Expanding the AI/ML Research Frontier: A Curated Compendium of Recent Innovations and Influential Works

Section 1: Introduction

1.1 Purpose and Scope of Research Expansion

This report details the systematic expansion of an initial Artificial Intelligence (AI) and Machine Learning (ML) research corpus. The primary objective is to broaden the existing research map by identifying and incorporating new, impactful academic papers. This recursive expansion focuses on literature that introduces significant innovations, inspires subsequent research, or represents paradigm shifts within the dynamic field of AI and ML.

The scope of this expansion prioritizes recent advancements, with a particular emphasis on works published between 2022 and 2025, especially those from 2024 and 2025. Concurrently, foundational works that continue to shape current research trajectories and provide essential context for understanding these recent developments are also acknowledged and included. The aim is to provide a comprehensive yet focused overview of the rapidly evolving landscape of AI/ML research.

1.2 Methodology for Paper Selection

The expansion process involved a meticulous analysis of the articles within the seed corpus. For each seed article, its bibliography and cited works were examined to identify potential new candidates for inclusion. This step was crucial for the recursive nature of the corpus expansion, ensuring that the research map grows organically from established and relevant literature.

Candidate papers were evaluated against several key criteria to ensure their relevance and impact:

- **State-of-the-Art (SOTA) Contributions:** Papers presenting novel methods, algorithms, or architectures that achieve benchmark-setting results in specific AI/ML sub-domains. These often represent the cutting edge of current capabilities.
- **High Citation Count or Influence:** Works recognized by the research community as significant or foundational. This is often indicated by a high number of citations shortly after publication or explicit mention in reputable surveys and review articles as "influential" or "highly-cited."
- **Recent Breakthroughs (2024–2025):** Cutting-edge research, including pre-prints from recognized archives (e.g., arXiv) and newly published peer-reviewed papers, that highlight the very latest developments and emerging trends.
- **Emerging Methods, Architectures, or Paradigms:** Papers introducing new conceptual

frameworks, algorithmic approaches, or system designs that suggest novel directions or potential paradigm shifts in AI/ML research.

A de-duplication step was performed to ensure that no articles already present in the original seed corpus were re-included in this expanded list. The selection process aimed for a balance between depth in specific innovative areas and breadth across the major themes currently driving AI/ML research.

1.3 Structure of the Report

This report is structured to present the findings of the research expansion in a clear and accessible manner.

Section 2 contains the core deliverable: the expanded list of research papers, designated as t000_dr04_article-list. This compendium is thematically organized to facilitate navigation and contextual understanding. Each entry includes essential metadata: title, authors, publication year, and a concise abstract or statement of the paper's primary contribution. Furthermore, a justification for its inclusion, based on the selection criteria outlined above, is provided.

Section 3 discusses key thematic observations and overarching trends derived from the analysis of the newly added literature. This section aims to synthesize the collective advancements and identify significant patterns in the evolution of AI/ML.

Section 4 provides concluding remarks, summarizing the value of the expanded corpus and suggesting potential future research trajectories indicated by the current state of the field.

This structured approach is intended to make the expanded research map not only a repository of important papers but also a tool for understanding the current dynamics and future potential of AI and Machine Learning.

Section 2: Expanded Research Compendium (t000_dr04_article-list)

This section presents the comprehensive list of newly identified research papers, forming the expanded research compendium. The papers are organized thematically to enhance usability and allow researchers to quickly navigate to areas of specific interest. Each entry includes the title, authors, publication year, a concise abstract or contribution statement, a justification for its inclusion based on the selection criteria, and, where traceable, the seed article or source that referenced it.

The thematic organization reflects major currents in AI/ML research. This structure facilitates a deeper understanding of the landscape and supports the identification of interconnected research efforts and emerging trends discussed later in this report.

Table 1: t000_dr04_article-list - Expanded Research Compendium

Paper ID (New)	Title	Authors	Publication Year	Abstract/Contribution Statement	Category Justification	Referenced In (Source Snippet)
2.1						

Emerging Architectures, Paradigms, and Foundational Concepts						
N001	Attention Is All You Need	Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.	2017	Introduced the Transformer architecture, which relies solely on self-attention mechanisms, dispensing with recurrence and convolutions entirely. This model achieves SOTA in machine translation and has become foundational for many subsequent NLP and vision models.	Foundational; Essential context for LLMs and ViTs. Widely cited as the basis for modern deep learning in sequence processing.	[1]
N002	Vision Transformer (ViT)	Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani,	2020 (published 2021)	Demonstrated that a pure Transformer architecture applied directly to sequences of image	Foundational for vision models; Paradigm shift in computer vision.	[3]

		M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.		patches can achieve SOTA results on image classification tasks, challenging the long-standing dominance of Convolutional Neural Networks (CNNs).		
N003	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.	2018 (published 2019)	Introduced BERT, a language representation model pre-trained using a masked language model objective and next sentence prediction. It achieved SOTA results on a wide array of NLP tasks and significantly influenced subsequent LLM development.	Foundational for NLP pre-training and many LLMs.	[5]
N004	Mamba: Linear-Time Sequence	Gu, A., & Dao, T.	2023	Introduces the Mamba architecture,	Emerging Architecture (2023);	[7]

	Modeling with Selective State Spaces			a state space model (SSM) that achieves Transformer-level performance with linear-time complexity in sequence length and faster inference. It uses a selection mechanism to allow context-dependent reasoning.	Highly cited; Potential paradigm shift for long-sequence modeling.	
N005	QLoRA: Efficient Finetuning of Quantized LLMs	Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L.	2023	Proposes QLoRA, an efficient finetuning approach that reduces memory usage significantly by quantizing a pretrained LLM to 4-bit and then finetuning a small set of Low-Rank Adapters (LoRA). Enables finetuning of very large	SOTA/Influential (2023) for LLM efficiency; Highly cited.	[7]

				models on limited hardware.		
N006	Sparks of Artificial General Intelligence: Early experiments with GPT-4	Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y.	2023	Provides an early, in-depth exploration of GPT-4's capabilities, suggesting it exhibits sparks of Artificial General Intelligence (AGI) through its performance on novel and difficult tasks across various domains without task-specific prompting.	Highly Cited/Influential (2023); Paradigm exploration regarding LLM capabilities.	[7]
2.2 State-of-the-Art (SOTA) Contributions and Recent Advances by AI Sub-Domain						
2.2.1 Computer Vision						
N007	Vision Transformers	Darcet, T., Cord, M.,	2024	ICLR 2024 Outstanding	SOTA/Influential (2024);	[8]

