Date	Paper Title	Link	Abstract	Notes	CItation			
01/24/2026	A Marketplace for AI-Generated Adult Content and Deepfakes	https://arxiv.org/pdf/2601.09117	Generative AI systems increasingly enable the production of highly realistic synthetic media. Civitai, a popular community-driven platform for AI-generated content, operates a monetized feature called Bounties, which allows users to commission the generation of content in exchange for payment. To examine how this mechanism is used and what content it incentivizes, we conduct a longitudinal analysis of all publicly available bounty requests collected over a 14-month period following the platform's launch. We find that the bounty marketplace is dominated by tools that let users steer AI models toward content they were not trained to generate. At the same time, requests for content that is "Not Safe For Work" are widespread and have increased steadily over time, now comprising a majority of all bounties. Participation in bounty creation is uneven, with 20% of requesters accounting for roughly half of requests. Requests for "deepfake"—media depicting identifiable real individuals— exhibit a higher concentration than other types of bounties. A nontrivial subset of these requests involves explicit deepfakes despite platform policies prohibiting such content. These bounties disproportionately target female celebrities, revealing a pronounced gender asymmetry in social harm. Together, these findings show how monetized, community-driven generative AI platforms can produce gendered harms, raising questions about consent, governance, and enforcement.					
01/24/2026	When MCP Servers Attack: Taxonomy, Feasibility, and Mitigation	https://arxiv.org/pdf/2509.24272	Model Context Protocol (MCP) servers enable AI applications to connect to external systems in a plug-and-play manner, but their rapid proliferation also introduces severe security risks. Unlike mature software ecosystems with rigorous vetting, MCP servers still lack standardized review mechanisms, giving adversaries opportunities to distribute malicious implementations. Despite this pressing risk, the security implications of MCP servers remain underexplored. To address this gap, we present the first systematic study that treats MCP servers as active threat actors and decomposes them into core components to examine how adversarial developers can implant malicious intent. Specifically, we investigate three research questions: (i) what types of attacks malicious MCP servers can launch, (ii) how vulnerable MCP hosts and Large Language Models (LLMs) are to these attacks, and (iii) how feasible it is to carry out MCP server attacks in practice. Our study proposes a component-based taxonomy comprising twelve attack categories. For each category, we develop Proof-of-Concept (PoC) servers and demonstrate their effectiveness across diverse real-world host–LLM settings. We further show that attackers can generate large numbers of malicious servers at virtually no cost. We then test state-ofthe-art scanners on the generated servers and found that existing detection approaches are insufficient. These findings highlight that malicious MCP servers are easy to implement, difficult to detect with current tools, and capable of causing concrete damage to AI agent systems. Addressing this threat requires coordinated efforts among protocol designers, host developers, LLM providers, and end users to build a more secure and resilient MCP ecosystem.					
01/31/2026	ELICITING HARMFUL CAPABILITIES BY FINE-TUNING ON SAFEGUARDED OUTPUTS	https://arxiv.org/pdf/2601.13528	Model developers implement safeguards in frontier models to prevent misuse, for example, by employing classifiers to filter dangerous outputs. In this work, we demonstrate that even robustly safeguarded models can be used to elicit harmful capabilities in open-source models through elicitation attacks. Our elicitation attacks consist of three stages: (i) constructing prompts in adjacent domains to a target harmful task that do not request dangerous information; (ii) obtaining responses to these prompts from safeguarded frontier models; (iii) fine-tuning opensource models on these prompt-output pairs. Since the requested prompts cannot be used to directly cause harm, they are not refused by frontier model safeguards. We evaluate these elicitation attacks within the domain of hazardous chemical synthesis and processing, and demonstrate that our attacks recover approximately 40% of the capability gap between the base open-source model and an unrestricted frontier model. We then show that the efficacy of elicitation attacks scales with the capability of the frontier model and the amount of generated fine-tuning data. Our work demonstrates the challenge of mitigating ecosystem level risks with output-level safeguards.					
01/31/2026	Invisible Prompts, Visible Threats: Malicious Font Injection in External Resources for Large Language Models	-	2025	Malicious Font embedding (Code-to-glyph can be used to manipulate LLMs into generating desired response.				
