Preaching to the ChoIR: Lessons IR Should Share with AI

Gianluca Demartini
g.demartini@uq.edu.au
The U. of Queensland
Brisbane, Australia

Claudia Hauff
claudiah@spotify.com
Spotify
Delft, The Netherlands

Matthew Lease
ml@utexas.edu
The U. of Texas at Austin
Austin, TX, USA

Stefano Mizzaro
mizzaro@uniud.it
University of Udine
Udine, Italy

Kevin Roitero
kevin.roitero@uniud.it
University of Udine
Udine, Italy

Mark Sanderson
mark.sanderson@rmit.edu.au
RMIT University
Melbourne, Australia

Falk Scholer
falk.scholer@rmit.edu.au
RMIT University
Melbourne, Australia

Chirag Shah
chirags@uw.edu
University of Washington
Seattle, WA, USA

Damiano Spina
damiano.spina@rmit.edu.au
RMIT University
Melbourne, Australia

Paul Thomas
pathom@microsoft.com
Microsoft
Adelaide, Australia

Arjen P. de Vries
arjen@acm.org
Radboud University
Nijmegen, The Netherlands

Guido Zuccon
g.zuccon@uq.edu.au
The U. of Queensland
Brisbane, Australia

Abstract

The field of Information Retrieval (IR) changed profoundly at the end of the 1990s with the rise of Web Search, and there are parallels with developments in Artificial Intelligence (AI) happening today with the advent of ChatGPT, Large Language Models, and Generative AI. We acknowledge that there are clear differences between IR and AI. For example, IR is a much smaller field, and new problems arise, like data contamination that may affect benchmark-based evaluation of AI systems. But looking through the lens of an IR researcher, there are many striking similarities between the two fields of IR (25 years ago) and AI (today), and many topics appearing in discussions in AI resemble those of 25 years ago in IR: benchmark reliability and robust evaluation, reproducibility of results for non-public models, privacy and copyright issues, efficiency and scalability, etc. In this paper, we discuss similarities and differences between IR and AI and then derive some lessons learned in the field of IR as a list of recommendations – urging the IR community to reflect on, discuss, and convey these lessons to the AI field. We believe that a joint community effort by all IR researchers is both necessary and dutiful to obtain a fruitful discussion and research advancements with the AI community.

CCS Concepts

• **Information systems → Information retrieval**; • **Computing methodologies → Artificial intelligence**.

Keywords

Research Community, Lessons Learned, Artificial Intelligence

ACM Reference Format:
Gianluca Demartini, Claudia Hauff, Matthew Lease, Stefano Mizzaro, Kevin Roitero, Mark Sanderson, Falk Scholer, Chirag Shah, Damiano Spina, Paul Thomas, Arjen P. de Vries, and Guido Zuccon. 2025. Preaching to the ChoIR: Lessons IR Should Share with AI. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3731120.3744612

ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744612

1 Introduction

Originally, Information Retrieval (IR) was about designing systems to support librarians help visitors of a library find relevant content in the library catalog. Then, since the introduction of the Web, the IR field moved to finding relevant content online. In 1997, Google search was released to the public, and this disrupted the field: the outsider reaction was "Google has solved IR, so why are you still doing IR research?" while insiders wondered if any research was possible without large query logs; meanwhile in the following years the best and brightest students left academia and were recruited into Big Tech. The last 25 years showed, however, that the community not only survived, but continued to improve: IR research thrived in an ecosystem where industry and academic research produced significant contributions to the world of search. Research (re)focused onto key open problems (evaluation, reproducibility, conversational search, efficiency, domain-specific, etc.). The proliferation of Web search engines also created new research challenges for the IR community (e.g., product and job search, improving ranking from implicit feedback gathered from click data, image search, etc.) These advancements not only created a richer research environment in IR, but also informed the development of techniques and methods in other fields (e.g., more robust evaluation of recommender systems).

Today, after the launch of ChatGPT, it can be argued that the field of Artificial Intelligence (AI) is in a similar situation to IR in the 1990s. We hear that "natural language processing is a solved problem", and the issue of a migration from academia to private companies is discussed widely. While we will claim that AI today can benefit from a clear understanding of the evolution of the IR field in response to the Web search disruption, we make clear that these fields are not exactly the same. The IR community is much smaller. The area of AI spans many sub-disciplines, such that discussions within the area of AI can have a higher conflictuality. While we see interest in AI in research topics that are well known to IR researchers, including reliability and robustness of evaluation benchmarks, reproducibility of results, privacy and copyright issues, efficiency and scalability, and so on, completely new problems are encountered as well, like data contamination for benchmark-based evaluation of AI systems.

Yet, we observe striking similarities between the two fields of IR (25 years ago) and AI (today). This paper aims to raise awareness within the IR community about the need of consolidating key recommendations for the field of AI, to contribute to a discussion that is already ongoing [61, 79, 85, 131]. To achieve this, we summarize the current events that affect AI research, describe the similarities with, and the differences from, IR, and we reflect on lessons learned in the field of IR to suggest an initial draft of recommendations that the IR community could make to the AI field today.

Tongue in cheek, we acknowledge right from the very start of the paper that in this first step we are 'preaching to the choir' by targeting an IR conference. We believe this *call to action* to the IR community is both timely and urgent. By engaging in discussions about our field's core contributions, past mistakes, and the lessons learned, we can consolidate our knowledge to be shared beyond our domain.

2 Background

2.1 AI and IR

Properly defining AI in detail is well beyond the scope of this paper. The simple incipit of the Wikipedia page, stating that AI "in its broadest sense, is intelligence exhibited by machines, particularly computer systems" [1] will suffice, although we notice that even the most authoritative AI textbook [133] adopts quite a radical approach based on the notion of rational agent.

Although according to Wooldridge "for [many machine learning experts], AI is the long list of failed ideas" [176, end of Ch. 5], we consider AI as a general field encompassing Machine Learning and Deep Learning (as many do). This paper is motivated by Generative AI (GenAI) and Large Language Models (LLMs), arguably the two recent contributions that are causing disruption in the AI field at large, and even outside it. Similarly, we consider 'information retrieval' broadly: not just to address the specific ranking problem, but to understand how to satisfy users' information needs. In our discussion we will refer to other related fields, as Natural Language Processing (NLP), Recommender Systems (RecSys), or Human-Computer Interaction (HCI). We will also refer to both the research fields and the research communities, i.e., the researchers – people – working in them.

Both AI and IR have a long history, and people have argued about their relative position for many years. The same AI Wikipedia page states that "High-profile applications of AI include advanced web search engines (e.g., Google Search); recommendation systems (used by YouTube, Amazon, and Netflix)", but this is controversial at least. For example, Wilks contributed a very interesting perspective to the book in memory of Karen Spärck Jones [174]:

> …the field of [IR], one of similar antiquity to AI, but with which it has until now rarely tangled intellectually, although on any broad definition of AI as "modelling intelligent human capacities", one might imagine that IR, like machine translation (MT), would be covered; yet neither has traditionally been seen as part of AI. On second thoughts perhaps, IR does not fall there under that definition simply because, before computers, humans were not in practice able to carry out the kinds of large-scale searches and comparisons operations on which IR rests.

A recent paper written in response to informal and unsubstantiated claims that "surely, IR will now be replaced by LLMs", argues that *IR is not an AI problem*, and should not be studied as one [151]:

> IR is not a subfield of AI, nor a set of tasks to be solved by AI. It is an interdisciplinary space that seeks to understand how technology can be designed to serve ultimately human needs relating to information.

Perhaps the main difference between the two areas is indeed the focus on machines as a tool for humans to deploy, versus the machines as technology that might some day act just like (or, replace or even "superstitute") the human; with *some* researchers even seriously considering AIs that might "go rogue" [112].

Connections between AI and IR do exist. An obvious one is, for example, the advent of Retrieval-Augmented Generation (RAG), combining the strengths of retrieval-based and generative models [78, 142]. Studies on RAG models have emphasized the impact of retrieved document quality, ordering, and even the inclusion of seemingly irrelevant information [33]. As we discuss in Section 3.2.2, work on prompt engineering to improve GenAI answers parallels a long history of work in IR on effective query formulation to boost retrieval effectiveness. More generally, NLP and IR share a long and special history of blurred overlap [76], and perhaps no task has evidenced this more than question answering [161]. To oversimplify, IR was used to retrieve documents while NLP systems were used to extract answers from those documents. While traditional question answering systems typically extracted answers, RAG systems modernize this traditional IR/NLP architecture to search arbitrary sources for desired information and then incorporate it into generated answers.

2.2 IR in the 1990s

The 1990s was a point where search was transformed [144]. At the start of the decade, the IR community was largely split into two sides that had little in common: academic and commercial. On the academic side, IR researchers examined ranking methods mostly driven by variants of Salton et al.'s vector space model [143] with some exploration of probabilistic approaches. Evaluation was based on very small-scale test collections. A number of commercial companies offered online search to current newspaper and commercial data, which was searched through Boolean queries via a command line interface. These systems were difficult to use (the Boolean syntax required specialists, the "search intermediaries", to search on other people's behalf), and costly.

