LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models

AIDA KOSTIKOVA, University of Bielefeld, Germany
ZHIPIN WANG, University of Technology Nuremberg, Germany
DEIDAMEA BAJRI, University of Mannheim, Germany
OLE PÜTZ, University of Bielefeld, Germany
BENJAMIN PAASSEN, University of Bielefeld, Germany
STEFFEN EGER, University of Technology Nuremberg, Germany

Large language model (LLM) research has grown rapidly, along with increasing concern about their limitations such as failures in reasoning, hallucinations, and limited multilingual capability. While prior reviews have addressed these issues, they often focus on individual limitations or consider them within the broader context of evaluating overall model performance. This survey addresses the gap by presenting a data-driven, semi-automated review of research on limitations of LLMs (LLLMs) from 2022 to 2025, using a bottom-up approach. From a corpus of 250,000 ACL and arXiv papers, we extract 14,648 relevant limitation papers using keyword filtering and LLM-based classification, validated against expert labels. Using topic clustering (via two approaches, HDBSCAN+BERTopic and LlooM), we identify between 7 and 15 prominent types of limitations discussed in recent LLM research across the ACL and arXiv datasets. We find that LLM-related research increases nearly sixfold in ACL and nearly fifteenfold in arXiv between 2022 and 2025, while LLLMs research grows even faster, by a factor of over 12 in ACL and nearly 28 in arXiv. Reasoning remains the most studied limitation, followed by generalization, hallucination, bias, and security. The distribution of topics in the ACL dataset stays relatively stable over time, while arXiv shifts toward safety and controllability (with topics like security risks, alignment, hallucinations, knowledge editing), and multimodality between 2022 and 2025. We offer a quantitative view of trends in LLM limitations research and release a dataset of annotated abstracts and a validated methodology, available at: github.com/a-kostikova/LLLMs-Survey.

CCS Concepts: • **Information systems → Clustering and classification**; • **Computing methodologies** → *Artificial intelligence*; *Natural language processing*; *Natural language generation*; *Information extraction.*

Additional Key Words and Phrases: Large Language Models, LLM Limitations, LLM Trend Analysis

1 Introduction

With the explosive growth of large language model (LLM) research and deployment [111], questions of the limitations of LLMs (LLLMs) have also gathered increased interest, ranging from reasoning failures [81], social bias [59], hallucinations [117], difficulty in handling long contexts [53], and many more. Understanding where LLMs fail is essential for knowing how and whether they can be safely and effectively used in real-world settings, especially as LLMs are increasingly deployed in safety-sensitive domains such as healthcare, education, finance, and law [14]. Moreover, tracking how these failure modes evolve over time helps reveal whether the fast-paced research landscape is addressing them, overlooking them, or exposing new ones, offering a clearer picture of where further research is most needed.

However, given the sheer size of LLM research, with thousands of published research papers every year (even when limited to highly rated outlets), it is challenging to maintain an up-to-date overview of LLLMs research using traditional, manual literature review techniques. Accordingly, prior reviews on LLLMs mostly focus on specific limitations, such as reasoning [57, 116], or examine

Authors' Contact Information: Aida Kostikova, University of Bielefeld, Bielefeld, Germany, aida.kostikova@uni-bielefeld.de; Zhipin Wang, University of Technology Nuremberg, Nuremberg, Germany; Deidamea Bajri, University of Mannheim, Mannheim, Germany; Ole Pütz, University of Bielefeld, Bielefeld, Germany; Benjamin Paaßen, University of Bielefeld, Bielefeld, Germany; Steffen Eger, University of Technology Nuremberg, Nuremberg, Germany.

limitations within the broader context of evaluating overall model capabilities [9, 96]. To date, the field still misses an overview that covers the more recent LLM research between 2022 and now and cuts across limitations. Our review is an attempt to provide this high-level overview.

To make our task feasible, we opt for a data-driven, bottom-up approach and build a partially automated, systematic literature review pipeline. Starting from an initial set of almost 250,000 crawled papers from ACL (2022-2024) and arXiv (2022 through early 2025), we extract 14,648 papers that discuss `LLLMs` (filtering for keywords first, then classifying the papers' abstracts with an LLM, validated against human expert classifications). Finally, we cluster the papers using two different methods (HDBSCAN+BERTopic and LlooM) to understand which particular limitations are researched. These approaches offer complementary strengths: the former provides single-label, density-based clustering, while the latter uses multi-label, LLM-based assignments, allowing us to cross-validate and reduce method-specific bias. Overall, our methods serve to apply quantitative methods to surveying this vast field.

We observe four main results. i) `LLLMs` research has grown rapidly, outpacing even the growth of LLM research overall. The number of LLM-related papers has grown by a factor of nearly 6 in ACL and nearly 15 in arXiv between 2022-2025, reaching almost 80% of all crawled ACL papers and roughly 30% of all crawled arXiv papers; `LLLMs` papers have increased even more sharply, by a factor of over 12 in ACL and 28 in arXiv, accounting for more than 30% of LLM papers in Q1 of 2025. ii) Within `LLLMs` research, *reasoning* limitations are the most prominent, with *generalization*, *hallucination*, *bias*, and *security* as further important concerns. iii) The distribution of limitations appears relatively stable in the ACL dataset, whereas the arXiv dataset shows a rise in concern for topics related to safety and controllability (e.g., *Security Risks*, *Alignment Limitations*, *Knowledge Editing*, *Hallucination*) as well as *Multimodality*. (iv) Despite substantial methodological differences between HDBSCAN and LlooM, we observe topical overlap in several of the biggest clusters (e.g., *Reasoning*, *Hallucination*, *Security Risks*) across both approaches, with broadly similar trend patterns, suggesting that these findings are reliable.

The contributions of this review to the field are i) a large-scale dataset of paper abstracts, tagged with limitation information, for further research,[1] ii) an LLM-based paper annotation methodology, validated against human experts, iii) most importantly, quantitative insights into the evolution of `LLLMs` research covering the entire period 2022-2024 and early 2025, providing the first comprehensive overview of `LLLMs` research for this period.

2 Related Work

2.1 Surveys of Large Language Models

A growing number of surveys have aimed to synthesize the rapid progress of LLMs, covering their architectures, training paradigms, applications, and broader impact. Notably, Zhao et al. [125] have become a widely adopted reference in the field, offering a structured overview along four key dimensions: pre-training, adaptation, utilization, and evaluation. Other comprehensive works expand on this foundation by discussing emerging areas such as multimodal LLMs, robotics, and system efficiency [33, 77], as well as reasoning and planning capabilities in large-scale models [69].

In parallel to cross-domain surveys, a number of studies have investigated how LLMs are being adopted and evaluated in specific fields. In the medical domain, surveys examine the effectiveness of LLMs in clinical summarization and diagnostic reasoning [46, 104], as well as challenges related to hallucination and factual consistency in medical question answering [83]. Other works explore the capabilities and limitations of LLMs in capturing cultural commonsense knowledge [93] and scientific research processes [22, 64]. In recommendation systems, LLMs have been studied as both

[1] github.com/a-kostikova/LLLMs-Survey

retrieval and generation engines [114], while in information retrieval, surveys highlight their use in query expansion, passage ranking, and answer synthesis [127]. LLMs have also been applied to software engineering tasks such as code generation and bug fixing, with systematic reviews discussing both their potential and practical limitations [36]. Beyond text-based applications, LLMs have been integrated with structured resources such as knowledge graphs [44, 80] and studied in the context of autonomous agent design [106].

While these surveys offer valuable perspectives on LLM usage and challenges within specific domains, the literature remains fragmented with respect to how limitations are identified, categorized, and compared. Our work is motivated by the need for a more systematic and scalable approach to mapping research focused explicitly on the limitations of LLMs across domains and tasks.

2.2 Surveys on Limitations of LLMs

As LLMs are increasingly deployed in real-world applications, a growing body of research has emerged to examine their limitations from different capability-oriented perspectives. One prominent area of concern are hallucinations, where recent surveys investigate underlying causes and mitigation strategies in both text generation [38, 39, 102] and multimodal contexts [62, 90, 91]. Another major focus is reasoning, with surveys analyzing the development of novel techniques[37, 84] such as chain-of-thought prompting [13], reinforced reasoning [57, 116], and mathematical problem-solving [3]. In parallel, the trustworthiness and reliability of LLMs have been studied through the lens of fairness, transparency, and calibration [98], while other works concentrate on security and privacy threats including adversarial vulnerabilities and data leakage [17, 119]. A related thread explores ethical risks in LLM-based agents, particularly concerning safety, misuse, and human interaction [26, 43, 47].

Despite these valuable contributions, existing reviews typically focus on specific capabilities or domains in isolation, often adopting distinct definitions, evaluation metrics, and analytical frameworks. As a result, the broader landscape of LLM limitations remains fragmented, making it difficult to compare findings or track emerging research trends. This underscores the need for a more systematic and scalable approach to identifying and organizing literature across different limitations of LLMs.

2.3 LLMs as Analytical Tools for Scientific Literature

We apply a partially automated pipeline, relying on LLMs to filter the papers included in our survey and providing embeddings for clustering. Such methods have to be applied with care to avoid being misled because of the very limitations this survey is supposed to study. In developing our methodology, we rely on a growing literature of LLMs being used as instruments for analyzing scientific literature [22]. Several recent approaches employ LLMs for topic modeling, semantic clustering, and concept induction, enabling more interpretable organization of large-scale corpora [20, 24, 48, 82, 124]. This has led to a growing number of systematic reviews that use BERTopic and related methods to map research landscapes across areas such as generative AI, information assurance, LLM applications, and research impact evaluation [6, 21, 27, 32, 50].

In addition to these analytical techniques, other efforts focus on automating synthesis of results across papers. Systems such as SurveyX and AutoSurvey generate draft surveys from large paper collections [58, 108], while tools like LitLLM [2] and PaSa [35] support LLM-based retrieval, summarization, and exploration of academic texts.

To enable a data-driven, bottom-up analysis of the vast field of `LLLMs` research, including several ten thousand papers, we opt to apply some of the aforementioned automation techniques. However, given the limitations of LLMs, we aim to validate each step of our method, either by comparing to a human gold standard, or by comparing the outputs of different methods (hence, we use two

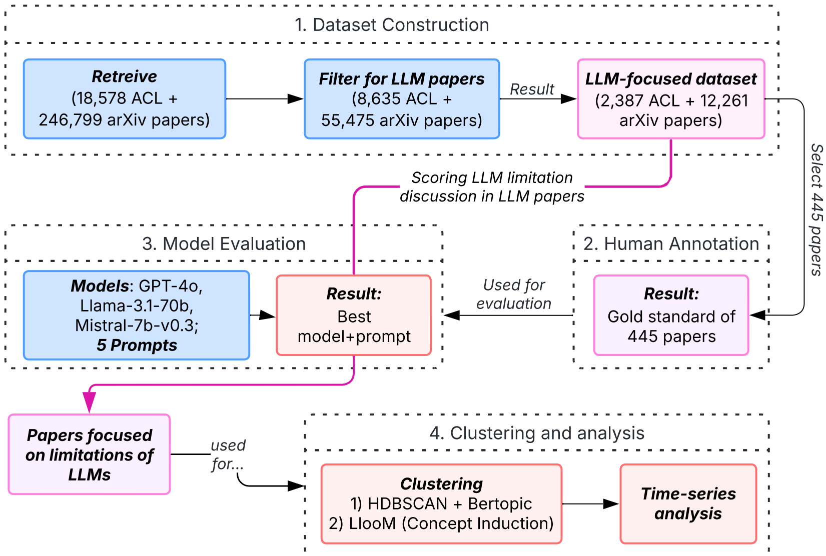


Fig. 1. Overview of the pipeline for our systematic literature review.

clustering approaches). As such, we aim to be more conservative in our utilization of LLMs in literature research compared to the prior work pointed out above.

3 Methodology

Fig. 1 illustrates the method for our systematic literature review. We begin by retrieving papers from arXiv and ACL (Section 3.1), filter according to keywords (Section 3.2), filter papers further by classifying their abstracts with an LLM (Section 3.3), and finally cluster the papers (Section 3.5). At each step of the analysis, we perform validations to ensure the robustness of our results: the keyword list is obtained with an iterative refinement procedure, the LLM classification step is validated against a gold standard of 445 human-annotated papers, and we use two distinct clustering methods for comparison (HDBSCAN+BERTopic and LlooM). In the remainder of this section, we describe each step in more detail.

3.1 Data Retrieval

Our initial dataset includes all academic papers published between January 2022 and March 2025, sourced from the ACL Anthology and arXiv. ArXiv captures (potentially) non–peer-reviewed research that closely tracks current developments, while ACL venues reflect peer-reviewed work and remain the primary publication outlets for NLP research, where much of the foundational work on LLMs originated. The time frame was chosen to capture the year preceding the release of ChatGPT as well as all subsequent research on LLMs [125].

For ACL Anthology, we scrape conference pages for AACL 2022–2023, ACL 2022–2024, EACL 2023–2024, EMNLP 2022–2024, ICLR 2022–2024, NAACL 2022 and 2024, and TACL 2022–2024 as the premier NLP venues.[2] For arXiv, we retrieve papers from the categories of Computation & Language (cs.CL), Machine Learning (cs.LG), Artificial Intelligence (cs.AI), and Computer Vision

[2]We exclude other tracks and venues such as workshops, system demonstrations, tutorials, shared tasks, and task-specific venues (e.g., SemEval, CoNLL, WMT) to maintain a focus on high-impact research from general-purpose NLP conferences.

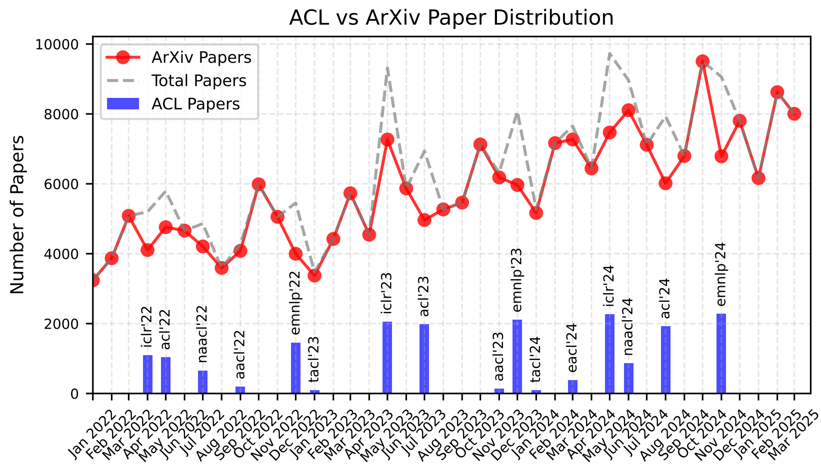


Fig. 2. Distribution of papers over time in the crawled dataset, showing ACL papers, arXiv papers, and the total count (ACL + arXiv).

(cs.CV) because these communities are closest to LLM research. However, we note that many papers are classified by the authors into multiple arXiv categories, so that research areas beyond these initial ones are covered as well (see Section 4.4.3 for more details). Each entry includes metadata such as title, publication date, author information, download link, and abstract, with arXiv papers also containing all assigned categories. We use titles and abstracts for keyword filtering and clustering, as they capture a paper's main claims and contributions and are well-suited for large-scale automated analysis. The final crawled dataset includes 245,835 papers (18,578 papers for ACL, 227,257 for arXiv). Fig. 2 shows raw numbers of crawled ACL and arXiv papers over time.

3.2 Keyword-Based Filtering

In a first filtering stage, we exclude papers if no LLM-related keyword occurs in their title or abstract. This step serves to avoid excessive resource needs in the later, more fine-grained filtering.

To identify keywords related to LLLMs research, we apply the following iterative approach:

(1) We use TNT-KID [68] to generate initial keywords for each abstract.
(2) Two manually reviewed sets of 50 LLM and 50 non-LLM papers are selected based on predefined seed terms (e.g., *LLM*, *large language model*).
(3) We compute log-likelihood ratio (LLR) scores for TNT-KID-generated keywords in both sets. Keywords with LLR $\geq$ 25 are added to the list.
(4) Using the updated list, we expand the dataset by adding 100 more papers to each set, maintaining balance across venues and years while avoiding duplicates.
(5) Steps 3 and 4 repeat until all papers are processed. As more keywords are added, the rate of new informative terms naturally decreases. To avoid including increasingly marginal or noisy keywords in later iterations, we raise the LLR threshold by 5% whenever fewer than 5% of the keywords in the current round are new. This keeps the keyword list focused on strongly distinctive terms as the process converges.

This results in a list of 90 keywords (19 unigrams, 44 bigrams, 16 trigrams, and 11 four-grams). Overall, the final keyword set covers key aspects of LLM research (see the full list in Section A supplementary material):

Table 1. Crawled vs. LLM-filtered paper counts across sources (2022–2025)

Source / Year	**2022**	**2023**	**2024**	**2025**
ACL	1,032 / 294	1,977 / 821	1,916 / 1,483	– / –
EACL	– / –	478 / 188	382 / 204	– / –
AACL	192 / 59	134 / 53	– / –	– / –
TACL	84 / 27	98 / 32	95 / 60	– / –
EMNLP	1,372 / 520	2,107 / 1,177	2,273 / 1,844	– / –
ICLR	1,094 / 143	1,573 / 251	2,260 / 674	– / –
NAACL	652 / 217	– / –	859 / 588	– / –
ArXiv	52,642 / 5,726	66,179 / 13,361	85,645 / 27,700	22,791 / 8,688
Yearly Total	**57,072 / 6,986**	**72,546 / 15,883**	**94,390 / 32,553**	**22,791 / 8,688**
Total (All Years)				**246,799 / 64,110**

Note: Each cell is `Crawled / LLM-filtered`. *ACL venues do not include data for 2025 (as of 31.03.2025).*

- terms related to LLMs, including *multimodal LLMs*, *small LMs* and *pre-trained LMs*;
- training methods: *LoRA*, *PEFT*, *instruction tuning*;
- capabilities (e.g. *mathematical*, *temporal*, and *commonsense reasoning*);
- limitations and risks, such as security vulnerabilities (*jailbreaking*, *prompt injection*, *data contamination*);
- model evaluation: *self-evaluation*, *benchmarking*;
- methods and techniques, such as reasoning paradigms (*chain of thought*, *self-reflection*, *tree of thoughts*), prompting strategies (*prompt optimization*, *prompt engineering*), augmentation methods (*retrieval-augmented generation*, *tool learning*).

We keep only those papers that contain at least one of the keywords in their title or abstract. This results in 64,110 papers (8,635 for ACL, 55,475 for arXiv). A breakdown of crawled and filtered papers across sources for each year is provided in Table 1. While this filtering step does not strictly isolate LLM-focused papers, we observe that the proportion of papers passing the filter increases over time, which may reflect a growing focus on LLMs in the broader NLP research landscape. We examine this pattern in more detail in Section 4.2, where we analyze trends in both LLM and `LLLMs` papers after additional filtering.

3.3 LLM-Based Filtering

In a second filtering stage, we apply an LLM to evaluate every abstract of the 64,110 papers left after the first filtering stage and 1) rate how much `LLLMs` are discussed on a scale from 0 to 5, as well as 2) extract text snippets that explicitly discuss limitations for papers rated 2 or higher. The text snippets will later form the basis for clustering.

However, before we apply this filtering, we set up a human-annotated gold standard dataset to check if LLMs are able to perform this filtering in the first place.

3.3.1 Human Annotation Task. For human annotation, we randomly select 445 papers from the keyword-filtered dataset, balancing the source (ACL or arXiv, ensuring conference representation within ACL) and publication year. Papers are manually annotated based on their titles and abstracts to assess whether they discuss `LLLMs`. The human annotators rated each paper on a scale from 0-5, reaching from no relation to LLMs (0) to exclusive focus on `LLLMs` (5). Refer to Table 2 for the detailed annotation guideline.

Table 2. Annotation scheme for LLM limitation discussion, including label descriptions and distribution of papers in the human annotated dataset.