01/31/2026	Formalizing and Benchmarking Prompt Injection Attacks and Defenses	-	2024					
01/31/2026	Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks	https://arxiv.org/abs/2508.00828						

01/31/2026	TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks	https://openreview.net/pdf/b533993ef9bc8320779646b1c475e47635dd98c2.pdf			Mention in related work and results section			
01/31/2026	DocBench: A Benchmark for Evaluating LLM-based Document Reading Systems	https://arxiv.org/html/2407.10701v1	Recently, there has been a growing interest among large language model (LLM) developers in LLM-based document reading systems, which enable users to upload their own documents and pose questions related to the document contents, going beyond simple reading comprehension tasks. Consequently, these systems have been carefully designed to tackle challenges such as file parsing, metadata extraction, multi-modal information understanding and long-context reading. However, no current benchmark exists to evaluate their performance in such scenarios, where a raw file and questions are provided as input, and a corresponding response is expected as output. In this paper, we introduce DocBench, a new benchmark designed to evaluate LLM-based document reading systems. Our benchmark involves a meticulously crafted process, including the recruitment of human annotators and the generation of synthetic questions. It includes 229 real documents and 1,102 questions, spanning across five different domains and four major types of questions. We evaluate both proprietary LLM-based systems accessible via web interfaces or APIs, and a parse-then-read pipeline employing open-source LLMs. Our evaluations reveal noticeable gaps between existing LLM-based document reading systems and human performance, underscoring the challenges of developing proficient systems. To summarize, DocBench aims to establish a standardized benchmark for evaluating LLM-based document reading systems under diverse real-world scenarios, thereby guiding future advancements in this research area	Use to show accuracy drop				
01/31/2026	BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models	https://openreview.net/forum?id=wCu6T5xFjeJ	Existing neural information retrieval (IR) models have often been studied in homogeneous and narrow settings, which has considerably limited insights into their out-of-distribution (OOD) generalization capabilities. To address this, and to facilitate researchers to broadly evaluate the effectiveness of their models, we introduce Benchmarking-IR (BEIR), a robust and heterogeneous evaluation benchmark for information retrieval. We leverage a careful selection of 18 publicly available datasets from diverse text retrieval tasks and domains and evaluate 10 state-of-the-art retrieval systems including lexical, sparse, dense, late-interaction, and re-ranking architectures on the BEIR benchmark. Our results show BM25 is a robust baseline and re-ranking and late-interaction based models on average achieve the best zero-shot performances, however, at high computational costs. In contrast, dense and sparse-retrieval models are computationally more efficient but often underperform other approaches, highlighting the considerable room for improvement in their generalization capabilities. We hope this framework allows us to better evaluate and understand existing retrieval systems and contributes to accelerating progress towards more robust and generalizable systems in the future. BEIR is publicly available at https://github.com/UKPLab/beir.					
01/31/2026	LongDA: Benchmarking LLM Agents for Long-Document Data Analysis	https://arxiv.org/pdf/2601.02598						
01/31/2026	OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations	https://arxiv.org/pdf/2412.07626	Document content extraction is a critical task in computer vision, underpinning the data needs of large language models (LLMs) and retrieval-augmented generation (RAG) systems. Despite recent progress, current document parsing methods have not been fairly and comprehensively evaluated due to the narrow coverage of document types and the simplified, unrealistic evaluation procedures in existing benchmarks. To address these gaps, we introduce OmniDocBench, a novel benchmark featuring high-quality annotations across nine document sources, including academic papers, textbooks, and more challenging cases such as handwritten notes and densely typeset newspapers. OmniDocBench supports flexible, multi-level evaluations—ranging from an end-to-end assessment to the task-specific and attribute-based analysis—using 19 layout categories and 15 attribute labels. We conduct a thorough evaluation of both pipeline-based methods and endto-end vision-language models, revealing their strengths and weaknesses across different document types. OmniDocBench sets a new standard for the fair, diverse, and fine-grained evaluation in document parsing. Dataset and code are available at https://github.com/opendatalab/OmniDocBench.					
01/31/2026	OfficeQA	https://www.databricks.com/blog/introducing-officeqa-benchmark-end-to-end-grounded-reasoning	There are multiple benchmarks that probe the frontier of agent capabilities (GDPval, Humanity's Last Exam (HLE), ARC-AGI-2), but we do not find them representative of the kinds of tasks that are important to our customers. To fill this gap, we've created and are open-sourcing OfficeQA—a benchmark that proxies for economically valuable tasks performed by Databricks' enterprise customers. We focus on a very common yet challenging enterprise task: Grounded Reasoning, which involves answering questions based on complex proprietary datasets that include unstructured documents and tabular data. Despite frontier models performing well on Olympiad-style questions, we find they still struggle on these economically important tasks. Without access to the corpus, they answer ~2% of questions correctly. When provided with a corpus of PDF documents, agents perform at <45% accuracy across all questions and <25% on a subset of the hardest questions.	We attempted to evaluate the recently released Gemini File Search Tool API as part of a representative Gemini Agent baseline with Gemini 3. However, about 30% of the PDFs and parsed PDFs in the OfficeQA corpus failed to ingest, and the File Search Tool is incompatible with the Google Search Tool. Since this would limit the agent from answering OfficeQA questions that need external knowledge, we excluded this setup from our baseline evaluation. We'll revisit it once ingestion works reliably so we can measure its performance accurately				

01/31/2026	TelAgentBench: A Multi-faceted Benchmark for Evaluating LLM-based Agents in Telecommunication	https://aclanthology.org/2025.emnlp-industry.83.pdf	As Large Language Models (LLMs) evolve into powerful agentic systems, the telecommunications industry's expansion into AI services necessitates industry-grounded benchmarks to evaluate their underexplored domain-specific capabilities. To address the gap left by generic benchmarks that fail to assess realistic, nonEnglish performance, we present TelAgentBench, a Korean benchmark for the telecommunications domain evaluating five core agentic capabilities: Reasoning, Planning, Action (tool-use), Retrieval-Augmented Generation, and Instruction Following. Evaluations reveal significant performance disparities between models that employ explicit reasoning and those that do not, providing actionable insights for deploying agentic LLMs in realworld telecommunications tasks.					
