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

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

Large language model (LLM) research has grown rapidly, along with increasing concern about their limitations. In this survey, we conduct a data-driven, semi-automated review of research on limitations of LLMs (LLLMs) from 2022 to early 2025 using a bottom-up approach. From a corpus of 250,000 ACL and arXiv papers, we identify 14,648 relevant papers using keyword filtering, LLM-based classification, validated against expert labels, and topic clustering (via two approaches, HDBSCAN+BERTopic and LlooM). We find that the share of LLM-related papers increases over fivefold in ACL and nearly eightfold in arXiv between 2022 and 2025. Since 2022, LLLMs research grows even faster, reaching over 30% of LLM papers by 2025. *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 *security risks*, *alignment*, *hallucinations*, *knowledge editing*, and *multimodality*. We offer a quantitative view of trends in LLLMs 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

ACM Reference Format:
Aida Kostikova, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger. 2026. LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models. *ACM Comput. Surv.* 58, 11, Article 282 (April 2026), 33 pages. https://doi.org/10.1145/3801096

This research was funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant no NW21-059A (SAIL). This project originated at the 2024 retreat of the Natural Language Learning & Generation (NLLG) Lab at TU Nürnberg.
Authors' Contact Information: Aida Kostikova (corresponding author), CITEC, Bielefeld University Faculty of Technology, Bielefeld, Germany; e-mail: aida.kostikova@uni-bielefeld.de; Zhipin Wang, University of Technology Nuremberg, Nuremberg, Germany; e-mail: wangzhipin@buaa.edu.cn; Deidamea Bajri, University of Mannheim, Mannheim, Germany; e-mail: deidamea.bajri@students.unimannheim.de; Ole Pütz, CITEC, Bielefeld University Faculty of Technology, Bielefeld, Germany; e-mail: ole.puetz@uni-bielefeld.de; Benjamin Paaßen, CITEC, Bielefeld University Faculty of Technology, Bielefeld, Germany; e-mail: bpaassen@techfak.unibielefeld.de; Steffen Eger, University of Technology Nuremberg, Nuremberg, Germany; e-mail: steffen.eger@utn.de.

ACM 0360-0300/2026/04-ART282
https://doi.org/10.1145/3801096

1 Introduction

With the explosive growth of **large language model** (**LLM**) research and deployment [87], questions of the **limitations of LLMs** (**LLLMs**) have also gathered increased interest, ranging from reasoning failures [63], social bias [47], hallucinations [91], difficulty in handling long contexts [44], 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 [11]. 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 [45, 90], or examine limitations within the broader context of evaluating overall model capabilities [8, 77]. 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 share of LLM-related papers has grown by a factor of over five in ACL and nearly eight 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.

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

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. [98] 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 [27, 60], as well as reasoning and planning capabilities in large-scale models [56].

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 [38, 82], as well as challenges related to hallucination and factual consistency in medical question answering [65]. Other works explore the capabilities and LLLMs in capturing cultural commonsense knowledge [74] and scientific research processes [18, 51]. In recommendation systems, LLMs have been studied as both retrieval and generation engines [89], while in information retrieval, surveys highlight their use in query expansion, passage ranking, and answer synthesis [99]. 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 [29]. Beyond text-based applications, LLMs have been integrated with structured resources such as knowledge graphs [36, 62] and studied in the context of autonomous agent design [84].

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 LLLMs across domains and tasks.

2.2 Surveys on LLLMs

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 [31, 32, 80] and multimodal contexts [49, 71, 72]. Another major focus is reasoning, with surveys analyzing the development of novel techniques[30, 66] such as chain-of-thought prompting [10], reinforced reasoning [45, 90], and mathematical problem-solving [3]. In parallel, the trustworthiness and reliability of LLMs have been studied through the lens of fairness, transparency, and calibration [78], while other works concentrate on security and privacy threats including adversarial vulnerabilities and data leakage [14, 93]. A related thread explores ethical risks in LLM-based agents, particularly concerning safety, misuse, and human interaction [20, 35, 39].

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 LLLMs.

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

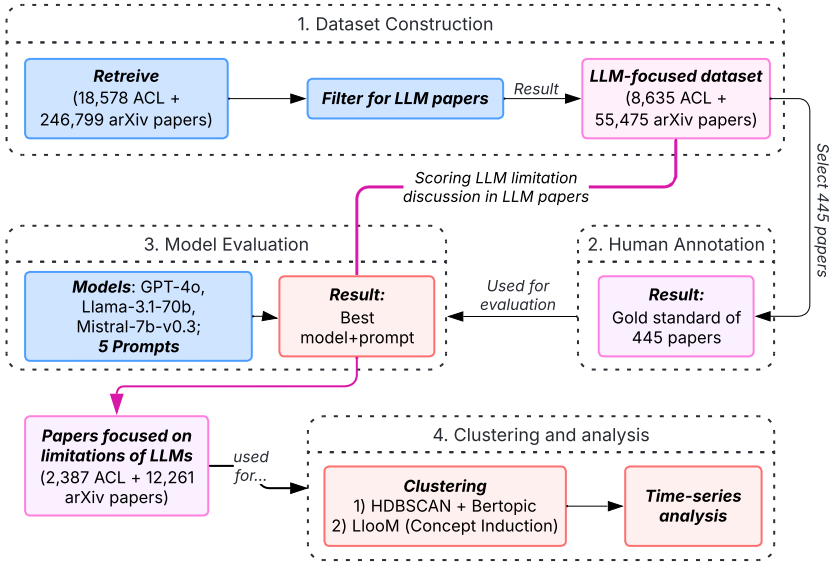


Fig. 1. Overview of the pipeline for our systematic literature review.

methodology, we rely on a growing literature of LLMs being used as instruments for analyzing scientific literature [18]. Several recent approaches employ LLMs for topic modeling, semantic clustering, and concept induction, enabling more interpretable organization of large-scale corpora [16, 19, 40, 64, 97]. 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 [5, 17, 21, 26].

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 [46, 85], while tools like LitLLM [2] and PaSa [28] support LLM-based retrieval, summarization, and exploration of academic texts.

Closer to our work, recent studies use automated methods to analyze research limitations at the paper or corpus level, primarily in support of peer review. Al Azher et al. [4] introduce datasets and methods for extracting and generating study limitations, while Azher et al. [6] apply LLM-assisted topic modeling to limitation sections. Xu et al. [92] evaluate whether LLMs can identify critical limitations in AI papers. In contrast, our work studies LLLMs themselves as the research object.

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 LLLMs, 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 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

Figure 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), complemented by stratified topic-level human evaluation of cluster assignments. In the remainder of this section, we describe each step in more detail.