Label	Description	Count
0	No mention of LLMs.	62
1	Mentions LLMs but not their limitations.	106
2	Briefly mentions a limitation, e.g., as justification for a new method.	169
3	Discusses one or two limitations in moderate detail but not as the primary focus.	62
4	Extensively discusses multiple limitations, making them a major focus.	37
5	Entirely focused on LLM limitations and challenges.	9

Table 3. Venue-year distribution of papers in the human-annotated dataset

Year	arxiv	acl	aacl	eacl	emnlp	iclr	naacl	tacl	Total
2022	17	15	4	0	21	2	14	0	73
2023	69	22	0	4	39	0	0	82	216
2024	62	19	0	6	25	4	13	27	156
Total	148	56	4	10	85	6	27	109	**445**

For papers rated 2-5, annotators highlighted textual evidence pointing to the limitation and its domain, hereafter referred to as "evidence". See Table 13 in the supplementary material for representative examples of annotated papers and highlighted evidence.

Limitation rating agreement. We measure annotator agreement using:

(1) standard Cohen's Kappa for raw agreement;
(2) quadratic weighted Cohen's Kappa, which accounts for the ordinal nature of the 0–5 scale by penalizing larger discrepancies more heavily.

The annotation process included multiple rounds, involving two professors (natural language processing and machine learning), one PhD student (NLP), and one Master's student (computer science). Initial annotations showed moderate inter-annotator agreement (0.27 standard Cohen's Kappa, and 0.62 weighted), which was improved by further rounds (0.57, and 0.75, respectively), indicating substantial agreement in the final version. Overall, the annotation process covered 445 samples (where 195 were annotated solely by a Master's student after all discussion rounds).

The final rating for each paper is determined by rounding the average of annotators' ratings. The number of papers assigned to each label in the human-annotated dataset is shown in Table 2. Table 3 displays the statistics of labels across years.

Evidence annotation agreement. We compare the evidence highlighted based on a representation as BIO-tagged sequences, where tokens are labeled as B-EVID, I-EVID, or O (beginning, inside or outside evidence, respectively). In our agreement analysis, we include any paper that is annotated by at least two annotators, which was the case for 250 of the 455 annotated papers. We compute agreement by means of the macro-averaged F1 score across annotator pairs, excluding cases where neither annotator selected evidence.

On average, evidence agreement across the full jointly annotated dataset (ratings 0–5, 250 papers) is 0.55 in terms of averaged pairwise F1. For papers explicitly discussing limitations (papers with ratings 3–5 as a final label, 48 jointly annotated papers), F1 score increases to 0.71. This score suggests reliable consistency, given the known difficulty of span-level annotation [19].

3.4 Models and Prompting Evaluation

We evaluate models on the human-annotated dataset to identify the most effective model-prompt configuration for scoring LLM limitation discussions and extracting supporting evidence. Considering both performance and cost, we select the best-performing one for full-dataset classification and clustering of papers by limitation topics described in Section 4.

To represent different families and sizes of models, we evaluated three models and selected the best performing one for full-scale annotation: `GPT-4o` selected as one of the best-performing models at the time of analysis [15], `Mistral-7B-Instruct-v0.3` [40] as a small-scale open-weight model and `Llama-3.1-70b-Instruct` [28] as a large-scale open-weight model.[3] We also compare the results against a Logistic Regression baseline with SBERT embeddings [86], using random sub-sampling validation with three 80/20 train–test splits, and applying SMOTE oversampling [11] to mitigate class imbalance, with results averaged across splits.

To account for the impact of prompting on model performance, we experiment with different strategies to determine the most effective approach:

- **Prompt 1**: zero-shot baseline (no explicit rating rules).
- **Prompt 2**: zero-shot with defined rating criteria.
- **Prompt 3**: few-shot with defined rating criteria and five examples for each rating which also include explanation for each rating.

Similar to human annotators highlighting evidence in abstracts to indicate discussions of `LLLMs`, models are prompted to extract supporting evidence from the text. Each prompt instructs the model: *"Please respond in the following format, providing a rating and supporting evidence for the discussion of LLM limitations in each abstract. Do not include explanations, only cite the evidence found in the abstract."* Prompt 3, included as the most comprehensive, is available in Figure 14 of the supplementary material.

Metrics and Evaluation. We compare model ratings to human ratings using weighted Cohen's Kappa for rating prediction. To evaluate evidence extraction, we compare the BIO sequence of human annotators to the BIO sequence of models using averaged pairwise F1. Only papers rated 3–5 are included in the evidence evaluation.

3.5 Clustering

To identify patterns in `LLLMs` discussions, we apply two clustering approaches: (i) HDBSCAN [70] + BERTopic [29] and (ii) LLooM concept induction [48], and compare their results. We select these algorithms in particular because they represent particularly distinct approaches to text clustering: HDBSCAN + BERTopic assigns each paper to at most one cluster and does so based on a hierarchical clustering in the embedding space. By contrast, LLooM derives topics first and then queries an LLM for each paper-topic-combination whether the respective paper belongs to that respective topic, thus permitting papers to belong to multiple clusters. We remain agnostic regarding the choice of clustering algorithm and focus on findings that are consistent across both approaches to enhance the robustness of our literature review. Below, we outline the data preparation process and describe each clustering pipeline in detail.

Data preparation. As clustering material, we use not the full abstract but only the passages that explicitly describe the LLM limitation the paper is concerned with, i.e. the evidence statements of papers rated 3-5 as extracted in Section 3.3. To enrich the text representation for clustering, we follow the approach of Viswanathan et al. [105] and generate keyphrases for each statement using

[3]For `Mistral-7B-Instruct-v0.3` and `Llama-3.1-70b-Instruct`, we set the temperature to 0.6 and top_k = 0.9; `Llama-3.1-70b-Instruct` is run with 4-bit quantization. GPT-4o is used in version `gpt-4o-2024-08-06`.

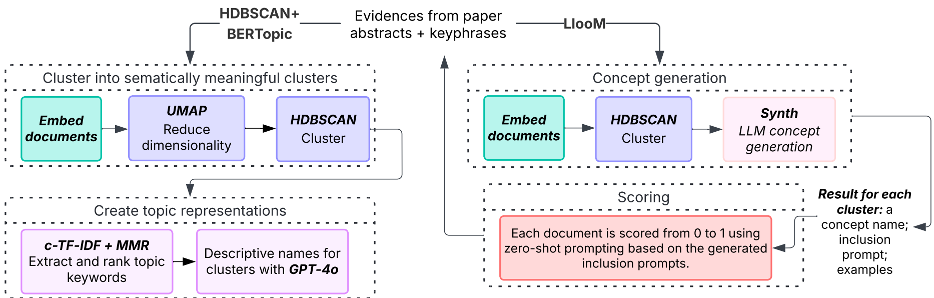


Fig. 3. Comparison of clustering steps in HDBSCAN+BERTopic and LLooM. Both methods take on evidence excerpts with appended keyphrases as an input. For LLooM, we omit the *Distill* step, which is typically used to summarize full documents, as our input already consists of concise excerpts.

GPT-4o. The model is prompted to "provide a comprehensive set of keyphrases describing the LLM limitations discussed in a paper", with no constraints on the number generated. For example:

Evidence: "We find that zero-shot CoT reasoning in sensitive domains significantly increases a model's likelihood to produce harmful or undesirable output [...]"[92]

Generated Keyphrases: "zero-shot CoT reasoning limitations", "increased harmful output", "sensitive domains challenges", "prompt format issues"

Each set of keyphrases is appended to the original evidence statement, and this combined text serves as input to both clustering approaches. Figure 3 illustrates the respective pipelines.

Clustering pipeline 1: HDBSCAN + BERTopic. We employ a density-based clustering approach using HDBSCAN+BERTopic. We follow the standard BERTopic pipeline [29]: we use OpenAI's `text-embedding-3-large` model to embed the combined evidence–keyphrase text, and reduce the embedding using UMAP [71], retaining 10 dimensions for ACL and 5 for arXiv. Such low dimensionality has been shown to be effective for HDBSCAN in prior work [89].

To ensure meaningful clusters and minimize spurious outliers, we tune UMAP, HDBSCAN and BERTopic parameters separately for ACL and arXiv. We further apply a distance-based outlier reassignment strategy after observing that many outliers were not true noise but semantically close to existing clusters, suggesting misclassification by HDBSCAN. Full parameter settings and details of the reassignment procedure are provided in Section B of the supplementary materials.

Finally, clusters are given descriptive names by GPT-4o, based on the top-ranked keywords extracted by BERTopic for each cluster.

Clustering pipeline 2: LlooM. For the second clustering approach, we adapt the LLooM concept induction method with modifications tailored to our use case. LlooM involves a process of summarization, clustering, and LLM-based synthesis. It first summarizes documents into bullet points (distill step), then clusters them using HDBSCAN. An LLM then generates a concept (a short, human-readable label that describes the theme of the cluster) and inclusion prompt for each cluster (synthesize step), which are used to score all documents for each concept via zero-shot prompting on a 0–1 scale (score step). For further implementation details, we refer the reader to the original LlooM paper. [48]

In our setup, we skip the distill step (summarization of input text into short bullet points), since our dataset already consists of concise quotes from research papers. We generate two concepts per

Table 4. Weighted Cohen's Kappa for limitation ratings and pairwise F1 scores for evidence extraction across models and prompts. The best weighted Kappa for each model is underlined, while the best score overall is in **bold**. Evidence extraction is measured in pairwise F1 between each annotator and the model, reported for the best-performing prompts.

Model	Prompt	Weighted Kappa	Evidence F1
`Mistral-7B-Instruct-v0.3`	Prompt 1	0.25	
	Prompt 2	0.60	
	Prompt 3	0.60	0.36
`Llama-3.1-70b-Instruct`	Prompt 1	0.60	
	Prompt 2	0.73	
	Prompt 3	**0.74**	0.65
GPT-4.o	Prompt 1	0.49	
	Prompt 2	0.68	
	Prompt 3	0.72	0.64
SBERT + log. regression	—	0.43	—

cluster, conducting two rounds of review to refine the concepts. The clustering step is performed using `text-embedding-3-large`, while `GPT-4o` is used for concept synthesis and iterative review. For final scoring, we employ `Llama-3.1-70b-Instruct`, and retain only those papers for which the model assigns a 75% to 100% confidence score for a given concept.

4 Results

In this section, we report the results of the LLM-based filtering stage and the clustering stage of our pipeline (Fig. 1). We begin, however, with the validation results of our LLM-based filtering stage, comparing LLM classifications of abstracts to human annotations.

4.1 LLM-based filtering evaluation

Table 4 summarizes how well different models with different prompts align with human expert annotations. We report quadratic weighted Cohen's Kappa for limitation ratings and pairwise F1 for evidence extraction, measured between each annotator and the model for the best-performing prompts.

As seen in the table, performance improves as prompts become more detailed across all models, with Prompt 3 consistently resulting in the highest Kappa scores but Prompt 2 performing similarly well across models, suggesting that models benefit from clear definitions but less from examples.

Overall, `Llama-3.1-70b-Instruct` shows the strongest agreement with human annotations, achieving the highest weighted Kappa (0.74) for rating assignment, as well as the highest evidence extraction F1 (0.65). For reference, human–human agreement reaches 0.75 for ratings and 0.71 for evidence extraction. `GPT-4o` follows closely, with a weighted Kappa of 0.72 and an evidence extraction F1 of 0.64. While `Mistral-7B` outperforms the baseline (weighted Kappa: 0.43), it lags behind Llama and `GPT-4o`, with a weighted Kappa of 0.60 and a much lower evidence extraction F1 of 0.36. Taken together, these results indicate that `Llama-3.1-70b` and `GPT-4o` can serve as reasonably reliable annotators in our setting, with agreement levels approaching those of human annotators.

Error analysis. To better understand Llama's performance in *limitation rating* beyond Kappa scores, we examine its confusion matrix. Figure 4 shows that both humans and the model often

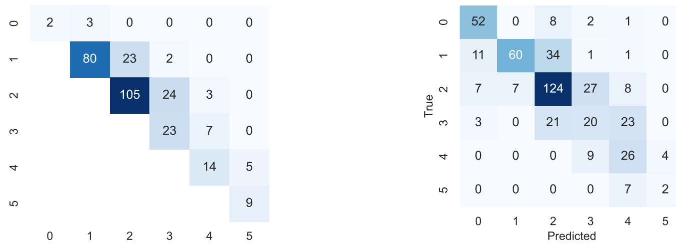
Table 5. Comparison of evidence extraction between human annotators and Llama-3.1-70b using Prompt 3. Text highlighted in green indicates parts which both the model and human annotators selected as evidence. Yellow highlights denote evidence selected only by human annotators, while blue highlights indicate evidence selected solely by the model.

Title	Abstract
(1) *"Unlocking Adversarial Suffix Optimization Without Affirmative Phrases: Efficient Black-box Jailbreaking via LLM as Optimizer"* [42], arXiv, August 2024	"Despite prior safety alignment efforts, mainstream LLMs can still generate harmful and unethical content when subjected to jailbreaking attacks. [...] In this paper, we present ECLIPSE, a novel and efficient black-box jailbreaking method utilizing optimizable suffixes. [...] Experimental results demonstrate that ECLIPSE achieves an average attack success rate (ASR) of 0.92 across three open-source LLMs and GPT-3.5-Turbo." **True label: 3, Predicted: 4**
(2) *"Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis"* [113], arXiv, October 2023	"[...] Our observation shows that, while GPT-4V demonstrates proficiency in distinguishing between medical image modalities and anatomy, it faces significant challenges in disease diagnosis and generating comprehensive reports. These findings underscore that while large multimodal models have made significant advancements in computer vision and natural language processing, it remains far from being used to effectively support real-world medical applications and clinical decision-making. [...]" **True label: 4, Predicted: 3**
(3) *"Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks"* [51], arXiv, June 2023	"We consider the questions of whether or not large language models (LLMs) have beliefs, and, if they do, how we might measure them. [...] We provide empirical results that show that these methods fail to generalize in very basic ways. We then argue that, even if LLMs have beliefs, these methods are unlikely to be successful for conceptual reasons. Thus, there is still no lie-detector for LLMs. [...]" **True label: 4, Predicted: 4**

confuse adjacent categories, such as 2 ↔ 3, 3 ↔ 4, which is expected given the ordinal nature of the labels. Overall, the model tends to overestimate rather than underestimate discussions of limitations. In some cases, it misses LLM mentions entirely, predicting label 0 where LLMs are discussed (21 cases). Despite some misclassifications, the model rarely confuses clearly high-rated papers (4–5) with clearly low-rated ones (0–2). This suggests that it can reliably separate papers that meaningfully discuss `LLLMs` from those that do not.

Concerning *evidence extraction*, in most cases, the model correctly identifies limitations when they are clearly stated and often fully matches human annotations exactly (Example 1 in Table 5). Disagreements primarily stem from the model's tendency to select only 1–2 key sentences, omitting longer arguments that annotators capture (Example 2), or when it chooses full statements instead of specific phrases (Example 3), occasionally even capturing content that humans overlook (for example, the sentence "...these methods fail to generalize in very basic ways" was missed by a human but selected by the model.)

As `Llama-3.1-70b-Instruct` performed best in our evaluation, we select it for the subsequent analysis. In the final classification step, `Llama-3.1-70b-Instruct` assigns each LLM-focused paper



(a) Annotators' agreement

(b) Llama-3.1-70b vs. Final Label

Fig. 4. Confusion matrices comparing human agreement (Figure 4a) and the predictions of Llama-3.1-70b against the final labels (Figure 4b). The human agreement matrix is aggregated over all pairwise annotator comparisons, making it symmetric.

Table 6. Distribution of ratings for ACL and arXiv papers. The *ACL Count (%)* and *arXiv Count (%)* columns show the number of papers with each rating and their percentage within the ACL and arXiv datasets, respectively. The *Total Count (%)* column combines both datasets. The last row sums papers with ratings 3–5, referred to as *Limitation Papers*.

Rating	ACL Count (%)	arXiv Count (%)	Total Count (%)
0	861 (10.0%)	11,416 (20.6%)	12,277 (19.2%)
1	1,463 (17.0%)	10,967 (19.8%)	12,430 (19.4%)
2	3,911 (45.4%)	20,810 (37.5%)	24,721 (38.6%)
3	1,274 (14.8%)	6,057 (10.9%)	7,331 (11.4%)
4	1,035 (12.0%)	5,723 (10.3%)	6,758 (10.5%)
5	78 (0.9%)	481 (0.9%)	559 (0.9%)
Limitation Papers (3–5)	**2,387 (27.7%)**	**12,261 (22.1%)**	**14,648 (22.9%)**

a rating from 0 to 5, with higher scores (3–5) indicating a deeper discussion of limitations. Table 6 summarizes the results of the large-scale classification by `Llama-3.1-70b-Instruct` across all ACL and arXiv papers. Most received a score of 2 or lower, with 2,338 ACL papers (27.4%) and 8,782 arXiv papers (20.9%) classified as discussing limitations in depth (ratings 3–5). These high-rated papers serve as input for the clustering analysis.

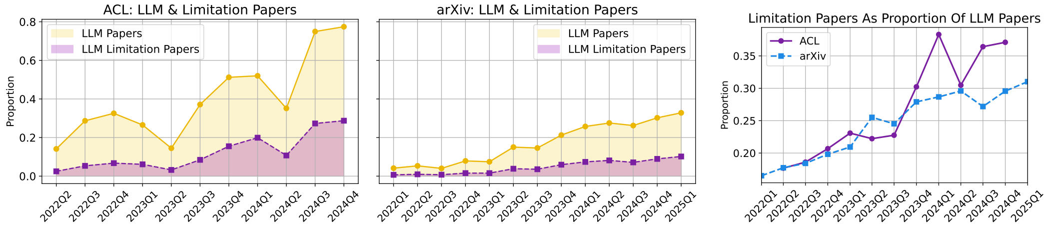
4.2 LLM and LLLMs Trends Over Time

Key Insights

- LLM research is growing rapidly: by late 2024, LLMs account for 75% of ACL papers and over 30% of arXiv papers.
- Research on LLLMs grew even more rapidly, with 1 in 3 LLM papers now addressing limitations.

Before we turn to clustering, we provide an analysis of the number of LLM-related papers (rating 1 or more) as well as limitations-focused papers (rating 3-5) over time. Figure 5 shows:

(i) the proportion of LLM-related and LLM limitation papers among all crawled papers, defined as $\frac{N_t^{\text{LLM}}}{N_t}$ and $\frac{N_t^{\text{Lim}}}{N_t}$, where N_t is the total number of papers at time t, N_t^{LLM} the number of LLM-related papers, and N_t^{Lim} the number of LLM limitation papers;



(i) LLM and limitation papers in ACL and arXiv datasets relative to all crawled papers.

(ii) Proportion of LLM limitation papers among all LLM papers.

Fig. 5. Trends in LLM and LLM limitation research over time. Figure 5i shows the share of LLM and limitation papers among all crawled papers, while Figure 5ii illustrates the proportion of limitation papers within LLM research. Note that the limitation trend in (ii) can rise even if it appears flatter in (i), as (ii) reflects growth relative to LLM research, not all papers.

(ii) the share of limitation papers among LLM-related papers, defined as $\frac{N_t^{\text{Lim}}}{N_t^{\text{LLM}}}$.

In both corpora, the *(i) overall share of LLM-related papers* has grown substantially since early 2023. This trend is particularly steep in ACL, where, by late 2024, over 75% of ACL papers are related to LLMs. This suggests a notable shift in NLP research, with LLMs becoming central to the field. In arXiv, growth is more moderate but consistent, hitting just above 30% of papers by the end of the same period. The lower proportion in arXiv might be due to different levels of engagement with LLM research across the categories in our study, as shown in Figure 17 in the supplementary material. In cs.CL, LLMs are widely discussed, reaching 80% by early 2025, similar to ACL, while in areas like cs.CV and cs.LG, their presence remains below 20%.

The *(ii) share of limitation papers among LLM-related work* has also grown notably. As shown in Figure 5ii, the proportion of LLLMs research has steadily increased in both venues. In ACL, this share climbs sharply through early 2024, peaking at nearly 38% before stabilizing around 35%. In arXiv, the rise is more gradual, reaching approximately 30% by the end of 2024.

Overall, as LLM research accelerates, so does work on their limitations, indicating that the community is not only developing or using new models but also, increasingly, engaging with their risks and shortcomings.

In the following sections, we refine this analysis and examine these emerging discussions in detail through topic clustering.

