, 2025

Upvote 61 | GitHub 176 | arXiv Page

A Survey of Reinforcement Learning for Large Reasoning Models

In this paper, we survey recent advances in Reinforcement Learning (RL) for reasoning with Large Language Models (LLMs). RL has achieved remarkable success in advancing the frontier of LLM capabilities, particularly in addressing complex logical tasks such as mathematics and coding. As a result, RL has emerged as a foundational methodology for transforming LL…

39 authors · Published on Sep 11, 2025

Upvote 166 | GitHub 1.42k | arXiv Page

DINOv3

Self-supervised learning holds the promise of eliminating the need for manual data annotation, enabling models to scale effortlessly to massive datasets and larger architectures. By not being tailored to specific tasks or domains, this training paradigm has the potential to learn visual representations from diverse sources, ranging from natural to aerial images —…

26 authors · Published on Aug 14, 2025

Upvote 261 | GitHub 7.31k | arXiv Page

MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe

Multimodal Large Language Models (MLLMs) are undergoing rapid progress and represent the frontier of AI development. However, their training and inference efficiency have emerged as a core bottleneck in making MLLMs more accessible and scalable. To address the challenges, we present MiniCPM-V 4.5, an 8B parameter model designed for high efficiency and…

34 authors · Published on Sep 17, 2025

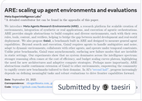
Upvote 43 | GitHub 22k | arXiv Page

InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

We introduce InternVL3, a significant advancement in the InternVL series featuring a native multimodal pre-training paradigm. Rather than adapting a text-only large language model (LLM) into a multimodal large language model (MLLM) that supports visual inputs, InternVL3 jointly acquires multimodal and linguistic capabilities from both diverse multimodal data…

47 authors · Published on Apr 15, 2025

Upvote 291 | GitHub 7.91k | arXiv Page



ARE: Scaling Up Agent Environments and Evaluations

We introduce Meta Agents Research Environments (ARE), a research platform for scalable creation of environments, integration of synthetic or real applications, and execution of agentic orchestrations. ARE provides simple abstractions to build complex and diverse environments, each with their own rules, tools, content, and verifiers, helping to bridge the gap…

24 authors · Published on Sep 22, 2025

Upvote 28
GitHub 234
arXiv Page

ScaleCUA: Scaling Open-Source Computer Use Agents with Cross-Platform Data

Vision-Language Models (VLMs) have enabled computer use agents (CUAs) that operate GUIs autonomously, showing great potential, yet progress is limited by the lack of large-scale, open-source computer use data and foundation models. In this work, we introduce ScaleCUA, a step toward scaling open-source CUAs. It offers a large-scale dataset spanning 6…

21 authors · Published on Sep 19, 2025

Upvote 101
GitHub 251
arXiv Page

Qwen3 Technical Report

In this work, we present Qwen3, the latest version of the Qwen model family. Qwen3 comprises a series of large language models (LLMs) designed to advance performance, efficiency, and multilingual capabilities. The Qwen3 series includes models of both dense and Mixture-of-Expert (MoE) architectures, with parameter scales ranging from 0.6 to 235 billion. A key…

60 authors · Published on May 14, 2025

Upvote 293
GitHub 24.8k
arXiv Page

Logics-Parsing Technical Report

Recent advances in Large Vision-Language models (LVLM) have spurred significant progress in document parsing task. Compared to traditional pipeline-based methods, end-to-end paradigms have shown their excellence in converting PDF images into structured outputs through integrated Optical Character Recognition (OCR), table recognition, mathematical…

10 authors · Published on Sep 24, 2025

Upvote 4
GitHub 45
arXiv Page

Collaborating Action by Action: A Multi-agent LLM Framework for Embodied Reasoning

Collaboration is ubiquitous and essential in day-to-day life -- from exchanging ideas, to delegating tasks, to generating plans together. This work studies how LLMs can adaptively collaborate to perform complex embodied reasoning tasks. To this end we introduce MINDcraft, an easily extensible platform built to enable LLM agents to control characters in the…