	Need Registers	Pérez-Pellitero, E., & Thome, N.		Paper Award winner. Demonstrates that adding a few extra "register" tokens to Vision Transformers can significantly improve their performance by providing a scratchpad space for global feature aggregation, without altering the core architecture.	ViT architectural refinement.	
N008	X-ray Imaging-driven Detection Network (XID-Net)	Not explicitly named, from paper summary.	2024	Proposes XID-Net, a network for X-ray prohibited item detection featuring a novel X-ray-specific augmentation strategy (Poisson blending for rare items) and contextual feature integration.	SOTA (2024) in a specialized CV domain (X-ray security).	9

				It significantly improves detection performance, outperforming popular SOTA methods by up to +17.2% in tail categories.		
N009	UniViTAR: A Unified Vision Transformer for Multi-Resolution Visual Recognition with Aspect-Ratio Preservation	Not explicitly named, from paper summary.	2024	Proposes UniViTAR, a Vision Transformer variant incorporating advanced modifications (2D Rotary Position Embedding, SwiGLU FFN, RMSNorm, QK-Norm) and a progressive training paradigm (resolution curriculum learning) to handle variable input resolutions and aspect ratios effectively.	Recent ViT advancement (2024); Emerging method for flexible visual input processing.	[4]
N010	ECViT: Efficient	Not explicitly named, from	2025	Introduces ECViT, a	Emerging Architecture	[10]

	Convolutional Vision Transformer with Local-Attention and Multi-scale Stages	paper summary.		hybrid architecture effectively combining CNN strengths (locality, translation invariance) and Transformers. Features Partitioned Multi-head Self-Attention (P-MSA) and Interactive Feed-forward Network (I-FFN) for optimal balance between performance and efficiency.	(2025); Focus on ViT efficiency and hybrid design.	
N011	SCHEME: Scalable CHannEl MixEr for Vision Transformers	Not explicitly named, from paper summary.	2023	Introduces SCHEME, a novel channel mixer for ViTs using a sparse Block Diagonal MLP (BD-MLP) structure and a parameter-free Channel Covariance Attention	Recent ViT advancement (2023); Efficiency and performance focus for channel mixers.	[12]

				(CCA) mechanism. This allows larger expansion ratios, improving accuracy/latency, especially for smaller transformers, without inference overhead from CCA.		
N012	AdaFace: Quality Adaptive Margin for Deep Face Recognition	Kim, M., Jain, A. K., & Liu, X.	2022	Proposes AdaFace, a loss function for face recognition that adaptively adjusts the margin based on image quality. This emphasizes easy samples less and hard samples more, improving performance, particularly for low-quality images.	SOTA/Influential (2022) in face recognition; Adaptive loss function.	[13]
N013	Vision Foundation Models in	Not explicitly named, from paper	2025	Reviews state-of-the-art research	Recent Survey (2025);	[3]

	Medical Image Analysis: Advances and Challenges	summary.		on adapting Vision Foundation Models (VFMs) like ViT and Segment Anything Model (SAM) to medical image segmentation. Discusses challenges and advancements in domain adaptation, model compression, federated learning, adapter-based improvements, knowledge distillation, and multi-scale contextual feature modeling.	Identifies SOTA/Emerging Trends in medical computer vision.	
N014	A Survey of Physics-Aware Generative AI in Computer Vision	Not explicitly named, from paper summary.	2025	Systematically reviews the emerging field of physics-aware generative AI in computer vision, categorizing methods	Recent Survey (2025); Emerging paradigm in generative computer vision.	[15]

				based on how they incorporate physical knowledge (explicit simulation or implicit learning) for generating realistic images, videos, and 3D/4D content.		
N015	Image Recognition with Online Lightweight Vision Transformer: A Survey	Not explicitly named, from paper summary.	2025	Surveys online strategies for generating lightweight Vision Transformers for image recognition, focusing on three key areas: Efficient Component Design, Dynamic Network, and Knowledge Distillation. Analyzes trade-offs on ImageNet-1K.	Recent Survey (2025) on Vision Transformer efficiency.	[16]
N016	Disentangled Source-Free Domain	Sharafi et al. (from FG 2025	2025	Introduces Disentangled Source-Free	Recent Breakthroug h (2025) in	[18]

	Adaptation (DSFDA) for Facial Expression Recognition	reference)		Domain Adaptation (DSFDA) for video-based Facial Expression Recognition (FER). DSFDA addresses missing target expression data by leveraging neutral target control video for end-to-end generation and adaptation, disentangling expression and identity features.	FER; Emerging method for SFDA.	
2.2.2 Natural Language Processing & Large Language Models (LLMs)						
N017	The Llama 3 Herd of Models	Grattafiori, A., et al.	2024	Presents Meta's Llama 3 family of open models (8B and 70B parameters), detailing their	SOTA/Influential (2024) LLM release; Open model advancement.	8

				architecture, extensive pretraining dataset (over 15T tokens), and improved performance, establishing new SOTA for open models of their size on various benchmarks.		
N018	Gemma: Open Models Based on Gemini Research and Technology	Mesnard, T., et al. (Google team)	2024	Introduces Google's Gemma family of open models (2B and 7B parameters), based on similar research and technology as the Gemini models. They outperform similarly sized open models on key benchmarks and include analysis of safety and responsibility aspects.	SOTA/Influential (2024) LLM release; Open model advancement.	[8]
N019	Why Larger Language Models Do	Min, Z., et al.	2024	Investigates how the scale of	Influential (2024) research on	[8]