3.1 Data Retrieval

Our initial dataset includes academic papers published between January 2022 and March 2025, sourced from arXiv as a broad, multi-field research corpus, and from the ACL Anthology as a narrower, NLP-focused set. ArXiv captures preprint research that closely tracks current developments and often appears months before formal publication, making it well suited for analyzing fast-moving research trends in LLMs, 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 and where consistent multi-year coverage enables longitudinal analysis.[2] The time frame was chosen to capture the year preceding the release of ChatGPT as well as all subsequent research on LLMs [98].

For ACL Anthology, we scrape conference pages for AACL 2022–2023, ACL 2022–2024, EACL 2023–2024, EMNLP 2022–2024, ICLR 2022–2024,[3] NAACL 2022 and 2024, and TACL 2022–2024 as the premier NLP venues.[4] For arXiv, we retrieve papers from the categories of Computation & Language (cs.CL), Machine Learning (cs.LG), Artificial Intelligence (cs.AI), and Computer Vision (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 4.3.3 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 consistently available across venues, making them well suited for large-scale automated analysis. Including additional sections (e.g., introductions or related work) could introduce noise, increase preprocessing complexity due to heterogeneous document structures, and raise the cost of large-scale LLM-based annotation.

The final crawled dataset includes 245,835 papers (18,578 papers for ACL, 227,257 for arXiv). Figure 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 [55] to generate initial keywords for each abstract. TNT-KID is a transformer-based keyword tagger that follows a two-stage training process: language model pretraining on large unlabeled text, followed by fine-tuning on labeled keyword data. As reported by Giarelis and Karacapilidis [22], TNT-KID is among the top-performing keyphrase extraction methods and provides ready-to-use pretrained models. We use the off-the-shelf TNT-KID model fine-tuned on the KP20k dataset of scientific abstracts and titles.

[2] In the arXiv dataset, some papers later appear in ACL venues; we retain them to preserve arXiv's role as a source of fast-moving, venue-agnostic research.

[3] Although ICLR is not an ACL venue nor part of the ACL Anthology, we include it in this group to represent a major peer-reviewed machine learning conference with substantial LLM research, and retain the "ACL" label for brevity.

[4] 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.

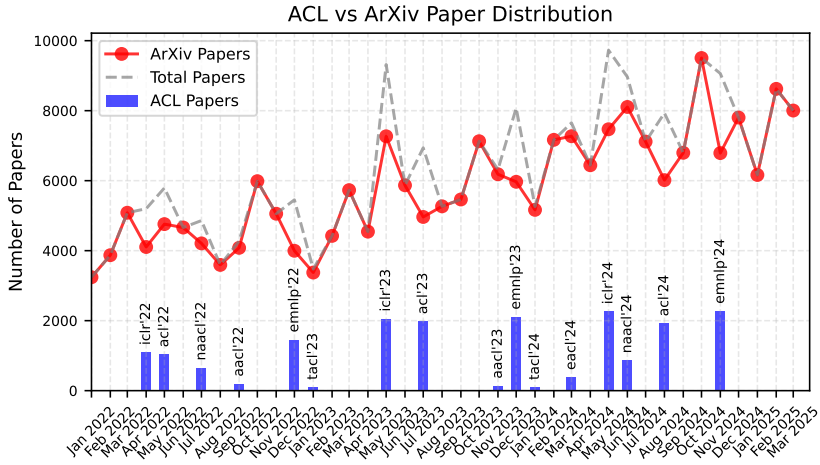


Fig. 2. Distribution of papers over time in the crawled dataset, showing ACL papers, arXiv papers, and the total count (ACL + arXiv).

(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 ≥ 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 of the online supplementary material):

— 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, it serves as an initial inclusion step; its output is refined by subsequent LLM-based classification. Even at this stage, we observe that the proportion of papers passing the

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).*

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

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.

For papers rated 2–5, annotators highlighted textual evidence pointing to the limitation and, where explicitly expressed in the text, its general type (i.e., the kind of limitation discussed), hereafter referred to as "evidence". See Table 2 in the online supplementary material for representative examples of annotated papers and highlighted evidence.

Limitation rating agreement. We measure annotator agreement using:

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**

(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.

In our agreement analysis, we include all papers annotated by at least two annotators (up to four). Agreement is computed pairwise for each annotator pair over the subset of papers they both annotated and averaged across all pairs.

The annotation process included three rounds, involving two professors (natural language processing and machine learning), one PhD student (NLP), and one Master's student (computer science), all of whom are fluent English speakers with at least C1 proficiency. Disagreements were addressed through discussion and refinement of the annotation guidelines between rounds, without per-instance adjudication. 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 highlighted evidence using BIO-tagged sequences, where tokens are labeled as B-EVID, I-EVID, or O (beginning, inside, or outside evidence, respectively). For each paper annotated by at least two annotators (250 of 455 abstracts), we compute pairwise inter-annotator F1 scores at the token level. For each annotator pair and paper, we compare BIO-tagged token sequences, treating B-EVID and I-EVID as positive labels and O as the negative label, and compute F1, excluding annotator pairs for which neither annotator selected any evidence. We report macro-average F1 across all valid annotator pairs.

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 [15].

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 [12], `Mistral-7B-Instruct-v0.3` [33] as a small-scale open-weight

model and `Llama-3.1-70b-Instruct` [23] as a large-scale open-weight model.[5] We also compare the results against a Logistic Regression baseline with SBERT embeddings [68], using random sub-sampling validation with three 80/20 train–test splits, and applying SMOTE oversampling [9] 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 1 of the online 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 [57] + BERTopic [24] and (ii) LLooM concept induction [40], 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. [83] and generate keyphrases for each statement using 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 [...]"[73]

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

[5]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`.

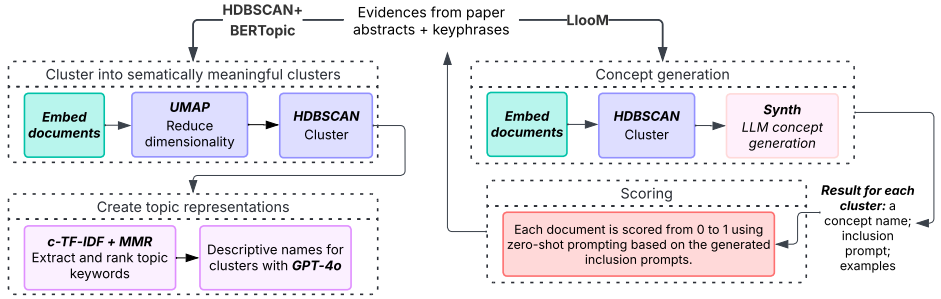


Fig. 3. Comparison of clustering steps in HDBSCAN+BERTopic and LLooM. Both methods take on evidence excerpts with appended keyphrases as an input. For LLooM, we omit the *Distill* step, which is typically used to summarize full documents, as our input already consists of concise excerpts.

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 [24]: we use OpenAI's `text-embedding-3-large` model to embed the combined evidence–keyphrase text, and reduce the embedding using UMAP [58], retaining 10 dimensions for ACL and 5 for arXiv. Such low dimensionality has been shown to be effective for HDBSCAN in prior work [70].