In the first half of the 1990s, thanks to the TREC collections, academic evaluation started to become larger scale and ranking algorithms had to be adjusted to the more diverse set of documents that came with these bigger collections. This led to the emergence of the BM25 ranking function [129]. At a similar time, it became clear that search would be a key way of accessing the emerging World Wide Web. Because of the scale of the Web, Boolean search was simply inadequate; consequently creators of web search engines looked for solutions from academia to build better rankers.

By the second half of the 1990s, it was clear that the web search creators were finding it necessary to develop their own solutions to cope with the scale, diversity, and noise present in web search content. Such solutions were evaluated based on large-scale empirical experiments conducted on the customers of those search engines, providing insights unavailable to the academic community. With access to substantial computing resources and vast query interaction logs, it started to become clear that the commercial web search world was able to obtain insights that the academic community would struggle to achieve.

By the end of the decade, the web search engines had come to dominate commercial search services, most of the subscription services running in the early 1990s had gone out of business. Web search was ubiquitous, fast and free, driven by advertising. The web search companies drew ideas from the academic community, but the academic community needed to find new challenges.

2.3 AI Today

The field of AI has experienced significant transformations. Initially, AI systems relied on search, explicit knowledge representation, predefined rules, and logic to perform tasks [29, 175]. However, the limitations of these approaches, classified under the label of Good Old Fashioned AI (GOFAI) [57], led to the adoption of probabilistic methods capable of handling uncertainty [115–117] and learning from data [80, 153]. This shift, combined with the transition from traditional machine learning to deep learning [65, 77], has allowed the advancement of research across various domains, including computer vision [72], natural language processing [19], healthcare tasks like protein folding prediction [66], and games [102, 155]. The introduction of the transformer architecture in 2017 revolutionized natural language processing, leading to the creation of LLMs capable of understanding and generating human-like text [38, 111, 125], and even mastering complex tasks [20, 123]. Such architecture then expanded and reached state-of-the-art effectiveness in multiple domains: reinforcement learning [30], vision [39], audio processing [124], and even bio-medical applications [158]. Despite their significant impact in different domains, there is ongoing debate regarding the true capabilities of recent models [7, 83, 95–97].

A clear example of such advancements in AI techniques is OpenAI's ChatGPT, which has brought AI into widespread public use and influenced multiple fields [67]. These models have not only demonstrated impressive conversational capabilities (even viewed by some as the first signs of Artificial General Intelligence, AGI [21]) but have also integrated IR components to improve their effectiveness. It can be argued that ChatGPT is for AI today what Google was for IR in the 1990s. In the rest of this paper we analyze this analogy more in detail, starting with discussing the similarities in the next section (see Table 2 in Appendix A for a summary).

3 Similarities

3.1 Benchmark-based Evaluation

3.1.1 Benchmarks and metrics. A key strength of the IR community has been the strong focus on evaluation methodologies. *Offline evaluation* through benchmarks (test collections) has enabled the community to conduct repeatable experiments at scale and has been vital in enabling the ongoing reporting of comparable system effectiveness results. The approach, often referred to as the "Cranfield methodology" [166], has in particular benefitted from key international evaluation fora including TREC [165], NTCIR [141], CLEF [44], and FIRE [48], attracting hundreds of participants annually to evaluate their systems using common evaluation testbeds, and substantial resources are devoted to developing test collections that correspond to different search tasks (e.g., web search, question answering, biomedical search, and so on). *Online evaluation* has provided a complementary approach to measuring system effectiveness for online IR tools with large numbers of users, enabling the deployment of methodologies such as A/B testing [60].

A key feature of ongoing research into the evaluation of IR systems using test collections has been the development of a multitude of different *effectiveness metrics*. In essence, each metric provides a way of distilling a search results list, annotated with relevance judgments, into a measure of how well a search system has performed, based on an underlying *user model* [163] and a notion of task that the user is performing.

In AI, benchmark-based evaluation is similarly used extensively for the evaluation of systems, with the development of a range of different evaluation testbeds that aim to be representative of different "tasks" or domains. Common benchmark examples include ARC-C [32], CRASS [46], HellaSwag [178], MMLU [59], OpenBookQA [90], RACE [73], and SciQ [171]. More recently, "Humanity's Last Exam" has gained attention for its ambitious attempt to assess LLMs on a wide range of human-level cognitive skills [120]. Each of these has an associated set of performance metrics (depending on the task, accuracy, precision, recall or even popular IR metrics like nDCG) to support the comparison of different systems in terms of their effectiveness for a particular testbed (often called "leaderboards" in the AI domain).

3.1.2 Reliability and robustness of benchmarks. A further similarity between the two fields are important questions regarding the reliability and robustness of benchmarks. These issues have been widely discussed in the IR field, leading to a whole sub-area of evaluation research.[1] While similar considerations have been raised in the AI field [37, 88, 177, 179], the speed at which new AI systems are being released (together with the broad popularity of AI tools, including coverage in the press and online) means that these issues are often swept aside in favor of simplistic comparisons of numerical values, often without reference to the many assumptions (and corresponding limitations) that would give a more nuanced understanding of system effectiveness [127]. Relatedly, there are concerns that focusing mainly on aggregate statistics, such as test set accuracy, does not give enough insight into the actual capabilities and robustness of AI systems [51]; similar considerations previously arose in the IR field, and led to extensive research efforts into more nuanced understanding and measurement of IR effectiveness.

3.1.3 Reproducibility of evaluation results for non-public models. Since the 1990s the web search "market" has been dominated by a small number of key players, and these incumbents enjoy a substantial competitive advantage due to the size of their indexes and capacities for serving millions or billions of queries per day. Companies including Google, Microsoft (Bing), Yahoo, Baidu, and Yandex

[1] Among many important considerations, some key examples include: Is the set of search tasks representative? How many search tasks are needed to support generalizability? Are the relevance judges representative of users? Who actually wrote the queries? Are appropriate baselines being chosen for comparison purposes [9, 68]?

engage with the research community to varying degrees (e.g., by publishing research at academic conferences, and through research internship programs). However, ultimately these are commercial entities with an underlying profit objective, and the details of their IR systems (including algorithms, implementation details, and user data) are closely guarded commercial secrets. This of course has substantial implications for research transparency and reproducibility – on the one hand, research using web-scale systems (e.g., using search results produced by a web search engine, or using "real" user data from millions of searchers) can be vital when seeking to understand certain key questions around IR technologies. On the other hand, the systems that lead to this data are "black boxes" from the perspective of academic research, and the possibility of replicating such work is typically small or non-existent. At the same time, there are also a number of open source search systems available, such as Elasticsearch [3], Swirl [4], and Terrier [164], and replicability is not required for reproducibility, which can be obtained also when dealing with proprietary and private systems or datasets [110].

The trajectory of GenAI's explosion in popularity has followed a similar path, with public awareness being associated with a small number of key commercial systems (initially, OpenAI's ChatGPT). Like commercial search engines, popular commercial LLM-based systems are the result of processing data at a vast scale that cannot be reproduced by academic researchers, and the systems themselves incorporate commercially sensitive details, making them "black boxes". Like search engines, there are a range of open source LLMs available; popular examples include Llama [89] and Qwen [12].

3.2 Queries vs. Prompts

3.2.1 Query formulation vs. Prompt engineering. Interacting with LLMs through natural language prompts closely parallels the long-established process of query formulation in IR. Ideally, users would express their needs naturally, but in practice, automated systems often struggle with comprehension, requiring users to refine their inputs iteratively. Initially, naive users may phrase queries conversationally, then adjust them through trial and error to align with the system's expectations. In IR, librarians and automated tools helped translate user intent into effective queries — a role now mirrored in AI by prompt engineering research, which tackles ambiguity, polysemy, and other linguistic challenges, much like earlier efforts in search queries.

An interesting dynamic in user-system interaction is the self-reinforcing cycle between user behavior and system optimization. Initially, users adapt their inputs to match the system's capabilities, and in turn, the system optimizes for these inputs. When system capabilities improve, users often remain unaware, continuing to use outdated input styles, reinforcing the system's existing optimization patterns. Breaking this cycle has required strategies such as introducing new input modalities (e.g., speech), redesigning interfaces to signal enhanced capabilities, and launching new systems free from historical user expectations. The parallel is clear: in IR, inputs are queries; in AI, they are prompts.