8 authors · Published on Apr 25, 2025

Upvote 5
GitHub 4.21k
arXiv Page

VoXtream: Full-Stream Text-to-Speech with Extremely Low Latency

We present VoXtream, a fully autoregressive, zero-shot streaming text-to-speech (TTS) system for real-time use that begins speaking from the first word. VoXtream directly maps incoming phonemes to audio tokens using a monotonic alignment scheme and a dynamic look-ahead that does not delay onset. Built around an incremental phoneme transformer, a…

3 authors · Published on Sep 19, 2025

Upvote 2
GitHub 119
arXiv Page

VolSplat: Rethinking Feed-Forward 3D Gaussian Splatting with Voxel-Aligned Prediction

Feed-forward 3D Gaussian Splatting (3DGS) has emerged as a highly effective solution for novel view synthesis. Existing methods predominantly rely on a pixel-aligned Gaussian prediction paradigm, where each 2D pixel is mapped to a 3D Gaussian. We rethink this widely adopted formulation and identify several inherent limitations: it renders the reconstruct…

10 authors · Published on Sep 24, 2025

Upvote 22
GitHub 64
arXiv Page

Youtu-GraphRAG: Vertically Unified Agents for Graph Retrieval-Augmented Complex Reasoning

Graph retrieval-augmented generation (GraphRAG) has effectively enhanced large language models in complex reasoning by organizing fragmented knowledge into explicitly structured graphs. Prior efforts have been made to improve either graph construction or graph retrieval in isolation, yielding suboptimal performance, especially when domain shifts occur. In thi…

9 authors · Published on Aug 27, 2025

Upvote 7
GitHub 689
arXiv Page

LIMI: Less is More for Agency

We define Agency as the emergent capacity of AI systems to function as autonomous agents actively discovering problems, formulating hypotheses, and executing solutions through self-directed engagement with environments and tools. This fundamental capability marks the dawn of the Age of AI Agency, driven by a critical industry shift: the urgent need for AI…

21 authors · Published on Sep 22, 2025

Upvote 87
GitHub 76
arXiv Page

Easy Dataset: A Unified and Extensible Framework for Synthesizing LLM Fine-Tuning Data from Unstructured Documents

Large language models (LLMs) have shown impressive performance on general-purpose tasks, yet adapting them to specific domains remains challenging due to the scarcity of high-quality domain data. Existing data synthesis tools often struggle to extract reliable fine-tuning data from heterogeneous documents effectively. To address this limitation, we propose…

7 authors · Published on Jul 5, 2025

Upvote 45
GitHub 10.9k
arXiv Page

MAPO: Mixed Advantage Policy Optimization

Recent advances in reinforcement learning for foundation models, such as Group Relative Policy Optimization (GRPO), have significantly improved the performance of foundation models on reasoning tasks. Notably, the advantage function serves as a central mechanism in GRPO for ranking the trajectory importance. However, existing explorations encounter…

14 authors · Published on Sep 23, 2025

Upvote 24
GitHub 34
arXiv Page

3D and 4D World Modeling: A Survey

World modeling has become a cornerstone in AI research, enabling agents to understand, represent, and predict the dynamic environments they inhabit. While prior work largely emphasizes generative methods for 2D image and video data, they overlook the rapidly growing body of work that leverages native 3D and 4D representations such as RGB-D imagery,…

23 authors · Published on Sep 5, 2025

Upvote 57
GitHub 565
arXiv Page

Is Diversity All You Need for Scalable Robotic Manipulation?