To ensure meaningful clusters and minimize spurious outliers, we tune UMAP, HDBSCAN, and BERTopic parameters separately for ACL and arXiv. This separation accounts for differences in corpus size and topical heterogeneity. 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 online supplementary material.

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 [40].

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 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.

Clustering Validation. We assess the *robustness* of the identified limitation topics using (i) cross-method comparison, which evaluates topic-level alignment (Jaccard overlap, AMI) and trend-level alignment (Kendall's Tau, Spearman's ρ) between HDBSCAN and LlooM, and the *validity* of these

Table 4. Weighted Cohen's Kappa for Limitation Ratings and Pairwise F1 Scores for Evidence Extraction Across Models and 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	—

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.

topics using (ii) human evaluation, in which two expert annotators independently re-annotate papers with limitation topics predefined by HDBSCAN and LlooM, and agreement with each clustering method is measured, based on a stratified sample of 50 ACL and 60 arXiv papers covering all topics identified by both methods. Detailed procedures and results are reported in Section 4.4.

4 Results

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

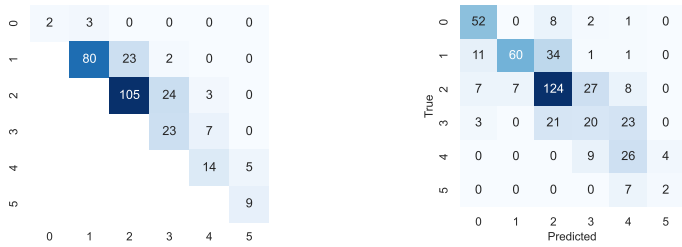
4.1 LLM-Based Filtering Evaluation

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 confuse adjacent categories, such as $2 \leftrightarrow 3$, $3 \leftrightarrow 4$, 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



(a) Annotators' agreement (b) Llama-3.1-70b vs. Final Label

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

Table 5. Comparison of Evidence Extraction between Human Annotators and Llama-3.1-70b Using Prompt 3

Title	Abstract
(1) *"Unlocking Adversarial Suffix Optimization Without Affirmative Phrases: Efficient Black-box Jailbreaking via LLM as Optimizer"* [34], 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"* [88], 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"* [42], 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**

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.

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).

Table 6. Distribution of Ratings for ACL and arXiv 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%)**

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*.

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 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

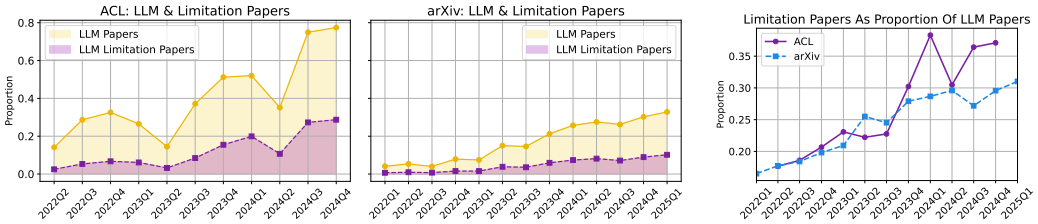
— LLM research is growing rapidly: by late 2024, LLMs account for 75% of ACL papers and over 30% of arXiv papers.
— Research on LLLMs grew even more rapidly, with 1 in 3 LLM papers now addressing limitations.

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;
(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 4 in the online supplementary material. In cs.CL, LLMs are widely discussed, reaching around 80% by early 2025, similar to ACL,



(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. (i) shows the share of LLM and limitation papers among all crawled papers, while (ii) 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.

while cs.AI shows a sharper but lower rise, peaking around 50–60%. In contrast, 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 5(ii), 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. We discuss topics identified with HDBSCAN and LlooM, continue with trends identified with LlooM and provide trends based on HDBSCAN trends in the online supplementary material.

4.3 Clustering Results

4.3.1 *Topics Identified within ACL and arXiv with HDBSCAN and LlooM.*

Key Insights

— Both methods consistently identify shared high-level limitation categories, including *Reasoning*, *Hallucination*, *Security Risks*, *Social Bias*, *Generalization*, and *Long-Context* limitations.
— Differences between HDBSCAN and LlooM primarily reflect methodological choices: HDBSCAN concentrates papers into fewer, broader clusters, whereas LlooM's multi-label assignments distribute papers across finer-grained limitation categories.

Figures 6 and 7 show topic distributions for ACL and arXiv under HDBSCAN+BERTopic and LlooM, respectively. HDBSCAN identifies 7 topics in ACL and 15 in arXiv (13.1% and 14.1% outliers), while LlooM produces comparable high-level topics with lower outlier rates (7.5% and 6.7%).

Both clustering methods identify *Reasoning* as the dominant limitation topic in ACL (36.4% under HDBSCAN vs. 26.3% under LlooM). HDBSCAN concentrates a larger share of papers in this single topic, while LlooM distributes papers more evenly across top limitations, including *Generalization* and *Knowledge Editing*, due to its multi-label assignment. In arXiv, both approaches show a broader topical spread than in ACL, with the most frequent themes being *Social Bias* (16.8%), *Security Risks*

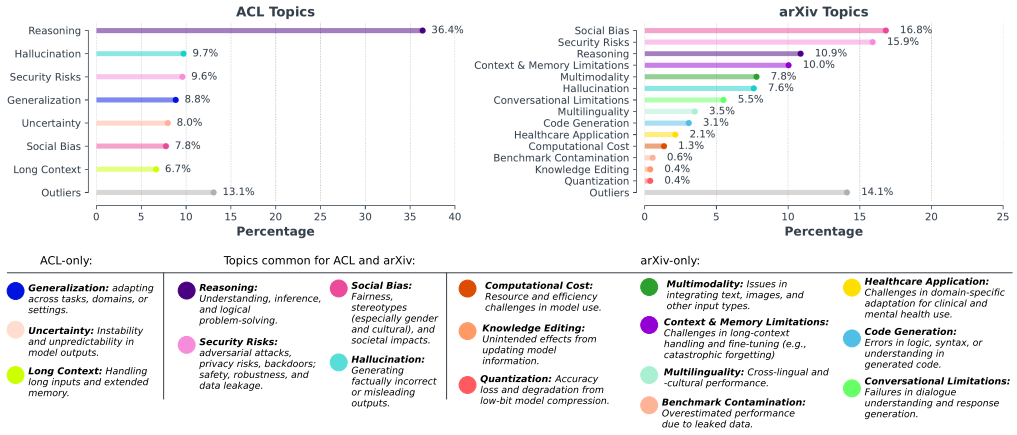


Fig. 6. Topics in ACL Anthology and arXiv, clustered using HDBSCAN + BERTopic. Percentages reflect each topic's proportion out of the total LLM limitation papers (2,387 in ACL and 12,261 in arXiv).