4.3 Clustering Results (HDBSCAN + BERTopic)

4.3.1 Topics identified within ACL and arXiv with HDBSCAN.

Key Insights

- HDBSCAN identifies 7 limitation topics in ACL and 15 in arXiv, with shared themes including *Reasoning*, *Social Bias*, *Security Risks*, and *Hallucinations*.
- ACL's limitation research is dominated by *Reasoning* (36.4%), while arXiv presents a broader topical spread, including specialized areas like *Healthcare*, *Code*, and *Quantization*.

Figure 6 presents the distribution of LLM limitation topics in the ACL Anthology and arXiv datasets, as identified by the HDBSCAN + BERTopic clustering approach. Across both corpora, the model identified 7 topics for ACL and 15 for arXiv, in addition to a set of outliers (13.1% in ACL, 14.1% in arXiv).

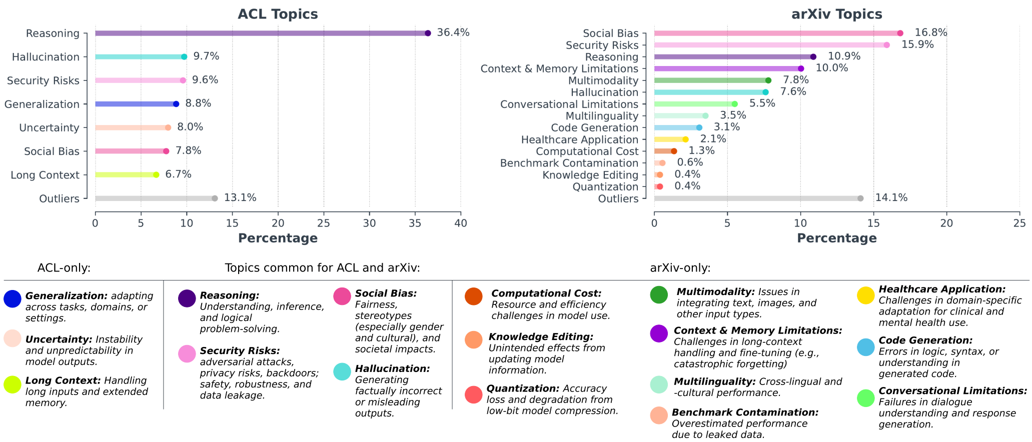


Fig. 6. Topics in ACL Anthology and arXiv, clustered using HDBSCAN + BERTopic. Percentages reflect each topic's proportion out of the total LLM limitation papers (2,387 in ACL and 12,261 in arXiv).

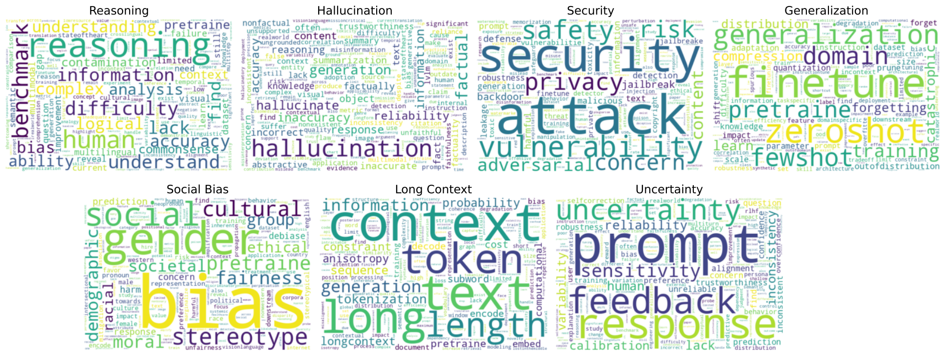
We find that, according to the HDBSCAN approach, ACL's limitation discussions are dominated by a single topic: Reasoning, which accounts for over one-third (36.4%) of all limitation-focused papers. In contrast, arXiv presents a more balanced topical spread, with the most frequent themes being *Social Bias* (16.8%), *Security Risks* (15.9%), and *Reasoning* (10.9%).

Several topics are shared across both corpora, including *Hallucination*, *Social Bias*, and *Security Risks*. However, arXiv also introduces a number of unique topics not observed in ACL, such as *Multimodality*, *Healthcare Applications*, *Multilinguality*, and *Computational Cost*.

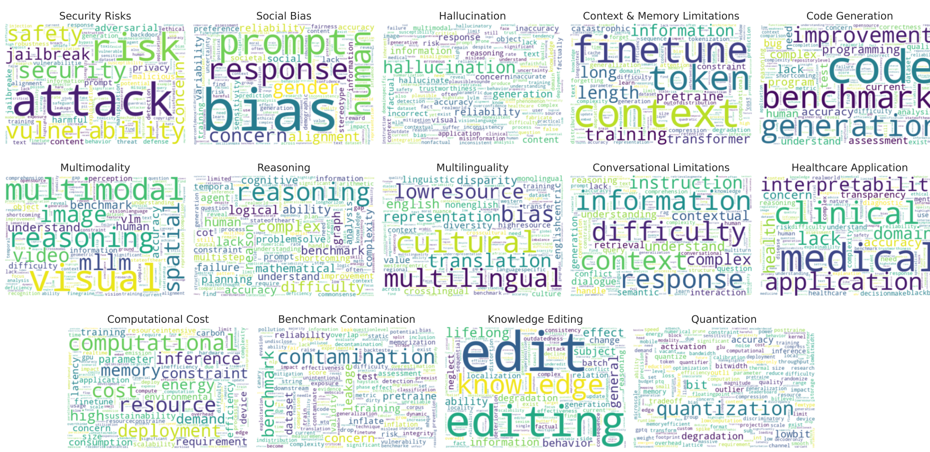
Figure 7i and Figure 7ii show the dominant terms for each limitation cluster in the ACL and arXiv datasets; additionally, Table 7 and Table 8 present representative paper examples for some of the ACL and arXiv clusters, respectively. In ACL, the *Reasoning* cluster includes core cognitive tasks such as natural language understanding (NLU), inference, and logical analysis and reflects a broad range of reasoning types (e.g., temporal, causal, etc.). It spans multimodal tasks (example 1 in Table 7), visual question answering, and overlaps with multilinguality and benchmark design. In arXiv, Reasoning appears more distinct from other topics, though it remains closely linked to NLU, as shown by keywords like *understanding* and *cognitive*. Additionally, it includes prompt-related terms such as *chain-of-thought* and *prompt*, pointing to the role of prompt engineering in enabling complex reasoning. The content differences between ACL (wider topic) and arXiv (narrower topic) may also explain why the reasoning cluster is much more prevalent in ACL compared to arXiv, though this may also be a clustering artefact from HDBSCAN handling datasets differently.

Several clusters focus on the correctness and reliability of model outputs. The *Hallucination* cluster addresses factual accuracy, including terms related to faithfulness, trust, and correctness (e.g., example 2). It also appears in multimodal contexts, with frequent references to hallucinations in image captioning and generation. The *Security* cluster highlights threats such as jailbreaks, adversarial prompts, and backdoors (e.g., example 3 shows adversarial attacks that evade detectors using prompt and synonym manipulation).

The *Social Bias* cluster focuses on fairness, representation, and demographic disparity, as evidenced by keywords such as *fairness*, *representation*, *gender*, *racial*, *cultural*, and *demographic*. It also includes terms tied to ethical and political considerations, including *ethical*, *moral*, *political*,



(i) ACL dataset



(ii) arXiv dataset

Fig. 7. Wordclouds for LLM limitation topics identified by HDBSCAN+Bertopic, generated based on TF-IDF scores computed over concatenated evidence and keyphrases grouped by topic.

and *societal*. In arXiv, the prominence of *prompt*, *sensitivity*, *reliability*, and *variability* suggests an emphasis on prompt-based bias mitigation [118].

Other clusters reflect technical constraints in model design and deployment. These include *Long Context* (ACL) and *Context & Memory Limitations* (arXiv), which address failures on long inputs and outputs, memory constraints, and, as a result, computational costs (example 6 in Table 7 and example 8 in Table 8). The *Generalization* cluster captures similar issues related to robustness and domain transfer (example 4).

Some clusters are specific to one dataset. In ACL, the *Uncertainty* cluster describes behavioral instability, including prompt sensitivity, calibration errors, and limited self-correction. In arXiv, additional clusters capture a wider range of specialized concerns: *Multimodality*, *Conversational Limitations*, *Multilinguality*, *Code Generation*, *Healthcare Application*, *Computational Cost*, *Benchmark Contamination*, *Knowledge Editing*, and *Quantization*. These clusters reflect both domain-specific challenges (e.g. *Healthcare Application*, which focuses on the use of LLMs in clinical settings and is primarily concerned with their black-box nature, see example 10) and implementation-level trade-offs. For example, *Benchmark Contamination* refers to test data leakage into training sets (example 12), *Knowledge Editing* focuses on updating model content without retraining (example 14), and *Quantization* and *Computational Cost* concern scaling and deployment efficiency (examples 13 and 11).

Table 7. Representative limitation-focused ACL papers, grouped by topic based on HDBSCAN clustering. Evidence is extracted from papers with a topic probability of 0.9 or higher. The numbers in parentheses in the "Topic" column indicate the example ID referenced in the main text.

Topic	Evidence
(1) Reasoning	"However, applying this [common-sense] reasoning to multimodal domains, where understanding text and images together is essential, remains a substantial challenge." (ACL 2024 [81])
(2) Hallucination	"Challenges on hallucination and factual inconsistency continue to impede their [LLMs'] wider real-world adoption. [...] However, challenges remain, particularly regarding... generating information not present in the evidence (hallucination)." (EACL 2024 [65])
(3) Security	"The prevalence and strong capability of large language models (LLMs) present significant safety and ethical risks if exploited by malicious users. [...] Experiments reveal that our attacks effectively compromise the performance of all detectors in the study with plausible generations [...]." (TACL 2024 [95])
(4) Generalization	"LLMs are generally trained on publicly available text and code and cannot be expected to directly generalize to domain-specific parsing tasks in a zero-shot setting [...]." (EMNLP 2023 [73])
(5) Social Bias	"We find that masked language models capture societal stigma about gender in mental health: models are consistently more likely to predict female subjects than male in sentences about having a mental health condition (32% vs. 19%) [...]." (EMNLP 2022 [59])
(6) Long Context	"However, they face challenges in managing long documents and extended conversations, due to significantly increased computational requirements, both in memory and inference time [...]." (EMNLP 2023 [53])
(7) Uncertainty	[...] "We find that methods aimed at improving usability, such as fine-tuning and chain-of-thought (CoT) prompting, can lead to miscalibration and unreliable natural language explanations." (NAACL 2024 [122])

While HDBSCAN provides an interpretable, unsupervised clustering approach, it has several limitations for trend analysis. HDBSCAN assigns each paper to only one cluster and becomes increasingly fragmented as more clusters are added (especially in ACL) making it difficult to capture overlapping limitations or analyze trends at finer topical granularity. To address these issues, we now turn to the LlooM approach, which supports multi-label assignment and allows for broader topical coverage. The following section presents the main findings using LlooM, including trend analysis of `LLLMs` topics; additional trend results from HDBSCAN are included in Section C in the supplementary material for completeness.

4.4 Clustering Results (LlooM)

4.4.1 Topics Identified within ACL and arXiv.

Table 8. Representative limitation-focused arXiv papers, grouped by topic based on HDBSCAN clustering. Only arXiv-specific clusters are shown; topics shared with ACL are excluded to avoid redundancy. Examples are selected based on a topic probability of 0.9 or higher. The numbers in parentheses in the "Topic" column indicate the example ID referenced in the main text.

Topic	**Evidence**
(8) Context & Memory	"Experiments on CoIN demonstrate that current powerful MLLMs still suffer catastrophic forgetting, and the failure in intention alignment assumes the main responsibility, instead of the knowledge forgetting." (arXiv, March 2024 [12])
(9) Conversational Limitations	"[...] Answers from LLMs can be improved with additional context. [...] Our results show multi-turn interactions are usually required for datasets which have a high proportion of incompleteness or ambiguous questions" (arXiv, March 2025 [76])
(10) Healthcare Application	"However, GPT3.5 performance falls behind BERT and a radiologist. [...] By analyzing the explanations of GPT3.5 for misclassifications, we reveal systematic errors that need to be resolved to enhance its safety and suitability for clinical use." (arXiv, June 2023 [99])
(11) Comput. Cost	"Training and deploying LLMs are expensive as it requires considerable computing resources and memory...[...]" (arXiv, November 2023 [123])
(12) Benchmark Contamination	"However, as LLMs are typically trained on vast amounts of data, a significant concern in their evaluation is data contamination, where overlap between training data and evaluation datasets inflates performance assessments." (arXiv, October 2023 [25])
(13) Quantization	"This study examines 4-bit quantization methods like GPTQ in large language models (LLMs), highlighting GPTQ's overfitting and limited enhancement in Zero-Shot tasks." (arXiv, December 2023 [115])
(14) Knowledge Editing	"[...] Extensive experiments on DepEdit show that existing knowledge editing methods are sensitive to the surface form of knowledge, and that they have limited performance in inferring the implications of edited facts." (arXiv, December 2023 [55])

Key Insights

- LlooM identifies a more fine-grained set of topics than HDBSCAN: 13 topics for ACL, including new topics like *Multimodality*, *Language and Culture*, and *Knowledge Editing*; and 15 for arXiv, adding concerns such as *Alignment*, *Trustworthiness*, and *Generalization*.
- *Reasoning*, *Generalization*, and *Hallucination* are top-ranking topics in both ACL and arXiv. Beyond these shared concerns, ACL places additional emphasis on *Knowledge Editing*, while arXiv is led by *Trustworthiness* and *Alignment*.

Figure 8 shows the distribution of LLM limitation topics in the ACL and arXiv datasets as identified by the LlooM method. The share of outliers is 7.5% in ACL and 6.7% in arXiv.

In ACL, *Reasoning* remains the most prominent topic (26.3%), followed by *Generalization* (24.8%) and *Knowledge Editing* (21.2%). Compared to HDBSCAN, the distribution across top limitations appears more balanced and can be explained by the fact that LlooM allows papers to belong to multiple clusters. ArXiv is led by *Trustworthiness* (21.0%), followed by *Reasoning* (13.2%), *Generalization* (10.1%), as well as *Alignment Limitations*, *Hallucination*, and *Bias and Fairness*, which appear at nearly equal levels (~7.9%).

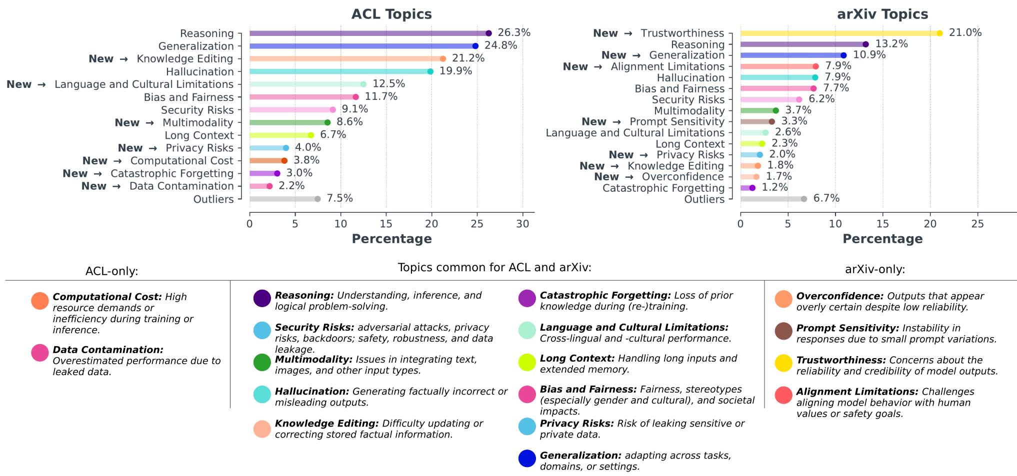


Fig. 8. Topics in ACL Anthology and arXiv, clustered using LlooM approach. Percentages reflect each topic's proportion out of the total LLM limitation papers (2,387 in ACL and 12,261 in arXiv). Since papers can be associated with multiple topics, the percentages may exceed 100% in total.

A number of topics are specific to LlooM. These are listed below, with explanations based on the prompts used by LlooM to guide topic assignment (see Table 14 and Table 15 in the supplementary material). For each cluster, we also provide representative paper examples in Table 9.

- *Trustworthiness* (arXiv): the largest cluster in the arXiv set (Figure 8). This category describes concerns about the reliability, transparency, and reproducibility of LLM outputs (see example 20 which refers to concerns about the reliability of outputs generated by LMs). As a broad category, it often overlaps with related issues like hallucination or alignment, as discussed further in Section D of the supplementary material.
- *Generalization* (arXiv): Previously ACL-only in HDBSCAN, this cluster now appears in arXiv as well. It captures failures across domains or tasks; e.g. example 19 shows reliance on shallow heuristics over true generalization.
- *Alignment Limitations* (arXiv): Highlights challenges in aligning LLMs with human values or safety protocols (see example 20 which discusses how models can generate outputs that are untruthful, toxic, or unhelpful despite alignment efforts).
- *Prompt Sensitivity* (arXiv): Highlights performance instability when prompts are minimally edited.
- *Language and Cultural Limitations* (ACL): difficulties in handling multilingual input, low-resource languages, or culturally specific content, previously grouped under the Reasoning cluster in HDBSCAN.
- *Overconfidence* (arXiv): Captures cases where LLMs express high certainty despite being incorrect, often due to poor calibration (example 22 shows how persuasive language can mask factual errors). This topic is closely related to the *Uncertainty* cluster identified in ACL under HDBSCAN.
- *Knowledge Editing* (ACL & arXiv): efforts to modify, correct, or update model knowledge post-training: for instance, example 15 (ACL) in Table 9 introduces methods for editing factual content without retraining.

- *Privacy Risks* (ACL & arXiv): previously part of the *Security* cluster in HDBSCAN for both datasets, this now is a distinct category in LlooM. It captures risks of leaking sensitive training data via model outputs or queries. Example 21 (arXiv) demonstrates privacy breaches from pretraining on sensitive data.
- *Multimodality* (ACL): challenges in integrating and reasoning over inputs from different modalities, such as text and images (example 16, ACL), previously grouped under the Reasoning cluster in HDBSCAN.
- *Computational Cost* (ACL): the high memory, compute, and energy demands of training or deploying LLMs (example 17).
- *Data Contamination* (ACL): inflated evaluation results caused by overlap between training and test datasets. While this topic appeared in arXiv under HDBSCAN (as *Benchmark Contamination*), LlooM identifies it only in ACL (see example 18 which highlights concerns about memorization skewing evaluation).

In summary, the LlooM clustering identifies a broader set of limitation types than HDBSCAN, including implementation-level issues (e.g., *Privacy*, *Computational Cost*) and behavioral factors (e.g., *Overconfidence*, *Prompt Sensitivity*) that were previously grouped under broader categories. It also reflects some dataset-specific shifts, such as *Generalization* appearing in arXiv and *Data Contamination* in ACL. In the next section, we examine how attention to these topics has changed over time.

4.4.2 Trend Analysis. In this section, we discuss three perspectives on topic dynamics over time:

(i) **LLM-wide share**, measured annually as $\frac{N^{\text{lim}}_{k,y}}{N^{\text{LLM}}_{y}}$, to reflect how often limitation topic k appears in LLM research in year y, relative to the total LLM papers. This shows whether a topic is gaining attention beyond limitations research and becoming part of the general LLM research agenda.

(ii) **Limitations share**, measured quarterly as $\frac{N^{\text{lim}}_{k,q}}{N^{\text{lim}}_{q}}$, to reflect the share of limitation-focused papers in quarter q that address topic k. Note the different denominator compared to the LLM-wide share (i): this metric is limited to the subset of limitation-focused papers to show the topic's visibility within the limitations-focused subfield.

(iii) Notable shifts in topic trajectories, such as spikes, dips, and periods of stabilization.

Here, $N^{\text{lim}}_{k,y}$ is the number of limitation papers on topic t in year y; N^{LLM}_{y} is the total number of LLM papers in that year; and $N^{\text{lim}}_{k,q}$, N^{lim}_{q} are the number of limitation papers on topic k and the total number of limitation papers, respectively, in quarter q.

(i) How are limitation topics represented in the broader growth of LLM research?