	In-context Learning Differently?			Large Language Models affects their in-context learning mechanisms. Shows that small language models are more robust to noise and less easily distracted than LLMs due to emphasis on a narrower selection of hidden features.	core LLM capabilities (in-context learning).	
N020	DLPO: Towards a Robust, Efficient, and Generalizable Prompt Optimization Framework from a Deep-Learning Perspective	Peng, D., Zhou, Y., Chen, Q., Liu, J., Chen, J., & Qin, L.	2025	Proposes DLPO, a framework that applies deep learning optimization techniques (e.g., Textual Learning Rate, Textual Dropout, Textual Simulated Annealing, Textual Momentum) to automated prompt optimization	Emerging Method (2025) in prompt engineering; SOTA improvement.	[19]

				for LLMs, significantly enhancing stability, efficiency, and generalization.		
N021	A Survey of Large Language Models	Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.	2023	A comprehensive survey on LLMs, covering background, milestones, key techniques (pre-training, adaptation, utilization, capacity evaluation), available resources, and diverse applications. Highly cited overview of the LLM landscape.	Highly Cited/Influential (2023) survey on LLMs.	[7]
N022	The Flan Collection: Designing Data and Methods for Effective Instruction Tuning	Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., & Roberts, A.	2023	Details the creation of the Flan dataset collection, comprising over 1800 NLP tasks formatted as instructions, and demonstrates methods	Highly Cited/Influential (2023) work on instruction tuning.	[7]

				for instruction tuning that significantly improve LLM generalization and performance on unseen tasks.		
N023	LIMA: Less Is More for Alignment	Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.	2023	Shows that LLMs can learn to produce high-quality responses from only a small set (around 1000) of carefully curated prompts and responses, without needing extensive reinforcement learning from human feedback (RLHF). Challenges the notion that massive alignment data is necessary.	Highly Cited/Influential (2023) work on LLM alignment.	[7]
N024	Phi-3.5-Vision-Instruct	Microsoft	2024	A lightweight multimodal model (MLLM) developed	Recent Breakthrough (Oct 2024) in MLLMs; Lightweight	[21]

				by Microsoft, designed for a wide range of vision–language tasks including general image understanding, OCR, chart/table comprehension, multi-image comparison, and video clip summarization. Uses a CLIP ViT-L/14 image encoder and a Phi-3.5-mini LLM.	MLLM.	
N025	Llama-3.2-11B-Vision-Instruct	Meta Platforms	2024	An 11-billion-parameter MLLM developed by Meta, designed to process both text and images simultaneously for multimodal conversations and visual reasoning	Recent Breakthrough (Sep 2024) in MLLMs.	[21]

				tasks. Built upon the Llama 3.1 architecture with a CLIP-based image encoder.		
N026	Pixtral-12B	Mistral AI	2024	A 12-billion-parameter MLLM by Mistral AI for understanding images and text simultaneously, enabling advanced multimodal reasoning. Comprises a CLIPA-based image encoder, a Mistral Nemo 12B LLM, and a multimodal decoder.	Recent Breakthrough (Oct 2024) in MLLMs.	[21]
N027	Language Modeling for the Future of Finance: A Quantitative Survey into Metrics, Tasks, and Data Opportunities	Tatarinov, N., Sukhani, S., Shah, A., & Chava, S.	2025	Reviews 374 NLP research papers (2017-2024) applied to finance, with a focused analysis of 221 papers. Identifies key trends such as increasing use of	Recent Survey (2025) on NLP/LLM applications in finance.	[22]

				general-purpose language models, progress in sentiment analysis and information extraction, and emerging efforts in explainability and privacy-preserving methods.		
N028	LLMs for Explainable AI: A Comprehensive Survey	Bilal, A., Ebert, D., & Lin, B.	2025	Provides a comprehensive overview of existing approaches for using LLMs to enhance Explainable AI (XAI), transforming complex machine learning outputs into easy-to-understand narratives and bridging the gap between model behavior and human interpretability. Discusses	Recent Survey (2025) on LLM applications for XAI.	[23]

				evaluation techniques, challenges, and applications.		
N029	LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models	Hie, B., Kim, S. Y., Huang, A., Brynjolfsson, E., & Zou, J.	2025	A data-driven review of research on LLM limitations (LLLMs) from 2022 to 2024, analyzing ~14,600 papers. Identifies key concerns like reasoning failures, hallucinations, safety, and controllability, and tracks their evolution in research focus.	Recent Survey (2025) on critical LLM aspects and research trends.	[25]
2.2.3 Reinforcement Learning & LLM Agents						
N030	Random Policy Enables In-Context Reinforcement Learning within Trust Horizons	Not explicitly named, from paper summary.	2025	Proposes State-Action Distillation (SAD), a novel approach to generate pretraining	Recent Breakthrough (May 2025) in RL; SOTA in ICRL.	[27]

				datasets for In-Context Reinforcement Learning (ICRL) using only random policies within a trust horizon. SAD distills outstanding state-action pairs and significantly outperforms existing SOTA ICRL algorithms.		
N031	Decision Pretrained Transformer (DPT)	Lee, K., et al. [27]	2024	An ICRL method that partially relaxes the requirement on context (can be gathered by random policies) but necessitates access to optimal policies to label optimal actions for query states.	SOTA/Recent (2024) in ICRL; Compared against by SAD.	[27]
N032	Decision Importance Transformer (DIT)	Dong, Q., et al. [27]	2024	An ICRL method aiming for ICRL without optimal policies by leveraging observed	SOTA/Recent (2024) in ICRL; Compared against by SAD.	[27]

				state-action pairs in context data as queries/labels, weighted by return-to-go. Still requires substantial context with a good portion from well-trained policies.		
N033	AutoConcierge	Zeng, Z., Zhang, R., Zhang, Y., Li, Y., & Nanas, N.	2024	An LLM-powered agent for conversational restaurant recommendation. It uses natural language conversations to understand user needs, collect preferences, and uses the LLM to understand and generate language, providing explainable personalized recommendations.	Emerging Method (2024) in LLM agents for Recommender Systems.	[29]
N034	AgentCF++: Memory-enh	Liu, J., Gu, S., Li, D.,	2025	An enhanced version of	Emerging Method	[31]