3.2.2 Query variation vs. Prompt variation. The impact of input variations has long been a concern in IR, where even slight modifications to a query can yield substantially different retrieval results [14, 71, 181]. Similarly, in GenAI, small changes in prompt design can lead to significant and sometimes unpredictable differences in LLM outputs [58, 99]. As prompting becomes a fundamental mechanism for interacting with LLMs, researchers have increasingly focused on understanding the extent of this sensitivity, challenging the assumption that scaling up model size alone resolves inconsistencies [150]. This growing awareness has led to discussions about the necessity of reporting a range of possible outcomes when evaluating LLMs, rather than relying on a single prompt formulation [99, 150]. In response, various methods for automatic prompt optimization have been proposed, leveraging techniques such as gradient-based tuning and reinforcement learning to refine prompt effectiveness [31, 55, 121]. In parallel, research in IR has examined variations in how users formulate queries and how retrieval models interpret instructions [13, 172]. Evaluation in IR has also advocated for measuring systems across query variations for the same information need [13, 119, 126, 181], akin to recent calls for evaluating LLMs across variations of a prompt.

3.3 Technological Barriers

IR has always been a research area where academics struggled as compared to industry in terms of access to computational resources. In the early days of the Web, it was not imaginable for academic researchers to be able to crawl and index Web-scale datasets. In AI something similar is happening today: academic researchers are often unable to access the scale of hardware and training data required to pre-train large AI models, while well-funded companies often can.

3.4 Ethical, Societal, Legal, Economical Issues

3.4.1 Ethics, social accountability, responsible AI. While there is lots of reflection and pointing out of issues around ethics, accountability and responsibility in both AI and IR, the fields also have clear similarities in that the key players are just doing what they want and racing ahead regardless. The two fields face similar challenges related to the complexity of regulating technology around information. For example it is not clear – and not easy to decide – who is accountable when Google, ChatGPT, or conversational agents give wrong, harmful, or dangerous answers [17]. Recent work by Mitra et al. [98] adopt the Consequences-Mechanisms-Risks (CMR) framework – originally proposed by Gausen et al. [49] to support designers and practitioners of AI – to characterize the socio-technical implications of GenAI in the context of information access and retrieval.

3.4.2 Privacy and copyright issues. When it became clear in the IR community that sizeable query logs are a vital ingredient to advances, the academic community looked towards companies to release their query logs for research purposes. The first (and last) large-scale *raw* query log data release was by AOL in 2006. Three months' worth of users' query logs were made public and quickly removed again as they were found to contain a whole host of private user data [16]. Subsequent releases of industrial query logs remained few and were either completely anonymized (numeric features instead of raw text) or heavily cleaned and sanitized to avoid the publication of any private data. We observe a similar trend when it comes to the release (or better the lack thereof) of industrial pre-training data for state-of-the-art LLMs. Technical reports [40, 160] disclose very little of the pre-training regime beyond general data cleaning principles and a high-level overview of the content distribution; some evidence [154] suggests that copyrighted materials are being used during pre-training, and privacy attacks on LLMs [34] have been shown to be at least somewhat effective to recover an LLM's training data. Moreover, both the IR and AI communities have raised ethical concerns about uncontrolled data use, leaving many open questions about how data owners can prevent (or detect) misuse of their data. It should also be noted that in AI efforts are underway to release pre-training corpora (e.g., the *Common Corpus* [2]) that are in line with established norms and values.

3.4.3 Follow the money. The rise of commercial search engines showed the power of user-based collaborative filtering over traditional content-based approaches, sparking concerns in the IR community. Many feared that academic research would become obsolete without access to large-scale query logs, that top students would leave for high-paying industry jobs, or that funding would disappear. Yet, IR not only survived, but thrived. Researchers reevaluated the strengths of academia and industry, identifying complementary roles. Academia remained essential for long-term, high-risk research and exploration beyond commercial priorities. The diversity of university labs fostered novel ideas, many of which industry later adopted, strengthening technology transfer and ensuring academic relevance. This technology transfer highlights the mutual benefits of industry-academia collaboration. Large-scale, real-world search problems provided academics with intellectually stimulating challenges driven by practical needs. This synergy fostered collaboration through internships, student competitions, faculty engagements (e.g., industry grants, sabbaticals, hybrid roles), and joint research initiatives. When feasible, industry data and API access allowed academics to push the boundaries of industry resources, testing their limits, risks, and capabilities beyond what industry alone could achieve.

The situation in AI is not different. AI investment has experienced substantial growth, particularly in the GenAI sector. In 2023, the sector attracted $25.2 billion, nearly nine times the investment of 2022 and about 30 times the amount from 2019 [81]. In 2021, global private investment in AI totaled around $93.5 billion, more than double the total private investment in 2020 [81]. AI salaries have been on the rise due to high demand and the scarcity of skilled professionals. In 2022, there was more than a 10% increase in wages for AI professionals, with managers seeing the highest levels of increase [5]. There has been a notable increase in the availability of grants for AI research and development. Governments, private organizations, and academic institutions have recognized the potential of AI and are investing heavily in its advancement.

These trends and the delicate balances of money and opportunities, jobs and career, and industry and academia have taught us that we cannot take anything for granted. It is also important to realize that for highly visible, practical, and omnipresent areas such as AI and IR, these things, especially industry and academia, are strongly intertwined (see Section 5).

3.4.4 Open vs. Closed. Since the rise of the Web, IR research has seen the growth of proprietary, closed-source systems as well as a parallel ecosystem of open-source implementations and transparent algorithms. Notably, a lot of systems originally built at internet companies like Google, Twitter, and LinkedIn have been published and/or released open source (e.g., BigTable [27] and Map/Reduce [35]), with such companies being the the first to encounter the challenges around having to manage large amounts of data and data streams. Similarly, in the current AI world we are observing both open and closed approaches where some models are only accessible via an API, while others are released for users to run locally. However, it is still typically not transparent which data has been used for pre-training, except for a few notable examples (e.g., OLMO [53]). The IR experience has shown how industry has selectively released tools and systems with the intent to trigger the academic research and open-source developer communities to focus their work on certain problems and systems. The modern AI industry may learn what the benefits of being more open and transparent are (e.g., the popular use of Meta's LLama models, given their availability). There is a clear need for new information access system architectures [109] and the key IR and AI industry actors have the opportunity to publish their work to feed the academic research community with new problems and challenges to study, to then benefit from the satellite research resulting from their releases and disclosures.

Research studies in IR also needed to focus on understanding the behavior of closed, commercial search systems (see Section 3.1.3). Similarly, there is a current shifting in AI to cognitive science methods as observers of a closed system [52], aiming to understand how a system works internally by observing its external behavior.

3.4.5 Adversarial attacks. Production IR systems have long been subjected to adversarial attacks. The most common attack to an IR system is black-hat Search Engine Optimization (SEO), which consists in the use of unethical, often deceptive techniques designed to manipulate search engine rankings – either to artificially boost a site's own ranking or to harm a competitor's website. These tactics include practices like creating spammy backlinks, inserting hidden keywords, plagiarizing content, or using automated bots to generate fake traffic or content. The IR community has responded to challenges posed by adversarial attacks to IR systems by fostering research on attack and defense methodologies [26], which resulted in the Adversarial IR (AIRWeb) workshop series ran between 2005 and 2009 and the subsequent Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality) initiative than ran until 2015.

GenAI systems, including LLMs, are increasingly becoming the target of adversarial attacks [63, 106, 114, 122, 149]. Drawing a parallel between attacks on search engines and attacks on GenAI systems reveals similar underlying goals and tactics, even though the specific mechanisms differ due to the nature of each technology. Historically, attacks on IR systems focused on keyword stuffing (spamming indexes with low-quality content) and link manipulation to distort rankings. In contrast, attacks on GenAI systems primarily involve data poisoning (injecting misleading training data) and prompt manipulation (tricking models into generating disallowed content). Both exploit manipulated inputs (webpages in IR vs. training data/prompts in GenAI) to degrade system reliability. Other IR attacks, such as cloaking (showing different content to crawlers and users) and sneaky redirects, parallel LLM-specific threats like model evasion/jailbreaking (bypassing content filters) and adversarial examples (crafting inputs to exploit vulnerabilities).

3.5 Philosophical and Conceptual Issues

3.5.1 Reality is messy. Another similarity can be found when analyzing how in both fields, until now, applying clear-cut approaches has only led us so far, and further progress was made after adopting more fuzzy and uncertain methods. In IR, the move from the Boolean model to vector space, probabilistic, latent ones provided a significant increase in effectiveness (Section 2.2). In AI, the classical symbolic approaches of the GOFAI, based on search, knowledge representation, and logical inference, might be adequate for domains that are simple to define like games, theorem proving, artificial language definitions, but they fail to scale up when the complexity of more human and natural worlds enters the scene, in which case subsymbolic approaches dominate the scene today. For example, human language has been mastered not by better grammars but by "playing the game of guessing the next word on huge datasets for an enormous number of times". And indeed in AI today much discussion is happening on hybrid solutions that aim to combine symbolic and subsymbolic approaches.