Data scaling has driven remarkable success in foundation models for Natural Language Processing (NLP) and Computer Vision (CV), yet the principles of effective data scaling in robotic manipulation remain insufficiently understood. In this work, we investigate the nuanced role of data diversity in robot learning by examining three critical dimensions-task (what to d…

10 authors · Published on Jul 9, 2025

Upvote 20
GitHub 2.46k
arXiv Page

VGGT: Visual Geometry Grounded Transformer

We present VGGT, a feed-forward neural network that directly infers all key 3D attributes of a scene, including camera parameters, point maps, depth maps, and 3D point tracks, from one, a few, or hundreds of its views. This approach is a step forward in 3D computer vision, where models have typically been constrained to and specialized for single tasks. It is also…

6 authors · Published on Mar 15, 2025

Upvote 31
GitHub 11.1k
arXiv Page

Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning

Despite the rapid growth of machine learning research, corresponding code implementations are often unavailable, making it slow and labor-intensive for researchers to reproduce results and build upon prior work. In the meantime, recent Large Language Models (LLMs) excel at understanding scientific documents and generating high-quality code. Inspired by…

4 authors · Published on Apr 24, 2025

Upvote 119
GitHub 3.37k
arXiv Page

SIM-CoT: Supervised Implicit Chain-of-Thought

Implicit Chain-of-Thought (CoT) methods present a promising, token-efficient alternative to explicit CoT reasoning in Large Language Models (LLMs), but a persistent performance gap has limited the application of implicit CoT. We identify a core latent instability issue by scaling the computational budget of implicit CoT approaches: as we increase the number of…

8 authors · Published on Sep 25, 2025

Upvote 32
GitHub 33
arXiv Page

MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm

We introduce MonkeyOCR, a vision-language model for document parsing that advances the state of the art by leveraging a Structure-Recognition-Relation (SRR) triplet paradigm. This design simplifies what would otherwise be a complex multi-tool pipeline (as in MinerU's modular approach) and avoids the inefficiencies of processing full pages with giant end-to-en…

10 authors · Published on Jun 6, 2025

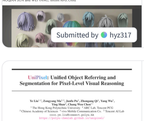
Upvote 2
GitHub 5.92k
arXiv Page

CHARM: Control-point-based 3D Anime Hairstyle Auto-Regressive Modeling

We present CHARM, a novel parametric representation and generative framework for anime hairstyle modeling. While traditional hair modeling methods focus on realistic hair using strand-based or volumetric representations, anime hairstyle exhibits highly stylized, piecewise-structured geometry that challenges existing techniques. Existing works often rely on…

9 authors · Published on Sep 25, 2025

Upvote 7
GitHub 7
arXiv Page



UniPixel: Unified Object Referring and Segmentation for Pixel-Level Visual Reasoning

Recent advances in Large Multi-modal Models (LMMs) have demonstrated their remarkable success as general-purpose multi-modal assistants, with particular focuses on holistic image- and video-language understanding. Conversely, less attention has been given to scaling fine-grained pixel-level understanding capabilities, where the models are expected to…

7 authors · Published on Sep 23, 2025

Upvote 3
GitHub 40
arXiv Page

MixGRPO: Unlocking Flow-based GRPO Efficiency with Mixed ODE-SDE

Although GRPO substantially enhances flow matching models in human preference alignment of image generation, methods such as FlowGRPO still exhibit inefficiency due to the necessity of sampling and optimizing over all denoising steps specified by the Markov Decision Process (MDP). In this paper, we propose MixGRPO, a novel framework that leverages t…

7 authors · Published on Jul 29, 2025

Upvote 13
GitHub 769
arXiv Page

FastVLM: Efficient Vision Encoding for Vision Language Models

Scaling the input image resolution is essential for enhancing the performance of Vision Language Models (VLMs), particularly in text-rich image understanding tasks. However, popular visual encoders such as ViTs become inefficient at high resolutions due to the large number of tokens and high encoding latency caused by stacked self-attention layers. At different…

11 authors · Published on Dec 18, 2024

Upvote 70
GitHub 6.65k
arXiv Page

VideoFrom3D: 3D Scene Video Generation via Complementary Image and Video Diffusion Models

In this paper, we propose VideoFrom3D, a novel framework for synthesizing high-quality 3D scene videos from coarse geometry, a camera trajectory, and a reference image. Our approach streamlines the 3D graphic design workflow, enabling flexible design exploration and rapid production of deliverables. A straightforward approach to synthesizing a video…

3 authors · Published on Sep 23, 2025

Upvote 22
GitHub 62
arXiv Page