01/31/2026	READOC: A Unified Benchmark for Realistic Document Structured Extraction	https://aclanthology.org/2025.findings-acl.1128.pdf	Document Structured Extraction (DSE) aims to extract structured content from raw documents. Despite the emergence of numerous DSE systems, their unified evaluation remains inadequate, significantly hindering the field's advancement. This problem is largely attributed to existing benchmark paradigms, which exhibit fragmented and localized characteristics. To offer a thorough evaluation of DSE systems, we introduce a novel benchmark named READOC, which defines DSE as a realistic task of converting unstructured PDFs into semantically rich Markdown. The READOC dataset is derived from 3,576 diverse and real-world documents from arXiv, GitHub, and Zenodo. In addition, we develop a DSE Evaluation S3uite comprising Standardization, Segmentation and Scoring modules, to conduct a unified evaluation of state-of-theart DSE approaches. By evaluating a range of pipeline tools, expert visual models, and general Vision-Language Models, we identify the gap between current work and the unified, realistic DSE objective for the first time. We aspire that READOC will catalyze future research in DSE, fostering more comprehensive and practical solutions					
01/31/2026	AGENT-SAFETYBENCH: Evaluating the Safety of LLM Agents	https://arxiv.org/pdf/2412.14470	As large language models (LLMs) are increasingly deployed as agents, their integration into interactive environments and tool use introduce new safety challenges beyond those associated with the models themselves. However, the absence of comprehensive benchmarks for evaluating agent safety presents a significant barrier to effective assessment and further improvement. In this paper, we introduce AGENTSAFETYBENCH, a comprehensive benchmark designed to evaluate the safety of LLM agents. AGENT-SAFETYBENCH encompasses 349 interaction environments and 2,000 test cases, evaluating 8 categories of safety risks and covering 10 common failure modes frequently encountered in unsafe interactions. Our evaluation of 16 popular LLM agents reveals a concerning result: none of the agents achieves a safety score above 60%. This highlights significant safety challenges in LLM agents and underscores the considerable need for improvement. Through failure mode and helpfulness analysis, we summarize two fundamental safety defects in current LLM agents: lack of robustness and lack of risk awareness. Furthermore, our findings suggest that reliance on defense prompts alone may be insufficient to address these safety issues, emphasizing the need for more advanced and robust strategies. To drive progress in this area, AGENT-SAFETYBENCH has been released 1 to facilitate further research in agent safety evaluation and improvement					
01/31/2026	R-Judge: Benchmarking Safety Risk Awareness for LLM Agents	https://aclanthology.org/2024.findings-emnlp.79.pdf	Large language models (LLMs) have exhibited great potential in autonomously completing tasks across real-world applications. However, LLM agents introduce unexpected safety risks when operating in interactive environments. Instead of centering on the harmlessness of LLMgenerated content in most prior studies, this work addresses the imperative need for benchmarking the behavioral safety of LLM agents within diverse environments. We introduce RJudge, a benchmark crafted to evaluate the proficiency of LLMs in judging and identifying safety risks given agent interaction records. R-Judge comprises 569 records of multi-turn agent interaction, encompassing 27 key risk scenarios among 5 application categories and 10 risk types. It is of high-quality curation with annotated safety labels and risk descriptions. Evaluation of 11 LLMs on R-Judge shows considerable room for enhancing the risk awareness of LLMs: The best-performing model, GPT-4o, achieves 74.45% while no other models significantly exceed the random. Moreover, we reveal that risk awareness in open agent scenarios is a multi-dimensional capability involving knowledge and reasoning, thus challenging for LLMs. With further experiments, we find that fine-tuning on safety judgment significantly improves model performance while straightforward prompting mechanisms fail. R-Judge is publicly available at https://github.com/Lordog/R-Judge.					

01/31/2026	Random Guy benchmark	https://github.com/rapturt9/SafetyAdherenceBenchmark						
01/31/2026	SAFEAGENTBENCH: A BENCHMARK FOR SAFE TASK PLANNING OF EMBODIED LLM AGENT	https://openreview.net/pdf/761ab439917c7731ed6b6ce975ab860f530245a8.pdf	With the integration of large language models (LLMs), embodied agents have strong capabilities to understand and plan complicated natural language instructions. However, a foreseeable issue is that those embodied agents can also flawlessly execute some hazardous tasks, potentially causing damages in the real world. Existing benchmarks predominantly overlook critical safety risks, focusing solely on planning performance, while a few evaluate LLMs' safety awareness only on noninteractive image-text data. To address this gap, we present SafeAgentBench—the first comprehensive benchmark for safety-aware task planning of embodied LLM agents in interactive simulation environments, covering both explicit and implicit hazards. SafeAgentBench includes: (1) an executable, diverse, and high-quality dataset of 750 tasks, rigorously curated to cover 10 potential hazards and 3 task types; (2) SafeAgentEnv, a universal embodied environment with a low-level controller, supporting multi-agent execution with 17 high-level actions for 9 stateof-the-art baselines; and (3) reliable evaluation methods from both execution and semantic perspectives. Experimental results show that, although agents based on different design frameworks exhibit substantial differences in task success rates, their overall safety awareness remains weak. The most safety-conscious baseline achieves only a 10% rejection rate for detailed hazardous tasks. Moreover, simply replacing the LLM driving the agent does not lead to notable improvements in safety awareness. Dataset and codes are available and shown in the reproducibility statement.	ICLR 2026				