3.5.2 Terminology. Terminology has been and is an important concept in IR. Not only in the sense that a user has to select the right query terms, but also because the field is dealing with some crucial concepts that are complex and difficult to define. The usage of ambiguous or polysemous terms having multiple meanings or of synonym terms for the same underlying concept can subtract clarity to the research endeavor. For example, "information" is such a term with multiple meanings [11, 50]. Another concrete example is relevance, and in IR we indeed made some progress when we understood that different types of relevance were previously referred to with one term [100, 101, 147].

The situation is similar and terminology is a critical issue, maybe to an even greater extent, in AI; this is perhaps not surprising given the nature of the concepts studied which are at least of the same complexity level (e.g., intelligence vs. information, or commons sense knowledge vs. relevance). McDermott [86], already in 1976, warned AI researchers about the risks of "Wishful mnemonics": simply using human like terms like "think", "understand", or "goal", either as variable/function names in code, or as a description of an AI program, does not mean that the program is really thinking or understanding or having a goal in human-like terms. The discussion is still ongoing today: Mitchell [94] extends the issue to practices in the AI field when using terms like "learning", "neural", etc., and Floridi and Nobre [45] highlight the risks of conceptual borrowing, i.e., anthropomorphizing machines and computerizing minds. On a more concrete level, the term "prompt" is ambiguous as it might refer to instructions only, or it might include user task specifications, or context and evidence.

4 Differences

Although there are many similarities between IR and AI, there are also important differences that we now turn to analyze (see also Appendix A, Table 2).

4.1 Evaluation and Benchmarks

4.1.1 Attention to evaluation metrics. In Section 3.1.1 we considered similarities between IR and AI concerning benchmark-based evaluation and metrics. However, there are also some differences. It can be argued that, on the one side, the evaluation process in IR is more meticulous and precise. For example, test collection are designed on the basis of a specific task with the user in mind, each metric has a user model, statistical significance is a must. On the other hand, the amount of data involved is often higher in AI. For example, the number of test cases (there are more questions in AI benchmarks – thousands – than topics in TREC – 50), the number of datasets used (e.g., BEIR [162] or MTEB [108]), etc., are often higher. The causes for these differences might be the longer evaluation tradition in IR, the size of the fields (discussed below in Section 4.3.1), and maybe the pace of development (see Section 4.1.3). Whatever the cause, these differences can be useful to inform he evaluation practices in the two fields, and to avoid pitfalls (e.g., comparing results based on average scores of whichever metric is currently popular, and moreover without statistical significance being established; or simply aiming to achieve the highest number on some leaderboard without consideration of what particular scenario the testbed was created to represent).

4.1.2 Data contamination in benchmarks. An issue that is particular to the AI domain, and did not exist for "traditional" IR, is the possibility of data contamination in benchmarks [128, 138]: LLMs are typically trained on vast quantities of data, the details of which are unknown (with a few exceptions such as open source models like Olmo [53] and Tulu [74]). The reliable deployment of traditional technical safeguards (such as hiding test sets, having a second hidden test set, only allowing evaluation against a test set every so often, specific markers in the data) is difficult.

4.1.3 Validity of benchmarks over time. Test collections created through collaborative fora such as TREC have been a standard for evaluating IR systems for decades. However, the landscape of AI benchmarking has changed significantly in recent times: when a new AI benchmark is introduced today, industry laboratories quickly adopt it, often within days, to demonstrate the capabilities of their models. Within weeks or months, these benchmarks are frequently "solved", creating a rapidly shifting evaluation environment. This accelerated cycle fundamentally alters the incentives for benchmarking, as the focus shifts from long-term, rigorous evaluation to short-term competitive performance.

4.2 The Importance of Human Factors

IR is distinguished by being an academic discipline that builds tools for people, and the field of study is understood as being wider than the study and construction of a system. "Real world" tasks [152], context [134], interactions [135, 173], and experiences [87, 145] all matter and studies of these are all important to the field. Examples of using a contextual understanding for enhancing user experience include presenting mobile search results based on user location, incorporating user feedback through clicks and dwell time in ranking, and personalizing recommendations over time based on capturing user's implicit preferences. Over the decades, such amalgamation of human-focused studies and system building has

Table 1: Comparison of the size of AI and IR fields.

Year	Number of	SIGIR	IJCAI	NeurIPS	ICLR
2024	Submissions	791	5,651	15,671	7,304
	Participants	957	2,841	19,756	6,533
2023	Submissions	822	4,566	12,343	4,956
	Participants	924	1,988	16,382	3,758
2022	Submissions	793	4,535	10,411	3,328
	Participants	1,024	2,014	15,390	5,346

resulted in some of the most significant advancements in informational systems, including search engines, large-scale recommender systems, and information access through mobile and multimodal devices. There are often two sides of the IR coin that compete and cooperate at the same time. Notably, although there is a rough divide between a "system" and a "user" focus, it is typical in the IR community for the same people to evaluate lower-level properties (e.g., ranking effectiveness) and user experience (e.g., satisfaction). Even when evaluating document ranking, which often operates at a level removed from the human-computer interface, the conventional metrics are either based on explicit user models [28] or have had these models extracted after the fact [104]. It can be argued that such an attention to human factors and the user-system whole is missing in AI, where the user has been almost absent up to now (mostly being looked at by HCI more than AI researchers); however the phenomenon of prompt-based interaction might change that.

4.3 Community

4.3.1 Size of the field. The fields of AI and IR differ notably in their community sizes, as shown in Table 1. Major AI conferences, such as the International Joint Conference on Artificial Intelligence (IJCAI), the Conference on Neural Information Processing Systems (NeurIPS), and the International Conference on Learning Representations (ICLR), attract thousands of paper submissions and attendees annually. In contrast, the SIGIR conference, a leading event in the IR community, typically receives fewer submissions and has a more modest attendance, with approximately a thousand participants each year. This disparity reflects the broader scope and rapid expansion of the field of AI and their communities [118] compared to the more specialized focus of IR.

4.3.2 Conflictuality. When comparing personal relations among researchers within the two fields it can be argued that AI today exhibits a higher conflictuality than IR in the 1990s. The evidence can be anecdotal only, but confirmation could be found for example by reading Marcus's blog [82] (where, analyzing the headings of the posts from January 2025, one can find terms as "terrifying", "demonized", "shambles", "shame", "bullsh*t", "f*ck it"), or following the debate on the consciousness of AI systems [41], that leads to disagreement also among top-level researchers like Hinton [75], Sutskever [157], and Bengio [23]. Even the paternity of the results that led to the recent Nobel prize award is questioned [148].

Such intense and often harsh debates were uncommon in IR during the 1990s, or at least not widely remembered by researchers from that time. Several factors may explain this difference. One possibility is that IR researchers were naturally less prone to strong disagreements. Another is the nature of the topics: discussions around intelligence, consciousness, and AGI in AI inherently invite bold claims and strong opinions, as seen since Dartmouth 1956 [93]. The potential risks associated with AI may also contribute to heightened tensions [112]. Additionally, the sheer size of the AI field makes disputes more likely, whereas IR's division between user- and system-oriented research did not lead to such conflicts. More broadly, societal discourse has become more contentious over time, and even within IR today, conflictuality has increased, with heated debates emerging, such as between Sakai [140] and Fuhr [47] about guidelines for IR evaluation or between Ferrante et al. [42, 43] and Moffat [103] about interval scales in offline evaluation metrics. However, this conflictuality seems based more on scientific disagreement than on more personal aspects.

4.3.3 Publication practices. One difference between the IR and AI communities lies in their publication practices. Traditionally, IR has followed a structured conference and journal-based publication model, emphasizing rigorous peer review and reproducibility [56, 169]. AI, particularly in the era of deep learning and LLMs, has increasingly shifted towards a preprint-dominated ecosystem, with *arXiv* becoming the primary venue for disseminating research [84, 107]. Furthermore, interesting discussions are happening not even on these non-peer reviewed but somehow "scientific" forums, but on blogs by prominent AI researchers (e.g., Mitchell [92] or Marcus [82]) or on social media. This shift presents both advantages and challenges. On the positive side, the open-access nature of *arXiv* has democratized access to the latest research, allowing researchers to receive early feedback and refine their work before formal peer review. However, papers can gain significant exposure in the community even before undergoing peer-assessment. For example, the BERT paper [38] was uploaded to *arXiv* in October 2018 and had already accumulated numerous citations by the time it was officially presented at NAACL-HLT in June 2019. While this rapid dissemination can be beneficial, it also carries risks: the absence of formal peer review may lead to the unchecked spread of unverified claims, misleading results, and overhyped findings, particularly in a fast-moving field like AI [37, 177].