Key Insights

- The presence of limitation topics in LLM research is increasing across both ACL and arXiv datasets. Topics like *Hallucination*, *Multimodality*, and *Long Context* surge in 2023 and 2024, while longer-standing ones like *Reasoning* grow more gradually and consistently over time.
- However, this rise may simply reflect the rapid growth of the limitation field itself.

We begin by examining how visible different limitation topics are across the broader LLM research field. Figure 9 shows the annual distribution of LLM limitation topics across ACL and arXiv, normalized by all LLM-focused papers. These proportions reflect the *overall visibility* of each topic. Additionally, to capture how visibility changes over time, we compute the *relative percentage*

Table 9. Representative limitation-focused papers from ACL and arXiv, grouped by topic. Only topics identified by LlooM and absent from HDBSCAN-based analyses are included. The numbers in parentheses in the "Topic" column indicate the example ID referenced in the main text.

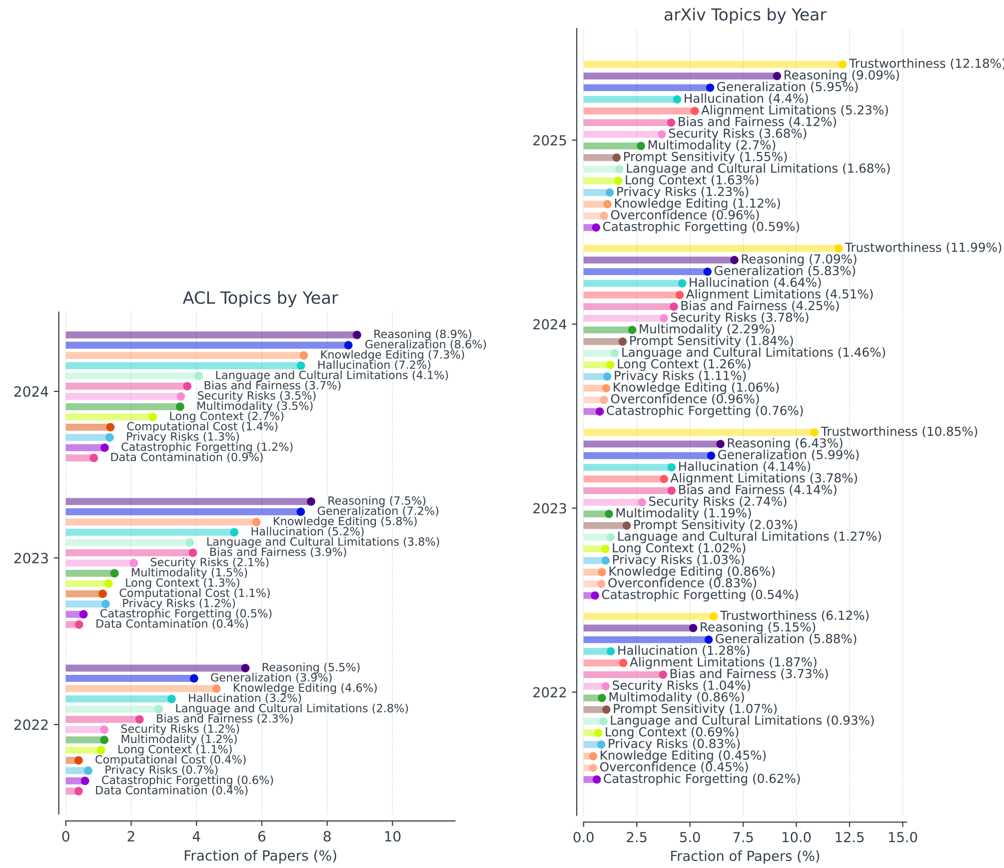
Topic	Evidence
(15) Knowledge Editing	"Even the most advanced language models remain susceptible to errors necessitating to modify these models without initiating a comprehensive retraining process." (EMNLP 2023 [5])
(16) Multimodality	"We experimented with state-of-the-art vision and LMs and found that the best (22%) performed substantially worse than humans (97%) in understanding figurative language." (EMNLP 2023 [120])
(17) Comput. Cost	"However, their large size and computational demands, coupled with privacy concerns in data transmission, limit their use in resource-constrained and privacy-centric settings." (NAACL 2024 [41])
(18) Data Contamination	"Data contamination in model evaluation has become increasingly prevalent with the growing popularity of LLMs. It allows models to cheat via memorisation instead of displaying true capabilities." (EMNLP 2024 [54])
(19) Generalization	"We suggest that the lack of generalization [...] means that the PLMs are currently not learning NLI, but rather spurious heuristics." (arXiv, January 2022 [30])
(20) Alignment & Trustworthiness	"LLMs can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. [...]" (arXiv, March 2022 [79])
(21) Privacy Risks	"LLMs [...] have been shown to memorize instances of training data thereby potentially revealing private information processed during pre-training." (arXiv, May 2022 [66])
(22) Overconfidence	"Despite their broad utility, LLMs tend to generate information that conflicts with real-world facts, and their persuasive style can make these inaccuracies appear confident and convincing." (arXiv, September 2024 [10])

change in LLM-normalized topic share from year y to $y+1$, defined as $\frac{\text{Share}_{y+1}-\text{Share}_{y}}{\text{Share}_{y}} \times 100$, where Share_{y} refers to the LLM-wide share of a given topic in year y, i.e., the proportion of all LLM papers that address that topic (see Table 16 in supplementary material).

As seen in Figure 9, most limitation topics show an increase in visibility within LLM research over the years. However, topics differ in how their share of LLM research changes over time. Some concerns surged in visibility within LLM research at specific moments (more than doubling in share), such as *Multimodality* (+133%), *Long Context* (+108%), *Catastrophic Forgetting* (+140%) in ACL 2024, and *Hallucination* (+223%), *Security Risks* (+163%), and *Alignment Limitations* (+102%) on arXiv in 2023, reflecting heightened attention to certain types of LLLMs following widespread LLM deployment. Others, like *Reasoning* and *Knowledge Editing*, show steadier growth across venues. Meanwhile, topics such as *Bias and Fairness* and *Language and Cultural Limitations* peaked in ACL 2023 but declined in 2024 (from +70% and +36% to −5% and +8%, respectively), while concerns like *Prompt Sensitivity* and *Overconfidence* on arXiv fell steadily after brief spikes. Since the data for 2025 data is incomplete, recent drops should be interpreted with caution.

Overall, LLM-normalized trends confirm that limitation topics are becoming more prevalent within the broader LLM research. Most topics show growth, some slower, some rapidly. However, a

Fig. 9. Distribution of LLM limitation topics over years for ACL and arXiv, based on clustering results with LlooM. Percentages reflect each topic's proportion out of the total LLM-focused papers (8,635 in ACL and 41,991 in arXiv).



(i) ACL Topic Distribution Per Year (LlooM clustering approach)

(ii) arXiv Topic Distribution Per Year (LlooM clustering approach)

topic's increasing presence in LLM research may simply reflect the overall expansion of limitation research, rather than increased relative focus on that specific topic. Therefore, in the next bullet point, we examine whether these trends hold within limitation-focused work.

(ii) What are the trends within LLM limitation research?

Key Insights

- Within LLLMs research, most limitation topics remain stable from 2022 to 2025 in both ACL and arXiv, with only a few showing significant shifts.
- *Long Context* increases significantly in ACL, while *Multimodality*, *Security Risks*, and *Alignment Limitations* rise in arXiv.
- *Generalization* and *Bias and Fairness* decline significantly in arXiv, but no topics show statistically significant decline in ACL.

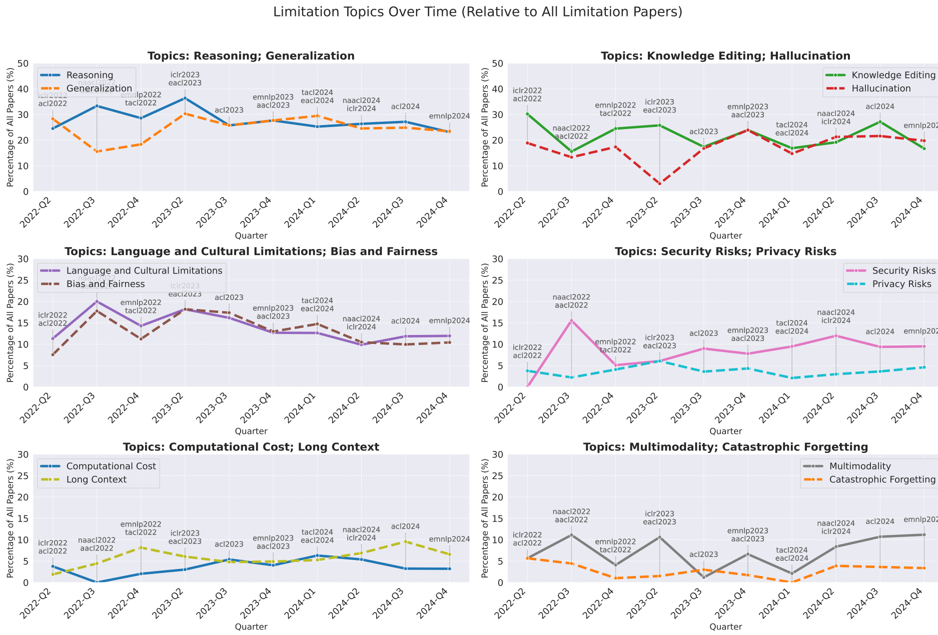


Fig. 10. LLLMs topics trends for the ACL dataset based on LLooM clustering approach. Note that y-axis limits vary across subplots to reflect differences in topic prevalence and improve visualization.

Figure 10 shows how the distribution of limitation topics within the ACL dataset has changed over time. We evaluate the significance of these trends using the Mann-Kendall test [45, 67] for monotonic trend detection.

- ↑ **Increasing:** *Long Context* shows an upward trend, rising from around 2% in 2022-Q2 to a peak of 10% by 2024-Q3. This trend is statistically significant according to the Mann-Kendall test ($\tau = 0.51$, $p = 0.0491$). *Security Risks* also shows a similar upward trajectory ($\tau = 0.47$), though it does not reach significance at the $p < 0.05$ threshold.
- ↓ **Decreasing:** No topics show statistically significant declines. However, *Bias and Fairness* drops from ∼17% to 10%, and *Language and Cultural Limitations* from ∼20% to 12% between 2023-Q2 and 2024-Q4. *Computational Cost*, after a steady rise in the period from late 2022 to late 2023, decreases from ∼5% in the late 2023 to below 5% by 2024-Q4. None of these changes are statistically significant.
- → **Stable or Fluctuating:** Most topics remain steady over time: *Reasoning*, *Generalization*, and *Hallucination* fluctuate between 10–35%, while *Knowledge Editing* varies more widely (18–39%) and *Multimodality* stays between 6–11%, with a slight increase after early 2024. *Privacy Risks* (3–6%), *Security Risks* (peaking at 15% in 2022-Q3 but mostly under 10%), and *Catastrophic Forgetting* (below 5%) remain consistently low.

For the arXiv dataset, we observe the following trends over time according to LlooM (Figure 11):

- ↑ **Increasing:** *Multimodality*, *Security Risks*, *Alignment Limitations*, and *Knowledge Editing* all show statistically significant upward trends according to the Mann-Kendall test ($p < 0.05$). *Multimodality* rises from approximately 2% in 2022 to nearly 10% by late 2024, *Security Risks* increases from around 9% to 11%, and *Alignment Limitations* grows from a post-2022 dip to around 18% by 2025. *Knowledge Editing* increases early on but stabilizes below 5%. Notably,

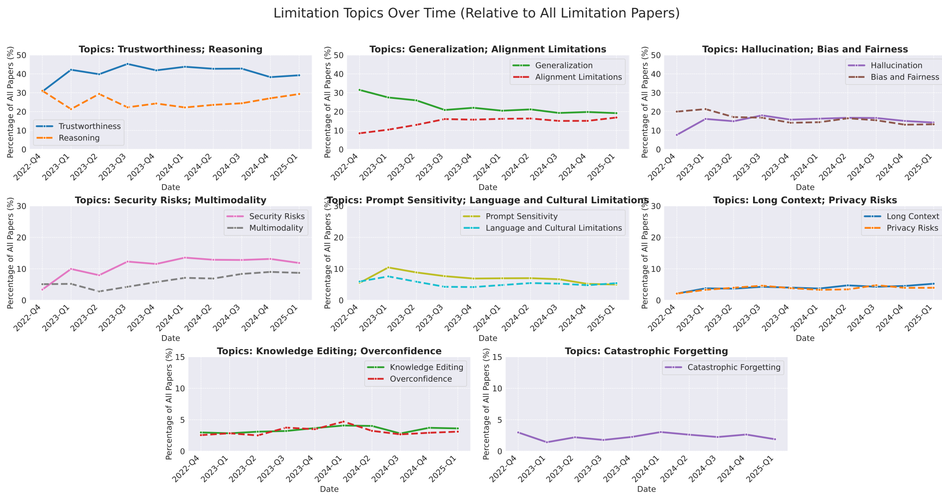


Fig. 11. LLLMs topics trends for the arXiv dataset based on LLooM clustering approach. Note that y-axis limits vary across subplots to reflect differences in topic prevalence and improve visualization.

Hallucination also shows a positive trend ($\tau = 0.38$), but this increase is not statistically significant under the Mann-Kendall test.

- ↓ **Decreasing:** *Generalization* declines from around 35% in the second quarter of 2022 to 20%, and *Bias and Fairness* drops from a peak near 25% to ~13%. Both of these downward trends are statistically significant ($p < 0.05$).
- → **Stable or Fluctuating:** Most remaining topics show no significant directional movement. *Reasoning* stays around 20–30%, *Trustworthiness* holds at 40–45%, and *Overconfidence*, *Long Context*, and *Privacy Risks* remain within narrow ranges. *Prompt Sensitivity* declines from 10% to ~5% (2023–2025), while *Hallucination* and *Language and Cultural Limitations* rise modestly and then level off. *Catastrophic Forgetting* stays consistently below 4%.

These results confirm that not all LLM-wide trends carry over into increased focus within limitation-specific research. While some topics, such as *Multimodality*, *Security Risks*, *Alignment Limitations*, and *Knowledge Editing*, do show consistent growth within limitation-focused work, the majority remain flat or variable despite gaining visibility across the broader LLM field, as discussed earlier. However, the Mann-Kendall test only captures consistent upward or downward trends, not short-term changes, which may explain why topics like *Security Risks* (ACL) and *Hallucination* (arXiv) show visible growth without being statistically significant. We examine these kinds of short-term spikes and dips in the next paragraph.

(iii) How do topics shift in ways not captured by overall trends?

Key Insights

- Limitation topics stabilize around 2023-Q2, either plateauing or beginning steady growth, after earlier volatility. For example, technical concerns (e.g., *Hallucination*, *Alignment Limitations*) rise and plateau, while social topics decline after 2023-Q2.
- This shift coincides with rising paper volume and the release of ChatGPT and other major models in early 2023.

Across both ACL and arXiv, 2023-Q2 marks a shift in the LLLMs research (Figure 10, Figure 11). Before this, topic shares are volatile (see e.g. a spike in *Security Risks* in 2022-Q3), particularly in ACL. After early 2023, topics begin to stabilize across both datasets: *Reasoning* levels off, *Generalization* remains flat or slightly declines before stabilizing, and newer concerns like *Security Risks*, *Hallucination*, and *Alignment Limitations* rise sharply and then plateau. *Multimodality* stabilizes somewhat later, starting a steady increase in ACL around early 2024, but earlier in arXiv (around Q2 2023). In contrast, socially focused topics such as *Bias and Fairness* and *Language and Cultural Limitations* decline in share after mid-2023, reflecting the integration of new concerns into the discourse.

The stabilization of topic trends coincides with a sharp rise in raw paper counts beginning in 2023 (see Figure 16), indicating not just increased research volume, but a shift toward a more coherent field. Before 2023-Q2, most topics appear in fewer than 25 ACL papers and under 100 in arXiv, making early signals harder to interpret. This growth aligns with the release of ChatGPT in November 2022, as well as the emergence of other major models like GPT-4 [1], PaLM [16], and LLaMA [103] between February and July 2023, which likely contributed to the expansion and differentiation of LLM limitations research during this period.

4.4.3 LLLMs Topics Distribution Across ArXiv Categories. Our analysis shows that LLLMs span a broad range of concerns, from reasoning and generalization to bias, safety, and multimodality. ArXiv's category system offers a way to examine how different research communities engage with these topics. We analyze topic distributions across categories to understand where this work is published and which concerns dominate in specific domains.

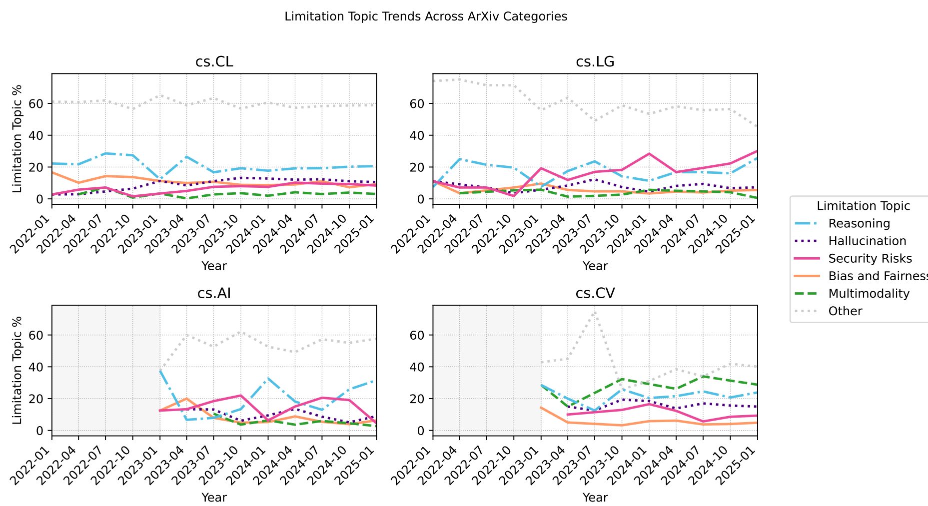
In our dataset, most LLLMs papers are concentrated in `cs.CL` (Computation & Language; 58.7%), followed by `cs.AI` (Artificial Intelligence; 8.7%), `cs.CV` (Computer Vision; 6.6%), and `cs.LG` (Machine Learning; 3.3%). This is expected, since these categories were used as our arXiv search criteria. Notably, we also observe papers with categories like `cs.CY` (cybersecurity), `cs.SE` (software engineering), and `cs.HC` (human-computer interaction), which appear as a result of multi-categorization.

Although many categories share topics such as *Trustworthiness*, *Reasoning*, *Generalization*, and *Alignment Limitations*, both their overall topic composition and temporal dynamics vary by field. Figure 18 (right) in the supplementary material, shows how topic shares differ across arXiv categories, while Figure 12 illustrates how topics in the four largest categories evolve over time. **Key trends in the largest categories are as follows:**

- `cs.CL` (Computation and Language) covers nearly all LLLMs, with *Reasoning* dominating across the full time range. *Bias and Fairness* rises mid-2023, but is overtaken by *Hallucination* by the end of the year. Other topics like *Security* and *Multimodality* remain marginal.
- `cs.LG` (Machine Learning) and `cs.AI` (Artificial Intelligence) follow similar distributions to `cs.CL`, but put more weight on *Security Risks* (10.8% in cs.LG, 9.1% in cs.AI). In cs.LG, this topic rises sharply in late 2023, reflecting growing concern with adversarial attacks. In contrast, cs.AI shows more fluctuation, alternating between a focus on *Reasoning* and *Security Risks*, indicating a split between safety and inference evaluation concerns.
- `cs.CV` (Computer Vision and Pattern Recognition) diverges from the others in its dominant focus on *Multimodality* (21.7%), which becomes the leading limitation category from early 2023 onward, driven by the rise of vision-language models. *Hallucination* and *Reasoning* remain present but secondary.

Beyond the four primary categories included in our analysis, several **smaller arXiv categories** show up, often exhibiting a clear focus on domain-specific concerns. Figure 13 highlights six such cases. `cs.CY` (Computers and Society) and `cs.HC` (Human-Computer Interaction) emphasize value alignment and societal impact, with high shares of *Social Bias* and *Alignment Limitations*,

Fig. 12. Limitation topic trends across four main arXiv categories (cs.CL, cs.LG, cs.AI, cs.CV). To maintain visual clarity, only the five most frequent topics (based on their overall percentage distribution across categories, as shown in Figure 18 in the supplementary material) are shown; others are grouped as "Other." Grey shading in cs.AI and cs.CV marks periods before 2023 with insufficient data for these categories.