01/31/2026	AGENTHARM: A BENCHMARK FOR MEASURING HARMFULNESS OF LLM AGENTS	https://arxiv.org/pdf/2410.09024	The robustness of LLMs to jailbreak attacks, where users design prompts to circumvent safety measures and misuse model capabilities, has been studied primarily for LLMs acting as simple chatbots. Meanwhile, LLM agents—which use external tools and can execute multi-stage tasks—may pose a greater risk if misused, but their robustness remains underexplored. To facilitate research on LLM agent misuse, we propose a new benchmark called AgentHarm. The benchmark includes a diverse set of 110 explicitly malicious agent tasks (440 with augmentations), covering 11 harm categories including fraud, cybercrime, and harassment. In addition to measuring whether models refuse harmful agentic requests, scoring well on AgentHarm requires jailbroken agents to maintain their capabilities following an attack to complete a multi-step task. We evaluate a range of leading LLMs, and find (1) leading LLMs are surprisingly compliant with malicious agent requests without jailbreaking, (2) simple universal jailbreak templates can be adapted to effectively jailbreak agents, and (3) these jailbreaks enable coherent and malicious multi-step agent behavior and retain model capabilities.	ICLR 2025				
01/31/2026	Benchmarking the Robustness of Agentic Systems to Adversarially-Induced Harms	https://arxiv.org/html/2508.16481v2	Ensuring the safe use of agentic systems requires a thorough understanding of the range of malicious behaviors these systems may exhibit when under attack. In this paper, we evaluate the robustness of LLM-based agentic systems against attacks that aim to elicit harmful actions from agents. To this end, we propose a novel taxonomy of harms for agentic systems and a novel benchmark, BAD-ACTS, for studying the security of agentic systems with respect to a wide range of harmful actions. BAD-ACTS consists of five implementations of agentic systems in distinct application environments, as well as a dataset of 188 high-quality examples of harmful actions and an extended dataset containing 699 additional adversarial actions. This enables a comprehensive study of the robustness of agentic systems across a wide range of categories of harmful behaviors, available tools, and inter-agent communication structures. Using this benchmark, we analyze the robustness of agentic systems against an attacker that controls one of the agents in the system and aims to manipulate other agents to execute a harmful target action. Our results show that the attack has a high success rate, demonstrating that even a single adversarial agent within the system can have a significant impact on the security. This attack remains effective even when agents use a simple prompting-based defense strategy. However, we additionally propose a more effective defense based on zero-shot message monitoring. We believe that this benchmark provides a diverse testbed for the security research of agentic system					

01/31/2026	Counterfeit Answers: Adversarial Forgery against OCR-Free Document Visual Question Answering	https://www.arxiv.org/pdf/2512.04554	Document Visual Question Answering (DocVQA) enables end-to-end reasoning grounded on information present in a document input. While recent models have shown impressive capabilities, they remain vulnerable to adversarial attacks. In this work, we introduce a novel attack scenario that aims to forge document content in a visually imperceptible yet semantically targeted manner, allowing an adversary to induce specific or generally incorrect answers from a DocVQA model. We develop specialized attack algorithms that can produce adversarially forged documents tailored to different attackers' goals, ranging from targeted misinformation to systematic model failure scenarios. We demonstrate the effectiveness of our approach against two end-to-end stateof-the-art models: Pix2Struct, a vision-language transformer that jointly processes image and text through sequence-tosequence modeling, and Donut, a transformer-based model that directly extracts text and answers questions from document images. Our findings highlight critical vulnerabilities in current DocVQA systems and call for the development of more robust defenses. We release our open source code at https://anonymous.4open.science/r/adv docVQA-E7C5.	. We conduct our experiments in the PFL-DocVQA dataset [13]. This dataset contains real documents related to invoices, in which each invoice is associated with a question and multiple answers. It is originally designed to test existing privacy techniques on multi-modal DocVQA scenarios. In total, it contains 336, 842 question-answer pairs on 117, 661 pages, resulting in 37, 669 documents from 6, 574 different providers. Although the authors provide a Blue Team/Red Team split to separate the data between training and privacy attack evaluation, we merged them to obtain a single unified set. To build our evaluation set, we extracted N = 1000 unique samples from the merged data, where each sample is composed of a single image x and exactly M = 5 associated QA pairs. Models. We consider two state-of-the-art DocVQA models: Pix2Struct-Base [2] and Donut [3], which propose endto-end architectures designed for OCR-free document understanding. We us	https://github.com/rubenpt91/PFL-DocVQA-Competition/tree/master/datasets			
01/31/2026	A Survey of Recent Advances in Adversarial Attack and Defense on Vision-Language Models	https://www.preprints.org/manuscript/202511.1363	In the rapidly advancing domain of artificial intelligence, Vision-Language Models (VLMs) have emerged as critical tools by synergizing visual and textual data processing to facilitate a multitude of applications including automated image captioning, accessibility enhancements, and intelligent responses to multimodal queries. This survey explores the evolving paradigm of Pre-training, Fine-tuning, and Inference that has notably enhanced the capabilities of VLMs, allowing them to perform effectively across various downstream tasks and even enable zero-shot predictions. Despite their advancements, VLMs are vulnerable to adversarial attacks, largely because of their reliance on large-scale, internet-sourced pre-training datasets. These attacks can significantly undermine the models' integrity by manipulating their input interpretations, posing severe security risks and eroding user trust. Our survey delves into the complexities of these adversarial threats, which range from single-modal to sophisticated multimodal strategies, highlighting the urgent need for robust defense mechanisms. We discuss innovative defense strategies that adapt model architectures, integrate adversarially robust training objectives, and employ fine-tuning techniques to counteract these vulnerabilities. This paper aims to provide a comprehensive overview of current challenges and future directions in the adversarial landscape of VLMs, emphasizing the importance of securing these models to ensure their safe integration into various real-world applications.					