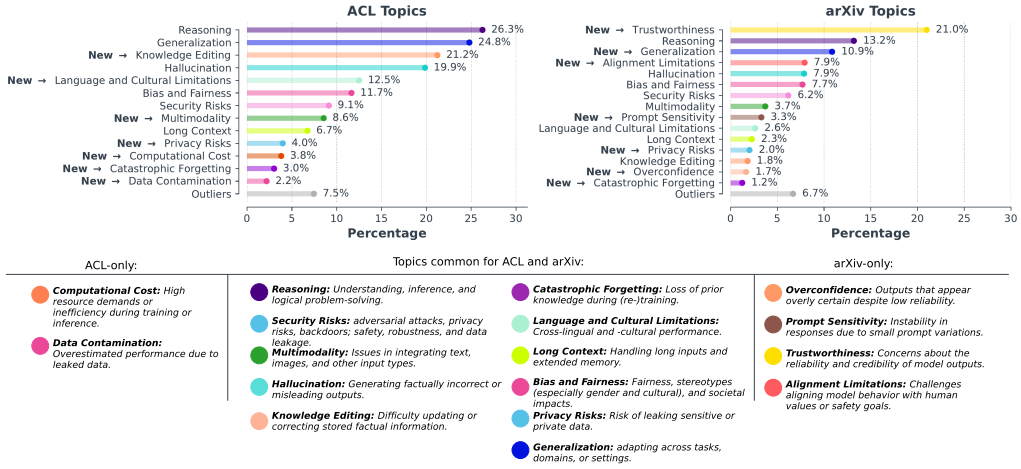


Fig. 7. Topics in ACL Anthology and arXiv, clustered using LlooM approach. Percentages reflect each topic's proportion out of the total LLM limitation papers (2,387 in ACL and 12,261 in arXiv). Since papers can be associated with multiple topics, the percentages may exceed 100% in total.

(15.9%), and *Reasoning* (10.9%) under HDBSCAN, and *Trustworthiness* (21.0%), *Reasoning* (13.2%) and *Generalization* (10.1%) under LlooM.

Topics shared across both methods. Table 7 presents representative papers for topics shared between HDBSCAN and LlooM in the ACL and arXiv datasets. Tables 6 and 7 in the online supplementary material report the dominant terms (top 20 keywords) for each limitation cluster identified with HDBSCAN.

Under both clustering methods, the *Reasoning* cluster captures core cognitive tasks such as natural language understanding (NLU), inference, and logical problem-solving, spanning a broad range of reasoning types (e.g., temporal and commonsense). It also includes prompt-related terms such as *chain-of-thought* and *prompt*, highlighting the role of prompt engineering in complex reasoning.

Table 7. Representative Limitation-Focused Papers for Topics That Are Shared by HDBSCAN and LlooM and That Occur in the ACL and/or arXiv Corpora

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 [63])
(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 [52])
(3) Security Risks	"The prevalence and strong capability of 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 [76])
(4) Generalization	"However, in-domain demonstrations are not always readily available in real scenarios, leading to cross-domain in-context learning. Besides, LLMs are still facing challenges in long-tail knowledge in unseen and unfamiliar domains." (EMNLP 2023 [50])
(5) Social Bias *(Bias & Fairness)*	"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 [47])
(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 [44])
(7) Multiling. *(Lang. & Cult.)*	"The automated systems [...] are primarily designed for and work far more effectively in English than in the world's other 7,000 languages. [...]" (arXiv, June 2023 [61])
(8) Knowledge Editing	"Existing editing methods lead to inevitable performance deterioration on general benchmarks. [...] When the number of edits is slightly large, the intrinsic knowledge structure of the model is disrupted or even completely damaged." (arXiv, October 2024 [43])
(9) 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 [94])
(10) Comput. Cost	"Training and deploying LLMs are expensive as it requires considerable computing resources and memory...[...]" (arXiv, November 2023 [96])

† Assigned to *Reasoning* under LlooM.

Topics that occur in both corpora are shown once, with dataset indicated by the venue/date preceding the citation. Unless marked with †, each paper is assigned to the conceptually corresponding topic by both methods. Topic labels follow HDBSCAN; parentheses show the LlooM label when it differs. Numbers in the "Topic" column indicate the example ID referenced in the main text.

Under HDBSCAN in ACL, however, the *Reasoning* cluster is broader, absorbing multimodal tasks (example 1 in Table 7), and overlapping with multilinguality and benchmark design. In contrast, in arXiv under HDBSCAN and in both datasets under LlooM, *Reasoning* is more distinct, as *Multilinguality* / *Language and Cultural Limitations* (difficulties in handling multilingual input, low-resource languages, or culturally specific content) and *Multimodality* (challenges in integrating and reasoning over inputs from different modalities, such as text and images, example 9) are separated into their own clusters. This difference likely contributes to the higher prevalence of *Reasoning* in ACL under HDBSCAN, though it may also partly reflect clustering artefacts.

Several shared clusters focus on the correctness and reliability of model outputs in both ACL and arXiv. The *Hallucination* cluster addresses failures in factual accuracy, with terms related to faithfulness, trust, and correctness (example 2), and also appears in multimodal contexts such as image captioning and generation. The *Security* cluster captures risks including jailbreaks, adversarial prompting, and backdoors (example 3). Under LlooM, security-related concerns further separate into distinct *Security* and *Privacy* topics. The *Social Bias* cluster focuses on fairness, representation, and demographic disparity (example 5), with keywords such as *stereotype*, *gender*, *cultural*, and *demographic*.

Other shared clusters reflect technical constraints in model design and deployment. *Long Context* / *Context & Memory Limitations* address failures in handling extended inputs and maintaining

information over long outputs (example 6 in Table 7), closely related to increased resource demands captured by the *Computational Cost* cluster (example 10). The *Generalization* cluster captures limitations in robustness and domain transfer (example 4). *Knowledge Editing* focuses on the difficulty of updating or correcting model knowledge without retraining, often leading to unintended performance degradation (example 8).

HDBSCAN-specific clusters. In addition to shared topics, HDBSCAN identifies several clusters that do not emerge under LlooM. Representative papers for these clusters are shown in Table 8 in the online supplementary material. In ACL, the *Uncertainty* cluster describes behavioral instability, including prompt sensitivity and calibration errors. In arXiv, additional clusters capture a wider range of specialized concerns: *Conversational Limitations Code Generation*, *Healthcare Application*, *Benchmark Contamination*, 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 13) and implementation-level tradeoffs. For example, *Benchmark Contamination* refers to test data leakage into training sets (example 14), and *Quantization* concerns scaling efficiency (example 15).

LlooM-specific clusters. 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 Tables 3 and 4 in the online supplementary material). For each LlooM-specific cluster, we also provide representative paper examples in Table 8 in the online supplementary material.