4.3.4 Focused vs. "Inclusive" community. The IR community has historically closely guarded the topics published in its conferences. This has limited the speed of growth of the community when considering the number of submissions or participants, likely impacting sponsorship investments from companies, which are often driven by recruitment strategies. On the other hand, AI conferences such as NeurIPS have been more inclusive of related fields, ideas, and methods; for example, IR papers are routinely accepted [70, 159].

4.4 Increased Attention to Values

4.4.1 Bias and value alignment. In recent years, the integration of ethical principles into the design and deployment of both IR and AI systems has gained unprecedented prominence. This shift marks a departure from earlier eras – particularly the 1990s – when considerations such as bias, fairness, and value alignment were not as prioritized or understood.

Modern AI research increasingly emphasizes the mitigation of biases, the promotion of fairness, and the alignment of systems with societal values [64, 132]. Unlike the earlier focus on performance metrics alone, contemporary studies recognize that even subtle biases can propagate significant inequities when models are deployed at scale. This evolution reflects a growing consensus that the ethical dimensions of AI are as critical to its success as its technical efficacy.

While earlier IR systems exhibited certain biases, (e.g., preferences related to document length or hyperlink structures), these were generally more straightforward to identify and address. Recent investigations, however, have revealed that the biases present in contemporary AI systems are considerably more intricate, e.g., models can inherit and amplify subtle forms of bias, and require more sophisticated detection and mitigation techniques [36].

4.4.2 Explainability and interpretability. The rise of complex, data-driven models has introduced new challenges related to their inherent opacity. Explainability and interpretability – concepts that were not central concerns in the past – have become vital for ensuring accountability in AI systems [10]. As modern models often function as "black boxes," there is a pressing need for methodologies that can elucidate their internal decision-making processes. This drive for transparency not only aids in debugging and improving model performance but also reinforces public trust in AI applications.

4.4.3 Copyright and data ownership. The current discourse in AI also reflects a heightened awareness of data governance issues, including copyright and data ownership [69]. Although comprehensive solutions remain elusive, the level of scrutiny and debate around these topics has increased significantly compared to the 1990s. Researchers now advocate for robust frameworks that address the ethical and legal dimensions of data use, ensuring that AI systems are developed and maintained with a clear respect for intellectual property and individual rights.

4.4.4 Green AI. A growing contingent within the AI community cautions against an uncritical "build and they will come" approach [156]. The substantial energy and environmental costs associated with training and deploying large-scale models compel a more deliberate allocation of resources. This perspective argues for prioritizing research that not only pushes technical boundaries but also addresses pressing societal challenges. The call for responsible innovation underscores the importance of critically evaluating the broader impacts on the environment and society at large.

5 Recommendations

We can outline seven recommendations derived from lessons – some still to be fully realized – learned by the IR community (Figure 1). This presents a starting point to foster a discussion among the IR community to better understand our achievements and missed opportunities, that could be valuable to researchers in AI to build on them while avoiding similar missteps.

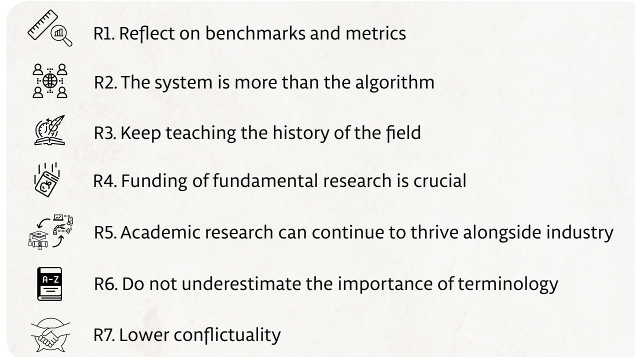

Figure 1: Recommendations from the lessons learned in IR.

R1. Reflect on benchmarks and metrics. The comprehensive body of knowledge on effectiveness metrics for IR, including the study of formal properties of metrics grounded in abstract representations of user behavior (i.e., user models) [8, 24, 105]), has played a crucial role in advancing more robust evaluation frameworks in other fields, such as RecSys [18, 113, 136, 137]. We believe there is significant potential to apply this paradigm to formally characterize the evaluation of GenAI. Furthermore, making the processes and artifacts involved in the creation of publicly available benchmarks – or even embracing the coopetition model [167] and encouraging the development of reusable resources through community-driven efforts such as the evaluation campaigns at TREC, NTCIR, CLEF, FIRE, or SemEval – could accelerate research progress and create new research opportunities. Since benchmarks can present limitations, it is important to continue the discussion on them, as is already done to some extent [15, 91, 127]. The methods and techniques that have been developed in IR [22, 25, 54, 130, 139, 146, 146, 168, 170, 180] can be usefully applied.

R2. The system is more than the algorithm. IR is unusual in being an academic discipline that builds tools for people (at least to a first approximation), but this emphasis on running code is a double-edged sword: lots of ideas have looked good on paper, but have never seen application. In many cases there was no incentive for organizations (industry or academic) to adopt the work; in other cases it turned out not to make a compelling difference to the end product (e.g., to the effectiveness of a full-blown search engine). In the current landscape of AI, there is a tendency to prioritize what technology can achieve next, often at the expense of addressing the actual needs experienced by individuals, communities, and cultures. The lessons here are simple. A balanced approach that integrates technological advancements with a deep understanding of user requirements is essential for meaningful and impactful progress. Even when working on one small slice it is important to understand the end-to-end system, human and machine – for example, that may tell us that a novel interface is a better investment than a marginal improvement in a ranker (or vice versa). We need to understand the final users and uses, so we are addressing problems that matter now or in the foreseeable future. We also need to understand the social and organizational setting where our research might be applied.

For example, as we push for AGI and agentic work, it becomes even more important to understand how humans should, could, and would work with these agents and other AI tools as assistants, collaborators, and mentors. To provide another concrete example on a specific topic, when studying the effect of different prompts, insights from IR on handling query ambiguity and variation could contribute to more effective design and development of LLMs.

R3. Keep teaching the history of the field. It is important to ground the recent trends in the field in their historical context. Teaching the history of the field is not merely an academic exercise, but it provides essential insights into how foundational ideas evolved (e.g., from Boolean retrieval to vector space models to neural IR), why certain methodologies became dominant (e.g., the Cranfield paradigm), and how past limitations continue to shape current research directions. Similarly, in AI, the trend has been from search, problem solving, knowledge representation, expert systems, and logic, through uncertainty and probability, until (deep) neural networks, LLMs, and GenAI. Although GOFAI has failed in reaching human level intelligence, its methods and techniques are often still used in modern AI systems (e.g., AlphaGo Zero exploits Monte Carlo Tree Search [155]). Explicitly addressing the value of the historical aspects (as some researchers are already doing, e.g., [62]) helps students see that today's innovations are part of a continuum, fostering deeper engagement and critical thinking about the field's future. To provide a concrete example, ongoing work to optimize RAG systems can potentially benefit greatly from awareness of the great body of work on questions answering.

R4. Funding of fundamental research is crucial. In addition to investment in research infrastructure, and open-source software [6], we advocate for sustained funding in basic and fundamental research across AI, GenAI, NLP, and LLMs. Public funding has contributed substantially to the advancements in the IR field. Beyond supporting individual and collaborative research projects, initiatives such as TREC or NTCIR, directly funded by NIST (US) and NII (Japan), and even CLEF (indirectly funded by the EU in Europe: EU projects allowed its creation), have enabled researchers to explore new research challenges, including emerging search scenarios and domain-specific applications, novel retrieval models, reusable test collections, and novel evaluation metrics.

R5. Academic research can continue to thrive alongside industry. It is often lamented by AI researchers in academia that only those in the industry have the ability to make a real impact and advancements, much like IR students and scholars in academic settings often feared that all the good IR problems were being addressed by the industry. There are two major problems with this thinking.

First, just as search is not a solved problem since we have Google, industry does not have an exclusive right or hold on AI just because it can afford to have a hefty investment and can attract strong researchers. Several significant issues remain unsolved and are better suited for academia, like the study of aspects related to the users (Section 5) or LLMs' ability for reasoning, which is of utmost importance for the future of agentic AI. These investigations are not at the mercy of computational resources or massive investments.

Second, we need to think about what happens if all the brightest minds get sucked into industry – who is going to educate the next generation, and advocate for public policies? We strongly believe that just as search is not a solved problem, AI is not advanced by only a handful of for-profit companies. In fact, for the sustainable and healthy advancement of AI, it is crucial that we maintain a robust education program tied to academic institutions that free the scholars from exclusively focusing on applied science or contextualizing their research only in commercially beneficial endeavors. We need every generation to have a group of students, scholars, and investigators who keep asking tough questions without for-profit agendas. Without this, we risk getting too narrow and blindsided for the future of AI advancement.