	anced LLM-based Agents for Popularity-aware Cross-domain Recommendations	Zhang, G., Han, M., Gu, H., Zhang, P., Lu, T., Shang, L., & Gu, N.		AgentCF that introduces a dual-layer memory architecture (domain-separated and domain-fused) and interest groups with group-shared memory to improve LLM-based agent decision-making in cross-domain recommendations and better capture popularity factor influences.	(2025) in LLM agents for Recommender Systems.	
N035	RecMind: Large Language Model Powered Agent For Recommendation	Wang, Y., et al.	2023 (arXiv) / 2024 (conference)	An innovative recommender agent fueled by LLMs, designed as an autonomous agent to furnish personalized recommendations through	Influential/Emerging (2023/2024) in LLM agents for Recommender Systems.	[33]

				strategic planning (using a Self-Inspiring algorithm that considers previously explored paths) and the utilization of external tools.		
N036	Agent4Rec	Zhang, J., et al.	2024a	A novel framework that uses LLM-based generative agents (equipped with persona, memory, and action capabilities) to mimic user interactions with recommender systems. This enables online evaluation of recommendation policies and investigation of phenomena like filter bubbles	Emerging Method (2024) in LLM agents for Recommender System evaluation and simulation.	[35]

				without real users.		
N037	LASER: LLM Agent with State-Space Exploration for Web Navigation	Ma, K., Zhang, H., Wang, H., Pan, X., Yu, W., & Yu, D.	2024	Proposes LASER, an LLM agent that models web navigation as state-space exploration. The agent transitions among pre-defined states by performing actions, enabling flexible backtracking and state-specific action spaces, significantly outperforming previous methods.	Emerging Method (2024) in LLM agents for web navigation.	[37]
N038	CoSearchAgent: A Lightweight Collaborative Search Agent with Large Language Models	Gong, P., Li, J., & Mao, J.	2024	A lightweight collaborative search agent powered by LLMs, designed as a Slack plugin. It supports collaborative search during multi-party conversations by	Emerging Method (2024) in LLM agents for collaborative search.	[39]

				understanding queries and context, searching the web via APIs, and responding with grounded answers or clarifying questions.		
N039	AVATAR: Optimizing LLM Agents for Tool Usage via Contrastive Reasoning	Wu, S., Zhao, S., et al.	2024	Introduces AVATAR, an automated framework with an actor LLM and a comparator LLM. The comparator generates holistic prompts by contrastively reasoning between positive and negative examples to teach the actor LLM more effective retrieval strategies and tool usage, improving performance on complex multimodal retrieval and	Emerging Method (2024) for LLM agent tool use optimization.	[36]

				QA.		
N040	USimAgent: Large Language Models for Simulating Search Users	Zhang, E., Wang, X., Gong, P., Lin, Y., & Mao, J.	2024	Introduces USimAgent, an LLM-based user search behavior simulator capable of simulating querying, clicking, and stopping behaviors to generate complete search sessions for specific tasks. Outperforms existing methods in query generation.	Emerging Method (2024) for user simulation with LLMs in search.	[41]
N041	A Survey of Large Language Model Empowered Agents for Recommendation and Search	Zhang, Y., Qiao, S., Zhang, J., Lin, T.-H., Gao, C., & Li, Y.	2025	Systematically reviews and classifies research on LLM agents in recommendation and search. Establishes a classification framework based on agent roles (e.g., user interaction, representati	Recent Survey (2025) on LLM agents for information retrieval.	[43]

				on optimization for RecSys; task decomposers, query rewriters for Search).		
N042	A Survey on LLM-powered Agents for Recommender Systems	Peng, Q., Liu, H., Huang, H., Yang, Q., & Shao, M.	2025	Presents a comprehensive review of LLM-powered agents for recommender systems, categorizing approaches into recommender-oriented, interaction-oriented, and simulation-oriented paradigms. Analyzes architectural components: profile construction, memory management, strategic planning, and action execution.	Recent Survey (2025) specifically on LLM agents for Recommender Systems.	[30]
N043	A Survey on Explainable Deep Reinforcement Learning	Not explicitly named, from paper summary.	2025	Provides a comprehensive review of Explainable Deep Reinforcement Learning	Recent Survey (2025) on Explainable Reinforcement Learning.	[45]

				(XRL) methods, their qualitative and quantitative assessment frameworks, and their role in policy refinement, adversarial robustness, and security. Also examines RLHF for AI alignment.		
2.2.4 General Machine Learning, Optimization, and Cross-Cutting Concerns						
N044	ZeroFlow: Overcoming Catastrophic Forgetting is Easier than You Think	Not explicitly named, from paper summary.	2025	Introduces the ZeroFlow benchmark to evaluate gradient-free optimization algorithms for overcoming catastrophic forgetting in continual learning. Finds that forward	Emerging Method/Benchmark (2025) in continual learning and gradient-free optimization.	[46]

				passes alone can be sufficient and reveals new optimization principles for managing task conflicts and memory demands.		
N045	Exposing Limitations of Language Model Agents in Sequential-Task Compositions on the Web (CompWoB)	Furuta, H., Matsuo, Y., Faust, A., & Gur, I.	2024	Introduces CompWoB, a benchmark with 50 compositional web automation tasks, to study LMA transferability to realistic sequential task compositions. Shows performance degradation on compositional tasks and trains HTML-T5++ which achieves SOTA zero-shot performance on CompWoB.	Recent Benchmark (2024) for Language Model Agents on web tasks.	[48]
N046	SlimLLM: An Effective and Fast	Not explicitly named, from paper	2024/2025	Proposes SlimLLM, an effective and	Emerging Method (2024/2025)	[50]