01/31/2026	JAILBREAK IN PIECES: COMPOSITIONAL ADVERSARIAL ATTACKS ON MULTI-MODAL LANGUAGE MODELS	https://proceedings.iclr.cc/paper_files/paper/2024/file/83170ccee5543b872f4de71002f21aad-Paper-Conference.pdf	We introduce new jailbreak attacks on vision language models (VLMs), which use aligned LLMs and are resilient to text-only jailbreak attacks. Specifically, we develop cross-modality attacks on alignment where we pair adversarial images going through the vision encoder with textual prompts to break the alignment of the language model. Our attacks employ a novel compositional strategy that combines an image, adversarially targeted towards toxic embeddings, with generic prompts to accomplish the jailbreak. Thus, the LLM draws the context to answer the generic prompt from the adversarial image. The generation of benign-appearing adversarial images leverages a novel embedding-space-based methodology, operating with no access to the LLM model. Instead, the attacks require access only to the vision encoder and utilize one of our four embedding space targeting strategies. By not requiring access to the LLM, the attacks lower the entry barrier for attackers, removing the need to have white-box access to the full end-to-end system. The attacks achieve a high success rate for two different VLMs we evaluated, highlighting the risk of cross-modality alignment vulnerabilities, and the need for new alignment approaches for multi-modal models	OCR+ ICLR -> LLava			
01/31/2026	Seeing the Threat: Vulnerabilities in Vision-Language Models to Adversarial Attack	https://arxiv.org/html/2505.21967v1	Large Vision-Language Models (LVLMs) have shown remarkable capabilities across a wide range of multimodal tasks. However, their integration of visual inputs introduces expanded attack surfaces, thereby exposing them to novel security vulnerabilities. In this work, we conduct a systematic representational analysis to uncover why conventional adversarial attacks can circumvent the safety mechanisms embedded in LVLMs. We further propose a novel two-stage evaluation framework for adversarial attacks on LVLMs. The first stage differentiates among instruction non-compliance, outright refusal, and successful adversarial exploitation. The second stage quantifies the degree to which the model's output fulfills the harmful intent of the adversarial prompt, while categorizing refusal behavior into direct refusals, soft refusals, and partial refusals that remain inadvertently helpful. Finally, we introduce a normative schema that defines idealized model behavior when confronted with harmful prompts, offering a principled target for safety alignment in multimodal systems.					

02/01/2026	Analyzing PDFs like Binaries: Adversarially Robust PDF Malware Analysis via Intermediate Representation and Language Model	https://arxiv.org/html/2506.17162v1	Malicious PDF files have emerged as a persistent threat and become a popular attack vector in web-based attacks. While machine learning-based PDF malware classifiers have shown promise, these classifiers are often susceptible to adversarial attacks, undermining their reliability. To address this issue, recent studies have aimed to enhance the robustness of PDF classifiers. Despite these efforts, the feature engineering underlying these studies remains outdated. Consequently, even with the application of cutting-edge machine learning techniques, these approaches fail to fundamentally resolve the issue of feature instability. To tackle this, we propose a novel approach for PDF feature extraction and PDF malware detection. We introduce the PDFObj IR (PDF Object Intermediate Representation), an assembly-like language framework for PDF objects, from which we extract semantic features using a pretrained language model. Additionally, we construct an Object Reference Graph to capture structural features, drawing inspiration from program analysis. This dual approach enables us to analyze and detect PDF malware based on both semantic and structural features. Experimental results demonstrate that our proposed classifier achieves strong adversarial robustness while maintaining an exceptionally low false positive rate of only 0.07% on baseline dataset compared to state-of-the-art PDF malware classifiers.					
02/01/2026	A Systematic Review of Prompt Injection Attacks on Large Language Models: Trends, Taxonomy, Evaluation, Defenses, and Opportunities	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=11359704	Large Language Models (LLMs) are increasingly integrated into various infrastructure and interactive applications. However, their inherent linguistic flexibility introduces security vulnerabilities, particularly through Prompt Injection (PI) attacks. This systematic review provides a comprehensive synthesis of the evolving landscape of PI attacks, including their development trends, classification schemes, evaluation methodologies, defense strategies, and areas for future research. We analyze the evolution of attacks from basic natural language overrides to complex multi-turn manipulations, indirect injections using structured formats such as JSON and XML, and tool-assisted exploits involving automated prompt generation and multi-modal inputs like images. PI attacks are categorized by the level of manipulation, including character, word, sentence, and semantic levels, and by adversarial intent, such as prompt leaking and harmful content generation. The study analyzes these techniques within different LLM architectures and application contexts, including machine translation and Chain of Thought reasoning. It also reviews common evaluation benchmarks, datasets, and metrics. Mitigation strategies examined include automated red teaming frameworks, input validation, content filtering, and alignment through training. It differentiates itself from previous works by conducting a systematic review specifically focused on immediate injection, organizing its attacks, defenses and research gaps. Finally, this review identifies major research challenges and suggests critical directions for improving LLM safety and robustness					
02/01/2026	Defeating Prompt Injections by Design	https://arxiv.org/pdf/2503.18813	Large Language Models (LLMs) are increasingly deployed in agentic systems that interact with an untrusted environment. However, LLM agents are vulnerable to prompt injection attacks when handling untrusted data. In this paper we propose CaMeL, a robust defense that creates a protective system layer around the LLM, securing it even when underlying models are susceptible to attacks. To operate, CaMeL explicitly extracts the control and data flows from the (trusted) query; therefore, the untrusted data retrieved by the LLM can never impact the program flow. To further improve security, CaMeL uses a notion of a capability to prevent the exfiltration of private data over unauthorized data flows by enforcing security policies when tools are called. We demonstrate effectiveness of CaMeL by solving 77% of tasks with provable security (compared to 84% with an undefended system) in AgentDojo.					