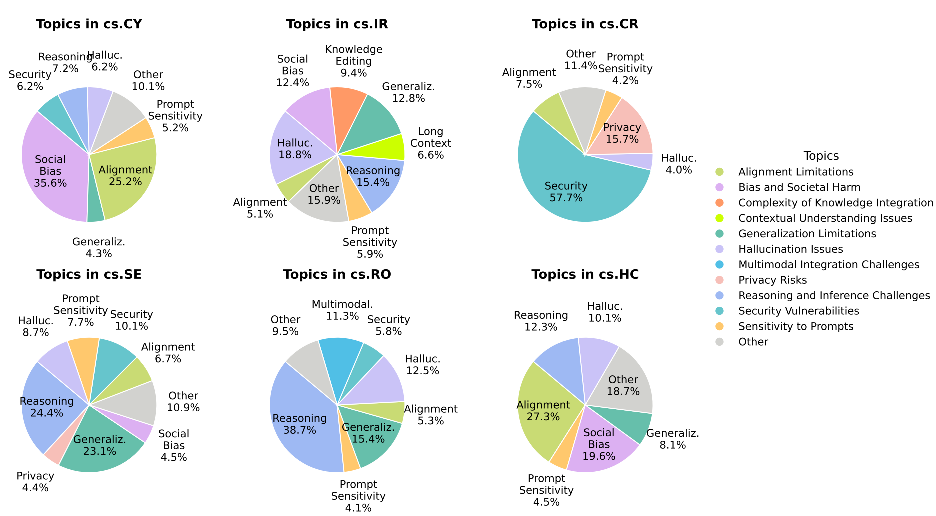


Fig. 13. Distribution of limitations-related topics in six arXiv categories with lower paper counts in our dataset, as shown in Figure 18. Each chart includes only papers where the category is assigned as the primary arXiv category. Topics that make up less than 3% are grouped under *Other*.

reflecting ethical and user-centered concerns. `cs.CR` (Cryptography and Security) is dominated by *Security Risks* (57.7%), consistent with its focus on adversarial threats and privacy vulnerabilities. `cs.IR` (Information Retrieval) distributes attention across *Hallucination*, *Reasoning*, and *Knowledge Editing*, likely due to challenges in document-grounded generation and factual consistency. `cs.SE` (Software Engineering) frequently discusses *Reasoning* and *Generalization*, which aligns with LLMs used in code generation and developer tooling. Finally, `cs.RO` (Robotics) highlights *Reasoning* and *Multimodality*, reflecting perception and control challenges in embodied settings. Though smaller in volume, these categories reflect more targeted concerns tied to specific application domains.

Together, these disciplinary patterns illustrate how research on `LLLMs` is not only growing, but also diversifying in focus based on domain needs.

4.5 Method and Output Comparison: HDBSCAN vs. LlooM

The aforementioned topic distributions differ depending on the clustering method. In this section, we compare the two approaches used, HDBSCAN+BERTopic (see Section C in the supplementary material) and LlooM (see Section 4), to identify which findings are stable and which may be method-specific. We first compare trend agreement, then examine methodological differences to better explain any observed divergences in their results.

To compare the trends, we report (i) Kendall's Tau for each individual topic (as determined in the trend analyses in Section C in the supplementary material and Section 4) to assess alignment in trend direction and significance, and (ii) the Spearman correlation of the time series between matched topics from HDBSCAN and LlooM to assess the similarity of overall trend shapes. To identify matching topics between HDBSCAN and LlooM, we select the best match for each cluster based on identical or semantically similar names. To validate these matches, we compute the Jaccard overlap between their associated paper sets. Specifically, for each cluster produced by HDBSCAN, we compute its Jaccard similarity with all LlooM topics, and vice versa. Given two paper sets X from HDBSCAN and Y from LlooM, the Jaccard similarity is defined as $\text{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$. We select the top-1 Jaccard score for each cluster, representing its highest similarity with any topic from the other method. Best-matching topic pairs are shown in Table 17 and Table 18 in the supplementary material, confirming that major topics identified by name also show substantial paper overlap.

Trend Alignment Between HDBSCAN and LlooM. Table 10 summarizes trend agreement between the two clustering approaches across matched topics in the ACL and arXiv datasets.

In the ACL dataset, 4 out of 6 matched topics (67%) share the same trend direction between HDBSCAN and LlooM based on Kendall's Tau. However, only half of these (33% overall) are also aligned in trend significance. Spearman ρ values for trend shape similarity are generally moderate to low and mostly not significant, with the exception of *Security/Security Risks*.

The arXiv dataset shows strong overall agreement between the HDBSCAN and LlooM clustering approaches. Most topics demonstrate matching trend directions according to Kendall's Tau (6 out of 8 topics, 75%), and the quarter-to-quarter fluctuations also correlate strongly, as reflected by high Spearman ρ values. For instance, *Multimodality* ($\rho = 0.86$) and *Multilinguality* ($\rho = 0.92$) achieve strong and statistically significant trend similarity across methods. Nonetheless, slight divergences remain: although trend directions often align, significance levels or trend shapes occasionally differ. For example, *Hallucination* trends upward in both methods but is statistically significant only in HDBSCAN, with a moderate but insignificant trend similarity ($\rho = 0.55$). One notable case of strong disagreement is *Long Context*, which displays both opposing trend directions and poor trend shape similarity ($\rho = -0.22$). These stronger results may reflect the greater reliability of the arXiv dataset due to its larger size, in contrast to the smaller ACL sample.

Sources of Divergence Between HDBSCAN and LlooM. Although trend agreement between HDBSCAN and LlooM is stronger in the arXiv dataset compared to ACL, it is still not fully consistent across all topics. These differences likely reflect methodological differences between the clustering pipelines. To better understand this, we compare HDBSCAN+BERTopic and LlooM in Table 11

Table 10. Comparison of limitation trends identified by HDBSCAN and LlooM for ACL and arXiv datasets. Trend direction (↑ *increasing*, → *flat*, ↓ *decreasing*) is based on Kendall's Tau from the Mann-Kendall test. Significance is indicated with an asterisk (*) and reported only for increasing or decreasing trends. Match symbols: ✓= *full agreement (direction and significance)*, ✓ = *partial agreement (direction only)*, ×= *disagreement.* For Spearman ρ values, asterisk (*) indicates $p > 0.05$ (significant correlation).

(i) ACL Dataset

HDBSCAN Topic	LlooM Topic	HDBSCAN	LlooM	Match	Spearman ρ
Security	Security Risks	→	↑	×	0.661*
Generalization	Generalization	→	→	✓	0.552
Social Bias	Bias and Fairness	→	→	✓	0.539
Hallucination	Hallucination	↑	→	×	0.539
Reasoning	Reasoning	↓	→	×	0.430
Long Context	Long Context	→	↑*	×	0.006

(ii) arXiv Dataset

HDBSCAN Topic	LlooM Topic	HDBSCAN	LlooM	Match	Spearman ρ
Multimodality	Multimodality	↑*	↑*	✓	0.855*
Hallucination	Hallucination	↑*	↑	✓	0.552
Context & Memory Lim.	Long Context	↓*	→	×	-0.224
Knowledge Editing†	Knowledge Editing†	↑	↑*	✓	0.632*
Security Risks	Security Risks	↑	↑*	✓	0.782*
Multilinguality	Language & Cultural Lim.	→	→	✓	0.927*
Social Bias	Bias & Fairness	→	↓*	×	0.758*
Reasoning	Reasoning	→	→	✓	0.624

† *Topics are matched by Jaccard overlap, except for Knowledge Editing, which was manually aligned based on topic names due to weak overlap.*

across both datasets. We report overall clustering characteristics (e.g., number of topics) and alignment metrics: (i) average top-1 Jaccard scores across clusters for topic-level similarity, and (ii) Adjusted Mutual Information (AMI) for structural agreement.[4]

Compared to HDBSCAN, LlooM achieves slightly higher coverage of papers across both datasets due to multi-topic assignment. Moreover, the topic-level Jaccard overlaps between methods are only moderate (0.313 for ACL, 0.244 for arXiv), and overall structural alignment, as measured by AMI, remains relatively low (0.229 for ACL, 0.221 for arXiv).

These results suggest that while LlooM and HDBSCAN identify similar broad limitation areas, they organize papers differently at a finer-grained level. This pattern is further supported by Table 17 and Table 18 in the supplementary material, which show the most closely aligned topics across methods. Some large limitation areas appear relatively stable across clustering strategies: e.g. topics such as *Reasoning*, *Hallucination*, *Security Risks*, and *Bias and Fairness* align well across datasets

[4]AMI compares how often data points are grouped similarly, while adjusting for the similarity that would be expected by random chance. To account for LlooM's multi-topic assignments, we compute a shuffle-based baseline: for each paper, we randomly select one of its LlooM topics and compare it to the HDBSCAN label. This process is repeated over 10 runs, and we report the mean and standard deviation.

Table 11. Comparison of HDBSCAN and LlooM clustering methods across key metrics.

Metric	**HDBSCAN+BERTopic**	**LlooM**
# of Topics	7 (ACL), 15 (arXiv)	13 (ACL), 15 (arXiv)
% of Papers Assigned	86.9% (ACL), 85.9% (arXiv)	92.5% (ACL), 93.5% (arXiv)
Avg. Topics per Paper	1	1.5 (ACL), 1.8 (arXiv)
Avg. Jaccard Overlap (top-1)	0.313 (ACL), 0.201 (arXiv)	0.239 (ACL), 0.244 (arXiv)
AMI Shuffled	0.2285 ± 0.0085 (ACL), 0.2206 ± 0.0028 (arXiv)	

and clustering approaches ($J > 0.4$), whereas smaller topics, such as *Overconfidence* and *Prompt Sensitivity*, are often merged into broader categories (e.g., *Hallucination*). This reflects differences in clustering granularity: LlooM tends to split topics into overlapping subcategories, while HDBSCAN merges related issues into broader clusters. For instance, HDBSCAN combines LlooM's *Security Risks*, *Privacy Risks*, and *Trustworthiness* into a single *Security* topic.

While broad limitation areas and general trend directions are reliably identified across HDBSCAN and LlooM, finer-grained topic structures and trend significance vary depending on the clustering method, which highlights that clustering choice impacts the interpretation of limitation trends. We return to these methodological considerations in Section 5.

5 Discussion and Conclusion

Based on the detailed results in Section 4, we conclude four major findings.

1. LLM limitation research grew rapidly in 2022-2025, outpacing even the overall growth of LLM research. LLM research now dominates NLP and increasingly influences neighboring fields: by the end of 2024, over 75% of ACL papers and more than 30% of arXiv submissions across cs.CL, cs.AI, cs.LG, and cs.CV focus on LLMs, with growth continuing into 2025. Within arXiv, LLM engagement in cs.CL closely mirrors ACL trends (reaching 80%), while areas like cs.CV and cs.LG remain below 20% but show steady growth. While only about 10% of LLM-related papers in early 2022 focused on limitations, the fraction increased to about one third by 2025. This growth in `LLLMs` research may indicate a maturation of LLM research: the very early enthusiasm for LLMs and their capabilities, driven by the public deployment of systems like ChatGPT, is now increasingly accompanied with a more critical perspective towards limitations [31]. Meta-analyses confirm this trend, showing a sharp rise in evaluation-focused papers from 2020 to 2023 [9].

2. Within `LLLMs` *research, reasoning is the most frequent topic, but research is diverse.* *Reasoning* is the most frequent limitation topic in ACL across both clustering approaches and remains among the top concerns in arXiv, ranking third in HDBSCAN and second (after *Trustworthiness*) in LlooM. Other prominent topics include *Generalization*, *Hallucination*, *Bias*, and *Security*.

Beyond these, LLM limitation research is notably diverse. Our clustering analyses (HDBSCAN and LlooM) reveal a broad spectrum of concerns, ranging from *code generation* and *benchmark contamination* to *prompt sensitivity* and *long context*. This breadth reflects the current state of LLM limitation research: a fast-growing, methodologically diverse field still defining its major challenges. Additionally, as shown in Section D in the supplementary materials, many papers address multiple limitations simultaneously, reflecting the complexity of emerging concerns.

3. The distribution of limitations appears relatively stable in the ACL dataset, whereas the arXiv dataset shows a rise in concern for topics related to safety and controllability. This contrast is nuanced and is reflected in two key trends, discussed below.

3.1 Emerging trends in LLLMs research. Our trend analysis reveals mixed dynamics within LLLMs research over the studied time period. Safety and controllability concerns (e.g., *Security Risks*, *Alignment Limitations*, *Knowledge Editing*, *Hallucination*), model capacity advances (e.g., *Long Context* in ACL), and *Multimodality* generally rise over time. In contrast, topics like *Bias and Fairness* decline, while others remain flat. Notably, we observe a shift around 2023-Q2, following the release of models like ChatGPT. After this point, early fluctuations diminish, topics that had been growing continue at a steadier pace, and the decline in certain areas becomes more pronounced.

These trends align with shifting priorities in the LLM community. The growing attention to *alignment* and *security* reflects their increasingly central role in both training and evaluation of LLMs. Though still relatively new and unsettled [94], these concerns became prominent in 2022 with the rise of Reinforcement Learning from Human Feedback (RLHF), which is now foundational in the training pipelines of major models [109]. Yet ensuring safety without compromising performance remains an open challenge [85, 100], making this an active and fast-moving research area.

3.2. Limitations persist as applications expand. As LLMs are deployed in high-stakes domains, interest in *hallucination* and *knowledge editing* is growing due to the increasing demand for factual accuracy and controllability. These remain deeply challenging: hallucination is increasingly seen as an inherent property of LLMs, rooted in model architecture itself [8, 117].

Finally, these challenges grow as models move beyond text. The rise of *multimodality*-related limitations suggests that LLMs not only inherit existing issues but also encounter new ones with inputs like images and audio [121]. This trend likely reflects growing interest caused by the release of GPT-4V [1], LLaVA [61], and other vision-language models in mid-2023. This and other trends discussed above coincide with broader shifts already noted in previous studies of the arXiv corpus: from early 2023 to late 2024, top-cited LLM papers increasingly came from cs.CV and cs.LG, with cs.CL seeing a relative decline [49]. Similarly, authorship diversified, with many newcomers from computer vision, security, and software engineering [75].

4. Despite methodological differences, HDBSCAN and LlooM identify overlapping high-frequency topics (e.g., Reasoning, Hallucination, Security Risks) and similar trend patterns, pointing to the stability of the main findings. We validate our results by comparing HDBSCAN+BERTopic (single-topic, density-based) with LlooM (LLM-based, multi-topic). Despite methodological differences, both clustering approaches identify the high-frequency topics, most notably *Reasoning*, *Hallucination*, and *Security Risks*, showing substantial agreement in both topic composition and trend trajectories. And although smaller topics (e.g., *Prompt Sensitivity*) and trend significance can vary, the main trends reported in this study appear stable across methods.

Limitations and Future Work

While our analysis involves multiple datasets and clustering approaches, several methodological and temporal constraints should be kept in mind when interpreting the results:

- Although Llama-3.1-70B performs near human level in annotating limitation relevance, it still slightly lags, particularly in extracting supporting evidence, possibly leading to missed or incorrect information. However, as noted in Section 4.1, human annotators also overlooked some cases, suggesting that some level of imprecision is inherent to the task.
- Both of our clustering approaches are prone to some instability. LlooM can be variable due to its reliance on LLM outputs, a limitation noted by its authors as well [48]. HDBSCAN+BERTopic might also show some run-to-run variability, stemming from the stochastic nature of UMAP and sensitivity to embedding changes. While high-level patterns are generally stable, topic composition and temporal trends may shift slightly. We mitigate these issues by validating results across both methods.

- While we adopt a broad definition of LLMs in both automated and human annotation (including transformer-based, foundational, multimodal models), this scope may still introduce bias, e.g., we may not capture newly emerging terms in fields such as computer vision.
- Our trend analysis for the arXiv dataset includes data up to early 2025. Therefore, apparent declines or plateaus in the latest quarter should be interpreted with caution. We also exclude data prior to 2022, even though interest in limitations of smaller-scale LMs had already been rising since the introduction of models like BERT in 2018 [87, 125].

Future work could refine and extend these findings in several directions. First, the limitation topics could be decomposed into subcategories, such as types of reasoning or specific forms of bias, using hierarchical or agglomerative clustering techniques. Second, extending the analysis to earlier years, especially the period following BERT's introduction in 2018, could clarify how concerns raised for smaller-scale PLMs have shifted, declined, or re-emerged with model scaling.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788* (2024).

[3] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157* (2024).

[4] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245* (2023).

[5] Afra Feyza Akyürek, Eric Pan, Garry Kuwanto, and Derry Wijaya. 2023. Dune: Dataset for unified editing. *arXiv preprint arXiv:2311.16087* (2023).

[6] Mudassar Hassan Arsalan, Omar Mubin, Abdullah Al Mahmud, Imran Ahmed Khan, and Ali Jan Hassan. 2025. Mapping Data-Driven Research Impact Science: The Role of Machine Learning and Artificial Intelligence. In *Metrics*, Vol. 2. 5. doi:10.3390/metrics2020005

[7] Thomas Ball, Shuo Chen, and Cormac Herley. 2024. Can we count on llms? the fixed-effect fallacy and claims of gpt-4 capabilities. *arXiv preprint arXiv:2409.07638* (2024).

[8] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746* (2024).

[9] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45. doi:10.1145/3641289

[10] Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning language models to emit linguistic expressions of uncertainty. *arXiv preprint arXiv:2409.12180* (2024).

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[12] Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Jingkuan Song, and Lianli Gao. 2024. CoIN: A Benchmark of Continual Instruction Tuning for Multimodel Large Language Models. In *Proceedings of the NeurIPS*, Vol. 37. 57817–57840.

[13] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567* (2025).

[14] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769* (2024).

[15] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113. doi:10.5555/3648699.3648939

[17] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *Comput. Surveys* 57, 6 (2025), 1–39. doi:10.1145/3712001

[18] Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Gotcha! Don't trick me with unanswerable questions! Self-aligning Large Language Models for Responding to Unknown Questions. *arXiv e-prints* (2024), arXiv–2402.

[19] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* (2019).

[20] Jairo Diaz-Rodriguez. 2025. k-LLMmeans: Summaries as Centroids for Interpretable and Scalable LLM-Based Text Clustering. *arXiv preprint arXiv:2502.09667* (2025).

[21] Qinxu Ding, Ding Ding, Yue Wang, Chong Guan, and Bosheng Ding. 2024. Unraveling the landscape of large language models: a systematic review and future perspectives. *Journal of Electronic Business & Digital Economics* 3, 1 (2024), 3–19.

[22] Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. 2025. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. *arXiv preprint arXiv:2502.05151* (2025).

[23] Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087* (2023).

[24] Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, and Kam-Fai Wong. 2024. LLMEdgeRefine: Enhancing Text Clustering with LLM-Based Boundary Point Refinement. In *Proceedings of the EMNLP*. 18455–18462. doi:10.18653/v1/2024.emnlp-main.1025

[25] Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen, and Fei Xia. 2024. Does Data Contamination Detection Work (Well) for LLMs? A Survey and Evaluation on Detection Assumptions. *arXiv preprint arXiv:2410.18966* (2024).

[26] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, and et al. Wang, Ting. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523* (2024).

[27] Bady Gana, Andrés Leiva-Araos, Héctor Allende-Cid, and José García. 2024. Leveraging llms for efficient topic reviews. *Applied Sciences* 14, 17 (2024), 7675. doi:10.3390/app14177675

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[29] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[30] Reto Gubelmann and Siegfried Handschuh. 2022. Uncovering more shallow heuristics: Probing the natural language inference capacities of transformer-based pre-trained language models using syllogistic patterns. *arXiv preprint arXiv:2201.07614* (2022).

[31] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).