- *Trustworthiness* (arXiv): the largest cluster in the arXiv set (Figure 7). This category describes concerns about the reliability, transparency, and reproducibility of LLM outputs (see example 19 and 20 which refer 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 online supplementary material.
- *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.
- *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.
- *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 16 (ACL) demonstrates privacy breaches from pretraining on sensitive data.
- *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).

The comparison of shared and method-specific clusters highlights both similarities and systematic differences between HDBSCAN and LlooM (we quantify these differences in Section 4.4). While HDBSCAN provides an interpretable, unsupervised clustering of limitations, it assigns each paper to a single cluster and becomes increasingly fragmented as more clusters are added, particularly in the ACL dataset. This makes it difficult to capture overlapping limitations and may complicate the

interpretation of trends. By contrast, LlooM allows for broader, non-mutually exclusive categories through multi-label assignment, which reduces fragmentation and provides a more stable basis for longitudinal analysis. Accordingly, we rely on LlooM for the trend analysis in the following section. Trend results derived from HDBSCAN are reported in Section C of the online supplementary material for completeness.

4.3.2 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.

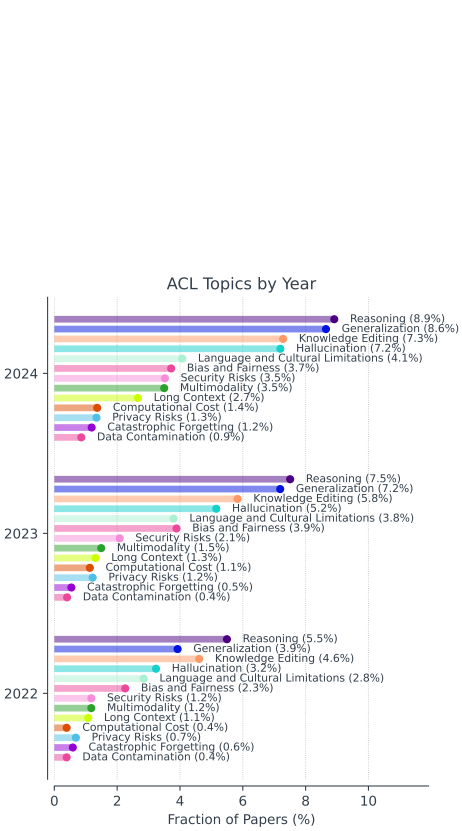
(i) How are limitation topics represented in the broader growth of LLM research?

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 8 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 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 5 in the online supplementary material).

As seen in Figure 8, 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.



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

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

Fig. 8. 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).

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 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.

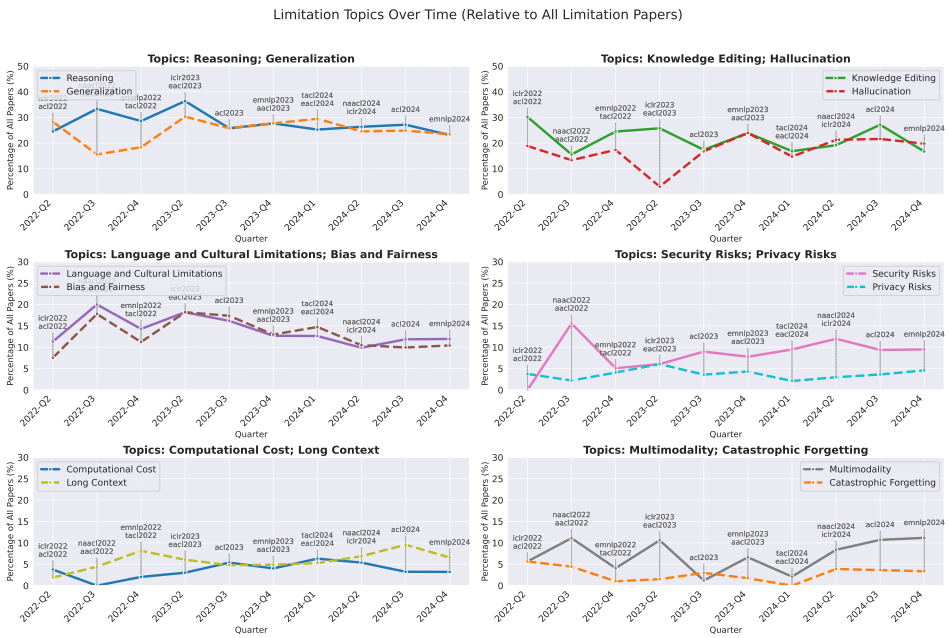


Fig. 9. LLLMs topics trends for the ACL dataset based on LLooM clustering approach. Note that y-axis limits vary across subplots to reflect differences in topic prevalence and improve visualization.

Figure 9 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 [37, 53] 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 10):

- ↑ **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,

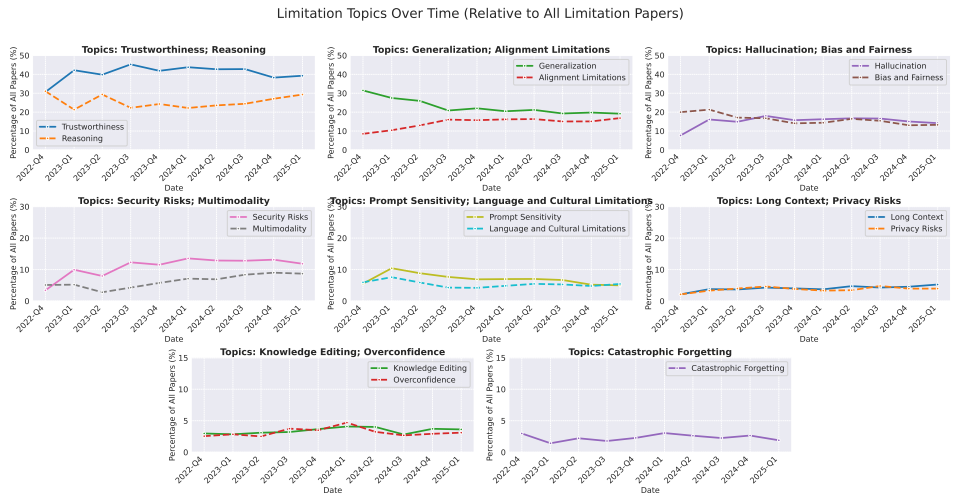


Fig. 10. LLLMs topics trends for the arXiv dataset based on LLooM clustering approach. Note that y-axis limits vary across subplots to reflect differences in topic prevalence and improve visualization.

Hallucination also shows a positive trend (τ = 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 (Figures 9, 10). 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 3 in the online supplementary material), 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 [13], and LLaMA [81] between February and July 2023, which likely contributed to the expansion and differentiation of LLM limitations research during this period.