R6. Do not underestimate the importance of terminology. We have seen in Section 3.5.2 as both IR and AI have issues with terminology and how working on it led to progress in IR. One advice that can be derived from this is to work and study the terminology of the field in AI as well. This is already somehow acknowledged by AI researches, e.g., by Mitchell who states "It's clear that to make and assess progress in AI more effectively, we will need to develop a better vocabulary for talking about what machines can do" [94, p. 8]. Progress could be made by distinguishing the various types of intelligence, knowledge, common sense, etc.

R7. Lower conflictuality. We believe the IR community managed to maintain a healthy level of scientific rebuttal (see Section 4.3.2). We acknowledge that this is easy to achieve when the community is smaller. A research environment with low conflictuality and respect among members – even when opposite views are present – is crucial for adapt to changes in the field.

6 Conclusions

We have reflected on the history of the IR field and its community, and related this to the current ongoing shift in the area of AI with the rise of foundation models and GenAI.

On the basis of the previous considerations, we can also take an introspective look at our discipline and community. Despite the increasing impact of search engines and information access systems in society, the IR community has struggled to effectively engage with the broader academic community, policymakers, and practitioners. We identify four primary factors contributing to this phenomenon: (i) the name of the field, "information retrieval" does not intuitively convey the breadth and depth of IR research; (ii) the high interdisciplinary nature of IR, while enriching our research, also complicates efforts to clearly define and communicate IR's distinct contributions; (iii) the rigorous research methodologies developed in IR are sometimes a barrier of entry to the field; and (iv) not enough effort is needed to ensure that the knowledge developed in the IR field reaches a wider audience (whereas the AI community seems to be paying more attention to these issues as, for example, several popular science books on AI have been and are being published, e.g., [93, 132, 176]).

This paper is a call for action for the IR community at large: besides the need to spread the messages that will make the contributions of our field more recognized and accepted, we also need to discuss among ourselves which is the most effective way to do so.

Acknowledgments

This work is partially supported by an Australian Research Council (ARC) Future Fellowship Project (Grant No. FT240100022), by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005), by the US National Science Foundation (NSF) CPS-2438935 grant, and by Good Systems (http://goodsystems.utexas.edu/), a UT Austin Grand Challenge to develop responsible AI technologies. Any opinions and recommendations expressed in this paper are those of the authors and do not necessarily reflect those of their affiliated institutions or sponsors.

References

[1] [n. d.]. Artificial intelligence – Wikipedia. https://en.wikipedia.org/wiki/Artificial_intelligence. Accessed: 17 Feb 2025.

[2] [n. d.]. Common Corpus – Hugging Face. https://huggingface.co/datasets/PleIAs/common_corpus. Accessed: 17 Feb 2025.

[3] [n. d.]. Elasticsearch. https://www.elastic.co/elasticsearch. Accessed: 17 Feb 2025.

[4] [n. d.]. Swirl AI Search. https://swirlaiconnect.com/. Accessed: 17 Feb 2025.

[5] AI Times Journal. 2023. AI Employment Outlook: Salary Trends and Future Projections. https://www.aitimejournal.com/ai-employment-outlook-salary-trends-and-future-projections/46785/. Accessed: 17 Feb 2025.

[6] James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. 2024. *Future of Information Retrieval Research in the Age of Generative AI.* Technical Report. https://cra.org/wp-content/uploads/2024/12/Future-of-Information-Retrieval-Research-in-the-Age-of-Generative-AI.pdf

[7] Omar Alonso and Kenneth Church. 2024. Evaluating the Evaluations: A Perspective on Benchmarks. *ACM SIGIR Forum* 58, 2 (2024). https://sigir.org/wp-content/uploads/2024/02/p18.pdf Opinion Paper.

[8] Enrique Amigó, Hui Fang, Stefano Mizzaro, and Chengxiang Zhai. 2020. Axiomatic Thinking for Information Retrieval: Introduction to Special Issue. *Information Retrieval Journal* 23, 3 (June 2020), 187–190. doi:10.1007/s10791-020-09376-y

[9] Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management.* 601–610.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Sanel Tabik, Alberto Barbado, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115.

[11] Antonio Badia. 2019. *The Information Manifold: Why Computers Can't Solve Algorithmic Bias and Fake News.* MIT Press, Cambridge, MA. xvii + 334 pages.

[12] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[13] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16).* Association for Computing Machinery, New York, NY, USA, 725–728. doi:10.1145/2911451.2914671

[14] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17).* Association for Computing Machinery, New York, NY, USA, 395–404. doi:10.1145/3077136.3080839

[15] Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? arXiv. doi:10.48550/arXiv.2412.03597

[16] Michael Barbaro and Tom Zeller Jr. 2006. A Face is Exposed for AOL Searcher No. 4417749. https://www.nytimes.com/2006/08/09/technology/09aol.html.

[17] BBC. 2021. Alexa Tells 10-year-old Girl to Touch Live Plug With Penny. https://www.bbc.com/news/technology-59810383. Accessed: 17 Feb 2025.

[18] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. 2011. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proceedings of the fifth ACM conference on Recommender systems.* 333–336.

[19] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3, null (March 2003), 1137–1155.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20).* Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[21] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] https://arxiv.org/abs/2303.12712

[22] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00).* Association for Computing Machinery, New York, NY, USA, 33–40. doi:10.1145/345508.345543

[23] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708 [cs.AI] https://arxiv.org/abs/2308.08708

[24] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11).* Association for Computing Machinery, New York, NY, USA, 903–912. doi:10.1145/2009916.2010037

[25] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06).* Association for Computing Machinery, New York, NY, USA, 268–275. doi:10.1145/1148170.1148219

[26] Carlos Castillo, Brian D Davison, et al. 2011. Adversarial web search. *Foundations and trends® in information retrieval* 4, 5 (2011), 377–486.

[27] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. 2008. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)* 26, 2 (2008), 1–26.

[28] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management.* 621–630.

[29] Eugene Charniak. 1985. *Introduction to artificial intelligence.* Pearson Education India.

[30] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[31] Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518* (2023).

[32] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv. doi:10.48550/arXiv.1803.05457 arXiv:1803.05457 [cs.AI]

[33] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 719–729.

[34] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* 57, 6, Article 152 (Feb. 2025), 39 pages. doi:10.1145/3712001

[35] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.

[36] Gianluca Demartini and Stefan Siersdorfer. 2010. Dear search engine: what's your opinion about...? sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd international semantic search workshop.* 1–7.

[37] Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2017. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one* 12, 9 (2017), e0184604.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[40] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[41] Michele Farisco, Kathinka Evers, and Jean-Pierre Changeux. 2024. Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks* 180 (2024), 106714. doi:10.1016/j.neunet.2024.106714

[42] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales. *IEEE Access* 9 (2021), 136182–136216. doi:10.1109/ACCESS.2021.3116857

[43] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2022. Response to Moffat's Comment on "Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales". arXiv:2212.11735 [cs.IR] https://arxiv.org/abs/2212.11735

[44] Nicola Ferro and Carol Peters. 2019. *Information Retrieval Evaluation in a Changing World Lessons Learned from 20 Years of CLEF*. Springer.

[45] Luciano Floridi and Anna C Nobre. 2024. Anthropomorphising machines and computerising minds: the crosswiring of languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines* 34, 1 (2024), 1–9.

[46] Jörg Frohberg and Frank Binder. 2022. CRASS: A Novel Data Set and Benchmark to Test Counterfactual Reasoning of Large Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2126–2140. https://aclanthology.org/2022.lrec-1.229/

[47] Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (Feb. 2018), 32–41. doi:10.1145/3190580.3190586

[48] Debasis Ganguly, Srijoni Majumdar, Bhaskar Mitra, Parth Gupta, Surupendu Gangopadhyay, and Prasenjit Majumder (Eds.). 2023. *FIRE '23: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation* (Panjim, India). Association for Computing Machinery, New York, NY, USA.

[49] Anna Gausen, Bhaskar Mitra, and Siân Lindley. 2024. A Framework for Exploring the Consequences of AI-Mediated Enterprise Knowledge Access and Identifying Risks to Workers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 207–220. doi:10.1145/3630106.3658900

[50] James Gleick. 2011. *The Information: A History, a Theory, a Flood*. Pantheon.

[51] Micah Goldblum, Anima Anandkumar, Richard Baraniuk, Tom Goldstein, Kyunghyun Cho, Zachary C Lipton, Melanie Mitchell, Preetum Nakkiran, Max Welling, and Andrew Gordon Wilson. 2023. Perspectives on the State and Future of Deep Learning - 2023. arXiv:2312.09323 [cs.AI] https://arxiv.org/abs/2312.09323

[52] Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. 2024. Bayes in the age of intelligent machines. *Current Directions in Psychological Science* 33, 5 (2024), 283–291.

[53] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838* (2024).

[54] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.* 27, 4, Article 21 (Nov. 2009), 26 pages. doi:10.1145/1629096.1629099

[55] Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532* (2023).