[32] Priyanka Gupta, Bosheng Ding, Chong Guan, and Ding Ding. 2024. Generative AI: A systematic review using topic modelling techniques. *Data and Information Management* 8, 2 (2024), 100066. doi:10.1016/j.dim.2024.100066

[33] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints* 1 (2023), 1–26. doi:10.36227/techrxiv.23589741.v8

[34] Fatemeh Haji, Mazal Bethany, Maryam Tabar, Jason Chiang, Anthony Rios, and Peyman Najafirad. 2024. Improving LLM Reasoning with Multi-Agent Tree-of-Thought Validator Agent. *arXiv preprint arXiv:2409.11527* (2024).

[35] Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, et al. 2025. PaSa: An LLM Agent for Comprehensive Academic Paper Search. *arXiv preprint arXiv:2501.10120* (2025).

[36] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* 33, 8 (2024), 1–79. doi:10.1145/3695988

[37] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

[38] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55. doi:10.1145/3703155

[39] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38. doi:10.1145/3571730

[40] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[41] Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. Trisum: Learning summarization ability from large language models with structured rationale. *arXiv preprint arXiv:2403.10351* (2024).

[42] Weipeng Jiang, Zhenting Wang, Juan Zhai, Shiqing Ma, Zhengyu Zhao, and Chao Shen. 2024. Unlocking adversarial suffix optimization without affirmative phrases: Efficient black-box jailbreaking via llm as optimizer. *arXiv preprint arXiv:2408.11313* (2024).

[43] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841* (2024).

[44] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 8622–8642. doi:10.1109/TKDE.2024.3469578

[45] Maurice George Kendall. 1948. Rank correlation methods. (1948).

[46] Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of Large Language Models in Clinical Problem-Solving Arising from Inflexible Reasoning. *arXiv preprint arXiv:2502.04381* (2025).

[47] Ashutosh Kumar, Shiv Vignesh Murthy, Sagarika Singh, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273* (2024).

[48] Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–28. doi:10.1145/3613904.3642830

[49] Christoph Leiter, Jonas Belouadi, Yanran Chen, Ran Zhang, Daniil Larionov, Aida Kostikova, and Steffen Eger. 2024. NLLG Quarterly arXiv Report 09/24: What are the most influential current AI Papers? *arXiv preprint arXiv:2412.12121* (2024).

[50] Andrés Leiva-Araos, Bady Gana, Héctor Allende-Cid, José García, and Manob Jyoti Saikia. 2025. Large scale summarization using ensemble prompts and in context learning approaches. *Scientific Reports* (2025). doi:10.1038/s41598-025-94551-8

[51] Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies* (2024), 1–27.

[52] Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. *arXiv preprint arXiv:2305.16572* (2023).

[53] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201* (2023).

[54] Yucheng Li, Frank Guerin, and Chenghua Lin. 2023. An open source data contamination report for large language models. *arXiv preprint arXiv:2310.17589* (2023).

[55] Zichao Li, Ines Arous, Siva Reddy, and Jackie CK Cheung. 2023. Evaluating dependencies in fact editing for language models: Specificity and implication awareness. *arXiv preprint arXiv:2312.01858* (2023).

[56] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. In *Proceedings of the EMNLP: Industry Track*. 881–893.

[57] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419* (2025).

[58] Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, et al. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776* (2025).

[59] Inna Wanyin Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. *arXiv preprint arXiv:2210.15144* (2022).

[60] Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Scott Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. In *Proceedings of the NeurIPS*, Vol. 37. 115588–115614.
[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
[62] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).
[63] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* (2023).
[64] Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. *arXiv preprint arXiv:2501.04306* (2025).
[65] Qitan Lv, Jie Wang, Hanzhu Chen, Bin Li, Yongdong Zhang, and Feng Wu. 2024. Coarse-to-fine highlighting: Reducing knowledge hallucination in large language models. *arXiv preprint arXiv:2410.15116* (2024).
[66] Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621* (2022).
[67] Henry B Mann. 1945. Nonparametric tests against trend. *Econometrica: Journal of the econometric society* (1945), 245–259.
[68] Matej Martinc, Blaž Škrlj, and Senja Pollak. 2022. TNT-KID: Transformer-based neural tagger for keyword identification. *Natural Language Engineering* 28, 4 (2022), 409–448.
[69] Andrea Matarazzo and Riccardo Torlone. 2025. A Survey on Large Language Models with some Insights on their Capabilities and Limitations. *arXiv preprint arXiv:2501.04040* (2025).
[70] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.
[71] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
[72] Jordan Meadows, Tamsin James, and Andre Freitas. 2024. Exploring the Limits of Fine-grained LLM-based Physics Inference via Premise Removal Interventions. *arXiv preprint arXiv:2404.18384* (2024).
[73] Dheeraj Mekala, Jason Wolfe, and Subhro Roy. 2022. Zerotop: Zero-shot task-oriented semantic parsing using large language models. *arXiv preprint arXiv:2212.10815* (2022).
[74] Amirkeivan Mohtashami, Florian Hartmann, Sian Gooding, Lukas Zilka, Matt Sharifi, et al. 2023. Social learning: Towards collaborative learning with large language models. *arXiv preprint arXiv:2312.11441* (2023).
[75] Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2023. Topics, authors, and institutions in Large Language Model research: trends from 17K arXiv papers. *arXiv preprint arXiv:2307.10700* (2023).
[76] Riya Naik, Ashwin Srinivasan, Estrid He, and Swati Agarwal. 2025. An Empirical Study of the Role of Incompleteness and Ambiguity in Interactions with Large Language Models. *arXiv preprint arXiv:2503.17936* (2025).
[77] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (2023).
[78] Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E Priebe, and Eric Horvitz. 2024. Medfuzz: Exploring the robustness of large language models in medical question answering. *arXiv preprint arXiv:2406.06573* (2024).
[79] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the NeurIPS*. 27730–27744.
[80] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100
[81] Brendan Park, Madeline Janecek, Naser Ezzati-Jivan, Yifeng Li, and Ali Emami. 2024. Picturing Ambiguity: A Visual Twist on the Winograd Schema Challenge. *arXiv preprint arXiv:2405.16277* (2024).
[82] Anup Pattnaik, Cijo George, Rishabh Tripathi, Sasanka Vutla, and Jithendra Vepa. 2024. Improving Hierarchical Text Clustering with LLM-guided Multi-view Cluster Representation. In *Proceedings of the EMNLP: Industry Track*. 719–727. doi:10.18653/v1/2024.emnlp-industry.54
[83] Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *arXiv preprint arXiv:2408.13808* (2024).
[84] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511* (2024).

[85] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693* (2023).

[86] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[87] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *TACL* 8 (2021), 842–866. doi:10.1162/tacl_a_00349

[88] Domenic Rosati, Giles Edkins, Harsh Raj, David Atanasov, Subhabrata Majumdar, Janarthanan Rajendran, Frank Rudzicz, and Hassan Sajjad. 2024. Defending against reverse preference attacks is difficult. *arXiv preprint arXiv:2409.12914* (2024).

[89] David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *14th International Workshop on Semantic Evaluation*. 187–193.

[90] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. [n. d.]. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the EMNLP*. 11709–11724. doi:10.18653/v1/2024.findings-emnlp.685

[91] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:2405.09589* (2024).

[92] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061* (2022).

[93] Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655* (2024).

[94] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025* (2023).

[95] Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. Red teaming language model detectors with language models. *TACL* 12 (2024), 174–189.

[96] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, and et al. Garriga-Alonso, Adrià. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[97] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of Thoughtlessness? An Analysis of CoT in Planning. In *Proceedings of the NeurIPS*.

[98] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* (2024).

[99] Salmonn Talebi, Elizabeth Tong, and Mohammad RK Mofrad. 2023. Beyond the hype: Assessing the performance, trustworthiness, and clinical suitability of gpt3. 5. *arXiv preprint arXiv:2306.15887* (2023).

[100] Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. 2025. Equilibrate RLHF: Towards Balancing Helpfulness-Safety Trade-off in Large Language Models. *arXiv preprint arXiv:2502.11555* (2025).

[101] Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. 2024. Limits of transformer language models on learning to compose algorithms. *Advances in Neural Information Processing Systems* 37 (2024), 7631–7674.

[102] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313* (2024).

[103] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[104] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine* 30, 4 (2024), 1134–1142. doi:10.1038/s41591-024-02855-5

[105] Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *TACL* 12 (2024), 321–333.

[106] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345. doi:10.1007/s11704-024-40231-1

[107] Xiaoye Wang, Nicole Xi Zhang, Hongyu He, Trang Nguyen, Kun-Hsing Yu, Hao Deng, Cynthia Brandt, Danielle S Bitterman, Ling Pan, Ching-Yu Cheng, et al. 2024. Safety challenges of AI in medicine. *arXiv preprint arXiv:2409.18968* (2024).

[108] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. 2024. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems* 37 (2024), 115119–115145.

[109] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216* (2024).

[110] Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023. What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics. In *Proceedings of ACL (Volume 1: Long Papers)*. 6790–6809. doi:10.18653/v1/2023.acl-long.375

[111] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models.

[112] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Boman, He He, and Shi Feng. 2024. Language Models Learn to Mislead Humans via RLHF. *arXiv preprint arXiv:2409.12822* (2024).

[113] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. 2023. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909* (2023).

[114] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60. doi:10.1007/s11280-024-01291-2

[115] Xiaoxia Wu, Haojun Xia, Stephen Youn, Zhen Zheng, Shiyang Chen, Arash Bakhtiari, Michael Wyatt, Reza Yazdani Aminabadi, Yuxiong He, Olatunji Ruwase, et al. 2023. Zeroquant (4+ 2): Redefining llms quantization with a new fp6-centric strategy for diverse generative tasks. *arXiv preprint arXiv:2312.08583* (2023).

[116] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686* (2025).

[117] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).

[118] Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. 2024. Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. *arXiv preprint arXiv:2403.09963* (2024).

[119] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211. doi:10.1016/j.hcc.2024.100211

[120] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. *arXiv preprint arXiv:2303.15445* (2023).

[121] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).

[122] Hanlin Zhang, Yi-Fan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. 2023. A study on the calibration of in-context learning. *arXiv preprint arXiv:2312.04021* (2023).

[123] Longteng Zhang, Xiang Liu, Zeyu Li, Xinglin Pan, Peijie Dong, Ruibo Fan, Rui Guo, Xin Wang, Qiong Luo, Shaohuai Shi, et al. 2023. Dissecting the runtime performance of the training, fine-tuning, and inference of large language models. *arXiv preprint arXiv:2311.03687* (2023).

[124] Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *arXiv preprint arXiv:2305.14871* (2023).

[125] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[126] Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795* (2023).

[127] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).

Supplementary Material

A Keyword List

The final keyword list used to filter LLM-focused papers is provided in Table 12.

Table 12. Final list of keywords sorted by N-grams

cot	*gpt*	*api*
rag	*judge*	*chat*
llms	*dpo*	*mllms*
llm	*lora*	*hallucination*
jailbreak	*speculative*	*self consistency*
agent	*cot prompting*	*llm agents*
model editing	*self correction*	*prompting*
self reflection	*function calling*	*language agents*
hallucination detection	*preference learning*	*long context*
language models	*data contamination*	*injection attacks*
instruction tuned	*prompt engineering*	*jailbreak attack*
preference alignment	*knowledge editing*	*text watermarking*
prompt optimization	*self evaluation*	*instruction tuning*
tree of thoughts	*evaluating llms*	*multi agent framework*
benchmarking llms	*retrieval augmented generation*	*direct preference optimization*
multimodal llms	*commonsense reasoning*	*chain of thought reasoning*
chain of thought prompting	*multi agent collaboration*	*augmented llms*
generated text detection	*jailbreaking attacks*	*jailbreak attacks*
prompting techniques		

Evaluate the following paper's title and abstract to determine if it discusses Large Language Models (LLMs) and whether it addresses their limitations. If it does, indicate specific evidence from the abstract or title that points to the limitations and identify the domain of the limitation where possible. Note that LMs and LLMs include pre-trained transformer-based models, foundational models, and multimodal models. Include all kinds of language models but exclude other, more general models.

Rate the depth of LLM limitations discussion (0-5) using the following scale:

0: No discussion of LLMs at all.
1: Discusses LLMs but does not mention any limitations of LLMs.
2: The abstract mentions one or more limitations of LLMs in passing or as a minor detail. The limitations are not explained, elaborated, or analyzed further and are primarily used to justify the paper's goals, methods, or contributions.
3: The abstract discusses one or two limitations of LLMs in moderate detail. These limitations are important but are not the primary focus of the abstract. The discussion provides some analysis, examples, or implications, but the abstract emphasizes the solution, methodology, or results more than the limitations.
4: The abstract dedicates significant attention to one or more limitations of LLMs, making them a major focus. The limitations are described in detail, with examples, analysis, or experimental evidence. While other aspects (e.g., solutions or results) may be discussed, the limitations play an equally or more important role in the narrative.
5: Entirely focused on LLM limitations and challenges.

Please look at the following examples alongside the explanations on why decided the respective ratings and rate the other abstracts from 0 to 5 accordingly by following the same logic as below:

Example 0:
Title: Simultaneous Selection and Adaptation of Source Data via Four-Level Optimization
Abstract: "In many NLP applications, to mitigate data deficiency in a target task, source data is collected... [...]"
Does it talk about LLMs: No
Rate Limitations of LLMs: 0
Evidence: No evidence of discussion of limitations of LLMs.
Output Explanation: [...] this paper does not talk about LLMs or any language model at all.

Example 1:
Title: SPAE Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs
Abstract: "In this work, we introduce Semantic Pyramid AutoEncoder (SPAE) for enabling frozen LLMs to perform both understanding and generation tasks. [...]"
Does it talk about LLMs: Yes
Rate Limitations of LLMs: 1
Evidence: No evidence of discussion of limitations of LLMs.
Output Explanation: [...] even though it talks about LLMs, it does not mention any explicit limitation of the models.

Example 2:
Title: Large Language Models for Conducting Advanced Text Analytics Information Systems Research
Abstract: "[...] Large Language Models (LLMs) have emerged... [...]. We also outline potential challenges and limitations in adopting LLMs for IS."
Does it talk about LLMs: Yes
Rate Limitations of LLMs: 2
Evidence: "We also outline potential challenges and limitations in adopting LLMs for IS."
Output Explanation: This abstract mentions just briefly one limitation of the Large Language Models without going into detail and focuses on other topics.

Example 3:
Title: Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning
Abstract: "In recent years, Large Language Models (LLMs) have shown remarkable performance... However, adapting these models to incorporate new, out-of-domain knowledge remains a challenge... [...]"
Does it talk about LLMs: Yes
Rate Limitations of LLMs: 3
Evidence: "However, adapting these models to incorporate new, out-of-domain knowledge remains a challenge [...]"; "[...] token-based scaling... may not provide uniform coverage of new knowledge."
Output Explanation: The limitation (difficulty in adapting to new knowledge) is mentioned but not explored in depth (e.g., why LLMs struggle with knowledge injection, etc.). The primary focus of the paper is on the proposed solution.

Example Output 4:
Title: Red Teaming Language Model Detectors with Language Models
Abstract: "The prevalence and strong capability of large language models (LLMs) present significant safety and ethical risks... [...]"
Does it talk about LLMs: Yes
Rate Limitations of LLMs: 4
Evidence: "The prevalence and strong capability of large language models (LLMs) present significant safety and ethical risks if exploited by malicious users."; "[...] our attacks effectively compromise the performance of all detectors in the study with plausible generations [...]."
Output Explanation: The paper extensively analyzes vulnerabilities in LLM detection systems under adversarial attacks. Limitation is a major focus, but mixed with technical contributions and experiments.

Example Output 5:
Title: LLMs for Relational Reasoning: How Far are We?
Abstract: "[...] there has been a surge of interest in investigating the reasoning ability of the LLMs. [...] it is hard to conclude that the LLMs possess strong reasoning ability by merely achieving positive results on these benchmarks... [...]"
Does it talk about LLMs: Yes
Rate Limitations of LLMs: 5
Evidence: "[...] it is hard to conclude that the LLMs possess strong reasoning ability. [...] Recent efforts have demonstrated that the LLMs are poor at solving sequential decision-making problems..."; "[...] the state-of-the-art LLMs are much poorer in terms of reasoning ability by achieving much lower performance and generalization [...]."
Output Explanation: The paper's primary focus is a detailed exploration of LLMs' reasoning limitations, rather than just using these limitations to motivate a solution or another contribution.

Please answer in the following format by providing the rating and a brief evidence for each abstract. Please do not give the respective Explanations, only the evidence found in the abstract.

Does it talk about LLMs: [Yes/No]
Rate Limitations of LLMs: [0-5]
Evidence: [the evidence text in the abstract or title].

Title: {title}
Abstract: {abstract}

Fig. 14. Prompt 3 used for LLM-based classification and evidence extraction (as discussed in 4.4.3). Abstract texts and evidence are shortened for brevity).

Table 13. Representative examples from the human rating annotation task, with brief explanation for abstracts' ratings on the 0–5 scale (defined in Table 2). In higher-rated papers (3–5), evidence of LLLMs is highlighted in bold.

Rating	Abstract	Explanation
0	"Multi-query attention (MQA), which only uses a single key-value head, drastically speeds up decoder inference. However, MQA can lead to quality degradation. [...]" [4], EMNLP 2023	No mention of LLMs or their limitations.
1	"We introduce the framework of 'social learning' in the context of LLMs, whereby models share knowledge with each other in a privacy-aware manner using natural language. [...]" [74], arXiv, 2023	Mentions LLMs but does not discuss any limitations.
2	"[...] **LLMs... often display a considerable level of overconfidence even when the question does not have a definitive answer.** [...] we propose a novel and scalable self-alignment method to utilize the LLM itself to enhance its response-ability to different types of unknown questions. [...]" [18], EMNLP 2024	Briefly identifies overconfidence as a limitation; used as motivation for the method.
3	"[...] **it remains difficult to distinguish effects of statistical correlation from more systematic logical reasoning grounded on the understanding of real world [in PLMs]**. [...] We find that models are consistently able to override real-world knowledge in counterfactual scenarios... however, we also find that **for most models this effect appears largely to be driven by simple lexical cues**. [...]" [52], ACL 2023	Discusses reasoning limitations in moderate detail, but not as the main focus.
4	"[...] we aim to assess the performance of OpenAI's newest model, GPT-4V(ision), specifically in the realm of multimodal medical diagnosis. [...] **While GPT-4V demonstrates proficiency in distinguishing between medical image modalities and anatomy, it faces significant challenges in disease diagnosis and generating comprehensive reports.** [...] while large multimodal models have made significant advancements in computer vision and NLP, **it remains far from being used to effectively support real-world medical applications and clinical decision-making.** [...]" [113], arXiv, 2023	Focuses extensively on GPT-4V's limitations in medical applications, but also explores strengths of the model.
5	"We analyze the capabilities of Transformer LMs on learning discrete algorithms. [...] **We observe that the compositional capabilities of state-of-the-art Transformer language models are very limited and sample-wise scale worse than relearning all sub-tasks for a new algorithmic composition. We also present a theorem in complexity theory, showing that gradient descent on memorizing feedforward models can be exponentially data inefficient.**" [101], arXiv, 2024	Entirely focused on identifying and analyzing LLM limitations in algorithmic learning.

B Clustering Details

For HDBSCAN+BERTopic, we perform a grid search over UMAP, HDBSCAN and BERTopic to produce meaningful clusters while minimizing the outliers. The final configurations are as follows:

- for ACL, UMAP is set to `n_neighbors` = 25, `n_components` = 10, `min_dist` = `0.0`; HDBSCAN to `min_cluster_size` = 25, `min_samples` = 10, and BERTopic to `min_topic_size` = 10.
- For ArXiv, UMAP is adjusted to `n_neighbors` = 15, `n_components` = 5, `min_dist` = `0.05`, while HDBSCAN and BERTopic are set to `min_cluster_size` = 40 and `min_topic_size` = 25.

Our initial runs revealed that many outliers were not true noise but rather positioned between clusters, causing HDBSCAN (used within BERTopic) to misclassify them. To address this, we implement a distance-based outlier reassignment strategy, where outliers identified by HDBSCAN and BERTopic are reassigned to their most probable topic based on the topic probability distribution, but only if they fall within a certain distance threshold to ensure proximity to an existing cluster.