4.3.3 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 5 (right) in the online supplementary material shows how topic shares differ across arXiv categories, while Figure 11 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 12 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*,

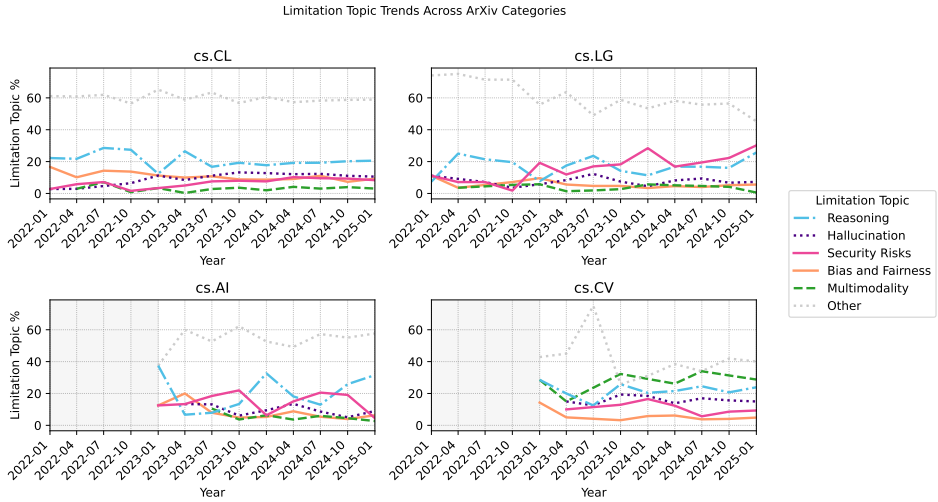


Fig. 11. Limitation topic trends across four main arXiv categories (cs.CL, cs.LG, cs.AI, cs.CV). To maintain visual clarity, only the five most frequent topics (based on their overall percentage distribution across categories, as shown in Figure 5 in the online supplementary material) are shown; others are grouped as "Other." Grey shading in cs.AI and cs.CV marks periods before 2023 with insufficient data for these categories.

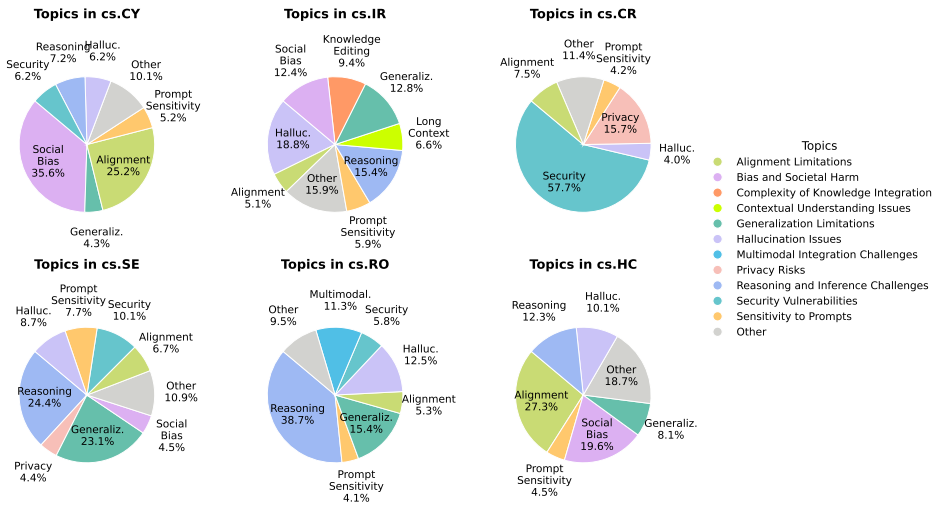


Fig. 12. Distribution of limitations-related topics in six arXiv categories with lower paper counts in our dataset, as shown in Figure 5 in the online supplementary material. Each chart includes only papers where the category is assigned as the primary arXiv category. Topics that make up less than 3% 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.

4.4 Clustering Methods Comparison

The aforementioned topic distributions differ depending on the clustering method. In this section, we compare HDBSCAN+BERTopic (Section C in the online supplementary material) and LlooM (Section 4.3) to identify which findings are stable and which may be method-specific. We first assess trend agreement, then explore methodological differences to explain any result divergences.

4.4.1 Cross-Method Comparison Between HDBSCAN and LlooM. 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 online supplementary material and Section 4.3) 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 first identify candidate matches for each cluster based on identical or semantically similar names and validate these by computing the Jaccard overlap between their associated paper sets.[6] 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 Tables 9 and 10 in the online supplementary material, confirming that major topics identified by name also show substantial paper overlap.

Trend Alignment Between HDBSCAN and LlooM. Table 8 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

[6] All matches except *Knowledge Editing* were confirmed by highest Jaccard overlap; *Knowledge Editing* was aligned manually.

Table 8. Comparison of Limitation Trends Identified by HDBSCAN and LlooM for ACL and arXiv Datasets

(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.*

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).*

Table 9. 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)	

pipelines. To better understand this, we compare HDBSCAN+BERTopic and LlooM in Table 9 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.[7]

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 show that LlooM and HDBSCAN identify similar broad limitation areas but organize papers differently at a finer-grained level. This is supported by Tables 9 and 10 in the online supplementary material, which report the closest topic alignments between methods. Large areas

[7]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 10. Inter-Annotator and Human–Method Agreement for HDBSCAN and LlooM on ACL and arXiv Datasets

Dataset	Method	Inter-Annotator	Human–Method
ACL	HDBSCAN	$\kappa = 0.79$	$\kappa = 0.74$
ACL	LlooM	boot-F1 = 0.70	boot-F1 = 0.61
arXiv	HDBSCAN	$\kappa = 0.71$	$\kappa = 0.55$–0.59
arXiv	LlooM	boot-F1 = 0.53	boot-F1 = 0.49–0.51

Cohen's κ is used for HDBSCAN (single-label), and boot-F1 is used for LlooM (multi-label).

such as *Reasoning*, *Hallucination*, *Security Risks*, and *Bias and Fairness* appear relatively stable and align well across methods and datasets ($J > 0.4$), whereas smaller topics, including *Overconfidence* and *Prompt Sensitivity*, are often absorbed into broader categories such as *Hallucination*. This reflects differences in clustering granularity: LlooM splits topics into overlapping subcategories, while HDBSCAN merges related issues into broader clusters. For example, HDBSCAN merges LlooM's *Security Risks*, *Privacy Risks*, and *Trustworthiness* into a single *Security* cluster.

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.

4.4.2 Human Evaluation. While a full gold-standard clustering is not scalable, we assess cluster validity through stratified human re-annotation of cluster assignments and measure both inter-annotator and human–method agreement for both HDBSCAN and LlooM.