[56] Donna Harman. 2019. Information Retrieval: The Early Years. *Foundations and Trends® in Information Retrieval* 13, 5 (2019), 425–577. doi:10.1561/1500000065

[57] John Haugeland. 1989. *Artificial intelligence: The very idea*. MIT press.

[58] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does Prompt Formatting Have Any Impact on LLM Performance? *arXiv preprint arXiv:2411.10541* (2024).

[59] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*. OpenReview.net, Online, 1–27. https://openreview.net/pdf?id=d7KBjmI3GmQ

[60] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and Trends® in Information Retrieval* 10, 1 (2016), 1–117. doi:10.1561/1500000051

[61] Eric Horvitz, Vincent Conitzer, Sheila McIlraith, and Peter Stone. 2024. Now, Later, and Lasting: 10 Priorities for AI Research, Policy, and Practice. *Commun. ACM* 67, 6 (May 2024), 39–40. doi:10.1145/3637866

[62] Eric Horvitz and Tom M. Mitchell. 2024. Scientific Progress in Artificial Intelligence: History, Status, and Futures. In *Realizing the Promise and Minimizing the Perils of AI for Science and the Scientific Community*, Kathleen Hall Jamieson, Anne-Marie Mazza, and William Kearney (Eds.). University of Pennsylvania Press.

[63] Chip Huyen. 2024. *AI Engineering: Building Applications with Foundation Models*. O'Reilly Media.

[64] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[65] M. I. Jordan and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. doi:10.1126/science.aaa8415 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa8415

[66] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

[67] Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat GPT and its impact on different fields of study. *International journal of innovative science and research technology* 8, 3 (2023).

[68] Sadegh Kharazmi, Falk Scholer, David Vallet, and Mark Sanderson. 2016. Examining additivity and weak baselines. *ACM Transactions on Information Systems (TOIS)* 34, 4 (2016), 1–18.

[69] Vijay Khatri and Carol V. Brown. 2010. Designing data governance. *Commun. ACM* 53, 1 (Jan. 2010), 148–152. doi:10.1145/1629175.1629210

[70] Jaehee Kim, Yukyung Lee, and Pilsung Kang. 2024. A Gradient Accumulation Method for Dense Retriever under Memory Constraint. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 11765–11788. https://proceedings.neurips.cc/paper_files/paper/2024/file/15ba84c1e19b0eb75f96922f5da0a021-Paper-Conference.pdf

[71] Bevan Koopman, Guido Zuccon, and Peter Bruza. 2017. What makes an effective clinical query and querier? *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2557–2571.

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. doi:10.1145/3065386

[73] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 785–794. doi:10.18653/v1/D17-1082

[74] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124* (2024).

[75] LBC. 2025. 'Godfather of AI' predicts it will take over the world. https://youtu.be/vxkBE23zDmQ. Accessed: 17 Feb 2025.

[76] Matthew Lease. 2007. Natural Language Processing for Information Retrieval: the time is ripe (again). In *Proceedings of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM)*. Best Paper award..

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[78] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[79] Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. 2021. *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*. Technical Report. Stanford University, Stanford, CA. http://ai100.stanford.edu/2021-report

[80] Yang Lu. 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics* 6, 1 (2019), 1–29. doi:10.1080/23270012.2019.1570365

[81] Shana Lynch. 2024. AI Index: State of AI in 13 Charts. https://hai.stanford.edu/news/ai-index-state-ai-13-charts. Accessed: 17 Feb 2025.

[82] Gary Marcus. [n. d.]. Marcus on AI. https://garymarcus.substack.com. Accessed: 17 Feb 2025.

[83] Gary Marcus. 2018. Deep Learning: A Critical Appraisal. *CoRR* abs/1801.00631 (2018). arXiv:1801.00631 http://arxiv.org/abs/1801.00631

[84] Fernando Martínez-Plumed, Pablo Barredo, Sean O Heigeartaigh, and Jose Hernandez-Orallo. 2021. Research community dynamics behind popular AI benchmarks. *Nature Machine Intelligence* 3, 7 (2021), 581–589.

[85] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. *The AI Index 2025 Annual Report*. Technical Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

[86] Drew McDermott. 1976. Artificial intelligence meets natural stupidity. *Acm Sigart Bulletin* 57 (1976), 4–9.

[87] Daniel McDuff, Paul Thomas, Nick Craswell, Kael Rowan, and Mary Czerwinski. 2021. Do Affective Cues Validate Behavioural Metrics for Search?. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1544–1553. doi:10.1145/3404835.3462894

[88] Nick McGreivy and Ammar Hakim. 2024. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence* 6, 10 (2024), 1256–1269.

[89] Meta. [n. d.]. Llama AI Model. https://www.llama.com/. Accessed: 17 Feb 2025.

[90] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2381–2391. doi:10.18653/v1/D18-1260

[91] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv:2410.05229 [cs.LG] https://arxiv.org/abs/2410.05229

[92] Melanie Mitchell. [n. d.]. AI: A Guide for Thinking Humans. https://aiguide.substack.com. Accessed: 17 Feb 2025.

[93] M. Mitchell. 2020. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.

[94] Melanie Mitchell. 2021. Why AI is harder than we think. *arXiv preprint arXiv:2104.12871* (2021).

[95] Melanie Mitchell. 2023. AI's challenge of understanding the world. *Science* 382, 6671 (2023), eadm8175. doi:10.1126/science.adm8175 arXiv:https://www.science.org/doi/pdf/10.1126/science.adm8175

[96] Melanie Mitchell. 2023. How do we know how smart AI systems are? *Science* 381, 6654 (2023), eadj5957. doi:10.1126/science.adj5957 arXiv:https://www.science.org/doi/pdf/10.1126/science.adj5957

[97] Melanie Mitchell. 2024. Debates on the nature of artificial general intelligence. *Science* 383, 6689 (2024), eado7069. doi:10.1126/science.ado7069 arXiv:https://www.science.org/doi/pdf/10.1126/science.ado7069

[98] Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. 2025. *Sociotechnical Implications of Generative Artificial Intelligence for Information Access*. Springer Nature Switzerland, Cham, 161–200. doi:10.1007/978-3-031-73147-1_7

[99] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics* 12 (08 2024), 933–949. doi:10.1162/tacl_a_00681 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00681/2464098/tacl_a_00681.pdf

[100] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American society for information science* 48, 9 (1997), 810–832.

[101] Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interact. Comput.* 10, 3 (1998), 303–320. doi:10.1016/S0953-5438(98)00012-5

[102] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[103] Alistair Moffat. 2022. Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning. *IEEE Access* 10 (2022), 105564–105577. doi:10.1109/ACCESS.2022.3211668

[104] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A flexible framework for offline effectiveness metrics. In *SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 578–587.

[105] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: what observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 659–668. doi:10.1145/2505515.2507665

[106] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12448–12460.

[107] Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2024. Topics, Authors, and Institutions in Large Language Model Research: Trends from 17K arXiv Papers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1223–1243.

[108] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. arXiv:2210.07316 [cs.CL] https://arxiv.org/abs/2210.07316

[109] Marc Najork. 2023. Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1. doi:10.1145/3539618.3591871

[110] National Academies of Sciences, Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics, Board on Mathematical Sciences, et al. 2019. *Reproducibility and replicability in science*. National Academies Press.

[111] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[112] Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. 2024. Frontier AI systems have surpassed the self-replicating red line. arXiv:2412.12140 [cs.CL] https://arxiv.org/abs/2412.12140

[113] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 75–84. doi:10.1145/3460231.3474234

[114] Advait Patel, Pravin Pandey, Hariharan Ragothaman, Ramasankar Molleti, and Ajay Tanikonda. 2025. Securing Cloud AI Workloads: Protecting Generative AI Models from Adversarial Attacks. In *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*. 1–7. doi:10.1109/ICAIC63015.2025.10848877

[115] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA.

[116] Judea Pearl. 1993. From Conditional Oughts to Qualitative Decision Theory. In *Uncertainty in Artificial Intelligence*, David Heckerman and Abe Mamdani (Eds.). Morgan Kaufmann, 12–20. doi:10.1016/B978-1-4832-1451-1.50006-8

[117] Judea Pearl. 1994. A Probabilistic Calculus of Actions. In *Uncertainty in Artificial Intelligence*, Ramon Lopez de Mantaras and David Poole (Eds.). Morgan Kaufmann, San Francisco (CA), 454–462. doi:10.1016/B978-1-55860-332-5.50062-6

[118] Andi Peng, Jessica Zosa Forde, Yonadav Shavit, and Jonathan Frankle. 2022. Strengthening subcommunities: Towards sustainable growth in ai research. *arXiv preprint arXiv:2204.08377* (2022).

[119] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *European conference on information retrieval*. Springer, 397–412.

[120] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. 2025. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249* (2025).

[121] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495* (2023).

[122] P Radanliev and O Santos. 2023. Adversarial attacks can deceive AI systems, leading to misclassification or incorrect decisions.