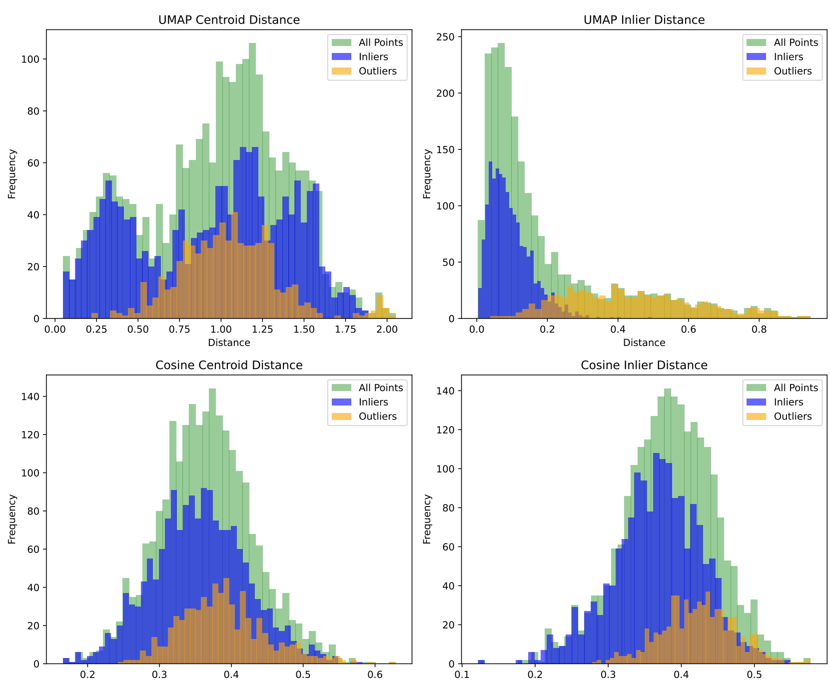


Fig. 15. Histograms of distance metrics used for outlier detection in ACL dataset clustering. Each subplot visualizes distances between a paper's embedding (based on evidence and keyphrases) and either a cluster centroid or nearest inlier, in both UMAP-reduced and original embedding spaces. These metrics help determine the threshold for reassigning outliers to the closest cluster.

To determine this threshold, we test several distance metrics, namely 1) the Euclidean distance to the closest cluster centroid in UMAP space, 2) the Euclidean distance to the closest inlier in UMAP space, 3) the cosine distance to the closest cluster centroid in the original embedding space, and 4) the cosine distance to the closest inlier in the original embedding space. As shown in Figure 15, the UMAP Inlier Distance provides the clearest separation between inliers and outliers, such that we selected it for subsequent analysis. We set the outlier threshold at the dip between the two peaks, using a value of 0.3.

Table 14. Prompt descriptions for each cluster in the ACL dataset identified by LlooM.

Cluster	Prompt
Reasoning	Does the text example highlight limitations of large language models in performing reasoning or inference tasks, such as multi-hop reasoning or deductive logic?
Generalization	Does the text example highlight limitations in a model's ability to understand context or generalize effectively across different domains and scenarios?
Knowledge Editing	Does the text discuss limitations in storing, encoding, or updating knowledge in large language models, including factual inaccuracies or difficulties in modifying learned information?
Hallucination	Does the text describe instances where language models generate factually incorrect information, describe non-existent content, or produce contextually unfaithful data?
Language and Cultural Limitations	Does the text example highlight limitations of large language models in handling multilingual tasks, language-specific or cross-cultural issues?
Bias and Fairness	Does the text address issues related to social biases or stereotyping within language models, including challenges in presence, amplification, and mitigation thereof?
Security Risks	Does this text example address the susceptibility of language models to adversarial attacks, including backdoor or jailbreak methods?
Multimodality	Does the text example address the struggles of large language models or multimodal models in integrating and reasoning across different modalities?
Long Context	Does the text example discuss challenges faced by large language models in handling long contexts or maintaining coherence over extended inputs?
Privacy Risks	Does the text example discuss privacy risks associated with language models, such as data leakage or memorization of sensitive information?
Computational Cost	Does the text example highlight the high computational cost or resource demands associated with using large language models?
Catastrophic Forgetting	Does the example discuss problems caused by models forgetting previously learned information during fine-tuning or continual learning?
Data Contamination	Does the text example discuss issues of training data containing test data, causing overestimation of model performance through memorization?

Table 15. Prompt descriptions for each cluster in the arXiv dataset identified by LlooM.

Cluster	**Prompt**
Trustworthiness	Does the text refer to concerns about the reliability or trustworthiness of outputs generated by language models?
Reasoning	Does the text example highlight limitations of large language models in performing reasoning or inference tasks, such as multi-hop reasoning or deductive logic?
Generalization	Does the text discuss the inability of language models to generalize across different tasks or inputs?
Long Context	Does this paper explore challenges faced by large language models (LLMs) in handling long context lengths, such as limitations in memory, performance, or understanding over extended inputs?
Bias and Fairness	Does the text describe how LLMs propagate biases, stereotypes, or misinformation, leading to potential societal harms?
Hallucination	Does the text mention inaccuracies or fabricated content in the outputs of large language models (also multimodal)?
Alignment Limitations	Does the text highlight any limitations or challenges in aligning large language models with human values or safety protocols?
Security Risks	Does the text address the security risks associated with adversarial attacks or exploits that can manipulate large language models?
Prompt Sensitivity	Does the text highlight how slight variations in prompts affect the performance of language models?
Multimodality	Does the text discuss difficulties faced by LLMs in effectively integrating or coordinating information across multiple modalities, such as text and images?
Language and Cultural Limitations	Does this text highlight struggles of Language Models (LLMs) in effectively handling low-resource languages?
Privacy Risks	Does the text discuss vulnerabilities related to the potential leakage of private or sensitive information by language models?
Knowledge Editing	Does the text discuss challenges in knowledge editing, such as knowledge distortion, struggles with updating specific knowledge types, or difficulty integrating new knowledge while maintaining coherence?
Overconfidence	Does the text describe challenges related to calibration or overconfidence in language models?
Catastrophic Forgetting	Does this text describe Large Language Models (LLMs) losing previously acquired knowledge when learning new information?

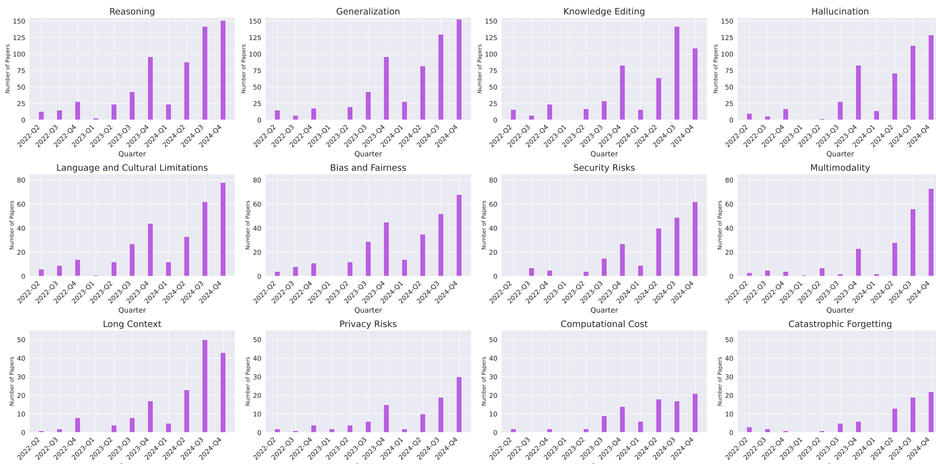
Table 16. Relative percentage growth of LLM-normalized shares for limitation topics across ACL and arXiv datasets (LlooM clustering). Values are based on relative change between consecutive years.

(a) ACL Dataset

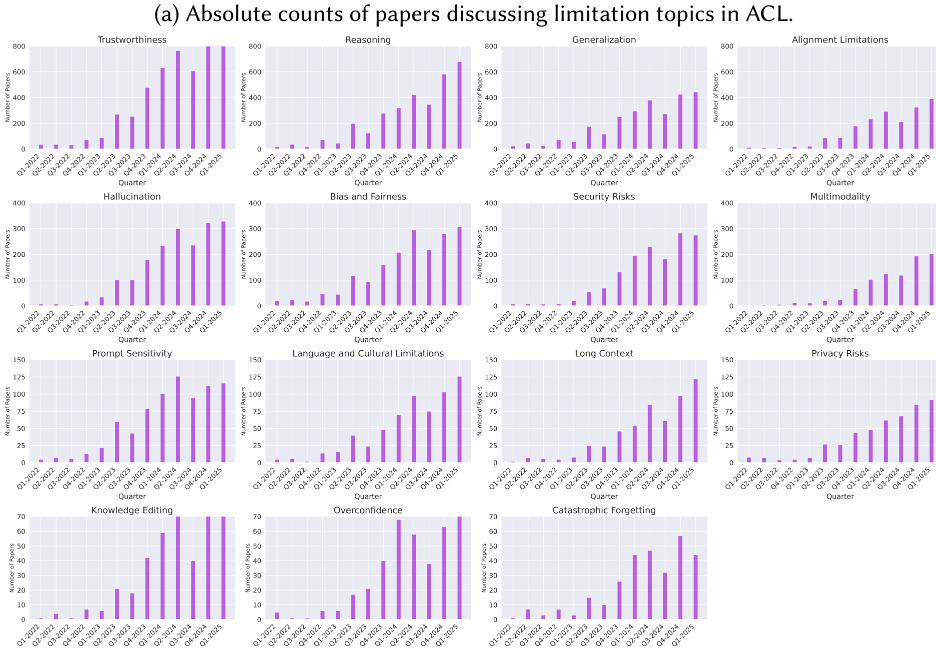
Topic	**Topic Shares (2022 → 2024)**	**→2023 (%)**	**→2024 (%)**
Reasoning	5.5 → 7.5 → 8.9	36.36%	18.67%
Generalization	3.9 → 7.2 → 8.6	84.62%	19.44%
Knowledge Editing	4.6 → 5.8 → 7.3	26.09%	25.86%
Hallucination	3.2 → 5.2 → 7.2	62.50%	38.46%
Language & Cultural Lim.	2.8 → 3.8 → 4.1	35.71%	7.89%
Bias and Fairness	2.3 → 3.9 → 3.7	69.57%	-5.13%
Security Risks	1.2 → 2.1 → 3.5	75.00%	66.67%
Multimodality	1.2 → 1.5 → 3.5	25.00%	133.33%
Long Context	1.1 → 1.3 → 2.7	18.18%	107.69%
Computational Cost	0.6 → 1.1 → 1.4	83.33%	27.27%
Privacy Risks	0.7 → 1.2 → 1.3	71.43%	8.33%
Catastrophic Forgetting	0.6 → 0.5 → 1.2	-16.67%	140.00%
Data Contamination	0.4 → 0.4 → 0.9	0.00%	125.00%

(b) arXiv Dataset

Topic	**Topic Shares (2022 → 2025)**	**→2023 (%)**	**→2024 (%)**	**→2025 (%)**
Trustworthiness	6.12 → 10.85 → 11.99 → 12.18	77.29%	10.51%	1.58%
Reasoning	5.15 → 6.43 → 7.09 → 9.09	24.85%	10.26%	28.21%
Generalization	5.88 → 5.99 → 5.83 → 5.95	1.87%	-2.67%	2.06%
Hallucination	1.28 → 4.14 → 4.64 → 4.40	223.44%	12.08%	-5.17%
Alignment Limitations	1.87 → 3.78 → 4.51 → 5.23	102.14%	19.31%	15.96%
Bias and Fairness	3.73 → 4.14 → 4.25 → 4.12	10.99%	2.66%	-3.06%
Security Risks	1.04 → 2.74 → 3.78 → 3.68	163.46%	37.96%	-2.65%
Multimodality	0.86 → 1.19 → 2.29 → 2.70	38.37%	92.44%	17.90%
Prompt Sensitivity	1.07 → 2.03 → 1.84 → 1.55	89.72%	-9.36%	-15.76%
Language & Cultural Lim.	0.93 → 1.27 → 1.46 → 1.68	36.56%	14.96%	15.07%
Long Context	0.69 → 1.02 → 1.26 → 1.63	47.83%	23.53%	29.37%
Privacy Risks	0.83 → 1.03 → 1.11 → 1.23	24.10%	7.77%	10.81%
Knowledge Editing	0.45 → 0.86 → 1.06 → 1.12	91.11%	23.26%	5.66%
Overconfidence	0.45 → 0.83 → 0.96 → 0.96	84.44%	15.66%	0.00%
Catastrophic Forgetting	0.62 → 0.54 → 0.76 → 0.59	-12.90%	40.74%	-22.37%



(a) Absolute counts of papers discussing limitation topics in ACL.



(b) Absolute counts of papers discussing limitation topics in arXiv.

Fig. 16. Absolute counts of papers discussing limitation topics in ACL and arXiv, as identified by LlooM clustering approach.

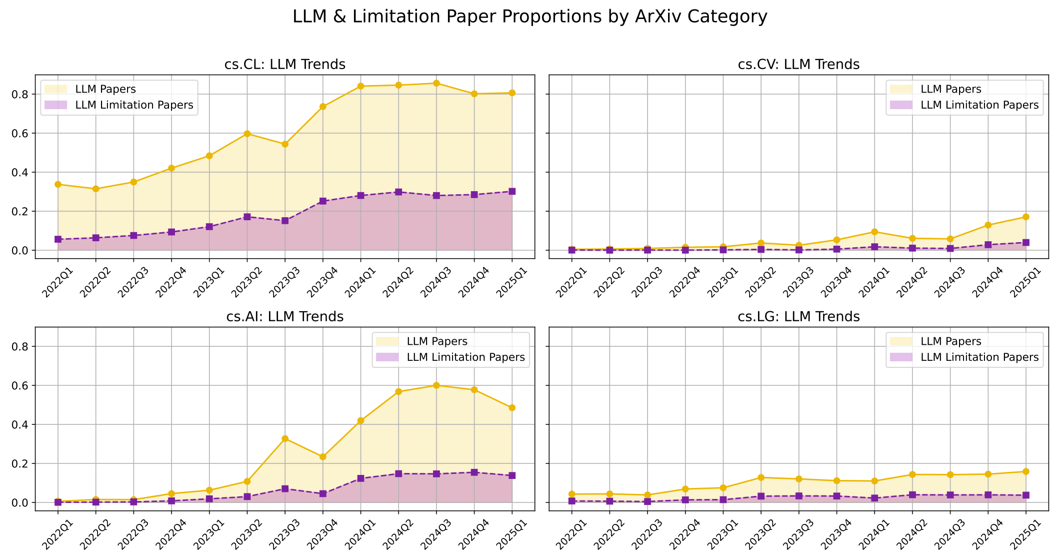


Fig. 17. Trends in LLM and LLM limitation research in arXiv over time, relative to all crawled papers, based on the main category only and broken down by category.

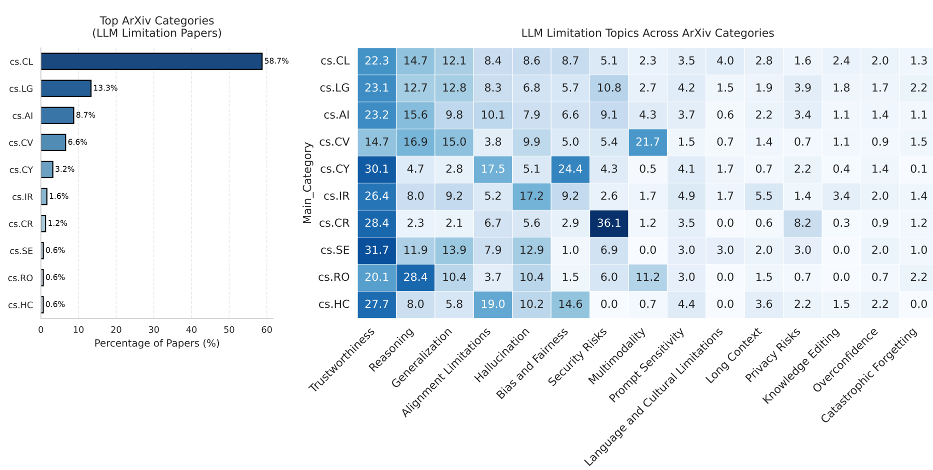


Fig. 18. Top nine arXiv (main) categories discussing LLLMs (left), and distribution of limitation topics within each category (right), shown as the percentage of papers in each category that mention a given limitation.

Table 17. Best-aligned LlooM topics for each HDBSCAN cluster based on Jaccard overlap of paper sets.

(a) ACL Dataset

HDBSCAN Topic	Best-Matching LlooM Topic	Jaccard Overlap
Social Bias	Bias and Fairness	0.498
Security	Security Risks	0.495
Hallucination	Hallucination	0.429
Reasoning	Reasoning	0.359
Long Context	Long Context	0.172
Generalization	Generalization	0.155
Uncertainty	Knowledge Editing	0.085
	Average Top Jaccard Overlap	**0.313**

(b) arXiv Dataset

HDBSCAN Topic	Best-Matching LlooM Topic	Jaccard Overlap
Security Risks	Security Risks	0.565
Social Bias	Bias and Fairness	0.288
Context & Memory Limitations	Long Context	0.150
Benchmark Contamination	Privacy Risks	0.011
Hallucination	Hallucination	0.477
Code Generation	Generalization	0.034
Reasoning	Reasoning	0.302
Conversational Limitations	Reasoning	0.070
Multimodality	Multimodality	0.379
Multilinguality	Language and Cultural Limitations	0.393
Healthcare Application	Alignment Limitations	0.027
Computational Cost	Privacy Risks	0.016
Knowledge Editing	Overconfidence	0.005
Quantization	Knowledge Editing	0.090
	Average Top Jaccard Overlap	**0.201**

Table 18. Best-aligned HDBSCAN topics for each LlooM topic based on Jaccard overlap of paper sets.

(a) ACL Dataset

LlooM Topic	**Best-Matching HDBSCAN Topic**	**Jaccard Overlap**
Generalization	Reasoning	0.194
Long Context	Long Context	0.172
Hallucination	Hallucination	0.429
Language and Cultural Limitations	Social Bias	0.204
Reasoning	Reasoning	0.359
Bias and Fairness	Social Bias	0.498
Security Risks	Security	0.495
Knowledge Editing	Reasoning	0.120
Catastrophic Forgetting	Generalization	0.155
Computational Cost	Generalization	0.114
Privacy Risks	Security	0.191
Multimodality	Reasoning	0.127
Data Contamination	Reasoning	0.042
	Average Top Jaccard Overlap	**0.239**

(b) arXiv Dataset

LlooM Topic	**Best-Matching HDBSCAN Topic**	**Jaccard Overlap**
Multimodality	Multimodality	0.379
Alignment Limitations	Hallucination	0.160
Hallucination	Hallucination	0.477
Trustworthiness	Security Risks	0.263
Generalization	Reasoning	0.116
Language and Cultural Limitations	Multilinguality	0.393
Bias and Fairness	Social Bias	0.288
Privacy Risks	Security Risks	0.152
Prompt Sensitivity	Hallucination	0.111
Reasoning	Reasoning	0.302
Knowledge Editing	Knowledge Editing	0.090
Overconfidence	Hallucination	0.081
Security Risks	Security Risks	0.565
Long Context	Context & Memory Limitations	0.150
Catastrophic Forgetting	Context & Memory Limitations	0.125
	Average Top Jaccard Overlap	**0.244**

C HDBSCAN + BERTopic Trend Analysis

In this section, we discuss LLLMs topic dynamics over time as identified by HDBSCAN + BERTopic. We apply the following perspectives:

(i) **LLM-wide share**, measured annually as $\frac{N_{t,y}^{\text{lim}}}{N_{y}^{\text{LLM}}}$, to reflect how often limitation topic t appears in LLM research in year y, relative to the total LLM papers. This shows whether a topic is gaining attention beyond LLLMs research and becoming part of the general LLM research agenda.

(ii) **Limitations share**, measured quarterly as $\frac{N_{t,q}^{\text{lim}}}{N_{q}^{\text{lim}}}$, to reflect the share of limitation-focused papers in quarter q that address topic t. This shows the topic's visibility within LLLMs subfield.

Here, $N_{t,y}^{\text{lim}}$ is the number of limitation papers on topic t in year y; N_{y}^{LLM} is the total number of LLM papers in that year; and $N_{t,q}^{\text{lim}}$, N_{q}^{lim} are the number of limitation papers on topic t and the total number of limitation papers, respectively, in quarter q.