For both ACL and arXiv, we construct a stratified human evaluation dataset (50 papers for ACL and 60 for arXiv) by sampling papers with approximately uniform per-topic sample sizes, including outliers, stratified with respect to HDBSCAN clusters. Each sampled paper is annotated with both its HDBSCAN and LlooM topic assignments. Because LlooM allows multi-label assignments, LlooM topic frequencies in the evaluation dataset are not explicitly controlled during sampling; however, all LlooM topics identified in the corpus are represented.[8]

The dataset was annotated by a professor (machine learning) and a PhD student (NLP). Annotators were provided with each paper's title, extracted evidence, and keyphrases, corresponding to the input used by the clustering methods (as described in Section 3.5). For HDBSCAN, annotators selected a single topic per paper. For LlooM, annotators could assign multiple topics per paper, reflecting its multi-label clustering setup. Annotators were shown representative BERTopic keyword lists and LlooM assignment prompts to convey cluster semantics.

We measure both inter-annotator agreement and human–method agreement, using Cohen's κ for HDBSCAN, which produces single-label assignments, and boot-F1 [54] for LlooM to account for chance agreement in the multi-label setting.

Table 10 shows substantial inter-annotator agreement for both datasets ($\kappa = 0.71$–0.79 for HDBSCAN and boot-F1 = 0.53–0.70 for LlooM), indicating consistent human judgments across both single-label and multi-label settings. Human–method agreement is also substantial for ACL ($\kappa = 0.74$, boot-F1 = 0.61) and moderate for arXiv ($\kappa = 0.55$–0.59, boot-F1 = 0.49–0.51), indicating that human annotations broadly correspond to the topics produced by the clustering methods.

[8]Except for the broad *Trustworthiness* cluster in arXiv, which was excluded from human annotation due to substantial overlap with other topics.

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 [25]. Meta-analyses confirm this trend, showing a sharp rise in evaluation-focused papers from 2020 to 2023 [8].

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 online supplementary material, 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 [75], 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 [86]. Yet ensuring safety without compromising performance remains an open challenge [67, 79], 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 [7, 91].

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 [95]. This trend likely reflects growing interest caused by

the release of GPT-4V [1], LLaVA [48], and other vision-language models in mid-2023. These and other trends discussed above coincide with broader shifts already noted in previous studies of the arXiv corpus: in particular, previous work finds that 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 [41]. Similarly, authorship diversified, with many newcomers from computer vision, security, and software engineering [59].

4. Despite methodological differences, HDBSCAN and LlooM identify overlapping high-frequency topics (e.g., Reasoning, Hallucination, Security Risks) and similar trend patterns, supporting 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 their differences, both methods recover the same dominant topics, especially *Reasoning*, *Hallucination*, and *Security Risks*, with strong agreement in topic composition and trend trajectories. And although smaller topics (e.g., *Prompt Sensitivity*) and trend significance can vary, the main trends 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 [40]. HDBSCAN+BERTopic can also vary across runs because UMAP is stochastic and sensitive 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. In particular, our reliance on abstract- and keyword-level coverage and on the extraction of explicitly stated limitation evidence may undercount limitations discussed using non-standard terminology, or newly emerging terms (e.g., in fields such as CV).
- 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 [69, 98].
- Given our automated survey design, trend analyses reflect the prevalence of limitation-related discussions in research rather than the severity of the discussed limitations. Although sustained research interest can reflect practical relevance, changes in topic prevalence should primarily be interpreted as shifts in research attention.
- This survey provides a large-scale overview of LLLMs research rather than a fine-grained taxonomy; while the identified categories do not capture detailed limitation subtypes, the topics are high-level to reflect dominant and stable themes across methods and datasets.

Future work can extend these findings in several ways. First, limitation topics can be decomposed into finer subcategories, such as specific reasoning types or bias forms, using hierarchical or agglomerative clustering. Second, expanding the analysis to earlier years, especially after BERT's introduction in 2018, could show how concerns about smaller PLMs evolved with model scaling.

Third, the automatic pipeline can be adapted to incorporate newly published papers, enabling regular updates of the dataset over time.

Acknowledgments

We thank Yanran Chen, Christoph Leiter, Ran Zhang, Daniil Larionov, and Jonas Belouadi for valuable early input and discussions that helped shape this work.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2023. Gpt-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[2] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy D. J. Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. arXiv:2402.01788. Retrieved from https://arxiv.org/abs/2402.01788

[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:2402.00157. Retrieved from https://arxiv.org/abs/2402.00157

[4] Ibrahim Al Azher, Miftahul Jannat Mokarrama, Zhishuai Guo, Sagnik Ray Choudhury, and Hamed Alhoori. 2025. BAGELS: Benchmarking the automated generation and extraction of limitations from scholarly text. arXiv:2505.18207. Retrieved from https://arxiv.org/abs/2505.18207

[5] 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. *Metrics*, 2, 2 (2025), 5. DOI: https://doi.org/10.3390/metrics2020005

[6] Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*. 1–12.

[7] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. arXiv:2409.05746. Retrieved from https://arxiv.org/abs/2409.05746

[8] 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 TIST* 15, 3 (2024), 1–45. DOI: https://doi.org/10.1145/3641289

[9] 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, 1 (2002), 321–357.

[10] 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:2503.09567. Retrieved from https://arxiv.org/abs/2503.09567

[11] 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:2405.01769. Retrieved from https://arxiv.org/abs/2405.01769

[12] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open Platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning* 235 (2024), 8359–8388.

[13] 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: https://doi.org/10.5555/3648699.3648939

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

[15] 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:1911.03429. Retrieved from https://arxiv.org/abs/1911.03429

[16] Jairo Diaz-Rodriguez. 2025. k-LLMmeans: Summaries as centroids for interpretable and scalable LLM-based text clustering. arXiv:2502.09667. Retrieved from https://arxiv.org/abs/2502.09667

[17] 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.

[18] 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:2502.05151. Retrieved from https://arxiv.org/abs/2502.05151

[19] 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: https://doi.org/10.18653/v1/2024.emnlp-main.1025

[20] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. arXiv:2411.09523. Retrieved from https://arxiv.org/abs/2411.09523

[21] 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: https://doi.org/10.3390/app14177675

[22] Nikolaos Giarelis and Nikos Karacapilidis. 2024. Deep learning and embeddings-based approaches for keyphrase extraction: A literature review. *Knowledge and Information Systems* 66, 11 (2024), 6493–6526.

[23] 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:2407.21783. Retrieved from https://arxiv.org/abs/2407.21783

[24] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. Retrieved from https://arxiv.org/abs/2203.05794

[25] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. arXiv:2310.19736. Retrieved from https://arxiv.org/abs/2310.19736

[26] 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: https://doi.org/10.1016/j.dim.2024.100066

[27] Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Syed Zohaib Hassan, et al. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv*. DOI: 10.36227/techrxiv.23589741.v8

[28] Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. PaSa: An LLM agent for comprehensive academic paper search. arXiv:2501.10120. Retrieved from https://arxiv.org/abs/2501.10120

[29] 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: https://doi.org/10.1145/3695988

[30] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. arXiv:2212.10403. Retrieved from https://arxiv.org/abs/2212.10403

[31] 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: https://doi.org/10.1145/3703155

[32] 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: https://doi.org/10.1145/3571730

[33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]. Retrieved from https://arxiv.org/abs/2310.06825

[34] 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:2408.11313. Retrieved from https://arxiv.org/abs/2408.11313

[35] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating llm ethics: Advancements, challenges, and future directions. arXiv:2406.18841. Retrieved from https://arxiv.org/abs/2406.18841

[36] 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: https://doi.org/10.1109/TKDE.2024.3469578

[37] Maurice G. Kendall. 1948. *Rank Correlation Methods*. Charles Griffin, London.