	Structured Pruning Method for Large Language Models	summary.		fast structured pruning method for LLMs. For channel and attention head pruning, it evaluates importance based on the entire channel or head, rather than aggregating individual element importance.	in LLM optimization/compression.	
N047	MoRE: Mixture of Low-Rank Experts for Multi-Task Parameter-Efficient Fine-Tuning	Not explicitly named, from paper summary.	2024/2025	Proposes Mixture of Low-Rank Experts (MoRE) for multi-task Parameter-Efficient Fine-Tuning (PEFT). It aligns different ranks of LoRA modules (low-rank experts) with different tasks using a novel adaptive rank selector,	Emerging Method (2024/2025) in PEFT and Multi-task learning.	[50]

				enhancing adaptability and efficiency.		
N048	On the Challenges and Opportunities in Generative AI	Dellaferrera, G., et al.	2025	Discusses fundamental shortcomings and unresolved challenges in large-scale generative AI models concerning expanding scope and adaptability (generalization, causality, heterogeneous data), optimizing efficiency (training, inference, evaluation), and ethical deployment (misinformation, privacy, fairness, interpretability, constraints).	Recent Survey (Mar 2025) on broad Generative AI challenges and opportunities.	[1]
2.3 Highly Cited and Influential Recent Works (2022-2025)						
N049	BLIP-2:	Li, J., Li, D.,	2023	Introduces	Highly	[7]

	Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models	Savarese, S., & Hoi, S.		BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. Achieves SOTA performance on various vision-language tasks with significantly fewer trainable parameters.	Cited/Influential (2023); SOTA in Vision-Language Pre-training.	
N050	InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning	Dai, W., Li, J., Li, D., Tiong, A. M., Zhao, J., Wang, W.,... & Hoi, S.	2023	Presents InstructBLIP, a vision-language instruction-tuning framework built upon BLIP-2. It introduces instruction-aware visual	Highly Cited/Influential (2023); SOTA in Vision-Language Instruction Tuning.	7

				feature extraction and corresponding instruction-tuned training, enabling strong zero-shot generalization on a wide range of vision-language tasks.		
N051	MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models	Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M.	2023	Proposes MiniGPT-4, which aligns a frozen visual encoder (ViT + Q-Former from BLIP-2) with an advanced frozen LLM (Vicuna) using just one trainable projection layer. Demonstrates capabilities similar to GPT-4, such as detailed image description generation and website creation from	Highly Cited/Influential (2023); Efficient MLLM alignment.	[7]

				handwritten drafts.		
N052	PaLM-E: An Embodied Multimodal Language Model	Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B.,... & Florence, P.	2023	Introduces PaLM-E, an embodied multimodal language model that integrates real-world continuous sensor modalities from robots (e.g., images, state estimation) directly into a language model. It demonstrates positive transfer learning for robotic tasks and visual-language tasks.	Highly Cited/Influential (2023); Paradigm for embodied AI and robotics.	[7]
N053	Visual Instruction Tuning (LLaVA)	Liu, H., Li, C., Wu, Q., & Lee, Y. J.	2023	Presents LLaVA (Large Language and Vision Assistant), an end-to-end trained large multimodal model that connects a vision encoder and an LLM for	Highly Cited/Influential (2023); SOTA in visual instruction following.	[7]

				general-purpose visual and language understanding. It uses language-only GPT-4 to generate multimodal language-image instruction-following data.		
2.4 Recent Breakthroughs (Primarily 2024–2025, especially pre-prints and newly published)						
This subsection highlights papers from 2024-2025 already listed above, emphasizing their recency as breakthroughs. Examples include:						
N007	*Vision Transformers Need Registers*	*Darcet, T., et al.*	*2024*	*ICLR 2024 Outstanding Paper; adds "register" tokens to*	*Recent Breakthrough (2024); SOTA/Influential.*	[8]

				ViTs for improved global feature aggregation.		
N008	*X-ray Imaging-driven Detection Network (XID-Net)*	*Not explicitly named*	*2024*	*Novel X-ray-specific augmentation and contextual feature integration for improved prohibited item detection.*	*Recent Breakthrough (2024); SOTA in specialized CV.*	[9]
N010	*ECViT: Efficient Convolutional Vision Transformer*	*Not explicitly named*	*2025*	*Hybrid CNN-Transformer with P-MSA and I-FFN for efficient image classification.*	*Recent Breakthrough (2025); Emerging Architecture.*	[10]
N016	*DSFDA for Facial Expression Recognition*	*Sharafi et al.*	*2025*	*Disentangled Source-Free Domain Adaptation for video-based FER using neutral target control video.*	*Recent Breakthrough (2025); Emerging SFDA method.*	[18]
N017	*The Llama 3 Herd of Models*	*Grattafiori, A., et al.*	*2024*	*Meta's Llama 3 family of open models, SOTA for*	*Recent Breakthrough (2024); SOTA/Influential LLM.*	[8]

				their size.		
N020	*DLPO: Prompt Optimization Framework*	*Peng, D., et al.*	*2025*	*Applies deep learning optimization techniques to automated prompt optimization for LLMs.*	*Recent Breakthrough (2025); Emerging Method in prompt engineering.*	[19]
N024, N025, N026	*Phi-3.5-Vision, Llama-3.2-Vision, Pixtral-12B*	*Microsoft, Meta, Mistral AI*	*2024*	*New lightweight and powerful MLLMs for vision-language tasks.*	*Recent Breakthroughs (Late 2024) in MLLMs.*	[21]
N030	*Random Policy Enables ICRL (SAD)*	*Not explicitly named*	*2025*	*State-Action Distillation (SAD) for ICRL using random policies, outperforming SOTA.*	*Recent Breakthrough (May 2025) in RL.*	[27]
N033, N034, N036, N037, N038, N039, N040	*Various LLM Agent Papers (AutoConcierge, AgentCF++, Agent4Rec, LASER, CoSearchAgent, AVATAR, USimAgent)*	*Various*	*2024/2025*	*Novel LLM agent frameworks for recommendation, search, web navigation, tool use, and user simulation.*	*Recent Breakthroughs (2024/2025) in LLM Agents.*	**
N044	*ZeroFlow: Overcoming Catastrophic Forgetting*	*Not explicitly named*	*2025*	*Benchmark for gradient-free optimization in continual learning,*	*Recent Breakthrough (2025); Emerging Method/Benchmark.*	[46]