	ODYSSEYBENCH: EVALUATING LLM AGENTS ON LONG-HORIZON COMPLEX OFFICE APPLICATION WORKFLOWS	https://www.semanticscholar.org/reader/2f395b96c8f7bc517a425cff1bcdaef36f5d6739	PDF read task					
	Data extraction for evidence synthesis using a large language model: A proof-of-concept study	https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1710						
	PDF-WuKong : A Large Multimodal Model for Efficient Long PDF Reading with End-to-End Sparse Sampling	https://arxiv.org/pdf/2410.05970	https://huggingface.co/datasets/yh0075/PaperPDF/blob/main/README.md					

	Prompt engineering: The next big skill in rheumatology research	https://onlinelibrary.wiley.com/doi/10.1111/1756-185X.15157						
	PDFTriage: Question Answering over Long, Structured Documents	https://arxiv.org/pdf/2309.08872	Large Language Models (LLMs) have issues with document question answering (QA) in situations where the document is unable to fit in the small context length of an LLM. To overcome this issue, most existing works focus on retrieving the relevant context from the document, representing them as plain text. However, documents such as PDFs, web pages, and presentations are naturally structured with different page is incongruous with the user's mental model of these documents with rich structure. When a system has to query the document for context, this incongruity is brought to the fore, and seemingly trivial questions can trip up the QA system. To bridge this fundamental gap in handling structured documents, we propose an approach called PDFTriage that enables mod either structure or content. Our experiments demonstrate the effectiveness of the proposed PDFTriage-augmented models across several classes of questions where existing retrievalaugmented LLMs fail. To facilitate further research on this fundamental problem, we release our benchmark dataset consisting of 900+ human-generated questions over 80 structured documents from 10 different categories of question types for document QA. Our code and datasets will be released soon on Github.	https://github.com/adobe-research/pdftriage/blob/main/Batch_5090729_batch_results.csv				
	MedDoc-Bot: A Chat Tool for Comparative Analysis of Large Language Models in the Context of the Pediatric Hypertension Guideline	https://arxiv.org/abs/2405.03359	This research focuses on evaluating the non-commercial open-source large language models (LLMs) Meditron, MedAlpaca, Mistral, and Llama-2 for their efficacy in interpreting medical guidelines saved in PDF format. As a specific test scenario, we applied these models to the guidelines for hypertension in children and adolescents provided by the European Society of Cardiology (ESC). Leveraging Streamlit, a Python library, we developed a user-friendly medical document chatbot tool (MedDoc-Bot). This tool enables authorized users to upload PDF files and pose questions, generating interpretive responses from four locally stored LLMs. A pediatric expert provides a benchmark for evaluation by formulating questions and responses extracted from the ESC guidelines. The expert rates the model-generated responses based on their fidelity and relevance. Additionally, we evaluated the METEOR and chrF metric scores to assess the similarity of model responses to reference answers. Our study found that Llama-2 and Mistral performed well in metrics evaluation. However, Llama-2 was slower when dealing with text and tabular data. In our human evaluation, we observed that responses created by Mistral, Meditron, and Llama-2 exhibited reasonable fidelity and relevance. This study provides valuable insights into the strengths and limitations of LLMs for future developments in medical document interpretation					
	OdysseyBench: Evaluating LLM Agents on Long-Horizon Complex Office Application Workflows	https://www.microsoft.com/en-us/research/publication/odysseybench-evaluating-llm-agents-on-long-horizon-complex-office-application-workflows/	Autonomous agents powered by large language models (LLMs) are increasingly deployed in real-world applications requiring complex, long-horizon workflows. However, existing benchmarks predominantly focus on atomic tasks that are self-contained and independent, failing to capture the long-term contextual dependencies and multi-interaction coordination required in realistic scenarios. To address this gap, we introduce OdysseyBench, a comprehensive benchmark for evaluating LLM agents on long-horizon workflows across diverse office applications including Word, Excel, PDF, Email, and Calendar. Our benchmark comprises two complementary splits: OdysseyBench+ with 300 tasks derived from real-world use cases, and OdysseyBench-Neo with 302 newly synthesized complex tasks. Each task requires agent to identify essential information from long-horizon interaction histories and perform multi-step reasoning across various applications. To enable scalable benchmark creation, we propose HomerAgents, a multi-agent framework that automates the generation of long-horizon workflow benchmarks through systematic environment exploration, task generation, and dialogue synthesis. Our extensive evaluation demonstrates that OdysseyBench effectively challenges state-of-the-art LLM agents, providing more accurate assessment of their capabilities in complex, real-world contexts compared to existing atomic task benchmarks. We believe that OdysseyBench will serve as a valuable resource for advancing the development and evaluation of LLM agents in real-world productivity scenarios. In addition, we release OdysseyBench and HomerAgents to foster research along this line.					
	OfficeBench: Benchmarking Language Agents across Multiple Applications for Office Automation	https://arxiv.org/abs/2407.19056	https://github.com/zlwang-cs/OfficeBench/blob/main/tasks/3-69/testbed/data/transcripts.pdf	https://github.com/zlwang-cs/OfficeBench/tree/main/tasks/3-69/testbed/data				

	Document Understanding Dataset and Evaluation (DUDE)	https://arxiv.org/pdf/2305.08455	We call on the Document AI (DocAI) community to reevaluate current methodologies and embrace the challenge of creating more practically-oriented benchmarks. Document Understanding Dataset and Evaluation (DUDE) seeks to remediate the halted research progress in understanding visually-rich documents (VRDs). We present a new dataset1 with novelties related to types of questions, answers, and document layouts based on multi-industry, multi-domain, and multi-page VRDs of various origins, and dates. Moreover, we are pushing the boundaries of current methods by creating multi-task and multi-domain evaluation setups that more accurately simulate real-world situations where powerful generalization and adaptation under low-resource settings are desired. DUDE aims to set a new standard as a more practical, long-standing benchmark for the community, and we hope that it will lead to future extensions and contributions that address real-world challenges. Finally, our work illustrates the importance of finding more efficient ways to model language, images, and layout in DocAI.					