[38] 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:2502.04381. Retrieved from https://arxiv.org/abs/2502.04381

[39] Ashutosh Kumar, Shiv Vignesh Murthy, Sagarika Singh, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms. arXiv:2401.12273. Retrieved from https://arxiv.org/abs/2401.12273

[40] 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: https://doi.org/10.1145/3613904.3642830

[41] 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:2412.12121. Retrieved from https://arxiv.org/abs/2412.12121

[42] Benjamin A. Levinstein and Daniel A. Herrmann. 2025. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies* 182, 7 (2025), 1539–1565.

[43] Qi Li, Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Xinglin Pan, and Xiaowen Chu. 2024. Should we really edit language models? On the evaluation of edited language models. *Advances in Neural Information Processing Systems* 37 (2024), 30850–30885.

[44] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. arXiv:2310.06201. Retrieved from https://arxiv.org/abs/2310.06201

[45] 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:2502.17419. Retrieved from https://arxiv.org/abs/2502.17419

[46] 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:2502.14776. Retrieved from https://arxiv.org/abs/2502.14776

[47] 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:2210.15144. Retrieved from https://arxiv.org/abs/2210.15144

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems* 36 (2023), 34892–34916.

[49] 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:2402.00253. Retrieved from https://arxiv.org/abs/2402.00253

[50] Quanyu Long, Wenya Wang, and Sinno Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6525–6542.

[51] Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A survey on large language models for scientific research. arXiv:2501.04306. Retrieved from https://arxiv.org/abs/2501.04306

[52] 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:2410.15116. Retrieved from https://arxiv.org/abs/2410.15116

[53] Henry B. Mann. 1945. Nonparametric tests against trend. *Econometrica* 13, 3 (July 1945), 245–259.

[54] Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3659–3668.

[55] 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.

[56] Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. arXiv:2501.04040. Retrieved from https://arxiv.org/abs/2501.04040

[57] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (2017), 205.

[58] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. Retrieved from https://arxiv.org/abs/1802.03426

[59] 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:2307.10700. Retrieved from https://arxiv.org/abs/2307.10700

[60] 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:2307.06435. Retrieved from https://arxiv.org/abs/2307.06435

[61] Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-English content analysis. arXiv:2306.07377. Retrieved from https://arxiv.org/abs/2306.07377

[62] 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: https://doi.org/10.1109/TKDE.2024.3352100

[63] Brendan Park, Madeline Janecek, Naser Ezzati-Jivan, Yifeng Li, and Ali Emami. 2024. Picturing ambiguity: A visual twist on the winograd schema challenge. arXiv:2405.16277. Retrieved from https://arxiv.org/abs/2405.16277

[64] 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: https://doi.org/10.18653/v1/2024.emnlp-industry.54

[65] Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. arXiv:2408.13808. Retrieved from https://arxiv.org/abs/2408.13808

[66] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. arXiv:2407.11511. Retrieved from https://arxiv.org/abs/2407.11511

[67] 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:2310.03693. Retrieved from https://arxiv.org/abs/2310.03693

[68] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv:1908.10084. Retrieved from https://arxiv.org/abs/1908.10084

[69] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *TACL* 8 (2020), 842–866. DOI: 10.1162/tacl_a_00349

[70] 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 *Proceedings of the 14th International Workshop on Semantic Evaluation*. 187–193.

[71] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 11709–11724. DOI: 10.18653/v1/2024.findings-emnlp.685

[72] 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:2405.09589. Retrieved from https://arxiv.org/abs/2405.09589

[73] 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:2212.08061. Retrieved from https://arxiv.org/abs/2212.08061

[74] 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:2405.04655. Retrieved from https://arxiv.org/abs/2405.04655

[75] 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:2309.15025. Retrieved from https://arxiv.org/abs/2309.15025

[76] 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.

[77] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and Garriga-Alonso, Adrià. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615. Retrieved from https://arxiv.org/abs/2206.04615

[78] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. arXiv:2401.05561. Retrieved from https://arxiv.org/abs/2401.05561

[79] 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:2502.11555. Retrieved from https://arxiv.org/abs/2502.11555

[80] S. M. Tonmoy, S. M. 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:2401.01313. Retrieved from https://arxiv.org/abs/2401.01313

[81] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971. Retrieved from https://arxiv.org/abs/2302.13971

[82] 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: https://doi.org/10.1038/s41591-024-02855-5

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

[84] 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: https://doi.org/10.1007/s11704-024-40231-1

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

[86] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. arXiv:2407.16216. Retrieved from https://arxiv.org/abs/2407.16216

[87] 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. *Transactions on Machine Learning Research* 2022 (2022). Available at https://openreview.net/forum?id=yzkSU5zdwD

[88] 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:2310.09909. Retrieved from https://arxiv.org/abs/2310.09909

[89] 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: https://doi.org/10.1007/s11280-024-01291-2

[90] 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:2501.09686. Retrieved from https://arxiv.org/abs/2501.09686

[91] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. arXiv:2401.11817. Retrieved from https://arxiv.org/abs/2401.11817

[92] Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. Can LLMs identify critical limitations within scientific research? A systematic evaluation on AI research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.), Association for Computational Linguistics, Vienna, Austria, 20652–20706. DOI: https://doi.org/10.18653/v1/2025.acl-long.1009

[93] 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* 4, 2 (2024), 100211. DOI: https://doi.org/10.1016/j.hcc.2024.100211

[94] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. arXiv:2303.15445. Retrieved from https://arxiv.org/abs/2303.15445

[95] 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:2401.13601. Retrieved from https://arxiv.org/abs/2401.13601

[96] Longteng Zhang, Xiang Liu, Zeyu Li, Xinglin Pan, Peijie Dong, Ruibo Fan, Rui Guo, Xin Wang, Qiong Luo, Shaohuai Shi, and et al. 2023. Dissecting the runtime performance of the training, fine-tuning, and inference of large language models. arXiv:2311.03687. Retrieved from https://arxiv.org/abs/2311.03687

[97] Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. arXiv:2305.14871. Retrieved from https://arxiv.org/abs/2305.14871

[98] 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:2303.18223. Retrieved from https://arxiv.org/abs/2303.18223

[99] 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:2308.07107. Retrieved from https://arxiv.org/abs/2308.07107

Received 26 May 2025; revised 12 February 2026; accepted 1 March 2026