				showing forward passes can be effective.		
N046, N047	*SlimLLM, MoRE*	*Not explicitly named*	*2024/2025*	*New methods for structured pruning and multi-task PEFT for LLMs.*	*Recent Breakthroughs (2024/2025) in LLM Optimization.*	[50]
Various Surveys	*Surveys on LLM Agents, VFMs in Medical Imaging, NLP in Finance, LLMs for XAI, LLLMs limitations, Physics-aware GenAI, Lightweight ViTs, Explainable DRL*	*Various*	*2025*	*Comprehensive overviews of rapidly advancing subfields.*	*Recent Breakthroughs (2025) in knowledge synthesis.*	**

Section 3: Key Thematic Observations from the Expanded Corpus

The process of expanding the research corpus has illuminated several dominant themes and trajectories within contemporary AI/ML research. These observations, derived from the newly incorporated literature, provide a clearer picture of the field's current momentum and future directions.

3.1 Dominance of Transformer-based Architectures and their Evolution

The foundational impact of the Transformer architecture, first introduced by Vaswani et al. (2017) [N001], continues to be a central narrative in AI/ML. This is evident in the proliferation and advancement of its direct descendants: Vision Transformers (ViTs) for computer vision tasks [N002] and Large Language Models (LLMs) for natural language processing and beyond [N003, N017, N018]. A significant portion of recent research, as reflected in the expanded corpus, is dedicated to refining these architectures. For instance, efforts to enhance ViT efficiency and capability are prominent, with innovations like the SCHEME channel mixer [N011], the addition of "register" tokens for improved global feature aggregation [N007], the development of UniViTAR for flexible input resolutions [N009], and the creation of hybrid CNN-ViT models like ECViT for better efficiency [N010]. Surveys dedicated to lightweight ViTs further underscore this trend [N015].

While Transformer-based models are undeniably dominant, the research landscape also reveals a growing exploration of alternative or complementary architectures designed to address specific limitations of Transformers. The Mamba architecture, for example, offers linear-time complexity for sequence modeling, presenting a potentially more scalable solution for very long sequences compared to the quadratic complexity of standard attention mechanisms [N004]. Similarly, hybrid models like ECViT aim to reintroduce beneficial inductive biases from CNNs into the ViT framework [N010]. This suggests a maturation of the field: after the initial widespread adoption of a powerful architecture, the community is now systematically identifying its weaknesses (e.g., computational cost, data hunger, lack of certain biases) and proposing targeted solutions or entirely new paradigms. This evolutionary process is critical for pushing the boundaries of what is practically achievable with deep learning models.

3.2 Rise of LLM-Powered Agentic Systems

A striking trend emerging from the recent literature is the rapid development and application of LLM-powered agentic systems. This marks a significant conceptual shift from models that primarily process information or generate content to systems designed to act, interact, and achieve goals autonomously within digital or even physical environments. Several recent surveys specifically map this burgeoning area, highlighting the architectural components (profile, memory, planning, action) and operational paradigms (recommender-oriented, interaction-oriented, simulation-oriented) of these agents [N041, N042].

The applications of these LLM agents are diverse and expanding. In recommender systems, agents like AutoConcierge [N033] and RecMind [N035] leverage LLMs for natural language interaction, preference understanding, and explainable recommendations. For information retrieval and search, agents such as CoSearchAgent [N038] facilitate collaborative search, while others focus on complex query decomposition and result synthesis [N041]. Web navigation is another domain where LLM agents like LASER demonstrate advanced capabilities in executing multi-step tasks [N037]. Furthermore, frameworks like AVATAR are being developed to optimize how these agents utilize external tools [N039], and systems like USimAgent employ LLMs to simulate complex user behaviors for evaluation purposes [N040]. The development of specialized benchmarks like CompWoB [N045] to test the compositional abilities of web agents indicates the field's drive towards more robust and generalizable agentic capabilities. This progression towards more autonomous and goal-oriented AI systems has profound implications for automation, human-computer interaction, and the very nature of intelligent systems.

3.3 Multimodality as a Key Frontier

The integration of information from diverse modalities—text, images, video, audio, and structured data—stands out as a critical frontier in AI research. The expanded corpus reflects intense activity in this area, particularly with the advent and refinement of Multimodal Large Language Models (MLLMs). Recent MLLM releases, such as Microsoft's Phi-3.5-Vision-Instruct [N024], Meta's Llama-3.2-11B-Vision-Instruct [N025], and Mistral AI's Pixtral-12B [N026], all appearing in late 2024, underscore the rapid advancements in creating models that can jointly process and reason about different types of data. These models often build upon strong unimodal foundation models, combining powerful vision encoders (frequently ViT-based) with capable LLMs.

The development of sophisticated MLLMs is enabling new applications and enhancing existing ones. Multimodal recommender systems, for instance, aim to leverage diverse data sources for more nuanced and accurate recommendations, a topic explored in recent surveys.[52] Foundational research in vision-language understanding, exemplified by highly influential works like BLIP-2 [N049], InstructBLIP [N050], MiniGPT-4 [N051], PaLM-E for embodied AI [N052], and LLaVA [N053], has laid the groundwork for these more integrated MLLMs. The core challenge in this domain often revolves around effective strategies for aligning and fusing representations from different modalities, ensuring that the combined information leads to emergent capabilities rather than a mere aggregation of unimodal strengths. The progress in unimodal foundation models (e.g., increasingly powerful LLMs and ViTs) directly fuels the potential of MLLMs, making the co-evolution of these areas a key dynamic in the field.

3.4 Emphasis on Efficiency, Scalability, and Responsible AI

As AI models, particularly LLMs and large ViTs, continue to grow in size and complexity, practical considerations related to their deployment and societal impact have become paramount. This is clearly reflected in the research trends observed in the expanded corpus.