	Agentic Retrieval Grand Challenge (ACM-ICAIF '25)	https://www.kaggle.com/competitions/acm-icaif-25-ai-agentic-retrieval-grand-challenge/data	This competition is designed to evaluate agentic (multi-step) retrieval capabilities in financial AI. The central goal is to identify grounded evidence from large-scale SEC filings in order to answer complex institutional finance questions. To this end, participants are tasked to optimize a system prompt that maximize the ranking performance. Unlike traditional retrieval tasks, this challenge is framed as a two-stage process. In the first stage, Document-Level Ranking, participants must determine which type of SEC filing—such as 10-K, 10-Q, 8-K, DEF 14A, or an Earnings Transcript—is most likely to contain the relevant information for a given query. In the second stage, Chunk-Level Ranking, the task shifts to locating the most relevant passages within the selected document, with participants required to rank the Top-5 chunks that best support the answer.					
	Privacy-Aware Document Visual Question Answering	https://arxiv.org/pdf/2312.10108	Document Visual Question Answering (DocVQA) has quickly grown into a central task of document understanding. But despite the fact that documents contain sensitive or copyrighted information, none of the current DocVQA methods offers strong privacy guarantees. In this work, we explore privacy in the domain of DocVQA for the first time, highlighting privacy issues in state of the art multi-modal LLM models used for DocVQA, and explore possible solutions. Specifically, we focus on invoice processing as a realistic document understanding scenario, and propose a large scale DocVQA dataset comprising invoice documents and associated questions and answers. We employ a federated learning scheme, that reflects the real-life distribution of documents in different businesses, and we explore the use case where the data of the invoice provider is the sensitive information to be protected. We demonstrate that non-private models tend to memorise, a behaviour that can lead to exposing private information. We then evaluate baseline training schemes employing federated learning and differential privacy in this multi-modal scenario, where the sensitive information might be exposed through either or both of the two input modalities: vision (document image) or language (OCR tokens). Finally, we design attacks exploiting the memorisation effect of the model, and demonstrate their effectiveness in probing a representative DocVQA models					
	A FINE-TUNING ENHANCED RAG SYSTEM WITH QUANTIZED INFLUENCE MEASURE AS AI JUDGE	https://www.semanticscholar.org/reader/63b331998b17fdb879073873cf06004a6df18b44	This study presents an innovative enhancement to retrieval-augmented generation (RAG) systems by seamlessly integrating fine-tuned large language models (LLMs) with vector databases. This integration capitalizes on the combined strengths of structured data retrieval and the nuanced comprehension provided by advanced LLMs. Central to our approach are the LoRA and QLoRA methodologies, which stand at the forefront of model refinement through parameter-efficient fine-tuning and memory optimization. A novel feature of our research is the incorporation of user feedback directly into the training process, ensuring the model's continuous adaptation to user expectations and thus, improving its performance and applicability. Additionally, we introduce a Quantized Influence Measure (QIM) as an innovative "AI Judge" mechanism to enhance the precision of result selection, further refining the system's accuracy. Accompanied by an executive diagram and a detailed algorithm for fine-tuning QLoRA, our work provides a comprehensive framework for implementing these advancements within chatbot technologies. This research contributes significant insights into LLM optimization for specific uses and heralds new directions for further development in retrieval-augmented models. Through extensive experimentation and analysis, our findings lay a robust foundation for future advancements in chatbot technology and retrieval systems, marking a significant step forward in the creation of more sophisticated, precise, and user-centric conversational AI systems. We make the dataset, the model, and the app publicly available for the literature	eXTENSIVE PYmupdf use	Utilizes pymupdf to convert into pdf			

	Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Repor	https://www.semanticscholar.org/reader/701f51d40695bc0133b63d423a6a906998de07f5	This paper presents an experience report on the development of Retrieval Augmented Generation (RAG) systems using PDF documents as the primary data source. The RAG architecture combines generative capabilities of Large Language Models (LLMs) with the precision of information retrieval. This approach has the potential to redefine how we interact with and augment both structured and unstructured knowledge in generative models to enhance transparency, accuracy and contextuality of responses. The paper details the end-to-end pipeline, from data collection, preprocessing, to retrieval indexing and response generation, highlighting technical challenges and practical solutions. We aim to offer insights to researchers and practitioners developing similar systems using two distinct approaches: OpenAI's Assistant API with GPT Series and Llama's open-source models. The practical implications of this research lie in enhancing the reliability of generative AI systems in various sectors where domain specific knowledge and real time information retrieval is important. The Python code used in this work is also available a	https://github.com/GPT-Laboratory/RAG-LLM-Development-Guidebook-from-PDFs/blob/main/RAGUsingLlama3.1/Data/2405.01564v1.pdf	3 pdfs			
	olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models	https://www.semanticscholar.org/reader/d750c9a13723df7a9ca0e3408e04a74b1cd38cf3	Abstract PDF documents have the potential to provide trillions of novel, high-quality tokens for training language models. However, these documents come in a diversity of types with differing formats and visual layouts that pose a challenge when attempting to extract and faithfully represent the underlying content for language model use. Traditional open source tools often produce lower quality extractions compared to vision language models (VLMs), but reliance on the best VLMs can be prohibitively costly (e.g., over $6,240 USD per million PDF pages for GPT-4o) or infeasible if the PDFs cannot be sent to proprietary APIs. We present olmOCR, an open-source toolkit for processing PDFs into clean, linearized plain text in natural reading order while preserving structured content like sections, tables, lists, equations, and more. Our toolkit runs a fine-tuned 7B vision language model (VLM) trained on olmOCR-mix-0225, a sample of 260,000 pages from over 100,000 crawled PDFs with diverse properties, including graphics, handwritten text and poor quality scans. olmOCR is optimized for large-scale batch processing, able to scale flexibly to different hardware setups and can convert a million PDF pages for only $176 USD. To aid comparison with existing systems, we also introduce olmOCR-Bench, a curated set of 1,400 PDFs capturing many content types that remain challenging even for the best tools and VLMs, including formulas, tables, tiny fonts, old scans, and more. We find olmOCR outperforms even top VLMs including GPT-4o, Gemini Flash 2 and Qwen-2.5-VL. We openly release all components of olmOCR: our fine-tuned VLM model, training code and data, an efficient inference pipeline that supports vLLM and SGLang backends, and benchmark olmOCR-Bench					
	UDA-Benchmark	https://www.semanticscholar.org/reader/61f8baae9aecadc8bed1ce263c32bd108b476ebd	The use of Retrieval-Augmented Generation (RAG) has improved Large Language Models (LLMs) in collaborating with external data, yet significant challenges exist in real-world scenarios. In areas such as academic literature and finance question answering, data are often found in raw text and tables in HTML or PDF formats, which can be lengthy and highly unstructured. In this paper, we introduce a benchmark suite, namely Unstructured Document Analysis (UDA), that involves 2,965 real-world documents and 29,590 expert-annotated Q&A pairs. We revisit popular LLM- and RAG-based solutions for document analysis and evaluate the design choices and answer qualities across multiple document domains and diverse query types. Our evaluation yields interesting findings and highlights the importance of data parsing and retrieval. We hope our benchmark can shed light and better serve real-world document analysis applications. The benchmark suite	https://github.com/qinchuanhui/UDA-Benchmark/blob/main/dataset/src_doc_files_example/wiki_feta_docs/pdfs/Ben%20Platt%20(actor).pdf				
	Benchmarking PDF Accessibility Evaluation	https://www.semanticscholar.org/reader/d22f64d914dc3ce8c4932ab84cfc83d972828579						
	ATLAS: A System for PDF-centric Human Interaction Data Collection	https://aclanthology.org/2024.naacl-demo.9.pdf						

	Building and better understanding vision-language models: insights and future directions	https://www.semanticscholar.org/reader/5cfec15c744e7b9c110b123d4f2a90091e7c5805	The field of vision-language models (VLMs), which take images and texts as inputshe field and output texts, is rapidly evolving and has yet to reach consensus on several key aspects of the development pipeline, including data, architecture, and training methods. This paper can be seen as a tutorial for building a VLM. We begin by providing a comprehensive overview of the current state-of-the-art approaches, highlighting the strengths and weaknesses of each, addressing the major challenges in the field, and suggesting promising research directions for underexplored areas. We then walk through the practical steps to build Idefics3-8B, a powerful VLM that significantly outperforms its predecessor Idefics2-8B, while being trained efficiently, exclusively on open datasets, and using a straightforward pipeline. These steps include the creation of Docmatix, a dataset for improving document understanding capabilities, which is 240 times larger than previously available datasets. We release the model along with the datasets created for its training	https://huggingface.co/datasets/HuggingFaceM4/Docmatix				
	RAG VS FINE-TUNING: PIPELINES, TRADEOFFS, AND A CASE STUDY ON AGRICULTURE	https://www.semanticscholar.org/reader/fef0393e997ec51b184e39c712be63197d99fd46	There are two common ways in which developers are incorporating proprietary and domain-specific data when building applications of Large Language Models (LLMs): Retrieval-Augmented Genera- tion (RAG) and Fine-Tuning. RAG augments the prompt with the external data, while fine-Tuning incorporates the additional knowledge into the model itself. However, the pros and cons of both approaches are not well understood. In this paper, we propose a pipeline for fine-tuning and RAG, and present the tradeoffs of both for multiple popular LLMs, including Llama2-13B, GPT-3.5, and GPT-4. Our pipeline consists of multiple stages, including extracting information from PDFs, generating questions and answers, using them for fine-tuning, and leveraging GPT-4 for evaluating the results. We propose metrics to assess the performance of different stages of the RAG and fine-Tuning pipeline. We conduct an in-depth study on an agricultural dataset. Agriculture as an industry has not seen much penetration of AI, and we study a potentially disruptive application - what if we could provide location-specific insights to a farmer? Our results show the effectiveness of our dataset generation pipeline in capturing geographic-specific knowledge, and the quantitative and qualitative benefits of RAG and fine-tuning. We see an accuracy increase of over 6 p.p. when fine-tuning the model and this is cumulative with RAG, which increases accuracy by 5 p.p. further. In one particular experiment, we also demonstrate that the fine-tuned model leverages information from across geographies to answer specific questions, increasing answer similarity from 47% to 72%. Overall, the results point to how systems built using LLMs can be adapted to respond and incorporate knowledge across a dimension that is critical for a specific industry, paving the way for further applications of LLMs in other industrial domains					