(i) How are limitation topics represented in the broader growth of LLM research?

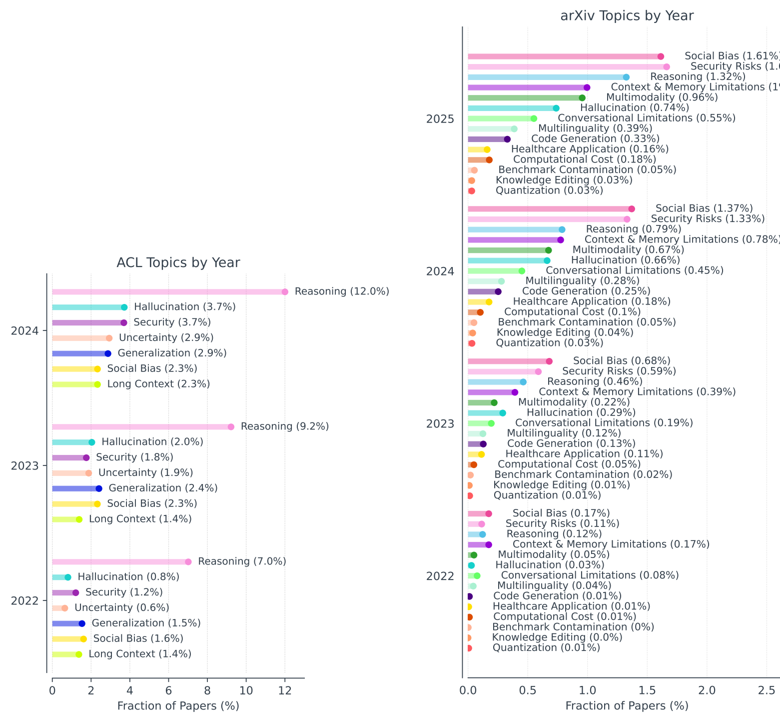
Key Insights

- Limitation topics show growing visibility within LLM research across both ACL and arXiv datasets. Some concerns, such as *Security* and *Long Context* in ACL, surge specifically in 2024, while others like *Reasoning* and *Hallucination* rise more steadily.
- On arXiv, the dramatic relative increases seen in 2023 (e.g., *Hallucination*, *Code Generation*, *Healthcare Application*) reflect the emergence of new concerns from a low baseline, but many of these remain small in absolute share.

Figure 19 shows the annual distribution of LLM limitation topics as a proportion of all LLM-focused papers in ACL and arXiv. Across both ACL and arXiv datasets, most limitation topics show grow in visibility across LLM research from 2022 to 2024 in ACL and 2022 to early 2025 in arXiv. However, as shown in Table 19 in the supplementary material, they do so at different pace. On ACL, some concerns surged sharply in 2024, such as *Security* (+106%) and *Long Context* (+64%), while others like *Reasoning* and *Hallucination* grew more steadily across years. In contrast, topics such as *Generalization* and *Social Bias* showed growth in 2023 but flattened or declined by 2024.

On arXiv, most topics were marginal in 2022, leading to dramatic relative growth in 2023 as new concerns entered the field: for example, *Hallucination* (+867%), *Code Generation* (+1200%), and *Healthcare Application* (+1000%). Despite steep increases, many remained low in absolute share (e.g., *Code Generation* at 0.13% in 2023). After this emergence, growth remained strong though less extreme, especially for *Reasoning*, *Security Risks*, *Multimodality*, and *Computational Cost*. A few topics declined by 2025, including *Benchmark Contamination*, *Quantization*, and *Knowledge Editing*. However, these topics are the least represented, and since 2025 data covers only one quarter, recent drops should be interpreted with caution.

These comparisons reflect annual, macro-level trends in topical focus and relative prominence. To examine shorter-term dynamics and account for changes in overall LLM publication volume, measure the percentage of papers on each topic, relative to all LLM limitation papers, for each quarter. Additionally, we apply the Mann-Kendall trend test [67] to identify whether each topic's share exhibits a consistent upward or downward trend across quarters. This non-parametric test is well-suited for detecting monotonic trends in time series without assuming linearity or normality.



(a) ACL Topic Distribution Per Year

(b) arXiv Topic Distribution Per Year

Fig. 19. Distribution of LLM limitation topics over years for ACL and arXiv, based on clustering results with HDBSCAN + BERT. Percentages reflect each topic's proportion out of the total LLM-focused papers (8,635 in ACL and 41,991 in arXiv).

However, non-monotonic changes (peaking behavior or decreases followed by increases) are not detected by this test, such that some observations below will be associated with high p-values.

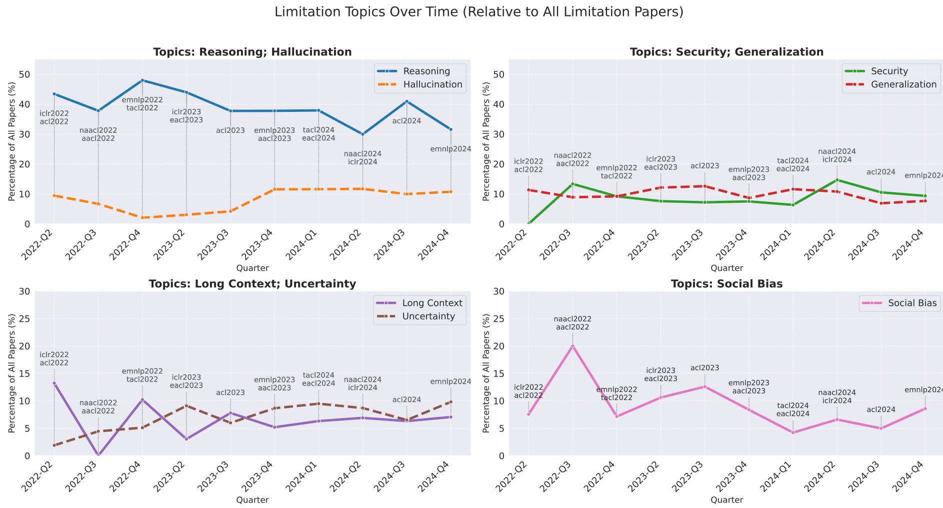
(ii) What are the internal trends within LLM limitation research?

Key Insights

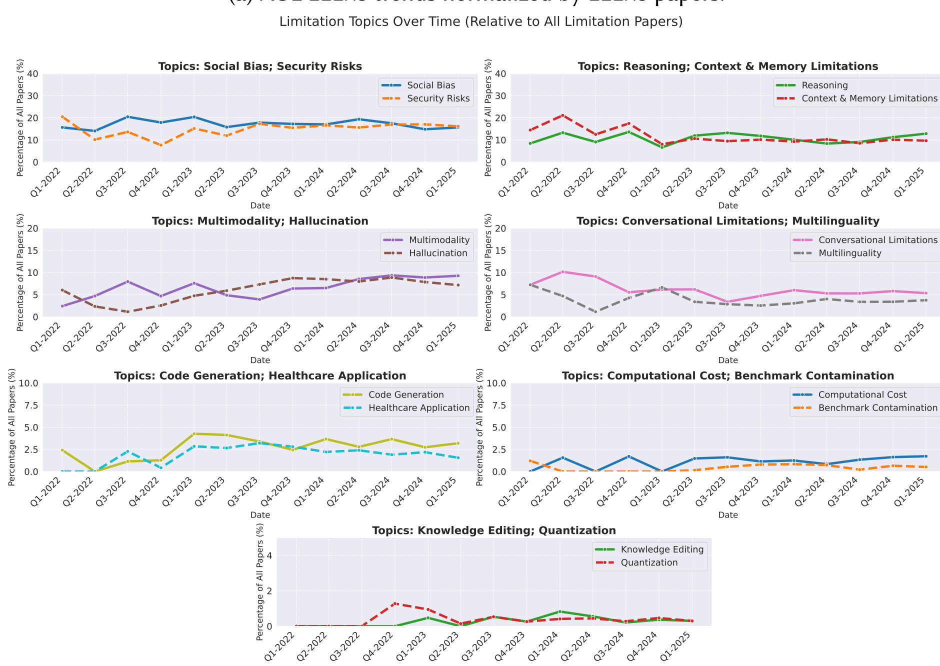
- On ACL, *Uncertainty* is the only topic with a statistically significant upward trend, while *Social Bias* and *Reasoning* show non-significant declines after early peaks. Most other topics remain relatively stable over time.
- On arXiv, *Multimodality* and *Hallucination* show significant growth, with the former likely linked to the rise of vision-language models. *Knowledge Editing* also increases, though not significantly.

Among ACL limitation topics (Figure 20a), we observe the following temporal patterns:

- ↑ **Increasing:** *Uncertainty* shows the clearest upward trend, rising from around 5–6% in mid-2022 to over 10% by late 2024. This trend is statistically significant according to the Mann-Kendall test ($\tau = 0.64$, $p = 0.0123$). *Hallucination* also trends upward but does not reach statistical significance ($\tau = 0.42$, $p = 0.107$).
- ↓ **Decreasing:** In contrast, *Social Bias* peaks sharply in 2022-Q3 (26%), then declines and stabilizes below 15% in subsequent quarters. Due to the inconsistency of this change, it is not



(a) ACL LLLMs trends normalized by LLLMs papers.



(b) arXiv LLLMs trends normalized by LLLMs papers.

Fig. 20. LLLMs topics trends for ACL and arXiv datasets based on HDBSCAN + BERTopic clustering approach (normalized by LLLMs papers). Note that y-axis limits vary across subplots to reflect differences in topic prevalence and improve visualization.

significant in the Mann-Kendal test ($p = 0.283$). *Reasoning* remains the most prevalent topic, consistently ranging between 30–40% per quarter. While the test indicates a downward tendency ($\tau = -0.38$), the trend is not statistically significant ($p = 0.152$).

Other topics remain relatively stable over time:

- → **Plateau / Stable:** *Hallucination*, *Security*, and *Generalization* each remain within a 5–15% range with no significant directional movement. *Long Context* also shows minor variation but no sustained change. These patterns are consistent with the Mann-Kendall test, which detects no significant trends for any of these topics.

Some topics also show sharp, isolated changes. *Social Bias* spikes in 2022-Q3 (26%), then drops to around 10% by early 2023. This may reflect heightened concern about bias in the early stages of LLM deployment, which is also visible in the annual distribution (Figure 19), where *Social Bias* is the second most prominent topic in 2022 but (relative to the overall expansion of LLLMs research) declines in later years. *Reasoning* peaks in 2022-Q4 and again in 2023-Q3 (35–40%), likely reflecting increased interest following the launch of ChatGPT and GPT-4 [1] in late 2022 and early 2023. *Security* sees a modest increase in 2024-Q2 and Q3, which might be related to growing concerns about jailbreaks and adversarial attacks.

Among arXiv limitation topics (Figure 20b), we observe the following temporal patterns:

- ↑ **Increasing:** *Multimodality* shows the most consistent upward trend, increasing from around 5% in early 2022 to 15% by early 2025. This pattern is statistically significant according to the Mann-Kendall trend test ($\tau = 0.59$, $p = 0.006$), and likely reflects heightened interest following the release of GPT-4V [1], LLaVA [61], and other vision-language models in mid-2023.
  Hallucination also grows slightly over time, from 10% to 13%, and this increase is statistically significant as well ($\tau = 0.51$, $p = 0.017$). *Knowledge Editing* also exhibits an upward trajectory, though this trend does not reach statistical significance ($\tau = 0.41$, $p = 0.051$).
- ↓ **Decreasing:** *Context & Memory Limitations* exhibits a clear and statistically significant decline, dropping from 25% in 2022-Q1 to around 12–13% by 2025 ($\tau = -0.46$, $p = 0.033$). This shift may reflect improved handling of long inputs through retrieval-augmented methods and the emergence of models with extended context capabilities [56].
  Conversational Limitations follows a similar downward trajectory ($\tau = -0.44$, $p = 0.044$), though its decline appears to plateau at the beginning of 2024.
- → **Stable / Plateau:** Other topics remain relatively stable: *Security Risks* appears to grow from 12% to 18%, and *Social Bias* gradually declines from 23% to 17%, but neither trend is statistically significant ($p = 0.127$ and $p = 0.502$, respectively). This may be because concerns regarding social bias increased immediately following the release of ChatGPT but then returned to the pre-ChatGPT level. *Reasoning*, *Multilinguality*, *Code Generation*, and *Healthcare Application* all fluctuate without a clear directional trend, while *Computational Cost*, *Benchmark Contamination*, *Knowledge Editing*, and *Quantization* remain low throughout.

Table 19. Relative percentage growth of LLM-normalized shares for limitation topics across ACL and arXiv datasets (HDBSCAN+BERTopic). Values are based on relative change between consecutive years. Percentage change is not reported (—) where the earlier year's value is zero.

(a) ACL Dataset

Topic	Topic Shares (2022 → 2024)	→2023 (%)	→2024 (%)
Reasoning	7.0 → 9.2 → 12.0	31.43%	30.43%
Hallucination	0.8 → 2.0 → 3.7	150.00%	85.00%
Security	1.2 → 1.8 → 3.7	50.00%	105.56%
Uncertainty	0.6 → 1.9 → 2.9	216.67%	52.63%
Generalization	1.5 → 2.4 → 2.9	60.00%	20.83%
Social Bias	1.6 → 2.3 → 2.3	43.75%	0.00%
Long Context	1.4 → 1.4 → 2.3	0.00%	64.29%

(b) arXiv Dataset

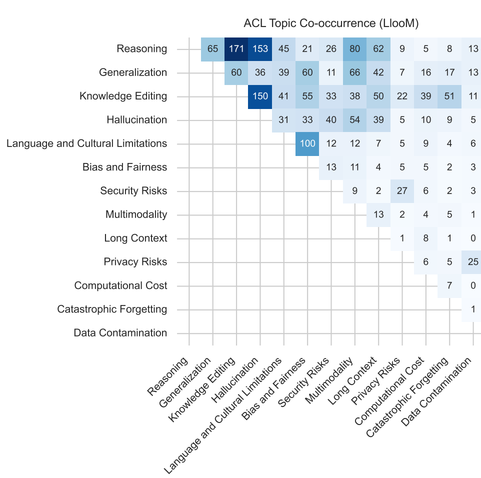
Topic	Topic Shares (2022 → 2025)	→2023 (%)	→2024 (%)	→2025 (%)
Social Bias	0.17 → 0.68 → 1.37 → 1.61	300.00%	101.47%	17.52%
Security Risks	0.11 → 0.59 → 1.33 → 1.66	436.36%	125.42%	24.81%
Reasoning	0.12 → 0.46 → 0.79 → 1.32	283.33%	71.74%	67.09%
Context & Memory Limitations	0.17 → 0.39 → 0.78 → 1.0	129.41%	100.00%	28.21%
Multimodality	0.05 → 0.22 → 0.67 → 0.96	340.00%	204.55%	43.28%
Hallucination	0.03 → 0.29 → 0.66 → 0.74	866.67%	127.59%	12.12%
Conversational Limitations	0.08 → 0.19 → 0.45 → 0.55	137.50%	136.84%	22.22%
Multilinguality	0.04 → 0.12 → 0.28 → 0.39	200.00%	133.33%	39.29%
Code Generation	0.01 → 0.13 → 0.25 → 0.33	1200.00%	92.31%	32.00%
Healthcare Application	0.01 → 0.11 → 0.18 → 0.16	1000.00%	63.64%	-11.11%
Computational Cost	0.01 → 0.05 → 0.1 → 0.18	400.00%	100.00%	80.00%
Benchmark Contamination	0.0 → 0.02 → 0.05 → 0.05	—	150.00%	0.00%
Knowledge Editing	0.0 → 0.01 → 0.04 → 0.03	—	300.00%	-25.00%
Quantization	0.01 → 0.01 → 0.03 → 0.03	0.00%	200.00%	0.00%

D LlooM Topic Co-Occurrence

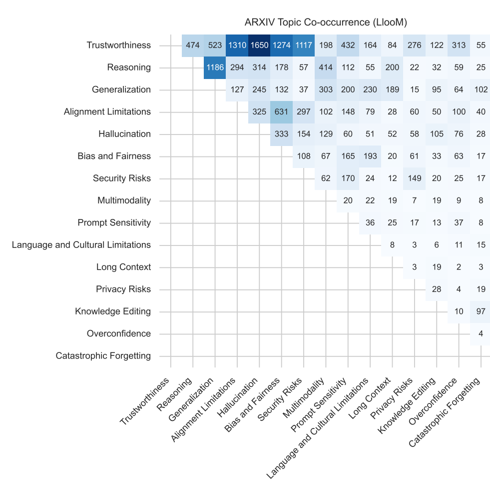
LLLMs are often studied in combination, as reflected in the high number of multi-topic papers: 43% in ACL and over 60% in arXiv, with some spanning up to eight topics (see Table 20). These overlaps often reflect shared task setups (e.g., multimodal hallucination) or related concerns (e.g., alignment and bias). We analyze these links via topic co-occurrence.

Table 20. Distribution of topic counts per paper in ACL and arXiv datasets

Number of Topics	ACL Papers	arXiv Papers
1 topic	1,093	3,317
2 topics	726	4,182
3 topics	253	2,307
4 topics	77	667
5 topics	9	159
6 topics	2	29
7 topics	—	5
8 topics	—	1



(i) ACL Topic Co-occurrence Matrix



(ii) arXiv Topic Co-occurrence Matrix

Fig. 21. Topic co-occurrence matrices for ACL and arXiv LLM limitation papers, clustered by LlooM approach.

In both ACL and arXiv papers, topic co-occurrences tend to concentrate around the largest clusters (see Figure 21). The main patterns are as follows:

- *Reasoning* shows the highest co-occurrence with other limitations, especially in ACL. It frequently overlaps with *Knowledge Editing* (171 ACL), *Hallucination* (153 ACL; 314 arXiv), *Multimodality* (80 ACL; 414 arXiv), and *Generalization* (65 ACL; 1186 arXiv), as well as *Long Context* and *Alignment Limitations*. These links reflect how flawed reasoning can lead to other issues, such as brittle updates in knowledge editing [126], incoherent multimodal responses [63], and hallucinations [72].

- *Trustworthiness* covers not only failures across *Alignment Limitations* (1310), *Security Risks* (1117), *Privacy Risks* (276), but also general output reliability, co-occurring with *Reasoning* (1650), and *Hallucination* (117). Trust breaks down in different ways: flawed reasoning yields faulty responses [34], misalignment leads to persuasive but incorrect outputs [112], and adversarial prompts can trigger data leaks [88].
- *Generalization* co-occurs with *Reasoning*, *Multimodality*, *Knowledge Editing*, *Prompt Sensitivity*, *Security Risks*, and *Bias and Fairness*, reflecting how it is stress-tested across prompts, tasks, and domains. For instance, models often fail under prompt perturbations [7], OOD reasoning [97], or clinical QA [78].
- *Alignment Limitations and Security Risks* cluster with issues focused on output control: *Bias and Fairness* (631), *Prompt Sensitivity* (170), *Privacy Risks* (149), and *Hallucination* (325 with alignment; 154 with security). These links reflect how attempts to control model behavior can introduce new failures: e.g. fine-tuning for alignment can increase hallucination rates [60]. In high-stakes domains like medicine, these concerns combine: hallucination, misalignment, and privacy risks all compromise safe deployment of LLMs in clinical settings [107].
- *Hallucination* appears with a wide range of issues: *Knowledge Editing*, *Language and Cultural Limitations*, and *Long Context* in ACL; *Bias and Fairness* (333), *Multimodality* (129), and *Generalization* in arXiv, highlighting its role across both technical and sociocultural settings.
- *Bias and Fairness* with sociocultural and safety concerns: *Language and Cultural Limitations* (100 ACL; 193 arXiv), *Prompt Sensitivity* (165), *Security Risks* (108), and *Multimodality* (67), forming a distinct subcluster focused on language representation and cultural alignment [23, 110].

These co-occurrence patterns show that many failures are often studied in combination. As a result, even if individual topics are not always the primary focus, they continue to appear in multi-topic work. This overlap likely contributes to the trend stability observed earlier: few topics disappear entirely because they remain relevant through their connections to others.