A significant body of work is dedicated to improving model efficiency and scalability. Techniques for efficient LLM fine-tuning, such as QLoRA [N005], allow large models to be adapted with substantially reduced memory footprints. Structured pruning methods like SlimLLM [N046] aim to remove redundant parameters without sacrificing performance, while innovations in Parameter-Efficient Fine-Tuning (PEFT) like MoRE [N047] explore more effective ways to adapt models for multiple tasks.

Alongside efficiency, the challenge of continual learning—enabling models to learn new information or tasks sequentially without catastrophically forgetting previous knowledge—remains a critical research area. The ZeroFlow benchmark and associated findings suggest that even gradient-free optimization methods can play a role in mitigating forgetting [N044].

Concurrently, there is a growing emphasis on responsible AI development. The need for Explainable AI (XAI) is increasingly recognized, with research exploring how LLMs themselves can be used to generate human-understandable explanations for complex model decisions, as detailed in a recent survey [N028]. Addressing fairness, identifying and mitigating biases in models and data, and ensuring user privacy are also crucial concerns. The proliferation of research on the limitations of LLMs, systematically reviewed in surveys like "LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models" [N029], and broader discussions on the challenges in generative AI [N048], highlight the community's commitment to understanding and mitigating potential negative impacts. This dual focus on pushing the boundaries of capability while simultaneously addressing the practical and ethical dimensions of AI deployment is a hallmark of a maturing field.

3.5 Proliferation of Specialized Surveys and Benchmarks

The rapid pace of innovation across various AI/ML subfields has led to a notable increase in the publication of specialized surveys and the development of new benchmarks. The expanded corpus includes several such surveys published in 2024 and 2025, each attempting to map and synthesize knowledge within specific, fast-evolving domains. Examples include surveys on LLM-powered agents for recommendation and search [N041, N042], Vision Foundation Models in medical imaging [N013], online lightweight Vision Transformers [N015], the application of LLMs in finance [N027], LLMs for XAI [N028], the limitations of LLMs [N029], physics-aware generative AI [N014], and explainable deep reinforcement learning [N043].

This proliferation of surveys indicates that many subfields are experiencing rapid growth and diversification, necessitating dedicated efforts to consolidate current knowledge, identify key challenges, and outline future research directions. Similarly, the emergence of new benchmarks, such as CompWoB for evaluating the compositional task abilities of language model agents [N045] and ZeroFlow for assessing continual learning strategies [N044], reflects the need for standardized evaluation methods to measure progress on specific capabilities or to address newly identified challenges. This trend is a positive sign of a vibrant research ecosystem where rapid advancements are quickly followed by efforts to structure the acquired knowledge and rigorously measure further progress, thereby fueling a cycle of focused innovation.

Section 4: Concluding Remarks

4.1 Summary of Expansion and Value

The systematic expansion of the AI/ML research corpus, as detailed in this report, has successfully incorporated a significant number of state-of-the-art, highly influential, and recent breakthrough papers. The inclusion of these works, spanning foundational concepts to cutting-edge applications from 2022 to 2025, provides a substantially enriched and more current research map. The thematic organization of the expanded compendium, t000_dr04_article-list, offers a structured lens through which to view the evolving landscape of AI and Machine Learning.

The value of this expanded map lies in its utility for researchers and practitioners seeking to inform their current research directions, identify key innovations across various sub-domains, and understand the emerging paradigms that are likely to shape the future of the field. By highlighting not only specific advancements but also the overarching thematic trends, this work serves as a curated guide to the forefront of AI/ML research.

4.2 Future Research Trajectories Suggested by the Corpus

The analysis of the expanded corpus points towards several particularly promising and, in some cases, urgent research trajectories:

- **Robust and Reliable LLM Agents:** While the capabilities of LLM agents are rapidly advancing, ensuring their reliability, controllability, and safety, especially when interacting with complex environments or external tools, remains a critical challenge. Future work will likely focus on more robust planning mechanisms, better error handling, and verifiable decision-making processes.
- **Scalable and Interpretable Multimodal Learning:** The integration of multiple data modalities is a key driver of innovation. However, developing MLLMs that are not only powerful but also computationally scalable and whose cross-modal reasoning processes are interpretable is an ongoing research endeavor. Efficient fusion techniques and methods for explaining multimodal predictions will be crucial.
- **Inherently Explainable and Fair AI Systems:** Moving beyond post-hoc explanations, there is a growing need for AI systems that are designed with explainability and fairness as intrinsic properties. This involves developing new architectures and training methodologies that promote transparency and mitigate biases from the outset.
- **Next-Generation Architectures:** While Transformers remain dominant, the exploration of alternative architectures (e.g., Mamba-like SSMs, novel graph neural networks, hybrid models) that offer advantages in terms of efficiency, scalability for extremely long contexts, or different inductive biases will continue to be an important research avenue.
- **Domain-Specific Foundation Models and Adaptation:** Tailoring foundation models to specific scientific, industrial, or societal domains (e.g., medicine [N013], finance [N027], X-ray analysis [N008]) and developing efficient techniques for domain adaptation will unlock significant real-world value.

4.3 Potential Next Steps for Corpus Curation

The curation of a research corpus is an ongoing process, especially in a field as dynamic as AI/ML. Potential next steps to further enhance and maintain the value of this expanded map include:

- **Deeper Dives into Specific Sub-Themes:** Conducting more focused literature reviews on rapidly emerging sub-themes identified in Section 3, such as specific types of LLM agent architectures or novel approaches to continual learning.
- **Continuous Updating with New Pre-prints and Publications:** Implementing a strategy for regularly scanning key archives (e.g., arXiv) and conference proceedings to incorporate the latest relevant pre-prints and peer-reviewed publications, ensuring the corpus remains current.
- **Expansion to Related Fields:** Exploring the inclusion of highly relevant papers from adjacent fields that increasingly intersect with AI/ML, such as cognitive science (for insights into agent design and human-like reasoning), neuroscience (for biologically inspired AI), robotics (for embodied intelligence), and specific scientific domains where AI is being applied to accelerate discovery.