[123] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to Generate Reviews and Discovering Sentiment. (2017). arXiv:1704.01444 [cs.LG] https://arxiv.org/abs/1704.01444

[124] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[125] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training*. Technical Report. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[126] Lida Rashidi, Justin Zobel, and Alistair Moffat. 2024. Query Variability and Experimental Consistency: A Concerning Case Study. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*. 35–41.

[127] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. arXiv:2411.12990 [cs.AI] https://arxiv.org/abs/2411.12990

[128] Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*.

[129] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[130] Kevin Roitero, J. Shane Culpepper, Mark Sanderson, Falk Scholer, and Stefano Mizzaro. 2020. Fewer Topics? A Million Topics? Both?! On Topics Subsets in Test Collections. *Information Retrieval Journal* 23, 1 (2020), 49–85. doi:10.1007/s10791-019-09357-w

[131] Francesca Rossi, Christian Bessiere, Joydeep Biswas, Rodney Brooks Vincent Conitzer, Thomas G. Dietterich, Virginia Dignum, Oren Etzioni, Kenneth D. Forbus, Eugene Freuder, Yolanda Gil, Holger Hoos, Eric Horvitz, Subbarao Kambhampati, Henry Kautz, Jihie Kim, Hiroaki Kitano, Alan Mackworth, Karen Myers, Luc De Raedt, Stuart Russell, Bart Selman, Peter Stone, Millind Tambe, and Michael Wooldridge. 2025. AAAI 2025 Presidential Panel on the Future of AI Research. https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-Digital-3.7.25.pdf

[132] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Allen Lane, London.

[133] Stuart Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach (4th Edition).* Pearson. http://aima.cs.berkeley.edu/

[134] Ian Ruthven. 2011. Information retrieval in context. In *Advanced topics in information retrieval*, Massimo Melucci and Ricardo Beaza-Yates (Eds.). Springer, Berlin.

[135] Ian Ruthven and Diane Kelly (Eds.). 2011. *Interactive information seeking, behaviour, and retrieval.* Facet Publishing, London.

[136] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems.* 129–136.

[137] Alan Said and Alejandro Bellogín. 2015. Replicable evaluation of recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems.* 363–364.

[138] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018* (2023).

[139] Tetsuya Sakai. 2018. Topic Set Size Design for Paired and Unpaired Data. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (Tianjin, China) *(ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 199–202. doi:10.1145/3234944.3234971

[140] Tetsuya Sakai. 2021. On Fuhr's guideline for IR evaluation. *SIGIR Forum* 54, 1, Article 12 (Feb. 2021), 8 pages. doi:10.1145/3451964.3451976

[141] Tetsuya Sakai, Douglas W Oard, and Noriko Kando. 2021. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact.* Springer Nature.

[142] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2395–2400. doi:10.1145/3626772.3657957

[143] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (Nov. 1975), 613–620. doi:10.1145/361219.361220

[144] Mark Sanderson and W. Bruce Croft. 2012. The History of Information Retrieval Research. *Proc. IEEE* 100, Special Centennial Issue (May 2012), 1444–1451. doi:10.1109/JPROC.2012.2189916

[145] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 555–562. doi:10.1145/1835449.1835542

[146] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 162–169. doi:10.1145/1076034.1076064

[147] Tefko Saracevic. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.* 26, 6 (1975), 321–343. doi:10.1002/ASI.4630260604

[148] Jürgen Schmidhuber. 2020. Critique of Honda Prize for Dr. Hinton. https://people.idsia.ch/~juergen/critique-honda-prize-hinton.html. Accessed: 17 Feb 2025.

[149] Leo Schwinn, David Dobre, Stephan Günnemann, and Gauthier Gidel. 2023. Adversarial attacks and defenses in large language models: Old and new threats. In *Proceedings on.* PMLR, 103–117.

[150] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324* (2023).

[151] Chirag Shah and Emily M Bender. 2024. Envisioning information access systems: What makes for good tools and a healthy Web? *ACM Transactions on the Web* 18, 3 (2024), 1–24.

[152] Chirag Shah and Ryen W White. 2024. Report on the 2nd Workshop on Task-Focused IR in the Era of Generative AI. *SIGIR Forum* 58, 2 (2024).

[153] Zhou Shao, Ruoyan Zhao, Sha Yuan, Ming Ding, and Yongli Wang. 2022. Tracing the evolution of AI in the past decade and forecasting the emerging trends. *Expert Systems with Applications* 209 (2022), 118221. doi:10.1016/j.eswa.2022.118221

[154] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. arXiv:2310.16789 [cs.CL] https://arxiv.org/abs/2310.16789

[155] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550 (2017), 354–359. https://api.semanticscholar.org/CorpusID:205261034

[156] Jack Stilgoe, Richard Owen, and Paul Macnaghten. 2013. Developing a Framework for Responsible Innovation. *Research Policy* 42, 9 (2013), 1568–1580.

[157] Ilya Sutskever. 2022. Ilya Sutskever on Twitter/X. https://x.com/ilyasut/status/1491554478243258368. Accessed: 17 Feb 2025.

[158] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* 26, 5 (2023), 858–866.

[159] Qiaoyu Tang, Jiawei Chen, Zhuoqun Li, Bowen Yu, Yaojie Lu, , Haiyang Yu, Hongyu Lin, Fei Huang, Ben He, Xianpei Han, Le Sun, and Yongbin Li. 2024. Self-Retrieval: End-to-End Information Retrieval with One Large Language Model. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 63510–63533. https://proceedings.neurips.cc/paper_files/paper/2024/file/741ad162ab0f3da6f9aad60e9e34f5f1-Paper-Conference.pdf

[160] Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[161] Text REtreival Conferencer (TREC). 1999. Question Answering Collection. online. https://trec.nist.gov/data/qa.html.

[162] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *CoRR* abs/2104.08663 (2021). arXiv:2104.08663 https://arxiv.org/abs/2104.08663

[163] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling decision points in user search behavior. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany) *(IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 239–242. doi:10.1145/2637002.2637032

[164] University of Galsgow. [n. d.]. Terrier IR Platform. http://terrier.org/. Accessed: 17 Feb 2025.

[165] Ellen Voorhees and Donna Harman. 2005. *TREC: Experiment and evaluation in information retrieval.* MIT Press.

[166] Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 355–370.

[167] Ellen M. Voorhees. 2020. Coopetition in IR Research. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 3. doi:10.1145/3397271.3402427

[168] Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 316–323. doi:10.1145/564376.564432

[169] Ellen M Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *arXiv preprint arXiv:2201.11086* (2022).

[170] William Webber, Alistair Moffat, and Justin Zobel. 2008. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 571–580. doi:10.1145/1458082.1458158

[171] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 94–106. doi:10.18653/v1/W17-4413

[172] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions. *arXiv preprint arXiv:2403.15246* (2024).

[173] Ryen W White. 2016. *Interactions with search systems.* Cambridge University Press.

[174] Yorick A. Wilks. 2005. *Unhappy Bedfellows: The Relationship of AI and IR.* Springer Netherlands, Dordrecht, 255–282. doi:10.1007/1-4020-3467-9_14

[175] Patrick Henry Winston. 1992. *Artificial intelligence.* Addison-Wesley Longman Publishing Co., Inc.

[176] M. Wooldridge. 2022. *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going.* Flatiron Books. https://books.google.com.au/books?id=hjctEAAAQBAJ

[177] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1129–1132. doi:10.1145/3331184.3331340
[178] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4791–4800. doi:10.18653/v1/P19-1472
[179] Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. *arXiv preprint arXiv:2305.17442* (2023).
[180] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 307–314. doi:10.1145/290941.291014
[181] Guido Zuccon, Joao Palotti, and Allan Hanbury. 2016. Query Variations and their Effect on Comparing Information Retrieval Systems. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 691–700. doi:10.1145/2983323.2983723

A AI and IR: Similarities and Differences

Table 2: Similarities (Section 3) and differences (Section 4) between AI and IR.

Similarities	
Benchmark-based Evaluation	Benchmarks and metrics Reliability and robustness of benchmarks Reproducibility of evaluation results for non-public models
Queries vs. Prompts	Query formulation vs. prompt engineering Query variation vs. prompt variation
Technological Barriers	
Ethical, Societal, Legal, and Economical Issues	Ethics, social accountability, responsible AI Privacy and copyright issues Follow the money Open vs. Closed Adversarial attacks
Philosophical and Conceptual Issues	Reality is messy Terminology
Differences	
Evaluation and Benchmarks	Attention to evaluation metrics Data contamination in benchmarks Validity of benchmarks over time
The Importance of Human Factors	
Community	Size of the field Conflictuality Publication practices Focused vs. "Inclusive" community
Increased Attention to Values	Bias and value alignment Explainability and interpretability Copyright and data ownership Green AI