- **Enhanced Metadata and Linkages:** Augmenting the corpus with richer metadata, such as links to publicly available code repositories, datasets used, and potentially building a citation graph to visualize relationships between papers more explicitly.

By pursuing these avenues, the research map can continue to evolve as a valuable resource for navigating and contributing to the advancing frontier of Artificial Intelligence and Machine Learning.

Works cited

1. On the Challenges and Opportunities in Generative AI - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2403.00025
2. Tips for Optimizing GPU Performance Using Tensor Cores | NVIDIA Technical Blog, accessed June 2, 2025, https://developer.nvidia.com/blog/optimizing-gpu-performance-tensor-cores/
3. Vision Foundation Models in Medical Image Analysis: Advances and Challenges - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.14584v2
4. UniViTAR: Unified Vision Transformer with Native Resolution - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.01792v1
5. From Deep Learning to LLMs: A survey of AI in Quantitative Investment - arXiv, accessed June 2, 2025, https://arxiv.org/html/2503.21422v1
6. MutBERT: Probabilistic Genome Representation Improves Genomics Foundation Models, accessed June 2, 2025, https://www.biorxiv.org/content/10.1101/2025.01.23.634452v1.full-text
7. Analyzing the homerun year for LLMs: the top-100 most cited AI ..., accessed June 2, 2025, https://www.zeta-alpha.com/post/analyzing-the-homerun-year-for-llms-the-top-100-most-cited-ai-papers-in-2023-with-all-medals-for-o
8. 5 of the Most Influential Machine Learning Papers of 2024 ..., accessed June 2, 2025, https://machinelearningmastery.com/5-most-influential-machine-learning-papers-2024/
9. arXiv:2411.18078v2 [cs.CV] 11 Mar 2025, accessed June 2, 2025, https://arxiv.org/pdf/2411.18078
10. ECViT: Efficient Convolutional Vision Transformer with Local-Attention and Multi-scale Stages - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.14825v1
11. [Papierüberprüfung] ECViT: Efficient Convolutional Vision ..., accessed June 2, 2025, https://www.themoonlight.io/de/review/ecvit-efficient-convolutional-vision-transformer-with-local-attention-and-multi-scale-stages
12. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2312.00412
13. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2412.02198
14. Vision Foundation Models in Medical Image Analysis: Advances and Challenges - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2502.14584
15. Generative Physical AI in Vision: A Survey - arXiv, accessed June 2, 2025,

https://arxiv.org/html/2501.10928v1
16. Image Recognition with Online Lightweight Vision Transformer: A Survey - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.03113v2
17. Image Recognition with Online Lightweight Vision Transformer: A Survey - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.03113v1
18. arXiv:2503.20771v3 [cs.CV] 5 Apr 2025, accessed June 2, 2025, https://arxiv.org/pdf/2503.20771?
19. arXiv:2503.13413v3 [cs.CL] 19 Mar 2025, accessed June 2, 2025, https://arxiv.org/pdf/2503.13413
20. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2503.13413
21. Advancing Multimodal Large Language Models: Optimizing Prompt Engineering Strategies for Enhanced Performance - MDPI, accessed June 2, 2025, https://www.mdpi.com/2076-3417/15/7/3992
22. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2504.07274
23. LLMs for Explainable AI: A Comprehensive Survey - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.00125v1
24. LLMs for Explainable AI: A Comprehensive Survey - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2504.00125
25. LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.19240v1
26. [Literature Review] LLLMs: A Data-Driven Survey of Evolving ..., accessed June 2, 2025, https://www.themoonlight.io/en/review/lllms-a-data-driven-survey-of-evolving-research-on-limitations-of-large-language-models
27. Random Policy Enables In-Context Reinforcement Learning within Trust Horizons - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2410.19982
28. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2410.19982
29. A Survey on LLM-powered Agents for Recommender Systems - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.10050v1
30. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2502.10050
31. AgentCF++: Memory-enhanced LLM-based Agents for Popularity-aware Cross-domain Recommendations - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.13843v2
32. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2502.13843
33. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review - arXiv, accessed June 2, 2025, https://arxiv.org/html/2402.18590v2
34. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2402.18590
35. Multi-agents based User Values Mining for Recommendation - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.00981
36. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2503.05659
37. LASER: LLM Agent with State-Space Exploration for Web Navigation - arXiv, accessed June 2, 2025, https://arxiv.org/html/2309.08172v2
38. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2309.08172

39. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2402.06360
40. AVATAR: Optimizing LLM Agents for Tool Usage via Contrastive Reasoning - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2406.11200?
41. USimAgent: Large Language Models for Simulating Search Users - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2403.09142
42. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2403.09142
43. A Survey of Large Language Model Empowered Agents for Recommendation and Search: Towards Next-Generation Information Retrieval - arXiv, accessed June 2, 2025, https://arxiv.org/html/2503.05659v1
44. arXiv:2502.10050v1 [cs.IR] 14 Feb 2025, accessed June 2, 2025, https://arxiv.org/pdf/2502.10050
45. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2502.06869
46. ZeroFlow: Overcoming Catastrophic Forgetting is Easier than You Think - arXiv, accessed June 2, 2025, https://arxiv.org/html/2501.01045v3
47. [Literature Review] ZeroFlow: Overcoming Catastrophic Forgetting is ..., accessed June 2, 2025, https://www.themoonlight.io/en/review/zeroflow-overcoming-catastrophic-forgetting-is-easier-than-you-think
48. Exposing Limitations of Language Model Agents in Sequential-Task Compositions on the Web - arXiv, accessed June 2, 2025, https://arxiv.org/html/2311.18751v3
49. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2311.18751
50. Machine Learning - arXiv, accessed June 2, 2025, https://arxiv.org/list/cs.LG/new
51. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2403.00025
52. A Survey on Large Language Models in Multimodal Recommender Systems - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.09777v1
53. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.09777